SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 1–25

A SPLIT GODUNOV SCHEME FOR SOLVING ONE-DIMENSIONAL HYPERBOLIC SYSTEMS IN A NONCONSERVATIVE FORM*

GISÈLE MOPHOU† AND PASCAL POULLET†

Abstract. In this paper, we developed a theoretical study for nonconservative sytems in one dimension in order to construct numerical schemes for solving the Riemann problem. The nonconservative form of our model system required the use of a well-adapted theory in order to give us a sense of our problem. We chose a framework of generalized functions for solving a scalar hyperbolic equation with a discontinuous coefficient $\sigma_t + u\sigma_x \approx 0$, where u is the velocity solution of a Burgers's equation. After an explicit solution of the Riemann problem, we derived Godunov split schemes for computing an approximate solution of the Cauchy problem. We applied our study to a system modeling elasticity and a system modeling gas dynamics. Some stability properties of a scheme and its convergence to a generalized solution are proved for the first model. Numerical experiments confirmed this convergence result. For the second model, calculations of flows containing weak-to-moderate shocks showed that conservation errors are reduced when the mesh is refined but were not entirely eliminated.

Key words. Riemann solver, Burgers's equation, splitting method, Godunov scheme, nonconservative system

AMS subject classifications. 46F10, 65M06, 76L05, 73C50, 35Q99

PII. S0036142900378637

1. Introduction.

For presenting our study, we first consider the simplified system modeling elasticity in Eulerian coordinates [5, 2]:

$$\left\{ \begin{array}{rcl} \partial_t u(x,t) + u(x,t)\partial_x u(x,t) - \partial_x \sigma(x,t) & \approx & 0, \\ \partial_t \sigma(x,t) + u(x,t)\partial_x \sigma(x,t) - k^2 \partial_x u(x,t) & \approx & 0, \end{array} \right. \tag{1.1}$$

which is a dynamic equation coupled with Hooke's law when the density of the solid is almost equal to one, with u representing the velocity, σ the stress, and k a positive fixed real number.

It is well known that such systems present numerical and physical solutions u and σ which are simultaneously discontinuous functions. Therefore, the term $u\,.\,\partial_x\sigma$ appears in the form of the meaningless product $Y.\delta$ of the Heaviside and the Dirac functions. To remove this ambiguity, J. F. Colombeau developed a mathematical theory by introducing the differential algebra $\mathcal{G}(\mathbb{R}\times\mathbb{R}^+)$ of generalized functions where discontinuous functions are represented by means of classes of $\mathcal{C}^\infty$ functions [4, 2, 5].

Although the Riemann problem (1.1)–(1.2) admits an infinite number of jump conditions for shocks [7], we are interested in seeking its shock solution when the velocity and the stress are written with the same Heaviside function. The decomposition of the Riemann problem (1.1) with the initial data of the form

$$(u_0(x),\, \sigma_0(x)) = \left\{ \begin{array}{ll} (u_l,\, \sigma_l) & \text{if} \quad x < 0, \\ (u_r,\, \sigma_r) & \text{if} \quad x > 0, \end{array} \right. \tag{1.2}$$

*Received by the editors September 22, 2000; accepted for publication (in revised form) November 13, 2001; published electronically April 12, 2002.
http://www.siam.org/journals/sinum/40-1/37863.html

†Laboratoire de Mathématiques, Université des Antilles et de la Guyane, Campus de Fouillole, F-97159 Pointe-à-Pitre Cedex, Guadeloupe, French West Indies (Gisele.Mophou@univ-ag.fr, Pascal.Poullet@univ-ag.fr).

achieved by means of a splitting technique [1, 8], permits us to obtain two other Riemann problems which are easier to solve. One of them (propagation terms) is associated with the following linear hyperbolic system:

$$\left\{ \begin{array}{rcl} \partial_t u(x,t) - \partial_x \sigma(x,t) & \approx & 0, \\ \partial_t \sigma(x,t) - k^2 \partial_x u(x,t) & \approx & 0. \end{array} \right. \tag{1.3}$$

The other one is linked to a nonlinear, nonstrictly hyperbolic system (convection terms):

$$\left\{ \begin{array}{rcl} \partial_t u(x,t) + u(x,t)\partial_x u(x,t) & \approx & 0, \\ \partial_t \sigma(x,t) + u(x,t)\partial_x \sigma(x,t) & \approx & 0. \end{array} \right. \tag{1.4}$$

Due to the nonstrictly hyperbolicity of the system (1.4), many authors have already attempted to solve the associated Riemann problem. M. Oberguggenberger in [11] made a theoretical study of the same system by simultaneously solving the two equations. In [6] Colombeau and Le Roux proposed a Godunov scheme for this system by assuming that the variables u and σ moved in phase on the shock. But the case of decreasing initial data when the velocity u has a variable sign was not considered.

Recently, in [13], Remaki studied hyperbolic equations with a discontinuous coefficient depending on the space variable x, which permits us to study the second equation in (1.4), u being known from the first equation. Moreover, a technique of perturbation for solving a degenerated system in conservative formulation has also been introduced by Le Roux and collaborators in [1].

In this paper, we develop two methods for studying the Riemann problem associated with the system of convection terms (1.4). The first one is the approach by perturbation mentioned above. It consists of viewing the system (1.4) as the limit when $\varepsilon \to 0,\ \varepsilon > 0$ of the following well-posed system:

$$\left\{ \begin{array}{rcl} \partial_t u(x,t) + u(x,t)\partial_x u(x,t) - \varepsilon\, \partial_x \sigma(x,t) & = & 0, \\ \partial_t \sigma(x,t) + u(x,t)\partial_x \sigma(x,t) - \varepsilon\, k^2 \partial_x u(x,t) & = & 0. \end{array} \right. \tag{1.5}$$

Proceeding in this way, one shows that its associated Riemann problem presents a shock wave solution when $u_l \geq u_r$; otherwise, a rarefaction wave solution is obtained. Moreover, the Riemann invariants for the systems of convection and propagation terms are the same as those obtained for the original system (1.1).

The second approach consists of substituting this velocity u, which is a solution of the inviscid Burgers's equation, in the second equation of (1.4). Then we study the second equation of the system as a scalar hyperbolic equation with a discontinuous coefficient depending on the variables x and t. Therefore, when we seek the discontinuous solutions of the latter equation, we obtain for $u_l \geq u_r$ a single shock solution and a two shock solution for $u_l < u_r$. (This last equation had not been considered by Colombeau and Le Roux [6].)

Following this analysis, we propose new Godunov schemes to compute an approximate solution of problem (1.4)–(1.2). Then by a splitting method, we construct new Godunov splitted schemes for the system (1.1). Adapting the technique mentioned in [3] for the scheme obtained by the first approach, we prove the stability of the L^∞-norm for the total variation in space and in time (in a Tonnelli–Cesari's sense) and that this scheme is convergent to a generalized solution.

Finally, we consider the gas dynamics system in the nonconservative form [6, 5]:

$$\left\{ \begin{array}{rcl} \partial_t v(x,t) + u(x,t)\partial_x v(x,t) - v(x,t)\partial_x u(x,t) & \approx & 0, \\ \partial_t u(x,t) + u(x,t)\partial_x u(x,t) + v(x,t)\partial_x p(x,t) & \approx & 0, \\ \partial_t p(x,t) + u(x,t)\partial_x p(x,t) + \gamma p(x,t)\partial_x u(x,t) & \approx & 0, \end{array} \right. \tag{1.6}$$

where u, p, and v denote, respectively, the velocity, the pressure, and specific volume.

By the same approach as mentioned above for solving the system of convection terms, we propose split Godunov schemes using the Riemann solver developed by Colombeau–Le Roux [6] for the propagation terms.

It has been proved in [5] that the approximate solution of system (1.6) can cause wrong variations of specific volume, velocity, and pressure. Therefore, we use the modified Godunov scheme for the system of propagation terms which consists of projecting the density, the momentum, and the total energy instead of the specific volume, the velocity, and the pressure.

Consequently, for decreasing initial data where $u_l > 0 > u_r$, our new schemes permit us to compute an acceptable approximate solution either for the elasticity problem or for the gas dynamics Riemann problem, while the Colombeau–Le Roux scheme fails.

Section 2 is devoted to presenting the theoretical and numerical studies of the Riemann problem for the degenerated hyperbolic system (1.4) with the perturbation and the substitution methods. Split schemes for the system modeling elasticity (1.1) are proposed, and the stability properties of the splitted scheme of elasticity obtained with the first approach are proved in section 3. A generalized solution is constructed from the approximate solution in section 4. In section 5 we apply the different approaches to the system of gas dynamics. Numerical results are given in section 6, and concluding remarks are made.

2. Convection problem for the elasticity.

2.1. Resolution by perturbation. The system (1.5) is strictly hyperbolic since the term $\varepsilon\, k^2$ does not vanish. Its matrix has two real eigenvalues $\lambda_1(u, \varepsilon, k) = u - \varepsilon k$ and $\lambda_2(u, \varepsilon, k) = u + \varepsilon k$ associated with two eigenvectors $r_1(k) = (1, k)^t$ and $r_2(k) = (1, -k)^t$, respectively.

As each characteristic field is genuinely nonlinear, we must deal with shock solutions or rarefaction wave solutions to solve the system (1.5). From now on, we set $W = (u,\, \sigma)^t$.

• **Shock wave solution.** If we look for a solution of the form

$$W_\varepsilon(x, t) = W_l + (W_r - W_l)H_\varepsilon(x - ct),$$

one gets $c = c_\varepsilon = \frac{u_l + u_r}{2} \mp \varepsilon\, k$.

When ε tends to 0, c_ε tends to $\frac{u_l + u_r}{2}$, and

$$W(x, t) = W_l + (W_r - W_l)H(x - ct). \tag{2.1}$$

Therefore, the limit of the function u_ε is a solution of Burgers's equation according to the Rankine–Hugoniot condition.

• **Rarefaction wave solution.** Setting $\zeta = \frac{x}{t}$, the self-similar solution of problem (1.5)–(1.2) must satisfy [8, 15] for $i = 1, 2$, the following ordinary differential system:

$$W'(\zeta) = r_i(k), \tag{2.2}$$

with

$$W(\lambda_i(W_l),\, \lambda_i(W_r)) = (W_l,\, W_r).$$

From (2.2) we find the relations

$$\sigma'(\zeta) + k\,u'(\zeta) = 0 \quad \text{or} \quad \sigma'(\zeta) - k\,u'(\zeta) = 0,$$

which give by integration the following Riemann invariants:

$$\left\{ \begin{array}{rcll} \sigma - k\,u & = & \sigma_l - k\,u_l & \text{if } u > u_l, \\ \sigma + k\,u & = & \sigma_r + k\,u_r & \text{if } u < u_r. \end{array} \right. \tag{2.3}$$

The characteristic rays passing through the origin and a general point (x,t) situated inside the fan have the following slopes: $\frac{x}{t} = u - \varepsilon\,k$ or $\frac{x}{t} = u + \varepsilon\,k$. Then, using the Riemann invariants (2.3), one gets as ε tends to zero the value inside the fan:

$$(u,\,\sigma) = \left(\frac{u_l + u_r}{2} + \frac{\sigma_r - \sigma_l}{2\,k},\; \frac{\sigma_l + \sigma_r}{2} + k\frac{u_r - u_l}{2} \right).$$

2.2. Resolution by substitution. We know that Burgers's equation presents discontinuous solutions. Therefore, let us consider in $\mathcal{G}_S(\mathbb{R}\times\mathbb{R}^+)$ the solution of the first equation of system (1.4) under the following form:

$$u(x,t) = u_l + (u_r - u_l)H(x - ct),$$

with the Rankine–Hugoniot condition $c = \frac{(u_l+u_r)}{2}$.

Then, system (1.4) can be written as

$$\left\{ \begin{array}{l} \partial_t\sigma(x,t) + u(x,t)\partial_x\sigma(x,t) \approx 0, \\ u(x,t) = u_l + (u_r - u_l)H(x - ct), \quad \text{with} \quad c = \dfrac{(u_l + u_r)}{2}. \end{array} \right. \tag{2.4}$$

When we add to (2.4) the initial condition

$$\sigma_0(x) = \sigma_l + (\sigma_r - \sigma_l)H(x), \tag{2.5}$$

we obtain a Riemann problem for a scalar equation with the discontinuous coefficient u. From Oberguggenberger [12], the hyperbolic equation (2.4) has a solution in $\mathcal{G}_S(\mathbb{R}\times\mathbb{R}^+)$.

Because u is a piecewise function, one recovers the case of scalar hyperbolic equation with constant coefficient (see [8]) outside the discontinuity. Thus, the characteristic curves are straight lines with slope $\frac{1}{u_l}$ if $x < ct$ and straight lines with slope $\frac{1}{u_r}$ for $x > ct$. To determine the discontinuous solutions of (2.4) we must deal with the sign of $u_l - u_r$.

• *If* $u_l - u_r \geq 0$, we have a shock wave solution

$$\sigma(x,t) = \sigma_l + (\sigma_r - \sigma_l)H(x - ct). \tag{2.6}$$

Moreover, placing (2.6) into (2.4) as $c = \frac{(u_l+u_r)}{2}$, we recover the following classical relation linking a generalized Heaviside function with its derivative:

$$H(x - ct)H'(x - ct) \approx \frac{1}{2}H'(x - ct).$$

• *If* $u_l - u_r < 0$, the function $\sigma(x,t)$ is equal to σ_l if $t > \frac{x}{u_l}$ and equal to σ_r for $t < \frac{x}{u_r}$ (see Figure 1). But in the sector $\frac{x}{u_r} < t < \frac{x}{u_l}$ (which is nonempty because

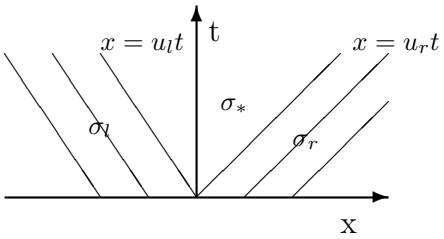


FIG. 1. *Characteristic straight lines when* $u_l - u_r < 0$.

$u_l - u_r < 0$) there is a lack of information. Proceeding as Remaki in [13], we consider a two shock wave solution with an intermediate step value σ_*:

$$\sigma(x,t) = \sigma_l + (\sigma_* - \sigma_l)H(x - u_l t) + (\sigma_r - \sigma_*)H(x - u_r t). \tag{2.7}$$

But, it appears that this function σ verifies (2.4) for all value σ_*.

In fact, a reasonable value of σ_* can derive from the initial condition. Indeed, placing (2.7) into (2.4) and making t tend to 0, one obtains

$$(\sigma_r - \sigma_*)H'(x) + H(x)\{(\sigma_r - \sigma_l)H'(x)\} \approx 0. \tag{2.8}$$

Since $H(x)H'(x) \approx \frac{1}{2}H'(x)$, the relation (2.8) gives $\frac{\sigma_r - \sigma_*}{\sigma_r - \sigma_l} = \frac{1}{2}$; therefore, $\sigma_* = \frac{\sigma_r + \sigma_l}{2}$.

2.3. The Godunov-type numerical schemes. We develop for both approaches a Godunov-type numerical scheme for solving system (1.4). We assume that u_0 and σ_0 are two initial conditions in the space of all bounded variation functions on $\mathbb{R}$. Let $h > 0$ be the space mesh size and $\Delta t = rh$ the time step, with $r > 0$. The discretization of the x-axis is performed by setting for all $i \in \mathbb{Z}$, $x_i = ih$ and $I_i = [x_{i-1/2}, x_{i+1/2}[$. For a given h, the value of the approximate pair (u_h, σ_h) at the point $x_i = ih$ and at the time $t_n = nrh$ is denoted by (u_i^n, σ_i^n), with $i \in \mathbb{Z}$ and $n \in \mathbb{N}$.

The quantities u_i^0 and σ_i^0 are obtained as the mean value of the initial data u_0 and σ_0, respectively, on each I_i. From the knowledge of (u_i^n, σ_i^n) we set $c_{i-1/2}^n = \frac{u_i^n + u_{i-1}^n}{2}$, $i \in \mathbb{Z}$. Then, according to the sign of u_i^n, $i \in \mathbb{Z}$, we compute the value of $(u_i^{n+1}, \sigma_i^{n+1})$ given either by one cell of Table 2.1 for both approaches, or by the formulas (2.17)–(2.18) for the first approach, or by the formula (2.19) for the second approach.

If $u_{i-1}^n < 0 < u_i^n$, then we define for the first approach

$$(\sigma_{i-1/2}^n, u_{i-1/2}^n) = \left(\frac{\sigma_i^n + \sigma_{i-1}^n}{2} + k\frac{u_i^n - u_{i-1}^n}{2}, \frac{u_i^n + u_{i-1}^n}{2} + \frac{\sigma_i^n - \sigma_{i-1}^n}{2\,k} \right),$$

and we obtain

$$\sigma_i^{n+1} = \sigma_i^n - r\left(\frac{u_i^n + u_{i-1/2}^n}{2}\right)(\sigma_i^n - \sigma_{i-1/2}^n), \tag{2.17}$$

$$u_i^{n+1} = u_i^n - r\left(\frac{u_i^n + u_{i-1/2}^n}{2}\right)(u_i^n - u_{i-1/2}^n). \tag{2.18}$$

We set for the second approach

$$\sigma_{i-1/2}^n = \frac{\sigma_i^n + \sigma_{i-1}^n}{2}$$

TABLE 2.1
Approximation of (u, σ) *according to the sign of* $c^n_{i-1/2}$ *and* $c^n_{i+1/2} i \in \mathbb{Z}$.

	$c^n_{i-1/2} \geq 0$	$c^n_{i-1/2} \leq 0$
$c^n_{i+1/2} \geq 0$	(2.9) $\sigma_i^{n+1} = \sigma_i^n - rc^n_{i-1/2}(\sigma_i^n - \sigma_{i-1}^n)$ (2.10) $u_i^{n+1} = \sigma_i^n - rc^n_{i-1/2}(u_i^n - u_{i-1}^n)$	(2.11) $\sigma_i^{n+1} = \sigma_i^n$ (2.12) $u_i^{n+1} = u_i^n$
$c^n_{i-1/2} < 0$	$\sigma_i^{n+1} = \sigma_i^n - rc^n_{i-1/2}(\sigma_i^n - \sigma_{i-1}^n)$ (2.13) $-rc^n_{i+1/2}(\sigma_{i+1}^n - \sigma_i^n)$ $u_i^{n+1} = u_i^n - rc^n_{i-1/2}(u_i^n - u_{i-1}^n)$ (2.14) $-rc^n_{i+1/2}(u_{i+1}^n - u_i^n)$	$\sigma_i^{n+1} = -rc^n_{i+1/2}(\sigma_{i+1}^n - \sigma_i^n) + \sigma_i^n$ (2.15) $u_i^{n+1} = -rc^n_{i+1/2}(u_{i+1}^n - u_i^n) + u_i^n$ (2.16)

and we get the following:

$$\sigma_i^{n+1} = \sigma_i^n - ru_i^n(\sigma_i^n - \sigma_{i-1/2}^n). \tag{2.19}$$

The Courant–Friedrichs–Lewy (CFL) condition, which ensures that two shock waves do not meet inside the mesh, is as follows:

$$\begin{aligned} \max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}(|u_i^n|, |u_{i+1/2}^n|) &< \frac{1}{2r} \quad \text{for the first approach,} \\ \max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|u_i^n| &< \frac{1}{2r} \quad \text{for the second approach.} \end{aligned}$$

Note that the Godunov scheme developed for the second approach is based on the sign of u_i^n while the one developed for the first approach is based on the left and right waves of, respectively, speed $C^n_{l,i+1/2} = \frac{u_i^n + u_{i+1/2}^n}{2}$ and $C^n_{r,i+1/2} = \frac{u_{i+1}^n + u_{i+1/2}^n}{2}$. The intermediate step of the velocity is defined above or equals u_i^n or u_{i+1}^n.

3. Godunov scheme by a splitting method for the elasticity. Let u_0, σ_0 be two initial conditions in $\mathcal{BV}(\mathbb{R})$. Discretizing the x-axis and the time space as in subsection 2.3, the procedure consists of two steps.

First, with the assumption that $(\mathbf{u}_i^n, \boldsymbol{\sigma}_i^n)$ is the approximate solution of the system (1.1) at time $t_n = nrh$, we compute an approximate solution $(u_i^{n,1}, \sigma_i^{n,1})$ of the system (1.4) of the convection problem at time $t_{n+1} = (n+1)rh$.

In the second step, solving the linear strictly hyperbolic system (1.3) with the initial data $(u_i^{n,1}, \sigma_i^{n,1})$, the shock waves solutions present two contact discontinuities of, respectively, propagation speed $-k$ and k.

Thanks to the Rankine–Hugoniot jump conditions, for each $i \in \mathbb{Z}$, the intermediate step value of u and σ is given by the same formulas as those obtained from the

Riemann invariants (2.3):

$$\begin{aligned} u_{i+1/2}^{n,1} &= \frac{u_i^{n,1}+u_{i+1}^{n,1}}{2}+\frac{\sigma_{i+1}^{n,1}-\sigma_i^{n,1}}{2k}, \\ \sigma_{i+1/2}^{n,1} &= k\frac{u_{i+1}^{n,1}-u_i^{n,1}}{2}+\frac{\sigma_{i+1}^{n,1}+\sigma_i^{n,1}}{2}. \end{aligned}$$

Then, by projection we construct the approximate solution $(\mathbf{u}_i^{n+1},\, \boldsymbol{\sigma}_i^{n+1})$ of the system (1.1) at time $t_{n+1}=(n+1)rh$:

$$\mathbf{u}_i^{n+1} = u_i^{n,1}+kr(u_{i+1/2}^{n,1}-u_i^{n,1})-kr(u_i^{n,1}-u_{i-1/2}^{n,1}), \tag{3.1}$$

$$\boldsymbol{\sigma}_i^{n+1} = \sigma_i^{n,1}+kr(\sigma_{i+1/2}^{n,1}-\sigma_i^{n,1})-kr(\sigma_i^{n,1}-\sigma_{i-1/2}^{n,1}), \tag{3.2}$$

where the notation $u_i^{n,1}$ and $\sigma_i^{n,1}$ represents the result of the convection step (obtained in subsection 2.3). Moreover, the constant $r>0$ has to fit so that one of the following inequalities holds:

$$\begin{aligned} \max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}(|u_i^n|,|u_{i+1/2}^n|,k) &< \frac{1}{2r} \quad \text{for the first approach,} \\ \max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}(|u_i^n|,k) &< \frac{1}{2r} \quad \text{for the second approach.} \end{aligned}$$

Let us denote by $\mathcal{B}V(\mathbb{R})$ the space of all bounded variation functions on $\mathbb{R}$ and by $\mathcal{T}V(u)$ the total variation of the function u.

THEOREM 1. *For $u_0,\sigma_0\in\mathcal{B}V(\mathbb{R})\cap L^\infty(\mathbb{R})$ and under the assumption that*

$$\frac{1}{k}\sup_{x\in\mathbb{R}}(|ku_0(x)+\sigma_0(x)|,|ku_0(x)-\sigma_0(x)|)=M, \tag{3.3}$$

$$\max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} r(|u_i^n|,|u_{i+1/2}^n|,k)<\frac{1}{2}, \tag{3.4}$$

the scheme previously introduced in section 3 is stable for $L^\infty(\mathbb{R})$-norm for the total variation in space and for the total variation in time in the Tonnelli–Cesari's sense.

Therefore, for any $T>0$ it follows from the Banach–Alaoglu–Bourbaki theorem that there exist sequences u_{h_m}, σ_{h_m} (with sequences $(h_m)_{m\in\mathbb{N}}$ tending to 0) which converge for the topology of $L^1_{loc}(\mathbb{R}\times]0,\,T[)$ toward the functions $u,\,\sigma\in L^1_{loc}(\mathbb{R}\times]0,\,T[)\cap\mathcal{B}V(\mathbb{R}\times]0,\,T[)$. The initial condition at $t=0$ is justified as in [3].

Proof of Theorem 1. Let us introduce, for each i, n,

$$R_i^n = ku_i^n+\sigma_i^n, \tag{3.5}$$

$$Q_i^n = ku_i^n-\sigma_i^n, \tag{3.6}$$

$$R_i^{n,1} = ku_i^{n,1}+\sigma_i^{n,1}, \tag{3.7}$$

$$Q_i^{n,1} = ku_i^{n,1}-\sigma_i^{n,1}. \tag{3.8}$$

From (3.1) and (3.2) one has

$$\begin{cases} R_i^{n+1} &= R_i^{n,1}(1-kr)+krR_{i+1}^{n,1}, \quad i\in\mathbb{Z}, \\ Q_i^{n+1} &= Q_i^{n,1}(1-kr)+krQ_{i-1}^{n,1}, \quad i\in\mathbb{Z}. \end{cases} \tag{3.9}$$

We are going to prove that there exists a constant A such that

$$|R_i^n| \leq kM \quad \forall\, i \in \mathbb{Z},\, n \in \mathbb{N}, \tag{3.10}$$

$$\sum_{i\in\mathbb{Z}} |R_{i+1}^n - R_i^n| \leq A \qquad \forall\, n \in \mathbb{N} \tag{3.11}$$

and

$$|Q_i^n| \leq kM \quad \forall\, i \in \mathbb{Z},\, n \in \mathbb{N}, \tag{3.12}$$

$$\sum_{i\in\mathbb{Z}} |Q_{i+1}^n - Q_i^n| \leq A \qquad \forall\, n \in \mathbb{N}. \tag{3.13}$$

We proceed by induction on n. For $n = 0$ we have (3.10) from (3.3), (3.11) from (3.5), and the bounded variation assumption on (u_0, σ_0).

We assume that (3.10) and (3.11) hold for some n. Then, from (3.4) and (3.9) one gets

$$\begin{aligned} |R_i^{n+1}| &\leq (1-kr)|R_i^{n,1}| + kr|R_{i+1}^{n,1}|, \quad i \in \mathbb{Z}, \\ \sum_{i\in\mathbb{Z}} |R_{i+1}^{n+1} - R_i^{n+1}| &\leq (1-kr)\sum_{i\in\mathbb{Z}} |R_{i+1}^{n,1} - R_i^{n,1}| + kr\sum_{i\in\mathbb{Z}} |R_{i+2}^{n,1} - R_{i+1}^{n,1}| \\ &\leq \sum_{i\in\mathbb{Z}} |R_{i+1}^{n,1} - R_i^{n,1}|. \end{aligned}$$

Now we are going to prove that

$$|R_i^{n,1}| \leq kM, \tag{3.14}$$

$$\sum_{i\in\mathbb{Z}} |R_{i+1}^{n,1} - R_i^{n,1}| \leq \sum_{i\in\mathbb{Z}} |R_{i+1}^n - R_i^n|. \tag{3.15}$$

The Riemann invariant $R_i^{n,1}$ is defined with the approximate solution of the convection problem (see subsection 2.3), constructed according to the sign of the velocity u. We start with the case where u has a constant sign; then we treat the case where the velocity has a variable sign.

When $u_i^n < 0$ for all $i \in \mathbb{Z}$, obviously $c_{i+1/2}^n < 0$, and one obtains by projection (see Figure 2)

$$\begin{aligned} \sigma_i^{n,1} &= \sigma_i^n - rc_{i+\frac{1}{2}}^n(\sigma_{i+1}^n - \sigma_i^n), \\ u_i^{n,1} &= u_i^n - rc_{i+\frac{1}{2}}^n(u_{i+1}^n - \sigma_i^n). \end{aligned}$$

Using (3.4) and then shifting the index, it happens that

$$\sum_{i\leq s-1} |R_{i+1}^{n,1} - R_i^{n,1}| \leq \sum_{i\leq s-1} |R_{i+1}^n - R_i^n| - rc_{s+\frac{1}{2}}^n |R_{s+1}^n - R_s^n|, \tag{3.16}$$

$$\sum_{i\geq s} |R_{i+1}^{n,1} - R_i^{n,1}| \leq \sum_{i\geq s} |R_{i+1}^n - R_i^n| + rc_{s+\frac{1}{2}}^n |R_{s+1}^n - R_s^n|, \tag{3.17}$$

and $|R_i^{n,1}| \leq (1 + rc_{i+1/2}^n)|R_i^n| - rc_{i+1/2}^n |R_{i+1}^n| \leq kM$ for all $i \in \mathbb{Z}$.

Moreover, (3.16) and (3.17) give

$$\sum_{i\in\mathbb{Z}} |R_{i+1}^{n,1} - R_i^{n,1}| = \sum_{i\leq s-1} |R_{i+1}^{n,1} - R_i^{n,1}| + \sum_{i\geq s} |R_{i+1}^{n,1} - R_i^{n,1}| \leq \sum_{i\in\mathbb{Z}} |R_{i+1}^n - R_i^n|.$$

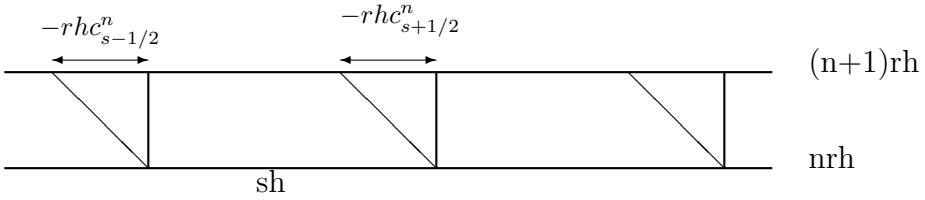


FIG. 2. *Type* 1. $rc^n_{s-1/2} < 0$ *and* $rc^n_{s+1/2} < 0$.

In the case $u^n_i \geq 0$ for all $i \in \mathbb{Z}$, corresponding to type 2, we similarly obtain the same inequalities (3.14) and (3.15).

Since we have chosen a monotonous solution of Burgers's equation, its sign can change only once. Consequently, when u has a variable sign, three cases (types 3, 4, and 5) corresponding to the three pictures Figures 3, 4, and 5 can occur.

Type 3 is the case where there exists $p \in \mathbb{Z}$ such that $u^n_i \leq 0$ for all $i \leq p$ and $u^n_i \geq 0$ for all $i \geq p+1$. We denote by $C^n_{l,i+1/2} = (u^n_i + u^n_{i+1/2})/2$ and $C^n_{r,i+1/2} = (u^n_{i+1} + u^n_{i+1/2})/2$, respectively, the left and the right sound speed between the cells I_p and I_{p+1}.

Thus, by projection (according to Figure 3) and using the CFL condition (3.4) we obtain

$$|R^{n,1}_p| \leq kM \quad \text{and} \quad |R^{n,1}_{p+1}| \leq kM. \tag{3.18}$$

Since the relation (3.14) is valid for $u^n_i \leq 0$ and $u^n_i \geq 0$ for all $i \in \mathbb{Z}$ (types 1 and 2), we have for $i \leq p$ and $i \geq p+2$, $|R^{n,1}_i| \leq kM$. Then with (3.18) one gets $|R^{n,1}_i| \leq kM$, $i \in \mathbb{Z}$.

Moreover,

$$\begin{aligned} |R^{n,1}_p - R^{n,1}_{p-1}| &\leq (1 + rc^n_{p-\frac{1}{2}})|R^n_p - R^n_{p-1}| - rC^n_{l,p+\frac{1}{2}}|R^n_{p+1} - R^n_p|, \\ |R^{n,1}_{p+1} - R^{n,1}_p| &\leq (1 + rC^n_{l,p+\frac{1}{2}})|R^n_{p+1} - R^n_p|, \\ |R^{n,1}_{p+2} - R^{n,1}_{p+1}| &\leq (1 - rc^n_{p+\frac{3}{2}})|R^n_{p+2} - R^n_{p+1}|, \end{aligned} \tag{3.19}$$

and in the case $u^n_i \geq 0$ for all $i \in \mathbb{Z}$,

$$\sum_{i \geq p+2} |R^{n,1}_{i+1} - R^{n,1}_i| \leq \sum_{i \geq p+2} |R^n_{i+1} - R^n_i| + rc^n_{p+\frac{3}{2}}|R^n_{p+2} - R^n_{p+1}|. \tag{3.20}$$

Adding these last estimates (3.20) to (3.19) and to (3.16) with $s = p-1$, one obtains

$$\sum_{i \in \mathbb{Z}} |R^{n,1}_{i+1} - R^{n,1}_i| \leq \sum_{i \in \mathbb{Z}} |R^n_{i+1} - R^n_i|.$$

Type 4 is the case where there exists p such that $u^n_i \leq 0$ for all $i \leq p$ and $u^n_i \geq 0$ for all $i \geq p+2$. Still by projection (according to Figure 4), one obtains

$$|R^{n,1}_{p+1}| = |R^n_{p+1}| \leq kM; \tag{3.21}$$

consequently, in the same way as previously (for type 3), the relation (3.14) is valid for all $i \in \mathbb{Z}$. Moreover,

$$\begin{aligned} |R^m_{p+1} - R^m_p| &= (1 + rc^n_{p+\frac{1}{2}})|R^n_{p+1} - R^n_p|, \\ |R^m_{p+2} - R^m_{p+1}| &= (1 - rc^n_{p+\frac{3}{2}})|R^n_{p+2} - R^n_{p+1}|; \end{aligned} \tag{3.22}$$

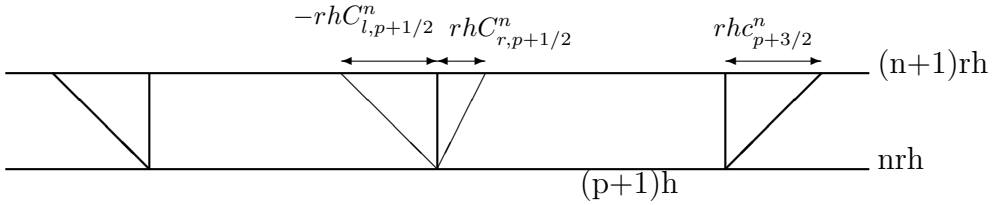


FIG. 3. *Type* 3. $u^n_{p+1} \geq 0$, *and* $u^n_p \leq 0$.

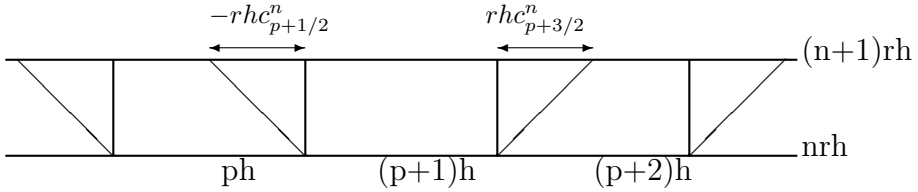


FIG. 4. *Type* 4. $rc^n_{p+1/2} \leq 0$, *and* $rc^n_{p+3/2} \geq 0$.

then adding the above estimates (3.22) to (3.16) for $s = p$ and to (3.20) one gets

$$\sum_{i\in\mathbb{Z}} |R^{n,1}_{i+1} - R^{n,1}_i| \leq \sum_{i\in\mathbb{Z}} |R^n_{i+1} - R^n_i|.$$

Type 5 is the case where there exists p such that $u^n_i \geq 0$ for all $i \leq p$ and $u^n_i \leq 0$ for all $i \geq p+2$. By projection (according to Figure 5), we obtain using the CFL condition (3.4)

$$|R^{n,1}_{p+1}| \leq |R^n_{p+1}|(1 - rc^n_{p+\frac{1}{2}} + rc^n_{p+\frac{3}{2}}) + rc^n_{p+\frac{1}{2}}|R^n_p| - rc^n_{p+\frac{3}{2}}|R^n_{p+2}| \leq kM. \tag{3.23}$$

As the relation (3.14) is valid for $u^n_1 \geq 0$ and $u^n_i \leq 0$ for all $i \in \mathbb{Z}$ (types 1 and 2), we have for $i \leq p$ and $i \geq p+2$, $| R^{n,1}_i | \leq kM$. Then, with (3.23) one gets $|R^{n,1}_i| \leq kM$, $i \in \mathbb{Z}$.

Besides, when $u^n_i \geq 0$ for all $i \in \mathbb{Z}$,

$$\sum_{i\leq p} |R^{n,1}_{i+1} - R^{n,1}_i| \leq \sum_{i\leq p} |R^n_{i+1} - R^n_i| - rc^n_{p+\frac{1}{2}}|R^n_{p+1} - R^n_p|, \tag{3.24}$$

and the following inequalities hold:

$$\begin{aligned} |R^m_{p+1} - R^m_p| &\leq (1 - rc^n_{p+\frac{1}{2}})|R^n_{p+1} - R^n_p| \\ &+ rc^n_{p-\frac{1}{2}}|R^n_p - R^n_{p-1}| - rc^n_{p+\frac{3}{2}}|R^n_{p+2} - R^n_{p+1}|, \\ |R^m_{p+2} - R^m_{p+1}| &\leq (1 + rc^n_{p+\frac{3}{2}})|R^n_{p+2} - R^n_{p+1}| \\ &+ rc^n_{p+\frac{1}{2}}|R^n_{p+1} - R^n_p| - rc^n_{p+\frac{5}{2}}|R^n_{p+3} - R^n_{p+2}|. \end{aligned} \tag{3.25}$$

Adding the above estimates (3.25) to (3.24) and to (3.17) for $s = p+1$, one obtains (3.15).

One similarly obtains the estimates

$$|Q^{n,1}_i| \leq kM \quad \text{and} \quad \sum_{i\in\mathbb{Z}} |Q^{n,1}_{i+1} - Q^{n,1}_i| \leq \sum_{i\in\mathbb{Z}} |Q^n_{i+1} - Q^n_i|, \tag{3.26}$$

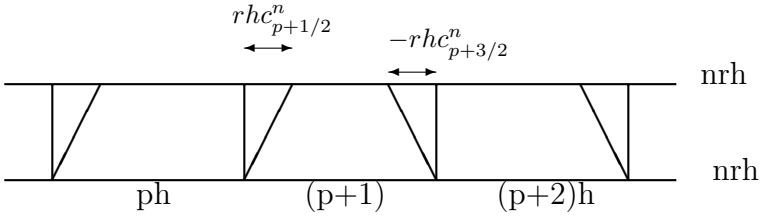


FIG. 5. *Type* 5. $c^n_{p+1/2} \geq 0$, *and* $c^n_{p+3/2} \leq 0$.

from which we deduce the inequalities (3.12) and (3.13).

With the estimations of each Riemann invariant (3.10) and (3.12), one has the stability for the L^∞-norm:

$$|u_i^n| \leq \frac{1}{2k}(|R_i^n| + |Q_i^n|) \leq M \quad \text{and} \quad |\sigma_i^n| \leq \frac{1}{2}(|R_i^n| + |Q_i^n|) \leq kM.$$

Further, from (3.11) and (3.13) one gets the stability for the total variation in space:

$$\begin{aligned} |u_{i+1}^n - u_i^n| &\leq \frac{1}{2k}\left(|R_{i+1}^n - R_i^n| + |Q_{i+1}^n - Q_i^n|\right) \leq \frac{1}{k}A, \\ |\sigma_{i+1}^n - \sigma_i^n| &\leq \frac{1}{2}\left(|R_{i+1}^n - R_i^n| + |Q_{i+1}^n - Q_i^n|\right) \leq A. \end{aligned}$$

To complete the proof of Theorem 1, an easy way to show the stability for total variation in the time property is to observe that the Riemann invariants (3.7) and (3.8), defined with the approximate solutions of the system (1.4), verify the following lemma.

LEMMA 1. *For all* $i \in \mathbb{Z} \quad \exists\, \alpha_i, \beta_i \in \{0, \frac{1}{4}, \frac{1}{2}\}$, *such that*

$$\begin{aligned} |R_i^{n,1} - R_i^n| &\leq \alpha_i |R_{i+1}^n - R_i^n| + \beta_i |R_i^n - R_{i-1}^n|, \\ |Q_i^{n,1} - Q_i^n| &\leq \alpha_i |Q_{i+1}^n - Q_i^n| + \beta_i |Q_i^n - Q_{i-1}^n|. \end{aligned}$$

Proof of Lemma 1. The CFL inequality (3.4) allows us to conclude. ∎

We have from (3.9), (3.26), (3.15), and Lemma 1 that

$$\begin{aligned} \sum_{i\in\mathbb{Z}} |u_i^{n+1} - u_i^n| &\leq \frac{1}{2k}\sum_{i\in\mathbb{Z}} \left(|R_i^{n+1} - R_i^n| + |Q_i^{n+1} - Q_i^n|\right) \\ &\leq \frac{1}{2k}\sum_{i\in\mathbb{Z}} \left(|R_i^{n+1} - R_i^{n,1}| + |R_i^{n,1} - R_i^n|\right) \\ &+ \frac{1}{2k}\sum_{i\in\mathbb{Z}} \left(|Q_i^{n+1} - Q_i^{n,1}| + |Q_i^{n,1} - Q_i^n|\right) \\ &\leq \frac{1}{2k}\sum_{i\in\mathbb{Z}} kr\, \left(|R_{i+1}^{n,1} - R_i^{n,1}| + |Q_i^{n,1} - Q_{i-1}^{n,1}|\right) \\ &+ \frac{1}{2k}\sum_{i\in\mathbb{Z}} (\alpha_i + \beta_{i+1}) \left(|R_{i+1}^n - R_i^n| + |Q_{i+1}^n - Q_i^n|\right) \\ &\leq \frac{1}{2k}\sum_{i\in\mathbb{Z}} (kr + \alpha_i + \beta_{i+1}) \left(|R_{i+1}^n - R_i^n| + |Q_{i+1}^n - Q_i^n|\right). \end{aligned}$$

As $\alpha_i + \beta_{i+1} \leq 1$ for all i in $\mathbb{Z}$ and $k\,r < 1/2$, it follows that

$$\sum_{i\in\mathbb{Z}} |u_{i+1}^{n+1} - u_i^n| \leq \frac{3}{4k}\sum_{i\in\mathbb{Z}} \left(|R_{i+1}^n - R_i^n| + |Q_{i+1}^n - Q_i^n|\right) \leq \frac{3}{2k}A.$$

An analogous estimate holds for $\sum_{i\in\mathbb{Z}} |\sigma_{i+1}^{n+1} - \sigma_i^n|$.

4. A generalized solution of the elasticity problem. Let ρ be a $\mathcal{C}^\infty$ function on $\mathbb{R}$ with compact support and nonnegative values such that $\int \rho(x)\,dx = 1$.

We define two mappings R_u and R_σ from $]0,\,1]\times\mathbb{R}^2$ into $\mathbb{R}$ by

$$\begin{aligned} R_u(\varepsilon,x,t) &= \iint u_\varepsilon(x-\varepsilon^3\zeta, t-\varepsilon^3 r\tau)\rho(\zeta)\rho(\tau)\,d\zeta d\tau, \\ R_\sigma(\varepsilon,x,t) &= \iint \sigma_\varepsilon(x-\varepsilon^3\zeta, t-\varepsilon^3 r\tau)\rho(\zeta)\rho(\tau)\,d\zeta d\tau \end{aligned} \tag{4.1}$$

if $\varepsilon = h_m$ for some m and by

$$R_u(\varepsilon,x,t) = R_u(h_m,x,t) \text{ and } R_\sigma(\varepsilon,x,t) = R_\sigma(h_m,x,t)$$

if $h_{m+1} \le \varepsilon < h_m$.

Since the scheme is stable for the L^∞-norm, there exists $A \in \mathbb{R}_+^*$ such that for all $t > 0$,

$$\sup_\varepsilon ||u_\varepsilon(.,\,t)||_{L^\infty} \le A, \quad \sup_\varepsilon ||\sigma_\varepsilon(.,\,t)||_{L^\infty} \le A,$$

and $R_u,\ R_\sigma \in \mathcal{E}_{M,S}(\mathbb{R}\times\mathbb{R}^+)$. Let U and Σ be the respective classes of R_u and R_σ in $\mathcal{G}_S(\mathbb{R}\times\mathbb{R}^+)$.

THEOREM 2. *The generalized functions* $U,\ \Sigma \in \mathcal{G}_S(\mathbb{R}\times\mathbb{R}^+)$ *are solutions of the system* (1.1) *renamed as follows:*

$$\partial_t U(x,t) + U(x,t)\partial_x U(x,t) - \partial_x \Sigma(x,t) \approx 0, \tag{4.2}$$

$$\partial_t \Sigma(x,t) + U(x,t)\partial_x \Sigma(x,t) - k^2 \partial_x U(x,t) \approx 0. \tag{4.3}$$

The connection between the $U_{|_{t=0}}$, $\Sigma_{|_{t=0}}$ and U, Σ, with the classical functions u_0, σ_0 and u, σ (the weak star limit of the subsequences u_{h_m}, σ_{h_m}), is given by the following proposition.

PROPOSITION 1. $\left(U_{|_{t=0}}, \Sigma_{|_{t=0}}\right)$ *and* (U,Σ) *have, respectively,* (u_0,σ_0) *and* (u,σ) *as macroscopic aspects.*

The proof of Proposition 1 can be found in [3, 2].

Proof of Theorem 2. In order to prove (4.3), by the definition of association we must prove that for all $\psi \in \mathcal{D}(\mathbb{R}\times]0,\,\infty[)$, the quantity

$$\begin{aligned} P_\varepsilon &= \int\int \left(\frac{\partial}{\partial t} R_\sigma(\varepsilon,x,t) - R_u(\varepsilon,x,t)\frac{\partial}{\partial x} R_\sigma(\varepsilon,x,t)\right)\psi(x,t)\,dx\,dt \\ &- \int\int k^2 \frac{\partial}{\partial x} R_u(\varepsilon,x,t)\psi(x,t)\,dx\,dt \end{aligned} \tag{4.4}$$

tends to 0 as $\varepsilon \to 0$.

We adapt the process developed by Cauret in [3] to the case where the velocity sign can change.

If there exists $p \in \mathbb{Z}$ such that $u_p < 0 < u_{p+1}$,

$$E_{p+1}^n = (\sigma_{p+1}^{n+1} - \sigma_{p+1}^n) + rC_{l,p+\frac{1}{2}}^n(\sigma_{p+\frac{1}{2}}^n - \sigma_p^n) + rC_{r,p+\frac{1}{2}}^n(\sigma_{p+1}^n - \sigma_{p+\frac{1}{2}}^n) - rk^2(u_{p+1}^n - u_p^n);$$

otherwise,

$$E_{p+1}^n = (\sigma_{p+1}^{n+1} - \sigma_{p+1}^n) + rc_{p+\frac{1}{2}}^n(\sigma_{p+1}^n - \sigma_p^n) - rk^2(u_{p+1}^n - u_p^n).$$

Let us denote for $i \neq p+1$

$$E_i^n = (\sigma_i^{n+1} - \sigma_i^n) + rc_{i-\frac{1}{2}}^n(\sigma_i^n - \sigma_{i-1}^n) - rk^2(u_i^n - u_{i-1}^n).$$

Setting $\psi_i^n = \psi((i-\frac{1}{2})\varepsilon, r(n-\frac{1}{2})\varepsilon)$, we obtain

$$\left| P_\varepsilon - \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} (\varepsilon - 2\varepsilon^3)\psi_i^n E_i^n \right| \leq C\varepsilon.$$

Therefore, P_ε tends to 0 when $\varepsilon \to 0$, if $\lim_{\varepsilon\to 0}\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} (\varepsilon - 2\varepsilon^3)\psi_i^n E_i^n = 0$.

To complete the proof of (4.3), we must prove that the approximate solution is such that the following inequality holds:

$$\begin{aligned}\left|(\varepsilon - 2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} \psi_i^n E_i^n\right| &\leq C\varepsilon^2 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |\sigma_{i+1}^n - \sigma_i^n| \left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} \\ &+ C\varepsilon^2 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |u_i^n - u_{i-1}^n| \left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} \\ &+ C\varepsilon^3 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |\sigma_i^{n,1} + ku_i^n| \left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}. \end{aligned} \tag{4.5}$$

Indeed, as the scheme is stable for the total variation in space and $\psi \in \mathcal{D}(\mathbb{R}\times]0,\infty[)$, one gets

$$\left|(\varepsilon - 2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} \psi_i^n E_i^n\right| \leq C\varepsilon.$$

Therefore,

$$(\varepsilon - 2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} \psi_i^n E_i^n \to 0 \text{ when } \quad \varepsilon \to 0.$$

Using the classification in different types introduced in the section 3, we are going to prove the estimation (4.5). For type 1 (see Figure 2), we have

$$\begin{aligned}\psi_i^n E_i^n &= \psi_i^n - rc_{i+\frac{1}{2}}^n(\sigma_{i+1}^n - \sigma_i^n)\psi_i^n + rc_{i-\frac{1}{2}}^n(\sigma_i^n - \sigma_{i-1}^n)\psi_i^n \\ &+ \frac{rk^2}{2}\left[-rc_{i+\frac{3}{2}}^n(u_{i+2}^n - u_{i+1}^n)\psi_i^n + rc_{i-\frac{1}{2}}^n(u_i^n - u_{i-1}^n)\psi_i^n\right] \\ &+ \frac{kr}{2}\left[(\sigma_{i+1}^{n,1} + ku_{i+1}^n)\psi_i^n - 2(\sigma_i^{n,1} + ku_i^n)\psi_i^n + (\sigma_{i-1}^{n,1} + ku_{i-1}^n)\psi_i^n\right].\end{aligned}$$

Then using the mean value theorem with the CFL condition (3.4) we achieve

$$\begin{aligned}\left|(\varepsilon - 2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} \psi_i^n E_i^n\right| &\leq C\varepsilon^2 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |\sigma_{i+1}^n - \sigma_i^n| \left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} \\ &+ C\varepsilon^2 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |u_i^n - u_{i-1}^n| \left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} \\ &+ C\varepsilon^3 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |\sigma_i^{n,1} + ku_i^n| \left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}. \end{aligned}$$

Similarly we have for type 2 ($u_i^n \geq 0$ for all $i \in \mathbb{Z}$)

$$\begin{aligned}\left|(\varepsilon - 2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} \psi_i^n E_i^n\right| &\leq C\varepsilon^2 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |u_i^n - u_{i-1}^n| \left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} \\ &+ C\varepsilon^3 \sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}} |\sigma_i^{n,1} + ku_i^n| \left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}.\end{aligned}$$

For type 3 (see Figure 3) we have

$$\begin{aligned}\sum_{\substack{i\leq p\\ n\in\mathbb{N}}} \psi_i^n E_i^n &= \sum_{\substack{i\leq p-1\\ n\in\mathbb{N}}} rc_{i+1/2}^n(\sigma_{i+1}^n - \sigma_i^n)(\psi_{i+1}^n - \psi_i^n) - rC_{l,p+1/2}^n(\sigma_{p+1/2}^n - \sigma_p^n)\psi_p^n \\ &+ \frac{rk^2}{2}\left[-rC_{l,p+1/2}^n(u_{p+1/2}^n - u_p^n)\psi_{p-1}^n - rC_{r,p+1/2}^n(u_{p+1}^n - u_{p+1/2}^n)\psi_p^n\right] \\ &+ \frac{kr}{2}\left[(\sigma_{p+1}^{n,1} + ku_{p+1}^n)\psi_p^n - 2(\sigma_p^{n,1} + ku_p^n)\psi_p^n + (\sigma_p^{n,1} + ku_p^n)\psi_{p-1}^n\right] \\ &+ \frac{kr}{2}\sum_{\substack{i\leq p-1\\ n\in\mathbb{N}}} (\sigma_i^{n,1} + ku_i^n)(\psi_{i+1}^n - 2\psi_i^n + \psi_{i-1}^n) \\ &+ \frac{rk^2}{2}\sum_{\substack{i\leq p-1\\ n\in\mathbb{N}}} rc_{i+1/2}^n(u_{i+1}^n - u_i^n)(\psi_{i+1}^n - \psi_{i-1}^n),\end{aligned} \tag{4.6}$$

$$\begin{aligned}\psi_{p+1}^n E_{p+1}^n &= rC_{l,p+1/2}^n(\sigma_{p+1/2}^n - \sigma_{p+1}^n)\psi_{p+1}^n \\ &+ \frac{rk^2}{2}\left[rC_{l,p+1/2}^n(u_{p+1/2}^n - u_p^n)\psi_{p+1}^n - rc_{p+3/2}^n(u_{p+2}^n - u_{p+1}^n)\psi_{p+1}^n\right] \\ &+ \frac{kr}{2}\left[(\sigma_{p+2}^{n,1} + ku_{p+2}^n)\psi_{p+1}^n - 2(\sigma_{p+1}^{n,1} + ku_{p+1}^n)\psi_{p+1}^n + (\sigma_p^{n,1} + ku_p^n)\psi_{p+1}^n\right],\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}\sum_{\substack{i\geq p+2\\ n\in\mathbb{N}}} \psi_i^n E_i^n &= \frac{rk^2}{2}\left[rC_{r,p+1/2}^n(u_{p+1}^n - u_{p+1/2}^n)\psi_{p+2}^n + rc_{p+3/2}^n(u_{p+2}^n - u_{p+1}^n)\psi_{p+3}^n\right] \\ &+ \frac{kr}{2}\left[\sum_{\substack{i\geq p+3\\ n\in\mathbb{N}}} (\sigma_i^{n,1} + ku_i^n)(\psi_{i+1}^n - 2\psi_i^n + \psi_{i-1}^n) + (\sigma_{p+1}^{n,1} + ku_{p+1}^n)\psi_{p+2}^n\right] \\ &+ \frac{kr}{2}\left[-2(\sigma_{p+2}^{n,1} + ku_{p+2}^n)\psi_{p+2}^n + (\sigma_{p+2}^{n,1} + ku_{p+2}^n)\psi_{p+3}^n\right] \\ &+ \frac{rk^2}{2}\left[\sum_{\substack{i\geq p+2\\ n\in\mathbb{N}}} rc_{i+1/2}^n(u_{i+1}^n - u_i^n)(\psi_{i+2}^n - \psi_i^n)\right].\end{aligned} \tag{4.8}$$

Therefore, summing (4.6), (4.7), (4.8) and using the mean value theorem with the

CFL condition (3.4) we get

$$\begin{aligned}\left|(\varepsilon-2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}\psi_i^n E_i^n\right| &\le C\varepsilon^2\sum_{\substack{i\le p\\ n\in\mathbb{N}}}|\sigma_{i+1}^n-\sigma_i^n|\left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^2\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|u_i^n-u_{i-1}^n|\left\|\frac{\partial}{\partial x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^3\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|\sigma_i^{n,1}+ku_i^n|\left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}.\end{aligned}$$

The computations for getting the estimation (4.5) for types 4 and 5 are similar to those presented for type 3. For type 4, one obtains

$$\begin{aligned}\left|(\varepsilon-2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}\psi_i^n E_i^n\right| &\le C\varepsilon^2\sum_{\substack{i\le p\\ n\in\mathbb{N}}}|\sigma_{i+1}^n-\sigma_i^n|\left\|\frac{\partial}{\partial_x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^2\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|u_i^n-u_{i-1}^n|\left\|\frac{\partial}{\partial_x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^3\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|\sigma_i^{n,1}+ku_i^n|\left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)},\end{aligned}$$

and for type 5 one gets

$$\begin{aligned}\left|(\varepsilon-2\varepsilon^3)\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}\psi_i^n {E_1}_i^n\right| &\le C\varepsilon^2\sum_{\substack{i\ge p+1\\ n\in\mathbb{N}}}|\sigma_{i+1}^n-\sigma_i^n|\left\|\frac{\partial}{\partial_x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^2\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|u_i^n-u_{i-1}^n|\left\|\frac{\partial}{\partial_x}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}\\ &+ C\varepsilon^3\sum_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}|\sigma_i^{n,1}+ku_i^n|\left\|\frac{\partial^2}{\partial x^2}\psi\right\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)}.\end{aligned}$$

Consequently,

$$P_\varepsilon\to 0 \quad \text{when} \quad \varepsilon\to 0.$$

The association (4.2) can be proved in the same way as (4.3). This ends the proof of Theorem 2. □

Following the study of the model of elasticity in section 2, we use the same approaches for the system of hydrodynamics in the nonconservative form (1.6).

5. A Godunov scheme by a splitting method for a model of hydrodynamics. Let v_0, u_0, σ_0 be the initial conditions in $\mathcal{B}V(\mathbb{R})$. Discretizing the x-axis and the time space and using the same notation as in subsection 2.3, we denote by v_i^0, u_i^0, and p_i^0 the mean value of the initial data v_0, u_0, and p_0, respectively, on each I_i. From our knowledge of (v_i^n, u_i^n, p_i^n), the solution of (1.6), we compute the value $(v_i^{n+1}, u_i^{n+1}, p_i^{n+1})$ in two steps.

First, with $(v_i^n,\, u_i^n,\, p_i^n)$ as initial data, we compute an approximate solution $(v_i^{n,1},\, u_i^{n,1},\, p_i^{n,1})$ of the Riemann problem associated with the convection terms (5.1):

$$\left\{\begin{array}{rcl} \partial_t v(x,t) + u(x,t)\partial_x v(x,t) &=& 0,\\ \partial_t u(x,t) + u(x,t)\partial_x u(x,t) &=& 0,\\ \partial_t p(x,t) + u(x,t)\partial_x p(x,t) &=& 0. \end{array}\right. \tag{5.1}$$

The formulas of the approximate solution for both approaches have the same form as those given in subsection 2.3 for the system (1.4). Only the intermediate step values change, when $u_l < 0 < u_r$.

Using the approach by perturbation, we obtain from the Riemann invariants the following value inside the rarefaction fan:

$$\begin{aligned} u^n_{i-\frac{1}{2}} &= u^n_{i-1} + \frac{u^n_i - u^n_{i-1} + \dfrac{2B^n_i}{\gamma-1}\left(\left(\dfrac{p^n_{i-1}}{p^n_i}\right)^{\frac{\gamma-1}{2\gamma}} - 1\right)}{\left(\dfrac{B^n_i}{B^n_{i-1}}\right)\left(\dfrac{p^n_{i-1}}{p^n_i}\right)^{\frac{\gamma-1}{2\gamma}} + 1},\\ p^n_{i-\frac{1}{2}} &= p^n_{i-1}\left(1 + \frac{\gamma-1}{2B^n_{i-1}}(u^n_{i-1} - u^n_{i-\frac{1}{2}})\right)^{\frac{2\gamma}{\gamma-1}},\\ v^n_{i-\frac{1}{2}} &= v^n_{i-1}\left((1 + \frac{\gamma-1}{2B^n_{i-1}}(u^n_{i-1} - u^n_{i-\frac{1}{2}})\right)^{\frac{2}{1-\gamma}}, \end{aligned}$$

where the sound speed $B^n_i = \sqrt{\gamma\, v^n_i\, p^n_i}$.

Using the approach by substitution, we study the first and the third equation of system (5.1) as the scalar hyperbolic equation (2.4). Then we obtain the following intermediate step values:

$$\begin{aligned} p^n_{i-1/2} &= \frac{p^n_{i-1} + p^n_i}{2},\\ v^n_{i-1/2} &= \frac{v^n_{i-1} + v^n_i}{2}. \end{aligned}$$

Therefore, we approximate the velocity u with the same formulas as in subsection 2.3. Then, as p and v play the role of σ, we substitute them and recover the expressions of the approximate specific volume and approximate pressure.

Second, with the value $(v_i^{n,1},\, u_i^{n,1},\, p_i^{n,1})$ as initial data and considering only the Hugoniot curves, we construct an approximate solution for the system of propagation waves:

$$\left\{\begin{array}{rcl} \partial_t v(x,t) - v(x,t)\partial_x u(x,t) &=& 0,\\ \partial_t u(x,t) + v(x,t)\partial_x p(x,t) &=& 0,\\ \partial_t p(x,t) + \gamma p(x,t)\partial_x u(x,t) &=& 0. \end{array}\right. \tag{5.2}$$

With the Riemann solver developed by Colombeau and Le Roux in [6, p. 30], one

computes the flux across the interface between the cells I_i and I_{i+1} to obtain

$$\begin{aligned}
u^{n,1}_{i+1/2} &= \frac{1}{Z^{n,1}_{l,i+1/2}+Z^{n,1}_{r,i+1/2}}(Z^{n,1}_{l,i+1/2}u^{n,1}_i + Z^{n,1}_{r,i+1/2}u^{n,1}_{i+1} + p^{n,1}_i - p^{n,1}_{i+1}),\\
v^{n,1}_{2,i+1/2} &= \frac{1}{Z^{n,1}_{r,i+1/2}}(u^{n,1}_{i+1} - u^{n,1}_{i+1/2}) + v^n_{i+1},\\
v^{n,1}_{1,i+1/2} &= \frac{1}{Z^{n,1}_{l,i+1/2}}(u^{n,1}_{i+1/2} - u^{n,1}_i) + v^n_i,\\
p^{n,1}_{i+1/2} &= p^n_i - Z^{n,1}_{l,i+1/2}(u^{n,1}_{i+1/2} - u^{n,1}_i),
\end{aligned}$$

where the impedances are as follows:

$$Z^{n,1}_{l,i+1/2} = \sqrt{\gamma\frac{p^{n,1}_{i+1/2}+p^{n,1}_i}{v^{n,1}_{1,i+1/2}+v^{n,1}_i}}, \quad Z^{n,1}_{r,i+1/2} = \sqrt{\gamma\frac{p^{n,1}_{i+1/2}+p^{n,1}_{i+1}}{v^{n,1}_{2,i+1/2}+v^{n,1}_{i+1}}}.$$

Then with the sound speeds

$$C^{n,1}_{l,i+1/2} = -\frac{v^{n,1}_i + \overline{v}^{n,1}_{1,i+1/2}}{2}Z^n_{l,i+1/2}, \quad C^{n,1}_{r,i+1/2} = \frac{v^{n,1}_{i+1} + \overline{v}^{n,1}_{2,i+1/2}}{2}Z^n_{r,i+1/2},$$

we compute by projection

$$\begin{aligned}
u^{n+1}_i &= u^{n,1}_i - rC^{n,1}_{l,i+1/2}(u^{n,1}_{i+1/2} - u^{n,1}_i) - rC^{n,1}_{r,i-1/2}(u^{n,1}_i - u^{n,1}_{i-1/2}),\\
v^{n+1}_i &= v^{n,1}_i - rC^{n,1}_{l,i+1/2}(v^{n,1}_{1,i+1/2} - v^{n,1}_i) - rC^{n,1}_{r,i-1/2}(v^{n,1}_i - v^{n,1}_{2,i-1/2}),\\
p^{n+1}_i &= p^{n,1}_i - rC^{n,1}_{l,i+1/2}(p^{n,1}_{i+1/2} - p^{n,1}_i) - rC^{n,1}_{r,i-1/2}(p^{n,1}_i - p^{n,1}_{i-1/2}).
\end{aligned} \tag{5.3}$$

This latter step is stable if for all $i \in \mathbb{Z},\ n \in \mathbb{N}$

$$1 - rC^{n,1}_{r,i-1/2} + rC^{n,1}_{l,i+1/2} > 0.$$

The CFL condition for the split scheme is ensured as soon as the constant $r > 0$ is adapted so that one of the following inequalities holds:

$$\begin{aligned}
\max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}(|C^{n,1}_{l,i+1/2}|, |C^{n,1}_{r,i-1/2}|, |u^n_i|, |u^n_{i+1/2}|) &< 1/2r \quad \text{for the first approach,}\\
\max_{\substack{i\in\mathbb{Z}\\ n\in\mathbb{N}}}(|C^{n,1}_{l,i+1/2}|, |C^{n,1}_{r,i-1/2}|, |u^n_i|) &< 1/2r \quad \text{for the second approach.}
\end{aligned}$$

The formula (5.3) can be at the origin of defects in conservation of mass, momentum, and energy. To lessen this drawback, we use the conservative projection developed by Colombeau in [5, p. 83], which consists of projecting the density instead of the specific volume in the second step.

6. Numerical results. Following the study of the Riemann problem associated with the system (1.1) and (1.6), we are going to present numerical results obtained by the different schemes. We denote by the SHS scheme and the NSHS scheme, respectively, the new schemes obtained with the convection problems by perturbation and by substitution.

From different initial data, we compare the solution obtained by the Colombeau–Le Roux scheme (C-L scheme) with those of the new schemes (SHS scheme and

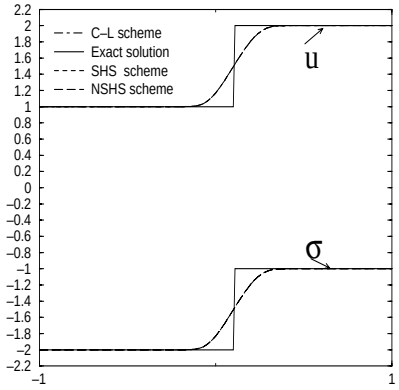


FIG. 6. *Solution of the Riemann problem associated with the system* (1.1)*, when* $(u_l, u_r) = (1, 2)$.

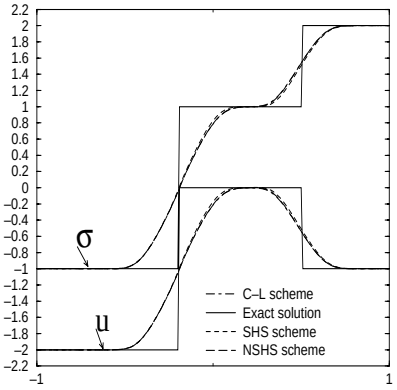


FIG. 7. *Solution of the Riemann problem associated with the system* (1.1)*, when* $(u_l, u_r) = (-1, 2)$.

NSHS scheme). Then, we conduct convergence analysis and study the behavior of conservation errors.

For the numerical tests of both systems, whenever it is not written, the space mesh size h is fixed to a convenient value $h_0 = 0.01$. From now on, the parameter $r = 0.2$ is a fixed real number such that time step $\Delta t = rh$. In order to compare the different schemes, we are interested in the approximate solution after 100 time steps (i.e., the simulation time $T = 0.2$). However, for the convergence tests, we fixed $T = 0.1$.

6.1. Numerical results for the system modeling elasticity. Let $(\sigma_l, \sigma_r) = (-2, -1)$ be the initial data for the stress. Since the exact solution for this problem is known [5], we use it to compare the different schemes between them.

When the velocity has a constant sign (Figure 6) or has a variable sign with an increasing initial data (Figure 7), the C-L scheme and the new schemes give similar numerical solutions.

When $u_l > 0 > u_r$, one can remark that the solution obtained with the C-L scheme is not valid, because none of the intermediate step values is reached (see Figure 8), whereas with the SHS and the NSHS schemes the intermediate step values for u and σ are obtained precisely (see Figure 9).

Because the SHS scheme gives the same solution as the NSHS scheme does, the convergence analysis of the new schemes is done only for the NSHS scheme. For this purpose, only the solution presenting two shocks is considered ($u_l > 0 > u_r$).

In this system modeling elasticity, only the momentum is conserved. Then, as the mass is constant (equals one), one can see that the L^1-norm of the velocity error (equal to the conservation error of the momentum) is reduced by at least one order of magnitude as the mesh decreased (see Figures 10 and 11). One can also observe that the error of the intermediate step value of the stress and that of the velocity converge towards zero (see Figures 11 and 12).

6.2. Numerical results for the system modeling hydrodynamics. Let us denote by AS the exact solution computed with the software CLAWPACK [10], solving the Euler equation in the conservative formulation. The computations are done for different initial data, given by Table 6.1 with $\gamma = 1.4$.

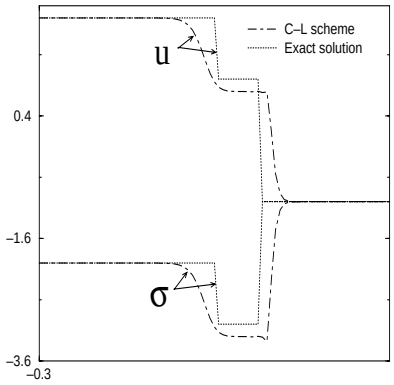


FIG. 8. *Solution of the Riemann problem associated with the system* (1.1), *when* $(u_l, u_r) = (2, -1)$.

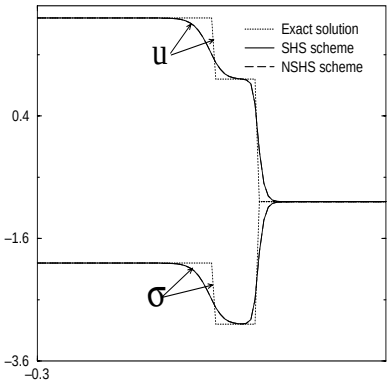


FIG. 9. *Solution of the Riemann problem associated with the system* (1.1), *when* $(u_l, u_r) = (2, -1)$.

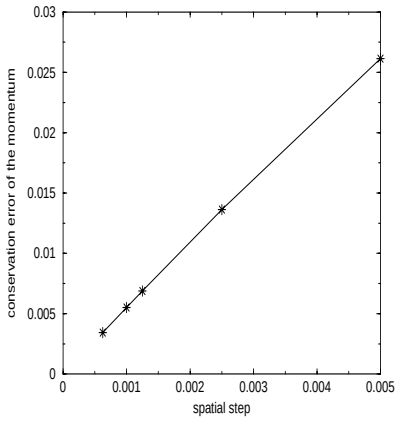


FIG. 10. *Decrease of the momentum conservation error,* $T = 0.1$.

h	$\|u_h - u_e\|_{L^1}$	$E(\sigma)$	$E(u)$
h_0	$4.7366E^{-2}$	0.02	0.04
$\frac{h_0}{2}$	$2.6146E^{-2}$	210^{-3}	2.10^{-3}
$\frac{h_0}{4}$	$1.3633E^{-2}$	310^{-5}	6.10^{-5}
$\frac{h_0}{8}$	$6.8765E^{-3}$	310^{-9}	7.10^{-10}
$\frac{h_0}{10}$	$5.5030E^{-3}$	10^{-10}	10^{-10}
$\frac{h_0}{16}$	$3.4397E^{-3}$	10^{-10}	10^{-10}

FIG. 11. *Relative error in the intermediate value versus mesh size.* u_h *is the computed momentum and* u_e *the exact momentum,* $T = 0.1$.

Test 1 is the so-called Sod test problem [14]. The other tests, test 2 and test 3, are more general but have been proposed with an aim of finding the better scheme. It is known that the Riemann solution problem of test 1 presents, successively, a left rarefaction wave, a contact discontinuity, and a right shock wave. While the solution of test 2 presents a left and a right rarefaction wave separated by a contact discontinuity. Test 3 has a solution consisting of two shock waves separated by a contact discontinuity.

The Figures 13, 15, and 17 present the solutions of test 1, computed by the Godunov-type schemes (NSHS, SHS, and C-L). Figures 14, 16, and 18 present the solution of test 1 computed by the modified Godunov schemes, obtained by a conservative projection. (We call them the NSHS2, SHS2, and C-L2 schemes.)

One can remark that the solutions of test 1 are similar, though the intermediate step value for the density and the velocity is calculated with more accuracy with the

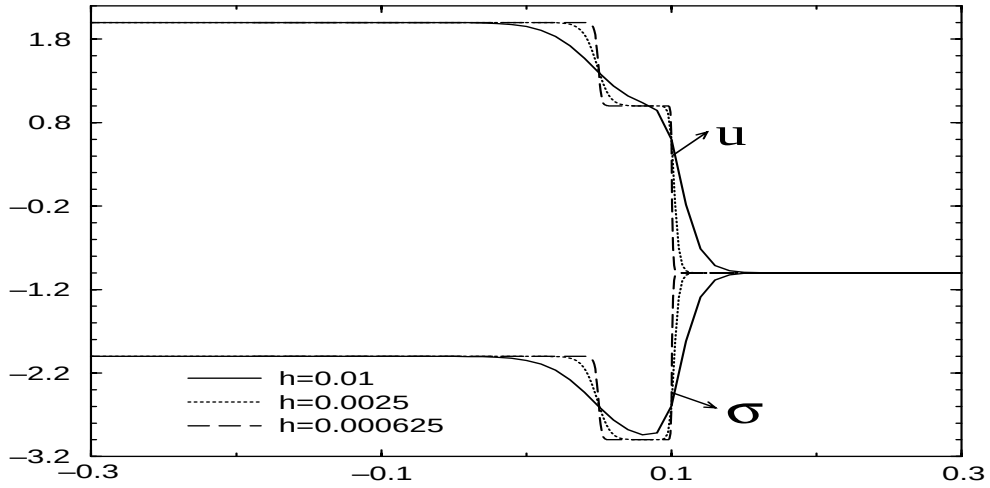


FIG. 12. *Numerical solutions of the Riemann problem with the NSHS scheme for different spatial step,* $(u_l, \sigma_l) = (2, -2)$ *and* $(u_r, \sigma_r) = (-1, -1)$*,* $T = 0.1$*.*

TABLE 6.1
Initial data for the system (1.6).

Initial data	(u_l, u_r)	(p_l, p_r)	(ρ_l, ρ_r)
Test 1	$(0.0, 0.0)$	$(1.0, 0.1)$	$(1.0, 0.125)$
Test 2	$(-1.0, 1.0)$	$(30.0, 10.0)$	$(10.0, 10.0)$
Test 3	$(2.0, -1.0)$	$(30.0, 10.0)$	$(5.0, 5.0)$

schemes using the conservative projection than the others.

We keep the same presentation for the other test problems. Figures 19, 21, and 23 (resp., Figures 25, 27, and 29) are proposed in order to compare the different Godunov-type schemes while the Figures 20, 22, and 24 (resp., Figures 26, 28, and 30) present the different conservative projection schemes.

For test 2, one observes the same behavior in each of the two splitted scheme families. For test 3, one remarks that the schemes using a conservative projection permit us to reach more precisely the intermediate step values. Moreover, this test shows how difficult it is to approximate the density (see Figure 29) by a Godunov-type scheme in the nonconservative formulation. Nevertheless, the new approach by the NSHS2 or SHS2 schemes gives us good enough results while the C-L2 scheme (Figure 30) is significantly less accurate than the new schemes. This phenomenon has already been observed in elasticity when the velocity has the same initial data. As the C-L scheme did not satisfy the CFL condition, we have computed the solution after 100 time steps with a smaller parameter $r = 0.12$ than that used for the other tests.

Unfortunately, one remarks that for each of the tests, the numerical solution computed either with the NSHS2 scheme or with the NSHS scheme presents a surplus of diffusion. And although the NSHS2 scheme is more accurate than the NSHS scheme, the two shocks of test 3 are less well captured when using the NSHS2 scheme (see Figures 25, 27, 29, and 26, 28, 30).

Using the "antidiffusion" process proposed by Colombeau and Le Roux, one reduces the amount of diffusion but the intermediate step values are not reached (cf. [6]).

Following the example of Karni in [9], we conducted a sequence of shock tube tests

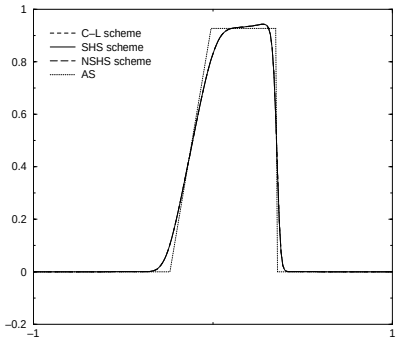


FIG. 13. *Velocity computed by the Godunov scheme for test* 1.

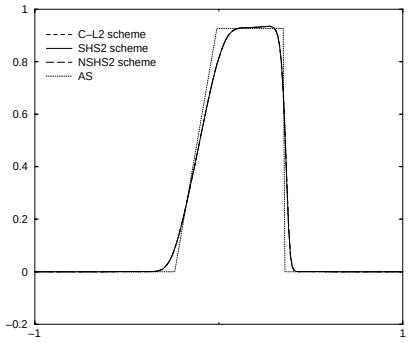


FIG. 14. *Velocity computed by the conservative projection scheme for test* 1.

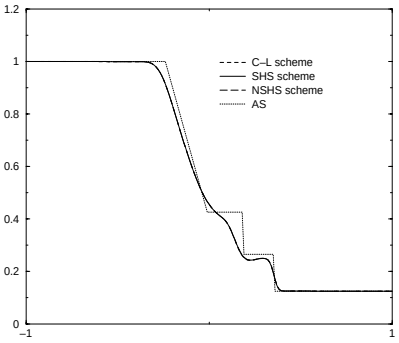


FIG. 15. *Density computed by the Godunov scheme for test* 1.

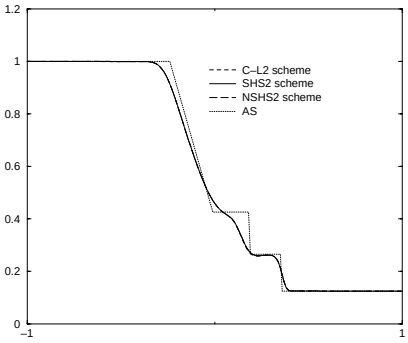


FIG. 16. *Density computed by the conservative projection scheme for test* 1.

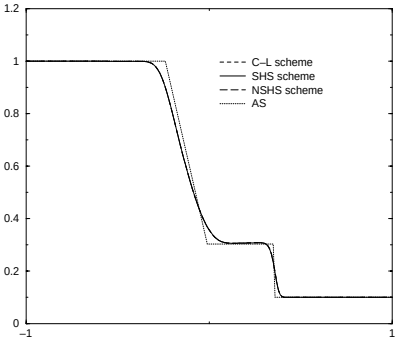


FIG. 17. *Pressure computed by the Godunov scheme for test* 1.

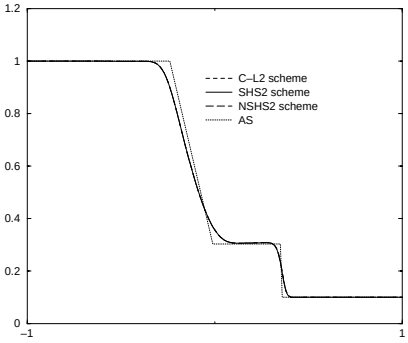


FIG. 18. *Pressure computed by the conservative projection scheme for test* 1.

for shock waves of varying strengths involving a single shock. We kept the initial right value $(\rho_r\,, u_r\,, P_r) = (0.125,\, 0.0,\, 0.1)$ and we chose different left values of the initial data for defining a range of shock Mach numbers approximately between 1.1 and 3.1. From Table 6.2, one observes that the conservation error is smaller for the split scheme using the conservative projection than those induced by the split Godunov scheme. Furthermore, the relative error is between 0.6 and 4% for the tests with the NSHS2 scheme, whereas with the NSHS scheme the error varies between 1.6 and 6%.

From Figures 31 and 32, we note that as the mesh decreases, conservation er-

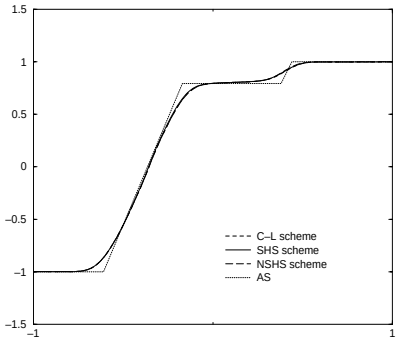


FIG. 19. *Velocity computed by the Godunov scheme for test* 2.

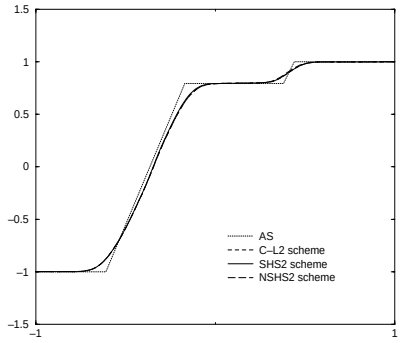


FIG. 20. *Velocity computed by the conservative projection for test* 2.

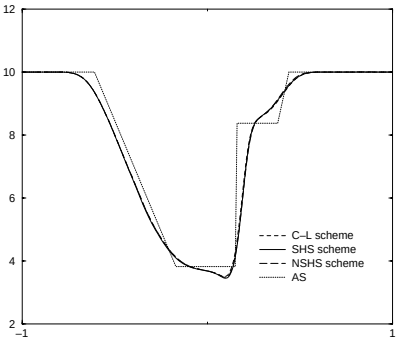


FIG. 21. *Density computed by the Godunov scheme for test* 2.

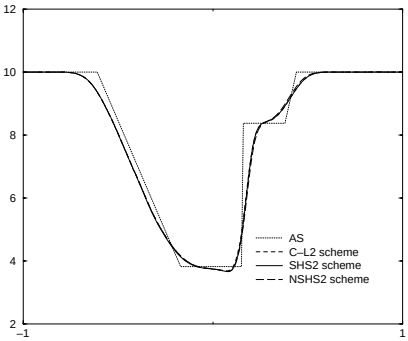


FIG. 22. *Density computed by the conservative projection scheme for test* 2.

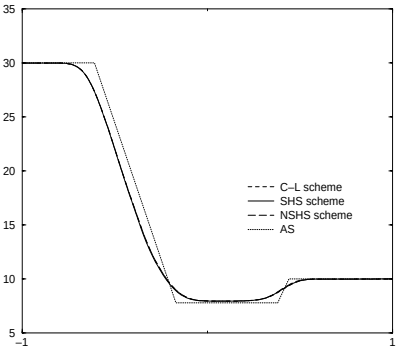


FIG. 23. *Pressure computed by the Godunov scheme for test* 2.

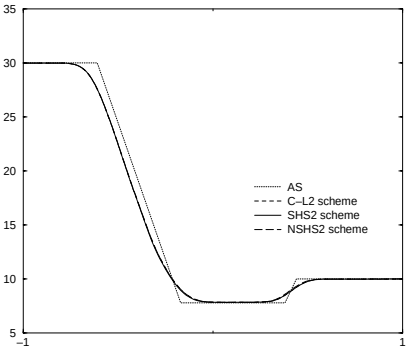


FIG. 24. *Pressure computed by the conservative projection scheme for test* 2.

rors decrease almost linearly but are not eliminated. Moreover, we remark that the limit value (with the mesh refinement) of the conservation errors depends on the shock strength. Indeed, the stronger the shock is, the bigger the limit value of the conservation errors becomes.

7. Concluding remarks. We have shown that in the case $u_l.u_r > 0$ our schemes give the same solution as the C-L or C-L2 schemes. In fact, the formula defining the approximate solution of each scheme is the same. But when the sign of

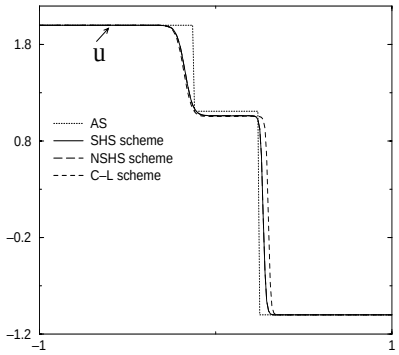


FIG. 25. *Velocity computed by the Godunov scheme for test* 3.

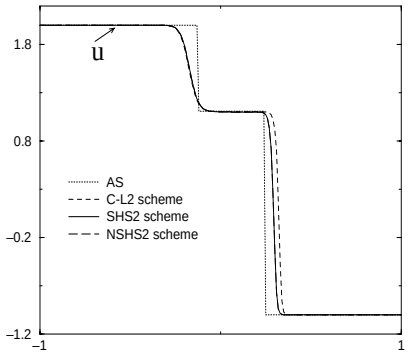


FIG. 26. *Velocity computed by the conservative projection scheme for test* 3.

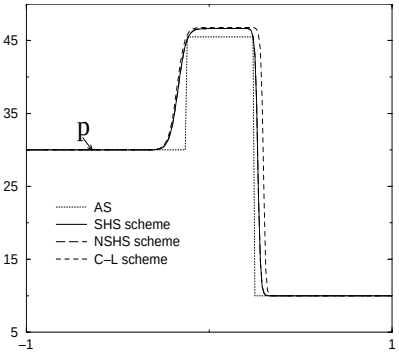


FIG. 27. *Pressure computed by the Godunov scheme for test* 3.

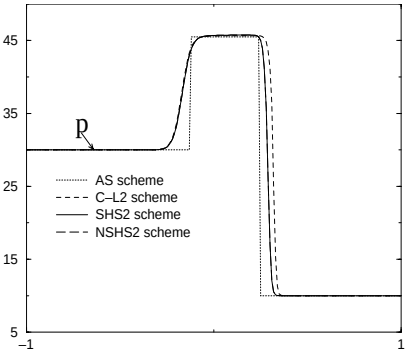


FIG. 28. *Pressure computed by the conservative projection scheme for test* 3.

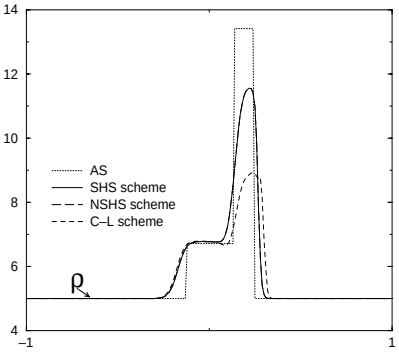


FIG. 29. *Density computed by the Godunov scheme for test* 3.

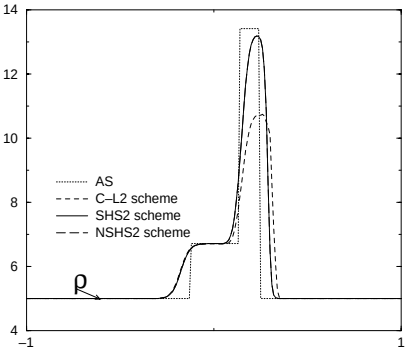


FIG. 30. *Density computed by the conservative projection scheme for test* 3.

the velocity changes, one observes two phenomena.

In the context of increasing initial data, the approximate solutions of the three schemes coincide, even though the scheme expressions are different.

We also have in the context of decreasing initial data that the approximate solution computed by our scheme is valid, whereas the C-L scheme does not permit us to calculate an acceptable solution.

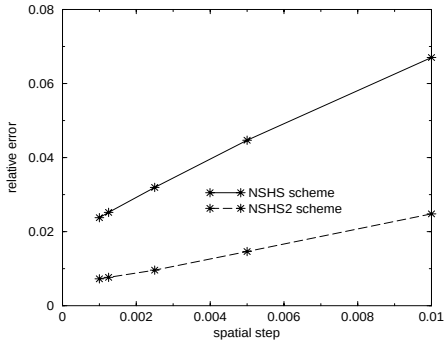


FIG. 31. *Conservation errors versus mesh size for the Sod tube test. The Mach shock number is* $Ms = 1.6556$, *the exact pressure behind the shock is* $P_e = 0.30313$. *the error belongs to the computed intermediate pressure value.*

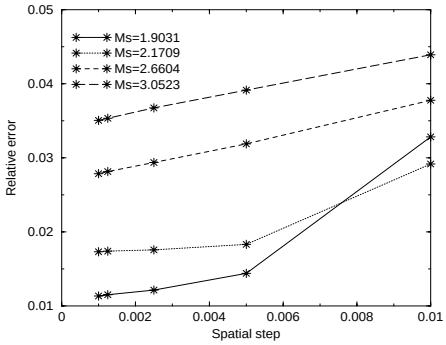


FIG. 32. *Conservation errors versus mesh size for NSHS*2 *scheme. The Mach shock numbers vary between* 1.5 *and* 3.1, *the error belongs to the computed intermediate pressure value.*

TABLE 6.2

Conservation errors versus shock strengths for shock tube tests. M_s *is the Mach number;* P_e *is the exact pressure behind the shock.* P_{NSHS} *and* P_{NSHS2} *are the intermediate pressure values computed by the NSHS and NSHS*2 *schemes;* $E(.)$ *are relative errors.* $h = h_0$, $T = 0.2$, $(\rho_r\,, u_r\,, P_r) =$ (0.125, 0.0, 0.1).

M_s	P_e	P_{NSHS2}	P_{NSHS}	$E(P_{NSHS2})$	$E(P_{NSHS})$
1.1905	0.1487	0.1497	0.1512	0.0068	0.0168
1.5932	0.2795	0.2826	0.2870	0.0111	0.0268
1.8955	0.4025	0.4090	0.4193	0.0161	0.0417
2.1709	0.5332	0.5450	0.5584	0.0221	0.0473
2.6604	0.8090	0.8354	0.8571	0.0325	0.0593
3.0523	1.0702	1.1127	1.1385	0.0396	0.0637

We prove that split Godunov schemes are more efficient when they use the conservative projection step. This is confirmed by conducting convergence tests with regard to the conservation errors. One can develop well-adapted splitted schemes by proceeding in the same way for other systems in nonconservative form—for example, nonconservative systems modeling elastoplasticity or multifluids.

Acknowledgments. The authors are pleased to acknowledge J. F. Colombeau, A.-Y. Le Roux, and A. Méril for their helpful advice and suggestions. They also want to thank an unknown referee who suggested adding a precise convergence analysis.

REFERENCES

[1] R. BARAILLE, G. BOURDIN, F. DUBOIS, AND A.-Y. LE ROUX, *Une version à pas fractionnaire du schéma de godunov pour l'hydrodynamique*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 147–152.

[2] H. BIAGIONI, *A Nonlinear Theory of Generalized Functions*, Lecture Notes in Math. 1421, Springer-Verlag, Berlin, 1990.

[3] J. CAURET, *Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations*, J. Math. Anal. Appl., 139 (1989), pp. 552–573.

[4] J.-F. COLOMBEAU, *New Generalized functions and Multiplication of Distributions*, North-Holland Math. Stud. 84, North-Holland, Amsterdam, 1984.

[5] J.-F. COLOMBEAU, *Multiplication of Distributions*, Lecture Notes in Math. 1532, Springer-Verlag, Berlin, 1992.

[6] J.-F. Colombeau and A.-Y. Leroux, *Numerical methods for hyperbolic systems in nonconservative form using product of distributions*, in Advances in Computer Methods for Partial Differential Equation VI, IMACS, North-Holland, Amsterdam, 1987, pp. 28–37.

[7] J.F. Colombeau, A.Y. Le Roux, A. Noussair, and B. Perrot, *Microscopic profiles of shock waves and ambiguities in multiplications of distributions*, SIAM J. Numer. Anal., 26 (1989), pp. 871–883.

[8] E. Godlewski and P. A. Raviard, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.

[9] S. Karni, *Viscous shock profiles and primitive formulations*, SIAM J. Numer. Anal., 26 (1992), pp. 1592–1609.

[10] R. LeVeque, J. Langseth, M. Berger, and S. Mitran, CLAWPACK. *Conservation Law Package,* http://www.amath.washington.edu/~claw/.

[11] M. Oberguggenberger, *Case study nonlinear non conservative non strictly hyperbolic system*, Nonlinear Anal., 19 (1992), pp. 53–79.

[12] M. Oberguggenberger, *Multiplication of Distributions and Applications to Partial Differential Equations*, Pitman Research Notes in Math. Series 259, Longman Scientific and Technical, Harlow, UK, 1992.

[13] L. Remaki, *Etude théorique et numérique d'équations quasi-linéaires scalaires à coefficients discontinus de l'acoustique linéaire* $2D$, Ph.D. thesis, Université Lyon I, Lyon, France, 1997.

[14] G. Sod, *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws*, J. Comput. Phys., 27 (1978), pp. 1–31.

[15] E. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 2nd ed., Springer-Verlag, Berlin, 1999.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 26–40

ORDER ESTIMATES IN TIME OF SPLITTING METHODS FOR THE NONLINEAR SCHRÖDINGER EQUATION*

CHRISTOPHE BESSE†, BRIGITTE BIDÉGARAY‡, AND STÉPHANE DESCOMBES§

Abstract. In this paper, we consider the nonlinear Schrödinger equation $u_t + i\Delta u - F(u) = 0$ in two dimensions. We show, by an operator-theoretic proof, that the well-known Lie and Strang formulae (which are splitting methods) are approximations of the exact solution of order 1 and 2 in time.

Key words. nonlinear Schrödinger equation, splitting methods

AMS subject classifications. 65M12, 35Q

PII. S0036142900381497

1. Introduction.

Let us consider the cubic nonlinear Schrödinger equation

$$\begin{cases} \dfrac{\partial u}{\partial t} + i\Delta u + i\varepsilon|u|^2 u = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}^2, \end{cases} \tag{1.1}$$

with $\varepsilon = \pm 1$. A large number of articles are devoted to the numerical study of this equation using many different time discretizations, with or without splitting. The later case is represented by Crank–Nicolson type [4], Runge–Kutta type [1], [12], symplectic (see, for example, [14], [15]), and relaxation [2] methods. Splitting methods are based on a decomposition of the flow of (1.1). More precisely, let us define the flow X^t of the linear Schrödinger equation

$$\begin{cases} \dfrac{\partial v}{\partial t} + i\Delta v = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ v(x,0) = v_0(x), & x \in \mathbb{R}^2, \end{cases}$$

and the flow Y^t for the differential equation

$$\begin{cases} \dfrac{\partial w}{\partial t} + i\varepsilon|w|^2 w = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ w(x,0) = w_0(x), & x \in \mathbb{R}^2. \end{cases}$$

The idea of splitting methods is to approximate the flow of (1.1) by combining the two flows X^t and Y^t. Two classical methods are the following: the Lie formula given by $Z_{\mathrm{L}}^t = X^t Y^t$ (or $Y^t X^t$) and the Strang formula [18] $Z_{\mathrm{S}}^t = X^{t/2} Y^t X^{t/2}$ (or $Y^{t/2} X^t Y^{t/2}$); we introduce these four definitions since it is sometimes better to exchange the role of X^t and Y^t when one of the two equations is nonsmooth [17].

*Received by the editors November 21, 2000; accepted for publication (in revised form) October 16, 2001; published electronically April 12, 2002.
http://www.siam.org/journals/sinum/40-1/38149.html

†Laboratoire de Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France (besse@mip.ups-tlse.fr).

‡Laboratoire de Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, INSA de Toulouse, 135 avenue de Rangeuil, 31077 Toulouse Cedex 4, France (bidegara@mip.ups-tlse.fr).

§Unité de Mathématiques Pures et Appliquées, CNRS UMR 5669, Ecole normale supérieure de Lyon, 46, Allée d'Italie, 69364 Lyon Cedex 07, France (sdescomb@umpa.ens-lyon.fr).

This leads to good numerical methods for the periodic problem since the linear part may be computed efficiently by the use of fast Fourier transforms and the nonlinear part is solved exactly [19], [20]. We are interested in showing that the Lie formula is a first order approximation of the flow of (1.1) and the Strang formula is a second order approximation of the flow of (1.1). This result could be obtained formally with the formal Lie algebra theory (explained in the book [14] and in [13]), but here we give a simple proof allowing us to have an idea of the size of the constants.

The linear case has already been studied in [11] and [7] and we extend these results to the nonlinear case.

Following an idea of Donnat [8], we restrict ourselves to the case where the nonlinearity is a Lipschitz function; this may be done by a truncation method on a time interval before a possible blow-up. Thus we consider u the solution to the continuous problem

$$\left\{\begin{array}{ll} \dfrac{\partial u}{\partial t} + i\Delta u - F(u) = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}^2, \end{array}\right. \tag{1.2}$$

where we assume that F is a Lipschitz function with constant K such that $F(0) = 0$ and the first four derivatives of F are bounded. We introduce the flow S^t, associated with (1.2) (that is, $u(t,\cdot) = S^t u_0$), and the two flows X^t and Y^t, solutions to

$$\left\{\begin{array}{ll} \dfrac{\partial v}{\partial t} + i\Delta v = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ v(x,0) = v_0(x), & x \in \mathbb{R}^2, \end{array}\right. \tag{1.3}$$

and

$$\left\{\begin{array}{ll} \dfrac{\partial w}{\partial t} - F(w) = 0, & x \in \mathbb{R}^2\,,\ t > 0, \\ w(x,0) = w_0(x), & x \in \mathbb{R}^2. \end{array}\right. \tag{1.4}$$

In what follows, we call Z^t any of the four splitting schemes when there is no ambiguity. Let us also recall that the semigroup X^t is a unitary operator on all classical Sobolev spaces $H^s = H^s(\mathbb{R}^2)$, $s \in \mathbb{R}$. Let us quote the main result of this article.

THEOREM 4.1. *For all u_0 in H^2 and for all $T > 0$, there exists C and h_0 such that for all $h \in (0, h_0]$, for all n such that $nh \leq T$*

$$\left\| \left(Z_L^h\right)^n u_0 - S^{nh} u_0 \right\| \leq C(\|u_0\|_{H^2}) h \|u_0\|_{H^2}.$$

Moreover, if u_0 belongs to H^4, then

$$\left\| \left(Z_S^h\right)^n u_0 - S^{nh} u_0 \right\| \leq C(\|u_0\|_{H^4}) h^2 \|u_0\|_{H^4}.$$

To prove the convergence order for each splitting scheme, for a small $h > 0$ and all integer n such that $nh \leq T$, we have to estimate the quantity $\|(Z^h)^n u_0 - S^{nh} u_0\|$, where $\|\cdot\|$ denotes the L^2 norm. As noticed in [5], the triangle inequality yields

$$\|(Z^h)^n u_0 - S^{nh} u_0\| \leq \sum_{j=0}^{n-1} \|(Z^h)^{n-j-1} Z^h S^{jh} u_0 - (Z^h)^{n-j-1} S^{(j+1)h} u_0\|.$$

In section 3 we prove that for all the studied schemes there exists a constant C_0 such that for w_0 and $w_0' \in L^2$ and all $t \in [0,1]$

$$\|Z^t w_0 - Z^t w_0'\| \leq (1 + C_0 t)\|w_0 - w_0'\|. \tag{1.5}$$

Therefore

$$\|(Z^h)^n u_0 - S^{nh} u_0\| \leq \sum_{j=0}^{n-1} (1 + C_0 h)^{n-j-1} \left\|(Z^h - S^h) S^{jh} u_0\right\|. \tag{1.6}$$

Thus we may restrict our study to the case for which at each time step the initial data are the same for the continuous model and the splitting scheme and is equal to $v_0 = S^{jh} u_0$. Classical results on solutions to the nonlinear Schrödinger equation allow us to state that $S^{jh} u_0$ is uniformly bounded in H^4 for $jh \leq T$. Now we may write a Duhamel formula for the continuous problem (1.2) that reads as

$$u(t) = X^t v_0 + \int_0^t X^{t-s} F(u(s))\, ds$$

and express the difference of the exact solution and the splitting solution $v(t) = Z^t v_0$ as

$$u(t) - v(t) = \int_0^t X^{t-s} \left[F(u(s)) - F(v(s))\right] ds + R(t),$$

where the fact that F is Lipschitz and X^t is unitary in L^2 leads to

$$\|u(t) - v(t)\| \leq K \int_0^t \|u(s) - v(s)\|\, ds + \|R(t)\|.$$

There remains to show that the remainder $R(t)$ may be estimated as $\|R(t)\| = O(t^{p+1})$ for t small and to use a Gronwall lemma to conclude that the scheme is of order p.

This paper is organized as follows: In section 2, we prove a Gronwall lemma and some estimates on X^t and Y^t. In section 3, we show that each scheme is Lipschitz continuous and we study the local error between Z^t and S^t. Section 4 is devoted to the proof of Theorem 4.1.

2. Some useful estimates.

2.1. A Gronwall lemma.

LEMMA 2.1 (Gronwall). *Let P be a polynomial with positive coefficients and no constant term. We assume that the function ϕ is such that there exists a constant $C \geq 0$ such that for all $t \geq 0$*

$$0 \leq \phi(t) \leq \phi(0) + P(t) + C \int_0^t \phi(s) ds.$$

Then for all $\alpha > 1$ there exists $t_0 > 0$ such that for all $0 \leq t \leq t_0$

$$\phi(t) \leq \phi(0) e^{Ct} + \alpha P(t).$$

Proof. Let us set

$$\psi(t) = \left(\phi(0) + P(t) + C \int_0^t \phi(s) ds\right) e^{-Ct}.$$

Then

$$\psi'(t) = \left(P'(t) + C\phi(t) - C\left(\phi(0) + P(t) + C\int_0^t \phi(s)ds\right)\right) e^{-Ct} \leq P'(t)e^{-Ct};$$

therefore,

$$\psi(t) - \psi(0) \leq \int_0^t P'(s)e^{-Cs}ds,$$

and since $P(0) = 0$, $\psi(0) = \phi(0)$. Hence, because P' is positive,

$$\phi(t) \leq \psi(t)e^{Ct} \leq \phi(0)e^{Ct} + \int_0^t P'(s)e^{C(t-s)}\ ds \leq \phi(0)e^{Ct} + e^{Ct_0}\int_0^t P'(s)\ ds.$$

We choose t_0 such that $e^{Ct_0} \leq \alpha$ and, for all $0 \leq t \leq t_0$,

$$\phi(t) \leq \phi(0)e^{Ct} + \alpha P(t). \qquad \square$$

2.2. Estimates on the Schrödinger flow X^t. From the definition of the Schrödinger flow we first state that

$$\dot{X}^t = i\Delta X^t = iX^t\Delta. \tag{2.1}$$

This leads to the following estimates.

LEMMA 2.2. 1. *For all $w \in H^2$ and all $t \geq 0$,*

$$\|X^t w - w\| \leq t\|w\|_{H^2}. \tag{2.2a}$$

2. *For all $w \in H^4$ and all $t \geq 0$,*

$$\|X^t w - w\|_{H^2} \leq t\|w\|_{H^4}. \tag{2.2b}$$

3. *Let $T > 0$; there exists a constant C such that, for all $w \in \mathcal{C}^1([0,T];H^2) \cap L^\infty([0,T],H^4)$ and $0 \leq t \leq T$,*

$$\left\|\int_0^t \left(X^{t-s}w(s) - X^{t/2}w(s)\right) ds\right\| \leq Ct^3(\|w\|_{\mathcal{C}^1([0,T];H^2)} + \|w\|_{L^\infty([0,T],H^4)}). \tag{2.2c}$$

4. *There exists a constant C such that for all $w \in H^4$,*

$$\left\|X^{t/2}w - \frac{1}{2}X^t w - \frac{1}{2}w\right\| \leq Ct^2\|w\|_{H^4}. \tag{2.2d}$$

Proof. 1. Let $w \in H^2$; we have

$$\|X^t w - w\| = \left\|\int_0^t \dot{X}^s w\ ds\right\| = \left\|\int_0^t X^s\Delta w\ ds\right\| \leq \int_0^t \|\Delta w\|ds \leq t\|w\|_{H^2}.$$

2. If we assume that $w \in H^4$, the estimate may be proved as the previous one replacing the L^2 norm by the H^2 norm.

3. A Taylor expansion gives

$$X^{t-s} - X^{t/2} = (t/2 - s)\dot{X}^{t/2} + \int_{t/2}^{t-s}(t - s - \sigma)\ddot{X}^\sigma\ d\sigma$$

and

$$(X^{t-s} - X^{t/2})w(s) = i(t/2-s)X^{t/2}\Delta w(s) - \int_{t/2}^{t-s}(t-s-\sigma)X^{\sigma}\Delta^2 w(s)\ d\sigma.$$

A simple change of variables implies that

$$\int_0^t (t/2-s)\Delta w(s)\ ds = \int_0^{t/2}(t/2-s)[\Delta w(s) - \Delta w(t-s)]\ ds.$$

The Lipschitz constant of the map $s \mapsto \Delta w(s)$ is estimated using $\|w\|_{\mathcal{C}^1([0,T],H^2)}$ and, therefore,

$$\begin{aligned}
&\left\| \int_0^t \left(X^{t-s}w(s) - X^{t/2}w(s) \right) ds \right\| \\
&\quad \le \left\| \int_0^{t/2} (t/2-s)[\Delta w(s) - \Delta w(t-s)]\ ds \right\| + \left\| \int_0^t \int_{t/2}^{t-s} (t-s-\sigma)X^{\sigma}\Delta^2 w(s) d\sigma ds \right\| \\
&\quad \le 2\|w\|_{\mathcal{C}^1([0,T],H^2)} \int_0^{t/2} (t/2-s)^2 ds + \|w\|_{L^\infty([0,T],H^4)} \int_0^t \int_{t/2}^{t-s} (t-s-\sigma)\ d\sigma\ ds \\
&\quad \le Ct^3(\|w\|_{\mathcal{C}^1([0,T],H^2)} + \|w\|_{L^\infty([0,T],H^4)}).
\end{aligned}$$

4. Once more, Taylor expansions yield

$$X^{t/2} - \frac{1}{2}X^t - \frac{1}{2}X^0 = -\frac{1}{2}\int_0^{t/2} \sigma(\ddot{X}^{\sigma} + \ddot{X}^{t-\sigma})\ d\sigma,$$

and the same arguments as for the last estimates show the result. ∎

2.3. Estimates on the nonlinear flow Y^t. The definition of the nonlinear flow Y^t may also read as

$$Y^t w = w + \int_0^t F(Y^s w)\ ds. \tag{2.3}$$

LEMMA 2.3. *Let $w \in H^2$; then there exists a constant C that depends only on $M = \|w\|_\infty$ such that for all $0 \le t \le 1$*

$$\|Y^t w\| \le e^{Kt}\|w\| \text{ and } \|Y^t w\|_{H^2} \le C\|w\|_{H^2}. \tag{2.4a}$$

Moreover, if $w \in H^4$, then there exists a constant C that depends only on $M = \|w\|_\infty$ such that for all $0 \le t \le 1$

$$\|Y^t w\|_{H^4} \le C\|w\|_{H^4}. \tag{2.4b}$$

Finally, for w_1, $w_2 \in L^2$, there exists a constant C that depends only on F such that for all $0 \le t \le 1$

$$\|Y^t w_1 - Y^t w_2\| \le (1+Ct)\|w_1 - w_2\|. \tag{2.4c}$$

Proof. Equation (2.3) first yields a L^∞ estimate, namely,

$$\|Y^t w\|_\infty \le \|w\|_\infty + K\int_0^t \|Y^s w\|_\infty\ ds.$$

Then the classical Gronwall lemma leads to

$$\|Y^t w\|_\infty \le e^{Kt}\|w\|_\infty.$$

A L^2 estimate also follows from (2.3):

$$\|Y^t w\| \le \|w\| + K\int_0^t \|Y^s w\|\ ds. \tag{2.5}$$

For all first order differential operators D

$$DY^t w = Dw + \int_0^t F'(Y^s w) D(Y^s w)\ ds,$$

and, denoting by M' the maximum for F', we obtain

$$\|DY^t w\| \le \|Dw\| + M'\int_0^t \|DY^s w\|\ ds.$$

Differentiating once more,

$$\Delta Y^t w = \Delta w + \int_0^t \left(F''(Y^s w) D(Y^s w)^2 + F'(Y^s w)\Delta Y^s w\right) ds,$$

and, denoting by M'' the maximum for F'',

$$\|\Delta Y^t w\| \le \|\Delta w\| + \int_0^t \left(M''\|DY^s w\|^2 + M'\|\Delta Y^s w\|\right) ds.$$

Using the Gagliardo–Nirenberg inequality,

$$\|DY^s w\|^2 \le \|Y^s w\|_{H^2}\|Y^s w\|_\infty$$

and

$$\|\Delta Y^t w\| \le \|\Delta w\| + \int_0^t \left(M''\|Y^s w\|_\infty + M'\right)\|Y^s w\|_{H^2}\ ds.$$

Therefore, using the L^∞ estimate, there exists a constant c such that

$$\|Y^t w\|_{H^2} \le \|w\|_{H^2} + c\int_0^t \left(1 + e^{Ks}\right)\|Y^s w\|_{H^2}\ ds.$$

Last, using the Gronwall lemma,

$$\|Y^t w\|_{H^2} \le \|w\|_{H^2} \exp\left(c\int_0^t \left(1 + e^{Ks}\right) ds\right).$$

Equation (2.5) also leads to

$$\|Y^t w\| \le e^{Kt}\|w\|.$$

For $t \le 1$, there exists a constant C such that

$$\exp\left(c\int_0^t \left(1 + e^{Ks}\right) ds\right) \le C$$

and estimate (2.4a) follows. The proof for (2.4b) is similar and left to the reader. Finally, estimate (2.4c) is a simple consequence of the Gronwall lemma. □

3. Lipschitz properties of Z^t and local errors. In this section we more specifically give precise estimates for Lie and Strang formulae. We first show Lipschitz properties on Z^t, i.e., that estimate (1.5) is valid. Next we estimate the remainder $R(t)$ defined in the introduction.

3.1. Lipschitz properties.

- *Lie approximation—case $Z^t = X^t Y^t$.*
  The solution to the Lie approximation with initial data $v_0 \in L^2$ reads as

$$v(t) = Z^t v_0 = X^t v_0 + \int_0^t X^t F(Y^s v_0)\, ds.$$

  Therefore the difference between two solutions for initial data w_0 and w_0' in L^2 is

$$Z^t w_0 - Z^t w_0' = X^t(w_0 - w_0') + \int_0^t X^t (F(Y^s w_0) - F(Y^s w_0'))\, ds,$$

  and using the fact that X^t is unitary in L^2, that F is Lipschitz, and estimate (2.4c), we obtain that there exists a constant C depending only on F such that for $0 \le t \le 1$

$$\|Z^t w_0 - Z^t w_0'\| \le (1 + Ct)\|w_0 - w_0'\|.$$

- *Lie approximation—case $Z^t = Y^t X^t$.*
  Since

$$v(t) = Z^t v_0 = X^t v_0 + \int_0^t F(Y^s X^t v_0)\, ds,$$

  the difference is

$$Z^t w_0 - Z^t w_0' = X^t(w_0 - w_0') + \int_0^t (F(Y^s X^t w_0) - F(Y^s X^t w_0'))\, ds;$$

  thus, using the same tools as above, we obtain that there exists a constant C depending only on F such that for $0 \le t \le 1$

$$\|Z^t w_0 - Z^t w_0'\| \le (1 + Ct)\|w_0 - w_0'\|.$$

- *Strang approximation—case $Z^t = X^{t/2} Y^t X^{t/2}$.*
  Since

$$v(t) = Z^t v_0 = X^t v_0 + \int_0^t X^{t/2} F(Y^s X^{t/2} v_0)\, ds,$$

  we have

$$\begin{aligned} Z^t w_0 - Z^t w_0' &= X^t(w_0 - w_0') \\ &\quad + \int_0^t X^{t/2}(F(Y^s X^{t/2} w_0) - F(Y^s X^{t/2} w_0'))\, ds, \\ \|Z^t w_0 - Z^t w_0'\| &\le (1 + Ct)\|w_0 - w_0'\|. \end{aligned}$$

- *Strang approximation—case* $Z^t = Y^{t/2}X^tY^{t/2}$.
  Since

$$v(t) = Z^t v_0 = X^t Y^{t/2} v_0 + \int_0^{t/2} F(Y^s X^t Y^{t/2} v_0)\, ds,$$

we have

$$\begin{aligned} Z^t w_0 - Z^t w_0' &= X^t Y^{t/2} w_0 - X^t Y^{t/2} w_0' \\ &\quad + \int_0^{t/2} (F(Y^s X^t X^{t/2} w_0) - F(Y^s X^t Y^{t/2} w_0'))\, ds, \\ \|Z^t w_0 - Z^t w_0'\| &\le (1 + Ct)\|w_0 - w_0'\|. \end{aligned}$$

3.2. Local errors.

- *Lie approximation—case* $Z^t = X^tY^t$.
  For $v_0 \in H^2$ and $0 \le t \le 1$, the remainder can be written as

$$R(t) = \int_0^t X^{t-s} F(X^s Y^s v_0)\, ds - \int_0^t X^t F(Y^s v_0)\, ds.$$

Let us define $R_1(s) = F(X^sY^sv_0) - X^sF(Y^sv_0)$; then, using the fact that F is Lipschitz and estimates (2.2a) and (2.4a),

$$\begin{aligned} R_1(s) &= F(X^sY^sv_0) - F(Y^sv_0) + F(Y^sv_0) - X^sF(Y^sv_0), \\ \|R_1(s)\| &\le K\|X^sY^sv_0 - Y^sv_0\| + \|F(Y^sv_0) - X^sF(Y^sv_0)\| \\ &\le s\,(K\|Y^sv_0\|_{H^2} + \|F(Y^sv_0)\|_{H^2}) \\ &\le Cs\|v_0\|_{H^2}. \end{aligned}$$

Therefore, since $R(t) = \int_0^t X^{t-s}R_1(s)ds$,

$$\|R(t)\| \le C\|v_0\|_{H^2} \int_0^t s\, ds = \frac{Ct^2}{2}\|v_0\|_{H^2}.$$

- *Lie approximation—case* $Z^t = Y^tX^t$.
  For $v_0 \in H^2$ and $0 \le t \le 1$, the remainder can be written as

$$R(t) = \int_0^t X^{t-s} F(Y^s X^s v_0)\, ds - \int_0^t F(Y^s X^t v_0)\, ds.$$

In this case $R(t) = \int_0^t R_1(s)ds$, where $R_1 = X^{t-s}F(Y^sX^sv_0) - F(Y^sX^tv_0)$, and using the fact that F is Lipschitz and estimates (2.2a), (2.4a), (2.4c), we obtain

$$\begin{aligned} R_1(s) &= X^{t-s}F(Y^sX^sv_0) - F(Y^sX^sv_0) + F(Y^sX^sv_0) - F(Y^sX^tv_0), \\ \|R_1(s)\| &\le (t-s)\|F(Y^sX^sv_0)\|_{H^2} + K\|X^sv_0 - X^tv_0\| \\ &\le C(t-s)\|v_0\|_{H^2}; \end{aligned}$$

hence

$$\|R(t)\| \le C\|v_0\|_{H^2} \int_0^t (t-s)\, ds = \frac{Ct^2}{2}\|v_0\|_{H^2}.$$

- *Strang approximation—case* $Z^t = X^{t/2}Y^tX^{t/2}$.
  For $v_0 \in H^4$ and $0 \leq t \leq 1$, the remainder can be written as

$$R(t) = \int_0^t X^{t-s}F(X^{s/2}Y^sX^{s/2}v_0)\ ds - \int_0^t X^{t/2}F(Y^sX^{t/2}v_0)\ ds.$$

  We may write $R(t)$ as $R(t) = \int_0^t R_1(s)\ ds + X^{t/2}\int_0^t R_2(s)\ ds$, where

$$R_1(s) = X^{t-s}w(s) - X^{t/2}w(s), \qquad w(s) = F(Y^sX^{t/2}v_0),$$

  and

$$R_2(s) = F(X^{s/2}Y^sX^{s/2}v_0) - F(Y^sX^{t/2}v_0).$$

  Using estimate (2.2c), we obtain that

$$\left\| \int_0^t R_1(s)\ ds \right\| \leq Ct^3\|v_0\|_{H^4}.$$

  A Taylor expansion yields that

$$\begin{aligned} R_2(s) = F'(v_0)\cdot(X^{s/2}Y^sX^{s/2}v_0 - Y^sX^{t/2}v_0) \\ + \int_0^1 (1-\theta)\Big[F''(v_0+\theta(X^{s/2}Y^sX^{s/2}v_0 - v_0)) \\ \cdot (X^{s/2}Y^sX^{s/2}v_0 - v_0)^2 \\ -F''(v_0+\theta(Y^sX^{t/2}v_0 - v_0))\cdot(Y^sX^{t/2}v_0 - v_0)^2\Big]d\theta. \end{aligned}$$

  Using triangle inequalities, estimates (2.2a), (2.4a), formulation (2.3), and the fact that F is Lipschitz, we obtain that

$$\|X^{s/2}Y^sX^{s/2}v_0 - v_0\| \leq Cs\|v_0\|_{H^2}$$

  and

$$\|Y^sX^{t/2}v_0 - v_0\| \leq Ct\|v_0\|_{H^2}.$$

  Besides, we recall that F'' is uniformly bounded by M'', and therefore, using that H^2 is an algebra,

$$\begin{aligned} \left\| \int_0^1 (1-\theta)\Big[F''(v_0+\theta(X^{s/2}Y^sX^{s/2}v_0 - v_0))\cdot(X^{s/2}Y^sX^{s/2}v_0 - v_0)^2 \right. \\ \left. - F''(v_0+\theta(Y^sX^{t/2}v_0 - v_0))\cdot(Y^sX^{t/2}v_0 - v_0)^2\Big]d\theta \right\| \leq Ct^2\|v_0\|^2_{H^4}. \end{aligned}$$

  Moreover, let us define $R_3(s) = X^{s/2}Y^sX^{s/2}v_0 - Y^sX^{t/2}v_0$; formulation (2.3) yields that

$$\begin{aligned} R_3(s) = X^sv_0 - X^{t/2}v_0 + \int_0^s (X^{s/2}F(Y^\sigma X^{s/2}v_0) - F(Y^\sigma X^{s/2}v_0))\ d\sigma \\ + \int_0^s (F(Y^\sigma X^{s/2}v_0) - F(Y^\sigma X^{t/2}v_0))\ d\sigma. \end{aligned}$$

A simple change of variable in lemma (2.2c) proves that

$$\left\| \int_0^t (X^s v_0 - X^{t/2} v_0)\ ds \right\| \leq C t^3 \|v_0\|_{H^4},$$

and using once more estimates (2.2a) and (2.4b) and the fact that F is Lipschitz, we have

$$\left\| \int_0^t \int_0^s (X^{s/2} F(Y^\sigma X^{s/2} v_0) - F(Y^\sigma X^{s/2} v_0))\ d\sigma ds \right\| \leq C t^3 \|v_0\|_{H^2}$$

and, also using (2.2c),

$$\left\| \int_0^t \int_0^s (F(Y^\sigma X^{s/2} v_0) - F(Y^\sigma X^{t/2} v_0))\ d\sigma ds \right\| \leq C t^3 \|v_0\|_{H^2}.$$

Finally, since $X^{t/2}$ is unitary,

$$\left\| X^{t/2} \int_0^t R_2(s)\ ds \right\| \leq C t^3 \|v_0\|_{H^4},$$

and the conclusion is that

$$\|R(t)\| \leq C(1 + \|v_0\|_{H^4}) t^3 \|v_0\|_{H^4}.$$

- *Strang approximation—case $Z^t = Y^{t/2} X^t Y^{t/2}$.*
For $v_0 \in H^4$ and $0 \leq t \leq 1$, the remainder can be written as

$$\begin{aligned} R(t) = &\int_0^t X^{t-s} F(Y^{s/2} X^s Y^{s/2} v_0)\ ds \\ &-\frac{1}{2} \int_0^t X^t F(Y^{s/2} v_0)\ ds - \frac{1}{2} \int_0^t F(Y^{s/2} X^t Y^{t/2} v_0)\ ds. \end{aligned}$$

Taylor expansions yield

$$\begin{aligned} F(Y^{s/2} X^s Y^{s/2} v_0) = &F(v_0) + F'(v_0) \cdot (Y^{s/2} X^s Y^{s/2} v_0 - v_0) \\ &+ \int_0^1 (1-\theta) F''(v_0 + \theta(Y^{s/2} X^s Y^{s/2} v_0 - v_0)) \\ &\qquad \cdot (Y^{s/2} X^s Y^{s/2} v_0 - v_0)^2 d\theta, \\ F(Y^{s/2} v_0) = &F(v_0) + F'(v_0) \cdot (Y^{s/2} v_0 - v_0) \\ &+ \int_0^1 (1-\theta) F''(v_0 + \theta(Y^{s/2} v_0 - v_0)) \\ &\qquad \cdot (Y^{s/2} v_0 - v_0)^2 d\theta, \\ F(Y^{s/2} X^t Y^{t/2} v_0) = &F(v_0) + F'(v_0) \cdot (Y^{s/2} X^t Y^{t/2} v_0 - v_0) \\ &+ \int_0^1 (1-\theta) F''(v_0 + \theta(Y^{s/2} X^t Y^{t/2} v_0 - v_0)) \\ &\qquad \cdot (Y^{s/2} X^t Y^{t/2} v_0 - v_0)^2 d\theta, \end{aligned}$$

and the same sort of estimates as above give

$$\begin{aligned} \|Y^{s/2} X^s Y^{s/2} v_0 - v_0\| &\leq Cs \|v_0\|_{H^4}, \\ \|Y^{s/2} v_0 - v_0\| &\leq Cs \|v_0\|_{H^4}, \\ \|Y^{s/2} X^t Y^{t/2} v_0 - v_0\| &\leq Ct \|v_0\|_{H^4}. \end{aligned}$$

Therefore, the time integral over the interval $[0,t]$ of the integral remainders may be estimated by $Ct^3\|v_0\|_{H^2}^2$. Besides, there remains to estimate $\int_0^t R_1(s)\,ds$ with

$$\begin{aligned}R_1(s) = &\left(X^{t-s} - \frac{1}{2}X^t - \frac{1}{2}\mathrm{Id}\right)F(v_0)\\ &+ (X^{t-s}-\mathrm{Id})F'(v_0)\cdot(Y^{s/2}X^sY^{s/2}v_0 - v_0)\\ &- \frac{1}{2}(X^t-\mathrm{Id})F'(v_0)\cdot(Y^{s/2}v_0 - v_0)\\ &+ F'(v_0)\cdot(Y^{s/2}X^sY^{s/2}v_0 - v_0)\\ &- \frac{1}{2}F'(v_0)\cdot(Y^{s/2}v_0 - v_0) - \frac{1}{2}F'(v_0)\cdot(Y^{s/2}X^tY^{t/2}v_0 - v_0).\end{aligned}$$

The first term is estimated by $Ct^3\|v_0\|_{H^4}$, combining estimates (2.2c) and (2.2d). The two next terms are, respectively, estimated by $CM'(t-s)s\|v_0\|_{H^4}$ and $CM'ts\|v_0\|_{H^2}$.
Last, since $F'(v_0)$ is a linear operator, we have to study $\int_0^t F'(v_0)R_2(s)\,ds$ with

$$\begin{aligned}R_2(s) &= Y^{s/2}X^sY^{s/2}v_0 - \frac{1}{2}Y^{s/2}v_0 - \frac{1}{2}Y^{s/2}X^tY^{t/2}v_0\\ &= X^sv_0 + \frac{1}{2}\int_0^s X^sF(Y^{\sigma/2}v_0)\,d\sigma + \frac{1}{2}\int_0^s F(Y^{\sigma/2}X^sY^{s/2}v_0)\,d\sigma\\ &\quad - \frac{1}{2}v_0 - \frac{1}{4}\int_0^s F(Y^{\sigma/2}v_0)\,d\sigma\\ &\quad - \frac{1}{2}X^tv_0 - \frac{1}{4}\int_0^t X^tF(Y^{\sigma/2}v_0)\,d\sigma - \frac{1}{4}\int_0^t F(Y^{\sigma/2}X^tY^{t/2}v_0)\,d\sigma,\end{aligned}$$

where we have used intensively formulation (2.3). Everywhere where F occurs we subtract and add $F(v_0)$. This leads to terms involving differences which may be estimated by $Ct^2\|v_0\|_{H^2}$, and therefore their time integral is bounded by $Ct^3\|v_0\|_{H^2}$. The only terms that remain are $R_3(s) = X^sv_0 - \frac{1}{2}v_0 - \frac{1}{2}X^tv_0$ and $R_4(s) = \frac{1}{2}\int_0^s X^sF(v_0)d\sigma - \frac{1}{4}\int_0^t X^tF(v_0)d\sigma$. We have

$$R_3(s) = (X^sv_0 - X^{t/2}v_0) + \left(X^{t/2}v_0 - \frac{1}{2}v_0 - \frac{1}{2}X^tv_0\right),$$

$$\left\|\int_0^t R_3(s)\,ds\right\| \le Ct^3\|v_0\|_{H^4};$$

$$R_4(s) = \frac{1}{2}\left(s(X^s - X^t)F(v_0) + \left(s - \frac{t}{2}\right)X^tF(v_0)\right),$$

$$\left\|\int_0^t R_4(s)\,ds\right\| = \left\|\int_0^t s(X^s - X^t)F(v_0)\,ds\right\| \le Ct^3\|v_0\|_{H^2}.$$

Finally, we obtain that

$$\|R(t)\| \le C(1+\|v_0\|_{H^4})t^3\|v_0\|_{H^4}.$$

This last estimate concludes the study of the remainders for the four schemes. Now a consequence of the Gronwall lemma 2.1 is the following lemma.

LEMMA 3.1. *Let $v_0 \in H^2$; there exists $t_0 > 0$ such that for all $0 \le t \le t_0$*

$$\|Z_{\mathrm{L}}^t v_0 - S^t v_0\| \le Ct^2,$$

where C depends on $\|v_0\|_{H^2}$. Moreover, if $v_0 \in H^4$, there exists $t_1 > 0$ such that for all $0 \le t \le t_1$

$$\|Z_{\mathrm{S}}^t v_0 - S^t v_0\| \le Ct^3,$$

where C depends on $\|v_0\|_{H^4}$.

Remark 3.2. In [11], Jahnke and Lubich have shown the first and second order approximation for a linear Schrödinger equation under the weaker regularity conditions $v_0 \in H^1$ and $v_0 \in H^2$. Unfortunately, it is not possible to keep exactly the same hypothesis for the nonlinear case for the following reason: Let us focus on the first order approximation; we can formally extend the results of Jahnke and Lubich in the nonlinear case using Lie commutators. However, the Lie commutator between the Laplace operator and the nonlinear term involves a term containing $(\partial v_0/\partial x)^2$ and $(\partial v_0/\partial y)^2$ (see [13] for more details). To control these two terms, we have two possibilities, either we assume that $v_0 \in H^2$ and we use a Gagliardo–Nirenberg inequality, or we assume that $v_0 \in H^1 \cap W^{1,+\infty}$. Thus, in our lemma, H^2 is not optimal if we also assume that $v_0 \in W^{1,+\infty}$.

4. Order estimate.

THEOREM 4.1. *For all u_0 in H^2 and for all $T > 0$, there exists C and h_0 such that for all $h \in (0, h_0]$, for all n such that $nh \le T$*

$$\left\| \left(Z_L^h\right)^n u_0 - S^{nh} u_0 \right\| \le C(\|u_0\|_{H^2}) h \|u_0\|_{H^2}.$$

Moreover, if u_0 belongs to H^4, then

$$\left\| \left(Z_S^h\right)^n u_0 - S^{nh} u_0 \right\| \le C(\|u_0\|_{H^4}) h^2 \|u_0\|_{H^4}.$$

Proof. As noticed in the introduction, the triangle inequality yields

$$\|(Z^h)^n u_0 - S^{nh} u_0\| \le \sum_{j=0}^{n-1} \|(Z^h)^{n-j-1} Z^h S^{jh} u_0 - (Z^h)^{n-j-1} S^{(j+1)h} u_0\|.$$

In section 3 we have proved that for all the studied schemes there exists a constant C_0 such that for w_0 and $w_0' \in L^2$ and all $t \in [0, 1]$

$$\|Z^t w_0 - Z^t w_0'\| \le (1 + C_0 t)\|w_0 - w_0'\|,$$

and therefore

$$\|(Z^h)^n u_0 - S^{nh} u_0\| \le \sum_{j=0}^{n-1} (1 + C_0 h)^{n-j-1} \|(Z^h - S^h) S^{jh} u_0\|.$$

For the Lie formula when u_0 belongs to H^2, for all j such that $jh \le T$, $S^{jh} u_0$ belongs to H^2 and is uniformly bounded in this space; thus we have

$$\|(Z_{\mathrm{L}}^h - S^h) S^{jh} u_0\| \le C(\|u_0\|_{H^2}) h^2 \|u_0\|_{H^2},$$

and we deduce that

$$\begin{aligned}\|(Z_{\mathrm{L}}^h)^n u_0 - S^{nh}u_0\| &\leq C(\|u_0\|_{H^2})\|u_0\|_{H^2}\sum_{j=0}^{n-1}\exp(C_0h)^{n-j-1}h^2\\ &\leq C(\|u_0\|_{H^2})\|u_0\|_{H^2}\exp(C_0T)nh^2\\ &\leq C(\|u_0\|_{H^2})\|u_0\|_{H^2}h.\end{aligned}$$

For the scheme Z_{S}^h, when u_0 belongs to H^4, for all j such that $jh \leq T$, $S^{jh}u_0$ belongs to H^4 and is uniformly bounded in this space, and we have

$$\begin{aligned}\|(Z_{\mathrm{S}}^h)^n u_0 - S^{nh}u_0\| &\leq C(\|u_0\|_{H^4})\|u_0\|_{H^4}\sum_{j=0}^{n-1}\exp(C_0h)^{n-j-1}h^3\\ &\leq C(\|u_0\|_{H^4})\|u_0\|_{H^4}\exp(C_0T)nh^3\\ &\leq C(\|u_0\|_{H^4})\|u_0\|_{H^4}h^2.\end{aligned}$$

This concludes the proof of Theorem 4.1. ∎

Remark 4.2. Theorem 4.1 shows that the Lie and Strang formulae are approximations of order one and two of the exact solution. We can notice that the proof can be extended to high order splitting formulae. In [9], it is shown that we can construct Nth order approximation ($N \geq 3$) by considering splitting schemes of the form

$$Z_{\mathrm{HO}}^t = X^{c_0 t}Y^{d_1 t}X^{c_1 t}Y^{d_2 t}\cdots Y^{d_{m-1}t}X^{c_{m-1}t}Y^{d_m t}X^{c_m t}, \tag{4.1}$$

but we have to assume that at least one of the coefficient $c_0, \ldots, c_m$ must be negative and at least one of the coefficient $d_1, \ldots, d_m$ must be negative. (This result generalized the fundamental result of [16].) The same result holds if we consider convex combinations of (4.1). For these kinds of formulae, the Lipschitz property is an immediate consequence of their forms, and we notice that we can still use some Taylor formulae for X^t and Y^t to show that the remainder may be estimated as $\|R(t)\| = O(t^{N+1})$ for t small; however, as we have seen for the last scheme studied in the previous proof, it would be very technical.

5. Numerical experiments. We proved in the previous sections that the order p of Lie and Strang formulations are, respectively, 1 and 2 for initial data in H^2 and H^4.

If the numerical order p_{num} given in Table 5.1 does confirm the theoretical orders, it is nevertheless difficult to force the desired regularity for a discretized initial datum. Typically, the regularity of the L^2 initial datum in Figure 5.1 is certainly slightly better.

Let us define $t_n = nh$ and let $\Omega = [-10, 10] \times [-10, 10]$ be the computational domain. The numerical order p_{num} is computed by

$$p_{\mathrm{num}} = \max_{t_n \in [0,T]} \frac{1}{\ln 2}\ln\left(\frac{\|u_2 - u_1\|_{L^2(\Omega)}}{\|u_3 - u_2\|_{L^2(\Omega)}}\right),$$

where u_1 is computed for the time step h, and u_2 and u_3 are, respectively, computed for time steps $h/2$ and $h/4$.

We use initial data displayed in Figure 5.1, and in order to avoid numerical reflections due to boundaries we choose periodic boundary conditions and a FFT method to invert the Laplacian.

TABLE 5.1
Computation of p_{num} for different initial data.

	Lie	Strang
H^2	1.000685	2.000072
H^1	1.001721	2.006374
L^2	1.014480	2.010045

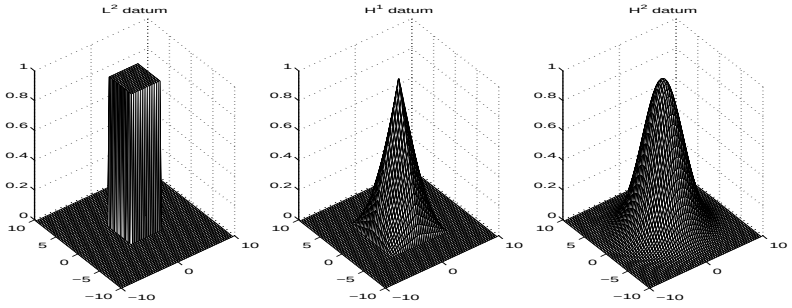


FIG. 5.1. *Initial data used for numerical experiments.*

TABLE 5.2
Computation of p_{num} for different time and space steps.

	$N = 64$	$N = 128$	$N = 256$
$h = 10^{-3}$	2.000016	2.000072	2.000289
$h = 10^{-2}$	2.001637	2.007160	2.030023

The results displayed in Table 5.1 are computed for $h = 10^{-3}$ and $N = 128$ points in both space dimensions.

This results are not much dependent on the choice for the time and space steps. Indeed, for the H^2 initial datum and the Strang formulation, we obtain the results of Table 5.2.

6. Conclusion. We have shown in this paper that, for the nonlinear Schrödinger equation, the Lie and Strang formulae are, respectively, approximations of order 1 and 2. This result could be extended to cover the case of the Schrödinger–Debye equations [3], where one can find a proof for the first order. The case of the nonlinear heat equation could also be treated with the same arguments because we have never used the group property but only the semigroup property of the flow of the linear Schrödinger equation; besides, we may write an equivalent of Lemma 2.2. In particular, this extends also the results of [6]. Our proof may also be extended to the Ginzburg–Landau equation, for which some splitting methods are also used (see, e.g., [10]) since it will use the fact that we are able to perform the proof for both the Schrödinger and the heat equation.

Our analysis does not give any hint on how to choose one splitting scheme among the others. The order of convergence is not the only criterion as stressed in the introduction: in case of stiff terms, the order of the different steps is of consequence. Namely, the last step should be the stiff one which is the nonlinear step Y^t in our context. This fact is hidden in our constants that depend on norms that grow with the size of the exact solution.

Acknowledgment. This work has been partially done during a session of the GDR "Equations d'Amplitudes et Propriétés Qualitatives" (GDR CNRS 2103: E.A.P.Q.).

REFERENCES

[1] G. Akrivis, V. A. Dougalis, and O. Karakashian, *Solving the systems of equations arising in the discretization of some nonlinear p.d.e.'s by implicit Runge-Kutta methods*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 251–287.

[2] C. Besse, *Schéma de relaxation pour l'équation de Schrödinger non linéaire et les systèmes de Davey et Stewartson*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1427–1432.

[3] C. Besse and B. Bidégaray, *Numerical study of self-focusing solutions to the Schrödinger-Debye system*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 35–56.

[4] M. Delfour, M. Fortin, and G. Payre, *Finite-difference solutions of a nonlinear Schrödinger equation*, J. Comput. Phys., 44 (1981), pp. 277–288.

[5] S. Descombes, *Convergence of a splitting method of high order for reaction-diffusion systems*, Math. Comp., 70 (2001), pp. 1481–1501.

[6] S. Descombes and M. Schatzman, *On Richardson Extrapolation of Strang Formula for Reaction-Diffusion Equations*, in Équations aux dérivées partielles et applications, Gauthier–Villars, Paris, 1998, pp. 429–452.

[7] S. Descombes and M. Schatzman, *Strang's formula for holomorphic semi-groups*, J. Math. Pures Appl., 81 (2002), pp. 93–114.

[8] P. Donnat, *Quelques contributions mathématiques en optique non linéaire*, Ph.D. thesis, École polytechnique, Palaiseau, France, 1994.

[9] D. Goldman and T. J. Kaper, *Nth-order operator splitting schemes and nonreversible systems*, SIAM J. Numer. Anal., 33 (1996), pp. 349–367.

[10] D. Goldman and L. Sirovich, *A novel method for simulating the complex Ginzburg-Landau equation*, Quart. Appl. Math., 53 (1995), pp. 315–333.

[11] T. Jahnke and C. Lubich, *Error bounds for exponential operator splittings*, BIT, 40 (2000), pp. 735–744.

[12] O. Karakashian, G. D. Akrivis, and V. A. Dougalis, *On optimal order error estimates for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 30 (1993), pp. 377–400.

[13] D. Lanser and J. G. Verwer, *Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling*, J. Comput. Appl. Math., 111 (1999), pp. 201–216.

[14] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.

[15] J. M. Sanz-Serna and J. G. Verwer, *Conservative and nonconservative schemes for the solution of the nonlinear Schrödinger equation*, IMA J. Numer. Anal., 6 (1986), pp. 25–42.

[16] Q. Sheng, *Global error estimates for exponential splitting*, IMA J. Numer. Anal., 14 (1994), pp. 27–56.

[17] B. Sportisse, *An analysis of operator splitting techniques in the stiff case*, J. Comput. Phys., 161 (2000), pp. 140–168.

[18] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[19] T. R. Taha and M. J. Ablowitz, *Analytical and numerical aspects of certain nonlinear evolution equations.* II. *Numerical, nonlinear Schrödinger equation*, J. Comput. Phys., 55 (1984), pp. 203–230.

[20] J. A. C. Weideman and B. M. Herbst, *Split-step methods for the solution of the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 23 (1986), pp. 485–507.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 41–65

SYMMETRIC COUPLING FOR EDDY CURRENT PROBLEMS*

R. HIPTMAIR†

Abstract. In this paper a novel symmetric finite element method-boundary element method-coupling for the **E**-based eddy current model is derived in a rigorous fashion. To that end, the properties of potentials and boundary integral operators arising from a Stratton–Chu-type representation formula for the electric field in the nonconducting region are thoroughly analyzed in a Hilbert-space setting. It yields a variational problem with symmetric bilinear form that is coercive in the natural function spaces. Unknowns are the electric field inside the conductor and the equivalent surface current related to the magnetic field. Existence and uniqueness of solutions and the convergence of a conforming finite element/boundary element Galerkin discretization immediately follows. In particular, schemes based on **curl**-conforming edge elements and divergence-conforming surface elements are examined, and some aspects of implementation are discussed.

Key words. eddy current problem, trace spaces, boundary integral operators, symmetric coupling, edge elements, boundary elements, stream functions, preconditioning

AMS subject classifications. 65N30, 35Q60, 45L10, 35C15

PII. S0036142900380467

1. Introduction.

The fundamental task in eddy current computation is the following: Given a time-dependent solenoidal exciting current and a conductor of prescribed shape, determine the induced eddy currents in the conductor. To put everything into a mathematical framework, let $\Omega_C \subset \mathbb{R}^3$ designate the space occupied by the conductor. It is to be a bounded domain with piecewise smooth Lipschitz-boundary $\Gamma := \partial\Omega_C$. In other words, Ω_C should be the union of a few (curved) Lipschitz-polyhedra [44, 1.5.2]. Inside Ω_C the conductivity $\sigma \in L^\infty(\Omega_C)$ is uniformly positive, i.e., $\sigma \geq \sigma_0 > 0$ almost everywhere in Ω_C, whereas σ vanishes outside Ω_C in the "air region" $\Omega_E := \mathbb{R}^3 \setminus \bar{\Omega}_C$. The second important material parameter, the magnetic permeability $\mu \in L^\infty(\mathbb{R}^3)$, is uniformly positive everywhere and constant in Ω_E. By a simple scaling we can always get nondimensional equations and achieve $\mu \equiv 1$ in Ω_E. Excitation is provided either by a divergence-free source current $\mathbf{j}_0 \in \boldsymbol{L}^2(\mathbb{R}^3)$ or by prescribing the total current circulating through a loop of the conductor. To begin with, the (generic) case of a given source current $\mathbf{j}_0$ will be the only one considered. Besides, for the sake of simplicity, we assume that $\mathrm{supp}(\mathbf{j}_0) \subset \bar{\Omega}_C$, that there is no flux of $\mathbf{j}_0$ through Γ, and that Ω_E is connected. I stress that it takes only slight alterations to adapt the considerations of this paper to a more general setting.

The eddy current model emerges from Maxwell's equations by formally dropping the displacement currents (magnetoquasistatic approximation). This amounts to neglecting capacitive effects (space charges) and provides a reasonable approximation in the low frequency range and in the presence of high conductivity [3, 40]. In the time harmonic case (frequency domain) the equations of the eddy current model read as

$$\mathbf{curl}\,\mathbf{E} = -i\omega\mu\mathbf{H}, \quad \mathbf{curl}\,\mathbf{H} = \sigma\mathbf{E} + \mathbf{j}_0 \quad \text{in } \mathbb{R}^3\,, \tag{1.1}$$

$$\mathbf{E}(\mathbf{x}) = O(|\mathbf{x}|^{-1}), \quad \mathbf{H}(\mathbf{x}) = O(|\mathbf{x}|^{-1}) \quad \text{for } |\mathbf{x}| \to \infty\,, \tag{1.2}$$

*Received by the editors November 6, 2000; accepted for publication (in revised form) October 18, 2001; published electronically April 12, 2002. This work was supported by DFG as part of SFB 382.

http://www.siam.org/journals/sinum/40-1/38046.html

†Sonderforschungsbereich 382, Universität Tübingen, Tübingen, Germany (ralf@hiptmair.de).

where $\omega > 0$ is a fixed angular frequency. Various reformulations of the eddy current problem have been suggested [1, 12], which differ in their choice of the primary unknown. This gives rise to a distinction between approaches zeroing in on the magnetic field ("**H**-based"), the electric field ("**E**-based"), or certain (vector) potentials. I am going to focus on the **E**-based formulation in the frequency domain, which, in the sense of distributions, reads as (cf. [3], $\epsilon_0 \equiv \text{const.} > 0$ in Ω_E)

$$\begin{aligned} &\mathbf{curl}\,\mu^{-1}\,\mathbf{curl}\,\mathbf{E} + i\omega\sigma\mathbf{E} = -i\omega\mathbf{j}_0 \quad \text{in } \mathbb{R}^3\ , \\ &\operatorname{div}(\epsilon_0\mathbf{E}) = 0 \quad \text{in } \Omega_E, \quad \int_{\Gamma_i} \mathbf{E}\cdot\mathbf{n}\,dS = 0\ , i = 1,\ldots,N_C\ , \\ &\mathbf{E}(\mathbf{x}) = O(|\mathbf{x}|^{-1}), \quad \mathbf{curl}\,\mathbf{E}(\mathbf{x}) = O(|\mathbf{x}|^{-1}) \quad \text{for } |\mathbf{x}| \to \infty\ . \end{aligned} \tag{1.3}$$

Here, $\{\Gamma_i\}_{i=1}^{N_C}$ stands for the finitely many connected components of $\Gamma := \partial\Omega_C$, and $\mathbf{n}$ denotes the unit normal vectorfield on Γ, defined almost everywhere and pointing from Ω_C into Ω_E.

In many practical applications the material parameters inside the conductor display substantial spatial variations. Moreover, the permeability μ may even depend on $\mathbf{H}$, which leads to a nonlinear problem. Thus, a viable numerical scheme for eddy current computation has to rely on a complete spatial discretization of the problem inside Ω_C. Starting from the variational formulation (1.3), a finite element scheme (FEM) based on edge elements offers the most attractive option [18, 63]. In fact, their use is mandatory in order to capture the singularities of the fields at material interfaces [37, 11, 19].

However, we also have to deal with the unbounded exterior domain. In many cases, due to the decay properties of the fields, one simply introduces homogeneous boundary conditions for the electric field some distance away from the conductor. Then by extending the finite element mesh to parts of the air region, a satisfactory approximation can be obtained. However, there are many conceivable shapes of Ω_C, where the number of required additional elements much exceeds the number needed to mesh Ω_C.

In this case, BEMs may be used to tackle the unbounded exterior domain, since in Ω_E we face a homogeneous problem with constant coefficients. Then the topic of this paper comes into focus, because the discrete boundary integral equations and the systems of equations arising from the finite element scheme have to be linked properly.

Formally, for second order elliptic problems, the coupling of a finite element discretization with some unbounded exterior domain is achieved by discretizing the Dirichlet-to-Neumann map of the exterior problem. For a self-adjoint differential operator, this map must be self-adjoint, too. However, in the traditional approach employing boundary elements, a nonsymmetric matrix emerges as discrete Dirichlet-to-Neumann map [52]. A breakthrough was accomplished by M. Costabel, who discovered that, by using the full Calderón-projector, symmetry can be restored [34]. In addition, coercive variational problems naturally emerge. The principle can be applied to a wide range of elliptic problems [25] and has recently been adapted to the full Maxwell's equations [4].

The idea of coupled FEM-BEM methods for eddy current computation is by no means new. An early reference is [58], where an **H**-ϕ-approach is discussed. It employs a scalar magnetic potential outside Ω_C, and the coupling makes use of the continuity of tangential components of $\mathbf{H}$ and that of the normal components of $\mu\mathbf{H}$. A similar idea was used in the context of an **A**-ϕ-model in [57] and, along with impedance boundary

conditions, in [75]. Yet this coupling is valid only for simply connected conductors [68, 56].

Field based methods avoid the fundamental difficulties encountered by potential based approaches in the case of complex topologies. Most prominently, the TRIFOU project adopted a coupling strategy based on **H** or **E** [70, 71, 16]. In the former case, scalar potentials still played a role, entailing cutting surfaces to deal with non-simply connected conductors. In the latter case, the concept of a vectorial Dirichlet-to-Neumann map bulked large [17, 15], which, though self-adjoint by nature, was tackled by an inherently unsymmetric indirect boundary integral formulation. A symmetrization of the resulting matrices is performed later on the level of the discrete linear systems, justified by heuristics drawing on the symmetry of the Dirichlet-to-Neumann map [18, sect. 7.4.5] and [72]. Yet, a rigorous mathematical analysis of the impact of this procedure is not available.

This paper is aimed at extending the "TRIFOU-**E**-approach" towards a genuinely symmetric formulation, which is to arise from a thorough mathematical examination of the coupled eddy current problem. We are rewarded by an asymptotically optimal a priori error estimate for the energy norm of the discretization error and linear systems of equations amenable to efficient preconditioning. No restrictions on the topology of Ω_C are imposed. Owing to the lack of strong ellipticity in the case of the eddy current problem, the formal approach to symmetric coupling in the case of second order elliptic boundary value problems cannot be simply copied.

The paper is organized as follows: The following section presents a weak formulation of the eddy current problem. After that, the crucial function spaces of tangential traces are introduced in section 3. Then I examine the relevant Dirichlet and Neumann data for the eddy current problem in Ω_C and Ω_E. The correct coupling conditions are established. In section 5 a representation formula is computed, comprised of several boundary potentials. Their properties and those of related boundary integral operators will be studied in sections 6 and 7. In section 8 I state the symmetric coupled problem in weak form and prove existence and uniqueness of a solution. Section 9 is dedicated to a Galerkin finite element discretization of the coupled problem and the use of discrete stream functions. In the final section the assembly of the discrete operators and iterative solution techniques are examined.

2. Variational formulation and gauging. As the eddy current problem is posed on an unbounded domain, its variational formulation has to rely on the weighted Sobolev space

$$\boldsymbol{W}(\mathbf{curl},\mathbb{R}^3) := \left\{ \mathbf{u} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)',\ \frac{\mathbf{u}(\mathbf{x})}{\sqrt{1+|\mathbf{x}|^2}} \in \boldsymbol{L}^2(\mathbb{R}^3),\ \mathbf{curl}\,\mathbf{u} \in \boldsymbol{L}^2(\mathbb{R}^3) \right\},$$

which corresponds to the classical Beppo–Levi spaces. A fairly complete treatment of these spaces is given in [41] and [65, sect. 2.5.4]. The constrained space

$$\mathbf{X}(\mathbb{R}^3) := \left\{ \mathbf{u} \in \boldsymbol{W}(\mathbf{curl},\mathbb{R}^3),\ \operatorname{div}\mathbf{u} = 0 \text{ in } \Omega_E,\ \int_{\Gamma_i} \mathbf{u}_{|\Omega_E} \cdot \mathbf{n}\, dS = 0,\ i = 1,\dots,N_C \right\}$$

offers a suitable setting for the weak eddy current problem: Seek $\mathbf{E} \in \mathbf{X}(\mathbb{R}^3)$ such that for all $\mathbf{v} \in \mathbf{X}(\mathbb{R}^3)$

$$a(\mathbf{E},\mathbf{v}) := \left(\mu^{-1}\,\mathbf{curl}\,\mathbf{E}, \mathbf{curl}\,\mathbf{v}\right)_{0;\mathbb{R}^3} + i\omega\,(\sigma\mathbf{E},\mathbf{v})_{0;\Omega_C} = -i\omega\,(\mathbf{j}_0,\mathbf{v})_{0;\Omega_C}\ . \tag{2.1}$$

THEOREM 2.1. *A solution of the variational problem* (2.1) *exists and is unique.*

Proof. First, I show that the bilinear form $a(\cdot,\cdot)$ of (2.1) is coercive (in the sense of [60, Thm. 2.34]) on

$$\widetilde{\mathbf{X}}(\mathbb{R}^3) := \{\mathbf{u} \in \boldsymbol{W}(\mathbf{curl},\mathbb{R}^3),\, \operatorname{div}\mathbf{u} = 0 \text{ in } \Omega_E\}\,. \tag{2.2}$$

To this end, pick $\mathbf{v} \in \tilde{\mathbf{X}}(\mathbb{R}^3)$, write $\mathbf{v}^C := \mathbf{v}_{|\Omega_C} \in \boldsymbol{H}(\mathbf{curl};\Omega_C)$ and $\tilde{\mathbf{v}} \in \boldsymbol{H}(\mathbf{curl};\mathbb{R}^3)$ for a divergence-free extension of $\mathbf{v}^C$ to $\mathbb{R}^3$. (Extension theorems for $\boldsymbol{H}(\mathbf{curl};\Omega)$ are presented in the next section.) Then define $\mathbf{w} := \mathbf{v} - \tilde{\mathbf{v}} \in \widetilde{\mathbf{X}}(\mathbb{R}^3)$ and observe that $\mathbf{w}$ has vanishing tangential components on Γ. Next, set

$$\mathbf{u} := \mathbf{curl}\,\mathbf{w} \in \boldsymbol{H}_0(\operatorname{div}0;\Omega_E) := \{\mathbf{q} \in \boldsymbol{H}(\operatorname{div};\Omega_E),\, \operatorname{div}\mathbf{q} = 0,\, \mathbf{q}\cdot\mathbf{n} = 0 \text{ on } \Gamma\}\,,$$

and retain the notation $\mathbf{u}$ for its extension by zero into the interior of Ω_C. This will yield $\mathbf{u} \in \boldsymbol{H}(\operatorname{div};\mathbb{R}^3)$ with $\operatorname{div}\mathbf{u} = 0$. According to [41, Thm. 2.5] we can find a unique vector potential $\boldsymbol{\Psi}$ in the vectorial Beppo–Levi space (cf. [65, sect. 2.5.4])

$$\boldsymbol{W}^{1,-1}(\mathbb{R}^3) := \left\{\boldsymbol{\Phi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)',\, \frac{\boldsymbol{\Phi}(\mathbf{x})}{\sqrt{1+|\mathbf{x}|^2}} \in \boldsymbol{L}^2(\mathbb{R}^3),\, D\boldsymbol{\Phi} \in (L^2(\mathbb{R}^3))^{3\times 3}\right\}$$

such that $\mathbf{curl}\,\boldsymbol{\Psi} = \mathbf{u}$, $\operatorname{div}\boldsymbol{\Psi} = 0$, and, with some universal constant $C > 0$,

$$\|\boldsymbol{\Psi}\|_{\boldsymbol{W}^{1,-1}(\mathbb{R}^3)} \le C\,\|\mathbf{u}\|_{\boldsymbol{L}^2(\mathbb{R}^3)}\,.$$

By definition, $\mathbf{curl}\,\boldsymbol{\Psi} = 0$ in Ω_C, which yields the representation

$$\boldsymbol{\Psi} = \operatorname{grad}\phi + \boldsymbol{\eta}, \quad \phi \in H^1(\Omega_C)/\mathbb{R},\, \boldsymbol{\eta} \in \mathbb{H}(\Omega_C)\,,$$

where $\mathbb{H}(\Omega_C) \subset \boldsymbol{L}^2(\Omega_C)$ is the *finite dimensional* space of harmonic Neumann vectorfields on Ω_C, $\boldsymbol{L}^2(\Omega_C)$-orthogonal to $\operatorname{grad} H^1(\Omega_C)$ [5, sect. 3.c]. Then solve the exterior Dirichlet problem

$$\Delta\mu = 0 \quad \text{in } \Omega_E, \quad \mu = \phi \quad \text{on } \Gamma$$

in the Beppo–Levi space $W^{1,-1}(\Omega_E)$ and observe that, with generic constants $C > 0$,

$$\|\mu\|_{W^{1,-1}(\Omega_E)} \le C\left\|\phi_{|\Gamma}\right\|_{H^{\frac12}(\Gamma)} \le C\,\|\phi\|_{H^1(\Omega_C)} \le C\,\|\boldsymbol{\Psi}\|_{\boldsymbol{H}^1(\Omega_C)} \le C\,\|\mathbf{curl}\,\mathbf{w}\|_{\boldsymbol{L}^2(\mathbb{R}^3)}\,.$$

This means for $\boldsymbol{\zeta} := \boldsymbol{\Psi} - \operatorname{grad}\mu - \mathbf{w}$ that

$$\mathbf{curl}\,\boldsymbol{\zeta} = 0 \quad \text{in } \Omega_E, \quad \operatorname{div}\boldsymbol{\zeta} = 0 \quad \text{in } \Omega_E\,. \tag{2.3}$$

On top of that, the tangential components of $\boldsymbol{\zeta}$ on Γ agree with those of vectorfields in $\mathbb{H}(\Omega_C)$. Along with (2.3) this implies that $\boldsymbol{\zeta}$ belongs to a finite dimensional space of harmonic vectorfields on Ω_E. Summing up, we get in Ω_E

$$\mathbf{v} = \mathbf{q} - \boldsymbol{\zeta}, \quad \mathbf{q} := \tilde{\mathbf{v}} + \boldsymbol{\Psi} - \operatorname{grad}\mu$$

and the estimate

$$\begin{aligned}\left\|\frac{\mathbf{q}(\mathbf{x})}{\sqrt{1+|\mathbf{x}|^2}}\right\|_{\boldsymbol{L}^2(\Omega_E)} &\le \|\boldsymbol{\Psi}\|_{\boldsymbol{W}^{1,-1}(\Omega_E)} + \|\tilde{\mathbf{v}}\|_{\boldsymbol{L}^2(\Omega_E)} + \|\mu\|_{W^{1,-1}(\Omega_E)}\\ &\le C\left(\|\mathbf{v}\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} + \|\mathbf{curl}\,\mathbf{v}\|_{\boldsymbol{L}^2(\Omega_E)}\right).\end{aligned}$$

We conclude that, with a constant $c > 0$ depending on the material parameters μ, σ, and on Ω_C,

$$|a(\mathbf{v},\mathbf{v})| \geq c \left(\left\| \frac{(\mathbf{v}+\boldsymbol{\zeta})(\mathbf{x})}{\sqrt{1+|\mathbf{x}|^2}} \right\|^2_{\boldsymbol{L}^2(\Omega_E)} + \|\mathbf{v}\|^2_{\boldsymbol{L}^2(\Omega_C)} + \|\mathbf{curl}\,\mathbf{v}\|^2_{\boldsymbol{L}^2(\mathbb{R}^3)} \right) .$$

Recalling that the norm on $\tilde{\mathbf{X}}(\mathbb{R}^3)$ agrees with the norm of $\boldsymbol{W}(\mathbf{curl},\mathbb{R}^3)$ and that $\boldsymbol{\zeta}$ belongs to a finite dimensional space, $a(\cdot,\cdot)$ is immediately seen to be $\tilde{\mathbf{X}}(\mathbb{R}^3)$-coercive, i.e., $\tilde{\mathbf{X}}(\mathbb{R}^3)$-elliptic modulo a compact perturbation.

According to [60, Thm. 2.34], existence of solutions of the weak eddy current problem thus follows from uniqueness. The latter is clear (cf. [3, Thm. 3.2]); as for $\mathbf{u} \in \widetilde{\mathbf{X}}(\mathbb{R}^3)$, the requirement $a(\mathbf{u},\mathbf{u}) = 0$ forces $\mathbf{u}$ to be an exterior Dirichlet vectorfield. However, those are ruled out by the additional constraint in the definition of $\mathbf{X}(\mathbb{R}^3)$. □

It is evident that the constraints in the definition of $\mathbf{X}(\mathbb{R}^3)$ merely serve to enforce the uniqueness of $\mathbf{E}$ outside Ω_C. Dispensing with them will affect neither the magnetic field $\mathbf{H} := -\frac{1}{i\omega\mu}\,\mathbf{curl}\,\mathbf{E}$ in $\mathbb{R}^3$ nor the eddy currents $\mathbf{j} = \sigma\mathbf{E}$ in the conductor. In other words, the constraints on $\mathbf{E}$ are not essential for the validity of the eddy current model. A meaningful model can also be stated in terms of *equivalence classes of electric fields*. It will still supply unique solutions for many interesting quantities. Indeed, frequently the constraints on $\mathbf{E}$ are dropped in numerical schemes [31]. In the spirit of [53], the constraints imposed in (1.3) represent *gauge conditions*, unless one wants to know the true $\mathbf{E}$ in Ω_E.

As my focus is on Ω_C, I am going to relax the constraints on $\mathbf{E}$ in what follows, considering the variational problem only on $\widetilde{\mathbf{X}}(\mathbb{R}^3)$: Seek $\mathbf{u} \in \widetilde{\mathbf{X}}(\mathbb{R}^3)$ such that

$$a(\mathbf{u},\mathbf{v}) = -i\omega\,(\mathbf{j}_0,\mathbf{v})_{0;\Omega_C} \quad \forall \mathbf{v} \in \widetilde{\mathbf{X}}(\mathbb{R}^3)\ . \tag{2.4}$$

Keep in mind that then $\mathbf{u}_{|\Omega_E}$ is merely unique up to contributions from $\mathbb{H}_D(\Omega_E)$, with

$$\mathbb{H}_D(\Omega) := \{\mathbf{v} \in \boldsymbol{L}^2(\Omega),\, \mathbf{curl}\,\mathbf{v} = 0,\, \mathrm{div}\,\mathbf{v} = 0,\, \mathbf{v}\times\mathbf{n} = 0 \text{ on } \partial\Omega\}$$

standing for the finite dimensional space of harmonic Dirichlet vectorfields (cf. [5, sect. 3]). To remind the reader that I am no longer dealing with a physical electric field, this "ungauged" solution will not be denoted by $\mathbf{E}$. Retreating to $\widetilde{\mathbf{X}}(\mathbb{R}^3)$ has to be accompanied by an important caveat: The numerical scheme proposed in the present paper has to be supplemented by some postprocessing in order to yield meaningful values for $\mathbf{E}$ in Ω_E.

Please recall that by means of expansions into spherical harmonics (cf. the proof of Proposition 3.1 in [3]) the following decay properties can be established for any solution $\mathbf{u}$ of (2.4):

$$\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|^{-1}), \quad \mathbf{curl}\,\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|^{-2}) \quad \text{uniformly for } |\mathbf{x}| \to \infty\ . \tag{2.5}$$

3. Traces. The boundary integral equations have to be considered in spaces that are closely related to traces of vectorfields in $\boldsymbol{H}(\mathbf{curl};\Omega)$ onto Γ. (Here and in the following, Ω stands for either Ω_C or Ω_E.) On smooth boundaries their theory is well established. I refer to the papers by Paquet [66] and Alonso and Valli [2] and, in particular, to the monographs by Cessenat [26, sect. 2.1] and Nédélec [65, sect. 5.4.1] for a comprehensive exposition. Only recently, these results have been successfully

extended to piecewise smooth boundaries by Buffa [21] and Buffa and Ciarlet [22, 23] and even to Lipschitz-domains in [24]. These articles provide the chief references for the results cited below. Loosely speaking, by judicious generalizations of norms and surface differential operators, the main results for regular surfaces carry over to a nonsmooth setting.

To begin with, we define two different tangential surface trace operators: The *tangential surface trace* $\gamma_{\mathbf{t}}$ is defined for $\mathbf{u} \in C(\bar{\Omega})^3$ by $\gamma_{\mathbf{t}}\mathbf{u}(\mathbf{x}) := \mathbf{n}(\mathbf{x}) \times (\mathbf{u}(\mathbf{x}) \times \mathbf{n}(\mathbf{x}))$ for almost all $\mathbf{x} \in \Gamma$. Accordingly, the *twisted tangential surface trace* $\gamma_{\mathbf{t}}^{\times}$ can be computed through $\gamma_{\mathbf{t}}^{\times}\mathbf{u}(\mathbf{x}) := \mathbf{u}(\mathbf{x}) \times \mathbf{n}(\mathbf{x})$.

Apart from the usual Hilbert spaces of scalar functions and functionals, $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$, we will make heavy use of the following Hilbert spaces of *tangential vectorfields* on Γ: $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$, $\boldsymbol{H}_{||}^{1/2}(\Gamma)$, and $\boldsymbol{H}_{\perp}^{1/2}(\Gamma)$, where the latter are defined in [22, sect. 1] (and generalized to spaces V_π and V_γ in [24, sect. 2]). The associated dual spaces will be denoted by $\boldsymbol{H}_{||}^{-1/2}(\Gamma)$ and $\boldsymbol{H}_{\perp}^{-1/2}(\Gamma)$, respectively. Sloppily speaking, $\boldsymbol{H}_{||}^{1/2}(\Gamma)$ contains the tangential surface vectorfields that are in $\boldsymbol{H}^{1/2}(F_i)$ for each smooth component F_i of Γ and feature a suitable "weak tangential continuity" across the edges of the F_i. A corresponding "weak normal continuity" is satisfied by surface vectorfields in $\boldsymbol{H}_{\perp}^{1/2}(\Gamma)$. These spaces occur as images of tangential traces of vectorfields.

THEOREM 3.1. *The tangential trace mapping $\gamma_{\mathbf{t}} : \boldsymbol{H}^1_{\mathrm{loc}}(\Omega) \mapsto \boldsymbol{H}_{||}^{1/2}(\Gamma)$ is continuous and surjective and possesses a continuous right inverse.*

The twisted tangential trace mapping $\gamma_{\mathbf{t}}^{\times} : \boldsymbol{H}^1_{\mathrm{loc}}(\Omega) \mapsto \boldsymbol{H}_{\perp}^{1/2}(\Gamma)$ is continuous and surjective and possesses a continuous right inverse.

Proof. See Proposition 1.7 in [22]. ∎

Please recall the definitions of the surface differential operators grad_Γ, $\mathbf{curl}_\Gamma$, curl_Γ, and div_Γ acting on tangential vectorfields (see [24, sect. 3], [22, sect. 2], [23, sect. 1, Defs. 1.1, 1.3], and [65, sect. 2.5.6] for smooth surfaces). I mention that $\mathrm{div}_\Gamma = -\mathrm{grad}_\Gamma^*$ and $\mathbf{curl}_\Gamma = \mathrm{curl}_\Gamma^*$, where $*$ denotes the adjoint with respect to the pivot space $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$. The surface differential operators play a role in the definition of the spaces $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma, \Gamma)$ and $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ introduced in [22] by

$$\begin{aligned}\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma, \Gamma) &= \{\mathbf{v} \in \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma),\ \mathrm{curl}_\Gamma \mathbf{v} \in H^{-\frac{1}{2}}(\Gamma)\}\ ,\\ \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) &= \{\boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma),\ \mathrm{div}_\Gamma \boldsymbol{\zeta} \in H^{-\frac{1}{2}}(\Gamma)\}\ .\end{aligned}$$

These spaces are endowed with the natural graph norms $\|\cdot\|_{\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma,\Gamma)}$ and $\|\cdot\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}$. They are significant as suitable trace spaces for vectorfields in $\boldsymbol{H}(\mathbf{curl}; \Omega)$.

THEOREM 3.2. *The trace mapping $\gamma_{\mathbf{t}} : \boldsymbol{H}(\mathbf{curl}; \Omega) \mapsto \boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma, \Gamma)$ is continuous and surjective with a continuous right inverse.*

The trace mapping $\gamma_{\mathbf{t}}^{\times} : \boldsymbol{H}(\mathbf{curl}; \Omega) \mapsto \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ is continuous and surjective with a continuous right inverse.

Proof. For simply connected Γ this result combines Theorems 2.7 and 2.8 from [22] and Theorem 4.5 in [23]. For general topology see [21]. The statement for smooth domains is made in [65, Thm. 5.4.2]. ∎

These traces occur in the definition of Dirichlet-type boundary conditions for $\boldsymbol{H}(\mathbf{curl}; \Omega)$-elliptic variational problems. Hence, we adopt the alternative notation γ_D for $\gamma_{\mathbf{t}}$ ("Dirichlet trace").

Another result in [23, sect. 4] is that $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$ and $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$ are dual to each other, when $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$ is used as pivot space. More precisely, the usual $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$-inner product can be extended to a duality pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau}}$ between $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$ and $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$.

A couple of weakly defined traces will also be needed in order to deal with Neumann-type boundary conditions: For

$$\mathbf{u}\in\boldsymbol{W}(\mathbf{curl}^2,\Omega):=\{\mathbf{v}\in\boldsymbol{W}(\mathbf{curl},\Omega),\ \mathbf{curl}\,\mathbf{curl}\,\mathbf{v}\in\boldsymbol{L}^2(\Omega)\},$$

we define $\gamma_N\mathbf{u}\in\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$ ("Neumann trace") by

$$\langle\gamma_N\mathbf{u},\gamma_{\mathbf{t}}\mathbf{v}\rangle_{\boldsymbol{\tau}}=\pm\int_\Omega\mathbf{curl}\,\mathbf{u}\cdot\mathbf{curl}\,\overline{\mathbf{v}}-\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}\cdot\overline{\mathbf{v}}\,d\mathbf{x}\quad\forall\mathbf{v}\in\boldsymbol{H}(\mathbf{curl};\Omega)\,,$$

where the positive sign applies to $\Omega=\Omega_C$, the negative to $\Omega=\Omega_E$. An overbar designates complex conjugation. The trace γ_N furnishes a continuous and surjective mapping $\gamma_N:\boldsymbol{W}(\mathbf{curl}^2,\Omega)\mapsto\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$, where $\boldsymbol{W}(\mathbf{curl}^2,\Omega)$ is equipped with the graph norm. Obviously $\gamma_N\mathbf{u}=\gamma_{\mathbf{t}}^{\times}(\mathbf{curl}\,\mathbf{u})$, as this holds for smooth fields.

LEMMA 3.3. *If* $\mathbf{u}\in\boldsymbol{W}(\mathbf{curl}^2,\Omega)$ *satisfies* $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}=0$ *in* Ω*, there holds*

$$\|\gamma_N\mathbf{u}\|_{\boldsymbol{H}_{||}^{-\frac12}(\mathrm{div}_\Gamma,\Gamma)}\le C\,\|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)}$$

with some constant $C>0$ *independent of* $\mathbf{u}$.

Proof. Set $P_\perp:\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)\mapsto\boldsymbol{H}(\mathbf{curl};\Omega)$ for the continuous right inverse of $\gamma_{\mathbf{t}}$ from Theorem 3.2. Then the definition of γ_N, straightforward manipulations, and the continuity of the tangential trace show

$$\begin{aligned}\|\gamma_N\mathbf{u}\|_{\boldsymbol{H}_{||}^{-\frac12}(\mathrm{div}_\Gamma,\Gamma)}&=\sup_{\mathbf{v}\in\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}\frac{\langle\gamma_N\mathbf{u},\gamma_D P_\perp\mathbf{v}\rangle_{\boldsymbol{\tau}}}{\|\mathbf{v}\|_{\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}}\\ =\sup_{\mathbf{v}\in\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}\frac{(\mathbf{curl}\,\mathbf{u},\mathbf{curl}\,P_\perp\mathbf{v})_{0;\Omega}}{\|\mathbf{v}\|_{\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}}&\le\sup_{\mathbf{v}\in\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}\frac{\|\mathbf{curl}\,P_\perp\mathbf{v}\|_{\boldsymbol{L}^2(\Omega)}\,\|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{L}^2(\Omega)}}{\|\mathbf{v}\|_{\boldsymbol{H}_{\perp}^{-\frac12}(\mathrm{curl}_\Gamma,\Gamma)}}\,.\end{aligned}$$

This amounts to the assertion of the lemma. ∎

The weak normal trace $\gamma_{\mathbf{n}}$ is defined for vectorfields $\mathbf{u}\in\boldsymbol{H}(\mathrm{div};\Omega)$ by

$$\langle\gamma_{\mathbf{n}}\mathbf{u},\phi\rangle_{1/2,\Gamma}=\int_\Omega\overline{\phi}\,\mathrm{div}\,\mathbf{u}+\mathbf{u}\cdot\mathrm{grad}\,\overline{\phi}\,d\mathbf{x}\quad\forall\phi\in H^1(\Omega)$$

with $\langle\cdot,\cdot\rangle_{1/2,\Gamma}$ as duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. $\gamma_{\mathbf{n}}:\boldsymbol{H}(\mathrm{div};\Omega)\mapsto H^{-1/2}(\Gamma)$ is continuous and surjective [42, Thm. 2.5], [65, Thm. 5.4.1].

4. Transmission conditions and Cauchy data. The interface transmission conditions for electromagnetic fields are

$$[\gamma_D\mathbf{u}]_\Gamma=0,\qquad\left[\mu^{-1}\gamma_N\mathbf{u}\right]_\Gamma=0\,. \tag{4.1}$$

I have adopted the notation $[\gamma\cdot]_\Gamma=\gamma^+\cdot-\gamma^-\cdot$ for the jump of some trace γ across Γ. Here, γ^+ refers to the trace from Ω_E and γ^- to the trace from inside. Please be aware

that, in physical terms, (4.1) represents the continuity of the tangential components of the electric and magnetic fields. This ensures the normal continuity of the energy flux given by the Poynting-vector $\mathbf{E} \times \overline{\mathbf{H}}$.

Remark. As only magnetic quantities matter, it is tempting to replace (4.1) by purely magnetic transmission conditions

$$[\gamma_D \mathbf{H}]_\Gamma = 0, \qquad [\gamma_\mathbf{n} \mathbf{B}]_\Gamma = 0 \, .$$

However, for nontrivial topology of Ω_C the resulting coupled problem fails to have a unique solution even for $\mathbf{H}$ [68, 56]: Loop currents in Ω_C can no longer be determined, because Faraday's law is not completely taken into account. Thus, in general, FEM-BEM-coupling for the eddy current problem cannot be based on $\mathbf{A}$-V-formulations [10], V a scalar magnetic potential, unless one is willing to grapple with integrals over cuts in the air region.

Next, we scrutinize the interior eddy current problem in Ω_C. Its strong forms are straightforward: The Dirichlet problem for Dirichlet-data $\mathbf{g} \in \boldsymbol{H}_\perp^{-1/2}(\mathrm{curl}_\Gamma, \Gamma)$ reads

$$\mathbf{curl}\, \frac{1}{\mu}\, \mathbf{curl}\, \mathbf{u} + i\omega\sigma \mathbf{u} = -i\omega \mathbf{j}_0 \quad \text{in } \Omega_C, \quad \gamma_D \mathbf{u} = \mathbf{g} \quad \text{on } \Gamma \, ,$$

whereas the Neumann problem for $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ can be stated as

$$\mathbf{curl}\, \mu^{-1}\, \mathbf{curl}\, \mathbf{u} + i\omega\sigma \mathbf{u} = -i\omega \mathbf{j}_0 \quad \text{in } \Omega_C, \quad \mu^{-1} \gamma_N \mathbf{u} = \boldsymbol{\lambda} \quad \text{on } \Gamma \, .$$

Reassuringly, the Cauchy data $(\gamma_D \mathbf{u}, \gamma_N \mathbf{u})$ exactly match the transmission conditions (4.1).

Now, we examine the part of the variational problem (1.3) that is associated with Ω_E. As pointed out in section 2, only relaxed constraints on $\mathbf{E}$ in Ω_E will be considered. Then the weak saddle point formulation of the exterior[1] Dirichlet problem corresponding to (2.4) reads as follows: Given $\mathbf{g} \in \boldsymbol{H}_\perp^{-1/2}(\mathrm{curl}_\Gamma, \Gamma)$, seek $\mathbf{u} \in \boldsymbol{W}(\mathbf{curl}, \Omega_E)$, $\gamma_\mathbf{t} \mathbf{u} = \mathbf{g}$, $p \in W_0^{1,0}(\Omega_E)$, where (cf. [65, sect. 2.5.4])

$$W_0^{1,0}(\Omega) := \{\phi \in L^2(\Omega),\, \sqrt{1+|\mathbf{x}|^2}\, \mathrm{grad}\, \phi(\mathbf{x}) \in \boldsymbol{L}^2(\Omega),\, \phi_{|\partial\Omega} = 0\} \, ,$$

such that

$$\begin{array}{rcccll} (\mathbf{curl}\, \mathbf{u}, \mathbf{curl}\, \mathbf{v})_{0;\Omega_E} & + & (\mathrm{grad}\, p, \mathbf{v})_{0;\Omega_E} & = & 0 & \forall \mathbf{v} \in \boldsymbol{W}_0(\mathbf{curl}, \Omega_E) \, , \\ (\mathbf{u}, \mathrm{grad}\, q)_{0;\Omega_E} & & & = & 0 & \forall q \in W_0^{1,0}(\Omega_E) \, . \end{array}$$

This gives rise to the boundary value problem

$$\mathbf{curl}\, \mathbf{curl}\, \mathbf{u} = 0, \quad \mathrm{div}\, \mathbf{u} = 0 \quad \text{in } \Omega_E, \quad \gamma_\mathbf{t} \mathbf{u} = \mathbf{g} \quad \text{on } \Gamma \, , \tag{4.2}$$

plus decay conditions as in (1.3).

Dropping boundary conditions in the function spaces leads to the Neumann problem: For $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$, find $\mathbf{u} \in \boldsymbol{W}(\mathbf{curl}, \Omega_E)$, $p \in W^{1,0}(\Omega_E)$ with

$$\begin{array}{rcccll} (\mathbf{curl}\, \mathbf{u}, \mathbf{curl}\, \mathbf{v})_{0;\Omega_E} & + & (\mathrm{grad}\, p, \mathbf{v})_{0;\Omega_E} & = & \langle \boldsymbol{\lambda}, \gamma_\mathbf{t} \mathbf{v} \rangle_{\boldsymbol{\tau}} & \forall \mathbf{v} \in \boldsymbol{W}(\mathbf{curl}, \Omega_E) \, , \\ (\mathbf{u}, \mathrm{grad}\, q)_{0;\Omega_E} & & & = & 0 & \forall q \in W^{1,0}(\Omega_E) \, . \end{array}$$

[1] I keep calling the problem "exterior" even if it may be posed on a bounded domain.

The related boundary value problem can be stated (without decay conditions) as

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} + \operatorname{grad} p = 0 \quad \text{in } \Omega_E, \quad \operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega_E\ , \tag{4.3}$$

$$\gamma_{\mathbf{n}}\mathbf{u} = 0 \quad \text{on } \Gamma, \qquad \gamma_N\mathbf{u} = \boldsymbol{\lambda} \quad \text{on } \Gamma\ . \tag{4.4}$$

DEFINITION 4.1. *A distribution $\mathbf{u} \in \boldsymbol{\mathcal{D}}(\Omega)'$ is called a solution of the "exterior" eddy current problem in Ω if $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} = 0$ and $\operatorname{div}\mathbf{u} = 0$. If the domain is not specified, $\Omega = \Omega_C \cup \Omega_E$ is implied.*

Remembering $\operatorname{div}_\Gamma^* = \operatorname{grad}_\Gamma$ and $\operatorname{grad}_\Gamma(\gamma_{\mathbf{t}}\boldsymbol{\Psi}) = \gamma_{\mathbf{t}}(\operatorname{grad}\boldsymbol{\Psi})$, we see from (4.3) that $\operatorname{grad} p = 0$ if and only if $\operatorname{div}_\Gamma\boldsymbol{\lambda} = 0$. This entails a constraint for the Neumann boundary values that produce solutions of the exterior eddy current problem in Ω_E. In sum, we have found $\boldsymbol{H}_\perp^{-1/2}(\operatorname{curl}_\Gamma,\Gamma)$ as the right space for the Dirichlet-data $\gamma_{\mathbf{t}}\mathbf{u}$. Conversely the Neumann-data $(\gamma_N,\gamma_{\mathbf{n}})$ should belong to $\boldsymbol{H}_{||}^{-1/2}(\operatorname{div}_\Gamma 0,\Gamma)\times H^{-1/2}(\Gamma)$, where

$$\boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma 0,\Gamma) := \{\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma,\Gamma),\ \operatorname{div}_\Gamma\boldsymbol{\lambda} = 0\}\ .$$

These are the domain and range, respectively, of the Dirichlet-to-Neumann map for the exterior eddy current problem. However, this map is not well defined in the case of relaxed constraints, as the normal trace $\gamma_{\mathbf{n}}$ is nonzero for harmonic Dirichlet fields. However, the trace γ_N is *not* affected and turns out unique for given Dirichlet-data.

The gist of these considerations is that a part of the Neumann data for the exterior eddy current problem is not captured by the transmission conditions. Fittingly, it is exactly that component that is not unique without gauging. This reflects that gauging is redundant as far as the solution in Ω_C is concerned. As we will see in the next section, we have to retain the divergence constraint, nevertheless.

5. Potentials. As usual, the correct boundary integral equations for the exterior eddy current problem emerge from a representation formula. It involves the singular function for the Laplacian in three dimensions:

$$E(\mathbf{x},\mathbf{y}) := \frac{1}{4\pi}\frac{1}{|\mathbf{x}-\mathbf{y}|}\ , \quad \mathbf{x},\mathbf{y} \in \mathbb{R}^3,\ \mathbf{x} \neq \mathbf{y}\ .$$

As in [33, sect. 6.2], for any Lipschitz-domain $\Omega \subset \mathbb{R}^3$ with exterior unit normal $\mathbf{n}$, we arrive at the following representation formula for vectorfields $\mathbf{u} \in C^2(\bar{\Omega})^3$ with $\operatorname{div}\mathbf{u}$ and $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}$ compactly supported and decaying like $\mathbf{u}(\mathbf{x}) = O(|\mathbf{x}|^{-1})$ and $\mathbf{curl}\,\mathbf{u}(\mathbf{x}) = o(|\mathbf{x}|^{-1})$ uniformly for $|\mathbf{x}| \to \infty$:

$$\begin{aligned}\mathbf{u}(\mathbf{x}) = &-\mathbf{curl}_{\mathbf{x}}\int_\Gamma (\mathbf{n}\times\mathbf{u})(\mathbf{y})E(\mathbf{x},\mathbf{y})\,dS(\mathbf{y}) + \int_\Gamma (\mathbf{n}\times\mathbf{curl}\,\mathbf{u})(\mathbf{y})E(\mathbf{x},\mathbf{y})\,dS(\mathbf{y})\\ &+\operatorname{grad}_{\mathbf{x}}\int_\Gamma (\mathbf{n}\cdot\mathbf{u})(\mathbf{y})E(\mathbf{x},\mathbf{y})\,dS(\mathbf{y}) + \int_\Omega \mathbf{curl}\,\mathbf{curl}\,\mathbf{u}(\mathbf{y})E(\mathbf{x},\mathbf{y})\,d\mathbf{y}\\ &-\int_\Omega \operatorname{div}\mathbf{u}(\mathbf{y})\operatorname{grad}_{\mathbf{x}}E(\mathbf{x},\mathbf{y})\,d\mathbf{y}, \qquad \mathbf{x}\in\Omega\ .\end{aligned} \tag{5.1}$$

It is an analogue to the Stratton–Chu formulas for the full Maxwell's equations (see [33, Thm. 6.1], [59, sect. 2], and [65, sect. 5.5]).

Evidently, information on the divergence is crucial in the derivation of the representation formula. This is the fundamental reason why we have to keep the zero divergence constraint for the exterior eddy current problem.

We recall the definition of the scalar single layer potential [27, Chap. 6]:

$$\Psi_V(\phi)(\mathbf{x}) := \int_\Gamma \phi(\mathbf{y}) E(\mathbf{x},\mathbf{y})\, dS(\mathbf{y}) , \quad \mathbf{x} \notin \Gamma . \tag{5.2}$$

It is well known that it can be extended to a continuous mapping $\Psi_V : H^{-1/2}(\Gamma) \mapsto H^1_{\mathrm{loc}}(\mathbb{R}^3)$, that it satisfies certain jump relations, and that it induces a Hermitian $H^{-1/2}(\Gamma)$-elliptic bilinear form [39, Chap. XI, sect. 2, Thm. 3]. We aim to establish similar results for the potentials occurring in (5.1), too. To this end we closely follow the approach in [35]. First, for a tangential vectorfield $\boldsymbol{\lambda}$ on Γ, we define the vectorial single layer potential by

$$\boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})(\mathbf{x}) := \int_\Gamma \boldsymbol{\lambda}(\mathbf{y}) E(\mathbf{x},\mathbf{y})\, dS(\mathbf{y}) , \quad \mathbf{x} \notin \Gamma . \tag{5.3}$$

Introducing the vectorial Newton-potential as

$$\mathbf{G}(\boldsymbol{\lambda})(\mathbf{x}) := \int_{\mathbb{R}^3} \boldsymbol{\lambda}(\mathbf{y}) E(\mathbf{x},\mathbf{y})\, d\mathbf{y} , \tag{5.4}$$

which defines a continuous mapping $\mathbf{G} : \boldsymbol{H}^{-1}_{\mathrm{comp}}(\mathbb{R}^3) \mapsto \boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$, we formally get for a smooth compactly supported test field $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)$

$$\begin{aligned}\int_{\mathbb{R}^3} \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})(\mathbf{x}) \cdot \overline{\boldsymbol{\Phi}(\mathbf{x})}\, d\mathbf{x} &= \int_{\mathbb{R}^3} \int_\Gamma E(\mathbf{x},\mathbf{y}) \boldsymbol{\lambda}(\mathbf{y}) \cdot \overline{\boldsymbol{\Phi}(\mathbf{x})}\, dS(\mathbf{y})\, d\mathbf{x} \\ &= \int_\Gamma \boldsymbol{\lambda}(\mathbf{y}) \cdot \overline{\mathbf{G}(\boldsymbol{\Phi})(\mathbf{y})}\, dS(\mathbf{y}) .\end{aligned}$$

We conclude

$$(\boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}), \boldsymbol{\Phi})_{0;\mathbb{R}^3} = (\boldsymbol{\lambda}, \gamma_{\mathbf{t}} \mathbf{G}(\boldsymbol{\Phi}))_{0;\Gamma} = (\gamma_{\mathbf{t}}{}^* \boldsymbol{\lambda}, \mathbf{G}(\boldsymbol{\Phi}))_{0;\mathbb{R}^3} = (\mathbf{G}(\gamma_{\mathbf{t}}{}^* \boldsymbol{\lambda}), \boldsymbol{\Phi})_{0;\mathbb{R}^3} ,$$

which means, by density,

$$\boldsymbol{\Psi}_{\mathbf{A}}^* = \gamma_{\mathbf{t}} \circ \mathbf{G} \quad \Leftrightarrow \quad \boldsymbol{\Psi}_{\mathbf{A}} = \mathbf{G} \circ \gamma_{\mathbf{t}}{}^* . \tag{5.5}$$

THEOREM 5.1. $\boldsymbol{\Psi}_{\mathbf{A}} : \boldsymbol{H}_{||}^{-1/2}(\Gamma) \mapsto \boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$ *is a continuous operator.*

Proof. By Theorem 3.1, the L^2-adjoint $\gamma_{\mathbf{t}}{}^* : \boldsymbol{H}_{||}^{-1/2}(\Gamma) \mapsto \boldsymbol{H}^{-1}(\mathbb{R}^3)$ is continuous. Then the assertion is immediate from (5.5). ∎

The next lemma supplies an important auxiliary relationship.

LEMMA 5.2. *For* $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma, \Gamma)$ *we have*

$$\mathrm{div}\, \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) = \Psi_V(\mathrm{div}_\Gamma \boldsymbol{\lambda}) \quad \textit{in } \boldsymbol{L}^2(\mathbb{R}^3) .$$

Proof. Cf. the proof of Lemma 2.3 in [59]. ∎

From the decay properties of the kernel we infer $\boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) \in W^{1,-1}(\mathbb{R}^3)$. Now, we confine ourselves to $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0, \Gamma)$. According to the previous lemma, this implies $\mathrm{div}\, \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) = 0$, and, since $\Delta \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) = 0$ away from Γ, $\mathbf{curl}\,\mathbf{curl}\, \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})$ is seen to vanish in $\mathbb{R}^3 \setminus \Gamma$ as well. Thus, $\gamma_N \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})$ is well defined from both sides of Γ.

Now we can apply Green's formula to Ω_C and Ω_E separately and add up the expressions: For all $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)$

$$\int_{\mathbb{R}^3} \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) \cdot \mathbf{curl}\,\mathbf{curl}\, \boldsymbol{\Phi}\, d\mathbf{x} = \langle [\gamma_D \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})]_\Gamma, \gamma_N \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} - \langle [\gamma_N \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})]_\Gamma, \gamma_D \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} .$$

On the other hand, using $\operatorname{div} \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) = 0$, we get

$$\begin{aligned}\int_{\mathbb{R}^3} \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})(\mathbf{x}) \cdot \mathbf{curl}\,\mathbf{curl}\,\boldsymbol{\Phi}(\mathbf{x})\,d\mathbf{x} &= -\int_{\mathbb{R}^3} \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})(\mathbf{x}) \cdot \boldsymbol{\Delta}\boldsymbol{\Phi}(\mathbf{x})\,d\mathbf{x} \\ &= -\left\langle \boldsymbol{\lambda}, (\boldsymbol{\Psi}_{\mathbf{A}}^* \circ \boldsymbol{\Delta})\boldsymbol{\Phi} \right\rangle_{\boldsymbol{\tau}} = \left\langle \boldsymbol{\lambda}, \gamma_D \boldsymbol{\Phi} \right\rangle_{\boldsymbol{\tau}}.\end{aligned}$$

We have made use of (5.5), i.e., $\boldsymbol{\Psi}_{\mathbf{A}}^* = \gamma_{\mathbf{t}} \circ \mathbf{G}$, and $\mathbf{G} \circ \Delta = -Id$. Next, we see from the regularity result of Theorem 5.1 that $[\gamma_D \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})]_\Gamma = 0$. Subtracting the previous equation, we establish

$$[\gamma_N \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})]_\Gamma = -\boldsymbol{\lambda} \quad \text{in } \boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma), \tag{5.6}$$

thanks to the density of $\boldsymbol{\mathcal{D}}(\mathbb{R}^3)_{|\Gamma}$ in $\boldsymbol{H}_{||}^{-1/2}(\operatorname{div}_\Gamma, \Gamma)$. This is the desired *jump relation* for the vectorial single layer potential.

The vectorial double layer potential for a tangential vectorfield $\mathbf{u}$ is given by

$$\boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u}) := \mathbf{curl}\,\boldsymbol{\Psi}_{\mathbf{A}}(R\mathbf{u}), \quad R\mathbf{u} := \mathbf{n} \times \mathbf{u}\,. \tag{5.7}$$

From the regularity of the vectorial single layer potential (Theorem 5.1) and the fact that $R : \boldsymbol{H}_{||}^{-1/2}(\Gamma) \mapsto \boldsymbol{H}_{\perp}^{-1/2}(\Gamma)$ is an isometry, we can infer that $\boldsymbol{\Psi}_{\mathbf{M}}$ is a continuous mapping from $\boldsymbol{H}_{\perp}^{-1/2}(\Gamma)$ to $\boldsymbol{H}(\operatorname{div}; \mathbb{R}^3)$.

Armed with these regularity results and the continuity of the trace mappings, we can extend the representation formula (5.1) to vectorfields $\mathbf{u} \in \boldsymbol{W}(\mathbf{curl}, \mathbb{R}^3)$ decaying according to (2.5) and satisfying $\operatorname{div} \mathbf{u}_{|\Omega_C} \in L^2(\Omega_C)$, $\operatorname{div} \mathbf{u}_{|\Omega_E} \in L^2(\Omega_E)$ compactly supported, $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}_{|\Omega_C} \in \boldsymbol{L}^2(\Omega_C)$, and $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}_{|\Omega_E} \in \boldsymbol{L}^2(\Omega_E)$ compactly supported: We get the *transmission formula*

$$\begin{aligned}\mathbf{u} = &-\mathbf{G}(\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}) + \operatorname{grad}\mathbf{G}(\operatorname{div}\mathbf{u}) + \boldsymbol{\Psi}_{\mathbf{M}}([\gamma_D\mathbf{u}]_\Gamma) \\ &- \boldsymbol{\Psi}_{\mathbf{A}}([\gamma_N\mathbf{u}]_\Gamma) - \operatorname{grad}\Psi_V([\gamma_{\mathbf{n}}\mathbf{u}]_\Gamma).\end{aligned} \tag{5.8}$$

Let us write $\mathfrak{S} : \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{W}(\mathbf{curl}^2, \Omega) \cap \boldsymbol{H}(\operatorname{div} 0; \Omega)$ for a continuous solution operator for the Dirichlet problem (4.2) on both sides of Γ. It fulfills (for $\mathbf{g} \in \boldsymbol{H}_{\perp}^{-1/2}(\operatorname{curl}_\Gamma, \Gamma)$)

$$\begin{aligned}\mathfrak{S}\,\mathbf{g} &= \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{g}) - \boldsymbol{\Psi}_{\mathbf{A}}(\gamma_N\,\mathfrak{S}\,\mathbf{g}) - \operatorname{grad}\Psi_V(\gamma_{\mathbf{n}}\,\mathfrak{S}\,\mathbf{g}) \quad \text{in } \Omega_E\,, \\ \mathfrak{S}\,\mathbf{g} &= -\boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{g}) + \boldsymbol{\Psi}_{\mathbf{A}}(\gamma_N\,\mathfrak{S}\,\mathbf{g}) + \operatorname{grad}\Psi_V(\gamma_{\mathbf{n}}\,\mathfrak{S}\,\mathbf{g}) \quad \text{in } \Omega_C\,.\end{aligned} \tag{5.9}$$

Of course, the introduction of $\mathfrak{S}$ entails some gauging of $\mathfrak{S}\,\mathbf{g}$ to render it well defined. As the kind of gauging involved ultimately turns out to be immaterial, we do not have to bother about its details.

THEOREM 5.3. *$\boldsymbol{\Psi}_{\mathbf{M}}$ is a continuous mapping from $\boldsymbol{H}_{\perp}^{-1/2}(\operatorname{curl}_\Gamma, \Gamma)$ to $\boldsymbol{H}(\operatorname{div} 0; \Omega) \cap \boldsymbol{H}(\mathbf{curl}; \Omega)$, for both $\Omega = \Omega_C$ and $\Omega = \Omega_E$.*

Proof. Equations (5.9) mean that, e.g., in Ω_E, $\boldsymbol{\Psi}_{\mathbf{M}} = (Id - \boldsymbol{\Psi}_{\mathbf{A}} \circ \gamma_N + \operatorname{grad} \circ \Psi_V \circ \gamma_{\mathbf{n}}) \circ \mathfrak{S}$. Employing the continuity properties of the other potentials and of the trace operators, we get the result. ∎

We may plug an (arbitrarily gauged) solution of the Dirichlet problem (4.2) into the transmission formula and extend $\mathbf{u}$ to the other side of Γ by solving a Neumann problem (4.3) such that $[\gamma_{\mathbf{n}}\mathbf{u}]_\Gamma = 0$ and $[\gamma_N\mathbf{u}]_\Gamma = 0$. In the end, we have $\mathbf{u} = \boldsymbol{\Psi}_{\mathbf{M}}([\gamma_D\mathbf{u}]_\Gamma)$ and can state $\operatorname{Im}(\boldsymbol{\Psi}_{\mathbf{M}}) \subset \boldsymbol{W}(\mathbf{curl}^2, \Omega_C) \times \boldsymbol{W}(\mathbf{curl}^2, \Omega_E)$. So it is legal to evaluate the Neumann trace γ_N for the vectorial double layer potential.

Next, we aim to establish the jump relations for $\boldsymbol{\Psi}_{\mathbf{M}}$. Note that from (5.5)

$$\boldsymbol{\Psi}_{\mathbf{M}} = \mathbf{curl} \circ \boldsymbol{\Psi}_{\mathbf{A}} \circ R = \mathbf{curl} \circ \mathbf{G} \circ {\gamma_{\mathbf{t}}}^* \circ R = \mathbf{curl} \circ \mathbf{G} \circ (\gamma_{\mathbf{t}}^{\times})^*,$$

where I used $R^* = -R$. By virtue of $\mathbf{curl}\,\mathbf{curl}\,\mathbf{curl} = -\,\mathbf{curl} \circ \boldsymbol{\Delta}$ and $\mathbf{curl}^* = \mathbf{curl}$, this implies

$$\mathbf{curl}\,\mathbf{curl}\,\boldsymbol{\Psi}_{\mathbf{M}} = -\,\mathbf{curl} \circ (\boldsymbol{\Delta} \circ \mathbf{G}) \circ (\gamma_{\mathbf{t}}^{\times})^* = (\gamma_{\mathbf{t}}^{\times} \circ \mathbf{curl})^* = \gamma_N^* \tag{5.10}$$

in the sense of distributions. As above, we get from Green's formula for $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)$

$$\int_{\mathbb{R}^3} \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u}) \cdot \mathbf{curl}\,\mathbf{curl}\,\overline{\boldsymbol{\Phi}}\, d\mathbf{x} = \langle [\gamma_D \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma, \gamma_N \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} - \langle [\gamma_N \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma, \gamma_D \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} .$$

Combining this with (5.10) confirms that for all $\boldsymbol{\Phi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^3)$

$$\langle [\gamma_D \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma - \mathbf{u}, \gamma_N \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} - \langle [\gamma_N \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma, \gamma_D \boldsymbol{\Phi} \rangle_{\boldsymbol{\tau}} = 0 . \tag{5.11}$$

As we can see along the lines of the proof of Lemma 1.5.39 in [43], $\boldsymbol{\mathcal{D}}(\Omega)$ is dense in $\boldsymbol{W}(\mathbf{curl}^2, \Omega)$. This means that we can equivalently state (5.11) to hold for all $\boldsymbol{\Phi} \in \boldsymbol{W}(\mathbf{curl}^2, \Omega)$. Finally, let $\mathbf{p} \in \boldsymbol{W}(\mathbf{curl}^2, \Omega)$ be a solution of the Dirichlet problem

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{p} = \mathbf{f} \text{ in } \Omega, \quad \operatorname{div} \mathbf{p} = 0 \text{ in } \Omega, \quad \gamma_D \mathbf{p} = 0 \text{ on } \Gamma ,$$

with $\mathbf{f} \in \boldsymbol{H}(\operatorname{div} 0; \Omega)$ compactly supported. On top of that, define $\boldsymbol{\eta} \in \boldsymbol{W}(\mathbf{curl}^2, \Omega)$ as a solution of the Dirichlet problem

$$\mathbf{curl}\,\mathbf{curl}\,\boldsymbol{\eta} = 0 \text{ in } \Omega, \quad \operatorname{div} \boldsymbol{\eta} = 0 \text{ in } \Omega, \quad \gamma_D \boldsymbol{\eta} = [\gamma_D \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma - \mathbf{u} \text{ on } \Gamma .$$

By using the definition of the Neumann trace operator γ_N,

$$\begin{aligned} 0 &= \langle \gamma_D \boldsymbol{\eta}, \gamma_N \mathbf{p} \rangle_{\boldsymbol{\tau}} = \langle \gamma_D \boldsymbol{\eta}, \gamma_N \mathbf{p} \rangle_{\boldsymbol{\tau}} - \langle \gamma_N \boldsymbol{\eta}, \gamma_D \mathbf{p} \rangle_{\boldsymbol{\tau}} \\ &= (\mathbf{curl}\,\mathbf{curl}\,\boldsymbol{\eta}, \mathbf{p})_{0;\Omega} - (\boldsymbol{\eta}, \mathbf{curl}\,\mathbf{curl}\,\mathbf{p})_{0;\Omega} = -\,(\boldsymbol{\eta}, \mathbf{f})_{0;\Omega} . \end{aligned}$$

Hence, $\boldsymbol{\eta}$ must vanish, as $\operatorname{div} \boldsymbol{\eta} = 0$. At this stage, we already know

$$[\gamma_D \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma = \mathbf{u} \quad \text{in } \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_\Gamma, \Gamma) , \tag{5.12}$$

and from (5.11) $[\gamma_N \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u})]_\Gamma = 0$ is readily available. This finishes the proof of the jump relations for the vectorial double layer potential.

6. Boundary integral operators. The regularity properties demonstrated in the previous section pave the way for defining related boundary integral operators

$$\begin{aligned} \mathbf{A}\boldsymbol{\lambda} &:= \gamma_D \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}), & \mathbf{B}\boldsymbol{\lambda} &:= \gamma_N^+ \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda}) , \\ \mathbf{C}\mathbf{u} &:= \gamma_D^+ \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u}), & \mathbf{N}\mathbf{u} &:= \gamma_N \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{u}) , \\ \mathbf{S}\phi &:= \gamma_D^+(\operatorname{grad} \Psi_V(\phi)). \end{aligned}$$

Alternatively, we could have written $\mathbf{S}\phi := \operatorname{grad}_\Gamma \Psi_V(\phi)$. The continuity of the potential mappings and that of the trace operators bear out the following theorem.

THEOREM 6.1. *The mappings*

$$\begin{aligned} &\mathbf{A} : \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma), & &\mathbf{B} : \boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma) , \\ &\mathbf{C} : \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_\Gamma, \Gamma), & &\mathbf{N} : \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_\Gamma, \Gamma) \mapsto \boldsymbol{H}_{||}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma) , \\ &\mathbf{S} : H^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma) \end{aligned}$$

are continuous.

Next, pick arbitrary $\mathbf{p},\mathbf{q} \in \boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$, $\boldsymbol{\zeta},\boldsymbol{\eta} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ and define solutions to the transmission problem for the exterior eddy current problem by

$$\begin{aligned}\mathbf{u} &:= \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{p}) + \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\zeta}), \quad \text{in } \Omega_C \cup \Omega_E\ ,\\ \mathbf{v} &:= \boldsymbol{\Psi}_{\mathbf{M}}(\mathbf{q}) + \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\eta}), \quad \text{in } \Omega_C \cup \Omega_E\ .\end{aligned}$$

By definition,

$$\begin{aligned}\gamma_D^+\mathbf{u} &= \mathbf{C}(\mathbf{p}) + \mathbf{A}(\boldsymbol{\zeta}), \quad \gamma_N^+\mathbf{u} = \mathbf{N}(\mathbf{p}) + \mathbf{B}(\boldsymbol{\zeta})\ ,\\ \gamma_D^+\mathbf{v} &= \mathbf{C}(\mathbf{q}) + \mathbf{A}(\boldsymbol{\eta}), \quad \gamma_N^+\mathbf{v} = \mathbf{N}(\mathbf{q}) + \mathbf{B}(\boldsymbol{\eta})\ ,\end{aligned}$$

and from the jump relations we get

$$\begin{aligned}\gamma_D^-\mathbf{u} &= \mathbf{C}(\mathbf{p}) - \mathbf{p} + \mathbf{A}(\boldsymbol{\zeta}), \quad \gamma_N^-\mathbf{u} = \mathbf{N}(\mathbf{p}) + \mathbf{B}(\boldsymbol{\zeta}) + \boldsymbol{\zeta}\ ,\\ \gamma_D^-\mathbf{v} &= \mathbf{C}(\mathbf{q}) - \mathbf{q} + \mathbf{A}(\boldsymbol{\eta}), \quad \gamma_N^-\mathbf{v} = \mathbf{N}(\mathbf{q}) + \mathbf{B}(\boldsymbol{\eta}) + \boldsymbol{\eta}.\end{aligned}$$

As $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} = 0$ and $\mathbf{curl}\,\mathbf{curl}\,\mathbf{v} = 0$ in $\mathbb{R}^3\backslash\Gamma$, Green's formula implies the identities

$$\left\langle \gamma_N^+\mathbf{v}, \gamma_D^+\mathbf{u}\right\rangle_{\boldsymbol{\tau}} = -\left(\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v}\right)_{0;\Omega_E} = \overline{\left\langle \gamma_N^+\mathbf{u}, \gamma_D^+\mathbf{v}\right\rangle_{\boldsymbol{\tau}}}\ , \tag{6.1}$$

$$\left\langle \gamma_N^-\mathbf{v}, \gamma_D^-\mathbf{u}\right\rangle_{\boldsymbol{\tau}} = \left(\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v}\right)_{0;\Omega_C} = \overline{\left\langle \gamma_N^-\mathbf{u}, \gamma_D^-\mathbf{v}\right\rangle_{\boldsymbol{\tau}}}\ . \tag{6.2}$$

By setting $\boldsymbol{\zeta} = 0$, $\boldsymbol{\eta} = 0$, we deduce

$$\begin{aligned}(6.1) \quad \Rightarrow \qquad \left\langle \mathbf{N}(\mathbf{q}), \mathbf{C}(\mathbf{p})\right\rangle_{\boldsymbol{\tau}} &= \overline{\left\langle \mathbf{N}(\mathbf{p}), \mathbf{C}(\mathbf{q})\right\rangle_{\boldsymbol{\tau}}}\ ,\\ (6.2) \quad \Rightarrow \qquad \left\langle \mathbf{N}(\mathbf{q}), (\mathbf{C}-Id)(\mathbf{p})\right\rangle_{\boldsymbol{\tau}} &= \overline{\left\langle \mathbf{N}(\mathbf{p}), (\mathbf{C}-Id)(\mathbf{q})\right\rangle_{\boldsymbol{\tau}}}\ ,\end{aligned}$$

which directly leads to

$$\left\langle \mathbf{N}(\mathbf{q}), \mathbf{p}\right\rangle_{\boldsymbol{\tau}} = \overline{\left\langle \mathbf{N}(\mathbf{p}), \mathbf{q}\right\rangle_{\boldsymbol{\tau}}}\ . \tag{6.3}$$

As we admitted any $\mathbf{p},\mathbf{q} \in \boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$, this shows that $\mathbf{N}$ is self-adjoint. In a parallel fashion, by setting $\mathbf{p} = 0$, $\mathbf{q} = 0$, it can be proved that

$$\left\langle \boldsymbol{\eta}, \mathbf{A}(\boldsymbol{\zeta})\right\rangle_{\boldsymbol{\tau}} = \overline{\left\langle \boldsymbol{\zeta}, \mathbf{A}(\boldsymbol{\eta})\right\rangle_{\boldsymbol{\tau}}} \quad \forall \boldsymbol{\zeta},\boldsymbol{\eta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0,\Gamma)\ . \tag{6.4}$$

Finally, we choose $\mathbf{p} = 0$, $\boldsymbol{\eta} = 0$ and observe

$$\begin{aligned}(6.1) \quad \Rightarrow \qquad \left\langle \mathbf{N}(\mathbf{q}), \mathbf{A}(\boldsymbol{\zeta})\right\rangle_{1/2,\Gamma} &= \overline{\left\langle \mathbf{B}(\boldsymbol{\zeta}), \mathbf{C}(\mathbf{q})\right\rangle_{\boldsymbol{\tau}}}\ ,\\ (6.2) \quad \Rightarrow \qquad \left\langle \mathbf{N}(\mathbf{q}), \mathbf{A}(\boldsymbol{\zeta})\right\rangle_{1/2,\Gamma} &= \overline{\left\langle (\mathbf{B}+Id)(\boldsymbol{\zeta}), (\mathbf{C}-Id)(\mathbf{q})\right\rangle_{\boldsymbol{\tau}}}\ ,\end{aligned}$$

which implies

$$\left\langle \mathbf{B}(\boldsymbol{\zeta}), \mathbf{q}\right\rangle_{\boldsymbol{\tau}} = \left\langle \boldsymbol{\zeta}, (\mathbf{C}-Id)\mathbf{q}\right\rangle_{\boldsymbol{\tau}} \quad \forall \mathbf{q} \in \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma,\Gamma), \boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0,\Gamma)\ . \tag{6.5}$$

The main properties of the operators $\mathbf{A}$ and $\mathbf{N}$, including "ellipticity," are summed up in the following theorems.

THEOREM 6.2. *The bilinear form on* $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ *induced by the operator* $\mathbf{A}$ *is symmetric and there is a constant* $c>0$ *depending on* Γ *such that*

$$\langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{\lambda}\rangle_{\boldsymbol{\tau}} \geq c\,\|\boldsymbol{\lambda}\|^2_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \qquad \forall \boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0,\Gamma)\ .$$

Proof. Symmetry can be concluded from (6.4). Pick $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ and set $\mathbf{u} := \boldsymbol{\Psi}_{\mathbf{A}}(\boldsymbol{\lambda})$, which provides a solution of the exterior eddy current problem in $\Omega_C \cup \Omega_E$. We apply Green's formula on both domains separately and also resort to the jump relations (5.6) for the vectorial simple layer potential:

$$\|\mathbf{curl}\,\mathbf{u}\|^2_{\boldsymbol{L}^2(\Omega_C\cup\Omega_E)} = -\langle [\gamma_N \mathbf{u}]_\Gamma, \gamma_D \mathbf{u}\rangle_{\boldsymbol{\tau}} = \langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{\lambda}\rangle_{1/2,\Gamma}\ .$$

Thus, Lemma 3.3 justifies the contention of the theorem. ∎

The examination of the operator $\mathbf{N}$ will rely on the following equivalent expression.

LEMMA 6.3. *For all* $\mathbf{u},\mathbf{v} \in \boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$

$$\langle \mathbf{N}(\mathbf{u}),\mathbf{v}\rangle_{\boldsymbol{\tau}} = -\langle \mathrm{curl}_\Gamma \mathbf{v}, V(\mathrm{curl}_\Gamma \mathbf{u})\rangle_{1/2,\Gamma}$$

holds, where $V : H^{-\frac{1}{2}}(\Gamma) \mapsto H^{\frac{1}{2}}(\Gamma)$ *is the ordinary scalar single layer potential operator on* Γ.

Proof. I refer to the derivation of formula (2.86) in [32]. ∎

THEOREM 6.4. *The bilinear form induced by the boundary integral operator* $\mathbf{N}$ *on* $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$ *is symmetric and negative semidefinite. In particular*

$$-\langle \mathbf{N}(\mathbf{u}),\mathbf{u}\rangle_{\boldsymbol{\tau}} \geq c\,\|\mathrm{curl}_\Gamma \mathbf{u}\|^2_{H^{-\frac{1}{2}}(\Gamma)} \qquad \forall \mathbf{u} \in \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma,\Gamma)\ ,$$

for some constant $c>0$.

Proof. Symmetry is a consequence of (6.3). For the estimate we simply rely upon the well-known $H^{-1/2}(\Gamma)$-ellipticity of V and use Lemma 6.3. ∎

Remark. The boundary integral operator that has been investigated in a Sobolev space setting in this section can also be considered in spaces of Hölder-continuous functions. This was done (for similar boundary integral operators) in [33, 32, 54].

7. Symmetric coupling. By setting $\mathbf{u} \equiv 0$ in Ω_C, the transmission formula (5.8) is turned into a representation formula for the (arbitrarily gauged) solutions $\mathbf{u}$ of the exterior eddy current problem in Ω_E:

$$\mathbf{u} = \boldsymbol{\Psi}_{\mathbf{M}}(\gamma_D^+\mathbf{u}) - \boldsymbol{\Psi}_{\mathbf{A}}(\gamma_N^+\mathbf{u}) - \mathrm{grad}\,\Psi_V(\gamma_{\mathbf{n}}^+\mathbf{u})\ . \tag{7.1}$$

Formally applying both trace operators γ_D^+ and γ_N^+ to (7.1), we arrive at

$$\gamma_D^+\mathbf{u} = \mathbf{C}(\gamma_D^+\mathbf{u}) - \mathbf{A}(\gamma_N^+\mathbf{u}) - \mathbf{S}(\gamma_{\mathbf{n}}^+\mathbf{u})\ , \tag{7.2}$$

$$\gamma_N^+\mathbf{u} = \mathbf{N}(\gamma_D^+\mathbf{u}) - \mathbf{B}(\gamma_N^+\mathbf{u})\ . \tag{7.3}$$

The first equation is set in $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$ and the second in $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$, which are the appropriate spaces of Dirichlet and Neumann data, as explained in section 3. The goal is to extract a *Calderón-projector* [27, sect. 4.5], [38] from (7.2) and (7.3). In a straightforward fashion, this is not possible, foiled by the presence of the extra Neumann data $\gamma_{\mathbf{n}}^+\mathbf{u}$. The idea is to get the equivalent of a weak Calderón-projector.

We attempt to achieve this goal by testing (7.2) against functions in $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ only. Apparently this sacrifices a lot of information: Whereas the dual space of $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$ is the entire space $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma,\Gamma)$, the dual of $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ is the orthogonal complement of $\mathrm{grad}_\Gamma H^{1/2}(\Gamma)$ in $\boldsymbol{H}_{\perp}^{-1/2}(\mathrm{curl}_\Gamma,\Gamma)$. (Orthogonality has to be understood with respect to the duality pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{\tau}}$.) On the other hand, the reward is clear from the observation

$$\langle \boldsymbol{\zeta}, \mathbf{S}(\phi)\rangle_{\boldsymbol{\tau}} = 0 \quad \forall \boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0,\Gamma) \tag{7.4}$$

because $\mathbf{S}(\phi) \in \mathrm{grad}_\Gamma H^{1/2}(\Gamma)$. It reveals that, after all, the test space $\boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ suppresses the impact of gauging because (7.4) can be used to get rid of the "artificial" Neumann data $\gamma_{\mathbf{n}}^+\mathbf{u}$ due to the gauge condition. Thus, we can derive the following variational equations from (7.2) and (7.3):

$$\begin{array}{rcccll} \langle \boldsymbol{\zeta}, \gamma_D^+\mathbf{u}\rangle_{\boldsymbol{\tau}} & = & \langle \boldsymbol{\zeta}, \mathbf{C}(\gamma_D^+\mathbf{u})\rangle_{\boldsymbol{\tau}} & - & \langle \boldsymbol{\zeta}, \mathbf{A}(\gamma_N^+\mathbf{u})\rangle_{\boldsymbol{\tau}} & \forall \boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma 0,\Gamma)\,, \\ \langle \gamma_N^+\mathbf{u}, \mathbf{w}\rangle_{\boldsymbol{\tau}} & = & \langle \mathbf{N}(\gamma_D^+\mathbf{u}), \mathbf{w}\rangle_{\boldsymbol{\tau}} & - & \langle \mathbf{B}(\gamma_N^+\mathbf{u}), \mathbf{w}\rangle_{\boldsymbol{\tau}} & \forall \mathbf{w} \in \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma,\Gamma)\,, \end{array} \tag{7.5}$$

which are satisfied by all solutions $\mathbf{u} \in \boldsymbol{W}(\mathbf{curl}^2,\Omega_E)$ of the exterior eddy current problem. This is the desired Calderón-projector in weak form.

Now, we can pursue the classical policy due to Costabel [34] in order to couple the interior problem in Ω_C with (7.5). Given a solution $\mathbf{u} \in \boldsymbol{W}(\mathbf{curl},\mathbb{R}^3) \cap \boldsymbol{W}(\mathbf{curl}^2,\Omega_E)$ of the full eddy current problem (1.3), we find through Green's formula and the coupling conditions (4.1) that

$$q(\mathbf{u},\mathbf{v}) - \langle \gamma_N^+\mathbf{u}, \gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} = -i\omega\,(\mathbf{j}_0,\mathbf{v})_{0;\Omega_C} \tag{7.6}$$

for all $\mathbf{v} \in \boldsymbol{H}(\mathbf{curl};\Omega_C)$. For the sake of brevity I have set

$$q(\mathbf{u},\mathbf{v}) := \left(\tfrac{1}{\mu}\,\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v}\right)_{0;\Omega_C} + i\omega\,(\sigma\mathbf{u},\mathbf{v})_{0;\Omega_C}\ .$$

Equation (7.6) with $\gamma_N^+\mathbf{u}$ replaced by $\boldsymbol{\lambda}$ is added to the second equation of (7.5) tested with $\gamma_D\mathbf{v}$. The first equation of (7.5) completes the coupled variational problem: Seek $\mathbf{u} \in \boldsymbol{H}(\mathbf{curl};\Omega_C)$, $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ such that

$$\begin{array}{rcccll} q(\mathbf{u},\mathbf{v}) - \langle \mathbf{N}(\gamma_D\mathbf{u}), \gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} & + & \langle \mathbf{B}\boldsymbol{\lambda}, \gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} & = & -i\omega\,(\mathbf{j}_0,\mathbf{v})_{0;\Omega_C}\,, \\ \langle \boldsymbol{\zeta}, (Id-\mathbf{C})\gamma_D\mathbf{u}\rangle_{\boldsymbol{\tau}} & + & \langle \boldsymbol{\zeta}, \mathbf{A}\boldsymbol{\lambda}\rangle_{\boldsymbol{\tau}} & = & 0 \end{array} \tag{7.7}$$

for all $\mathbf{v} \in \boldsymbol{H}(\mathbf{curl};\Omega_C)$, $\boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$.

For ease of presentation, I am going to write $\mathbf{d}(\cdot,\cdot)$ for the sesquilinear form on $\mathcal{W} := \boldsymbol{H}(\mathbf{curl};\Omega_C) \times \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$ spawning the variational problem (7.7). The right-hand side will be designated by $f \in \boldsymbol{H}(\mathbf{curl};\Omega_C)'$. Existence and uniqueness of a solution $(\mathbf{u},\boldsymbol{\lambda})$ of (7.7) follow from the next theorem by the Lax–Milgram lemma.

THEOREM 7.1. *The sesquilinear form* $\mathbf{d}$ *is* $\mathcal{W}$*-elliptic and continuous.*

Proof. The continuity of the sesquilinear form is an immediate consequence of the continuity of the boundary integral operators (Theorem 6.1) and that of the traces (Theorem 3.2). Owing to (6.5), the operator in (7.7) is block skew-symmetric:

$$\begin{aligned}
&|\mathbf{d}((\mathbf{v},\boldsymbol{\zeta}),(\mathbf{v},\boldsymbol{\zeta}))|\\
&= \left|q(\mathbf{v},\mathbf{v}) - \langle \mathbf{N}(\gamma_D\mathbf{v}),\gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} + \langle \mathbf{B}\boldsymbol{\zeta},\gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} + \langle \boldsymbol{\zeta},(Id-\mathbf{C})\gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} + \langle \boldsymbol{\zeta},\mathbf{A}\boldsymbol{\zeta}\rangle_{1/2,\Gamma}\right|\\
&= \left|q(\mathbf{v},\mathbf{v}) - \langle \mathbf{N}(\gamma_D\mathbf{v}),\gamma_D\mathbf{v}\rangle_{\boldsymbol{\tau}} + \langle \mathbf{A}\boldsymbol{\zeta},\boldsymbol{\zeta}\rangle_{1/2,\Gamma}\right|\\
&\geq c\left(\left(\tfrac{1}{\mu}\,\mathbf{curl}\,\mathbf{v},\mathbf{curl}\,\mathbf{v}\right)_{0;\Omega_C} + (\sigma\mathbf{v},\mathbf{v})_{0;\Omega_C} + \|\mathrm{curl}_\Gamma\,\gamma_D\mathbf{v}\|^2_{H^{-\frac{1}{2}}(\Gamma)} + \|\boldsymbol{\zeta}\|^2_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)}\right)\\
&\geq c\left(\|\mathbf{v}\|^2_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} + \|\boldsymbol{\zeta}\|^2_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\right),
\end{aligned}$$

with generic constants $c > 0$ independent of $\mathbf{v} \in \boldsymbol{H}(\mathbf{curl};\Omega_C)$ and $\boldsymbol{\zeta} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0,\Gamma)$. The estimates relied on Theorems 6.2 and 6.4 and used the uniform boundedness of μ and σ (the latter in Ω_C). ∎

Uniqueness of the solution for the coupled problem ultimately justifies our approach. We know that, taking $\mathbf{u} := \mathbf{E}$ from (2.1), $\mathbf{u}_{|\Omega_C}$ and $\gamma_N\mathbf{u}$ are unique (even independent of gauging). The derivation of (7.7) ensures that a solution of (2.1) will always give rise to a corresponding solution of the coupled problem. Uniqueness guarantees that we get the one correct answer.

Remark. The unknown $\boldsymbol{\lambda}$ corresponds to the equivalent surface current density $\mathbf{H}\times\mathbf{n}$ scaled by $-i\omega$.

Remark. In the case of good conductors, smooth surfaces, and high frequency ω, a skin-effect approximation is feasible, which neglects the interior of the conductors. Instead, impedance boundary conditions $\gamma_{\mathbf{t}}\mathbf{E} = \eta\gamma_{\mathbf{t}}^{\times}\mathbf{H}$ are imposed on Γ, where $\eta := (1+i)\sqrt{\omega\mu/2\sigma} \in L^\infty(\Gamma)$ is the surface impedance expressed through the material parameters that are expected to prevail on Γ [75] and [32, sect. 4.7]. In this case we should replace (7.6) with

$$\left(\eta^{-1}\gamma_D\mathbf{u},\mathbf{v}\right)_{0;\Gamma} - \left\langle\gamma_N^+\mathbf{u},\gamma_D\mathbf{v}\right\rangle_{\boldsymbol{\tau}} = 0 \quad \forall\mathbf{v}\in\boldsymbol{H}_\perp^{-\frac{1}{2}}(\mathrm{curl}_\Gamma,\Gamma)\;. \tag{7.8}$$

This will give an analogue of the coupled problem (7.7) with $\boldsymbol{\lambda}$ and $\gamma_D\mathbf{u}$ as unknowns. Of course, the sources $\mathbf{j}_0$ are no longer meaningful, but prescribing total currents in conducting loops continues to make sense.

Yet, (7.8) is pointless without higher regularity of $\gamma_D\mathbf{u}$ so that we have to seek $\gamma_D\mathbf{u}$ in the space $\boldsymbol{L}^2(\Gamma)$. After this alteration, all the above considerations carry over to the skin-effect model, because $\Re\eta > 0$.

8. Discretization. The domain Ω_C is equipped with a triangulation Ω_h in the sense of [29, Chap. 2, sect. 2.2], which may consist of tetrahedra, hexahedra, and prisms. It induces a surface mesh Γ_h of Γ composed of triangles and quadrangles.

As conforming finite element space for approximation in $\boldsymbol{H}(\mathbf{curl};\Omega_C)$, we use edge elements [64, 14] and adopt the notation $\boldsymbol{\mathcal{NED}}_1(\Omega_h)$ for the resulting finite element space. Edge elements can be regarded as discrete 1-forms [13] and they match the tangential continuity of the electric field. In the presence of re-entrant edges of Ω_C or discontinuous material parameters their use is mandatory, because the fields might not belong to $\boldsymbol{H}^1(\Omega_C)$ [37, 36].

The Neumann data $\boldsymbol{\lambda} \in \boldsymbol{H}_{||}^{-1/2}(\mathrm{div}_\Gamma 0, \Gamma)$ require an approximation by means of solenoidal div_Γ-conforming surface finite elements. They are supplied by the space $\boldsymbol{\mathcal{RT}}_0^0(\Gamma_h) := \{\boldsymbol{\lambda}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h),\, \mathrm{div}_\Gamma \boldsymbol{\lambda}_h = 0\}$ of divergence-free lowest order Raviart–Thomas elements on Γ_h [67]. It is worth noting that both kinds of finite elements form affine equivalent families in the sense of [29] based on suitable co- and contravariant transformations of vectorfields [64, 20]. These transformations also make it possible to extend the definition of the finite element spaces to curved elements of Ω_h and Γ_h [9].

Setting $\mathcal{W}_h := \boldsymbol{\mathcal{ND}}_1(\Omega_h) \times \boldsymbol{\mathcal{RT}}_0^0(\Gamma_h) \subset \mathcal{W}$ for the discrete approximation space, the discrete problem can be stated as follows: Seek $(\mathbf{u}_h, \boldsymbol{\lambda}_h) \in \mathcal{W}_h$ such that

$$\mathbf{d}((\mathbf{u}_h, \boldsymbol{\lambda}_h), (\mathbf{v}_h, \boldsymbol{\zeta}_h)) = f(\mathbf{v}_h) \quad \forall (\mathbf{v}_h, \boldsymbol{\zeta}_h) \in \mathcal{W}_h \,. \tag{8.1}$$

As before, we get existence and uniqueness of a discrete solution $(\mathbf{u}_h, \boldsymbol{\lambda}_h)$, and from Theorem 7.1 we conclude the a priori error estimate

$$\begin{aligned} &\|\mathbf{u} - \mathbf{u}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \\ &\qquad \leq C \inf_{(\mathbf{v}_h, \boldsymbol{\zeta}_h) \in \mathcal{W}_h} \left\{ \|\mathbf{u} - \mathbf{v}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} + \|\boldsymbol{\lambda} - \boldsymbol{\zeta}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \right\} , \end{aligned} \tag{8.2}$$

with a constant $C > 0$ depending only on the geometry and the material parameters σ and μ. A more concrete estimate is also available.

THEOREM 8.1. *Let $h > 0$ be the meshwidth of Ω_h and assume that the solution $\mathbf{u}$ of the eddy current problem is as regular as $\mathbf{u} \in \boldsymbol{H}^s(\Omega_C)$, $\mathbf{curl}\,\mathbf{u} \in \boldsymbol{H}^s(\Omega_C)$, $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} \in \boldsymbol{H}^s(\Omega_C)$ for some $s > \frac{1}{2}$. Then*

$$\begin{aligned} &\|\mathbf{u} - \mathbf{u}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \\ &\qquad \leq C' h^{\min\{s,1\}} \left(\|\mathbf{u}\|_{\boldsymbol{H}^s(\Omega_C)} + \|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{H}^s(\Omega_C)} + \|\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{H}^s(\Omega_C)} \right) , \end{aligned}$$

where $C' > 0$ depends on C from (8.2) *and the shape-regularity constants* (*cf.* [29, *Chap.* 3, *sect.* 3.1]) *of the triangulation Ω_h.*

Proof. Recall the canonical interpolation operators Π and Π^Γ for the spaces $\boldsymbol{\mathcal{ND}}_1(\Omega_h)$ and $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ [64, 20]. Π is well defined for vectorfields satisfying the assumptions of the theorem (cf. [5, Lemma 4.7]) and there holds (cf. [30, Lemmas 3.2, 3.3])

$$\|\mathbf{u} - \Pi\,\mathbf{u}\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} \leq C h^{\min\{s,1\}} \left(\|\mathbf{u}\|_{\boldsymbol{H}^s(\Omega_C)} + \|\mathbf{curl}\,\mathbf{u}\|_{\boldsymbol{H}^s(\Omega_C)} \right) , \tag{8.3}$$

with a constant merely depending on the shape-regularity of the mesh. Hence, we may choose $\mathbf{v}_h = \Pi\,\mathbf{u}$ on the right-hand side of (8.2).

We set $\mathbf{q}_h := \Pi\,\mathbf{curl}\,\mathbf{u}$, which is also well defined for the given $\mathbf{u}$. We may then pick $\boldsymbol{\zeta}_h := \gamma_\mathbf{t}^\times \mathbf{q}_h$ because we have the commuting relationship $\gamma_\mathbf{t}^\times \circ \Pi = \Pi^\Gamma \circ \gamma_\mathbf{t}^\times$ for sufficiently smooth vectorfields (cf. [49]). Since $\mathrm{div}_\Gamma(\gamma_N \mathbf{u}) = 0$ on Γ, $\gamma_\mathbf{n}(\mathbf{curl}\,\mathbf{curl}\,\mathbf{u}) = 0$, too. Then we may use the commuting diagram property [49, Thm. 13]

$$\mathrm{div}_\Gamma \circ \Pi^\Gamma \circ \gamma_\mathbf{t}^\times = Q^\Gamma \circ \mathrm{div}_\Gamma \circ \gamma_\mathbf{t}^\times = Q^\Gamma \circ \gamma_\mathbf{n} \circ \mathbf{curl}\,,$$

where Q^Γ stands for the $L^2(\Gamma)$-orthogonal projection onto the space of piecewise constants with respect to Γ_h. In short, we get $\mathrm{div}_\Gamma \boldsymbol{\zeta}_h = 0$; i.e., $\boldsymbol{\zeta}_h \in \boldsymbol{\mathcal{RT}}_0^0(\Gamma_h)$.

Next, denote by $P : \boldsymbol{H}_{||}^{1/2}(\Gamma) \mapsto \boldsymbol{H}^1(\Omega_C)$ the continuous extension operator mentioned in Theorem 3.1. Then we can evaluate $\|\boldsymbol{\lambda} - \boldsymbol{\zeta}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)}$ by

$$\|\boldsymbol{\lambda} - \boldsymbol{\zeta}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} = \sup_{\boldsymbol{\Psi}\in\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)} \frac{(\boldsymbol{\lambda} - \boldsymbol{\zeta}_h, \boldsymbol{\Psi})_{0;\Gamma}}{\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)}} = \sup_{\boldsymbol{\Psi}\in\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)} \frac{\left(\gamma_N \mathbf{u} - \gamma_{\mathbf{t}}^\times \mathbf{q}_h, \gamma_{\mathbf{t}} P\boldsymbol{\Psi}\right)_{0;\Gamma}}{\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)}}$$

$$= \sup_{\boldsymbol{\Psi}\in\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)} \frac{1}{\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)}} \int_{\Omega_C} \mathbf{curl}(P\overline{\boldsymbol{\Psi}}) \cdot (\mathbf{curl}\,\mathbf{u} - \mathbf{q}_h) - \mathbf{curl}(\mathbf{curl}\,\mathbf{u} - \mathbf{q}_h) \cdot (P\overline{\boldsymbol{\Psi}})\, d\mathbf{x}$$

$$\leq \sup_{\boldsymbol{\Psi}\in\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)} \frac{\|P\boldsymbol{\Psi}\|_{\boldsymbol{H}^1(\Omega_C)}}{\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)}} \|\mathbf{curl}\,\mathbf{u} - \mathbf{q}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} \leq C \|\mathbf{curl}\,\mathbf{u} - \mathbf{q}_h\|_{\boldsymbol{H}(\mathbf{curl};\Omega_C)} .$$

Applying (8.3) for $\mathbf{curl}\,\mathbf{u}$ instead of $\mathbf{u}$ and using (8.2) finishes the proof. □

Still one issue looms large, namely, how to deal with the subspace $\boldsymbol{\mathcal{RT}}_0^0(\Gamma_h)$ of solenoidal surface Raviart–Thomas vectorfields. If Γ was simply connected, we could exploit $\boldsymbol{\mathcal{RT}}_0^0(\Gamma_h) = \mathbf{curl}_\Gamma \mathcal{S}_1(\Gamma_h)$, where $\mathcal{S}_1(\Gamma_h)$ is the space of scalar, continuous, piecewise linear finite element functions on Γ_h [49] ("stream functions," "loop currents" [51]). Yet, as we allowed more general topology of Ω_C, harmonic surface vectorfields can also contribute to the kernel of div_Γ.

The attack on the problem starts from the results of [55, sect. III] and [45, Chap. 3]: They show that there exists an orientable cutting surface $\Sigma_1, \ldots, \Sigma_L$ in Ω_E such that in $\widetilde{\Omega}_E := \Omega_E \setminus (\Sigma_1 \cup \cdots \cup \Sigma_L)$ each irrotational vectorfield has a single valued scalar potential (i.e., the cohomology group $H^1(\widetilde{\Omega}_E, \mathbb{R})$ is trivial). Their number L equals the first Betti-number of Ω_E. Moreover, $\partial\Sigma_k \subset \Gamma$, $k = 1, \ldots, L$, and none of the Σ_k is tangent to Γ. In addition, under the current assumptions on Ω_C, these surfaces can be chosen to be piecewise smooth and Lipschitz-continuous, and their boundaries can be 1-cycles (i.e., closed paths) of edges in Γ_h. Collect these paths in the set $\mathfrak{C} := \{\gamma_1, \ldots, \gamma_L\}$.

As Ω_E is connected, the first Betti-numbers of Ω_E and Ω_C coincide. Thus we can find orientable and piecewise smooth cutting surfaces $\Sigma_1', \ldots, \Sigma_L'$ that render the cohomology group $H^1(\widetilde{\Omega}_C, \mathbb{R})$ trivial: $\widetilde{\Omega}_C := \Omega_C \setminus (\Sigma_1' \cup \cdots \cup \Sigma_L')$. Their boundaries can also be 1-cycles of edges in Γ_h, again, and give the set $\mathfrak{C}' := \{\gamma_1', \ldots, \gamma_L'\}$. The union $\mathfrak{C} \cup \mathfrak{C}'$ is a set of generators of the homology group $H_1(\Gamma_h, \mathbb{Z})$, when Γ_h is regarded as cell complex (cf. [45, Chap. 2]).

The cutting surfaces can be chosen in a way that renders $\mathfrak{C}$ and $\mathfrak{C}'$ dual to each other. This means that $\mathfrak{C} \cup \mathfrak{C}' \setminus \{\gamma_k'\}$ is a set of generators of $H_1(\Gamma_h \setminus \{\gamma_k\}, \mathbb{Z})$, and $\mathfrak{C} \cup \mathfrak{C}' \setminus \{\gamma_k\}$ a set of generators of $H_1(\Gamma \setminus \{\gamma_k'\}, \mathbb{Z})$, $k = 1, \ldots, L$. The simplest example for dual 1-cycles are the "large" and "small" circle of a torus. Without loss of generality, dual edge cycles may intersect in one point only.

Let me cite a simple observation: As $\gamma_k \in \mathfrak{C}$, $k = 1, \ldots, L$, is the boundary of $\Sigma_k \subset \Omega_E$, we get from Stokes' theorem (Σ_k is orientable!)

$$\int_{\gamma_k} R\gamma_N \mathbf{u} \cdot d\vec{s} = \int_{\Sigma_k} \mathbf{curl}\,\mathbf{curl}\,\mathbf{u} \cdot \mathbf{n}\, dS = 0 ,$$

where R is defined in (5.7). We have this, first, for smooth solutions of the exterior eddy current problem and then, by continuity, for all. This means that in the discrete variational problem (8.1) we can confine ourselves to those $\boldsymbol{\lambda}_h \in \boldsymbol{\mathcal{RT}}_0^0(\Gamma_h)$ that satisfy

$$\int_\gamma R\boldsymbol{\lambda}_h \cdot d\vec{s} = 0 \quad \forall \gamma \in \mathfrak{C} \, .$$

The space of vectorfields in $\boldsymbol{\mathcal{RT}}_0^0(\Gamma_h)$ complying with this constraint will be denoted by $\widehat{\boldsymbol{\mathcal{RT}}}_0^0(\Gamma_h)$.

For each $\gamma_k \in \mathfrak{C}$, $k = 1, \ldots, L$, pick a function $\phi_k \in L^2(\Gamma)$ that is piecewise linear with respect to Γ_h and continuous except for a jump of size 1 across γ_k. Then, setting $\boldsymbol{\eta}_k := \widetilde{\mathbf{curl}}_\Gamma \phi_k$, where $\widetilde{\mathbf{curl}}_\Gamma$ is the surface-**curl** on $\Gamma \setminus \gamma_k$, we find $\boldsymbol{\eta}_k \in \boldsymbol{\mathcal{RT}}_0^0(\Gamma_h)$. Duality of the cycles implies

$$\int_{\gamma_k'} R\boldsymbol{\eta}_k \cdot d\vec{s} = [\phi_k]_{\gamma_k} = 1 \, .$$

A $\boldsymbol{\kappa}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ is in $\mathbf{curl}_\Gamma \mathcal{S}_1(\Gamma_h)$ if and only if $\int_\gamma R\boldsymbol{\kappa}_h \cdot d\vec{s} = 0$ for all closed paths γ: Just define the stream function by $\phi_h(\mathbf{x}) := \int_{\gamma(\mathbf{x}_0,\mathbf{x})} R\boldsymbol{\kappa}_h \cdot d\vec{s}$, $\mathbf{x}, \mathbf{x}_0 \in \Gamma$ ($\mathbf{x}_0$ arbitrary, but fixed), where the path $\gamma(\mathbf{x}_0, \mathbf{x})$ links $\mathbf{x}$ and $\mathbf{x}_0$. Then $\mathbf{curl}_\Gamma \phi_h = \boldsymbol{\kappa}_h$ and $\phi_h \in \mathcal{S}_1(\Gamma_h)$ by plain computations.

Now, given $\boldsymbol{\zeta}_h \in \widehat{\boldsymbol{\mathcal{RT}}}_0^0(\Gamma_h)$, the circulation of $R\tilde{\boldsymbol{\zeta}}_h$ with

$$\tilde{\boldsymbol{\zeta}}_h := \boldsymbol{\zeta}_h - \sum_{k=1}^{L} \int_{\gamma_k'} R\boldsymbol{\zeta}_h \cdot d\vec{s} \; \cdot \; \boldsymbol{\eta}_k$$

vanishes for all closed paths, because any closed path is bounding in Γ relative to $\mathfrak{C} \cup \mathfrak{C}'$. Finally, we have found

$$\widehat{\boldsymbol{\mathcal{RT}}}_0^0(\Gamma_h) = \mathbf{curl}_\Gamma \mathcal{S}_1(\Gamma_h) + \operatorname{Span}\{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L\} \, . \tag{8.4}$$

We can now phrase the discrete variational problem (8.1): Seek $\mathbf{u}_h \in \boldsymbol{\mathcal{ND}}_1(\Omega_h)$, $\phi_h \in \mathcal{S}_1(\Gamma_h)$, $\alpha_1, \ldots, \alpha_L \in \mathbb{C}$ such that

$$\begin{aligned}
q(\mathbf{u}_h, \mathbf{v}_h) + (V(\operatorname{curl}_\Gamma \gamma_{\mathbf{t}} \mathbf{u}_h), \operatorname{curl}_\Gamma \gamma_{\mathbf{t}} \mathbf{v}_h)_{0;\Gamma} + (\mathbf{B}(\mathbf{curl}_\Gamma \phi_h), \gamma_{\mathbf{t}} \mathbf{v}_h)_{0;\Gamma} & \\
+ \sum_{k=1}^{L} \alpha_k \, (\mathbf{B}\boldsymbol{\eta}_k, \gamma_{\mathbf{t}} \mathbf{v}_h)_{0;\Gamma} &= -i\omega \, (\mathbf{j}_0, \mathbf{v})_{0;\Omega_C} \, , \\
- (\mathbf{B}(\mathbf{curl}_\Gamma \psi_h), \gamma_{\mathbf{t}} \mathbf{u}_h)_{0;\Gamma} + (\mathbf{A}(\mathbf{curl}_\Gamma \phi_h), \mathbf{curl}_\Gamma \psi_h)_{0;\Gamma} & \\
+ \sum_{k=1}^{L} \alpha_k \, (\mathbf{A}\boldsymbol{\eta}_k, \mathbf{curl}_\Gamma \psi_h)_{0;\Gamma} &= 0 \, , \\
- \big(\mathbf{B}\boldsymbol{\eta}_j, \gamma_{\mathbf{t}} \mathbf{u}_h\big)_{0;\Gamma} + \big(\mathbf{A}(\mathbf{curl}_\Gamma \phi_h), \boldsymbol{\eta}_j\big)_{0;\Gamma} & \\
+ \sum_{k=1}^{L} \alpha_k \, \big(\mathbf{A}\boldsymbol{\eta}_k, \boldsymbol{\eta}_j\big)_{0;\Gamma} &= 0
\end{aligned} \tag{8.5}$$

for all $\mathbf{v}_h \in \boldsymbol{\mathcal{ND}}_1(\Omega_h)$, $\psi_h \in \mathcal{S}_1(\Gamma_h)$, and $j = 1, \ldots, L$. Of course, the solution for ϕ_h is unique only up to a constant.

Remark. Exciting total loop currents in the conductor can be easily coped with: Let the cycles $\gamma_1, \ldots, \gamma_K \in \mathfrak{C}$ belong to those loops. Write I_l, $l = 1, \ldots, K$, for the total current in loop $\#l$. Then, thanks to Ampere's law,

$$I_l = \int_{\Sigma_l'} \mathbf{j} \cdot \mathbf{n} \, dS = \int_{\gamma_l'} -\tfrac{1}{\mu i\omega} \, \mathbf{curl}\, \mathbf{E} \cdot d\vec{s} = -\tfrac{1}{i\omega} \int_{\gamma_l'} R(\tfrac{1}{\mu} \gamma_N^- \mathbf{E}) \cdot d\vec{s} = -\tfrac{1}{i\omega} \int_{\gamma_l'} R\boldsymbol{\lambda} \cdot d\vec{s} \, .$$

Hence, the jump of the stream function across γ_l must be set to $-i\omega I_l$. This amounts to fixing $\alpha_l = -i\omega I_l$, $l = 1, \ldots, K$, in (8.5), while $-i\omega\,(\mathbf{j}_0, \mathbf{v})_{0;\Omega_C}$ is dropped. Only $L - K$ jumps $\alpha_{K+1}, \ldots, \alpha_L$ remain as unknowns.

9. Implementation. The choice of local basis functions for the spaces $\boldsymbol{\mathcal{NP}}_1(\Omega_h)$ and $\mathcal{S}_1(\Gamma_h)$ is canonical. They are associated with the edges of Ω_h and the vertices of Γ_h, respectively. The assembly of the linear system of equations related to (8.5) poses no unusual difficulties. First, we note that $\operatorname{curl}_\Gamma \gamma_{\mathbf{t}}\mathbf{v}_h$, $\operatorname{curl}_\Gamma \gamma_{\mathbf{t}}\mathbf{u}_h$, $\mathbf{curl}_\Gamma\psi_h$, $\mathbf{curl}_\Gamma\phi_h$, and $\boldsymbol{\eta}_k$ are piecewise constant on Γ_h and $\gamma_{\mathbf{t}}\mathbf{u}_h$, $\gamma_{\mathbf{t}}\mathbf{v}_h$ piecewise (bi-)linear. On top of that, all boundary integral operators are structurally equal to those for second order elliptic problems. For instance (see [68]),

$$\begin{aligned}(\mathbf{B}\boldsymbol{\zeta}_h, \mathbf{v}_h)_{0;\Gamma} = \quad & \int_\Gamma \int_\Gamma \boldsymbol{\zeta}_h(\mathbf{y}) \cdot \overline{\mathbf{v}}_h(\mathbf{x}) \frac{\partial E(\mathbf{x},\mathbf{y})}{\partial \mathbf{n}(\mathbf{x})}\, dS(\mathbf{y}) dS(\mathbf{x}) \\ & - \int_\Gamma \int_\Gamma \operatorname{grad}_{\mathbf{x}} E(\mathbf{x},\mathbf{y}) (\boldsymbol{\zeta}_h(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x})) \cdot \overline{\mathbf{v}}_h(\mathbf{x})\, dS(\mathbf{y}) dS(\mathbf{x}) \\ & - \frac{1}{2} \int_\Gamma \boldsymbol{\zeta}_h(\mathbf{x}) \cdot \overline{\mathbf{v}}_h(\mathbf{x})\, dS(\mathbf{x})\ .\end{aligned}$$

Even more striking, [39, Chap. XI, sect. 2, Thm. 7] reveals that

$$(\mathbf{A}(\mathbf{curl}_\Gamma\phi_h), \mathbf{curl}_\Gamma\psi_h)_{0;\Gamma} = \langle D\phi_h, \psi_h\rangle_{1/2,\Gamma} \quad \forall \phi_h, \psi_h \in \mathcal{S}_1(\Gamma_h)\ ,$$

where $D : H^{1/2}(\Gamma) \mapsto H^{-1/2}(\Gamma)$ is the hypersingular operator for the Laplacian.

This permits us to apply all the sophisticated techniques developed for Galerkin boundary element methods for second order elliptic problems. In particular, the same quadrature rules [73] and panel-clustering techniques [47] may be used in an implementation.

When the linear system has been built, it must be solved iteratively, because the sheer dimension of $\boldsymbol{\mathcal{NP}}_1(\Omega_h)$ will usually overwhelm any direct solver. Let us first study the case of trivial topology of Ω_C, i.e., $L = 0$. Then, in compact notation, the variational problem related to (8.5) is characterized by the bilinear form

$$\widetilde{\mathbf{d}}((\mathbf{u}_h, \phi_h), (\mathbf{v}_h, \psi_h)) := \mathbf{d}((\mathbf{u}_h, \mathbf{curl}_\Gamma\phi_h), (\mathbf{v}_h, \mathbf{curl}_\Gamma\psi_h))\ ,$$

with $\mathbf{u}_h, \mathbf{v}_h \in \boldsymbol{\mathcal{NP}}_1(\Omega_h)$, $\phi_h, \psi_h \in \mathcal{S}_1(\Gamma_h)$. Setting $\boldsymbol{\mathcal{U}}_h := \boldsymbol{\mathcal{NP}}_1(\Omega_h) \times \mathcal{S}_1(\Gamma_h)$ with seminorm

$$\|(\mathbf{v}_h, \psi_h)\|^2_{\boldsymbol{\mathcal{U}}} := \left(\frac{1}{\mu}\,\mathbf{curl}\,\mathbf{v}_h, \mathbf{curl}\,\mathbf{v}_h\right)_{0;\Omega_C} + \omega\,(\sigma\mathbf{v}_h, \mathbf{v}_h)_{0;\Omega_C} + \langle D\psi_h, \psi_h\rangle_{1/2,\Gamma}\,,$$

the proof of Theorem 7.1 teaches that

$$\frac{1}{\sqrt{2}}\,\|(\mathbf{u}_h, \phi_h)\|_{\boldsymbol{\mathcal{U}}} \le \|\mathcal{P}_h(\mathbf{u}_h, \phi_h)\|_{\boldsymbol{\mathcal{U}}'} \le \overline{c}\,\|(\mathbf{u}_h, \phi_h)\|_{\boldsymbol{\mathcal{U}}} \quad \forall (\mathbf{u}_h, \phi_h) \in \boldsymbol{\mathcal{U}}_h\ . \tag{9.1}$$

Here, $\mathcal{P}_h : \boldsymbol{\mathcal{U}}_h \mapsto \boldsymbol{\mathcal{U}}_h'$ is the operator associated with $\widetilde{\mathbf{d}}$, and $\overline{c}$ is a positive constant that is utterly independent of the choice of conforming finite element/boundary element spaces. More precisely, as can be seen from the proof of Theorem 7.1, it depends only on the norm bound for the boundary integral operators $\mathbf{N}, \mathbf{B}, \mathbf{C}$.

As exposed in [6, sect. 7] and [28, sect. 4], (9.1) means that solving discrete variational problems connected with the inner product on $\boldsymbol{\mathcal{U}}$ will supply a good preconditioner for $\mathcal{P}_h$. Thus we define the symmetric positive definite operator $\widehat{\mathcal{P}}_h : \boldsymbol{\mathcal{U}}_h' \mapsto \boldsymbol{\mathcal{U}}_h$ by $\widehat{\mathcal{P}}(f_\Re + if_\Im, \rho_\Re + i\rho_\Im) = (\mathbf{v}_\Re + i\mathbf{v}_\Im, \psi_\Re + i\psi_\Im) \in \boldsymbol{\mathcal{U}}_h$ with $(* = \Re, \Im)$

(9.2)
$$\begin{aligned} \left(\tfrac{1}{\mu}\,\mathbf{curl}\,\mathbf{v}_*, \mathbf{curl}\,\mathbf{q}_h\right)_{0;\Omega_C} + \omega\,(\sigma\mathbf{v}_*, \mathbf{q}_h)_{0;\Omega_C} &= f_*(\mathbf{q}_h) \quad \forall \mathbf{q}_h \in \boldsymbol{\mathcal{NP}}_1(\Omega_h)\;, \\ \langle D\psi_*, \eta_h\rangle_{1/2,\Gamma} &= \rho_*(\eta_h) \quad \forall \eta_h \in \mathcal{S}_1(\Gamma_h)\;. \end{aligned}$$

Here, subscripts $\Re$ and $\Im$ tag the (real valued) real and imaginary parts of functionals and functions. Consequently, the ansatz and test functions in the above variational formulation are real valued, too. Besides, we have to demand that ρ_* vanishes for constants. As the right-hand sides of the discrete systems are consistent and the preconditioner is applied to residuals, this trivially holds. Altogether, we infer from (9.1) that the spectrum of $\widehat{\mathcal{P}_h}\mathcal{P}_h$ is contained in $[-\overline{c}, -\frac{1}{2}\sqrt{2}] \cup \{0\} \cup [\frac{1}{2}\sqrt{2}, \overline{c}]$. This carries over to the spectrum of the product $\widehat{\mathsf{P}}\mathsf{P}$ of the matrices representing $\widehat{\mathcal{P}}_h$ and $\mathcal{P}_h$ with respect to some basis of $\boldsymbol{\mathcal{U}}_h$.

However, solving either equation in (9.2) is still prohibitively expensive. A second stage is required in the following way: To determine an approximation for $\mathbf{v}_*$, I suggest to use a symmetric V-cycle of the multigrid method for edge elements presented in [50, sect. 6], if a hierarchy of nested regularly refined meshes, whose finest is Ω_h, is at one's disposal. This amounts to solving the modified variational problem

$$b(\mathbf{v}_*, \mathbf{q}_h) = f_*(\mathbf{q}_h) \quad \forall \mathbf{q}_h \in \boldsymbol{\mathcal{NP}}_1(\Omega_h)\;,$$

with a bilinear form $b : \boldsymbol{\mathcal{NP}}_1(\Omega_h) \times \boldsymbol{\mathcal{NP}}_1(\Omega_h) \mapsto \mathbb{R}$ that fulfills

$$\underline{c}_b\,\|(\mathbf{v}_h, 0)\|^2_{\boldsymbol{\mathcal{U}}} \le b(\mathbf{v}_h, \mathbf{v}_h) < \|(\mathbf{v}_h, 0)\|^2_{\boldsymbol{\mathcal{U}}} \quad \forall \mathbf{v}_h \in \boldsymbol{\mathcal{NP}}_1(\Omega_h)\;. \tag{9.3}$$

Here, $\underline{c}_b$ is uniformly bounded away from zero (cf. [50, sect. 5]), way above $\frac{1}{2}$, as numerical evidence suggests. In case a hierarchy of meshes is not available, one may resort to algebraic multigrid methods [8, 69]. Yet, theoretical estimates like (9.3) are elusive in this case.

The second equation of (9.2) can be tackled following the recipe of [74]. The idea is to use the scalar single layer potential operator V as a preconditioner for D (see also [62, 61]): We know

$$\underline{k}\,\|\psi_h\|^2_{L^2(\Gamma)} \le \langle VD\psi_h, \psi_h\rangle_{1/2,\Gamma} \le \overline{k}\,\|\psi_h\|^2_{L^2(\Gamma)} \quad \forall \psi_h \in \mathcal{S}_1(\Gamma_h)/\mathbb{R}$$

with constants $\underline{k}, \overline{k} > 0$ that can be chosen independently of $\mathcal{S}_1(\Gamma_h)$. $\mathcal{S}_1(\Gamma_h)/\mathbb{R}$ designates the subspace of $\mathcal{S}_1(\Gamma_h)$ containing only functions with zero mean. Thus, an efficient way to determine an approximation for ψ_* from (9.2) is to use a small number of steps of a damped linear or (nonlinear) gradient-type iteration preconditioned by V. As alternatives, domain decomposition or multilevel methods might be used [76, 48].

This practical preconditioner can then be used to accelerate the conjugate residual method [46, sect. 9.5] applied to the symmetric linear system

$$\mathsf{P}\vec{\mathbf{x}} := \begin{pmatrix} \mathsf{M}_\Re & -\mathsf{M}_\Im & \mathsf{B}^T & 0 \\ -\mathsf{M}_\Im & -\mathsf{M}_\Re & 0 & -\mathsf{B}^T \\ \mathsf{B} & 0 & -\mathsf{D} & 0 \\ 0 & -\mathsf{B} & 0 & \mathsf{D} \end{pmatrix} \begin{pmatrix} \vec{\mathbf{u}}_\Re \\ \vec{\mathbf{u}}_\Im \\ \vec{\phi}_\Re \\ \vec{\phi}_\Im \end{pmatrix} = \begin{pmatrix} \vec{f}_\Re \\ \vec{f}_\Im \\ \vec{\rho}_\Re \\ \vec{\rho}_\Im \end{pmatrix} [=: \vec{\mathbf{y}}]. \tag{9.4}$$

Here, $\vec{\mathbf{u}}_\Re, \vec{\mathbf{u}}_\Im, \vec{\phi}_\Re, \vec{\phi}_\Im$ are the vectors of the real/imaginary components of degrees of freedom in $\boldsymbol{\mathcal{ND}}_1(\Omega_h)$ and $\mathcal{S}_1(\Gamma_h)$, respectively. The matrices $\mathsf{M}_\Re$, $\mathsf{M}_\Im$, B, and D arise from a discretization of $\widetilde{\mathbf{d}}$ in a straightforward manner. Thus, $\mathsf{D} = \mathsf{D}^T$ and $\mathsf{M}_* = \mathsf{M}_*^T$. The right-hand sides can be deduced from (8.5). Note that, compared to (8.5), signs in the second equation have been flipped to achieve a symmetric system.

If $L > 0$, we face the system

$$\begin{pmatrix} \mathsf{P} & \mathsf{F} \\ \mathsf{F}^T & \mathsf{H} \end{pmatrix} \begin{pmatrix} \vec{\mathbf{x}} \\ \vec{\alpha} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{y}} \\ \vec{\rho} \end{pmatrix} ,$$

where $\mathsf{H} \in \mathbb{R}^{2L,2L}$ is symmetric, positive definite. Switching to the Schur-complement system

$$\mathsf{S}\vec{\mathbf{x}} := (\mathsf{P} - \mathsf{F}\mathsf{H}^{-1}\mathsf{F}^T)\vec{\mathbf{x}} = \vec{\mathbf{y}} - \mathsf{F}\mathsf{H}^{-1}\vec{\rho},$$

we realize that it can be considered a rank-$2L$ perturbation of (9.4). The spectrum of $\widehat{\mathsf{P}}\mathsf{S}$ will still lie in $[-\bar{c}, -\frac{1}{2}\sqrt{2}] \cup \{0\} \cup [\frac{1}{2}\sqrt{2}, \bar{c}]$, except for $2L$ eigenvalues. However, as is typical of gradient-type schemes [7], the convergence of the conjugate residual method after at most $2L$ steps will no longer be affected by these eigenvalues. As L is moderate (typically 1 or 2), that many steps are not too expensive to conduct.

REFERENCES

[1] R. Albanese and G. Rubinacci, *Fomulation of the eddy–current problem*, IEE Proc. A, 137 (1990), pp. 16–22.

[2] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of $H(\mathrm{rot};\Omega)$ and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.

[3] H. Ammari, A. Buffa, and J.-C. Nédélec, *A justification of eddy currents model for the Maxwell equations*, SIAM J. Appl. Math., 60 (2000), pp. 1805–1823.

[4] H. Ammari and J. Nédélec, *Couplage éléments finis-équations intégrales pour la résolution des équations de Maxwell en milieu hétérogene*, in Equations aux derivees partielles et applications, Gauthier-Villars, Paris, 1998, pp. 19–33.

[5] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three–dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[6] D. Arnold, R. Falk, and R. Winther, *Preconditioning in $H(\mathrm{div})$ and applications*, Math. Comp., 66 (1997), pp. 957–984.

[7] O. Axelson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[8] R. Beck, *Algebraic Multigrid by Component Splitting for Edge Elements on Simplicial Triangulations*, Technical report SC 99-40, ZIB, Berlin, Germany, 1999.

[9] A. Bendali, J. Dominguez, and S. Gallic, *A variational approach for the vector potential formulation of the Stokes and Navier–Stokes problems in three dimensions*, J. Math. Anal. Appl., 107 (1985), pp. 537–560.

[10] O. Biro and K. Richter, *CAD in electromagnetism*, in Advances in Electronics and Electron Physics, Vol. 82, P. Hawkes, ed., Academic Press, San Diego, 1991, pp. 1–96.

[11] A.-S. Bonnet-Ben Dhia, C. Hazard, and S. Lohrengel, *A singular field method for the solution of Maxwell's equations in polyhedral domains*, SIAM J. Appl. Math., 59 (1999), pp. 2028–2044.

[12] A. Bossavit, *Two dual formulations of the 3D eddy–currents problem*, COMPEL, 4 (1985), pp. 103–116.

[13] A. Bossavit, *Mixed finite elements and the complex of Whitney forms*, in The Mathematics of Finite Elements and Applications VI, J. Whiteman, ed., Academic Press, London, 1988, pp. 137–144.

[14] A. Bossavit, *A rationale for edge elements in 3D field computations*, IEEE Trans. Mag., 24 (1988), pp. 74–79.

[15] A. Bossavit, *On various representations of fields by potentials and their use in boundary integral methods*, COMPEL, 9 (1990), pp. 31–36.

[16] A. Bossavit, *The computation of eddy–currents in dimension* 3 *by using mixed finite elements and boundary elements in association*, Math. Comput. Modelling, 15 (1991), pp. 33–42.
[17] A. Bossavit, *The scaler Poincare Steklov operator and the vector one: Algebraic structures which underlie their duality*, in Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, Y. Kuznetsov, G. Meurant, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, 1991, pp. 19–26.
[18] A. Bossavit, *Computational Electromagnetism. Variational Formulation, Complementarity, Edge Elements*, in Electromagnetism 2, Academic Press, San Diego, 1998.
[19] A. Bossavit, *On the Lorenz gauge*, COMPEL, 18 (1999), pp. 323–336.
[20] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math., Springer-Verlag, New York, 1991.
[21] A. Buffa, *Hodge decompositions on the boundary of a polyhedron: The multiconnected case*, Math. Mod. Meth. Appl. Sci., 11 (2001), pp. 1491–1504.
[22] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* I*: An integration by parts formula in Lipschitz polyhedra.*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.
[23] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* II*: Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.
[24] A. Buffa, M. Costabel, and D. Sheen, *On Traces for* $\mathbf{H}(\mathbf{curl}, \Omega)$ *in Lipschitz Domains*, J. Math. Anal. Appl., to appear.
[25] C. Carstensen and P. Wriggers, *On the symmetric boundary element method and the symmetric coupling of boundary elements and finite elements*, IMA J. Numer. Anal., 17 (1997), pp. 201–238.
[26] M. Cessenat, *Mathematical Methods in Electromagnetism*, Advances in Mathematics for Applied Sciences 41, World Scientific, Singapore, 1996.
[27] G. Chen and J. Zhou, *Boundary Element Methods*, Academic Press, New York, 1992.
[28] S.-H. Chou, D. Y. Kwak, and P. S. Vassilevski, *Mixed covolume methods for elliptic problems on triangular grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1850–1861.
[29] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.
[30] P. Ciarlet, Jr., and J. Zou, *Fully discrete finite element approaches for time–dependent Maxwell equations*, Numer. Math., 82 (1999), pp. 193–219.
[31] M. Clemens and T. Weiland, *Transient eddy current calculation with the FI–method*, IEEE Trans. Magnetics, 35 (1999), pp. 1163–1166.
[32] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Pure Appl. Math., John Wiley & Sons, New York, 1983.
[33] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Heidelberg, 1998.
[34] M. Costabel, *Symmetric methods for the coupling of finite elements and boundary elements*, in Boundary Elements IX, C. Brebbia, W. Wendland, and G. Kuhn, eds., Springer-Verlag, Berlin, 1987, pp. 411–420.
[35] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.
[36] M. Costabel and M. Dauge, *Singularities of Maxwell's equations on polyhedral domains*, in Analysis, Numerics and Applications of Differential and Integral Equations, Pitman Res. Notes Math. Ser. 379, M. Bach, ed., Addison-Wesley, Harlow, UK, 1998, pp. 69–76.
[37] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M^2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.
[38] M. Costabel and W. Wendland, *Strong ellipticity of boundary integral operators*, J. Reine Angew. Math., 372 (1986), pp. 39–63.
[39] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 4, Springer-Verlag, Berlin, Heidelberg, New York, 1990.
[40] H. Dirks, *Quasi–stationary fields for microelectronic applications*, Electrical Engineering, 79 (1996), pp. 145–155.
[41] V. Girault, *The divergence, curl and stokes operators in exterior domains of* R^3, in Recent Developments in Theoretical Fluid Mechanics, Pitman Res. Notes Math. Ser. 291, G. Galdi, ed., Longman Scientific & Technical, Harlow, UK, 1993, pp. 34–77.
[42] V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[43] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[44] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Springer-Verlag, New York, 1992.
[45] P. Gross, *Efficient Finite Element-Based Algorithms for Topological Aspects of* 3*-Dimensional Magnetoquasistatic Problems*, Ph.D. thesis, College of Engineering, Boston University, Boston, 1998.
[46] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Appl. Math. Sci. 95, Springer-Verlag, New York, 1993.
[47] W. Hackbusch and Z. Nowak, *On the fast matrix multiplication in the boundary element method by panel clustering*, Numer. Math., 54 (1989), pp. 463–491.
[48] N. Heuer and E. P. Stephan, *Iterative substructuring for hypersingular integral equations in* $\mathbb{R}^3$, SIAM J. Sci. Comput., 20 (1998), pp. 739–749.
[49] R. Hiptmair, *Canonical construction of finite elements*, Math. Comp., 68 (1999), pp. 1325–1346.
[50] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.
[51] K. Ishibashi, *Eddy current analysis by integral equation method utilizing loop electric and surface magnetic currents as unknowns*, IEEE Trans. Magnetics, 34 (1998), pp. 2585–2588.
[52] C. Johnson and J. Nédélec, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980).
[53] L. Kettunen, K. Forsman, and A. Bossavit, *Gauging in Whitney spaces*, IEEE Trans. Magnetics, 35 (1999), pp. 1466–1469.
[54] A. Kirsch, *Surface gradients and continuity properties for some integral operators in classical scattering theory*, Math. Methods Appl. Sci., 11 (1989), pp. 789–804.
[55] P. Kotiuga, *On making cuts for magnetic scalar potentials in multiply connected regions*, J. Appl. Phys., 61 (1987), pp. 3916–3918.
[56] P. Kotiuga, *Topological considerations in coupling magnetic scalar potentials to stream functions describing surface currents*, IEEE Trans. Magnetics, 25 (1989), pp. 2925–2927.
[57] M. Kuhn, U. Langer, and J. Schöberl, *Scientific computing tools for* 3D *magnetic field problems*, in The Mathematics of Finite Elements and Applications X, J. Whiteman, ed., Elsevier, Amsterdam, 2000, pp. 78–99.
[58] I. Mayergoyz, 3D *eddy current problems and the boundary integral equation method*, in Computational Electromagnetics, Z. Cendes, ed., Elsevier, Amsterdam, 1986, pp. 163–171.
[59] R. McCamy and E. Stephan, *Solution procedures for three-dimensional eddy-current problems*, J. Math. Anal. Appl., 101 (1984), pp. 348–379.
[60] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[61] W. McLean and O. Steinbach, *Boundary element preconditioners for a hypersingular integral equations on an interval*, Adv. Comput. Math., 11 (1999), pp. 271–286.
[62] W. McLean and T. Tran, *A preconditioning strategy for boundary element Galerkin methods*, Numer. Methods Partial Differential Equations, 13 (1997), pp. 283–301.
[63] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.
[64] J. Nédélec, *Mixed finite elements in* R^3, Numer. Math., 35 (1980), pp. 315–341.
[65] J.-C. Nédélec, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Appl. Math. Sci. 44, Springer-Verlag, Berlin, 2001.
[66] L. Paquet, *Problemes mixtes pour le systeme de maxwell*, Ann. Fac. Sci. Toulouse VI. Ser. Math (1982), pp. 103–141.
[67] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Elements of Finite Element Methods, Lecture Notes in Math. 606, I. Galligari and E. Mageres, eds., Springer-Verlag, New York, 1977, pp. 292–315.
[68] M. Reissel, *On a transmission boundary-value problem for the time-harmonic Maxwell equations without displacement currents*, SIAM J. Math. Anal., 24 (1993), pp. 1440–1457.
[69] S. Reitzinger and J. Schöberl, *An algebraic multigrid for finite element discretizations with edge elements*, Numer. Linear Algebra Appl., to appear.
[70] Z. Ren, F. Bouillault, A. Razek, A. Bossavit, and J. Vérité, *A new hybrid model using electric field formulation for* 3D *eddy–current problems*, IEEE Trans. Magnetics, 36 (1990), p. 473.
[71] Z. Ren, F. Bouillault, A. Razek, and J. Verité, *Comparison of different boundary integral formulations when coupled with finite elements in three dimensions*, IEE Proc. A, 135 (1988), pp. 501–505.

[72] Z. Ren and A. Razek, *New techniques for solving three–dimensional multiply connected eddy–current problems*, IEE Proc. A, 137 (1990), pp. 135–140.

[73] S. Sauter and A. Krapp, *On the effect of numerical integration in the Galerkin boundary element method*, Numer. Math., 74 (1996), pp. 337–359.

[74] O. Steinbach and W. Wendland, *The construction of some efficient preconditioners in the boundary element method*, Adv. Comput. Math., 9 (1998), pp. 191–216.

[75] O. Sterz and C. Schwab, *A scalar BEM for time harmonic eddy current problems with impedance boundary conditions*, in Scientific Computing in Electrical Engineering, Lect. Notes Comput. Sci. Eng. 18, Springer-Verlag, Berlin, 2001, pp. 129–136.

[76] T. Tran, E. Stephan, and P. Mund, *Hierarchical basis preconditioners for first kind integral equations*, Appl. Anal., 65 (1997), pp. 353–372.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 66–86

NATURAL BOUNDARY ELEMENT METHODS FOR THE ELECTRIC FIELD INTEGRAL EQUATION ON POLYHEDRA*

R. HIPTMAIR† AND C. SCHWAB‡

Abstract. We consider the electric field integral equation on the surface of polyhedral domains and its Galerkin discretization by means of divergence-conforming boundary elements. With respect to a Hodge decomposition, the continuous variational problem is shown to be coercive. However, this does not immediately carry over to the discrete setting, as discrete Hodge decompositions fail to possess essential regularity properties. Introducing an intermediate semidiscrete Hodge decomposition, we can bridge the gap and come up with asymptotically optimal a priori error estimates. Until now, those had been elusive, in particular for nonsmooth boundaries.

Key words. electric field integral equation, Rumsey's principle, Raviart–Thomas elements, Hodge decomposition, discrete coercivity

AMS subject classifications. 65N12, 65N38, 78M15

PII. S0036142901387580

1. Introduction.

One of the main tasks in computational electromagnetism is the computation of the scattering of electromagnetic waves at a perfectly conducting body $\Omega \subset \mathbb{R}^3$. It boils down to solving the time-harmonic Maxwell's equations in the exterior $\Omega' = \mathbb{R}^3 \backslash \overline{\Omega}$ of Ω for a fixed frequency, subject to a vanishing tangential trace of the total electric field on the surface of the scatterer and the Silver–Müller radiation conditions at ∞. It is known that the exterior scattering problem for Maxwell's equations has a unique solution (see, e.g., [26, Chap. 6] and [23]). In most technical applications, the boundary Γ of Ω will be only piecewise smooth.

Starting from the Stratton–Chu representation formulas in Ω' (see, e.g., [19, sect. 3]), an indirect method yields the well-known electrical field boundary integral equation (EFIE) for the unknown jump $\mathbf{j}$ of the magnetic field [19, sect. 4]. Cast in variational form, this integral equation is sometimes referred to as Rumsey's principle [9] and reads as follows: Find a complex amplitude $\mathbf{j} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ such that

$$\langle V_\varsigma \mathrm{div}_\Gamma \mathbf{j}, \mathrm{div}_\Gamma \mathbf{v} \rangle_{\frac{1}{2},\Gamma} - \varsigma^2 \langle \mathbf{A}_\varsigma \mathbf{j}, \mathbf{v} \rangle_{||,\Gamma} = f(\mathbf{v}) \quad \forall \mathbf{v} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma). \tag{1.1}$$

Here, $\varsigma \in \mathbb{R}_+$ is the nondimensional wavenumber, the continuous linear functional $f : \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{curl}_\Gamma, \Gamma) \mapsto \mathbb{C}$ represents the excitation due to an incident wave, and $V_\varsigma, \mathbf{A}_\varsigma$ stand for scalar and vectorial single layer potential integral operators, respectively. In (1.1), the (sesquilinear) duality pairings $\langle \cdot, \cdot \rangle_{\frac{1}{2},\Gamma}$ and $\langle \cdot, \cdot \rangle_{||,\Gamma}$ coincide with the usual $H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ duality when Γ is smooth. On polyhedra, however, there are several nonequivalent definitions of these dualities. Details will be explained below. For the well-posedness of the integral equation formulation (1.1), we adopt the following assumption throughout this paper.

ASSUMPTION 1.1. *The wavenumber ς is bounded away from the spectrum of the interior Maxwell's problem.*

*Received by the editors April 9, 2001; accepted for publication (in revised form) November 19, 2001; published electronically April 12, 2002.

http://www.siam.org/journals/sinum/40-1/38758.html

†Nachwuchsgruppe SFB 382, Universität Tübingen, 72076 Tübingen, Germany (hiptmair@na.uni-tuebingen.de).

‡Seminar für Angewandte Mathematik, ETH Zürich, CH-8092 Zurich, Switzerland (christoph.schwab@sam.math.ethz.ch).

This implies the injectivity of the boundary integral operator in (1.1).

Recalling the derivation of (1.1), the unknown $\mathbf{j}$ emerges as the jump of tangential traces $\mathbf{H}\times\mathbf{n}$ of magnetic field solutions for the full Maxwell's equations in the interior and exterior of Ω. When Maxwell's equations are concisely stated in the language of differential forms [6, 13], the magnetic field is modeled by a twisted 1-form. The same will hold for its trace on Γ. This suggests that two-dimensional discrete twisted 1-forms built upon a triangulation of Γ should be used to approximate $\mathbf{j}$. Those are provided by the boundary element counterparts of the two-dimensional Raviart–Thomas $\boldsymbol{H}(\mathrm{div};\Omega)$-conforming finite elements. We could also reason in an entirely discrete setting: it is no longer a moot point that a suitable discretization of magnetic fields is provided by $\boldsymbol{H}(\mathbf{curl};\Omega)$-conforming edge elements [44], which are discrete 1-forms in three dimensions. Taking a look at their tangential traces, again, we discover Raviart–Thomas elements mapped onto the surface [35]. Thus, we argue that the latter offer a "natural" boundary element discretization of (1.1) as follows: Find $\mathbf{j}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ such that

$$\langle V_\varsigma \mathrm{div}_\Gamma \mathbf{j}_h, \mathrm{div}_\Gamma \mathbf{v}_h\rangle_{\frac{1}{2},\Gamma} - \varsigma^2 \langle \mathbf{A}_\varsigma \mathbf{j}_h, \mathbf{v}_h\rangle_{||,\Gamma} = f(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h), \tag{1.2}$$

where $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ denotes the lowest order Raviart–Thomas boundary element space on a triangulation Γ_h on Γ.

The Galerkin discretization (1.2) is commonplace in engineering codes. The first convergence analysis of this scheme was given by Bendali in [7, 8] and was based on a saddle point formulation and elliptic regularization, which is inherently confined to smooth surfaces. Using parametric variants of the Raviart–Thomas boundary elements, he established asymptotic a priori convergence estimates. Attempts to adapt his approach to nonsmooth surfaces have not been successful. Recently, Buffa, Costabel, and Schwab [19] succeeded in showing convergence of a mixed discretization of (1.1) which, however, is different from the "natural" scheme (1.2) used in engineering.

Obstructions of convergence analysis on nonsmooth surfaces are threefold. First, the correct function spaces and relevant surface differential operators have to be properly characterized. For smooth domains, using smooth charts and trace theorems for the entire scale of Sobolev spaces, this is not hard to do [2, 23]. It becomes a challenge in a nonsmooth setting, as is vividly conveyed in the introduction of [20]. The situation on polyhedral boundaries Γ was clarified only recently by Buffa and Ciarlet [17, 18, 15], and general Lipschitz boundaries were presented in [20]. We emphasize that it is these results only that made possible the progress reported in the current paper.

Second, with (1.1) we recognize the typical difficulty faced when dealing with variational problems arising from Maxwell's equations: owing to the large kernel of the surface divergence operator div_Γ, it becomes impossible to assign one term the role of a principal part and, thus, the sesquilinear form of (1.1) fails to be coercive. A remedy was first found in the case of Maxwell's differential equations [42, 39] and it is marked by the use of *Hodge decompositions*. Also, for boundary integral equations this idea is fruitful and in fact was exploited many times to recover coercive problems [31, 19, 3].

Unfortunately, Hodge decompositions and the divergence-conforming boundary elements do not match easily. This is the third obstacle and it is also faced in the analysis of $\boldsymbol{H}(\mathbf{curl};\Omega)$-conforming finite element schemes. In that context, a solution has been devised relying on judiciously combining discrete and continuous Hodge decompositions. This idea was successful in the analysis of multigrid methods for

edge elements [36, 4, 38] and in the numerical analysis of schemes for Maxwell's eigenproblems that are free of spurious modes [10, 12, 11, 43, 22].

It is this idea that permits us to launch a successful attack on Galerkin discretizations of the EFIE (1.1) on polyhedral domains. Yet, numerous adjustments of this technique are necessary to cope with the low regularity of the function spaces of interest on Γ. Whereas for problems on domains $\subset \mathbb{R}^3$ all fields are at least square integrable, here we find that surface vectorfields in $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ do not necessarily have this property. In this paper we aim to elucidate how to handle this difficulty. Christiansen [24] pursues a policy partly similar to ours, but with a different objective, suboptimal results, and confinement to smooth domains. Other related publications are [21] and [16].

The paper is organized as follows. In the next section we summarize important results about spaces of tangential vectorfields on polyhedra. The third and fourth sections establish the coercivity of the continuous variational problem with respect to a Hodge decomposition of $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$. In the fifth section we introduce divergence-conforming boundary elements and review their main properties. In the sixth section we define and scrutinize mappings that create a link between discrete and continuous Hodge decompositions. The seventh section is dedicated to proving a discrete inf-sup condition. The final section covers asymptotic a priori estimates of the discretization error.

It was our objective to keep the treatment as focused and self-contained as possible. To that end we forgo any generalizations and investigate only the lowest order Raviart–Thomas boundary elements and Lipschitz polyhedra with plane faces. By and large, generalizations are straightforward. Numerical experiments are skipped, since the Galerkin discretization of the EFIE by Raviart–Thomas boundary elements is widely and successfully used in electrical engineering.

2. Spaces. The domain $\Omega \subset \mathbb{R}^3$ is assumed to be a Lipschitz polyhedron (cf. the introduction of [29]). In particular, we assume that the Lipschitz boundary Γ can be written as a union of a finite number of plane faces Γ_j, $j = 1, \ldots, N_\Gamma$, i.e., $\bar{\Gamma} = \bigcup_i \bar{\Gamma}_j$. For each face Γ_j we find a constant unit normal vector $\mathbf{n}_i$ pointing into the exterior of Ω. These vectors can be blended into an exterior unit normal vectorfield $\mathbf{n} \in L^\infty(\Gamma)$, defined almost everywhere on Γ. In addition, we can fix two orthogonal unit vectors $\mathbf{e}_j^1, \mathbf{e}_j^2$ that span the tangential plane for Γ_i. It goes without saying that each Γ_j can be identified with a bounded subset of $\mathbb{R}^2$.

Next, we introduce two different tangential surface trace operators [20, sect. 2]. The tangential components trace $\pi_{\mathbf{t}}$ is defined for $\mathbf{u} \in \boldsymbol{C}^\infty(\bar{\Omega})$ by $\pi_{\mathbf{t}}\mathbf{u}(\mathbf{x}) := \mathbf{n}(\mathbf{x}) \times (\mathbf{u}(\mathbf{x}) \times \mathbf{n}(\mathbf{x}))$ for almost all $\mathbf{x} \in \Gamma$. Accordingly, the tangential surface trace $\gamma_{\mathbf{t}}$ can be computed through $\gamma_{\mathbf{t}}\mathbf{u}(\mathbf{x}) := \mathbf{u}(\mathbf{x}) \times \mathbf{n}(\mathbf{x})$. The same traces from Ω' are ${\pi_{\mathbf{t}}}'$ and $\gamma_{\mathbf{t}}'$. To begin with, the trace operators supply functions in

$$\boldsymbol{L}_{\mathbf{t}}^2(\Gamma) := \{\mathbf{u} \in (L^2(\Gamma))^3, \mathbf{u} \cdot \mathbf{n} = 0\}.$$

The usual Sobolev spaces of scalar functions and related functionals, $H^s(\Gamma)$ and $H^{-s}(\Gamma)$, can be defined invariantly for $0 \le s \le 1$ [34, sect. 1.3.3]. For indices $s > 1$ we resort to the piecewise definition

$$H^s(\Gamma) := \{u \in H^1(\Gamma),\, u_{|\Gamma_j} \in H^s(\Gamma_j),\, j = 1, \ldots, N_\Gamma\}.$$

We equip this space with the graph norm

$$\|u\|^2_{H^s(\Gamma)} := \|u\|^2_{H^1(\Gamma)} + \sum_{j=1}^{N_\Gamma} \|u\|^2_{H^s(\Gamma_j)}\,.$$

Using the local coordinate systems introduced above, spaces of tangential vectorfields that feature certain Sobolev regularity in a piecewise sense are readily available:

$$\boldsymbol{H}^s_{\mathbf{t}}(\Gamma) := \{\mathbf{u} \in \boldsymbol{L}^2_{\mathbf{t}}(\Gamma),\, \mathbf{u}_{|\Gamma_j} \cdot \mathbf{e}^i_j \in H^s(\Gamma_j),\, j = 1,\ldots,N_\Gamma,\, i = 1,2\}.$$

By localization to Γ_j we can define the tangential surface gradient $\mathbf{grad}_\Gamma$ [20, Def. 3.1]. Its continuity as a mapping $H^{s+1}(\Gamma) \mapsto \boldsymbol{H}^s_{\mathbf{t}}(\Gamma)$, $s \geq 0$, is straightforward. The surface divergence is obtained as a formal $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$-adjoint $\mathrm{div}_\Gamma : \boldsymbol{L}^2_{\mathbf{t}}(\Gamma) \mapsto H^{-1}_*(\Gamma)$. Its range space is

$$H^{-s}_*(\Gamma) := \{\phi \in H^{-s}(\Gamma),\, \langle \chi, \phi\rangle_{s,\Gamma} = 0\, \forall \chi \in Z\}, \tag{2.1}$$

where Z is the space of piecewise constants on connected components of Γ and $\langle\cdot,\cdot\rangle_{s,\Gamma}$ denotes the $H^s(\Gamma) \times H^{-s}(\Gamma)$ duality pairing.

The two operators can be used to define the surface Laplace–Beltrami operator $\Delta_\Gamma : H^1(\Gamma) \mapsto H^{-1}_*(\Gamma)$ by $\Delta_\Gamma := \mathrm{div}_\Gamma \mathbf{grad}_\Gamma$. This will be a key tool, because it possesses the following lifting property shown in Theorem 5.3 of [19].

THEOREM 2.1. *If $f \in H^s_*(\Gamma)$ for $s \geq -1$, then the (unique) solution $u \in H^1(\Gamma)/Z$ of $-\Delta_\Gamma u = f$ belongs to $H^{1+r}(\Gamma)$ for $0 \leq r \leq \min\{s+1, s^*\}$, where $s^* > 0$ depends on the geometry of Γ in neighborhoods of vertices only.*

In other words, with $\tilde{C} = \tilde{C}(t,\Gamma)$ and $0 \leq r \leq \min\{s+1, s^*\}$,

$$f \in H^s_*(\Gamma), \quad -\Delta_\Gamma u = f \quad \Rightarrow \quad \|u\|_{H^{r+1}(\Gamma)/Z} \leq \tilde{C}\, \|f\|_{H^s(\Gamma)}\,. \tag{2.2}$$

We adopt the convention that C and c stand for generic positive constants, whose values might be different between different occurrences, but must not depend on any concrete function. When tagged with a tilde on top, they may depend only on ς, continuous function spaces, and the geometry of Γ.

Note that there exist polyhedral vertices for which $s^* > 0$ is arbitrarily small (see [19]). Nevertheless, reasonable geometries will allow for s^* to be well bounded above zero. For instance, if only three edges meet at a vertex O, we find $s^* = 2\pi/(\varphi_1+\varphi_2+\varphi_3) - \varepsilon$ for any $\varepsilon > 0$, where $\varphi_1, \varphi_2, \varphi_3$ are the opening angles at vertex O of the three plane faces Γ_j meeting at O.

Owing to Theorem 2.1, the space

$$H^{-\frac{1}{2}}(\Delta_\Gamma, \Gamma) := \{u \in H^1(\Gamma),\, \Delta_\Gamma u \in H^{-\frac{1}{2}}(\Gamma)\}$$

will actually be embedded in $H^{1+r}(\Gamma)$ for all $0 \leq r \leq \min\{\frac{3}{2}, s^*\}$. Based on div_Γ, we get the Hilbert spaces ($s \geq 0$)

$$\boldsymbol{H}^s(\mathrm{div}_\Gamma; \Gamma) := \{\mathbf{u} \in \boldsymbol{H}^s_{\mathbf{t}}(\Gamma),\, \mathrm{div}_\Gamma \mathbf{u} \in H^s(\Gamma)\}.$$

Tangential traces of vectorfields in $\boldsymbol{H}^1_{\mathrm{loc}}(\Omega)$ form the spaces $\boldsymbol{H}^{\frac{1}{2}}_{||}(\Gamma)$ and $\boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$ which were characterized in [17, Prop. 1.6]. Loosely speaking, $\boldsymbol{H}^{\frac{1}{2}}_{||}(\Gamma)$ contains the tangential surface vectorfields that are in $\boldsymbol{H}^{\frac{1}{2}}(\Gamma_i)$ for each smooth component Γ_i of Γ and feature

a suitable "weak tangential continuity" across the edges of the Γ_i. A corresponding "weak normal continuity" is satisfied by surface vectorfields in $\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$. For smooth Γ these spaces coincide with the spaces of tangential surface vectorfields in $\boldsymbol{H}_{\mathbf{t}}^{\frac{1}{2}}(\Gamma)$. The associated dual spaces will be denoted by $\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma)$, respectively, where the duality pairings are taken with $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$ as pivot space. Further, we denote by $\langle\cdot,\cdot\rangle_{||,\Gamma}$ and $\langle\cdot,\cdot\rangle_{\perp,\Gamma}$ the respective duality pairings. A fundamental result of [17] asserts that the tangential trace mapping $\pi_{\mathbf{t}} : \boldsymbol{H}_{\mathrm{loc}}^1(\Omega) \mapsto \boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)$ is continuous, surjective, and possesses a continuous right inverse (see Proposition 1.7 in [17]).

One of the crucial insights gained in [17] and [20] was that the tangential surface gradient $\mathbf{grad}_\Gamma : H^1(\Gamma) \mapsto \boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$ can be both extended and restricted to continuous, *closed*, and *injective* linear operators

$$\mathbf{grad}_\Gamma : \tilde{H}^{\frac{3}{2}}(\Gamma)/Z \mapsto \boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma), \quad \mathbf{grad}_\Gamma : H^{\frac{1}{2}}(\Gamma)/Z \mapsto \boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma)$$

(cf. Propositions 3.4 and 3.6 in [20]), where $\tilde{H}^{\frac{3}{2}}(\Gamma)$ is the space of traces of functions in $H^2(\Omega)$. Consequently, div_Γ also can be read as a continuous and *surjective* operator

$$\mathrm{div}_\Gamma : \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma) \mapsto \tilde{H}_*^{-\frac{3}{2}}(\Gamma), \quad \mathrm{div}_\Gamma : \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \mapsto H_*^{-\frac{1}{2}}(\Gamma).$$

This is important for the definition of the space $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, introduced in [17], as

$$\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) = \{\mathbf{v} \in \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma),\, \mathrm{div}_\Gamma \mathbf{v} \in H^{-\frac{1}{2}}(\Gamma)\}.$$

It is endowed with the natural graph norm $\|\cdot\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}$.

The key role of *Hodge decompositions* was emphasized in the introduction. The following theorem reveals the nature of the Hodge decomposition that we will need. More details are given in [20, sect. 5], [18], and [15].

THEOREM 2.2. *The space* $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ *has the direct and stable decomposition*

$$\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) := \mathbf{grad}_\Gamma\, H^{-\frac{1}{2}}(\Delta_\Gamma,\Gamma) \oplus (\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma) \cap \mathrm{Ker}(\mathrm{div}_\Gamma)).$$

Moreover, when restricted to $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma) \cap \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ *the decomposition is* $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$*-orthogonal.*

Proof. Any function in $\mathbf{grad}_\Gamma\, H^{-\frac{1}{2}}(\Delta_\Gamma,\Gamma) \cap \mathrm{Ker}(\mathrm{div}_\Gamma)$ must be the gradient of a function in the kernel of Δ_Γ on Γ. The latter contains only piecewise constants with respect to the connected components of Γ and, therefore, the decomposition is direct.

Next, pick some $\mathbf{v} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$. Since $\mathrm{div}_\Gamma : \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \mapsto H_*^{-\frac{1}{2}}(\Gamma)$ is surjective, we can find $\boldsymbol{\psi} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ such that $\mathrm{div}_\Gamma \boldsymbol{\psi} = \mathrm{div}_\Gamma \mathbf{v} \in H_*^{-\frac{1}{2}}(\Gamma)$. Define φ by

$$\varphi \in H^1(\Gamma)/Z: \quad (\mathbf{grad}_\Gamma \varphi, \mathbf{grad}_\Gamma \eta)_{0;\Gamma} = (\boldsymbol{\psi}, \mathbf{grad}_\Gamma \eta)_{0;\Gamma} \quad \forall \eta \in H^1(\Gamma)/Z,$$

that is, as the unique weak solution of $\Delta_\Gamma \varphi = \mathrm{div}_\Gamma \boldsymbol{\psi}$. This yields the decomposition

$$\mathbf{v} = \mathbf{grad}_\Gamma \varphi + (\boldsymbol{\psi} - \mathbf{grad}_\Gamma \varphi + \mathbf{v} - \boldsymbol{\psi}),$$

whose second part is readily seen to be divergence free. Since div_Γ is surjective, the open mapping theorem ensures that $\boldsymbol{\psi}$ can be chosen such that

$$\|\boldsymbol{\psi}\|_{\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)} \leq \tilde{C}\, \|\mathrm{div}_\Gamma \boldsymbol{\psi}\|_{H^{-\frac{1}{2}}(\Gamma)}\,.$$

This implies

$$\|\mathbf{grad}_\Gamma\,\varphi\|_{\boldsymbol{L}^2(\Gamma)} \le \|\boldsymbol{\psi}\|_{\boldsymbol{L}^2(\Gamma)} \le \|\boldsymbol{\psi}\|_{\boldsymbol{H}^{\frac12}_\perp(\Gamma)} \le \tilde{C}\,\|\mathrm{div}_\Gamma\boldsymbol{\psi}\|_{H^{-\frac12}(\Gamma)}\,,$$

which confirms the stability of the decomposition. For $\mathbf{v} \in \boldsymbol{L}^2_{\mathbf{t}}(\Gamma) \cap \boldsymbol{H}^{-\frac12}(\mathrm{div}_\Gamma,\Gamma)$, the $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$-orthogonality is immediate from the definition of div_Γ. ∎

In what follows, we write

$$\mathbf{X} := \mathbf{grad}_\Gamma\, H^{-\frac12}(\Delta_\Gamma,\Gamma) \quad \text{and} \quad \mathbf{N} := \boldsymbol{H}^{-\frac12}(\mathrm{div}_\Gamma,\Gamma) \cap \mathrm{Ker}(\mathrm{div}_\Gamma).$$

From the stability of the Hodge decomposition, we conclude that both $\mathbf{X}$ and $\mathbf{N}$ are closed subspaces of $\boldsymbol{H}^{-\frac12}(\mathrm{div}_\Gamma,\Gamma)$.

LEMMA 2.3. *If* $\mathbf{v} \in \mathbf{X}$ *satisfies* $\mathrm{div}_\Gamma\mathbf{v} \in H^s(\Gamma)$ *for some* $s \ge -\frac12$*, then for all* $0 \le r \le \min\{s+1, s^*\}$*,*

$$\mathbf{v} \in \boldsymbol{H}^r_{\mathbf{t}}(\Gamma) \quad \textit{and} \quad \|\mathbf{v}\|_{\boldsymbol{H}^r(\Gamma)} \le \tilde{C}\,\|\mathrm{div}_\Gamma\mathbf{v}\|_{H^s(\Gamma)}\,,$$

with a constant $\tilde{C} = \tilde{C}(r,s)$ *and* $s^* > 0$ *as in Theorem* 2.1.

Proof. $\mathbf{v} \in \mathbf{X}$ means $\mathbf{v} = \mathbf{grad}_\Gamma\,\varphi$ for some $\varphi \in H^1(\Gamma)$. By definition of $\mathbf{X}$, we see $\Delta_\Gamma\varphi = \mathrm{div}_\Gamma\mathbf{v}$, and the assertion follows from Theorem 2.1. ∎

In particular, we conclude that

$$\|\mathbf{v}\|_{\boldsymbol{H}^{-\frac12}_{||}(\Gamma)} \le \|\mathbf{v}\|_{\boldsymbol{L}^2(\Gamma)} \le \tilde{C}\,\|\mathrm{div}_\Gamma\mathbf{v}\|_{H^{-\frac12}(\Gamma)} \quad \forall\mathbf{v} \in \mathbf{X}. \tag{2.3}$$

3. Continuous variational problem. We recall the scalar single layer potential $\Psi^V_\varsigma : H^{-\frac12}(\Gamma) \mapsto H^1_{\mathrm{loc}}(\mathbb{R}^3)$ for the Helmholtz operator $\Delta + \varsigma^2$. Its relative, the vectorial Helmholtz single layer potential $\boldsymbol{\Psi}^{\mathbf{A}}_\varsigma(\mathbf{v})$ for $\mathbf{v} \in \boldsymbol{H}^{-\frac12}_{||}(\Gamma)$, is given by

$$\boldsymbol{\Psi}^{\mathbf{A}}_\varsigma(\mathbf{v})(\mathbf{x}) := \int_\Gamma \Phi_\varsigma(\mathbf{x},\mathbf{y})\mathbf{v}(\mathbf{y}), \quad \Phi_\varsigma(\mathbf{x},\mathbf{y}) := \frac{\exp(i\varsigma|\mathbf{x}-\mathbf{y}|)}{4\pi|\mathbf{x}-\mathbf{y}|}\,.$$

For every $\mathbf{v} \in \boldsymbol{H}^{-\frac12}_{||}(\Gamma)$, it defines a function in $\mathbf{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$ and, as a consequence of the trace theorem for $\pi_{\mathbf{t}}$, we can introduce the vectorial single layer boundary operator

$$\mathbf{A}_\varsigma : \boldsymbol{H}^{-\frac12}_{||}(\Gamma) \mapsto \boldsymbol{H}^{\frac12}_{||}(\Gamma), \quad \mathbf{A}_\varsigma := \pi_{\mathbf{t}} \circ \boldsymbol{\Psi}^{\mathbf{A}}_\varsigma\,,$$

in analogy to the scalar single layer integral operator

$$V_\varsigma : H^{-1/2}(\Gamma) \mapsto H^{\frac12}(\Gamma), \quad V_\varsigma := \gamma \circ \Psi^V_\varsigma,$$

where $\gamma : H^1_{\mathrm{loc}}(\mathbb{R}^3) \mapsto H^{\frac12}(\Gamma)$ is the standard trace operator. In the static case, i.e., at wavenumber $\varsigma = 0$, these operators are coercive.

LEMMA 3.1. *The operators* V_0 *and* $\mathbf{A}_0$ *are continuous, self-adjoint, and elliptic; i.e., there are constants* $\tilde{c}_1, \tilde{c}_2 > 0$ *depending only on* Γ *such that for all* $\mu \in H^{-\frac12}(\Gamma)$ *and all* $\boldsymbol{\mu} \in \boldsymbol{H}^{-\frac12}_{||}(\Gamma)$*,* $\mathrm{div}_\Gamma\boldsymbol{\mu} = 0$*,*

$$\langle V_0\mu,\mu\rangle_{\frac12,\Gamma} \ge \tilde{c}_1\,\|\mu\|^2_{H^{-\frac12}(\Gamma)}\,, \quad \langle \mathbf{A}_0\boldsymbol{\mu},\boldsymbol{\mu}\rangle_{||,\Gamma} \ge \tilde{c}_2\,\|\boldsymbol{\mu}\|^2_{\boldsymbol{H}^{-\frac12}_{||}(\Gamma)}\,.$$

Proof. See Corollary 8.13 in [41], or Theorem 3 in [30, Vol. IV, Chap. XI, sect. 2], and Theorem 6.2 in [37] or Proposition 4.1 in [19]. ∎

Along with the following result, this lemma yields the coercivity of V_ς and $\mathbf{A}_\varsigma$ (cf. the proof of Theorem 4.4 in [19]).

LEMMA 3.2. *The following operators are compact:*

$$\delta V_\varsigma := V_\varsigma - V_0 : H^{-\frac{1}{2}}(\Gamma) \mapsto H^{\frac{1}{2}}(\Gamma), \qquad \delta \mathbf{A}_\varsigma := \mathbf{A}_\varsigma - \mathbf{A}_0 : \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma) \mapsto \boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma).$$

Proof. We write G_ς for the Green's operator in $\mathbb{R}^3$ for the Helmholtz equation defined by $(G_\varsigma\varphi)(\mathbf{x}) = \int_{\mathbf{y}\in\mathbb{R}^3} \Phi_\varsigma(\mathbf{x},\mathbf{y})\varphi(\mathbf{y})d\mathbf{y}$ for $\varphi \in C_0^\infty(\mathbb{R}^3)$. With γ^* denoting a right inverse of the trace operator and appealing to the continuity of the trace map $\gamma : H^s_{loc}(\mathbb{R}^3) \to H^{s-1/2}(\Gamma)$, $s \in (1/2, 3/2)$ (see Lemma 3.6 in [27]), we find $\delta V_\varsigma = \gamma\circ(G_\varsigma - G_0)\circ\gamma^* : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ compactly, since the kernel $\Phi_\varsigma(\mathbf{x},\mathbf{y}) - \Phi_0(\mathbf{x},\mathbf{y})$ of the operator $G_\varsigma - G_0$ has essentially bounded derivatives in $\mathbf{x}$ and $\mathbf{y}$. The vectorial case follows in the same way. ∎

The main tool in the analysis of variational problem (1.1) is Hodge decompositions according to Theorem 2.2 (cf. [19, sect. 4.3]). Based on Theorem 2.2, we Hodge decompose $\mathbf{j} := \mathbf{j}^\perp + \mathbf{j}^0$, $\mathbf{j}^\perp \in \mathbf{X}$, $\mathbf{j}^0 \in \mathbf{N}$, and $\mathbf{v} := \mathbf{u}^\perp + \mathbf{v}^0$, $\mathbf{v}^\perp \in \mathbf{X}$, $\mathbf{v}^0 \in \mathbf{N}$, in (1.1). In this way, we end up with the following equivalent variational problem: Find $\mathbf{j}^\perp \in \mathbf{X}$, $\mathbf{j}^0 \in \mathbf{N}$ such that for all $\mathbf{v}^\perp \in \mathbf{X}$, $\mathbf{v}^0 \in \mathbf{N}$,

$$\begin{aligned} \left\langle V_\varsigma \mathrm{div}_\Gamma \mathbf{j}^\perp, \mathrm{div}_\Gamma \mathbf{v}^\perp \right\rangle_{\frac{1}{2},\Gamma} - \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^\perp \right\rangle_{||,\Gamma} - \quad & \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^\perp \right\rangle_{||,\Gamma} \;=\; f(\mathbf{v}^\perp), \\ \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^0 \right\rangle_{||,\Gamma} \qquad\qquad + \quad & \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^0 \right\rangle_{||,\Gamma} \;=\; -f(\mathbf{v}^0). \end{aligned} \tag{3.1}$$

Remember that $\langle .,. \rangle_{||,\Gamma}$ stands for the $\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma) \times \boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)$ duality pairing.

The natural setting for formulation (3.1) is the Hilbert space $\boldsymbol{\mathcal{G}} := \mathbf{X} \oplus \mathbf{N}$ endowed with the graph norm

$$\left\| (\mathbf{v}^\perp, \mathbf{v}^0) \right\|_{\boldsymbol{\mathcal{G}}}^2 := \left\| \mathbf{v}^\perp \right\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}^2 + \left\| \mathbf{v}^0 \right\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)}^2, \quad (\mathbf{v}^\perp, \mathbf{v}^0) \in \boldsymbol{\mathcal{G}}.$$

Thanks to Theorem 2.2, the space $\boldsymbol{\mathcal{G}}$ is isomorphic to $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ algebraically and topologically. The sesquilinear form $a : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ that belongs to (3.1) reads

$$\begin{aligned} a((\mathbf{j}^\perp, \mathbf{j}^0), (\mathbf{v}^\perp, \mathbf{v}^0)) := & \left\langle V_\varsigma \mathrm{div}_\Gamma \mathbf{j}^\perp, \mathrm{div}_\Gamma \mathbf{v}^\perp \right\rangle_{\frac{1}{2},\Gamma} \\ & - \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^\perp \right\rangle_{||,\Gamma} - \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^\perp \right\rangle_{||,\Gamma} \\ & + \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^0 \right\rangle_{||,\Gamma} + \varsigma^2 \left\langle \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^0 \right\rangle_{||,\Gamma} \end{aligned} \tag{3.2}$$

and is continuous, i.e.,

$$|a(\boldsymbol{\varphi}, \boldsymbol{\eta})| \le \tilde{C}_a \, \|\boldsymbol{\varphi}\|_{\boldsymbol{\mathcal{G}}} \, \|\boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \quad \forall \boldsymbol{\varphi}, \boldsymbol{\eta} \in \boldsymbol{\mathcal{G}}. \tag{3.3}$$

Using the form $a(\cdot,\cdot)$, we can express variational problem (3.1) succinctly as follows: Find $\boldsymbol{\iota} \in \boldsymbol{\mathcal{G}}$ such that

$$a(\boldsymbol{\iota}, \boldsymbol{\eta}) = f(\boldsymbol{\eta}) \quad \forall \boldsymbol{\eta} \in \boldsymbol{\mathcal{G}}, \tag{3.4}$$

where $f(\boldsymbol{\eta}) := f(\mathbf{v}^\perp) - f(\mathbf{v}^0)$, $\boldsymbol{\eta} := (\mathbf{v}^\perp, \mathbf{v}^0)$. We point out that (3.4) is entirely equivalent to (1.1) in the sense that, if $\mathbf{j} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$ is a solution of (1.1), then $\boldsymbol{\iota} := (\mathbf{j}^\perp, \mathbf{j}^0) \in \boldsymbol{\mathcal{G}}$ will solve (3.4). In particular, assertions on existence and uniqueness of solutions of (1.1) instantly carry over to (3.4) and vice versa.

4. Strong ellipticity. To establish strong ellipticity of the form $a(\cdot,\cdot)$ in (3.4), we proceed as in [19] and write $a = a_0 - k_0$, where $k_0 : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ reads

$$k_0((\mathbf{j}^\perp,\mathbf{j}^0),(\mathbf{v}^\perp,\mathbf{v}^0)) := -\left\langle \delta V_\varsigma \mathrm{div}_\Gamma \mathbf{j}^\perp, \mathrm{div}_\Gamma \mathbf{v}^\perp \right\rangle_{\frac{1}{2},\Gamma} + \varsigma^2 \left\langle \delta \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^\perp \right\rangle_{||,\Gamma} + \varsigma^2 \left\langle \delta \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^\perp \right\rangle_{||,\Gamma} - \varsigma^2 \left\langle \delta \mathbf{A}_\varsigma \mathbf{j}^\perp, \mathbf{v}^0 \right\rangle_{||,\Gamma} - \varsigma^2 \left\langle \delta \mathbf{A}_\varsigma \mathbf{j}^0, \mathbf{v}^0 \right\rangle_{||,\Gamma},$$

and where $a_0 : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ emerges from a by replacing V_ς with V_0 and A_ς with A_0. The next lemma is crucial for establishing the strong ellipticity of variational problem (3.4).

LEMMA 4.1. *The operator* $L : \mathbf{X} \mapsto \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)'$, *defined by* $L\mathbf{u}^\perp(\mathbf{z}) := \left\langle \mathbf{A}_0 \mathbf{u}^\perp, \mathbf{z} \right\rangle_{||,\Gamma}$ *for all* $\mathbf{u}^\perp \in \mathbf{X}$, $\mathbf{z} \in \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, *is compact.*

Proof. Consider a bounded sequence $(\mathbf{u}_n^\perp)_{n\in\mathbb{N}}$ in $\mathbf{X}$. By Lemma 2.3 it is also bounded in $\boldsymbol{H}_{\mathbf{t}}^{\frac{1}{2}}(\Gamma)$. By Rellich's theorem we can find a subsequence, also designated by $(\mathbf{u}_n^\perp)_n$, that converges in $\boldsymbol{L}_{\mathbf{t}}^2(\Gamma)$. Observe that, due to the continuity of the vectorial single layer boundary integral operator,

$$\left\| L\mathbf{z}^\perp \right\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)'} = \sup_{\mathbf{v}\in\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \frac{|(L\mathbf{z}^\perp)(\mathbf{v})|}{\|\mathbf{v}\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}} \leq \sup_{\mathbf{v}\in\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \frac{|\left\langle \mathbf{A}_0 \mathbf{z}^\perp, \mathbf{v} \right\rangle_{||,\Gamma}|}{\|\mathbf{v}\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)}} \leq \left\| \mathbf{A}_0 \mathbf{z}^\perp \right\|_{\boldsymbol{H}_{||}^{\frac{1}{2}}(\Gamma)} \leq \tilde{C} \left\| \mathbf{z}^\perp \right\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \leq \tilde{C} \left\| \mathbf{z}^\perp \right\|_{\boldsymbol{L}^2(\Gamma)}.$$

Thus $(L\mathbf{u}_n^\perp)_n$ will converge in $\mathbf{X}'$. ∎

At once, from $\mathbf{X} \subset \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ and $\mathbf{N} \subset \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$, we deduce that $L : \mathbf{X} \mapsto \mathbf{X}'$ and $L : \mathbf{X} \mapsto \mathbf{N}'$ are compact as well.

To establish the strong ellipticity of the form $a(\cdot,\cdot)$, we accordingly split $a_0(\cdot,\cdot)$ as $a_0 = d - k_1$, where the sesquilinear form $k_1 : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ is defined by

$$k_1((\mathbf{j}^\perp,\mathbf{j}^0),(\mathbf{v}^\perp,\mathbf{v}^0)) := \varsigma^2 \left\langle \mathbf{A}_0 \mathbf{j}^\perp, \mathbf{v}^\perp \right\rangle_{||,\Gamma} + \varsigma^2 \left\langle \mathbf{A}_0 \mathbf{j}^0, \mathbf{v}^\perp \right\rangle_{||,\Gamma} - \varsigma^2 \left\langle \mathbf{A}_0 \mathbf{j}^\perp, \mathbf{v}^0 \right\rangle_{||,\Gamma},$$

and the definite part $d : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ reads

$$d((\mathbf{j}^\perp,\mathbf{j}^0),(\mathbf{v}^\perp,\mathbf{v}^0)) := \left\langle V_0 \mathrm{div}_\Gamma \mathbf{j}^\perp, \mathrm{div}_\Gamma \mathbf{v}^\perp \right\rangle_{\frac{1}{2},\Gamma} + \varsigma^2 \left\langle \mathbf{A}_0 \mathbf{j}^0, \mathbf{v}^0 \right\rangle_{||,\Gamma}.$$

THEOREM 4.2. *The sesquilinear form* $a : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$ *is coercive; that is, it can be written as the difference of a* $\boldsymbol{\mathcal{G}}$*-elliptic sesquilinear form* d *and a compact sesquilinear form* $k : \boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}} \mapsto \mathbb{C}$.

Proof. Recall $a = a_0 - k_0$. Lemma 3.2 reveals that k_0 is a compact perturbation of a_0. Further, $a_0 = d - k_1$, and Lemma 4.1 implies that k_1 is a compact perturbation of d. From the ellipticity of the single layer boundary integral operators in Lemma 3.1, we immediately get

$$|d((\mathbf{v}^\perp,\mathbf{v}^0),(\mathbf{v}^\perp,\mathbf{v}^0))| = |\left\langle V_0 \mathrm{div}_\Gamma \mathbf{v}^\perp, \mathrm{div}_\Gamma \mathbf{v}^\perp \right\rangle_{\frac{1}{2},\Gamma} + \varsigma^2 \left\langle \mathbf{A}_0 \mathbf{v}^0, \mathbf{v}^0 \right\rangle_{||,\Gamma}| \geq \tilde{c}_1 \left\| \mathrm{div}_\Gamma \mathbf{v}^\perp \right\|^2_{H^{-\frac{1}{2}}(\Gamma)} + \tilde{c}_2 \varsigma^2 \left\| \mathbf{v}^0 \right\|^2_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)}$$

for all $(\mathbf{v}^\perp,\mathbf{v}^0) \in \boldsymbol{\mathcal{G}}$. Now, we can appeal to (2.3) and obtain

$$|d(\boldsymbol{\varphi},\boldsymbol{\varphi})| \geq \tilde{c}_d \|\boldsymbol{\varphi}\|^2_{\boldsymbol{\mathcal{G}}} \quad \forall \boldsymbol{\varphi} \in \boldsymbol{\mathcal{G}}.$$

Setting $k = k_0 + k_1$ yields $a = d - k$ with a principal part d, which is positive on $\boldsymbol{\mathcal{G}}$, and a compact perturbation k, as claimed. ∎

The strong ellipticity of the form $a(\cdot,\cdot)$ together with its injectivity ensured by Assumption 1.1 implies, as usual, the unique solvability of the EFIE (3.4) (and hence of (1.1)) for any admissible right-hand side. Moreover, there holds the continuous inf-sup condition for $a(\cdot,\cdot)$,

$$\sup_{\boldsymbol{\nu}\in\boldsymbol{\mathcal{G}}} \frac{|a(\boldsymbol{\varphi},\boldsymbol{\nu})|}{\|\boldsymbol{\nu}\|_{\boldsymbol{\mathcal{G}}}} \geq \tilde{c}_a \|\boldsymbol{\varphi}\|_{\boldsymbol{\mathcal{G}}} \quad \forall \boldsymbol{\varphi} \in \boldsymbol{\mathcal{G}}. \tag{4.1}$$

5. Boundary element spaces. We equip Γ with a family of shape-regular, quasi-uniform triangulations $(\Gamma_h)_{h>0}$ [25] comprising only flat triangles. The parameter h designates the meshwidth, that is, the length of the longest edge. Let $\mathbb{H}$ stand for the collection of meshwidths occurring in $(\Gamma_h)_{h\in\mathbb{H}}$ and assume that $\mathbb{H} \subset \mathbb{R}^+$ forms a decreasing sequence converging to zero. The set $\mathcal{T}_h$ will include all triangles of Γ_h, and $\mathcal{E}_h$ stands for the set of edges of Γ_h.

Using the local coordinate systems on the faces Γ_j, $j = 1,\ldots,N_\Gamma$, each $T \in \mathcal{T}_h$ can be embedded in $\mathbb{R}^2$. Then we can define the local spaces (cf. [45])

$$\boldsymbol{\mathcal{RT}}_0(T) := \{\mathbf{x} \mapsto \mathbf{a} + \beta\mathbf{x},\, \mathbf{a} \in \mathbb{C}^2,\, \beta \in \mathbb{C}\}, \quad T \in \mathcal{T}_h.$$

They give rise to the global boundary element space

$$\boldsymbol{\mathcal{RT}}_0(\Gamma_h) := \{\mathbf{v} \in \boldsymbol{H}(\mathrm{div}_\Gamma;\Gamma),\, \mathbf{v}_{|T} \in \boldsymbol{\mathcal{RT}}_0(T)\, \forall T \in \mathcal{T}_h\}.$$

Keep in mind that this definition is based on a weak notion of div_Γ. So Green's formula applied to the surface triangles can be used to confirm that the "edge-normal" components of the tangential vectorfields in $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ must be continuous across interelement edges. This renders the following degrees of freedom well-defined:

$$\phi_e : \boldsymbol{\mathcal{RT}}_0(\Gamma_h) \mapsto \mathbb{C}, \qquad \phi_e(\mathbf{v}_h) := \int_e (\mathbf{v}_h \times \mathbf{n}_j) \cdot d\vec{s}, \quad e \in \mathcal{E}_h\ ,$$

where $\mathbf{n}_j$ is the normal of a face Γ_j in whose closure e is contained. Given the degrees of freedom, we have nodal interpolation operators Π_h onto $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ at our disposal. To begin with, those can be declared for $\{\Gamma_j\}$-piecewise continuous tangential surface vectorfields whose edge-normal components are continuous, too. It turns out that this is not enough, and we badly need to apply Π_h to less regular surface vectorfields. A first step towards this goal is the following lemma (cf. formula (3.40) in [14]).

LEMMA 5.1. *For any $s > 0$ the local interpolation operator $\Pi_T : \boldsymbol{H}^s(T) \cap \boldsymbol{H}(\mathrm{div};T) \mapsto \boldsymbol{\mathcal{RT}}_0(T)$, $T \in \mathcal{T}_h$, is continuous.*

Proof. Only the case $s \leq \frac{1}{2}$ is of interest. We consider a single degree of freedom on T as follows: Pick an edge $e \subset \partial T$ and regard its characteristic function χ_e as an element in $W_q^{1-\frac{1}{q}}(e)$ for $q := 1+s$. As $1 < q < 2$, Theorem 1.4.5.2 of [34] reveals that extension by zero of χ_e onto all of ∂T will provide a function $\tilde{\psi}$ in $W_q^{1-\frac{1}{q}}(\partial T)$. Then we can use the trace theorem [34, Thm. 1.5.1.3] to extend $\tilde{\psi}$ to a function $\psi \in W_q^1(T)$ in a continuous fashion. Using Green's formula, extended by continuity, we estimate for any smooth vectorfield $\mathbf{v}$ that

$$\begin{aligned}\int_e \mathbf{v}\cdot\mathbf{n}_e\, ds &= \int_{\partial T} \tilde{\psi}\mathbf{v}\cdot\mathbf{n}\, ds = \int_T (\mathbf{grad}\,\psi\cdot\mathbf{v} + \psi\,\mathrm{div}\,\mathbf{v})\, d\mathbf{x}\\ &\leq \|\mathbf{grad}\,\psi\|_{L^q(T)}\,\|\mathbf{v}\|_{L^p(T)} + \|\psi\|_{L^2(T)}\,\|\mathrm{div}\,\mathbf{v}\|_{\boldsymbol{L}^2(T)}\,,\end{aligned}$$

where p is the exponent conjugate to q, i.e., $p^{-1} + q^{-1} = 1$. The Sobolev embedding theorem [1, Thm. 4.5] gives the continuous inclusions $W_q^1(T) \hookrightarrow L^2(T)$ and $\boldsymbol{H}^s(T) \hookrightarrow \boldsymbol{L}^p(T)$. This implies, with $\tilde{C} = \tilde{C}(s,T)$,

$$\int_e \mathbf{v} \cdot \mathbf{n}_e \, ds \leq \tilde{C} \left(\|\mathbf{grad}\, \psi\|^2_{L^q(T)} + \|\psi\|^2_{W_q^1(T)} \right)^{\frac{1}{2}} \left(\|\mathbf{v}\|^2_{\boldsymbol{H}^s(T)} + \|\mathrm{div}\, \mathbf{v}\|^2_{\boldsymbol{L}^2(T)} \right)^{\frac{1}{2}}$$

for all $\mathbf{v} \in \boldsymbol{H}^s(T) \cap \boldsymbol{H}(\mathrm{div}; T)$ and the assertion of the lemma, since ψ is fixed. ∎

The importance of the interpolation operators Π_h can be traced back to the *commuting diagram property* as follows [14, Prop. 3.7]:

$$\mathrm{div}_\Gamma \Pi_h \mathbf{v} = Q_h \mathrm{div}_\Gamma \mathbf{v} \quad \forall \mathbf{v} \in \boldsymbol{H}(\mathrm{div}; \Gamma) \cap \mathrm{Dom}(\Pi_h), \tag{5.1}$$

where Q_h is the $L^2(\Gamma)$-orthogonal projection onto the space

$$\mathcal{Q}_0(\Gamma_h) := \{\mu \in L^2(\Gamma),\, \mu_{|T} = \text{const.}\, \forall T \in \mathcal{T}_h\}.$$

Identity (5.1) is a simple consequence of the definition of the degrees of freedom and Gauss's theorem applied to elements. An important consequence is that

$$\mathrm{div}_\Gamma \mathbf{v} = 0 \quad \wedge \quad \mathbf{v} \in \mathrm{Dom}(\Pi_h) \qquad \Rightarrow \quad \mathrm{div}_\Gamma(\Pi_h \mathbf{v}) = 0.$$

The relationship (5.1) also reveals that $\mathrm{div}_\Gamma \boldsymbol{\mathcal{RT}}_0(\Gamma_h) = \mathcal{Q}_0(\Gamma_h)$.

Remark. The reader should be aware that we have restricted ourselves to lowest order Raviart–Thomas elements only for the sake of simplicity. All other $\boldsymbol{H}(\mathrm{div}; \Omega)$-conforming finite elements in two dimensions that provide valid discrete 1-forms could be used as well. A rich collection is offered in [14, sect. III.3]. All arguments in what follows will carry over to these elements with only slight alterations.

The Raviart–Thomas elements form an affine family of finite elements in the sense of [25] with respect to Piola's transformation [14, sect. III.1.3],

$$\mathfrak{P}_T : \boldsymbol{L}^2(\widehat{T}) \mapsto \boldsymbol{L}^2_{\mathbf{t}}(T)\,, \quad \mathfrak{P}_T(\widehat{\mathbf{v}}_h)(\mathbf{x}) := |\det D\boldsymbol{\Phi}_T|^{-1} D\boldsymbol{\Phi}_T \widehat{\mathbf{v}}_h(\boldsymbol{\Phi}_T^{-1}(\mathbf{x})), \quad \mathbf{x} \in T,$$

where $\widehat{T}$ is the reference triangle $\widehat{T} := \{\mathbf{x} \in \mathbb{R}^2,\, x_1, x_2 > 0,\, x_1 + x_2 < 1\}$, $T \in \mathcal{T}_h$, and $\boldsymbol{\Phi}_T$ the unique affine mapping that takes $\widehat{T}$ to T. The Piola transform preserves the values of degrees of freedom. Shape regularity and quasi-uniformity guarantee that $|\det D\boldsymbol{\Phi}_T| \asymp h^2$ and $\|D\boldsymbol{\Phi}_T\| \asymp h$ uniformly in $T \in \mathcal{T}_h$ and $h \in \mathbb{H}$. Here and in what follows, we use the symbol $\asymp$ to indicate equivalence up to constants that may depend on Γ and the shape regularity of $\{\Gamma_h\}_h$, but are independent of h. The same is understood of all generic constants unless they bear a tilde.

Now, using standard affine equivalence techniques, the effect of Piola's transform on fractional Sobolev norms can be controlled as shown below.

LEMMA 5.2. *The Piola transform $\mathfrak{P}_T$, $T \in \mathcal{T}_h$, satisfies, for $0 \leq s \leq 1$,*

$$|\widehat{\mathbf{u}}|_{\boldsymbol{H}^s(\widehat{T})} \asymp h^s \, |\mathfrak{P}_T \widehat{\mathbf{u}}|_{\boldsymbol{H}^s(T)} \quad \forall \widehat{\mathbf{u}} \in \boldsymbol{H}^s(\widehat{T}),$$

with constants depending only on the shape regularity of T.

Proof. See Lemma 3 in [45] for the cases $s = 0$ and $s = 1$. The rest follows by interpolation. ∎

Remark. Using Piola's transform, one easily constructs parametric divergence-conforming surface elements [8, 32] for piecewise smooth Γ. Thus, our approach can be instantly extended to curved Lipschitz polyhedra.

6. Hodge mapping. Coercivity of the sesquilinear form related to (1.1) could only be established in the split space $\boldsymbol{\mathcal{G}}$ arising from the Hodge decomposition. This means that, though the boundary element spaces $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ are conforming and natural for the EFIE (1.1), Theorem 4.2, i.e., the validity of a Gårding inequality on the continuous level, gives no immediate information about the convergence of the Galerkin discretization. The reason is that we would need conforming finite element subspaces of both $\mathbf{X}$ and $\mathbf{N}$ in order to apply the usual results (cf., e.g., [47, sect. 2.3]) about the convergence of Galerkin schemes for strongly elliptic variational problems.

A discrete $\boldsymbol{L}^2_{\mathbf{t}}(\Gamma)$-orthogonal Hodge decomposition

$$\boldsymbol{\mathcal{RT}}_0(\Gamma_h) = \mathbf{X}_h \oplus \mathbf{N}_h, \quad \mathbf{N}_h := \mathrm{Ker}(\mathrm{div}_\Gamma) \cap \boldsymbol{\mathcal{RT}}_0(\Gamma_h), \tag{6.1}$$

yields $\mathbf{N}_h \subset \mathbf{N}$, but generally we cannot expect $\mathbf{X}_h \subset \mathbf{X}$. In short, $\mathbf{X}_h$ provides only a *nonconforming* discretization of $\mathbf{X}$, and a discrete inf-sup condition does not follow from the strong ellipticity of the continuous problem. On the other hand, no modification of the sesquilinear form $a(\cdot,\cdot)$ is necessary if we consider the variational problem (3.4) over $\boldsymbol{\mathcal{G}}_h := \mathbf{X}_h \times \mathbf{N}_h$. This is simply due to the fact that everything remains perfectly conforming in $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$. In particular, $\boldsymbol{\mathcal{G}}_h$ can be equipped with the graph norm $\|\cdot\|_{\boldsymbol{\mathcal{G}}}$ of $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) \times \boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$. However, embedding and regularity properties of $\mathbf{X}$ (cf. Lemmas 2.3 and 4.1) are crucial and the space $\mathbf{X}_h$ lacks them. We deal with this by introducing semidiscrete spaces arising from the *continuous* Hodge decomposition of the discrete boundary element space as follows: We split $\mathbf{v}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ in two ways,

$$\mathbf{v}_h = \mathbf{v}_h^\perp + \mathbf{v}_h^0, \quad \mathbf{v}_h^\perp \in \mathbf{X}_h,\, \mathbf{v}_h^0 \in \mathbf{N}_h \quad \text{and} \quad \mathbf{v}_h = \mathbf{v}^\perp + \mathbf{v}^0, \quad \mathbf{v}^\perp \in \mathbf{X},\, \mathbf{v}^0 \in \mathbf{N}.$$

The discrete field $\mathbf{v}_h^\perp$ is a genuine boundary element function, but only the *semidiscrete* field $\mathbf{v}^\perp$ has the desired properties. We have labeled it semidiscrete because $\mathrm{div}_\Gamma \mathbf{v}^\perp = \mathrm{div}_\Gamma \mathbf{v}_h$ is still piecewise constant and, hence, $\mathbf{v}^\perp$ still depends on the triangulation. To bridge the gap between $\mathbf{v}_h^\perp$ and $\mathbf{v}^\perp$, we need the following device (cf. Def. 4.1 in [38]).

DEFINITION 6.1. *We define the* Hodge mapping $H_h : \boldsymbol{\mathcal{RT}}_0(\Gamma_h) \mapsto \mathbf{X}$ *by*

$$H_h \mathbf{v}_h \in \mathbf{X}: \quad \mathrm{div}_\Gamma H_h \mathbf{v}_h := \mathrm{div}_\Gamma \mathbf{v}_h, \quad \mathbf{v}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h).$$

Owing to (2.3), H_h is well defined. The Hodge mappings are uniformly continuous with respect to $h \in \mathbb{H}$. They create the desired link between $\mathbf{X}_h$ and $\mathbf{X}$ (cf. Lemma 4.2 in [38]) as follows.

LEMMA 6.2. *For any $s \geq -\frac{1}{2}$, the Hodge mappings satisfy the estimate*

$$\|\mathbf{v}_h - H_h \mathbf{v}_h\|_{\boldsymbol{L}^2(\Gamma)} \leq C h^r \, \|\mathrm{div}_\Gamma \mathbf{v}_h\|_{H^s(\Gamma)} \quad \forall \mathbf{v}_h \in \mathbf{X}_h \; \forall h \in \mathbb{H}$$

with $0 \leq r \leq \min\{s+1, 1, s^*\}$ and constants depending only on s, r, Γ, and the shape regularity of the surface triangulations.

Proof. We follow the proof of Lemma 4.2 from [38], pick $\mathbf{u}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)$, and focus on a single triangle $T \in \Gamma_h$. Take $H_h \mathbf{u}_{h|T}$ to the reference element and set $\widehat{\mathbf{w}} := \mathfrak{P}_T^{-1} H_h \mathbf{u}_h$. By (2.3), we see $\widehat{\mathbf{w}} \in \boldsymbol{H}^r(\widehat{T})$ so that the assumptions of Lemma 5.1 are satisfied. We have for any $r > 0$

$$\|\hat{\Pi}\widehat{\mathbf{w}}\|_{\boldsymbol{L}^2(\widehat{T})} \leq \tilde{C}(r) \left(\|\widehat{\mathbf{w}}\|_{\boldsymbol{H}^r(\widehat{T})} + \|\mathrm{div}\, \widehat{\mathbf{w}}\|_{L^2(\widehat{T})} \right),$$

where $\hat{\Pi}$ is the local interpolation operator on $\widehat{T}$. Remember that $\operatorname{div}_\Gamma H_h \mathbf{u}_h$ is piecewise constant, which renders $\operatorname{div} \widehat{\mathbf{w}}$ constant. Exploiting the equivalence of all norms on finite-dimensional spaces, we can easily bound $\|\operatorname{div} \widehat{\mathbf{w}}\|_{L^2(\widehat{T})}$ and arrive at

$$\|\hat{\Pi}\widehat{\mathbf{w}}\|_{\boldsymbol{L}^2(\widehat{T})} \leq \tilde{C}(r) \, \|\widehat{\mathbf{w}}\|_{\boldsymbol{H}^r(\widehat{T})} .$$

Constant vectorfields on $\widehat{T}$ are preserved by the interpolation $\hat{\Pi}$. Thus, for any $\mathbf{p} \in \mathbb{C}^2$,

$$\|\widehat{\mathbf{w}} - \hat{\Pi}\widehat{\mathbf{w}}\|_{\boldsymbol{L}^2(\widehat{T})} = \|\widehat{\mathbf{w}} - \mathbf{p} - \hat{\Pi}(\widehat{\mathbf{w}} - \mathbf{p})\|_{\boldsymbol{L}^2(\widehat{T})} \leq \|\widehat{\mathbf{w}} - \mathbf{p}\|_{\boldsymbol{L}^2(\widehat{T})} + \tilde{C}(r) \, \|\widehat{\mathbf{w}} - \mathbf{p}\|_{\boldsymbol{H}^r(\widehat{T})} .$$

From the definition of the fractional Sobolev norm [34, Def. 1.3.2.1] and $0 \leq r \leq 1$, it is immediate that

$$\|\widehat{\mathbf{w}} - \mathbf{p}\|^2_{\boldsymbol{H}^r(\widehat{T})} = \|\widehat{\mathbf{w}} - \mathbf{p}\|^2_{\boldsymbol{L}^2(\widehat{T})} + |\widehat{\mathbf{w}}|^2_{\boldsymbol{H}^r(\widehat{T})} .$$

As, according to Proposition 6.1 in [33], a Bramble–Hilbert-type estimate of the form

$$\inf_{c \in \mathbb{R}} \|f - c\|_{L^2(\widehat{T})} \leq \tilde{C}(r) \, |f|_{H^r(\widehat{T})} \quad \forall f \in H^r(\widehat{T})$$

also holds in fractional Sobolev spaces, we end up with the estimate

$$\|\widehat{\mathbf{w}} - \hat{\Pi}\widehat{\mathbf{w}}\|_{\boldsymbol{L}^2(\widehat{T})} \leq \tilde{C}(r) \, |\widehat{\mathbf{w}}|_{H^r(\widehat{T})} .$$

Since interpolation and the Piola transform commute, we may use Lemma 5.2 to pull the estimate back to the element T,

$$\|H_h \mathbf{u}_h - \Pi_h H_h \mathbf{u}_h\|_{\boldsymbol{L}^2(T)} \leq C h^r \, \|H_h \mathbf{u}_h\|_{\boldsymbol{H}^r(T)} .$$

At this stage, shape regularity starts affecting the constants. Squaring and summing over all elements yields

$$\|H_h \mathbf{u}_h - \Pi_h H_h \mathbf{u}_h\|_{\boldsymbol{L}^2(\Gamma)} \leq C h^r \, \|H_h \mathbf{u}_h\|_{\boldsymbol{H}^r(\Gamma)} ,$$

which, in light of Lemma 2.3, involves

$$\|H_h \mathbf{u}_h - \Pi_h H_h \mathbf{u}_h\|_{\boldsymbol{L}^2(\Gamma)} \leq C h^r \, \|\operatorname{div}_\Gamma \mathbf{u}_h\|_{H^s(\Gamma)} . \tag{6.2}$$

By the commuting diagram property of Π_h, we conclude from $\operatorname{div}_\Gamma(\mathbf{u}_h - H_h \mathbf{u}_h) = 0$ that also $\operatorname{div}_\Gamma(\mathbf{u}_h - \Pi_h H_h \mathbf{u}_h) = 0$. This means $\mathbf{u}_h - \Pi_h H_h \mathbf{u}_h \in \mathbf{N}_h$ and makes it possible for us to apply Nédélec's trick [44, sect. 3.3],

$$\begin{aligned} \|\mathbf{u}_h - H_h \mathbf{u}_h\|^2_{\boldsymbol{L}^2(\Gamma)} &= (\mathbf{u}_h - H_h \mathbf{u}_h, \mathbf{u}_h - \Pi_h H_h \mathbf{u}_h + \Pi_h H_h \mathbf{u}_h - H_h \mathbf{u}_h)_{0;\Gamma} \\ &= (\mathbf{u}_h - H_h \mathbf{u}_h, \Pi_h H_h \mathbf{u}_h - H_h \mathbf{u}_h)_{0;\Gamma} . \end{aligned}$$

Together with (6.2) this shows the assertion of the lemma. ∎

Now, we *fix* $t \leq \min\{\frac{1}{2}, s^*\}$ and keep it constant for the remainder of this paper. A legal choice for r in the previous lemma is $r = t$ for $s = -\frac{1}{2}$, and we denote the associated constant by C_3.

LEMMA 6.3. *The decomposition* $\boldsymbol{\mathcal{RT}}_0(\Gamma_h) = \mathbf{X}_h \oplus \mathbf{N}_h$ *is uniformly* $\boldsymbol{H}^{-\frac{1}{2}}(\operatorname{div}_\Gamma, \Gamma)$*-stable.*

Proof. For $\mathbf{u}_h \in \mathbf{X}_h$ we can use the Hodge mapping and the previous lemma to estimate

$$\|\mathbf{u}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \leq \|\mathbf{u}_h - H_h\mathbf{u}_h\|_{\boldsymbol{L}^2(\Gamma)} + \|H_h\mathbf{u}_h\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \leq C(h^t+1)\,\|\mathrm{div}_\Gamma \mathbf{u}_h\|_{H^{-\frac{1}{2}}(\Gamma)},$$

as $H_h\mathbf{u}_h \in \mathbf{X}$. Since $\mathrm{div}_\Gamma H_h\mathbf{u}_h = \mathrm{div}_\Gamma \mathbf{u}_h$ and $\mathbb{H}$ is bounded, the proof is finished. ∎

We also shall require the following right inverse of the Hodge mapping.

DEFINITION 6.4. *We define the linear continuous mappings $T_h : \mathbf{X} \mapsto \mathbf{X}_h$, $h \in \mathbb{H}$, by*

$$T_h\mathbf{w} \in \mathbf{X}_h : \quad \mathrm{div}_\Gamma T_h\mathbf{w} = Q_h^{-\frac{1}{2}}\mathrm{div}_\Gamma \mathbf{w} \quad \forall \mathbf{w} \in \mathbf{X},$$

where $Q_h^{-\frac{1}{2}} : H^{-\frac{1}{2}}(\Gamma) \mapsto \mathcal{Q}_0(\Gamma_h)$ are the $H^{-\frac{1}{2}}(\Gamma)$-orthogonal projections.

Note that it is only due to the preceding stability result that this definition makes real sense. In addition, Lemma 6.3 guarantees that the family of operators $(T_h)_{h\in\mathbb{H}}$ is uniformly continuous, as

$$\|T_h\mathbf{w}\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \leq C\,\|\mathrm{div}_\Gamma T_h\mathbf{w}\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C\,\|\mathrm{div}_\Gamma \mathbf{w}\|_{H^{-\frac{1}{2}}(\Gamma)}.$$

LEMMA 6.5. *For any fixed $\mathbf{w} \in \mathbf{X}$ we have*

$$\lim_{h\to 0} \|\mathbf{w} - T_h\mathbf{w}\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} = 0.$$

Proof. We resort to the same trick as in the proof of Lemma 6.3 and use $H_hT_h\mathbf{w} - \mathbf{w} \in \mathbf{X}$,

$$\begin{aligned}
\|T_h\mathbf{w} - \mathbf{w}\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} &\leq \|H_hT_h\mathbf{w} - T_h\mathbf{w}\|_{\boldsymbol{L}^2(\Gamma)} + \|H_hT_h\mathbf{w} - \mathbf{w}\|_{\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)} \\
&\leq C_3h^t\,\|\mathrm{div}_\Gamma T_h\mathbf{w}\|_{H^{-\frac{1}{2}}(\Gamma)} + \tilde{C}\,\|\mathrm{div}_\Gamma(H_hT_h\mathbf{w} - \mathbf{w})\|_{H^{-\frac{1}{2}}(\Gamma)} \\
&\leq Ch^t\,\|\mathrm{div}_\Gamma \mathbf{w}\|_{H^{-\frac{1}{2}}(\Gamma)} + \tilde{C}\inf_{\mu_h\in\mathcal{Q}_0(\Gamma_h)} \|\mathrm{div}_\Gamma \mathbf{w} - \mu_h\|_{H^{-\frac{1}{2}}(\Gamma)}.
\end{aligned}$$

As $\bigcup_{h\in\mathbb{H}} \mathcal{Q}_0(\Gamma_h)$ is dense in $L^2(\Gamma)$, which, in turn, is dense in $H^{-\frac{1}{2}}(\Gamma)$, the lemma holds true. ∎

Remark. The operator T_h of Definition 6.4 is closely related to the so-called *Fortin projector* for edge elements introduced by Boffi [10].

7. Stability of the Galerkin scheme. Galerkin discretization of (3.4) leads to the following discrete variational problem: Seek $\boldsymbol{\iota}_h \in \boldsymbol{\mathcal{G}}_h$ such that

$$a(\boldsymbol{\iota}_h, \boldsymbol{\eta}_h) = f(\boldsymbol{\eta}_h) \quad \forall \boldsymbol{\eta}_h \in \boldsymbol{\mathcal{G}}_h. \tag{7.1}$$

The discrete Hodge decomposition (6.1) shows (7.1) to be equivalent to the Galerkin discretization (1.2) of the EFIE (1.1). From Theorem 4.2, we saw that problem (3.4) is strongly elliptic, i.e., $a = d - k$, with a $\boldsymbol{\mathcal{G}}$-elliptic sesquilinear form d and a $\boldsymbol{\mathcal{G}}$-compact form k. Discretization of (3.4) by a dense and conforming family of finite-dimensional subspaces would therefore imply quasi-optimal asymptotic convergence of the approximate solutions. The problem here is that $\boldsymbol{\mathcal{G}}_h$ is generally nonconforming, i.e., $\boldsymbol{\mathcal{G}}_h \not\subset \boldsymbol{\mathcal{G}}$. Therefore, coercivity in the discrete setting must be established by

a separate argument. For the proof, we draw on an idea of Schatz [46]. A similar strategy is pursued in [21, sect. 4].

To get compact formulas, we replace bilinear forms by the associated Riesz operators. First, $A : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}'$ is associated to the sesquilinear form a. Next, the operator $K : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}'$ is associated with the sesquilinear form k defined in the proof of Theorem 4.2. Both operators are continuous from $\boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}'$. However, since $\boldsymbol{\mathcal{G}}_h$ is nonconforming, these operators are not defined on $\boldsymbol{\mathcal{G}}_h$ a priori. To extend them, we use Hodge mappings on $\boldsymbol{\mathcal{G}}_h$ which are defined through

$$\mathbf{H}_h : \boldsymbol{\mathcal{G}}_h \mapsto \boldsymbol{\mathcal{G}}, \quad \mathbf{H}_h(\mathbf{v}_h^\perp, \mathbf{v}_h^0) := (H_h \mathbf{v}_h^\perp, \mathbf{v}_h^0) \in \boldsymbol{\mathcal{G}}, \quad (\mathbf{v}_h^\perp, \mathbf{v}_h^0) \in \boldsymbol{\mathcal{G}}_h.$$

Lemma 6.2 ensures the uniform boundedness in h of this family of operators. We also define the extension $\mathbf{T}_h : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}_h$ to $\boldsymbol{\mathcal{G}}_h$ of the right inverses T_h, $h \in \mathbb{H}$, of the Hodge mappings from Definition 6.4:

$$\mathbf{T}_h(\mathbf{v}^\perp, \mathbf{v}^0) := (T_h \mathbf{v}^\perp, \mathbf{Q}_h^{-\frac{1}{2}} \mathbf{v}^0) \in \boldsymbol{\mathcal{G}}_h, \quad (\mathbf{v}^\perp, \mathbf{v}^0) \in \boldsymbol{\mathcal{G}},$$

where $\mathbf{Q}_h^{-\frac{1}{2}}$ is the $\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)$-orthogonal projection $\mathbf{N} \mapsto \mathbf{N}_h$. The operator $\mathbf{T}_h$ is well defined, since $\mathbf{N}_h \subset \mathbf{N}$ and $\mathbf{N}$ is a closed subspace of $\boldsymbol{H}_{||}^{-\frac{1}{2}}(\Gamma)$. Density of $\bigcup_{h\in\mathbb{H}} \mathbf{N}_h$ in $\mathbf{N}$ and Lemma 6.5 confirm pointwise convergence

$$\lim_{h\to 0} \|\boldsymbol{\varphi} - \mathbf{T}_h \boldsymbol{\varphi}\|_{\boldsymbol{\mathcal{G}}} = 0 \quad \forall \boldsymbol{\varphi} \in \boldsymbol{\mathcal{G}}. \tag{7.2}$$

Next, we consider the operator $S : \boldsymbol{\mathcal{G}}' \mapsto \boldsymbol{\mathcal{G}}$ defined as the solution operator of the $\boldsymbol{\mathcal{G}}$-elliptic variational problem

$$\boldsymbol{\eta}' \in \boldsymbol{\mathcal{G}}' : \quad d(S\boldsymbol{\eta}', \boldsymbol{\varphi}) = \boldsymbol{\eta}'(\boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \boldsymbol{\mathcal{G}}.$$

Continuity and ellipticity of the sesquilinear form d ensure that S is well defined and give

$$\tilde{C}_d^{-1} \|\boldsymbol{\eta}'\|_{\boldsymbol{\mathcal{G}}'} \le \|S\boldsymbol{\eta}'\|_{\boldsymbol{\mathcal{G}}} \le \tilde{c}_d^{-1} \|\boldsymbol{\eta}'\|_{\boldsymbol{\mathcal{G}}'} \quad \forall \boldsymbol{\eta}' \in \boldsymbol{\mathcal{G}}', \tag{7.3}$$

where $\tilde{C}_d := \|d\|$. Note also that the operator S is confined to the continuous setting.

LEMMA 7.1. *There is a decreasing function $b : \mathbb{H} \mapsto \mathbb{R}^+$ with $b(h) \to 0$ as $h \to 0$ such that*

$$\|(\mathbf{T}_h - Id)SK\boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \le b(h) \|\boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \quad \forall \boldsymbol{\eta} \in \boldsymbol{\mathcal{G}}.$$

Proof. We follow the ideas of the proof of Corollary 10.4 in [40]. Set $B_1(\boldsymbol{\mathcal{G}}) := \{\boldsymbol{\varphi} \in \boldsymbol{\mathcal{G}} : \|\boldsymbol{\varphi}\|_{\boldsymbol{\mathcal{G}}} \le 1\}$. As $K : \boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}'$ is compact, the set $KB_1(\boldsymbol{\mathcal{G}})$ is precompact in $\boldsymbol{\mathcal{G}}'$. Thanks to the continuity of S, the closure w.r.t. the $\|\cdot\|_{\boldsymbol{\mathcal{G}}}$-norm $M := \overline{SKB_1(\boldsymbol{\mathcal{G}})}$ is compact in $\boldsymbol{\mathcal{G}}$. Pick some $\epsilon > 0$ and write $B_\epsilon(\boldsymbol{\nu})$ for the ϵ-neighborhood of $\boldsymbol{\nu}$ in $\boldsymbol{\mathcal{G}}$. We can find finitely many $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_L$, $L = L(\epsilon) \in \mathbb{N}$, in M such that $M \subset \bigcup_l B_\epsilon(\boldsymbol{\nu}_l)$. From (7.2) we learn that there is $h_0 = h_0(\epsilon) \in \mathbb{H}$ such that

$$\|\mathbf{T}_h \boldsymbol{\nu}_l - \boldsymbol{\nu}_l\|_{\boldsymbol{\mathcal{G}}} \le \epsilon \quad \forall h < h_0, \qquad l = 1, \ldots, L.$$

For any $\boldsymbol{\eta} \in M$, there exists a $\boldsymbol{\nu}_l$ such that $\boldsymbol{\eta} \in B_\epsilon(\boldsymbol{\nu}_l)$. Hence,

$$\|\mathbf{T}_h \boldsymbol{\eta} - \boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \le \|\mathbf{T}_h \boldsymbol{\eta} - \mathbf{T}_h \boldsymbol{\nu}_l\|_{\boldsymbol{\mathcal{G}}} + \|\mathbf{T}_h \boldsymbol{\nu}_l - \boldsymbol{\nu}_l\|_{\boldsymbol{\mathcal{G}}} + \|\boldsymbol{\nu}_l - \boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \le (\|\mathbf{T}_h\|_{\boldsymbol{\mathcal{G}} \mapsto \boldsymbol{\mathcal{G}}} + 2)\epsilon$$

if $h < h_0$. Undoing the substitutions, we get

$$\|(\mathbf{T}_h - Id)SK\boldsymbol{\eta}\|_{\boldsymbol{\mathcal{G}}} \leq (\|\mathbf{T}_h\|_{\boldsymbol{\mathcal{G}}\mapsto\boldsymbol{\mathcal{G}}} + 2)\epsilon \quad \forall \boldsymbol{\eta} \in B_1(\boldsymbol{\mathcal{G}}), \qquad h < h_0.$$

A homogeneity argument finishes the proof. □

Next, we prove the discrete inf-sup condition for the form $a(\cdot,\cdot)$: Given $\boldsymbol{\eta}_h \in \boldsymbol{\mathcal{G}}_h$, we set

$$\boldsymbol{\varphi}_h := (Id - \mathbf{T}_h SK\mathbf{H}_h)\boldsymbol{\eta}_h \in \boldsymbol{\mathcal{G}}_h.$$

The uniform boundedness with respect to h of the operators involved ensures that there is $C_4 > 0$ independent of $h \in \mathbb{H}$ and $\boldsymbol{\eta}_h$ such that

$$\|\boldsymbol{\varphi}_h\|_{\boldsymbol{\mathcal{G}}} \leq C_4 \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}}. \tag{7.4}$$

We therefore estimate

$$\begin{aligned}
|a(\boldsymbol{\eta}_h, \boldsymbol{\varphi}_h)| &= |a(\boldsymbol{\eta}_h, (Id - \mathbf{T}_h SK\mathbf{H}_h)\boldsymbol{\eta}_h)| \\
&= |a(\boldsymbol{\eta}_h, ((Id - \mathbf{T}_h)(SK\mathbf{H}_h) + (Id - SK\mathbf{H}_h))\boldsymbol{\eta}_h)| \\
&\geq |a(\boldsymbol{\eta}_h, (Id - SK\mathbf{H}_h)\boldsymbol{\eta}_h)| - |a(\boldsymbol{\eta}_h, (Id - \mathbf{T}_h)SK\mathbf{H}_h\boldsymbol{\eta}_h)| \\
&\geq |a(\boldsymbol{\eta}_h, (Id - SK\mathbf{H}_h)\boldsymbol{\eta}_h)| - \tilde{C}_a \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \|(Id - \mathbf{T}_h)SK\mathbf{H}_h\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \\
&\geq |a(\boldsymbol{\eta}_h, (Id - SK\mathbf{H}_h)\boldsymbol{\eta}_h)| - b(h)\tilde{C}_a \|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}},
\end{aligned}$$

the final inequality being a consequence of Lemma 7.1. Further, we estimate the first term,

$$\begin{aligned}
|a(\boldsymbol{\eta}_h, (Id - SK\mathbf{H}_h)\boldsymbol{\eta}_h)| &= |a(\boldsymbol{\eta}_h, ((Id - \mathbf{H}_h) + (Id - SK)\mathbf{H}_h)\boldsymbol{\eta}_h)| \\
&\geq |a(\boldsymbol{\eta}_h, (Id - SK)\mathbf{H}_h\boldsymbol{\eta}_h)| - |a(\boldsymbol{\eta}_h, (Id - \mathbf{H}_h)\boldsymbol{\eta}_h)| \\
&\geq |a(\boldsymbol{\eta}_h, S(S^{-1} - K)\mathbf{H}_h\boldsymbol{\eta}_h)| - \tilde{C}_a \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \|(Id - \mathbf{H}_h)\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \\
&\geq |a(\boldsymbol{\eta}_h, S(S^{-1} - K)\mathbf{H}_h\boldsymbol{\eta}_h)| - \tilde{C}_a C_3 h^t \|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}},
\end{aligned}$$

by Lemma 6.2. Now, we note that $\boldsymbol{\psi} := S(S^{-1} - K)\boldsymbol{\lambda} \in \boldsymbol{\mathcal{G}}$, $\boldsymbol{\lambda} \in \boldsymbol{\mathcal{G}}$, satisfies

$$d(\boldsymbol{\psi}, \boldsymbol{\nu}) = \big\langle (S^{-1} - K)\boldsymbol{\lambda}, \boldsymbol{\nu} \big\rangle = d(\boldsymbol{\lambda}, \boldsymbol{\nu}) - k(\boldsymbol{\lambda}, \boldsymbol{\nu}) = a(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

for all $\boldsymbol{\nu} \in \boldsymbol{\mathcal{G}}$ ($\langle\cdot,\cdot\rangle$ stands for the duality pairing $\boldsymbol{\mathcal{G}} \times \boldsymbol{\mathcal{G}}' \mapsto \mathbb{C}$). In short, $S(S^{-1} - K) = SA$. This enables us to continue the estimates

$$\begin{aligned}
|a(\boldsymbol{\eta}_h, S(S^{-1} - K)\mathbf{H}_h\boldsymbol{\eta}_h)| &= |a(\boldsymbol{\eta}_h - \mathbf{H}_h\boldsymbol{\eta}_h + \mathbf{H}_h\boldsymbol{\eta}_h, SA\mathbf{H}_h\boldsymbol{\eta}_h)| \\
&\geq |a(\mathbf{H}_h\boldsymbol{\eta}_h, SA\mathbf{H}_h\boldsymbol{\eta}_h)| - \tilde{C}_a \|(Id - \mathbf{H}_h)\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \|SA\mathbf{H}_h\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \\
&\geq |d(SA\mathbf{H}_h\boldsymbol{\eta}_h, SA\mathbf{H}_h\boldsymbol{\eta}_h)| - \tilde{C}_a^2 \tilde{c}_d^{-1} C_3 h^t \|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}}.
\end{aligned}$$

For the last time we target the first term

$$\begin{aligned}
|d(SA\mathbf{H}_h\boldsymbol{\eta}_h, SA\mathbf{H}_h\boldsymbol{\eta}_h)| &\geq \tilde{c}_d \|SA\mathbf{H}_h\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}} \geq \tilde{c}_d \tilde{C}_d^{-1} \|A\mathbf{H}_h\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}'} \\
&\geq \tilde{c}_d \tilde{C}_d^{-1} \tilde{c}_a \|\boldsymbol{\eta}_h - (Id - \mathbf{H}_h)\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}} \\
&\geq \tilde{c}_4 (\|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}} - \|(Id - \mathbf{H}_h)\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}}) \\
&\geq \tilde{c}_4 \|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}} - \tilde{c}_4 C_3^2 h^{2t} \|\boldsymbol{\eta}_h\|^2_{\boldsymbol{\mathcal{G}}},
\end{aligned}$$

with $\tilde{c}_4 := \tilde{c}_d \tilde{C}_d^{-1} \tilde{c}_a$. Summing up, we have obtained

$$|a(\boldsymbol{\eta}_h, \boldsymbol{\varphi}_h)| \geq \left(\tilde{c}_4 - (\tilde{c}_4 C_3 h^t + \tilde{C}_a^2 \tilde{c}_d^{-1} + \tilde{C}_a) C_3 h^t - \tilde{C}_a b(h) \right) \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}}^2 .$$

If $h < h_*$ with $(\tilde{c}_4 C_3 h_*^t + \tilde{C}_a^2 \tilde{c}_d^{-1} + \tilde{C}_a) C_3 h_*^t + \tilde{C}_a b(h_*) < \frac{1}{2} \tilde{c}_4$, then we obtain the lower bound

$$|a(\boldsymbol{\eta}_h, \boldsymbol{\varphi}_h)| \geq \frac{1}{2} \tilde{c}_4 \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}}^2 \quad \forall h < h_*.$$

This is valid for any $\boldsymbol{\eta}_h$. Recalling (7.4), an immediate consequence is the discrete inf-sup condition

$$\sup_{\boldsymbol{\varphi}_h \in \boldsymbol{\mathcal{G}}_h} \frac{|a(\boldsymbol{\eta}_h, \boldsymbol{\varphi}_h)|}{\|\boldsymbol{\varphi}_h\|_{\boldsymbol{\mathcal{G}}}} \geq \frac{\tilde{c}_4}{2C_4} \|\boldsymbol{\eta}_h\|_{\boldsymbol{\mathcal{G}}} \quad \forall \boldsymbol{\eta}_h \in \boldsymbol{\mathcal{G}}_h, \qquad h < h_*. \tag{7.5}$$

Based on this discrete stability condition, stability (4.1) of the continuous problem, and the continuity (3.3) of the bilinear forms involved, we obtain quasi-optimal asymptotic convergence for the sequence of Galerkin solutions.

THEOREM 7.2. *There exists a meshwidth $h_* \in \mathbb{H}$ depending only on Γ, ς, and on the shape regularity of the triangulations Γ_h such that for any $h < h_*$ the Galerkin discretization of variational problem* (1.1) *in $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ possesses a unique solution $\mathbf{j}_h$. It converges quasi-optimally according to*

$$\|\mathbf{j} - \mathbf{j}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)} \leq C \inf_{\mathbf{v}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)} \|\mathbf{j} - \mathbf{v}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)}$$

with $C > 0$ independent of $\mathbf{j}$ and h.

Proof. We denote by $\hat{a}(\mathbf{j}, \mathbf{v}) := \langle V_\varsigma \mathrm{div}_\Gamma \mathbf{j}, \mathrm{div}_\Gamma \mathbf{v} \rangle_{\frac{1}{2}, \Gamma} - \varsigma^2 \langle \mathbf{A}_\varsigma \mathbf{j}, \mathbf{v} \rangle_{||, \Gamma}$ the bilinear form in (1.1) and (1.2). Since $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ is $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$-conforming, the $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)$-stable Hodge decompositions $\boldsymbol{\mathcal{RT}}_0(\Gamma_h) = \mathbf{X}_h \oplus \mathbf{N}_h$, $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma) = \mathbf{X} \oplus \mathbf{N}$ and the equivalence of (1.1) with (3.4) and of (1.2) with (7.1) imply that for every $(\mathbf{v}_h^\perp, \mathbf{v}_h^0) \in \boldsymbol{\mathcal{G}}_h$, $(\mathbf{z}_h^\perp, \mathbf{z}_h^0) \in \boldsymbol{\mathcal{G}}_h$, the sums $\mathbf{v}_h := \mathbf{v}_h^\perp + \mathbf{v}_h^0, \mathbf{z}_h := \mathbf{z}_h^\perp + \mathbf{z}_h^0$ belong to $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$. Further, there holds

$$a((\mathbf{z}_h^\perp, \mathbf{z}_h^0), (\mathbf{v}_h^\perp, -\mathbf{v}_h^0)) = \hat{a}(\mathbf{j}, \mathbf{v}_h). \tag{7.6}$$

From Lemma 6.3, we conclude the h-uniform equivalence of norms

$$\left\|(\mathbf{z}_h^\perp, \mathbf{z}_h^0)\right\|_{\boldsymbol{\mathcal{G}}} \asymp \|\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)}, \quad \left\|(\mathbf{v}_h^\perp, \mathbf{v}_h^0)\right\|_{\boldsymbol{\mathcal{G}}} \asymp \|\mathbf{v}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)}$$

with constants independent of h and the functions. Thus, by (7.6), the h-uniform discrete inf-sup condition for $\hat{a}$ on $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$,

$$\sup_{\mathbf{v}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h)} \frac{|\hat{a}(\mathbf{z}_h, \mathbf{v}_h)|}{\|\mathbf{v}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)}} \geq C \|\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma, \Gamma)} \quad \forall \mathbf{z}_h \in \boldsymbol{\mathcal{RT}}_0(\Gamma_h), \tag{7.7}$$

immediately follows from (7.5). Taking the cue from the standard approach of [5], we

resort to Galerkin orthogonality $\hat{a}(\mathbf{j}-\mathbf{j}_h,\mathbf{v}_h)=0$ for all $\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$ and get

$$\begin{aligned}\|\mathbf{j}-\mathbf{j}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} &\le \|\mathbf{j}-\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} + \|\mathbf{z}_h-\mathbf{j}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\\ &\le \|\mathbf{j}-\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} + C\sup_{\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)}\frac{|\hat{a}(\mathbf{z}_h-\mathbf{j}_h,\mathbf{v}_h)|}{\|\mathbf{v}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}}\\ &\le \|\mathbf{j}-\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} + C\sup_{\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)}\frac{|\hat{a}(\mathbf{z}_h-\mathbf{j},\mathbf{v}_h)|}{\|\mathbf{v}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}}\\ &\le (1+C\|\hat{a}\|)\,\|\mathbf{j}-\mathbf{z}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}\end{aligned}$$

for any $\mathbf{z}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$. Here, $\|\hat{a}\|$ stands for the norm of $\hat{a}$ on $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$. Thus, we get the assertion of the theorem. ∎

8. Convergence rates. Quantitative estimates of rates of convergence of $\mathbf{j}_h$ towards $\mathbf{j}$ hinge on extra smoothness of $\mathbf{j}$. Under corresponding assumptions the asymptotic decay of the error of best $\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$-approximations in $\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)$ can be quantified.

LEMMA 8.1. *If $\mathbf{z}\in\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)$, $\sigma>0$, then for any $\epsilon>0$,*

$$\inf_{\mathbf{u}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)}\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \le Ch^{\min\{\frac{3}{2}-\epsilon,\sigma+\frac{1}{2}-\epsilon,1+s^*,\sigma+s^*\}}\,\|\mathbf{z}\|_{\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)}\,,$$

with $C>0$ depending on Γ, s^*, ϵ, and the shape regularity of the meshes Γ_h.

Proof. The proof is based on duality techniques used to deal with the negative norms occurring in the definition of $\|\cdot\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)}$ as follows: Pick any $\mathbf{z}\in\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)$ and $\epsilon>0$. We consider the following mixed variational problem: Seek $\mathbf{u}\in\boldsymbol{H}(\mathrm{div};\Gamma)$, $p\in\boldsymbol{L}^2(\Gamma)/Z$ such that

$$\begin{array}{ccccccc}(\mathbf{u},\mathbf{v})_{0;\Gamma} & + & (\mathrm{div}_\Gamma\mathbf{v},p)_{0;\Gamma} & = & (\mathbf{z},\mathbf{v})_{0;\Gamma} & & \forall\mathbf{v}\in\boldsymbol{H}(\mathrm{div};\Gamma),\\ (\mathrm{div}_\Gamma\mathbf{u},q)_{0;\Gamma} & & & = & (\mathrm{div}_\Gamma\mathbf{z},q)_{0;\Gamma} & & \forall q\in L^2(\Gamma)/Z.\end{array}\tag{8.1}$$

The usual techniques employed for the analysis of mixed variational formulations for second order elliptic boundary value problems [14, sect. IV.1.2] in conjunction with Theorem 2.1 confirm the existence and uniqueness of solutions of this variational problem. Evidently, we have

$$\mathbf{u}=\mathbf{z}\quad\text{and}\quad p=0.$$

Now, we look at a conforming Galerkin discretization of (8.1). Seek $\mathbf{u}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)$, $p_h\in\mathcal{Q}_0(\Gamma_h)/Z$ with

$$\begin{array}{ccccccc}(\mathbf{u}_h,\mathbf{v}_h)_{0;\Gamma} & + & (\mathrm{div}_\Gamma\mathbf{v}_h,p_h)_{0;\Gamma} & = & (\mathbf{z},\mathbf{v}_h)_{0;\Gamma} & & \forall\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h),\\ (\mathrm{div}_\Gamma\mathbf{u}_h,q_h)_{0;\Gamma} & & & = & (\mathrm{div}_\Gamma\mathbf{z},q_h)_{0;\Gamma} & & \forall q_h\in\mathcal{Q}_0(\Gamma_h)/Z.\end{array}$$

For each $q_h\in\mathcal{Q}_0(\Gamma_h)/Z$ we can find a solution $\psi\in H^{1+s^*}(\Gamma)/Z$, s^* from Theorem 2.1, of $\Delta_\Gamma\psi=q_h$ on Γ. By Lemma 5.1, the interpolant $\Pi_h\,\mathbf{grad}_\Gamma\,\psi$ is well defined. The commuting diagram property (5.1) involves

$$\mathrm{div}_\Gamma\Pi_h\,\mathbf{grad}_\Gamma\,\psi=Q_h\mathrm{div}_\Gamma\,\mathbf{grad}_\Gamma\,\psi=Q_hq_h=q_h$$

and continuity means

$$\|\Pi_h \,\mathbf{grad}_\Gamma\,\psi\|_{\boldsymbol{L}^2(\Gamma)} \leq C\left(\|\mathbf{grad}_\Gamma\,\psi\|_{\boldsymbol{H}^{s^*}(\Gamma)} + \|\mathrm{div}_\Gamma\,\mathbf{grad}_\Gamma\,\psi\|_{L^2(\Gamma)}\right) \leq \tilde{C}\,\|q_h\|_{L^2(\Gamma)}\,.$$

We owe the second estimate to Theorem 2.1. Now we are in a position to establish the crucial uniform discrete Ladyshenskaya–Babuška–Brezzi condition [14, sect. II.2.1]

$$\begin{aligned}\sup_{\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)} \frac{(\mathrm{div}_\Gamma\mathbf{v}_h, q_h)_{0;\Gamma}}{\|\mathbf{v}_h\|_{\boldsymbol{H}(\mathrm{div};\Gamma)}} &\geq \frac{(\mathrm{div}_\Gamma\Pi_h\,\mathbf{grad}_\Gamma\,\psi, q_h)_{0;\Gamma}}{\sqrt{\|\Pi_h\,\mathbf{grad}_\Gamma\,\psi\|^2_{\boldsymbol{L}^2(\Gamma)} + \|\mathrm{div}_\Gamma\Pi_h\,\mathbf{grad}_\Gamma\,\psi\|^2_{L^2(\Gamma)}}}\\ &\geq \frac{1}{1+\tilde{C}}\,\|q_h\|_{L^2(\Gamma)}\,.\end{aligned}$$

Along with $\mathrm{div}_\Gamma\boldsymbol{\mathcal{RT}}_0(\Gamma_h) = \mathcal{Q}_0(\Gamma_h)/Z$, this settles the issue of existence, uniqueness, and asymptotic quasi-optimality of discrete solutions $\mathbf{u}_h$, p_h [14, Prop. 2.6]. In particular, plain interpolation error estimates based on affine equivalence techniques and a Bramble–Hilbert-type result (cf. the proof of Lemma 6.2) give us

$$\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}(\mathrm{div};\Gamma)} \leq C \inf_{\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h)} \|\mathbf{z}-\mathbf{v}_h\|_{\boldsymbol{H}(\mathrm{div};\Gamma)} \leq Ch^{\min\{1,\sigma\}}\,\|\mathbf{z}\|_{\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)}\,. \tag{8.2}$$

Next, pick $\mathbf{q}\in\boldsymbol{H}^s(\Gamma)$, $s := \max\{0, \frac{1}{2}-\epsilon\}$, and fix $\varphi\in H^1(\Gamma)/Z$ by

$$-\Delta_\Gamma\varphi = \mathrm{div}_\Gamma\mathbf{q} \in H^{s-1}_*(\Gamma).$$

From Theorem 2.1, we learn that $\varphi\in H^{r+1}(\Gamma)$, $r := \min\{s, s^*\}$, $0 < r < \frac{1}{2}$, and

$$\|\varphi\|_{H^{r+1}(\Gamma)} \leq C\,\|\mathrm{div}_\Gamma\mathbf{q}\|_{H^{s-1}(\Gamma)}\,.$$

Set $\mathbf{w} := \mathbf{q} + \mathbf{grad}_\Gamma\,\varphi \in \boldsymbol{H}^r(\Gamma)$ and observe $\mathrm{div}_\Gamma\mathbf{w} = 0$. Immediately we have the estimate

$$\begin{aligned}\|\mathbf{w}\|_{\boldsymbol{H}^r(\Gamma)} &\leq \|\mathbf{q}\|_{\boldsymbol{H}^r(\Gamma)} + \|\mathbf{grad}_\Gamma\,\varphi\|_{\boldsymbol{H}^r(\Gamma)}\\ &\leq \|\mathbf{q}\|_{\boldsymbol{H}^r(\Gamma)} + C\,\|\mathrm{div}_\Gamma\mathbf{q}\|_{H^{s-1}(\Gamma)} \leq C\,\|\mathbf{q}\|_{\boldsymbol{H}^s(\Gamma)}\,.\end{aligned} \tag{8.3}$$

The ultimate goal is to get information about the error $\mathbf{z}-\mathbf{u}_h$. To that end we note that Galerkin orthogonality implies

$$\begin{array}{rcrcll}(\mathbf{z}-\mathbf{u}_h, \mathbf{v}_h)_{0;\Gamma} & - & (\mathrm{div}_\Gamma\mathbf{v}_h, p_h)_{0;\Gamma} & = & 0 & \forall\mathbf{v}_h\in\boldsymbol{\mathcal{RT}}_0(\Gamma_h),\\ (\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h), q_h)_{0;\Gamma} & & & = & 0 & \forall q_h\in\mathcal{Q}_0(\Gamma_h)/Z.\end{array} \tag{8.4}$$

This has two obvious consequences. First, we see from the usual interpolation estimates for the $L^2(\Gamma)$-orthogonal projections Q_h that for all $\eta\in H^t(\Gamma)$, $0\leq t\leq 1$,

$$\begin{aligned}|\,(\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h), \eta)_{0;\Gamma}\,| &= |\,(\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h), \eta - Q_h\eta)_{0;\Gamma}\,|\\ &\leq Ch^t\,\|\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h)\|_{L^2(\Gamma)}\,\|\eta\|_{H^t(\Gamma)}\,.\end{aligned} \tag{8.5}$$

This immediately yields the duality estimate

$$\|\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h)\|_{H^{-\frac{1}{2}}(\Gamma)} \leq Ch^{\frac{1}{2}}\,\|\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h)\|_{L^2(\Gamma)}\,. \tag{8.6}$$

Second, (8.4) and the error estimates for Π_h lead to

$$\begin{aligned} |\,(\mathbf{z}-\mathbf{u}_h,\mathbf{w})_{0;\Gamma}\,| &= |\,(\mathbf{z}-\mathbf{u}_h,\Pi_h\mathbf{w})_{0;\Gamma}+(\mathbf{z}-\mathbf{u}_h,\mathbf{w}-\Pi_h\mathbf{w})_{0;\Gamma}\,| \\ &= |\,(\mathrm{div}_\Gamma\Pi_h\mathbf{w},p_h)_{0;\Gamma}+(\mathbf{z}-\mathbf{u}_h,\mathbf{w}-\Pi_h\mathbf{w})_{0;\Gamma}\,| \\ &\le Ch^r\,\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{L}^2(\Gamma)}\,\|\mathbf{w}\|_{\boldsymbol{H}^r(\Gamma)}\,. \end{aligned}$$

The last step relied on the commuting diagram property (5.1), which implies $\mathrm{div}_\Gamma\Pi_h\mathbf{w} = Q_h\mathrm{div}_\Gamma\mathbf{w} = 0$. Estimate (8.5)—more precisely, the case $t=1$ with η replaced by φ—and (8.3) can be combined with what we have just obtained. This yields

$$\begin{aligned} |\,(\mathbf{z}-\mathbf{u}_h,\mathbf{q})_{0;\Gamma}\,| &= |\,(\mathbf{z}-\mathbf{u}_h,\mathbf{w}-\mathbf{grad}_\Gamma\,\varphi)_{0;\Gamma}\,| \\ &= |\,(\mathbf{z}-\mathbf{u}_h,\mathbf{w})_{0;\Gamma}+(\mathrm{div}_\Gamma(\mathbf{z}-\mathbf{u}_h),\varphi)_{0;\Gamma}\,| \\ &\le Ch^r\,\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}(\mathrm{div};\Gamma)}\,\|\mathbf{q}\|_{\boldsymbol{H}^s(\Gamma)}\,. \end{aligned}$$

By definition of the dual norm for $0\le s<\frac{1}{2}$, this means

$$\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}^{-s}(\Gamma)} = \sup_{\mathbf{q}\in\boldsymbol{H}^s(\Gamma)}\frac{(\mathbf{z}-\mathbf{u}_h,\mathbf{q})_{0;\Gamma}}{\|\mathbf{q}\|_{\boldsymbol{H}^s(\Gamma)}} \le Ch^r\,\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}(\mathrm{div};\Gamma)}\,.$$

To finish the proof we have only to recall (8.2) and (8.6). ∎

Combined with Theorem 7.2, the preceding lemma instantly translates into an asymptotic convergence estimate for the electric field integral equation discretized by means of lowest order Raviart–Thomas elements.

THEOREM 8.2. *Provided that* $\mathbf{j}\in\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)$, $\sigma>0$, *and* $h<h_*$, *the discretization error encountered when using lowest order Raviart–Thomas surface elements for the Galerkin discretization of the electric field integral equation* (1.1) *behaves like*

$$\|\mathbf{j}-\mathbf{j}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}(\mathrm{div}_\Gamma,\Gamma)} \le Ch^{\min\{\frac{3}{2}-\epsilon,\sigma+\frac{1}{2}-\epsilon,1+s^*,\sigma+s^*\}}\,\|\mathbf{j}\|_{\boldsymbol{H}^\sigma(\mathrm{div}_\Gamma;\Gamma)}\,,$$

where $C>0$ *may depend only on* Γ, ς, $\epsilon>0$, *and the shape regularity of the surface triangulations.*

Remark. From the discussion of s^* after Theorem 2.1, it is clear that $s^*\ge\frac{1}{2}$ if at most three edges of Γ meet at a vertex.

Remark. As $\mathbf{j}$ is the jump of traces of magnetic field solutions of Maxwell's equations on both sides of Γ, the regularity theory of [28] illustrates that the regularity requirements of Theorem 8.2 will always be satisfied. However, $\sigma\ge\frac{1}{2}$ cannot be expected for nonsmooth Γ, since any edge of Γ will appear as a re-entrant when seen from either inside or outside.

Remark. Why is the case $\epsilon=0$ not covered by Lemma 8.1 in general? The reason is that in this case we would have to choose $\mathbf{q}\in\boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$ in order to achieve $\mathrm{div}_\Gamma\mathbf{q}\in H^{-\frac{1}{2}}(\Gamma)$. However, we would need $\mathbf{q}\in\boldsymbol{H}^{\frac{1}{2}}_{||}(\Gamma)$ to gain information about the dual norm $\|\mathbf{z}-\mathbf{u}_h\|_{\boldsymbol{H}^{-\frac{1}{2}}_{||}(\Gamma)}$. This mismatch foils the proof for $\epsilon=0$. Yet, for smooth surface Γ the spaces $\boldsymbol{H}^{\frac{1}{2}}_{||}(\Gamma)$ and $\boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$ coincide, and the best approximation estimate remains true for $\epsilon=0$.

Remark. Lemma 8.1 easily can be extended to kth order Raviart–Thomas boundary elements, $k\ge0$, yielding asymptotic convergence of order $\min\{k+\frac{3}{2}-\epsilon,\frac{1}{2}+\sigma-\epsilon,1+k+s^*,\sigma+s^*\}$.

Remark. In [19] an equivalent mixed formulation of (1.1) is proposed that takes the variational problem to classical Sobolev spaces. In this setting, duality techniques are available that give slightly better rates of asymptotic convergence (not reduced by ϵ), provided that s^* is sufficiently large. In [16], an alternative approach to the error estimates of section 8 is presented which shows that the convergence estimates in Theorem 8.2 are not limited by s^* but only by the regularity of the solution $\mathbf{j}$.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of $H(\mathrm{rot};\Omega)$ and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.

[3] H. Ammari and J.-C. Nédélec, *Coupling of finite and boundary element methods for the time-harmonic Maxwell equations.* II: *A symmetric formulation*, in The Maz'ya Anniversary Collection. Vol. 2, J. Rossmann, ed., Oper. Theory Adv. Appl. 110, Birkhäuser, Basel, 1999, pp. 23–32.

[4] D. Arnold, R. Falk, and R. Winther, *Multigrid in $H(\mathrm{div})$ and $H(\mathbf{curl})$*, Numer. Math., 85 (2000), pp. 175–195.

[5] I. Babuška, *Error bounds for the finite element method*, Numer. Math., 16 (1971), pp. 322–333.

[6] D. Baldomir, *Differential forms and electromagnetism in 3-dimensional Euclidean space $\mathbb{R}^3$*, Proc. IEE-A, 133 (1986), pp. 139–143.

[7] A. Bendali, *Numerical analysis of the exterior boundary value problem for time harmonic Maxwell equations by a boundary finite element method. Part* 1: *The continuous problem*, Math. Comp., 43 (1984), pp. 29–46.

[8] A. Bendali, *Numerical analysis of the exterior boundary value problem for time harmonic Maxwell equations by a boundary finite element method. Part* 2: *The discrete problem*, Math. Comp., 43 (1984), pp. 47–68.

[9] A. Bendali, M. Fares, and J. Gay, *A boundary-element solution of the Leontovitch problem*, IEEE Trans. Antennas and Propagation, 47 (1999), pp. 1597–1605.

[10] D. Boffi, *Discrete compactness and Fortin operator for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[11] D. Boffi, *A note on the discrete compactness property and the de Rham complex*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[12] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[13] A. Bossavit, *On the geometry of electromagentism* IV: *"Maxwell's house,"* J. Japan Soc. Appl. Electromagnetics and Mech., 6 (1998), pp. 318–326.

[14] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer–Verlag, New York, 1991.

[15] A. Buffa, *Hodge decompositions on the boundary of a polyhedron: The multiconnected case*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1491–1504.

[16] A. Buffa and S. Christiansen, *The Electric Field Integral Equation on Lipschitz Screens: Definition and Numerical Approximation*, Tech. Report 1216, Institute of Numerical Analysis, C.N.R. Pavia, Pavia, Italy, July 2001.

[17] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* I: *An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[18] A. Buffa and P. Ciarlet, *On traces for functional spaces related to Maxwell's equations. Part* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.

[19] A. Buffa, M. Costabel, and C. Schwab, *Boundary element methods for Maxwell's equations on non-smooth domains*, Numer. Math., to appear.

[20] A. Buffa, M. Costabel, and D. Sheen, *On Traces for $\mathbf{H}(\mathbf{curl},\Omega)$ in Lipschitz Domains*, Preprint IAN-CNR 1185, IAN, University of Pavia, Pavia, Italy, 2000.

[21] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell equations on Lipschitz domains*, Numer. Math., submitted.

[22] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.
[23] M. Cessenat, *Mathematical Methods in Electromagnetism*, Adv. Math. Appl. Sci. 41, World Scientific, Singapore, 1996.
[24] S. Christiansen, *Discrete Fredholm Properties and Convergence Estimates for the EFIE*, Tech. Report 453, CMAP, Ecole Polytechique, Paris, France, 2000.
[25] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.
[26] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 13, Springer-Verlag, Heidelberg, 1998.
[27] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.
[28] M. Costabel and M. Dauge, *Singularities of Maxwell's equations on polyhedral domains*, in Analysis, Numerics and Applications of Differential and Integral Equations, M. Bach, ed., Pitman Res. Notes Math. Ser. 379, Addison Wesley Longman, Harlow, 1998, pp. 69–76.
[29] M. Costabel and M. Dauge, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.
[30] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 4: *Integral Equations and Numerical Methods,* Springer-Verlag, Berlin, Heidelberg, New York, 1990.
[31] A. de La Bourdonnaye, *Some formulations coupling finite element and integral equation methods for Helmholtz equation and electromagnetism*, Numer. Math., 69 (1995), pp. 257–268.
[32] F. Dubois, *Discrete vector potential representation of a divergence-free vector field in three-dimensional domains: Numerical analysis of a model problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1103–1141.
[33] T. Dupont and R. Scott, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp., 34 (1980), pp. 441–463.
[34] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[35] R. Hiptmair, *Canonical construction of finite elements*, Math. Comp., 68 (1999), pp. 1325–1346.
[36] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.
[37] R. Hiptmair, *Symmetric coupling for eddy current problems*, SIAM J. Numer. Anal., 40 (2002), pp. 41–65.
[38] R. Hiptmair and A. Toselli, *Overlapping and multilevel Schwarz methods for vector valued elliptic problems in three dimensions*, in Parallel Solution of Partial Differential Equations, P. Bjorstad and M. Luskin, eds., IMA Vol. Math. Appl. 120, Springer, Berlin, 1999, pp. 181–202.
[39] A. Kirsch and P. Monk, *A finite element/spectral method for approximating the time-harmonic Maxwell system in* $\mathbb{R}^3$, SIAM J. Appl. Math., 55 (1995), pp. 1324–1344.
[40] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer-Verlag, Berlin, 1989.
[41] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[42] P. Monk, *A finite element method for approximating the time-harmonic Maxwell equations*, Numer. Math., 63 (1992), pp. 243–261.
[43] P. Monk and L. Demkowicz, *Discrete compactness and the approximation of Maxwell's equations in* $\mathbb{R}^3$, Math. Comp., 70 (2001), pp. 507–523.
[44] J. Nédélec, *Mixed finite elements in* R^3, Numer. Math., 35 (1980), pp. 315–341.
[45] P.-A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.
[46] A. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[47] W. Wendland, *Boundary element methods for elliptic problems*, in Mathematical Theory of Finite and Boundary Element Methods, A. Schatz, V. Thomée, and W. Wendland, eds., DMV Seminar 15, Birkhäuser, Basel, 1990, pp. 219–276.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 87–113

ERROR ANALYSIS FOR APPROXIMATION OF STOCHASTIC DIFFERENTIAL EQUATIONS DRIVEN BY POISSON RANDOM MEASURES*

ERIKA HAUSENBLAS†

Abstract. Let X_t be the solution of a stochastic differential equation (SDE) with starting point x_0 driven by a Poisson random measure. Additive functionals are of interest in various applications. Nevertheless they are often unknown and can only be found by simulation on computers. We investigate the quality of the Euler approximation. Our main emphasis is on SDEs driven by an α-stable process, $0 < \alpha < 2$, where we study the approximation of the Monte Carlo error $\mathbb{E}[f(X_T)]$, f belonging to L^∞. Moreover, we treat the case where the time equals $T \wedge \tau$, where τ is the first exit time of some interval.

Key words. stochastic differential equations, Euler scheme, Poisson random measure, α-stable process, Malliavin calculus, first exit time

AMS subject classifications. 60H07, 60H10, 60H30, 65C05

PII. S0036142999360275

1. Introduction.

Let X_t be a real valued process and solution to

$$X_t(x_0) = x_0 + \int_0^t \int \sigma(X_{s^-}, z)(\mu - \gamma)(dz, ds) + \int_0^t b(X_{s^-})ds, \tag{1.1}$$

where μ is a Poisson random measure satisfying certain conditions and γ is its compensator. Assume that $b : \mathbb{R} \mapsto \mathbb{R}$ and $\sigma : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ are Lipschitz continuous in x. Then the stochastic differential equation (SDE) admits a unique solution and the solution is a semimartingale. Cinlar and Jacod [7] have shown that up to a random time change, every Hunt process (right continuous and quasi-left continuous) can be represented as a solution of an SDE driven by a Wiener process, a Lebesgue measure, and a compensated Poisson random measure. Thus, a large class of stochastic processes can be covered by considering SDEs driven by Brownian motion and Poisson random measures.

In contrast to the Brownian case, the Poissonian case is barely investigated. Kurtz and Protter [14] have studied the convergence in law of the normalized error for the path-by-path Euler scheme, and $\mathcal{L}^p$ estimates of the Euler scheme are given by Kohatsu-Higa and Protter [13]. Protter and Talay [16] investigate the weak error $\mathbb{E}[f(X_T)]$, which has to be evaluated at a fixed time T; the diffusion coefficient $\sigma(x, z)$ is of the form $\sigma_0(x)h(z)$; and f, σ, and b are supposed to be four times differentiable. In contrast, we assume f to be only measurable but $\sigma : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$ to be five times differentiable. Additionally, we consider the approximation of $\mathbb{E}[f(X_{T\wedge\tau})]$, where τ denotes the first hitting time of zero ($\wedge$ denotes the minimum).

We proceed as in Bally and Talay [1], in which the Brownian case is treated. Moreover, Bally and Talay [2] give an expansion of the density for the Brownian case

*Received by the editors August 16, 1999; accepted for publication (in revised form) December 5, 2001; published electronically April 12, 2002. This work was partially supported by the grant APART 700 of the Austrian Academy of Science.

http://www.siam.org/journals/sinum/40-1/36027.html

†Department of Mathematics, University of Salzburg, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria (erika.hausenblas@sbg.ac.at).

in terms of $\frac{1}{n}$. We do not treat this case, but we conjecture that the same procedure can be applied to treat the Poissonian case to get an expansion.

Let X_t^n be the approximation of X_t by the Euler scheme with step size $\frac{1}{n}$ defined by

$$\left\{ \begin{array}{rcl} X_0^n & = & x_0, \\ X_t^n & = & X_{[t]_n}^n + b(X_{[t]_n}^n)(t-[t]_n) + \int_{[t]_n}^t \int \sigma(X_{[t]_n}^n, z)(\mu-\gamma)(ds,dz), \end{array} \right.$$

where $[t]_n = [tn]/n$. The entity $\mathbb{E}[f(X_T)]$ will be approximated by a finite sum over a large number N of independent trajectories, i.e.,

$$\mathbb{E}[f(X_T)] \quad \approx \quad \frac{1}{N}\sum_{i=1}^N f(X_T^n(\omega_i)).$$

The resulting error $e(n,N)$ depends on the sample size N and on the step size $\frac{1}{n}$, i.e.,

$$e(n,N) \leq \left| \frac{1}{N}\sum_{i=1}^N f(X_T^n(\omega_i)) - \mathbb{E}[f(X_T^n)] \right| + \left| \mathbb{E}[f(X_T^n)] - \mathbb{E}[f(X_T)] \right| \ = \mathrm{I} + \mathrm{II}.$$

If the driving process has finite variance, an upper bound for (I) can be found by the central limit theorem or deviation results. The main result of our paper is an error bound for the entity (II) under appropriate hypotheses for σ and b.

We suppose that μ is a random measure generated by a Poisson point process whose characteristic measure is Lebesgue, γ its compensator. Let X_t be a solution of (1.1).

Definition 1.1 (Bass and Cranston [3, p. 513]). *Let us call $\sigma(x,z)$ quasi-stable of order k between the indices α^- and α^+ if there exist $0 \leq z_0 < \infty$ and $0 < c_1, c_2 < \infty$ such that*

$$c_1|z|^{-\frac{1}{\alpha^-}-i} \leq |\partial_z^i \sigma(x,z)| \leq c_2|z|^{-\frac{1}{\alpha^+}-i} \tag{1.2}$$

for $i = 0, \ldots, k$, $|z| > z_0$, and all x.

Theorem 1.2. *Let X_t be a solution of the SDE (1.1), where $\sigma(x,z)$ is quasi-stable of order five between the indices α^- and α^+, $0 < \alpha^- \leq \alpha^+ < 2$, such that $\sigma_z(x,z) \geq 0$. Moreover, assume that there exist constants $1 \leq M < \infty$ and $1 \leq m_b \ll M$ such that σ and b satisfy each of the following hypotheses:*

- (H0) *For j, $0 < j \leq 5$, $x \in \mathbb{R}$, and $i = 1, \ldots, 5-j$, either $\partial_x^j \sigma(x,z) = 0$ or estimate (1.2) holds for $\partial_x^j \sigma(x,z)$.*
- (H1) *For all x, z the quantities $|\partial_z^i \partial_x^j \sigma(x,z)|$ are bounded uniformly by M in z and x for all i and j, $i + j \leq 5$, $j \neq 0$.*
- (H2) *$\sup_x |\partial_x^i b(x)|$, $i = 1, \ldots, 5$, is bounded by m_b and $\sup_x |\partial_x^i \sigma(x,z)|$, $i = 1, \ldots, 5$, is bounded by $h_\sigma(z)$ such that $|h_\sigma|_p \leq M$ for all $p \geq 2$.*
- (H3) *Let $\bar{z} = \sup_z\{|\sigma_x(x,z)| > \frac{1}{4} \text{ for all } x\}$. The functions $\left(\partial_x + \frac{\sigma_x(x,z)}{\sigma_z(x,z)}\partial_z\right)^i \sigma_x(x,z)$, $i = 1,2,3,4$, are then uniformly bounded by M in x for all $|z| < \bar{z}$ with the convention that $0/0 = 0$.*

If X_t is approximated by the Euler scheme, i.e., by X_t^n, then we have for $f \in L^\infty$

$$|E[f(X_T)] - E[f(X_T^n)]| \leq C(T) \cdot \frac{1}{n} \cdot M^{21} \cdot \left(1 + \exp(M^{16})\right).$$

REMARK 1.1. *If $\sigma(x,z)$ is quasi-stable of order six, and the indices five and four, respectively, are replaced by six and five, respectively, in* (H0), (H1), (H2), *and* (H3), *then for a Dirac-function $f=\delta$, we have*

$$|E[\delta(X_T)]-E[\delta(X_T^n)]| \le C\cdot\frac{1}{n}\cdot M^{28}\cdot\left(1+\exp(M^{16})\right).$$

If $\sigma(x,z)$ is quasi-stable of order i, $i=2,3,4$, and the indices five and four, respectively, are replaced by i and $i-1$, respectively, in (H0), (H1), (H2), *and* (H3), *then we have for $f\in C_b^{5-i}(\mathbb{R})$*

$$|E[f(X_T)]-E[f(X_T^n)]| \le C\cdot|f^{(5-i)}|_\infty\cdot\frac{1}{n}\cdot M^{c_i}\cdot\left(\exp(M^{2^{5-i}})+\exp(M^{16})\right),$$

and if $2/\alpha^+ > (i-1)/\alpha^- - 1$,

$$|E[f(X_T)]-E[f(X_T^n)]| \le C\cdot|f^{(5-i)}|_\infty\cdot\frac{1}{n}\cdot M^{c_i}\cdot\exp(M^{2^{5-i}}),$$

where $c_2=6$ and $c_i=c_{i-1}+i+1$ for $i>2$.

If the driving process has infinite variance, e.g., if it is an α-stable process, the case is more complicated. (I) is given by the percentiles of an α-stable variable, and to handle (II), we first truncate the driving process, i.e., we throw away all jumps larger than an integer M, and then we apply the result stated above to get the following error bounds:

- Let $f\in C_b^i(\mathbb{R})$, $i=2,3,4$. Then we have for any integer $M\ge 1$

$$|(II)| \le C_1(T)\cdot\frac{1}{n}\cdot M^{c_i}\left(1+\exp(M^{2^{5-i}})\right)+C_2\cdot\left(1-\exp(TM^{-\alpha})\right),$$

  where c_i is defined by $c_2=6$, $c_i=c_{i-1}+i+1$ for $i>2$, and $\alpha>i-3$.
- $f\in L^\infty$, and $\lim_{x\to\infty}\frac{|\sigma(x,z)|}{x^p} < C|z|^{-\frac{1}{\alpha}}$, where $C<\infty$, for all $p\ge 1$: the error bound of (II) is given by $|(II)|\le C\,\frac{1}{n}$.

C_1, C_2, and C are constants. Let $\tau=\inf_{t\ge0}\{X_t=0\}$ and $\tau^n=\inf_{t\ge0}\{X_t^n=0\}$ and $M=1$. Finally we show for $2/\alpha^+ > (i-1)/\alpha^- - 1$ and $f\in C_b^{5-i}(\mathbb{R})$, $i=2,3,4$,

$$|E[f(X_{T\wedge\tau})]-E[f(X^n_{T\wedge\tau^n})]| \le C_1(T)1/\sqrt{n}+C_2(T)|f^{(5-i)}|_\infty\cdot\frac{1}{n}M^{c_i}\exp(M^{2^{5-i}}).$$

The paper is organized as follows: In the second section, we give some preliminaries on point processes and Malliavin calculus. The third section is concerned with the main result, i.e., the error bound for $f\in L^\infty$. After that we give some additional remarks and consider the remaining cases.

2. Preliminaries. In this section we recall some basic facts about point processes, α-stable processes, and the Malliavin calculus. For details on point processes, α-stable processes, and semimartingales, see Bertoin [4] and Cinlar et al. [8].

Let $\mathcal{F}_t$ be a filtration satisfying the usual conditions. Let $\mathcal{Z}$ be a measurable space. A point process with state space $\mathcal{Z}$ is a countable collection of adapted random variables $(Z_i,T_i)\in\mathcal{Z}\times\mathbb{R}^+$. Given a point process, one usually works with the associated random measure μ defined by $\mu(A\times[0,t])(\omega)=\sum_{T_i\le t}1_A(Z_i(\omega))$. A random measure μ has a random measure γ as compensator if γ is predictable and $\mu(A\times[0,t])-\gamma(A\times[0,t])$ is a local martingale in t for all Borel sets A such that $\mathbb{E}[\gamma(A\times[0,t])]<\infty$ for all $t>0$.

A point process is a Poisson point process with characteristic measure ν on $\mathcal{Z}$ if for each Borel set A with $\nu(A) < \infty$ and for each t the counting measure of the set $A \times [0,t]$ is Poisson with parameter $\nu(A)\,t$. It follows that μ has independent increments and that $\mu(A \times [0,t])$ and $\mu(B \times [0,t])$ are independent for $A \cap B = \emptyset$. If the compensator $\gamma(dz,ds)$ of the random measure μ is of the form $\gamma(dz,ds) = \nu(dz)ds$ for some σ-finite measure ν on $\mathcal{Z}$, then the process $t \mapsto \mu(A \times [0,t])$ is called a Poisson point process with characteristic measure ν mentioned above, and the law of the point process is uniquely determined.

In analogy to the Wiener case, we can define a stochastic integral with respect to a Poisson measure. Let $h(s,z,\omega)$ be a simple, càdlàg, and predictable process; i.e., $h(s,z,\omega) = \sum_{i=1}^{n} 1_{(t_{i-1},t_i]}(s) 1_{A_i}(z) H_i(\omega)$, where $0 = t_1 < \cdots < t_n < \infty$ is a finite sequence of stopping times, H_i is bounded and adapted to $\mathcal{F}_{t_i}$, and $\nu(A_i) < \infty$ for all $i = 1, \ldots, n$. The stochastic integral is defined by the Stieltjes integral

$$\int_0^t \int_{\mathcal{Z}} h(s,z,\omega)(\mu-\gamma)(dz,ds) = \sum_{i=1}^{n} H_i(\omega)(\mu-\gamma)(A_i \times (t_{i-1} \wedge t, t_i \wedge t]).$$

The above definition can be extended by $\mathcal{L}^2$ limits to $\mathcal{M}^2 = \{h : h \text{ predictable}, \mathbb{E}[\int_0^t \int_{\mathcal{Z}} h^2(s,z,\omega)\gamma(dz,ds)] < \infty\}$. In what follows we omit ω and $\mathcal{Z}$ for simplicity if there is no danger of confusion.

A purely discontinuous martingale is one where $\mathbb{E}[M_t^2] - \mathbb{E}[M_0^2] = \mathbb{E}[\sum_{s \le t} \Delta M_s^2]$ with $\Delta M_t = M_t - M_{t^-}$. In this case, let $[M,M]_t = \sum_{s \le t} \Delta M_s^2$. One can show that $M_t = \int_0^t \int h(s,z)(\mu-\gamma)(dz,ds)$ is a purely discontinuous (local) martingale with $[M,M]_t = \int_0^t \int h^2(s,z)\mu(dz,ds)$ for $h \in \mathcal{M}^2$. In particular,

$$\mathbb{E}[M_t^2] = \mathbb{E}\left[\int_0^t \int h^2(s,z)\mu(dz,ds)\right] = \mathbb{E}\left[\int_0^t \int h^2(s,z)\gamma(dz,ds)\right].$$

We will suppose throughout the remainder of this paper that $\mathcal{Z} = \mathbb{R} \setminus \{0\}$, and μ is a random measure generated by a Poisson point process whose characteristic measure is Lebesgue, denoted by λ, and γ is its compensator.

REMARK 2.1. *Suppose that $Z_t = \int_0^t \int h(z)\mu(dz,ds)$ is of finite variation on compacts, i.e., the Lévy measure defined by $\nu_h([y,\infty)) = \lambda\{h(z) > y\}$ for $y > 0$ and $\nu_h(x,(-\infty,y]) = \lambda\{h(z) < y\}$ for $y < 0$ satisfies $\int (1 \wedge |x|)\nu_h(dx) < \infty$. It follows (see, e.g., Bertoin* [4, *Proposition* III.8]*) that $t^{-1}\,Z_t$ tends to a constant a.s. as t tends to zero. By Billingsley* [6, *p.* 25] *we know that $t^{-1}\|Z_t\|_p$ is uniformly bounded as t tends to zero for all Z_t, where $\|Z_t\|_p < \infty$.*

REMARK 2.2. *Let $Z_t = \int_0^t \int h(z)\mu(dz,ds)$ with Lévy measure ν_h (cf. Remark* 2.1*). If $\overline{\{z|h(z) \neq 0\}}$ is compact in $\mathcal{Z}$, ν_h has finite total mass and Z_t is a compound Poisson process. Let us define $Y_t = \int_0^t \int \sqrt{|h(z)|}\,\mu(dz,ds)$. It follows that Y_t is also a compound Poisson process and we have a.s. $Z_t \le \sum_{s \le t} |\Delta Z_s| = \sum_{s \le t} \Delta Y_s^2 = [Y,Y]_t$. Since Y_t has finite variation on compacts, $t^{-p}\|Y_t\|_p^p$ is bounded for $t \to 0$ (see Remark* 2.1*). On the other hand, we have for $p \ge 1$ (see Barlow, Jacka, and Yor* [11]*) $C(p)\,t^{-2p}\|Y_t\|_{2p}^{2p} \ge t^{-2p}\|[Y,Y]_t\|_p^p \ge t^{-2p}\|Z_t\|_p^p$ a.s. and therefore $t^{-2}\|Z_t\|_p$ is bounded as t tends to zero. By iterating we see that $t^{-n}\|Z_t\|_p$ is bounded as $t \to 0$ for $n = 2^k$ $k \in \mathbb{N}$.*

REMARK 2.3. *Suppose that $Z_t = \int_0^t \int h(z)\mu(dz,ds)$ has finite Lévy measure with $\nu_h(\mathcal{Z}) = C_h$ and $\sup_z |h(z)| = h_{max} < \infty$. We are interested in an upper bound for $t^{-1}\|Z_t\|_p$. Note that $Z_t \le h_{max}R$, where $R = \int_0^t \int \mu(dz,ds)$ is an exponential*

distributed random variable with parameter $t\,\nu_h(A)$ *and moments* $(t\,\nu_h(A))^k$, $k \in \mathbb{N}$. *It follows that*

$$t^{-p}\|Z_t\|_p^p \le h_{max}^p \mathbb{E}\left[\frac{R^p}{t^p}\right] = t^{-p}\, h_{max}^p \nu_h(A)^p,$$

and therefore $\sup_{0<t\le T} t^{-1}\|Z_t\|_p \le h_{max}\nu_h(A)$.

2.1. Stable processes. In Definition 1.1 we introduced the notion of quasi-stable processes to the class of processes which are the solution to $X_t = \int_0^t \int \sigma(X_{s^-}, z)$ $(\mu-\gamma)(dz, ds)$ for some $\sigma : \mathbb{R}\times\mathcal{Z}\to\mathbb{R}$. Assume that X_t is well defined and consider the process $Z_t = \int_0^t \int_{\mathcal{Z}} \sigma^Z(X_{s^-}, z)\mu(dz, ds)$, where $|z|^{-\frac{1}{\alpha^-}} \le |\sigma^Z(x, z)| \le |z|^{-\frac{1}{\alpha^+}}$ for $0 < \alpha^- \le \alpha^+ < 2$ and for all z, where $|z| \ge z_0$ for some $0 < z_0 < \infty$. Obviously, the driving process $\sigma^Z(X_{s^-}, z)$ behaves like a quasi-stable process of order zero with indices α^- and α^+ but does not necessarily belong to the class of processes for which Definition 1.1 is given. Thus, in order to also consider these cases, we extend the notation of quasi-stable processes to a more general class of processes.

Definition 2.1. *Let* $h : \mathbb{R}^+\times\mathcal{Z}\times\Omega\to\mathbb{R}$ *be adapted. We call* $h(t, z, \omega)$ *quasi-stable of order* k *between the indices* α^- *and* α^+ *if there exist* $z_0 < \infty$ *and* $c_1, c_2 < \infty$ *such that*

$$c_1|z|^{-\frac{1}{\alpha^-}-k} \le |\partial_z^k h(t, z, \omega)| \le c_2|z|^{-\frac{1}{\alpha^+}-k}$$

for $|z| > z_0$ *uniformly for all* $\omega\in\Omega$ *and* $t\in\mathbb{R}^+$.

Remark 2.4. *In Definition* 1.1 *we introduced quasi-stable processes for the solution to* $X_t = \int_0^t \int \sigma(X_{s^-}, z)\,(\mu-\gamma)(dz, ds)$. *Now* $\sigma(X_{s^-}, z)$ *can also be written as* $h(s, z, \omega) = \sigma(X_{s^-}(\omega), z)$, *i.e.,* $X_t = \int_0^t \int h(s, z, \omega)(\mu-\gamma)(dz, ds)$. *The above* $h(s, z, x)$ *can be regarded as quasi-stable in the sense of both Definition* 1.1 *and Definition* 2.1.

Remark 2.5. *In contrast to Bass and Cranston* [3] *we include the* k*th derivative. The number* k *depends on the highest order of the Malliavin derivative involved.*

Since the function $h : \mathbb{R}\to\mathbb{R} : z\mapsto |z|^{-1/\alpha}$ does not belong to $L^2(\mathcal{Z})$ for $0 < \alpha < 2$, we define the truncated α-stable process Z_t^m by throwing away all jumps larger than m, i.e., for $0 < \alpha < 2$, $Z_t^m = \int_0^t \int (|z|^{-1/\alpha}\wedge m)(\mu-\gamma)(dz, ds)$. Furthermore, if $0 < \alpha < 1$, we can also define the truncated α-stable subordinator by

$$Z_t^m = \int_0^t \int (|z|^{-1/\alpha}\wedge m)\mu(dz, ds). \tag{2.1}$$

Proposition 2.1. *Let* $Z_t = \int_0^t \int h(s, z, \omega)\mu(dz, ds)$, *where* h *is bounded and belongs to* $\mathcal{M}^2$ *such that* $h(s, z, \omega)$ *is a quasi-stable process of order zero between the indices* α^- *and* α^+ *with* $0 < \alpha^- \le \alpha^+ < 1$.

(i) *Define* $\sigma^+ := \sup_{(s,z,\omega)} |z|^{\frac{1}{\alpha^+}}\,|h(s, z, \omega)|$, *and*

$$m = (\sigma^+)^{-1} \sup_{(s,z,\omega)} \left(1\vee|z|^{\frac{1}{\alpha}}\right)|h(s, z, \omega)|,$$

where $\vee$ *denotes the maximum. Let* $Z_t^{m,+}$ *be defined by* (2.1), *where* α *is replaced by* α^+. *Then* $\|Z_t\|_p \le \sigma^+\|Z_t^{m,+}\|_p$ *for* $p\ge 1$ *and* $t\in\mathbb{R}^+$.

(ii) *Assume that* $h(s, z, \omega)$ *is positive a.s. and define* $\sigma^- = \inf_{(s,z,\omega)} |z|^{\frac{1}{\alpha^-}}\, h(s, z, \omega)$. *Then* $\sigma^-\|Z_t^{m,-}\|_p \le \|Z_t\|_p$ *for* $p\ge 1$ *and* $t\in\mathbb{R}^+$, *where* $Z_t^{m,-} = \int_0^t \int 1_{(-\infty,-1]\cup[1,\infty)}(z)\; |z|^{-\frac{1}{\alpha^-}}\; \mu(dz)$.

Proof. For (i) note that from $|h(s,z,\omega)| \leq \sigma^+(m \wedge |z|^{-\frac{1}{\alpha^+}})$ it follows that $|\Delta Z_s| \leq \Delta Z_s^{m,+}$ and therefore

$$|Z_t| = \Big|\sum_{s\leq t} \Delta Z_s\Big| \leq \sum_{s\leq t} |\Delta Z_s| \leq \sigma^+ \sum_{s\leq t} \Delta Z_s^{m,+} \leq \sigma^+ Z_t^{m,+} \quad a.s.$$

Thus, we obtain $\|Z_t\|_p \leq \sigma^+ \|Z_t^{m,+}\|_p$. Part (ii) can be proved in an analogous way because $h(s,z,\omega) \geq \sigma^- |z|^{-\frac{1}{\alpha^-}}$. □

PROPOSITION 2.2. *Let $Z_t^m = \int_0^t \int_{\mathcal{Z}} (|z|^{-\frac{1}{\alpha}} \wedge m)\, \mu(dz,ds)$, where $0 < \alpha < 1$ and $0 < t \leq 1$, and let $V_t = \int_0^t \int 1_{(-\infty,-1]\cup[1,\infty)}(z)\, |z|^{-\frac{1}{\alpha}} \mu(dz,ds)$. Then the following conditions are satisfied:*

1. *for $\beta > \alpha$, $t^{-\frac{1}{\beta}} \|Z_t^m\| \leq c(p)\,(m + \|Z_1^m\|_p)$, and*
2. *for $\beta \leq \alpha$, $V_1 \leq t^{-\frac{1}{\beta}} Z_t^m$.*

Proof. To show (i), we decompose Z_t into two processes, i.e.,

$$Z_t^m = \sum_{s\leq t} 1_{\{\Delta Z_s^m < t^{\frac{1}{\beta}}\}} \Delta Z_s^m + \sum_{s\leq t} 1_{\{\Delta Z_s^m \geq t^{\frac{1}{\beta}}\}} \Delta Z_s^m =: \tilde{V}_t^t + \tilde{K}_t^t,$$

where $\tilde{V}_s^t$ and $\tilde{K}_s^t$ are defined by $\tilde{V}_s^t = \int_0^s \int_{\mathcal{Z}} 1_{(-\infty,-t^{-\frac{\alpha}{\beta}})\cup(t^{-\frac{\alpha}{\beta}},\infty)}(z)\, |z|^{-\frac{1}{\alpha}}\, \mu(dz,dr)$ and $\tilde{K}_s^t = \int_0^s \int_{\mathcal{Z}} 1_{[-t^{-\frac{\alpha}{\beta}}, t^{-\frac{\alpha}{\beta}}]}(z)\, (m \wedge |z|^{-\frac{1}{\alpha}})\, \mu(dz,dr)$. Note that since $\mathcal{Z} = \mathbb{R} \setminus \{0\}$, the process $\tilde{K}_s^t$ is well defined. A short calculation shows that the Laplace transform $\Psi(\lambda)$ of $\tilde{V}_t^t / t^{\frac{1}{\beta}}$ is $\Psi(\lambda) = \exp(t \int_0^{t^{\frac{1}{\beta}}} e^{-\lambda x / t^{\frac{1}{\beta}}}\, \bar{\nu}(dx))$, where

$$\bar{\nu}(x) = \nu(x,\infty) = \begin{cases} x^{-\alpha}, & x < m^{-\alpha}, \\ 0, & m^{-\alpha} \leq x. \end{cases}$$

Substitution yields $\Psi(\lambda) = \exp(t \int_0^1 e^{-\lambda x} \bar{\nu}(t^{\frac{1}{\beta}} dx))$. It is easy to see that $\bar{\nu}(tx) \leq t^{-\alpha} \bar{\nu}(x)$ for $0 < x < m$. Thus it follows $\Psi(\lambda) \leq \exp(t^{1-\frac{\alpha}{\beta}} \int_0^1 e^{-\lambda x} \bar{\nu}(dx))$, and therefore $\tilde{V}_t^t / t^{\frac{1}{\beta}} \leq \tilde{V}_1^1 = V_1 \leq Z_1^m$ a.s. It remains to calculate $t^{-\frac{1}{\beta}} \|\tilde{K}_t^t\|_p$. We take k such that $k - 1 < \frac{1}{\beta} \leq k$. Fix $k_0 = 2^{n_0}$ for some fixed $n_0 \in \mathbb{N}$ such that $k_0(1 - \frac{\alpha}{\beta}) \geq k$. Let

$$K_s^{i,t} = \int_0^s \int_{\mathcal{Z}} 1_{[-t^{-\frac{\alpha}{\beta}}, t^{-\frac{\alpha}{\beta}}]}(z) \left(m^{\frac{1}{2^i}} \wedge |z|^{-\frac{1}{2^i \alpha}}\right) \mu(dz,dr), \quad i = 1, \ldots, n_0.$$

Note that $K_s^{i,t}$ has Lévy measure with total mass $C\, t^{-\frac{\alpha}{\beta}}$. Therefore, $K_s^{i,t}$ is a compound Poisson process for $0 < t \leq 1$ and Remark 2.2 implies, for $p \geq 1$,

$$\left\|\tilde{K}_s^t\right\|_p^p = \left\|\left[K^{1,t}, K^{1,t}\right]_s\right\|_p^p \leq c_1(p) \left\|K_s^{1,t}\right\|_{2p}^{2p}.$$

Iteration yields

$$\begin{aligned} \left\|\tilde{K}_s^t\right\|_p &\leq c_1(p) \left\|K_s^{1,t}\right\|_{2p}^2 \leq \cdots \leq c_{i-1}(p) \left\|K_s^{i-1,t}\right\|_{2^{i-1}p}^{2^{i-1}} \\ &= c_{i-1}(p) \left\|\left[K^{i,t}, K^{i,t}\right]_s\right\|_{2^{i-1}p}^{2^{i-1}} \\ &\leq c_i(p) \left\|K_s^{i,t}\right\|_{2^i p}^{2^i} \leq \cdots \leq c_{n_0}(p) \left\|K_s^{n_0,t}\right\|_{k_0 p}^{k_0}. \end{aligned}$$

By letting $s = t$ and Remark 2.3 we see that $t^{-1}\|K_t^{n_0,t}\|_p \le m^{\frac{1}{k_0}} t^{-\frac{\alpha}{\beta}}$ and therefore

$$t^{-\frac{1}{\beta}}\|\tilde{K}_t^t\|_p \le t^{-k}\|\tilde{K}_t^t\|_p \le c_{n_0}(p)\, t^{-k}\, \|K_s^{n_0,t}\|_{k_0 p}^{k_0} = m\, t^{-k_0\frac{\alpha}{\beta}-k+k_0} \le m,$$

since $-k + k_0 - \frac{\alpha}{\beta}k_0 \ge 0$. Thus $t^{-\frac{1}{\beta}}\|Z_t^m\|_p \le C(p)\,(m + \|Z_1^m\|_p) \le C(p)\, m$. To show (ii), let $\beta \le \alpha$. Proceeding as above, we see that the Laplace transform of $\tilde{V}_t^t/t^{\frac{1}{\beta}}$ is $\Psi(\lambda) = \exp(t^{1-\frac{\alpha}{\beta}} \int_0^1 e^{-\lambda x}\bar{\nu}(dx))$. Because $\tilde{K}_t^t$ has only positive jumps, $V_1 \le \tilde{V}_t^t/t^{\frac{1}{\beta}} \le t^{-\frac{1}{\beta}} Z_t^m$ for $\beta \le \alpha$. □

REMARK 2.6. *Let A_t be defined as in Bass and Cranston* [3, Lemma 6.1]*, i.e., $A_t = \int_0^t \int \exp(-|z|/p_0)\mu(dz, ds)$. Since $|z|^{-1/\alpha} \ge \exp(-|z|/p_0)$ for all $|z| \ge z_0$, z_0 large enough, $CZ_t \ge A_t$ a.s. for a constant $C > 0$, and therefore $Z_t^{-1} \le CA_t^{-1}$ a.s. Because A_t^{-1} is in $\mathcal{L}^{p_0}$, the inverse Z_t^{-1} is also in $\mathcal{L}^{p_0}$. Analyzing the proof of Bass and Cranston, we see for a truncated, quasi-stable subordinator (see* (2.1)*) that $\|Z_t^{m-1}\|_p \le C\,\frac{1}{\alpha}\,\Gamma(\frac{1+p}{\alpha} - 1)$, and therefore we can give a threshold of $\|Z_t^{m-1}\|_p$ independent of m.*

REMARK 2.7. *Let Z_t^m be defined as in* (2.1)*. Combining Proposition* 2.2 *and Remark* 2.6*, we obtain an estimate for the inverse of Z_t^m, i.e., $\|(Z_t^m)^{-1}\|_p \le C\, t^{-\frac{1}{\beta}}$, where $\beta \le \alpha$.*

2.2. The Doléans–Dade exponential. Let us now introduce the stochastic exponential, or Doléans–Dade exponential, and its generalization. Since it is necessary for our computations, we list some properties. The proofs can be found in Protter [15, Chapter II.8].

Let X_t be a semimartingale whose martingale part is purely discontinuous. The stochastic exponential $\mathcal{E}(X)_t$ is defined by

$$\mathcal{E}(X)_t = \exp(X_t - X_0) \prod_{s\le t} [(1 + \Delta X_s)\exp(-\Delta X_s)].$$

Before giving an estimate of $\|\mathcal{E}(X)_t\|_p$, $p \ge 2$, we state the following lemmas.

LEMMA 2.2 (Bass and Cranston [3, Lemma 5.2]). *Let $n \ge 1$ and $p = 2^n$. Suppose that $h(s,z)$ is predictable and $|h(s,z)| \le K_s \bar{h}(z)$, where $\bar{h}$ is a deterministic bounded function that is in $L^2(\nu)$. Suppose that $Z_t = \int_0^t \int h(s,z)(\mu - \gamma)(dz, ds)$ and let $Z_t^* = \sup_{0\le s\le t} |Z_s|$. Then we have $\mathbb{E}[Z_t^{*p}] \le c^*(p, \bar{h}, t) \int_0^t \mathbb{E}[\,|K_s|^p]\, ds$ and $\mathbb{E}[Z_t^p] \le c(p, \bar{h}, t) \int_0^t \mathbb{E}[|K_s|^p]\, ds$.*

REMARK 2.8. *Analyzing the proof of* [3, Lemma 5.2]*, we see that*

$$c^*(p, \bar{h}, t) \le \bar{c}(t) c(2^n) \Big(2^{p-1} c(2^{n-1}) \Big(2^{p-1} \dots \Big(2^{p-1} c(4) \times \Big(c(2)|\bar{h}|_{2^n}^{2^n} + |\bar{h}|_{2^{n-1}}^{2^n}\Big) + |\bar{h}|_{2^{n-2}}^{2^n}\Big) + \cdots\Big) + |\bar{h}|_2^{2^n}\Big),$$

where the constant $c(p)$ arises by the Burkholder–Gundy inequality, $\bar{c}(t) \ge 1$ and increasing in t. Thus, if $|\bar{h}|_q \le \bar{m}$ for all $q \ge 2$, then $c^*(p, \bar{h}, t) \le c(p,t)\,\bar{m}^p$ for a constant $c(p,t)$ depending only on p and t. Further, we have $c(p, \bar{h}, t) \le c(t)\, 2^{np} \sum_{k=1}^n |h|_{2^k}^{2^n}$.

Proof. For the first part see [3, Lemma 5.2]. For the second part, the only difference is that we use the isometry of the stochastic integral instead of the Burkholder inequality. □

COROLLARY 2.3. *Let X_t be a solution of $X_t = x_0 + \int_0^t \int \sigma(X_{s^-}, z)\,(\mu-\gamma)(dz, ds) + \int_0^t b(X_{s^-})\,ds$, where $\sup_x \sigma(x,z)$ is bounded by a bounded deterministic function $\bar{h}(z)$ in $L^2(\nu)$. Furthermore, assume that $b(x)$ is bounded by m_b. Then for $n \geq 1$ and $p = 2^n$,*

$$\|\mathcal{E}(X)_t^*\|_p^p \leq 2^{p-1}\,\exp(x_0 p)\,\exp\left(2^{p-1}\left[c^*(p,\bar{h},t) + m_b^p\right]\,t\right)$$

and

$$\|\mathcal{E}(X)_t\|_p^p \leq 2^{p-1}\,\exp(x_0 p)\,\exp\left(2^{p-1}\left[c(p,\bar{h},t) + m_b^p\right]\,t\right),$$

where the constants $c^*(p,\bar{h},t)$ and $c(p,\bar{h},t)$ coincide with those in Lemma 2.2.

Proof. Note that $\mathbb{E}\left[X_t^{*p}\right]^{\frac{1}{p}}$ is a norm for $p \geq 2$ (see Protter [15, Chapter V.2]). By Lemma 2.2 we have

$$\mathbb{E}\left[\left|\mathcal{E}(X)_t - \exp(x_0) - \int_0^t \mathcal{E}(X)_{s^-} b(X_{s^-})\,ds\right|^{*\,p}\right] \leq c^*(p,\bar{h},t)\int_0^t \mathbb{E}\left[|\mathcal{E}(X)_s|^p\right]\,ds.$$

The triangle inequality yields

$$\mathbb{E}\left[\mathcal{E}(X)_t^{*p}\right]^{\frac{1}{p}} \leq \exp(x_0) + m_b\left(\int_0^t \mathbb{E}\left[\mathcal{E}(X)_s^{*p}\right]\,ds\right)^{\frac{1}{p}} + \left(c^*(p,\bar{h},t)\int_0^t \mathbb{E}\left[\mathcal{E}(X)_s^{*p}\right]\,ds\right)^{\frac{1}{p}}$$

and therefore

$$\mathbb{E}\left[|\mathcal{E}(X)_t|^{*p}\right] \leq 2^{p-1}\left(\exp(px_0) + \left(c^*(p,\bar{h},t) + m_b^p\right)\int_0^t \mathbb{E}\left[\mathcal{E}(X)_s^{*p}\right]\,ds\right).$$

Gronwall's lemma then yields the assertion. The proof of the second inequality is in analogy. □

Let H_t be a càdlàg and adapted semimartingale and suppose that X_t satisfies the assumption of Corollary 2.3. The generalization of the stochastic exponential is the solution of the SDE $Z_t = H_t + \int_0^t Z_{s^-}\,dX_s$, which is explicitly given by $Z_t = \mathcal{E}_H(X)_t = \mathcal{E}(X)_t(H_0 + \int_0^t \mathcal{E}(X)_s^{-1} dH_s)$. Suppose that $H_0 = 0$. As above, we can show that for $n \geq 1$ and $p = 2^n$ we have

$$\left\|Z_t - H_t - \int_0^t b(X_{s^-})Z_{s^-}\,ds\right\|_p^p \leq c(p,\bar{h},t)\int_0^t \|Z_s\|_p^p\,ds.$$

The triangle inequality implies

$$\|\mathcal{E}_H(X)_t\|_p = \|Z_t\|_p \leq \|H_t\|_p + m_b\left(\int_0^t \|Z_s\|_p^p\,ds\right)^{\frac{1}{p}} + \left(c(p,\bar{h},t)\int_0^t \|Z_s\|_p^p\,ds\right)^{\frac{1}{p}}.$$

Therefore we get

$$\|\mathcal{E}_H(X)_t\|_p^p \leq 2^{p-1}\|H_t\|_p^p + 2^{p-1}\left(c(p,\bar{h},t) + m_b^p\right)\int_0^t \|\mathcal{E}_H(X)_s\|_p^p\,ds. \tag{2.2}$$

Proceeding as above, we also obtain

$$\|\mathcal{E}_H(X)_t^*\|_p^p \leq 2^{p-1}\|H_t\|_p^p + 2^{p-1}\left(c^*(p,\bar{h},t) + m_b^p\right)\int_0^t \|\mathcal{E}_H(X)_s^*\|_p^p\,ds. \tag{2.3}$$

If $\|H_t\|_p$ in (2.2) or (2.3) is of polynomial growth in t, i.e., $\|H_t\|_p = O(t^\delta)$ for some $\delta \in \mathbb{R}$, $\delta \geq 0$, we can apply the following modification of Gronwall's lemma to obtain an estimate of $\mathcal{E}_H(X)_t$.

LEMMA 2.4 (modification of Gronwall's lemma). *Suppose that g is a continuous function satisfying $0 \leq g(t) \leq a\,t^\delta + \beta \int_0^t g(s)ds$ for $0 \leq t \leq T$ with $\beta > 0$, $\delta \in \mathbb{R}$, $\delta \geq 0$. Then $g(t) \leq a\,t^\delta \exp(\beta t)$ for $0 \leq t \leq T$.*

Proof. Using Gronwall's lemma and integration by parts, the proof is done by direct calculations. □

Now assume that $\|H_t\|_p = O(t^\delta)$ for some $\delta \in \mathbb{R}$, $\delta \geq 0$. A combination of Lemma 2.4 and (2.2) leads to the estimate

$$\|\mathcal{E}_H(X)_t\|_p^p \leq 2^{p-1}\, t^{\delta p} \exp\left(2^{p-1}\ \left(c(p,\bar{h},t) + m_b^p\right)\ t\right). \tag{2.4}$$

We now investigate the inverse of $\mathcal{E}(X)_t$ and $\mathcal{E}_H(X)_t$, where X_t satisfies the assumption of Corollary 2.3. First we have to make sure that the jump sizes are not too large. For this we additionally assume that there exists a $0 < \rho \leq 1$ such that there are no jumps ΔX_s smaller than $-(1-\rho)$ and larger than one. Since for fixed ω there is only a finite number of s such that $|\Delta X_s| \geq \frac{1}{2}$ on each compact interval, it is sufficient to show that

$$V_t = \prod_{0<s\leq t} \left(1 + \Delta X_s 1_{\{|\Delta X_s| \leq \frac{1}{2}\}}\right) \exp\left(-\Delta X_s 1_{\{|\Delta X_s| \leq \frac{1}{2}\}}\right)$$

converges and is of finite variation. But since $|\ln(1+x) - x|, |\ln(1-x) + x| \leq x^2/2$ for $|x| \leq 1/2$, it follows that both V_t and V_t^{-1} are bounded by $\exp([X,X]_t)$ (see also Bass and Cranston [3, p. 510]). Now, consider

$$U_t = \prod_{0<s\leq t} \left(1 + \Delta X_s 1_{\{|\Delta X_s| > \frac{1}{2}\}}\right) \exp\left(-\Delta X_s 1_{\{|\Delta X_s| > \frac{1}{2}\}}\right).$$

Because $\left(1 + \Delta X_s 1_{\{|\Delta X_s| > \frac{1}{2}\}}\right)^{-1} \leq \rho^{-1}$ and $|x| \leq 2x^2$ for $|x| \geq \frac{1}{2}$, U_t^{-1} is bounded by $\exp(2[X,X]_t)/\rho$. Thus we have $\mathcal{E}(X)_t^{-1} \leq C \cdot \exp(|X|_t^*) \exp(2[X,X]_t)$ a.s. Since $\sup_x \sigma(x,z)$ is bounded by a bounded deterministic function, Lemma 2.1 [3] shows that $\|\mathcal{E}(X)_t^{-1}\|_p$ is finite. Now, let $Z_t = \mathcal{E}_H(X)_t$, where X_t is defined in Corollary 2.3 and the jumps are bounded from below by $-(1-\rho)$, $0 < \rho < 1$, and from above by one. Then we have (see Bass and Cranston [3, p. 510]) $Z_t \geq \mathcal{E}(X)_t (\inf_{s\leq t} \mathcal{E}(X)_s^{-1}) H_t$, and, if H_t is invertible, $Z_t^{-1} \leq \mathcal{E}(X)_t^{-1} \mathcal{E}(X)_t^* (H_t)^{-1}$. Assume furthermore that X_t and σ, respectively, are quasi-stable of order zero between the indices α^+ and α^-, $0 < \alpha^- \leq \alpha^+ < 2$. We then obtain

$$\|\mathcal{E}(X)_t^{-1}\|_p^p \leq C(t, q_1, \bar{h}, m_b)\, \|H_t^{-1}\|_{q_2},$$

where $1/q_1 + 1/q_2 = 1/p$.

2.3. Malliavin calculus. In this section we recall briefly some main features of the Malliavin calculus for Poisson random measures. For details see Bass and Cranston [3] (we use the notation of this article) or the book of Bichteler, Gravereaux, and Jacod [5].

In the Wiener case, the key ingredient in the Malliavin calculus, or calculus of variations, is the introduction of a symmetric linear operator $\mathcal{L}$ defined on a dense subspace of the Hilbert space of $\mathcal{L}^2$ functionals, together with a bilinear form $\Gamma(\cdot,\cdot)$

defined by $\Gamma(X,Y) := \mathcal{L}(XY) - \mathcal{L}XY - X\mathcal{L}Y$, X, Y in the domain of $\mathcal{L}$. These operations satisfy that if X and Y are in the domain of $\mathcal{L}$ and ϕ is a C^1 function, then $\phi \circ X$ is also in the domain of $\mathcal{L}$ and $\mathbb{E}[\Gamma(\phi(X), Y)] = -\mathbb{E}[\phi'(X)\Gamma(X,Y)]$. This formula is also called the "integration by parts setting," and if $\Gamma(X,Y)$ is invertible, we can give an estimate of the quantity $\mathbb{E}[\phi'(X)F]$ for some $\mathcal{L}^2$ functional F. Bass and Cranston [3] transferred this idea to the Poissonian case to prove the existence of the local time for certain Lévy processes. To show the convergence of the Euler scheme we have to show that $\mathbb{E}[\phi^{(IV)}(X)Y]$ is bounded, where X is the solution to (1.1) and $Y \approx X - x_0$. In the following we recall the main results of the Malliavin calculus for quasi-stable processes.

One approach of the Malliavin calculus is due to Bismut and is based on a perturbation argument. In contrast to the Wiener case, infinitely many jumps have to be perturbed simultaneously—this essentially requires that the jump times of μ be left unchanged and that only the jump size be modified. We can do it as follows.

Define $\mathcal{M}_\infty^2 = \{h : h \text{ is predictable, and we have for some bounded deterministic function } H(z) \in \mathcal{L}^2(\mu)\colon\ |h(s,z,\omega)| \le H(z) \text{ for all } s,\ z \text{ a.s.}\}$. Suppose that $l \in \mathcal{M}_\infty^2$ with $|l(s,z)| \le 1$ a.s., and let

$$v(s,z) = \int_0^z l(s,y)dy. \tag{2.5}$$

Now the process X_t will be perturbed in direction l by shifting the random measure μ as $\mu^\epsilon(B \times [0,T]) = \int_0^t \int 1_B(z + \epsilon v(s,z))\mu(dz,ds)$, and

$$L_t = \int_0^t \int l(s,z)(\mu - \gamma)(dz,ds). \tag{2.6}$$

By the Girsanov transformation we can construct a martingale M_t^ϵ such that $\mu^\epsilon(B,[0,t]) - t\nu(B)$ is a local martingale. The martingale M_t^ϵ is given by the stochastic exponential of ϵL_t. The associated probability measure Q_t^ϵ is defined by its Radon–Nikodým derivative.

In the Wiener case, the Malliavin derivative is the Fréchet derivative of the perturbed path with respect to ϵ at $\epsilon = 0$. Analogously, the derivative for a functional G of μ in the direction of l is given in the following definition.

DEFINITION 2.5. *A functional G of μ will be called $\mathcal{L}^p(P)$ smooth with derivative $D_lG(\mu) \in \mathcal{L}^p(P)$ if for every $l \in \mathcal{M}_\infty^2$, $\mathbb{E}[|\epsilon^{-1}[G(\mu^\epsilon) - G(\mu) - \epsilon D_lG(\mu)]|^p] \longrightarrow 0$ as $\varepsilon \to 0$.*

Then the following theorem gives the integration by parts setting.

THEOREM 2.6 (Bass and Cranston [3, Theorem 3.4]). *Suppose that G is an $\mathcal{L}^1(P)$ smooth functional of μ and belongs to $\mathcal{L}^p(P)$ for some $p > 1$. Suppose that $l \in \mathcal{M}_\infty^2$, and L_t is defined by* (2.6). *Then $\mathbb{E}[G(\mu)L_t] = -\mathbb{E}[D_lG(\mu)]$.*

Let $h \in C^2$ with compact support and set $G(\mu) = h(Y)$. Suppose that Y is $\mathcal{L}^1$ smooth. From

$$\begin{aligned}|h(Y(\mu^\varepsilon)) - h(Y) - \epsilon h'(Y)D_lY| \le\ & |h(Y(\mu^\varepsilon)) - h(Y) - h'(Y)(Y(\mu^\epsilon) - Y)| \\ & + \|h'\| \cdot |Y(\mu^\epsilon) - Y - \epsilon D_lY|,\end{aligned}$$

and since $\epsilon^{-1}(Y(\mu^\epsilon) - Y(\mu))$ tends to $D_lY(\mu)$ as ϵ tends to zero in $\mathcal{L}^1$, we can conclude that $h(Y)$ is $\mathcal{L}^1(P)$ smooth with derivative $h'(Y)D_lY$. Moreover, thanks to Theorem 2.6 we know $\mathbb{E}[h'(Y)D_lY] = -\mathbb{E}[h(Y)L]$. Assume that D_lY is strictly positive. Let

F be a functional of μ. Setting $h_0(\mu) = h(Y(\mu))\ F(\mu)\ (D_lY)^{-1}(\mu)$ we obtain

$$\begin{aligned}\mathbb{E}[h_0(\mu)L] &= -\mathbb{E}[D_l h_0(\mu)] \\ &= -\mathbb{E}[h'(Y)F(\mu) + h(Y)D_lF(\mu)(D_lY)^{-1} + h(Y)F(\mu)(D_lY)^{-2}D_l^2Y],\end{aligned}$$

and therefore

$$\mathbb{E}[h'(Y)F(\mu)] = -\mathbb{E}[h(Y)H_Y\,[F(\mu)]], \tag{2.7}$$

where $H_Y\,[F(\mu)] := (D_lY)^{-1}((D_lY)^{-1}D_l^2YF(\mu) + F(\mu)L + D_lF(\mu))$.

Suppose that X_t satisfies the hypothesis of Theorem 5.1 [3] and l belongs to $\mathcal{M}_\infty^2$. Then X_t is $\mathcal{L}^p(P)$ smooth for all $p \geq 1$ and the derivative D_lX_t is a solution to

$$\begin{aligned}D_lX_t = &\int_0^t \int \sigma_x(X_{s^-}, z)D_lX_{s^-}(\mu - \gamma)(dz, ds) + \int_0^t b_x(X_{s^-})D_lX_{s^-}ds \\ &+ \int_0^t \int \sigma_z(X_{s^-}, z)v(s, z)\mu(dz, ds),\end{aligned} \tag{2.8}$$

where v is given by (2.5).

Fix $k_0 \in \mathbb{N}$. Suppose in the next paragraph that $D_l^kX_t$ is well defined for $0 \leq k \leq k_0$ and $D_lX_t^{-1}$ exists. Until now we have investigated the forward variable. But considering $\mathbb{E}^x[f(X_t)] =: \mathbb{E}[f(X_t(x))]$ it is easy to see how the Malliavin calculus can be applied to the study of the backward variable x. We use the method of Gihman and Skorokhod [9, Chapter II.2.8] (see also Bichteler, Gravereaux, and Jacod [5, Proof of Theorem 28, resp., Chapter 4-c]), i.e.,

$$\partial_k^x\mathbb{E}[f(X_t)] = \sum_{i=1}^k \mathbb{E}^x\left[(\partial_{k-i+1}^y f)(y)\big|_{y=X_t}\nabla^i X_t\right],$$

where ∇X_t satisfies

$$\nabla X_t = 1 + \int_0^t \int \sigma_x(X_{s^-}, z)\nabla X_{s^-}(\mu - \gamma)(dz, ds) + \int_0^t b_x(X_{s^-})\nabla X_{s^-}\,ds, \tag{2.9}$$

and $\nabla^i X_t = \nabla^{i-1}\nabla X_t$. Now it follows from (2.7) that

$$\partial_k^x\mathbb{E}^x[f(X_t)] = \sum_{i=1}^k \mathbb{E}^x\left[f(X_t)H_X^{k-i+1}\left[\nabla_l^i X_t\right]\right]. \tag{2.10}$$

Now, assume that X_t is a solution of (1.1), where σ is quasi-stable of order one between indices α^- and α^+ with $\sigma_z(x, z)$ positive. Furthermore, assume that $\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$ satisfy the following hypotheses for some constants M and m_b (compare also Bass and Cranston [3, Theorem 4.4]):

- (H0*) The function $\sigma_x(x, z)$ satisfies inequality (1.2) uniformly in x for $i = 0$ and $|z| > z_0$, $0 \leq z_0 < \infty$.
- (H1*) $\sup_{(x,z)} |\sigma_z(x, z)| \leq M$.
- (H2*) (i) $\sup_x |b_x(x)| \leq m_b$.
  (ii) $\sup_{x,z} |\sigma_x(x, z)| \leq M$ and $\sup_x |\sigma_x(x, \cdot)|_p \leq M$.
- (H3*) Let $z_0(x) = \sup_z\{|\sigma_x(x, z)| > \frac{1}{4}\}$. Then $|\sigma_x(x, z)/\sigma_z(x, z)|$ is bounded by M for all $|z| < z_0(x)$, with the convention that $0/0 = 0$.

Next we have also to specify the direction of the derivative, i.e., the function l. Of course, there are many choices of l, but l should be chosen so that $D_l X_t$ is invertible and $\|D_l X_t\|_p$ is not too large. Our choice is

$$l(s,z) = \frac{\partial}{\partial z}\left(\frac{z^2}{1+|z|^{1+\frac{1}{p_0}}}\right) - \frac{\partial}{\partial z}\left(\frac{\sigma_x(X_{s^-},z)}{\sigma_z(X_{s^-},z)}\varphi\left(\sigma_x(X_{s^-},z)\right)\right) D_l X_{s^-}, \tag{2.11}$$

where $\frac{1}{2} < p_0 < \frac{2}{3}$ and $\varphi \in C^3(\mathbb{R})$ such that $\varphi^{(i)}(x) \leq 40$, $i = 1,2,3,4$, and

$$\varphi(x)\begin{cases} = & 0, & x \in \left(-\frac{1}{4}, \frac{3}{4}\right), \\ \in & (0,1), & x \in \left(-\frac{3}{4}, -\frac{1}{4}\right) \cup \left(\frac{3}{4}, 1\right), \\ = & 1, & x \in \left(-\infty, -\frac{3}{4},\right) \cup (1, \infty). \end{cases}$$

In fact, $l(s,z)$ does not belong to $\mathcal{M}_\infty^2$. But as in Bass and Cranston [3, Proof of Theorem 4.4, p. 509] we can show under the hypothesis (H0*), (H1*), (H2*), and (H3*) and a limit argument that this choice is valid. For simplicity, we write $v(s,z) = v_1(z) + v_2(s,z)$.

Remark 2.9. *Defining $\phi(x,z) := (\varphi \circ \sigma_x)(x,z)$ and taking* (H0*) *into account, we see that the set $\cup_{x\in\mathbb{R}}\{z \in \mathbb{R};\, \phi(x,z) \neq 0\}$ has finite mass. Analogously, defining $\phi'(x,z) := (\varphi' \circ \sigma_x)(x,z)$, the set $\cup_{x\in\mathbb{R}}\{z \in \mathbb{R};\, \phi'(x,z) \neq 0\}$ has finite mass.*

The process $D_l X_t = Y_t$ can also be written as a Doléans–Dade exponential, i.e., $Y_t = \int_0^t Y_{s^-}\, dK_s + H_t$, where K_t and H_t are given by

$$\begin{aligned} K_t &= \int_0^t \int \sigma_x(X_{s^-},z)(\mu-\gamma)(dz,ds) + \int_0^t b_x(X_{s^-})ds \\ &\quad - \int_0^t \int \sigma_x(X_{s^-},z)\varphi(\sigma_x(X_{s^-},z))\mu(dz,ds), \end{aligned} \tag{2.12}$$

$$H_t = \int_0^t \int \sigma_z(X_{s^-},z)v_1(z)\mu(dz,ds). \tag{2.13}$$

Our next objective is to give an upper bound of the derivative $D_l X_t$. For clarity, we set $\alpha = \alpha^+$. Since $\gamma(dz,ds) = \nu(dz)\,ds = dz\,ds$, we can write for K_t

$$\begin{aligned} K_t &= \int_0^t \int \sigma_x(X_{s^-},z)\left[1-\varphi(\sigma_x(X_{s^-},z))\right](\mu-\gamma)(dz,ds) \\ &\quad - \int_0^t \int \sigma_x(X_{s^-},z)\varphi(\sigma_x(X_{s^-},z))\,dz\,ds + \int_0^t b_x(X_{s^-})ds. \end{aligned} \tag{2.14}$$

Let $h_K(z) := \sup_x \sigma_x(x,z)\left[1-\varphi(\sigma_x(x,z))\right]$. Note that $h_K(z)$ is bounded from below by $-\frac{3}{4}$ and from above by one. Moreover, by Remark 2.9 the function $\sigma_x(x,\cdot)\varphi(\sigma_x(x,\cdot))$ has a uniformly bounded support in x, and since $\sup_{x,z}|\sigma_x(x,z)|$ is bounded, the integral $\int \sigma_x(x,z)\varphi(\sigma_x(x,z))\,dz$ is bounded uniformly for all x. Setting $C_a = \sup_x\left|\int \sigma_x(x,z)\ \varphi(\sigma_x(x,z))\,dz\right|$, we obtain by (2.3)

$$\|D_l X_t^*\|_p^p \leq 2^{p-1}\|H_t\|_p^p + 2^{p-1}\left(c^*(p,h_K,t) + (m_b + C_a)^p\right)\int_0^t \|D_l X_s^*\|_p^p\,ds,$$

where $p = 2^n$ for some $n \geq 1$ and H_t is given by (2.13). Let us define

$$\sigma_H^+ = \sup_{x,z}|z|^{-\frac{1}{\alpha}-\frac{1}{p_0}}|\sigma_z(x,z)v_1(z)| \quad \text{and} \quad m_H = {\sigma_H^+}^{-1}\sup_{x,z}\left(|\sigma_z(x,z)v_1(z)| \vee 1\right).$$

Since $\sigma(x,z)$ is quasi-stable of order one, we know that $\sigma_H^+ < \infty$ and σ_H^+ is strictly larger than zero. Therefore and by (H1*), it follows that $m_H < \infty$. Proposition 2.1 leads to $\|H_t\|_p \le \sigma_H^+ \|Z_t^H\|_p$, where

$$Z_t^H = \int_0^t \int \left(m_H \wedge |z|^{-\frac{1}{\alpha}-\frac{1}{p_0}} \right) \mu(dz,ds).$$

By Proposition 2.2 we obtain $t^{-\delta_H}\|Z_t^H\|_p \le C(T)\,\sigma_H^+ m_H \le C(T)\,M$ for $0<t<T$ and $1/p_0 + 1/\alpha > \delta_H$. It follows by a generalization of Gronwall's lemma 2.4 (see also (2.4)) that

$$t^{-\delta_H}\|D_l X_t^*\|_p^p \le C(T)\,M^p\,2^{p-1}\,\exp\left(2^{p-1}\left(c^*(p,h_K,t) + (m_b + C_a)^p\right)\,t\right). \tag{2.15}$$

Assume $\sigma(x,z)$ is a quasi-stable process with indices α^+ and α^- of order five satisfying the hypotheses of Theorem 1.2. Our next objective is to compute the Malliavin derivative $D_l^2 X_t$. The second derivative $D_l^2 X$ satisfies the SDE

$$\begin{aligned}
D_l^2 X_t = &\int_0^t \int \left[\sigma_{xx}(X_{s^-},z) D_l X_{s^-}^2 + \sigma_x(X_{s^-},z) D_l^2 X_{s^-}\right](\mu-\gamma)(dz,ds) \\
&+ \int_0^t \left[b_{xx}(X_{s^-}) D_l X_{s^-}^2 + b_x(X_{s^-}) D_l^2 X_{s^-}\right] ds \\
&- \int_0^t \int \partial_x\left(\sigma_x(X_{s^-},z)\varphi(\sigma_x(X_{s^-},z))\right) D_l X_{s^-}^2\,\mu(dz,ds) \\
&- \int_0^t \int \sigma_x(X_{s^-},z)\varphi(\sigma_x(X_{s^-},z)) D_l^2 X_{s^-}\,\mu(ds,dz) \\
&+ \int_0^t \int \sigma_{xz}(X_{s^-},z) D_l X_{s^-} v_1(z)\mu(dz,ds) \\
&+ \int_0^t \int \Big[\sigma_{xz}(X_{s^-},z) D_l X_{s^-} + \partial_z\left(\sigma_x(X_{s^-},z)\varphi(\sigma_x(X_{s^-},z))\right) D_l X_{s^-} \\
&\qquad + \partial_z\left(\sigma_z(X_{s^-},z) v_1(z)\right)\Big] v(s,z)\,\mu(dz,ds).
\end{aligned}$$

Analogously to $D_l X_t$, we can write $D_l^2 X_t$ as a generalized stochastic exponential with $\bar{H}$ and K defined by (2.14) and

$$\begin{aligned}
\bar{H}_t = &\int_0^t \int \sigma_{xx}(X_{s^-},z) D_l X_{s^-}^2 (\mu-\gamma)(dz,ds) + \int_0^t b_{xx}(X_{s^-}) D_l X_{s^-}^2\,ds \\
&+ \int_0^t \int 2 D_l X_{s^-}\sigma_{xz}(X_{s^-},z) v_1(z)\mu(dz,ds) \\
&+ \int_0^t \int \sigma_z(X_{s^-},z) v_1(z) v_{1\,z}(z) + \sigma_{zz}(X_{s^-},z) v_1^2(z)\,\mu(dz,ds) \\
&+ \int_0^t \int \left\{a_1(X_{s^-},z) D_l X_{s^-}^2 + a_2(X_{s^-},z) D_l X_{s^-}\right\}\mu(dz,ds),
\end{aligned} \tag{2.16}$$

where

$$\begin{aligned}
a_1(x,z) &= \frac{\sigma_{xz}(x,z)}{\sigma_z(x,z)}\sigma_x(x,z)\phi(x,z) - \partial_x\big(\sigma_x(x,z)\phi(x,z)\big) \\
&\qquad + \partial_z\big(\sigma_x(x,z)\phi(x,z)\big)\frac{\sigma_x(x,z)}{\sigma_z(x,z)}\phi(x,z), \\
a_2(x,z) &= \partial_z\big(\sigma_z(x,z)v_1(z)\big)\frac{\sigma_x(x,z)}{\sigma_z(x,z)}\phi(x,z).
\end{aligned}$$

Note that we set $\phi(x,z) := (\varphi \circ \sigma_x)(x,z)$ and $\phi'(x,z) = (\varphi' \circ \sigma_x)(x,z)$.

REMARK 2.10. *Set*

$$A_{1t} = \int_0^t \int a_1(X_{s^-}, z)\, \mu(dz, ds),$$
$$A_{2t} = \int_0^t \int a_2(X_{s^-}, z)\, \mu(dz, ds).$$

By Remark 2.9 *we know the processes* A_{1t} *and* A_{2t} *are compound Poisson processes. Thanks to condition* (H3)*, the quantities* $\sup_{x,z} |a_1(x,z)|$ *and* $\sup_{x,z} |a_2(x,z)|$ *are bounded by* CM^2 *and by* M*. Combining Remarks* 2.2 *and* 2.3 *yields* $t^{-n}\|A_{1t}\|_p \leq \frac{3}{4}\, C(T)\, M^2$ *and* $t^{-n}\|A_{2t}\|_p \leq \frac{3}{4}\, C(T)\, M$ *for* $0 < t \leq T$ *and* $n = 2^k$*,* $k \in \mathbb{N}$.

Analogous to (2.3), we obtain

$$\|D_l^2 X_t^*\|_p^p \leq 2^{p-1}\, \|\bar{H}_t\|_p^p + 2^{p-1}\, (c^*(p, h_K, t) + (m_b + C_a)^p) \int_0^t \|D_l^2 X_s^*\|_p^p\, ds. \tag{2.17}$$

Defining $J_t = \int_0^t \int \sigma_{xz}(X_{s^-}, z) v_1(z)\, \mu(dz, ds)$ and $V_t = \int_0^t \int \sigma_z(X_{s^-}, z) v_1(z) v_{1z}(z) + \sigma_{zz}(X_{s^-}, z) v_1^2(z)\, \mu(dz, ds)$, we can write for $\bar{H}_t$

$$\begin{aligned}\bar{H}_t = &\int_0^t \int \sigma_{xx}(X_{s^-}, z) D_l X_{s^-}^2 (\mu - \gamma)(dz, ds) + \int_0^t b_{xx}(X_{s^-}) D_l X_{s^-}^2\, ds \\ &+ \int_0^t D_l X_{s^-}\, dJ_t + V_t + \int_0^t D_l X_{s^-}^2\, dA_{1s} + \int_0^t D_l X_{s^-}\, dA_{2s}.\end{aligned}$$

Applying Lemma 2.2 to the first summand and Emery's inequality to the third, fifth, and sixth summands, we obtain

$$\begin{aligned}\|\bar{H}_t\|_p^p \leq (c\,(p,t)\; M^p + m_b^p) &\int_0^t \|D_l X_s^2\|_p^p\, ds + \|D_l X_t^*\|_{q_1}^p \|J_t\|_{q_2}^p + \|V_t\|_p^p \\ &+ \|D_l X_t^{*2}\|_{2q_1}^p \|A_{1t}\|_{q_2}^p + \|D_l X_t^*\|_{q_1}^p \|A_{2t}\|_{q_2}^p,\end{aligned}$$

where $1/q_1 + 1/q_2 = 1/p$. Let σ_J^+ and σ_V^+ be two constants given by $\sigma_J^+ = \sup_{(x,z)} |z|^{\frac{1}{\alpha} + \frac{1}{p_0}} |\sigma_{xz}(x,z) v_1(z)|$ and $\sigma_V^+ = \sup_{(x,z)} |z|^{\frac{1}{\alpha} + \frac{2}{p_0}} |\sigma_z(x,z) v_1(z) v_{1z}(z) + \sigma_{zz}(x,z) v_1^2(z)|$, respectively. The suprema exist and are finite, because we have first (see (2.11)) $\sup_z |z|^{\frac{1}{p_0}} v_1(z) \leq C \cdot z$ for some $C > 0$ and second, by condition (H1), $\sup_{(x,z)} |z|^{\frac{1}{\alpha}+1} \sigma_{xz}(x,z) \leq C$. Note that the constant C may vary from line to line. Thus, we have $\sup_{(x,z)} |z|^{\frac{1}{\alpha} + \frac{1}{p_0}} \sigma_{xz}(x,z) v_1(z) \leq C < \infty$, and therefore σ_J^+ exists and is finite. To show the existence of σ_V^+, we have to take into account that $\sigma(x,z)$ is quasi-stable at least of order two. Let

$$m_J = \big(\sigma_J^+\big)^{-1} \sup_{x,z} \big(|\sigma_{xz}(x,z) v_1(z)| \vee 1\big),$$

and

$$m_V = (\sigma_Z^+)^{-1} \sup_{x,z} \big(|\sigma_z(x,z) v_1(z) v_{1z}(z) + \sigma_{zz}(x,z) v_1^2(z)| \vee 1\big).$$

Setting $h_J(s,z,\omega) = \sigma_{xz}(X_{s^-}(\omega), z) v_1(z)$ and

$$h_V(s,z,\omega) = \sigma_z(X_{s^-}(\omega), z) v_1(z) v_{1z}(z) + \sigma_{zz}(X_{s^-}(\omega), z) v_1^2(z),$$

we know

$$\sup_{(s,z,\omega)} |z|^{\frac{1}{\alpha}+\frac{1}{p_0}} h_J(s,z,\omega) \le \sigma_J^+ \quad \text{and} \quad \sup_{(s,z,\omega)} |z|^{\frac{1}{\alpha}+\frac{2}{p_0}} h_V(s,z,\omega) \le \sigma_V^+,$$

respectively, and therefore h_J and h_V are quasi-stable of order zero with above indices $(1/\alpha + 1/p_0)^{-1}$ and $(1/\alpha + 2/p_0)^{-1}$, respectively. Further, let $Z_t^J = \int_0^t \int (m_J \wedge |z|^{-\frac{1}{\alpha}-\frac{1}{p_0}})\mu(dz,ds)$ and $Z_t^V = \int_0^t \int (m_V \wedge |z|^{-\frac{1}{\alpha}-\frac{2}{p_0}})\mu(dz,ds)$. Since $(1/\alpha+1/p_0)^{-1} < 1$ and $(1/\alpha+2/p_0)^{-1} < 1$, we can apply Proposition 2.1 in order to get $\|J_t\|_p \le \sigma_J^+ \|Z_t^J\|_p$ and $\|V_t\|_p \le \sigma_V^+ \|Z_t^V\|_p$, respectively. Proposition 2.2 yields $t^{-\delta_J}\|Z_t^J\|_p \le C(T)\,\sigma_J^+ m_J \le C(T)\,M$ for $1/p_0 + 1/\alpha > \delta_J = \delta_H$ and $t^{-\delta_V}\|Z_t^V\|_p \le C(T)\,\sigma_V^+ m_V \le C(T)\,M$ for $2/p_0 + 1/\alpha > \delta_V$ and $0 < t \le T$, respectively. Thus, we can conclude for $\bar{H}_t$ and $0 < t \le T$ that

$$\begin{aligned}
t^{-\delta_V p}\|\bar{H}_t\|_p^p &\le \left(c(p,t)\,M^p + m_b^p\right)\, t^{-\delta_V p}\, 2^{2p-1} \\
&\quad \times \int_0^t s^{2\delta_H p} \exp\left(2^{2p-1}\left(c(2p,h_K,t) + (m_b+C_a)^{2p}\right)s\right)\, ds \\
&\quad + M^p\, 2^{2p-1} \exp\left(2^{2p-1}\left(c^*(2p,h_K,t) + (m_b+C_a)^{2p}\right)t\right)\, \underbrace{t^{-\delta_J p}\|J_t\|_p^p}_{\le C(T)M^p} + \underbrace{t^{-\delta_V p}\|V_t\|_p^p}_{\le C(T)M^p} \\
&\quad + \|D_l X_t^{*2}\|_{2q_1}^p \underbrace{t^{-\delta_V p}\|A_{1t}\|_{q_2}^p}_{\le C(T)\,M^{2p}} + \|D_l X_t^*\|_{q_1}^p \underbrace{t^{-\delta_V p}\|A_{2t}\|_{q_2}^p}_{\le C(T)\,M^p}.
\end{aligned}$$

Substituting $\bar{H}_t$ in (2.17) and applying Lemma 2.4 we get for $q_1 = q_2 = 2p$

$$\begin{aligned}
t^{-\delta_V p}\|D_l^2 X_t^*\|_p^p &\le C\left(M^{4p} + m_b^p\right) \\
&\quad \times \exp\left(2^{2p-1}\left(c^*(4p,h_K,t) + (m_b+C_a)^{4p}\right)t\right). \qquad (2.18)
\end{aligned}$$

Therefore $t^{-\delta_V}\|D_l^2 X_t^*\|_p$ is uniformly bounded by $C(T)\,M^4$ in t on $(0,T]$. Iterating (2.18) we obtain that $D_l^3 X_t$ equals $\mathcal{E}_{\hat{H}}(K)_t$, where K_t is given by (2.12) and

$$\begin{aligned}
\hat{H}_t &= \int_0^t \int \left[3\sigma_{xx}(X_{s^-},z) D_l X_{s^-} D_l^2 X_{s^-} + \sigma_{xxx}(X_{s^-},z) D_l X_{s^-}^3\right](\mu-\gamma)(dz,ds) \\
&+ \int_0^t \left[3b_{xx}(X_{s^-}) D_l X_{s^-} D_l^2 X_{s^-} + b_{xxx}(X_{s^-}) D_l X_{s^-}^3\right] ds \\
&+ \int_0^t \int \left\{3\sigma_{xz}(X_{s^-},z) v_1(z) D_l^2 X_{s^-} + \cdots + \sigma_z(X_{s^-},z) v_{1z}^2\right\}\mu(dz,ds) \\
&+ \int_0^t \int \left[\hat{a}_1(X_{s^-},z) D_l X_{s^-}^3 + \hat{a}_2(X_{s^-},z) D_l X_{s^-} D_l^2 X_{s^-} + \hat{a}_3(X_{s^-},z) D_l^2 X_{s^-}\right. \\
&\qquad \left. + \hat{a}_4(X_{s^-},z) D_l X_{s^-}^2 + \hat{a}_5(X_{s^-},z) D_l X_{s^-}\right]\mu(dz,ds), \qquad (2.19)
\end{aligned}$$

where $\hat{a}_i(x,z)$ $i = 1,\ldots,5$, have bounded support (see condition (H3)). Observe that $\hat{a}_1(x,z) = \partial_x^2(\sigma_x(x,z)\phi(x,z)) + \cdots \le C\,M^3$, $\hat{a}_2(x,z) \le C\,M^3$, $\hat{a}_3(x,z), \hat{a}_4(x,z) \le C\,M^2$, and $\hat{a}_5(x,z) \le C\,M$. Further we have the identity $D_l^4 X_t = \mathcal{E}_{\tilde{H}}(K)_t$, where K_t is given in (2.12) and

$$\tilde{H}_t = \int_0^t \int \left\{4\sigma_{xx}(X_{s^-},z) D_l^3 X_{s^-} D_l X_{s^-} + \cdots + \sigma_z(X_{s^-},z) v_1(z) v_{1z}^3(z)\right\}\mu(dz,ds)$$

$$+\int_0^t\int\Big[\tilde a_1(X_{s^-},z)D_lX_{s^-}^4+\tilde a_2(X_{s^-},z)D_lX_{s^-}^3D_lX_{s^-}\\ +\cdots+\hat a_9(X_{s^-},z)D_lX_{s^-}\Big]\mu(dz,ds).$$

Next, we investigate in the norm of $D_l^3X_t$. Defining $\hat A_{it}=\int_0^t\int\hat a_1(X_{s^-},z)\mu(dz,ds)$, we know

$$\begin{aligned}\|\hat H_t\|_p^p\le C\ (M^p+m_b^p)\ \Big[\|D_lX_t^3\|_p^p+\|D_lX_t^2\|_{2p}^p\|D_lX_t\|_{2p}^p\Big]\,ds\\ +\|D_lX_t^{2*}\|_{2p}^p\ \|J_t^1\|_{2p}^p\|D_l^2X_t^*\|_{2p}^p\ \|J_t^2\|_{2p}^p\\ +\|D_lX_t{}^*\|_{2p}^p\ \|V_t\|_{2p}^p+\|W_t\|_p^p+\|D_lX_t^{3*}\|_{2p}^p\|\hat A_{1t}\|_{2p}^p\\ +\|D_l^2X_t{}^*\|_{3p}^p\|D_lX_t{}^*\|_{3p}^p\ \|\hat A_{2t}\|_{3p}^p+\cdots+\|D_lX_t{}^*\|_{2p}^p\ \|\hat A_{9t}\|_{2p}^p,\end{aligned}\tag{2.20}$$

where $J_t^1=\int_0^t\int 2\sigma_{xxz}(X_{s^-},z)v_1(z)\mu(dz,ds)$, $J_t^2=\int_0^t\int 3\sigma_{xz}(X_{s^-},z)v_1(z)\mu(dz,ds)$, $V_t=\int_0^t\int 3\big(\sigma_{xzz}(X_{s^-},z)v_1(z)+\sigma_{xz}(X_{s^-},z)v_{1z}(z)\big)v_1(z)\mu(dz,ds)$, and

$$\begin{aligned}W_t&=\int_0^t\int\partial_z\big(\partial_z\big(\sigma_z(X_{s^-},z)v_1(z)\big)v_1(z)\big)v_1(z)\mu(dz,ds)\\ &=:\int_0^t\int h_W(s,z,\omega)\mu(dz,ds).\end{aligned}\tag{2.21}$$

By the same arguments as above and Proposition 2.1, we can conclude that $\|J_t^i\|_p\le C_i(T)\,\|Z_t^{m_i}\|_p$ for $i=1,2$, for $0<t\le T$, and for some C_i and m_i such that $C_im_i\le M$. Applying Proposition 2.2 yields that $t^{-\delta_H}\|J_t^i\|_p$ and $t^{-\delta_V}\|V_t\|_p^p$ are uniformly bounded by $C(T)\,M$ as t tends to zero, $i=1,2$. It remains to tackle W_t. Since $\sigma(x,z)$ is quasi-stable at least of order three, we know that $\sup_{x,z}|z|^{\frac1\alpha}\sigma_{zzz}(x,z)\le C\,|z|^3$. By definition of $v_1(z)$, we know that $\sup_z|z|^{\frac1{p_0}}|v_1(z)|\le C|z|$ and $\sup_z|z|^{\frac1{p_0}}|v_{1z}(z)|\le C|z|^2$. Therefore we have $\sup_{x,z}|z|^{\frac3{p_0}+\frac1\alpha}|\sigma_{zzz}(v_1^2(z)+v_{1z}(z))v_1(z)|\le C<\infty$. Analogously, we can show $\sup_{x,z}|z|^{\frac3{p_0}+\frac1\alpha}|\sigma_{zz}(X_{s^-},z)\partial_z(v_1^2(z)+v_{1z}(z))v_1(z)|\le C<\infty$. Thus, $\sigma_W^+=\sup_{x,z}|z|^{\frac3{p_0}+\frac1\alpha}|\partial_z(\partial_z(\sigma_z(x,z)v_1(z))v_1(z))v_1(z)|$ is finite and strictly larger than zero. Let $m_W=\sigma_W^{+\,-1}(\sup_{x,z}|\partial_z(\partial_z(\sigma_z(x,z)v_1(z))v_1(z))v_1(z)|\vee1)$. By Proposition 2.1 we know $W_t\le\sigma_W^+\int_0^t\int\big(m_W\wedge|z|^{-\frac3{p_0}-\frac1\alpha}\big)\mu(dz,ds)$, and by Proposition 2.2 $t^{-\delta_W}\|W_t\|_p$ is uniformly bounded on $[0,T]$ by $C(T)\,M$ for $3/p_0+1/\alpha>\delta_W$.

REMARK 2.11. *Note, concerning the processes $\hat A_{it}$, that the process which possesses the highest exponent in M is $\int_0^t\int D_lX_{s^-}D_l^2X_{s^-}d\hat A_{2\,s}$. To be more precise, we know $t^{-3p}\ \|\tilde A_{2t}\|_{3p}\ \|D_lX_t{}^*\|_{3p}\|D_l^2X_t{}^*\|_{3p}\le C\,M^8$ and that the other summands are at least of the same order.*

Combining (2.19), (2.4), and (2.20) we get

$$\begin{aligned}&t^{-\delta_W p}\|\hat H_t\|_p^p\le 2^{4p-1}\,C_1\ (M+m_b^p)\int_0^t\ \exp\Big(2^{4p-1}\Big(c^*(4p,h_K,t)+m_b^{4p}\Big)\,s\Big)\,ds\\ &+C_2\,M^{4p}\ \exp\Big(2^{4p-1}\Big(c(4p,h_K)+m_b^{4p}\Big)\,t\Big)\ t^{-i\delta_Hp}\ \big(\|J_t^1\|_{2p}^p+\|J_t^2\|_{2p}^p\big)\\ &+M^p\,C_3\ \exp\Big(2^{2p-1}\Big(c(2p,h_K)+m_b^{2p}\Big)\,t\Big)\ t^{-\delta_Vp}\|V_t\|_{2p}^p+C_4\ t^{-\delta_Wp}\|W_t^3\|_p^p\\ &+C_5M^{8p}\ \exp\Big(2^{2p-1}\Big(c(6p,h_K)+m_b^{2p}\Big)\,t\Big)\end{aligned}$$

as t tends to zero. Thus, we have

$$t^{-\delta_W p}\|D_l^3 X_t^*\|_p^p \le C(T)\, M^{8p}\ \exp\left(2^{8p-1}\left(c^*(8p,h_k,t)+m_b^{4p}\right)t\right). \tag{2.22}$$

REMARK 2.12. *Observe that the function h_K is bounded from above by one and from below by $-\frac{3}{4}$. Therefore since $|\sigma_x(x,z)| \le C|z|^{-\frac{1}{\alpha}}$ for z large enough, $|h_K|_p \le C(p)$ and the term $c^*(8p,h_K,t)$ is bounded by a constant $C(p,t)$ depending only on p and t.*

REMARK 2.13. *Analyzing the proof of the estimates* (2.15), (2.18), *and* (2.22) *we see (see also Remark* 2.11*) that the worst term is $\int_0^t D_l^{i-1} X_{s^-} D_l X_{s^-} d\dot{A}_{2s}$, where $\dot{A}$ equals A in the case of $i=2$, equals $\hat{A}$ in the case of $i=3$, and equals $\tilde{A}$ and in the case of $i=4$. A short calculation shows that the norm is bounded by M^{k_i}, where $k_i = k_{i-1}+1+i$, $k_2=4$.*

Analogously to the computations of $\|D_l^4 X_t^*\|$ and taking into account Remarks 2.12 and Remark 2.13, we obtain

$$\|D_l^4 X_t^*\|_p^p \le C(t)\, M^{13p}\ \exp\left(2^{16p}\left(c^*(16p,h_K,t)+m_b^{16p}\right)t\right) \le C(p,t,m_b)M^{13p}, \tag{2.23}$$

$$\|D_l^5 X_t^*\|_p^p \le C(t)\, M^{19p}\ \exp\left(2^{32p}\left(c^*(32p,h_K,t)+m_b^{32p}\right)t\right) \le C(p,t,m_b)M^{19p}. \tag{2.24}$$

Finally we have to investigate some estimates of $\|D_l^i \nabla^j X_t\|_p$, where $i+j \le 4$ and $\nabla^j X_t$ is given in (2.9). First, ∇X_t is the Doléans–Dade exponential $\mathcal{E}(K)_t$ with

$$K_t = \int_0^t \int \sigma_x(X_{s^-},z)(\mu-\gamma)(dz,ds) + \int_0^t b_x(X_{s^-})\,ds,$$

and norm $\|\mathcal{E}(K)_t^*\|_p \le \exp(\frac{1}{p}(c^*(p,t,\sigma_x)+m_b)t) \le \exp\left(c(p,t)\left(M^p+m_b\right)\right)$ for $p \ge 2$. Recursively applied, (2.9) yields $\nabla^i X_t = \mathcal{E}_{H_i}(K)_t$, where

$$\begin{aligned} H_{it} = &\sum \int_0^t \int \partial_x^j \sigma(X_{s^-},z)\nabla^{|I_1|}X_{s^-}\ldots\nabla^{|I_n|}X_{s^-}\,(\nu-\gamma)(dz,ds)\\ &+\sum \int_0^t \partial_x^j b(X_{s^-},z)\nabla^{|I_1|}X_{s^-}\ldots\nabla^{|I_n|}X_{s^-}\,ds, \end{aligned}$$

where the sum runs over the set of all partitions with length larger than 1 of $\{1,\ldots,i\} = I_1 \cup \cdots \cup I_\nu$ and $|\cdot|$ denotes the length. A short calculation yields

$$\begin{aligned} H_{2t} &= \int_0^t \int \sigma_{xx}(X_{s^-},z)\nabla X_{s^-}^2\,(\mu-\gamma)(dz,ds) + \int_0^t b_{xx}(X_{s^-},z)\nabla X_{s^-}^2\,ds,\\ &\ldots \quad \ldots\\ H_{4t} &= \int_0^t \int \left(\sigma_{xxxx}(X_{s^-},z)\nabla X_{s^-}^4 + \cdots + \sigma_{xx}(X_{s^-},z)\nabla^3 X_{s^-}\nabla X_{s^-}\right)(\mu-\gamma)(dz,ds)\\ &\quad + \int_0^t \left(b_{xxxx}(X_{s^-},z)\nabla X_{s^-}^4 + \cdots + b_{xx}(X_{s^-},z)\nabla^3 X_{s^-}\nabla X_{s^-}\right)ds. \end{aligned}$$

Applying [3, Lemma 2.1] we get for $p \ge 2$

$$\begin{aligned} &\|\nabla^2 X_t\|_p \le \exp\left(\frac{1}{p}(c(p,t,\sigma_x)+m_b)t\right)\cdot\|\nabla X_t^2\|_p \le \exp\left(C(p,t)(M^{2p}+m_b)\right),\\ &\|\nabla^3 X_t\|_p \le \exp\left(C(p,t)(M^{4p}+m_b)\right),\\ &\|\nabla^4 X_t\|_p \le \exp\left(C(p,t)(M^{8p}+m_b)\right). \end{aligned} \tag{2.25}$$

Obviously $D_l\nabla X_t$ is the generalized Doléans–Dade exponential $\mathcal{E}_H(K)_t$ with

$$H_t = \int_0^t \int \sigma_{xx}(X_{s^-},z) D_l X_{s^-} \nabla X_{s^-} (\mu-\gamma)(dz,ds)$$
$$+ \int_0^t \int \sigma_{xz}(X_{s^-},z) \left(v_1(z) + D_l X_{s^-} \frac{\sigma_x(X_{s^-},z)}{\sigma_z(X_{s^-},z)} \varphi(\sigma_x(X_{s^-},z)) \right) \nabla X_{s^-} v(s,z) \mu(dz,ds).$$

Remark 2.12 and hypotheses (H0), (H1), (H2), and (H3) lead to

$$\|H_t\| \leq C(p,t) \cdot M \cdot \int_0^t \|D_l X_s\|_{2p} \|\nabla X_s\|_{2p}\, ds,$$

and therefore

$$\|D_l \nabla X_t\|_p \leq C(p,t)\, M \, \exp\left(\frac{1}{p} \left(c^*(p,t,\sigma_x) + m_b \right) t \right) \cdot \|D_l X_t^*\|_{2p} \cdot \|\nabla X_t^*\|_{2p}$$
$$\leq C(t,p) \cdot M \cdot \exp\left(c(t,p)(M^{2p} + m_b) \right). \tag{2.26}$$

Iterating (2.18) yields

$$\|D_l^2 \nabla X_t\|_P \leq \exp\left(\frac{1}{p} \left(c(p,t,\sigma_x) + m_b \right) t \right) \|\nabla X_t\|_{2p} \left(M^2 \|D_l X_t^2\|_{2p} + M \|D_l^2 X_t\|_{2p} \right),$$
$$\leq C(t,p) M^5 \exp\left(c(t,p)(M^{2p} + m_b) \right),$$
$$\|D_l^3 \nabla X_t\|_P \leq C(t,p) M^9 \exp\left(c(t,p)(M^{2p} + m_b) \right),$$

and

$$\|D_l^4 \nabla X_t\|_P \leq C(t,p) M^{14} \exp\left(c(t,p)(M^{2p} + m_b) \right)$$

for $M \geq 1$. Finally we have for $D_l \nabla^4 X_t$

$$\|D_l \nabla^4 X_t\|_p \leq C(p,t) M \exp\left(c(t,p)(M^{8p} + m_b) \right). \tag{2.27}$$

As our last point we give an estimate of $\|L_t\|_p$ and $\|D_l^i L_t\|_p$, $i = 1, \ldots, 4$. Combining (2.6) and (2.11), we obtain

$$L_t = \int_0^t \int \frac{\partial}{\partial z} \left(\frac{z^2}{1+|z|^{1+\frac{1}{p_0}}} \right) (\mu-\gamma)(dz,ds) - \int_0^t \int \phi(x,z) D_l X_{s^-} \, (\mu-\gamma)(dz,ds)$$
$$=: L_t^1 - L_t^2.$$

A short calculation shows $t^{-\frac{1}{p_0}} \|L_t^1\|_p \leq C(p)$. Further, by the Emery and Burkholder–Gundy inequality it follows for L_t^2 and $p \geq 2$ that $t^{-\frac{1}{p_0}} \|L_t^2\|_p \leq M \|D_l X_t^*\|_p$ and therefore

$$t^{-\frac{1}{p_0}} \|L_t\|_p \leq (M^2 + 1). \tag{2.28}$$

Evaluating $D_l L_t$ we get

$$D_l L_t = \int_0^t \int \frac{\partial}{\partial x} \phi(x,z) D_l X_{s^-}^2 \, (\mu-\gamma)(dz,ds) + \int_0^t \int \phi(x,z) D_l^2 X_{s^-}^2 \, (\mu-\gamma)(dz,ds)$$
$$+ \int_0^t \int \frac{\partial^2}{\partial z^2} \left(\frac{z^2}{1+|z|^{1+\frac{1}{p_0}}} \right) \left(v_1(z) + \phi(s,z) D_l X_{s^-} \right) \mu(dz,ds),$$

and therefore

$$\|D_l L_t\|_p \leq c_1(p) M \|D_l X_t^{*\,2}\|_p + c_2(p) M \|D_l^2 X_t^*\|_p + c_3(p) + c_4(p) M \|D_l X_t^*\|_p$$
$$\leq C(p)\, M^5. \tag{2.29}$$

3. Proof of Theorem 1.2. Suppose that X_t is a solution of the SDE (1.1) and X_t^n is its Euler approximation, and for the sake of simplicity, let $T = \frac{m}{n}$. Following Bally and Talay [1], we define the functional $u(x,t) = \mathbb{E}_x[f(X_{T-t}^n)]$. Then we can write

$$
\begin{aligned}
(3.1)\qquad & |\mathbb{E}_{x_0}[f(X_T)] - \mathbb{E}_{x_0}[f(X_T^n)]| = |\mathbb{E}_{x_0}[u(X_T,T) - u(x_0,0)]| \\
&= \left|\sum_{k=1}^{m} \mathbb{E}_{x_0}\left[u\left(X_{\frac{k}{n}}, \frac{k}{n}\right) - u\left(X_{\frac{k-1}{n}}, \frac{k-1}{n}\right)\right]\right| =: \left|\sum_{k=1}^{m} \mathbb{E}_{x_0}[\delta_k^n]\right|.
\end{aligned}
$$

For simplicity we consider only δ_1^n and omit the index x_0. Applying the Dynkin formula, we get

$$
\begin{aligned}
&\mathbb{E}[\delta_1^n] = \mathbb{E}\left[u\left(X_{\frac{1}{n}}, \frac{1}{n}\right) - u(x_0,0)\right] \\
&= \mathbb{E}\left[\int_0^{\frac{1}{n}} \left(\partial_x u(X_{s^-},s)b(X_{s^-}) + R(u(X_{s^-},s),X_{s^-}) + \partial_t u(X_{s^-},t)\Big|_{t=s}\right) ds\right],
\end{aligned}
$$

where $R(f(\cdot),x) = \int_{|y|\le 1}(f(\cdot+y) - f(\cdot) - f'(\cdot)y)\nu(x,dy)$, with $\nu(x,[y,\infty)) = \inf_z\{\sigma(x,z) \le y\}$. Applying the Dynkin formula to $\partial_t u(x,t)$, we get

$$
\begin{aligned}
\mathbb{E}[\delta_1^n] &= \mathbb{E}\Bigg[\int_0^{\frac{1}{n}} \partial_x u(X_{s^-},s)b(X_{s^-}) - \partial_x u(X_{s^-},s)b(x_0) + R(u(X_{s^-},s),X_{s^-}) \\
&\qquad - R(u(X_{s^-},s),x_0)\, ds\Bigg] \\
&= \mathbb{E}\Bigg[\int_0^{\frac{1}{n}} \partial_x u(X_{s^-},s)b(X_{s^-}) - \partial_x u(x_0,0)b(x_0) - \partial_x u(X_{s^-},s)b(x_0) \\
&\qquad + \partial_x u(x_0,0)b(x_0) + R(u(X_{s^-},s),X_{s^-}) - R(u(X_{s^-},s),x_0) \\
&\qquad + R(u(x_0,0),x_0) - R(u(X_{s^-},s),X_{s^-})\, ds\Bigg].
\end{aligned}
$$

Applying the Dynkin formula a third and fourth time, we get

$$
\begin{aligned}
&\mathbb{E}[\delta_1^n] = \mathbb{E}\Bigg[\int_0^{\frac{1}{n}}\int_0^{s} \partial_x\Big(\partial_x u(X_{r^-},r)(b(X_{r^-}) - b(x_0))\Big)(b(X_{r^-}) - b(x_0)) \\
&+ R\Big(\partial_x u(X_{r^-},r)(b(X_{r^-}) - b(x_0)), X_{r^-}\Big) \\
&- R\Big(\partial_x u(X_{r^-},r)(b(X_{r^-}) - b(x_0)), x_0\Big) \\
&\partial_x\Big(R(u(X_{r^-},r),X_{r^-}) - R(u(X_{r^-},r),x_0)\Big)(b(X_{r^-}) - b(X_0)) \\
&+ R\Big(R(u(X_{r^-},r),X_{r^-}) - R(u(X_{r^-},r),x_0), X_{r^-}\Big) \\
(3.2)\qquad &- R\Big(R(u(X_{r^-},r),X_{r^-}) - R(u(X_{r^-},r),x_0), x_0\Big)\, dr\, ds\Bigg].
\end{aligned}
$$

Before continuing, we investigate the operator R. By the Taylor formula, we know that for the remainder term we have

$$
f(\cdot + y) - f(\cdot) - f'(\cdot)\cdot y = \int_0^y z f''(\cdot + y + z)dz.
$$

Set $h_x(z) = \sigma(x,z)$. A short calculation shows that $\nu(x,dy) = g(x,y)dy$, where $g(x,y) = 1/[h'_x(h_x^{-1}(y)]$. What happens if the operator R is applied to the product $u(x)g(x)$? Set $f(x) = (ug)(x)$. Then we have

$$\begin{aligned} f(x+y) - f(x) - f'(x)\cdot y &= \big(u(x+y) - u(x) - u_x(x)\cdot y\big)g(x) \\ &\quad + u(x+y)\big(g(x+y) - g(x)\big) - g_x(x)u(x)y \\ &= \Big(\int_0^y z u_{xx}(x+y+z)dz\Big)\, g(x) \\ &\quad + \big(u(x+y)\big(g(x+y) - g(x)\big) - g_x(x)u(x)y\big). \end{aligned} \tag{3.3}$$

Set $x = X_{r^-}$ and $g(x) = b(x) - b(x_0)$ or $g(x) = g(x,y) - g(x_0,y)$. Taking the expectation, obviously the second summand of (3.3) is bounded, because no derivative of u appears. Thus, we have

$$\begin{aligned} \mathbb{E}[\delta_1^n] = \int_0^{\frac{1}{n}} \int_0^s \Big\{ &\mathbb{E}\big[u_{xx}(X_{r^-},r)\big(b(X_{r^-}) - b(x_0)\big)^2 \\ &\quad + u_x(X_{r^-},r)b_x(X_{r^-})\big(b(X_{r^-}) - b(x_0)\big)\big] \\ &+ \int_{|y|\le 1}\int_0^y \mathbb{E}\big[2u_{xxx}(X_{r^-}+y+z,r)\cdot z \\ &\qquad \times \big(b(X_{r^-}) - b(x_0)\big)\big(g(X_{r^-},y) - g(x_0,y)\big)\big]\, dy\, dz \\ &+ \int_{|y_1|\le 1}\int_0^{y_1}\int_{|y_2|\le 1}\int_0^{y_2} \mathbb{E}[u_{xxxx}(X_{r^-}+y_1+z_1+y_2+z_2,r)\cdot z \\ &\quad \times \big(g(X_{r^-},y_1) - g(x_0,y_1)\big)\big(g(X_{r^-},y_2) - g(x_0,y_2)\big)\big]\, dy_1\, dz_1\, dy_2\, dz_2 \\ &+ C\Big\}\, dr\, ds. \end{aligned} \tag{3.4}$$

Due to (H1) we know that $g(x,y)$ is differentiable with respect to x; i.e., $g_x(x,y) = -h'_x(z)^{-2}\ (\sigma_{xz}(x,z)\ +\sigma_{zz}(x,z)\partial_x z)$, where $z = h_x^{-1}(y)$. Thus $g(x,y) - g(x_0,y) = (x-x_0)g_x(\xi_{x,x_0},y)$ for some $\xi_{x,x_0} \in (x,x_0)$. It follows for $R(\cdot,\cdot)$ that

$$R(f(\cdot),x) - R(f(\cdot),x_0) = (x-x_0)\int_{|y|\le 1}\int_0^y z f''(\cdot + y - z) g_x(\xi_{x,x_0},y)\, dz\, dy.$$

Setting $j=1$ in condition (H3), we know $\int |y^2 g_x(x,y)|dy$ to be uniformly bounded in x and therefore $|R(f(\cdot),x) - R(f(\cdot),x_0)| \le C\,|x-x_0|$. Going back to (3.4) we have to show that the terms

(1) $\mathbb{E}\big[u_{xx}(X_{r^-},r)\big(b(X_{r^-}) - b(x_0)\big)^2 + u_x(X_{r^-},r)b_x(X_{r^-})\big(b(X_{r^-}) - b(x_0)\big)\big]$,
(2) $\mathbb{E}\big[2u_{xxx}(X_{r^-}+y,r)\big(b(X_{r^-}) - b(x_0)\big)\big(g(X_{r^-},y) - g(x_0,y)\big)\big]$,
(3) $\mathbb{E}\big[u_{xxxx}(X_{r^-}+y,r)\big(g(X_{r^-},y_1) - g(x_0,y_1)\big)\big(g(X_{r^-},y_2) - g(x_0,y_2)\big)\big]$

are bounded for all y with $|y| \le 2$.

Next, analogously to Bally and Talay we distinguish in the following small t from large t. If t is large, i.e., $t/2 \le t \le T$, we get rid of the derivatives of $u(t,x)$ using Malliavin's integration by parts formula with respect to the functional $X_t(x)$ and apply formula (2.7). For small t, i.e., $0 \le t \le t/2$, we apply formula (2.10).

The case $T/2 \le t \le T$. Picking up the worst case, we must show $|\mathbb{E}[\partial_x^4 \bar{f}(X_t))]| \le C(T)M^{21}$ for $\bar{f}(x) = u(x,t)$. Define $H_X[Y] = (D_lX)^{-1}(YL + D_l^2 X D_l X^{-1} Y + D_l Y) = D_l(D_l X^{-1} Y) + LY D_l X^{-1}$, where L is defined by (2.6) and (2.11). First, note that we have chosen l so that $D_l X_t$ is invertible (see Remarks 2.6 and 2.7). By the integration by parts formula (2.7), we know $\mathbb{E}[\partial_x g(X)Y] = \mathbb{E}[g(X)H_X[Y]]$. Applying integration by parts and $u(x,\cdot) \le |f|_\infty$ to the inner part of (3.4), we get

$$
\begin{aligned}
|\mathbb{E}[\delta_1^n]| \le \int_0^{\frac{1}{n}} \int_0^t \big(& |f|_\infty \mathbb{E}\big[|H_X^4[(b(X) - b(x_0))^2]_s + H_X^4[(g(\cdot, X) - g(\cdot, x_0))^2]_s \\
& + 2\, H_X^4[(b(X) - b(x_0))(g(\cdot, X) - g(\cdot, x_0))|\big]_s + C \big)\, ds\, dt.
\end{aligned} \tag{3.5}
$$

The worst case is due to the summand $H_X^4[(g(\cdot, X) - g(\cdot, x_0))^2]$. Thus we restrict ourselves to this case and define $Y_t = b(X_t) - b(x_0)$ and $R(\cdot, X_t) - R(\cdot, x_0)$. For clarity, we omit the index t if it is obvious and write G for $L + D_l X^{-1} D_l^2 X$. Define the operator $\widehat{H}_X[Y] = D_l Y + (L + D_l X^{-1} D_l^2 X)Y$. Obviously, we have $H_X[Y] = D_l X^{-1} \widehat{H}_X[Y]$. Next we evaluate $H_X^4[Y^2]$:

$$
\widehat{H}_X^4[Y^2] = D_l^2(\hat{H}_X^2[Y^2]) + 2G D_l(\hat{H}_X^2[Y^2]) + (D_l G + G^2)\hat{H}_X^2[Y^2].
$$

After some computations we get

$$
\begin{aligned}
\widehat{H}_X^4[Y^2] &= Y^2\big(G^4 + D_l^3 G + 6G^2 D_l G + 3(D_l G)^2 + 4G D_l^2 G\big) \\
&\quad + 8 D_l Y Y \big(3 D_l G G + G^3 + D_l^2 G\big) + 12\big(D_l^2 Y Y + (D_l Y)^2\big)(G^2 + D_l G) \\
&\quad + 8G\big(D_l^3 Y Y + 3 D_l^2 Y D_l Y\big) + 2\big(D_l^4 Y Y + 4 D_l^3 Y D_l Y + 3(D_l^2 Y)^2\big) \\
&= L^4 Y^2 + 6L^2 Y^2 D_l L + 3Y^2 D_l L^2 + 8L^3 Y D_l Y + \cdots \\
&\qquad + 8 D_l X^{-1} Y D_l Y D_l^4 X + 2Y D_l^4 Y + D_l X^{-1} Y^2 D_l^5 X.
\end{aligned} \tag{3.6}
$$

Set $Y = b(X) - b(x_0)$. Since b is four times boundedly differentiable, $D_l Y = b'(X) D_l X, \ldots, D_l^4 Y = \sum b^{(4)}(X) D_l X^4 + \cdots + b^{(1)}(X) D_l^4 X$ are bounded by $D_l X, \ldots, \sum_{i=1}^4 D_l^i X$. In the case of the operator R, we can set $Y \sim X$.

Next we investigate (3.6) term by term and show that the terms are, when multiplied by $D_l X_t^{-4}$, bounded by $C(T)M^{21}$. Note that thanks to Remark 2.6 we have an upper bound of $\|D_l X_t^{-1}\|_p$ independent of M. The first term we consider is $Y^2 L^4$. Using (2.28) we get

$$
\begin{aligned}
\|Y_t^2 L_t^4 D_l X_t^{-4}\|_1 &\le \|Y_t^2\|_4 \|L_t^4\|_4 \|D_l X_t^{-4}\|_2 \\
&\le t^{-Q}\, C(t) \cdot M^2 \cdot M^8 \;=\; t^{-Q}\, C(t) \cdot M^{10}
\end{aligned}
$$

for $t \in [T/2, T]$ and $Q < \infty$. Since $T > t > T/2$, the right side is uniformly bounded by $C(T)\, M^{10}$ for all $t \in [T/2, T]$. Analyzing term by term we get, by Remark 2.12 and the estimates (2.15), (2.18), (2.22), (2.23), (2.24), (2.28), and (2.29),

$$
\begin{aligned}
\|D_l X_t^{-4} L_t^2 Y_t^2 D_l L_t\|_1 &\le \|D_l X_t^{-4}\|_3 \cdot \|L_t^2\|_4 \cdot \|Y_t^2\|_6 \|D_l L_t\|_4 \\
&\le t^{-Q}\, C(t) \cdot M^4 \cdot M^2 \cdot M^5 \;=\; t^{-Q}\, C(t) \cdot M^{11}, \\
\|D_l X_t^{-4} Y_t^2 D_l L_t^2\|_1 &\le \|D_l X_t^{-4}\|_4 \cdot \|Y_t^2\|_4 \cdot \|D_l L_t^2\|_2 \\
&\le t^{-Q}\, C(t) \cdot M^2 \cdot M^{10} \;\le\; t^{-Q}\, C(t) \cdot M^{14}, \\
&\quad \vdots \\
\cdots &\quad \vdots \quad \cdots
\end{aligned}
$$

$$\begin{aligned}\|D_l X_t^{-4} Y_t\big(8D_l X_t^{-1} D_l Y_t D_l^4 Y_t + 2D_l^4 Y_t + D_l X_t^{-1} Y_t D_l^5 X_t\big)\|_1 \\ &\le t^{-Q}\, C(t)\cdot M\cdot\big(MM^{13}+M^{13}+MM^{19}\big)\\ &\le t^{-Q}\, C(t)\cdot M^{21}\,.\end{aligned}$$

Collecting all together, we get for $t\in[T/2,T]$

$$\mathbb{E}[|H^4[Y^2]_t|] = \|H^4[Y^2]_t\|_1 \le C(T)\cdot M^{21}.$$

The case $0 < t \le T/2$. Here, we have to apply formula (2.10) with time $r = T-t > T/2$. A short calculation shows that the terms $\mathbb{E}\left[\mathbb{E}\left[H_X[\nabla^4 X_r]|\mathcal{F}_t\right] Y_t^2\right]$, $\mathbb{E}\left[\mathbb{E}\left[H_X^2[\nabla^2 X_r^2]|\mathcal{F}_t\right] Y_t^2\right]$, $\mathbb{E}\left[\mathbb{E}\left[H_X^2[\nabla^3 X_t \nabla X_r]|\mathcal{F}_t\right] Y_t^2\right]$, $\mathbb{E}\left[\mathbb{E}\left[H_X^3[\nabla X_r^2 \nabla^2 X_r]|\mathcal{F}_t\right] Y_t^2\right]$, and $\mathbb{E}\left[\mathbb{E}\left[H_X^4[\nabla X_r^4]|\mathcal{F}_t\right] Y_t^2\right]$ have to be bounded by $r^{-Q}\, C(r)\cdot M^4\cdot\exp(M^{16})$ for some $Q<\infty$ and $C(r)$ is uniformly bounded on $[0,T]$. But the estimates given in equations (2.15), (2.18), and (2.22)–(2.29) lead to

$$\begin{aligned}\|H_X[\nabla^4 X_r]\|_1 &\le \|D_l X_r^{-1}\big(D_l\nabla^4 X_r + \nabla^4 X_r L_t + \nabla^4 X_r D_l^2 X_r D_l X_r^{-1}\big)\|_1\\ &\le r^{-Q}\, C(r)\cdot\exp(M^{16})\cdot\big(M+M^2+M^4\big)\,,\\ \|H_X^2[\nabla^2 X_r^2]\|_1 &\le \|D_l X_r^{-2}\big(L_r^2\nabla^2 X_r^2 + D_l L_r \nabla^2 X_r^2 + \cdots + D_l X_r^{-1}\nabla^2 X_r^2 D_l^3 X_r\big)\|_1\\ &\le r^{-Q}\, C(r)\cdot\exp(M^8)\cdot\big(M^4+M^5+\cdots+\exp(M^4)+M^8\big)\,,\\ &\cdots\quad\cdots\\ \|H_X^4[\nabla X_r^4]\|_t &\le \|D_l X_r^{-4}\big(L_r^4\nabla X_r^4 + 6L_r^2\nabla X_r^4 D_l L_t + \cdots + \nabla X_t^4 D_l^5 X_t\big)\|_1\\ &\le\ r^{-Q}\, C(r)\cdot(\exp(M^{16})M^8 + 6\exp(M^{16})M^4M^5 + \cdots + M^{19}\exp(M^{16})).\end{aligned}$$

Integration of (3.4) and applying the same localization argument as Bally and Talay [1] completes the proof. □

4. Conclusions and additional remarks.

REMARK 4.1. *To handle the case where f is a Delta function, we have to investigate the upper bound of $\|H^5[Y^2]\|_1$, $M^2\,\|H_X^5[\nabla X_t^5]_t\|_1$, and $M^2\|H_X[\nabla^5 X]_t\|_1$. But analyzing the preceding proof it is obvious that the worst term in the first case is $\|D_l^6 X_t Y_t^2 D_l X_t^{-5}\|_1$, which is of order M^{28}. In the second and third case the worst term is $\|H_X^5[\nabla X^5]\|M^2$, which is of order $M^{28}\exp(M^{32})$.*

REMARK 4.2. *Assume that f belongs to $C_b^1(\mathbb{R})$ and consider $u(x,t)=\mathbb{E}^x[f(X_{T-t}^n)]$ $\approx \mathbb{E}^x[f(X_{T-t})]$. To give an upper bound of $\partial_{x^4}^4 u(x,t)$, we have again to distinguish between small t and large t.*

Let $T/2\le t\le T$. Note that formula (2.10) leads to $\partial_x u(x,t)=\mathbb{E}^x[f'(X_{T-t})\nabla X_{T-t}]$ and therefore $|\partial_x u(x,t)|_\infty \le |f'|_\infty \sup_x \|\nabla X_{T-t}(x)\|_1 \le |f'|_\infty \exp\big((T-t)M^2\big)$. To get rid of the remaining derivatives, we have to apply the operator $H_X[\cdot]$ only three times to Y^2:

$$\begin{aligned}H_X^3[Y^2]_t = D_l X_t^{-3}\Big\{&L_t^3 Y_t^2 + 3L_t Y_t^2 D_l L_t + 6L_t^2 Y_t D_l Y_t + \cdots\\ &+ 6Y_t D_l Y_t D_l^3 X_t D_l X_t^{-2} + 2Y_t D_l^3 Y_t + Y_t^2 D_l^4 X_t D_l X_t^{-1}\Big\}.\end{aligned}\tag{4.1}$$

The worst factor is $Y_t^2 D_l^4 X_t D_l X_t^{-4}$. But $\|Y_t^2 D_l^4 X_t D_l X_t^{-4}\|_1 \le M^2 M^{13} \le C(T)M^{15}$, and therefore $\mathbb{E}[\partial_x^4 u(t,x)] \le |f'|_\infty \exp\big(TM^2\big)\, M^{15}$.

Let $0 \le t \le T/2$. Then we have

$$\begin{aligned}
&\left|\partial_x^4 u(t,x)\right| = \left|\mathbb{E}^x\left[f^{(4)}(X_{T-t})\nabla X_t^4 + \cdots + f'(X_{T-t})\nabla^4 X_t\right]\right| \\
&= \left|\mathbb{E}^x\left[f'(X_{T-t})H_X^3[\nabla X^4]_t + \cdots + f'(X_{T-t})\nabla^4 X_t\right]\right| \\
&\le |f'|_\infty\big(\|H_X^3[\nabla X^4]_t\|_1 + \cdots + \|\nabla^4 X_t\|_1\big) \\
&\le C\,|f'|_\infty\big(\|L_t^3\nabla X_t^4 D_l X_t^{-3}\|_1 + \cdots + \|\nabla X_t^4 D_l^4 X_t D_l X_t^{-3}\|_1 + \cdots + \|\nabla^4 X_t\|_1\big) \\
&\le C\,|f'|_\infty\big(M^6\exp(M^{16}) + \cdots + M\exp(2M^8) + \exp(M^{16})\cdot M^{13} + \cdots + \exp(M^{16})\big).
\end{aligned}$$

Collecting all together, we get as an error bound

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \le |f'|_\infty \frac{C}{n} M^{15}\,\left(\exp(M^2) + \exp(M^{16})\right).$$

REMARK 4.3. *Assume that $2/\alpha^+ > 3/\alpha^- - 1$ and $f \in C^1(\mathbb{R})$. Without any restriction on α^+ and α^-, we can see by the scaling properties that the right-hand side of* (4.1) *tends to infinity as t tends to zero. But with the restriction above on α^+ and α^-, the right-hand side of* (4.1) *is of order $O(t^\delta)$, where $\delta > -1$. Integration leads to the error bound*

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \le |f'|_\infty \frac{C}{n} M^{15}\,\exp(M^2).$$

It remains to show the right-hand side of (4.1) *is of order $O(t^\delta)$, i.e., $\|H_X^3[Y^2]_t\|_1 \le O(t^\delta)$. Analyzing term by term by the estimates* (2.15), (2.18), (2.22), (2.23), (2.24), (2.28), *and* (2.29) *we see that the leading term in* (4.1) *is the worst. Thus, we pick it up and show $\|D_l X_t^{-3} Y_t^2 L_t^3\|_1 \le O(t^\delta)$, where $\delta > -1$. As t tends to zero we know by Proposition* 2.2 *and Remark* 2.7 *that $\|D_l X_t^{-3}\|_p \le c(p)\,t^{-\frac{3}{\alpha^-}-\frac{3}{p_0}}$ and $\|Y_t^2\|_p \le C(M)\,t^{\delta_Y}$, where $\alpha^+ > 2\delta_Y$. Thanks to estimate* (2.28), *we see that $\int_0^T \|H_X^3[Y^2]_t\|_1\,dt$ is bounded for $\frac{2}{\alpha^+} > \frac{3}{\alpha^-} - 1$. To get the exact threshold of $\|H_X^3[Y^2]_t\|_1$, we have to investigate the last summand of* (4.1), *i.e., $\|D_l X_t^{-4} D_l^4 X_t Y_t^2\|_1$, which is of $O(M^{15})$.*

REMARK 4.4. *Assume that f is in $C_b^2(\mathbb{R})$. To show*

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \le |f''|_\infty \frac{C}{n} M^{10}\,\left(\exp(M^4) + \exp(M^{16})\right), \tag{4.2}$$

we have again to distinguish between large and small t.

Let $T/2 \le t \le T$. Then we have $\partial_x^2 u(x,t) = \mathbb{E}^x[f''(X_{T-t})\nabla^2 X_{T-t}]$ and $|\partial_x^2 u(x,t)|_\infty \le |f''|_\infty \exp\big((T-t)M^4\big)$. To get rid of the two remaining derivatives, we have to apply the operator $H_X[\cdot]$ only twice to Y^2, i.e.,

$$H_X^2[Y^2]_t = D_l X_t^{-2}\Big\{L_t^2 Y_t^2 + Y_t^2 D_l L_t + \cdots + 2Y_t D_l^2 Y_t - Y_t^2 D_l^3 X_t D_l X_t^{-1}\Big\}. \tag{4.3}$$

But we have $\|Y_t^2 D_l^3 X_t D_l X_t^{-3}\|_t \le C(T)\,M^{10}$.

Let $0 \le t \le T/2$ and set $r = T - t$. Analogous to Remark 4.2 *we have*

$$\begin{aligned}
&\left|\partial_x^4 u(t,x)\right| = \left|\mathbb{E}^x\left[f^{(4)}(X_{T-t})\nabla X_t^4 + \cdots + f'(X_{T-t})\nabla^4 X_t\right]\right| \\
&= \left|\mathbb{E}^x\left[f^{(2)}(X_{T-t})H_X^2[\nabla X^4]_t + \cdots + f^{(2)}(X_{T-t})\nabla^4 X_t\right]\right| \\
&\le |f''|_\infty\big(\|H_X^2[\nabla X^4]_t\|_1 + \cdots + \|\nabla^4 X_t\|_1\big) \le C\,|f''|_\infty\exp(M^{16})\cdot M^8.
\end{aligned}$$

Collecting all together, we get (4.2).

REMARK 4.5. *Assume that* $2/\alpha^+ > 3/\alpha^- - 1$ *and* $f \in C_b^2(\mathbb{R})$. *Without any restriction on* α^+ *and* α^- *we can see by the scaling properties that the right-hand side of* (4.1) *tends to infinity as* t *tends to zero. But with the restriction above on* α^+ *and* α^-, *the right-hand side of* (4.3) *is of order* $O(t^\delta)$, *where* $\delta > -1$. *Integration leads to the error bound*

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \leq |f''|_\infty \frac{C}{n} M^{10} \exp(M^4).$$

It remains to show the right-hand side of (4.3) *is of order* $O(t^\delta)$, *i.e.,* $\|H_X^2[Y^2]_t\|_1 \leq O(t^\delta)$. *The worst term in* (4.3) *is the leading term. Thus, we have to show* $\|D_l X_t^{-2} Y_t^2 L_t^2\|_1 \leq O(t^\delta)$, *where* $\delta > -1$. *As* t *tends to zero we know by Proposition* 2.2 *and Remark* 2.7 *that* $\|D_l X_t^{-2}\|_p \leq c(p)\, t^{-\frac{2}{\alpha^-} - \frac{2}{p_0}}$ *and* $\|Y_t^2\|_p \leq C(M)\, t^{\delta_Y}$, *where* $\alpha^+ > 2\delta_Y$. *Thanks to estimate* (2.28), *we see that* $\int_0^T \|H_X^3[Y^2]_t\|_1\, dt$ *is bounded for* $\frac{2}{\alpha^+} > \frac{2}{\alpha^-} - 1$. *To get the exact threshold of* $\|H_X^2[Y^2]_t\|_1$, *we have to investigate the last summand of* (4.3), *which is of order* $O(M^{10})$.

REMARK 4.6. *Analogous to Remarks* 4.2 *and* 4.4 *we can show*

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \leq |f'''|_\infty \frac{C}{n} M^6 \left(\exp(M^8) + \exp(M^{16})\right)$$

for $f \in C_b^3(\mathbb{R})$. *If* $\frac{2}{\alpha^+} > \frac{1}{\alpha^-} - 1$, *the same consideration as in Remarks* 4.3 *and* 4.5 *yields*

$$|\mathbb{E}\left[f(X_t)\right] - \mathbb{E}\left[f(X_t^n)\right]| \leq |f'''|_\infty \frac{C}{n} M^6 \exp(M^8).$$

REMARK 4.7. *If the driving process is* α*-stable, i.e.,* $\sigma(x,z)$ *can be written as* $\sigma_0(x) z^{-1/\alpha}$, *we have to truncate the process before applying Theorem* 1.2. *To be more exact, we call* X_t^m *the truncated process if* X_t^m *is a solution to* (1.1), *where* $\sigma(x,z)$ *is replaced by* $\sigma_0(x)(z^{-1/\alpha} \wedge m)$. *The error now splits into three parts:*

$$\begin{aligned}|\mathbb{E}[f(X_T) - f(X_t^n)]| \leq{}& |\mathbb{E}[f(X_T) - f(X_T^m)]| + |\mathbb{E}[f(X_T^m) - f(X_T^{m,n})]| \\ &+ |\mathbb{E}[f(X_T^m) - f(X_T^{m,n})]| =: \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Given an estimate of (II), *we can apply Theorem* 1.2. *Further, we have*

$$(\mathrm{I}) \leq |\mathbb{E}[f(X_T) - f(X_T^m) 1_{\{T^m \leq T\}}]| \leq 2|f|_\infty \mathbb{P}(T^m \leq T),$$

where $T^m = \inf_{t>0}\{|\Delta Z| > m\}$. *Hence the counting process of* $[m,\infty)$ *is Poisson distributed with parameter* $\mu([m,\infty))$ *(see section* 2 *or Protter and Talay* [15, Proposition 4.5]*); it follows that* $\mathbb{P}(T^m \leq T) = (1 - \exp(Tm^{-\alpha}))$. *Therefore, setting* $M = m$, *the error is given by*

$$|\mathbb{E}[f(X_T) - f(X_T^n)]| \leq C(T) \frac{1}{n} \cdot \left(m^{21} + m^4 \exp(m^{16})\right) + \left(1 - \exp\left(Tm^{-\alpha}\right)\right).$$

REMARK 4.8. *Let* X_t *be a solution to* (1.1), *where* σ *and* b *satisfy the assumption of Theorem* 1.2 *with* $M = 1$. *Proceeding as Kanagawa* [12] *(see also Talay* [17, Proposition 2.1]*) we can show that* $\|X_t^n - X_t\|_2 \leq C/\sqrt{n}$. *Let* $t = [t]_n$. *Using condition* (H3), *an induction on* k *shows that for any* $n \in \mathbb{N}$,

$$\mathbb{E}\left[\sup_{t\in[0,T]} |X_t^n|^2\right] \leq K(T)(1 + x_0^2)\exp(K(T)) \tag{4.4}$$

for some increasing function $K(\cdot)$. For $t \in (k/n, (k+1)/n]$ consider the process

$$\varepsilon_t := X_{\frac{k}{n}} - X^n_{\frac{k}{n}} + \int_{\frac{k}{n}}^t (b(X_{s^-}) - b(X^n_{\frac{k}{n}}))ds + \int_{\frac{k}{n}}^t (\sigma(X_{s^-}, z) - \sigma(X^n_{\frac{k}{n}}, z))(\mu - \gamma)(dz, ds).$$

Apply the Itô formula to $(\varepsilon_t)^2$ between $t = k/n$ and $t = (k+1)/n$: Standard computations, condition (H3)*, and* (4.4) *show that we have, for an increasing function $K(\cdot)$,*

$$\mathbb{E}[\varepsilon^2_{\frac{k+1}{n}}] \le \mathbb{E}[\varepsilon^2_{\frac{k}{n}}] \left(1 + \frac{K(T)}{n}\right) + \frac{1}{n^2}.$$

Noting that $\varepsilon_0 = 0$, an induction on k provides the estimate

$$\sup_{0 \le k \le n} \mathbb{E}\left[\varepsilon^2_{\frac{k+1}{n}}\right] \le C \exp(K(T)).$$

To conclude, it remains to use (4.4) *again.*

REMARK 4.9. *Let X_t be a solution to* (1.1)*, where σ and b satisfy the assumption of Theorem* 1.2 *with $M = 1$ and $f \in C^1_b(\mathbb{R})$, $2/\alpha^+ < 3/\alpha^- - 1$. Let $T > 0$ fixed, $\tau = \inf_{t \le 0}\{X_t = 0\}$, and $\tau^n = \inf_{t \le 0}\{X^n_t = 0\}$. Now, the error taken at the stopping time $T \wedge \tau$ equals*

$$\begin{aligned}|\mathbb{E}[f(X_{\tau \wedge T})] - \mathbb{E}[f(X^n_{\tau \wedge T})]| \le{}& |\mathbb{E}[f(X_{\tau \wedge T})] - \mathbb{E}[f(X^n_{\tau \wedge T})]| \\ &+ |\mathbb{E}[f(X^n_{\tau \wedge T})] - \mathbb{E}[f(X^n_{\tau^n \wedge T})]|.\end{aligned}$$

Replacing the first summand we can write

$$|E1| = \left|\sum_{k=1}^m \mathbb{E}_{x_0}\left[u\left(X_{\frac{k}{n} \wedge \tau}, \frac{k}{n} \wedge \tau\right) - u\left(X_{\frac{k-1}{n} \wedge \tau}, \frac{k-1}{n} \wedge \tau\right)\right]\right| =: \left|\sum_{k=1}^m \mathbb{E}_{x_0}[\delta^n_k]\right|,$$

and therefore we have to show that

$$|\mathbb{E}[\delta^n_1]| \le \int_0^{\frac{1}{n} \wedge \tau} \int_0^s \mathbb{E}\left[\sum_{i=2}^4 C_i H^i[Y^2]_t + C\right] dr\, ds$$

is bounded. But in Remark 4.3 *we have seen that $H^i[Y^2]_t$, $i = 2, 3$, is uniformly bounded in $t \in [0, T]$ for $Y = b(X) - b(x_0)$ or $Y = R(u, X) - R(u, x_0)$, respectively (see Remarks* 4.2 *and* 4.4*). It remains to investigate $E2$. Suppose that $F : \mathbb{R} \to \mathbb{R}$ is Lipschitz with $F(x) > 0$ for $x > 0$ and $F(x) = 0$ for $x \le 0$. By the same arguments as Gobet* [10]*, we can conclude for the second summand $S2$ that*

$$|S2| \le |f|_\infty \int_0^T |\mathbb{P}(t < \tau) - \mathbb{P}(t < \tau^n)|\, dt \;=\; |f|_\infty T\, |\mathbb{1}_{T<\tau} - \mathbb{1}_{t<\tau^n}|.$$

Continuing we get

$$\begin{aligned}|\mathbb{1}_{T<\tau} - \mathbb{1}_{t<\tau^n}| = {}& \big(\mathbb{1}_{\inf_{t\in[0,T]} F(X_t)=0} \mathbb{1}_{\inf_{t\in[0,T]} F(X^n_t)>0} + \mathbb{1}_{\inf_{t\in[0,T]} F(X_t)>0} \\ & \mathbb{1}_{\inf_{t\in[0,T]} F(X^n_t)=0}\big) \cdot \big(\mathbb{1}_{\inf_{t\in[0,T]} F(X_t)>\delta} + \mathbb{1}_{\inf_{t\in[0,T]} F(X_t)\le\delta}\big) \times \big(\mathbb{1}_{\inf_{t\in[0,T]} F(X^n_t)>\delta} \\ & + \mathbb{1}_{\inf_{t\in[0,T]} F(X^n_t)\le\delta}\big) \le |\mathbb{1}_{|\inf_{t\in[0,T]} F(X_t) - \inf_{t\in[0,T]} F(X^n_t)|>\delta}\end{aligned}$$

$$+ \mathbb{1}_{\inf_{t\in[0,T]} F(X_t)=0} \cdot \mathbb{1}_{0<\inf_{t\in[0,T]} F(X_t^n)\leq\delta} + \mathbb{1}_{\inf_{t\in[0,T]} F(X_t^n)=0} \cdot \mathbb{1}_{0<\inf_{t\in[0,T]} F(X_t)\leq\delta}$$
$$\leq \mathbb{P}\left(\left|\inf_{t\in[0,T]} F(X_t) - \inf_{t\in[0,T]} F(X_t^n)\right| > \delta\right) + \mathbb{P}\left(0 < \inf_{t\in[0,T]} F(X_t) \leq \delta\right)$$
$$+ \mathbb{P}\left(0 < \inf_{t\in[0,T]} F(X_t^n) \leq \delta\right).$$

Using the Chebyshev inequality and Remark 4.8 *we get*

$$\mathbb{P}\left(\left|\inf_{t\in[0,T]} F(X_t) - \inf_{t\in[0,T]} F(X_t^n)\right| > \delta\right) \leq \mathbb{P}\left(\inf_{t\in[0,T]} |F(X_t) - F(X_t^n)| > \delta\right)$$
$$\leq \frac{1}{\delta} \cdot \|F(X_t) - F(X_t^n)\|_2^2 \leq \frac{1}{\delta} \cdot \frac{1}{n}.$$

Moreover, since X_t *has a continuous density (see Bass and Cranston* [3, *Theorem* 4.1]*) we have* $\mathbb{P}(0 < \inf_{t\in[0,T]} F(X_t) \leq \delta) \leq C\delta$. *Set* $\delta \sim n^{-1/2}$. *Collecting all together, we get*

$$|\mathbb{E}[f(X_{\tau\wedge T})] - \mathbb{E}[f(X^n_{\tau\wedge T})]| \leq C_1(T)\,\frac{1}{\sqrt{n}} + |f'|_\infty\, C_2(T)\,\frac{M^{15}\exp(M^2)}{n}.$$

REMARK 4.10. *If an interval* I *is bounded, the jumps of an* α*-stable process are bounded by the length of* I *and we can apply Remark* 1.1*, setting* $M = |I|$ *to handle the quality of approximation for the first exit time.*

REMARK 4.11. *If* X_t *is driven by an* α*-stable process, the solution* X_t *does not belong to* $\mathcal{L}^2$ *in general. But, e.g., if* σ *is flat of order* q *at infinity, i.e.,* $\lim_{x\to\infty}\sigma(x,z)/x^q < C|z|^{-\frac{1}{\alpha}}$*, the solution* X *belongs to* $\cap_{p<\alpha(q+1)}\mathcal{L}^p$. *Thus, if* σ *is flat at infinity of all orders, the stochastic exponential belongs to* $\mathcal{L}^p$, $p \geq 1$, *and we do not need to truncate the driving process before applying Theorem* 1.2.

REFERENCES

[1] V. BALLY AND D. TALAY, *The law of the Euler schema for stochastic differential equations.* I. *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[2] V. BALLY AND D. TALAY, *The law of the Euler scheme for stochastic differential equations.* II. *Convergence rate of the density*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

[3] R.F. BASS AND M. CRANSTON, *The Malliavin calculus for pure jump processes and applications to local time*, Ann. Probab., 14 (1986), pp. 490–532.

[4] J. BERTOIN, *Lévy Processes*, Cambridge Tracts in Math. 121, Cambridge University Press, Cambridge, UK, 1996.

[5] K. BICHTELER, J. GRAVEREAUX, AND J. JACOD, *Malliavin Calculus for Processes with Jumps*, Vol. 2, Gordon and Breach, New York, 1987.

[6] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.

[7] E. CINLAR AND J. JACOD, *Representation of semimartingale Markov processes in terms of Wiener processes and Poisson random measures*, in Seminar on Stochastic Processes, Birkhäuser, Boston, 1981, pp. 159–242.

[8] E. CINLAR, J. JACOD, P. PROTTER, AND M.J. SHARPE, *Semimartingales and Markov processes*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 161–219.

[9] I.I. GIHMAN AND A.V. SKOROKHOD, *Stochastic Differential Equations*, K. Wickwire, trans., Springer-Verlag, New York, 1972.

[10] E. GOBET, *Weak approximation of killed diffusion*, Stochastic Process. Appl., 87 (2000), pp. 167–197.

[11] M.T. BARLOW, S.D. JACKA, AND M. YOR, *Inequalities for a pair of processes stopped at a random time*, Proc. London Math. Soc. (3), 52 (1986), pp. 142–172.

[12] S. KANAGAWA, *The rate of convergence for approximate solutions of stochastic differential equations*, Tokyo J. Math., 12 (1989), pp. 33–48.

[13] A. Kohatsu-Higa and P. Protter, *The Euler scheme for SDE's driven by semimartingales*, in Stochastic Analysis on Infinite Dimensional Spaces, Pitman Res. Notes Math. Ser. 310, Longman Scientific Technical, Harlow, UK, 1994, pp. 141–151.

[14] T. Kurtz and P. Protter, *Wong-Zakai corrections, random evolutions, and simulation schemes for SDE's*, in Stochastic Analysis, Academic Press, Boston, 1991, pp. 331–346.

[15] P. Protter, *Stochastic Integration and Differential Equations. A New Approach*, Vol. 21, Springer-Verlag, Berlin, 1990.

[16] P. Protter and D. Talay, *The Euler scheme for Lévy driven stochastic differential equations*, Ann. Probab., 25 (1997), pp. 393–423.

[17] D. Talay, *Elements of probabilistic numerical methods for partial differential equations*, in Probabilistic Models for Nonlinear Partial Differential Equations, Lectures Notes in Math. 1627, D. Talay and L. Tubaro, eds., Springer-Verlag, Berlin, 1996, pp. 148–196.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 114–133

THE ZEROS OF SPECIAL FUNCTIONS FROM A FIXED POINT METHOD*

JAVIER SEGURA†

Abstract. A scheme for the computation of the zeros of special functions and orthogonal polynomials is developed. We study the structure of the first order difference-differential equations (DDEs) satisfied by two fundamental sets of solutions of second order ODEs $y_n''(x)+A_n(x)y_n(x)=0$, n being the order of the solutions and $A_n(x)$ a family of continuous functions. It is proved that, with a convenient normalization of the solutions, $T_{\pm1}(z)=z\pm\operatorname{sign}(d)\arctan(y_n(x(z))/y_{n\pm1}(x(z)))$ are globally convergent iterations with fixed points $z(x_n^{(i)})$, $x_n^{(i)}$ being the zeros of $y_n(x)$; d is one of the coefficients in the DDEs and $z(x)$ is a primitive of d. The structure of the DDEs is also used to set global bounds on the differences between adjacent zeros of functions of consecutive orders and to find iteration steps which guarantee that all the zeros inside a given interval can be found with certainty. As an illustration, we describe how to implement this scheme for the calculation of the zeros of arbitrary solutions of the Bessel, Coulomb, Legendre, Hermite, and Laguerre equations.

Key words. special functions, zeros, fixed point iteration, second order ODE, recurrence relations

AMS subject classifications. 33XX, 65H05

PII. S0036142901387385

1. Introduction.

The literature concerning the evaluation of the real zeros of special functions is extensive, with special emphasis on the zeros of Bessel functions [2, 11, 20, 21, 23, 27, 28, 24]. There exist several algorithms with different characteristics [13, 22, 28] and a great variety of papers dealing with properties of the zeros of cylinder functions (see, for instance, [3, 5, 9, 17]). However, little is known regarding the real zeros of other special functions. Methods for the computation of the zeros of first kind Bessel functions [2, 10, 20, 21, 22, 24, 28], second kind Bessel functions [20, 22, 24, 28], Airy functions [26, 22, 6], and regular Coulomb wave functions [10, 2] have been described. Only matrix methods [10] have been applied for finding zeros of special functions other than Bessel or related functions (Airy functions).

The main task when evaluating zeros of a given function is usually the bracketing of the roots (with the exception of matrix eigenvalue methods [10]). After this, bisection or other standard methods can be used. However, the distribution of zeros changes drastically from one family of functions to another and from one set of values or parameters (like the order of Bessel functions) to another. In addition, the distribution of zeros may be highly nonuniform. Thus, bracketing is generally a difficult issue which requires, for instance, the use of the concept of topological degree [12]. Bracketing can be avoided if initial approximations for the roots [24, 20, 19] are available which allow local methods (such as the Newton–Raphson method) to converge.

We will build a single and simple numerical method which can be applied to all the previously mentioned functions as well as to more general cases. No bracketing is needed; instead we provide iterative relations between consecutive zeros. The method

*Received by the editors April 5, 2001; accepted for publication (in revised form) December 19, 2001; published electronically April 12, 2002. This research was supported in part by Generalitat Valenciana under project GV99-146-1-04.

http://www.siam.org/journals/sinum/40-1/38738.html

†Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911-Leganés, Madrid, Spain (jsegura@math.uc3m.es).

applies for special functions (and orthogonal polynomials) which are solutions of a second order ODE and satisfy systems of difference-differential equations (DDEs) of the type

$$\begin{aligned} y_n'(x) &= a_n(x)y_n(x) + d_n(x)y_{n-1}(x), \\ y_{n-1}'(x) &= b_n(x)y_{n-1}(x) + e_n(x)y_n(x), \end{aligned} \tag{1.1}$$

with continuous coefficients a_n, b_n, d_n, e_n. The method is able to find the zeros of any solution of the Bessel equation (including Airy functions), of any Coulomb wave function (not necessarily the regular Coulomb wave function, as matrix methods require [10, 2]), of any Legendre function, and so on.

Our method makes use of information regarding the distribution of the zeros that is carried by the coefficients of the DDEs. The main ingredients of the method consist of globally convergent fixed point iterations (FPIs) together with prescriptions, based on global bounds on the distance between adjacent zeros, to compute with certainty all the zeros inside a given interval. These global bounds on differences of consecutive and adjacent zeros apply to a broad family of special functions and orthogonal polynomials.

The structure of the paper is as follows. In section 2 we describe the properties of the coefficients in the DDEs, satisfied by the solutions of a second order ODE $y_k'' + A_k(x)y_k = 0$, under the assumption that such coefficients are continuous functions. The DDEs relating y_n' and y_{n-1}' with y_n and y_{n-1} for two sets of fundamental solutions exist and are unique [15]. The properties shown will be thus quite general. With these properties, we will show (section 3) that the ratios $y_n(x(z))/y_{n\pm1}(x(z))$ satisfy nonlinear first order ODEs resembling the equation for the function $\tan(z(x))$; $z = z(x)$ is a change of variables given by a primitive of one of the coefficients of the DDEs. We will use this similarity to build a globally convergent FPI in section 4. This FPI, together with bounds on the distance between the zeros of y_n and the adjacent zeros of $y_{n\pm1}$, will be used to build a global method to find with certainty all the roots inside a given interval (section 5). The method is monotonically convergent except for two possible exceptions which are treated in section 6. In section 7, we give explicit algorithms to compute with certainty all the zeros inside a given interval. Finally (section 8), we compile the analytical information needed to apply the method to arbitrary solutions of the Bessel, Coulomb, Legendre, Hermite, and Laguerre differential equations.

2. Structure of the DDEs. We will consider the solutions of second order ODEs in normal form,

$$y_k''(x) + A_k(x)y_k(x) = 0, \tag{2.1}$$

which are defined for all x in an interval I, where $A_k(x)$ is a family of continuous functions. The solutions $y_k(x)$ have continuous second derivative in I. This is a common situation: most special functions and orthogonal polynomials satisfy second order differential equations, reducible to the normal form (2.1), in which $A_k(z)$ is an analytic function in the complex plane except for a few singularities. The results in this paper apply to real intervals free of singularities of A_k.

Given a solution corresponding to a given value $k = n$, y_n, we will say that n is the order of the solution.

Let $\{y_k^{(1)}(x)\}$ and $\{y_k^{(2)}(x)\}$ be two families of independent solutions of (2.1) in I, i.e., two families of solutions with Wronskians different from zero,

$$W[y_k^{(1)}, y_k^{(2)}] = y_k^{(1)} y_k^{(2)\prime} - y_k^{(1)\prime} y_k^{(2)} \equiv w_k \neq 0. \tag{2.2}$$

It is obvious from (2.1) that the Wronskians are constant.

Following [15, Thm. 1], it can be proved that for every fixed pair of sequences $\{y_k^{(1)}(x)\}$, $\{y_k^{(2)}(x)\}$ of fundamental solutions of (2.1), there exists a unique pair of DDEs, relating y_k with y_{k-1} and its first derivative and y_{k-1} with y_k and its first derivative, which is valid for both fundamental sets of solutions (and hence for any linear combination of them with constant coefficients).

Let us consider two orders $k = n, n-1$ and write these relations in the form

$$\begin{aligned} y_n'(x) &= a_n(x) y_n(x) + d_n(x) y_{n-1}(x), \\ y_{n-1}'(x) &= b_n(x) y_{n-1}(x) + e_n(x) y_n(x). \end{aligned} \tag{2.3}$$

Let us now denote by $Z_n(x)$ the determinant

$$Z_n(x) = \begin{vmatrix} y_n^{(1)} & y_{n-1}^{(1)} \\ y_n^{(2)} & y_{n-1}^{(2)} \end{vmatrix}. \tag{2.4}$$

It is easy to prove that $Z_n(x)$ cannot be identically zero [15, Thm. 1]; see also [14]. By expressing the coefficients a_n, b_n, d_n, e_n in terms of $y_n^{(1)}$, $y_n^{(2)}$, $y_{n-1}^{(1)}$, $y_{n-1}^{(2)}$, and their derivatives, the following result can be easily proved.

LEMMA 2.1. *Let relations* (2.3) *be satisfied by solutions* $\{y_n^{(1)}, y_{n-1}^{(1)}\}$ *and* $\{y_n^{(2)}, y_{n-1}^{(2)}\}$, *where* $\{y_n^{(1)}, y_n^{(2)}\}$ *and* $\{y_{n-1}^{(1)}, y_{n-1}^{(2)}\}$ *are independent solutions of second order ODEs in normal form. Then*

1. $d_n(x) \neq 0$, $e_n(x) \neq 0$.
2. $d_n(x)/e_n(x) \equiv c_n \neq 0$ *is a constant.*
3. $a_n(x)$,$b_n(x)$,$d_n(x)$,$e_n(x)$ *are continuous* $\iff$ $a_n(x)$,$b_n(x)$,$d_n(x)$,$e_n(x)$ *are differentiable* $\iff$ $Z_n(x) \neq 0\ \forall x \in I$.

Proof. Writing the first equation of (2.3) for $\{y_n^{(1)}, y_{n-1}^{(1)}\}$ and $\{y_n^{(2)}, y_{n-1}^{(2)}\}$ and solving the resulting system, we get $Z_n(x) d_n(x) = w_n$, w_n being the Wronskian (2.2). Similarly, from the second equation in (2.3) we get $Z_n(x) e_n(x) = -w_{n-1}$. Considering these relations, the first two properties follow because the solutions of (2.1) and their first derivatives are continuous in I and the Wronskians are constant and different from zero. It is also evident that d_n and e_n are continuous in I if and only if $Z_n(x)$ does not vanish in I; obviously, the same is true for a_n and b_n. In addition, the coefficients are differentiable when $Z_n(x) \neq 0$, because the solutions of the differential equation are at least twice differentiable in I. ∎

By taking into account that both sets of solutions satisfy the second order ODE (2.1) we can obtain relations between the coefficients appearing in relations (2.3).

LEMMA 2.2. *Let relations* (2.3), *with continuous coefficients in* I, *be satisfied by solutions* $\{y_n^{(1)}, y_{n-1}^{(1)}\}$ *and* $\{y_n^{(2)}, y_{n-1}^{(2)}\}$, *where* $\{y_n^{(1)}, y_n^{(2)}\}$ *and* $\{y_{n-1}^{(1)}, y_{n-1}^{(2)}\}$ *are independent solutions of second order ODEs in normal form* (2.1). *Then, the following relations hold:*

1. $d_n/e_n = c_n \neq 0$ *with* c_n *constant and* $d_n(x) \neq 0\, \forall x$.
2. $a_n + b_n = -d_n'/d_n$.

3. $W[d_n, g_n] \equiv d_n g_n' - d_n' g_n = (A_n - A_{n-1}) d_n$, *where* $g_n \equiv b_n - a_n$.
4. $a_n' + a_n^2 = -A_n - e_n d_n$

Conversely, given two DDEs verifying relations 1, 2, 3, *and* 4 *for all* n, *there exist two sets of independent solutions* $\{y_k^{(1)}\}$, $\{y_k^{(2)}\}$ *which verify the DDEs and are solutions of the differential equation* (2.1) *for each* k.

Proof. The first relation was already shown in Lemma 2.1. Taking the derivative of the first DDE and using $y_{n-1} = \frac{1}{d_n}(y_n' - a_n y_n)$ to eliminate y_{n-1} (recall that $d_n \neq 0$), we find

$$y_n'' = \left[a_n + \frac{d_n'}{d_n} + b_n\right] y_n' + \left[a_n' + d_n e_n - a_n \frac{d_n'}{d_n} - a_n b_n\right] y_n,$$

and given that y_n satisfies $y_n'' + A_n y_n = 0$, by subtracting we have

$$\left[a_n + \frac{d_n'}{d_n} + b_n\right] y_n' + \left[a_n' + d_n e_n - a_n \frac{d_n'}{d_n} - a_n b_n + A_n\right] y_n = 0.$$

We demand that this equation be satisfied by two independent functions $y_n^{(1)}$ and $y_n^{(2)}$, and therefore

$$\begin{aligned} &a_n + b_n + \frac{d_n'}{d_n} = 0, \\ &a_n' + d_n e_n - a_n \frac{d_n'}{d_n} - a_n b_n + A_n = 0. \end{aligned} \tag{2.5}$$

Proceeding in a similar way, starting from the second equation in (2.3), we find

$$\begin{aligned} &a_n + b_n + \frac{e_n'}{e_n} = 0, \\ &b_n' + d_n e_n - b_n \frac{e_n'}{e_n} - a_n b_n + A_{n-1} = 0. \end{aligned} \tag{2.6}$$

Subtracting the first equations in (2.5) and (2.6), we obtain

$$\frac{d_n'}{d_n} - \frac{e_n'}{e_n} = 0, \tag{2.7}$$

which implies, as already known, that d_n/e_n is a constant (and d_n and e_n never vanish, as we know from Lemma 2.1).

Combining the two equations in (2.5), we have

$$a_n' + a_n^2 = -A_n - e_n d_n, \tag{2.8}$$

and proceeding in a similar way with (2.6), obtain

$$b_n' + b_n^2 = -A_{n-1} - e_n d_n. \tag{2.9}$$

Subtracting these last two equations and using the first equation in (2.5), we get the third equation of the theorem.

Conversely, given the two DDEs with coefficients satisfying relations 1, 2, 3, and 4, the sets of functions satisfying such DDEs automatically satisfy second order ODEs in normal form. Furthermore, one can always generate two sets of fundamental solutions by iterative application of the DDEs: take two sets of independent solutions of the ODE for one order (let's say $y_{m-1}^{(1)}$ and $y_{m-1}^{(2)}$) and use the second DDE to obtain $y_m^{(1)}$

and $y_m^{(2)}$, which necessarily satisfy the first DDE on account of the relations between coefficients. $y_m^{(1)}$ and $y_m^{(2)}$ are necessarily independent because $d_m/e_m = -w_m/w_{m-1}$, $d_m \neq 0$, $e_m \neq 0$, and $w_{m-1} = W[y_{m-1}^{(1)}, y_{m-1}^{(2)}] \neq 0$; this implies that $w_m \neq 0$. □

Further properties of the coefficients of the DDEs are obtained in Corollary 2.3 and Lemma 2.4.

COROLLARY 2.3. *Let A_n, A_{n-1}, a_n, b_n, d_n, and e_n be continuous in I. If $A_n(x) - A_{n-1}(x) \neq 0\, \forall x \in I$, then $a_n - b_n$ can have only one zero in I.*

Proof. Let us define the function $f(x) = (b_n(x) - a_n(x))/d_n(x)$, which is differentiable in I. By taking the derivative and considering relation 3 of Lemma 2.2 we get $f' = (A_n - A_{n-1})/d_n$.

Thus, f is monotonic in I because neither d_n nor $A_n - A_{n-1}$ change sign in I. Therefore, f can have no more than one zero in I. □

An additional property satisfied by the coefficients is the fact that the coefficients d_n and e_n have opposite signs. This property is a key ingredient of the global method that we will describe.

LEMMA 2.4. *Let y_n and y_{n-1} be two nontrivial solutions of* (2.1), *for $k = n$ and $k = n-1$, such that they satisfy* (2.3) *with continuous d_n. If one of these functions has, at least, two zeros in I, then*

1. *the zeros of y_n, y_{n-1} are simple.*
2. *y_n and y_{n-1} cannot vanish simultaneously.*
3. *the zeros of y_n and y_{n-1} are interlaced.*
4. *$e_n(x)d_n(x) < 0\ \forall x \in I$.*

Proof. 1. This is an immediate consequence of the existence and uniqueness theorem for linear homogeneous ODEs.

2. If $y_n(x_0) = y_{n-1}(x_0) = 0$, then by virtue of the DDEs and the continuity of the coefficients, $y_n'(x_0) = y_{n-1}'(x_0) = 0$. Then both solutions would be trivial: $y_n(x) = y_{n-1}(x) = 0$.

3. Let x_1, x_2 be two consecutive zeros of y_n. Given that y_n is differentiable, we have $y_n'(x_1)y_n'(x_2) < 0$. By virtue of the first DDE (2.3) and the fact that $d_n(x)$ does not change sign, we get $y_{n-1}(x_1)y_{n-1}(x_2) < 0$; however, y_{n-1} is continuous and therefore there is at least one $\bar{x}_1 \in (x_1, x_2)$ such that $y_{n-1}(\bar{x}_1) = 0$. Since we can prove in a similar way (using the second difference-differential relation) that between two zeros of y_{n-1} there is at least one zero of y_n, then there can be no more zeros of y_{n-1} in (x_1, x_2): if there were two zeros of y_{n-1} in (x_1, x_2), then there would be a zero of y_n in this interval, and x_1 and x_2 would not be consecutive zeros, in contrast to our hypothesis.

This proves that between two zeros of y_n there is exactly one zero of y_{n-1}. In the same way, between two zeros of y_{n-1} there is exactly one zero of y_n.

4. Let x_1, x_2, and $\bar{x}_1$ be defined as before. Let us take $y_n(\bar{x}_1) > 0$ (for $y_n(\bar{x}_1) < 0$ the proof is analogous). This assumption implies that $y_n'(x_1) > 0$ and $y_n'(x_2) < 0$ because the zeros are simple. Considering the first DDE, we get $\operatorname{sign}(d_n)y_{n-1}(x_1) > 0$, $\operatorname{sign}(d_n)y_{n-1}(x_2) < 0$, and so

$$\operatorname{sign}(d_n)y_{n-1}'(\bar{x}_1) < 0.$$

Considering the second DDE and given that we assumed that $\operatorname{sign}(y_n(\bar{x}_1)) > 0$, we have

$$\operatorname{sign}(y_{n-1}'(\bar{x}_1)) = \operatorname{sign}(e_n).$$

Therefore, $e_n d_n < 0$. □

The following theorem summarizes the main results of this section.

THEOREM 2.5. *Let $y_k'' + A_k(x)y_k(x) = 0$, where $A_k(x)$ are continuous functions in I. Let $\{y_n^{(1)}, y_n^{(2)}\}$,$\{y_{n-1}^{(1)}, y_{n-1}^{(2)}\}$ be two sets of independent solutions in I for $k = n$ and $k = n-1$. Then, both sets $\{y_n^{(1)}, y_{n-1}^{(1)}\}$ and $\{y_n^{(2)}, y_{n-1}^{(2)}\}$ satisfy the relations*

$$\begin{aligned} y_n'(x) &= a_n(x)y_n(x) + d_n(x)y_{n-1}(x), \\ y_{n-1}'(x) &= b_n(x)y_{n-1}(x) + e_n(x)y_n(x), \end{aligned} \tag{2.10}$$

where the coefficients $a_n(x)$, $b_n(x)$, $d_n(x)$, and $e_n(x)$ have the same discontinuities in I. If these coefficients are continuous, then the following properties hold:

1. *The coefficients are differentiable in I.*
2. *$d_n(x) \neq 0\ \forall x \in I$, $d_n/e_n = c_n \neq 0$, c_n constant.*
3. *$a_n + b_n = -d_n'/d_n$.*
4. *$W[d_n, g_n] \equiv d_n g_n' - d_n' g_n = (A_n - A_{n-1})d_n$, where $g_n \equiv b_n - a_n$.*
5. *$a_n' + a_n^2 = -A_n - e_n d_n$.*
6. *If at least one of the solutions (for n or $n-1$) has two (or more) zeros in I, we can set $d_n(x) = -e_n(x)$ after a convenient normalization of the solutions.*
7. *If $A_n(x) - A_{n-1}(x)$ does not change sign in I, then $a_n(x) - b_n(x)$ can have only one zero in I.*

Conversely, given two DDEs (2.3) *verifying the first* 5 *conditions for all n, there exist two sets of fundamental solutions $\{y_k^{(1)}\}$, $\{y_k^{(2)}\}$ which are solutions of the differential equation* (2.1) *for each k.*

Proof. It is known that the DDEs relating y_n' and y_{n-1}' with y_n and y_{n-1} for two sets of fundamental solutions of the ODE exist and are unique [15]. The rest of the theorem, except property 6, is a compilation of the results presented in Lemmas 2.1, 2.2, 2.4, and Corollary 2.3.

The only thing left to prove is that from the initial fundamental sets of solutions, one of them having at least two zeros in I, one can build other fundamental sets such that $d_n = -e_n$. Note that $e_n d_n < 0$ (Lemma 2.4) and that d_n/e_n is a constant. Therefore, the relation $d_n = -e_n$ can be indeed trivially accomplished by renormalizing the solutions through constant factors. Of course, such a renormalization does not change the a_n and b_n coefficients or any other result in this theorem and the previous ones. ∎

Although we have considered equations in normal form so far, these results apply more generally. The following result for ODEs in canonical form can be easily proved.

COROLLARY 2.6. *Let $y_n''(x) + B(x)y_n'(x) + A_n(x)y_n(x) = 0$ with $A_n(x)$ continuous and $B(x)$ not depending on n and with continuous derivative in I. If the coefficients of the DDEs are continuous in I, then the following properties hold:*

1. *The coefficients of the DDEs are differentiable in I.*
2. *$d_n(x) \neq 0\ \forall x \in I$, $d_n/e_n = c_n \neq 0$, c_n constant.*
3. *If there is a solution of order n or $n-1$ with at least two zeros in I, then $d_n(x) = -e_n(x)$ with a convenient normalization of the solutions.*
4. *$W[d_n, g_n] \equiv d_n g_n' - d_n' g_n = (A_n - A_{n-1})d_n$ being $g_n \equiv b_n - a_n$.*
5. *If $A_n(x) - A_{n-1}(x)$ does not change sign in I, then $a_n(x) - b_n(x)$ can have only one zero in I.*

Proof. Transform the ODE to the normal form by the change $y(x) = \nu(x)\bar{y}(x)$, $\nu(x) = \exp(-\frac{1}{2}\int B(x)dx)$. In this way,

$$\bar{y}_n'' + \bar{A}_n \bar{y}_n = 0 \text{ with } \bar{A}_n = A_n - B^2/4 - B'/2,$$

and $\bar{A}_n$ is continuous in I. Then apply Theorem 2.5. □

The rest of the properties in Theorem 2.5 must be slightly modified but the main properties that will be used in what follows (2, 3, and 5 of Corollary 2.6) remain invariant. Hence, the methods that we will develop can be directly applied for solutions of these types of equations.

On the other hand, if $B(x)$ also depends on the order n, these results are no longer valid. For instance, d_n/e_n is not necessarily a constant. However, one can transform the ODEs to the normal form by applying changes of the dependent variables depending on the order. We would write $y_k = \nu_k(x)\bar{y}_k$, $k = n, n-1$, where $\nu_k = \exp\left(-\frac{1}{2}\int B_k(x)dx\right)$ and the DDEs for the functions $\{\bar{y}_k\}$, with the same zeros as the functions $\{y_k\}$, satisfy all the relations between coefficients of Theorem 2.5.

3. Ratios of consecutive functions and the change of variables $z(x)$. From now on, we will consider functions normalized in such a way that $d_n = -e_n$ (Theorem 2.5).

We define

$$H_n(x) = \text{sign}(d_n)y_n/y_{n-1}. \tag{3.1}$$

$H_n(x)$ has the same zeros as y_n because the zeros of y_n and y_{n-1} are interlaced. This ratio will be the basis of our root-finding scheme.

Taking the derivative and using the DDEs, we have

$$H_n'(x) = \text{sign}(d_n)\left(\frac{y_n'}{y_{n-1}} - \frac{y_n}{y_{n-1}}\frac{y_{n-1}'}{y_{n-1}}\right) = |d_n| + (a_n - b_n)H_n + |d_n|H_n^2 \ .$$

If relations (2.3), with the conditions of Theorem 2.5, hold with the replacement $n \to n+1$, which is the usual situation, we can write

$$\begin{aligned} y_{n+1}'(x) &= a_{n+1}(x)y_{n+1}(x) + d_{n+1}(x)y_n(x), \\ y_n'(x) &= b_{n+1}(x)y_n(x) - d_{n+1}(x)y_{n+1}(x), \end{aligned} \tag{3.2}$$

and we can define a second ratio with the same roots as y_n:

$$H_{n,+1}(x) = -\text{sign}(d_{n+1})\frac{y_n(x)}{y_{n+1}(x)} \ . \tag{3.3}$$

Taking the derivative, we obtain

$$H_{n,+1}'(x) = |d_{n+1}|(1 + H_{n,+1}^2) - (a_{n+1} - b_{n+1})H_{n,+1}. \tag{3.4}$$

In a more compact form, we define

$$H_{n,i}(x) = -i\text{sign}(d)\frac{y_n}{y_{n+i}} \tag{3.5}$$

and we have that

$$H_{n,i}'(x) = |d|(1 + H_{n,i}^2) + i(b - a)H_{n,i}, \tag{3.6}$$

where $i = \pm 1$, and the orders of the coefficients a, b, and d are n for $i = -1$ and $n+1$ for $i = +1$.

We observe that the functions $H_{n,i}(x)$ satisfy the same differential equation as $\tan x$ if $|d| = 1$ and if the linear term is missing. This linear term prevents the $H_{n,i}$ functions from being monotonic in general. Monotonic functions (except at the zeros of y_{n+i}) can be obtained from the functions $H_{n,i}$, as similarly shown in [21], by exponentiation of the coefficient of the linear term:

$$f_{n,i}(x) = \exp\left(i \int (a-b)dx\right) H_{n,i}(x).$$

These functions can be used to build globally convergent Newton iterations. However, we will see how the properties satisfied by the function $H_{n,i}$ are enough to find a fixed point method which implies, as a weaker result, the global convergence of the Newton iterations based on $f_{n,i}$.

Let us introduce the changes of variables (one different change for each value of i):

$$z(x) = \int |d(x)|\, dx. \tag{3.7}$$

Denoting

$$H_i(z) \equiv H_{n,i}(x(z)) \tag{3.8}$$

and $\frac{dH_i}{dz} \equiv \dot{H}_i$, we find that

$$\dot{H}_i = (1 + H_i^2) - 2\eta_i H_i, \tag{3.9}$$

where

$$\eta_i = i\frac{a-b}{2|d|}. \tag{3.10}$$

These expressions are quite similar to those obtained for the particular case of the Bessel functions [21, 23], the main difference being that $(b-a)/d$ may change sign (only once if $A_n - A_{n-1}$ never vanishes), while this phenomenon was absent for Bessel functions. η_i are monotonic functions if $A_n - A_{n-1}$ never cancels (see the proof of Corollary 2.3).

4. Globally convergent fixed point method. The starting point of the fixed point method for the evaluation of the zeros of $Y_n(z) \equiv y_n(x(z))$ is the first derivative of H_i (see (3.9)).

We will discuss the evaluation of the zeros of $Y_n(z) \equiv y_n(x(z))$. Obviously, if z_n is a zero of Y_n, then $z^{-1}(z_n)$ is a zero of y_n and vice versa. We will assume that all hypotheses of Theorem 2.5 are met.

If we were given a guess value z_0 to obtain a zero z_n of Y_n and η_i was identically zero (which is the case for Bessel functions of order $n = 1/2$), the obvious answer would be

$$z_n = T_i(z_0) = z_0 - \arctan(H_i(z_0)).$$

When $\eta_i \neq 0$, we will show that $T_i(z)$ can be used to evaluate zeros by taking successive iterations of this function.

It is important to bear in mind that, if $A_n(x) - A_{n-1}(x) \neq 0 \, \forall x \in I$, η_i can change sign only once in I (see Theorem 2.5, properties 2 and 7, and (3.10)). The value $z \equiv z_\eta$ for which $\eta_i(z_\eta) = 0$ will be called the transition point (TP).

We will first prove the global convergence for the iteration in case η_i does not change sign, and we will consider the other situation later.

Let us first introduce some notation:

1. We denote by Y_k the functions $Y_k(z) \equiv y_k(x(z))$.
2. Let us consider that z_n is a zero of Y_n. We will denote by $z_{n+i}^{(j)}$ the closest zero of Y_{n+i} to z_n which is smaller than z_n (if it exists). Similarly, we denote by $z_{n+i}^{(j+1)}$ the closest zero of Y_{n+i} to z_n which is greater than z_n (if it exists).
3. We will denote by J_i an open interval such that there is exactly one zero of Y_n in J_i ($z_n \in J_i$) and such that J_i is the largest interval in which there are no zeros of Y_{n+i}.

It is important to bear in mind the following result.

LEMMA 4.1. *Let J_i be an interval as defined above. Then H_i is continuous in J_i. If $z \in J_i$, then* $\operatorname{sign}(H_i) = \operatorname{sign}(z - z_n)$. *If $z_{n+i}^{(j)}$ exists, then* $\lim_{z \to z_{n+i}^{(j)+}} H_i = -\infty$. *If $z_{n+i}^{(j+1)}$ exists, then* $\lim_{z \to z_{n+i}^{(j+1)-}} H_i = +\infty$.

Proof. H_i is continuous in J_i because Y_{n+i} does not vanish in I. In addition, (3.9) implies that $\dot{H}_i(z_n) = 1$ because $H_i(z_n) = 0$. Given that the only zero of Y_n in J_i is z_n and H_i is increasing at z_n, then H_i is positive for $z > z_n$ and negative for $z < z_n$ for values of z in J_i. ∎

The following lemma will lead to the definition of subintervals of monotonic convergence.

LEMMA 4.2. *Let $T_i(z) = z - \arctan(H_i(z))$.*

1. *If $\eta_i > 0 \; \forall z_0 \in (z_n, z') \subset J_i$, then $z_n < T_i(z_0) < z_0$ and $\lim_{p\to\infty} T_i^{(p)}(z_0) = z_n$ $\forall z_0 \in (z_n, z')$.*
2. *If $\eta_i < 0 \; \forall z_0 \in (z'', z_n) \subset J_i$, then $z_0 < T_i(z_0) < z_n$ and $\lim_{p\to\infty} T_i^{(p)}(z_0) = z_n \; \forall z_0 \in (z'', z_n)$.*

In both cases the FPI converges monotonically.

Proof. The only fixed point of the iteration T_i in J_i is z_n. Therefore, convergence to z_n is guaranteed once we have proved that the successive iterations of T_i form monotonic sequences which approach z_n and are bounded by z_n.

From (3.9) we have

$$\dot{H}_i = (1 + H_i^2) - 2\eta_i H_i \; . \tag{4.1}$$

Rearranging (4.1) gives

$$\operatorname{sign}(\eta_i)\operatorname{sign}(H_i) \left(\frac{\dot{H}_i}{1 + H_i^2} - 1 \right) = -2|\eta_i| \frac{|H_i|}{1 + H_i^2} \leq 0, \tag{4.2}$$

where the equality holds only for $z = z_n$.

Now, since $\operatorname{sign}(H_i(z)) = \operatorname{sign}(z - z_n) \; \forall z \in J_i$ (see Lemma 4.1),

$$\operatorname{sign}(\eta_i) \int_z^{z'} \left(\frac{\dot{H}_i}{1 + H_i^2} - 1 \right) dz < 0 \tag{4.3}$$

for all $z, z' \in J_i$ such that $z_n \leq z < z'$ or $z_n \geq z > z'$.

Integrating, we obtain

$$\operatorname{sign}(\eta_i)(\arctan(H_i(z')) - \arctan(H_i(z)) - (z' - z)) < 0, \tag{4.4}$$

and taking $z = z_n$ gives

$$\operatorname{sign}(\eta_i)(\arctan(H_i(z')) - (z' - z_n)) < 0, \qquad z' \in J_i \setminus \{z_n\}. \tag{4.5}$$

Therefore, if $\eta_i > 0$ $\forall z \in (z_n, z') \equiv K \subset J_i$, then we have that $H_i(z) > 0$ in K (Lemma 4.1), and given a value $z_0 \in K$, we have $T_i(z_0) = z_0 - \arctan(H_i(z_0)) < z_0$; in addition, from (4.5) it follows that $z_n < T_i(z_0)$. This proves the first case of the lemma; the second case can be proved in a similar way. ∎

This lemma suggests the following definition.

DEFINITION 4.3. *When $\eta_i > 0$ in $(z_n, z_{n+i}^{(j+1)}) \subset J_i$ (first case of Lemma 4.2), we say that z_n has a subinterval of monotonic convergence (SMC) on the right. The interval $(z_{n+i}^{(j)}, z_n) \subset J_i$ when $\eta_i > 0$ is called a subinterval of conditioned convergence (SCC).*

z_n has an SMC on the left when $\eta_i < 0$ holds in an interval $(z_{n+i}^{(j)}, z_n)$ (second case of Lemma 4.2). There is an SCC on the right if $\eta_i < 0$ in $(z_n, z_{n+i}^{(j+1)})$.

Lemma 4.2 can be used to show that, whenever $z_{n+i}^{(j)}$ and $z_{n+i}^{(j+1)}$ exist and η_i does not change sign in $(z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$, the FPI converges for any starting value in such an interval. Before this, it is convenient to set bounds on the lengths of SMCs and SCCs.

COROLLARY 4.4 (lengths of SCCs and SMCs). *Let $z_{n+i}^{(j)}$ and $z_{n+i}^{(j+1)}$ be two consecutive zeros of Y_{n+i} and let z_n be the zero of Y_n in $J_i = (z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$. Let $\eta_i \neq 0$ $\forall z \in J_i$. Then*

1. *the length of the SMC is larger than $\pi/2$,*
2. *the length of the SCC is smaller than $\pi/2$.*

Proof. Considering (4.5) and taking into account Lemma 4.1, we have

$$\operatorname{sign}(\eta_i)\operatorname{sign}(z' - z_n)(|\arctan(H_i(z'))| - |z' - z_n|) < 0 \ \forall z' \in J_i \setminus z_n. \tag{4.6}$$

Taking the limit $z' \to z_{n+i}$, where $z_{n+i} = z_{n+i}^{(j)}$ or $z_{n+i} = z_{n+i}^{(j+1)}$, gives

$$\operatorname{sign}(\eta_i)\operatorname{sign}(z_{n+i} - z_n)\left(\frac{\pi}{2} - |z_{n+i} - z_n|\right) < 0. \tag{4.7}$$

However, the values between z_n and z_{n+i} form an SMC if and only if $\operatorname{sign}(\eta_i)\operatorname{sign}(z_{n+i} - z_n) > 0$, and they are an SCC if and only if $\operatorname{sign}(\eta_i)\operatorname{sign}(z_{n+i} - z_n) < 0$ (see Lemma 4.2 and Definition 4.3). ∎

COROLLARY 4.5. *The distance between two consecutive zeros of Y_n or Y_{n+i} is greater than $\pi/2$.*

We can now prove the global convergence in the intervals $(z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$.

THEOREM 4.6. *Let $z_{n+i}^{(j)}$ and $z_{n+i}^{(j+1)}$ be two consecutive zeros of Y_{n+i} and let z_n be the zero of Y_n on $J_i = (z_{n+1}^{(j)}, z_{n+1}^{(j+1)})$. Assume that η_i does not change sign in J_i. Let $T_i(z) = z - \arctan(H_i)$. Then*

$$\lim_{p \to \infty} T_i^{(p)}(z_0) = z_n \ \forall z \in (z_{n+i}^{(j)}, z_{n+i}^{(j+1)}).$$

Proof. Given that η_i does not change sign, there is an SMC on either the left or on the right of z_n (Lemma 4.2). Let us consider that the SMC is on the right, which means that $\eta_i > 0$ (the case $\eta_i < 0$ is proved in a similar way). From (4.5) we have that $z_n < T_i(z) = z - \arctan(H_i(z))\,\forall z \in J_i$. As discussed in Lemma 4.2 this implies monotonic convergence to z_n for all the starting values z_0 in the SMC $(z_n, z_{n+i}^{(j+1)})$. For initial values z_0 in the SCC $(z_{n+i}^{(j)}, z_n)$, given that $z_0 < z_n < z_1 \equiv T_i(z_0) < z_0 + \pi/2$ (see (4.5)), the value of z after the first iteration (z_1) is greater than z_n and in fact lies in the SMC; this is so because the length of the SMC, $z_{n+i}^{(j+1)} - z_n$, is larger than $\pi/2$ while $z_1 - z_0 < \pi/2$. The next iterations starting from z_1 will therefore converge monotonically. ∎

The preceding proof illustrates how the convergence for starting values in an SCC inside an interval J_i is conditioned to the existence of an adjacent SMC in J_i.

Remark. Given that the zeros of Y_n and Y_{n+i} are interlaced, the preceding theorem guarantees (except for a trivial exception) convergence to a zero of Y_n for any starting value in an interval (z_{n+i}, z'_{n+i}), z_{n+i} and z'_{n+i} being any two zeros of Y_{n+i}, not necessarily consecutive. The trivial exception to this result comes from the fact that $H_i(z)$ is not defined at the zeros of Y_{n+i}. However, this "overflow problem" in the evaluation of H_i is easy to handle numerically. Let us avoid this circumstance by defining $T_i(z_{n+i}) = \pi/2$; then, convergence in intervals (z_{n+i}, z'_{n+i}) is global without exception.

The convergence theorems established so far, as well as all the following results, also apply to the Newton iteration based on the monotonic functions f_i. This iteration is given by

$$\hat{T}_i(z) = z - \frac{f_i}{\dot{f}_i} = z - \frac{H_i}{1 + H_i^2} \equiv z - R_i(z).$$

However, $\operatorname{sign}(R_i(z)) = \operatorname{sign}(\arctan(H_i(z)))$ and $|R_i(z)| < |\arctan(H_i(z))|$ and therefore this Newton iteration is also globally convergent.

5. Finding the zeros inside an interval. In order to compute with certainty all the zeros of Y_n inside an interval $I \subset (z_{n+i}, z'_{n+i})$, with z_{n+i}, z'_{n+i} zeros of Y_{n+i}, we need to find a step Δz such that, once a zero of Y_n (z_n) is known, the starting value $z_0 = z_n + \Delta z$ (or $z_0 = z_n - \Delta z$) will lead to convergence to a zero Y_n consecutive to z_n. This is provided by the following proposition.

PROPOSITION 5.1. *Let $I = [z_1, z_2] \subset (z_{n+i}^{(1)}, z_{n+i}^{(M)})$, where $z_{n+i}^{(1)}$, $z_{n+i}^{(2)}$, …, $z_{n+i}^{(M)}$ denote M consecutive zeros of Y_{n+i}; η_i is such that $\eta_i \neq 0\,\forall z \in I$. All the zeros of Y_n in I can be found by applying one of the following algorithms:*

1. *If $\eta_i < 0\,\forall z \in I$, then $z_n^{(1)} = \lim_{p\to\infty} T_i^{(p)}(z_1)$ is a zero of Y_n and the successive consecutive zeros $z_n^{(1)} < z_n^{(2)} \ldots$ are recursively obtained in the following way: $z_n^{(k+1)} = \lim_{p\to\infty} T_i^{(p)}(z_n^{(k)} + \pi/2)$. The convergence to the zeros $z_n^{(k)}$, $k > 1$, is always monotonic. The process is repeated until an m is reached for which $z_n^{(m)} + \pi/2 > z_2$.*
2. *If $\eta_i > 0\,\forall z \in I$, then $z_n^{(1)} = \lim_{p\to\infty} T_i^{(p)}(z_2)$ is a zero of Y_n and the successive consecutive zeros $z_n^{(1)} > z_n^{(2)} \ldots$ are recursively obtained in the following way: $z_n^{(k+1)} = \lim_{p\to\infty} T_i^{(p)}(z_n^{(k)} - \pi/2)$. The convergence to the zeros $z_n^{(k)}$, $k > 1$, is always monotonic. The process is repeated until an m is reached for which $z_n^{(m)} - \pi/2 < z_1$.*

Proof. 1. If $\eta_i < 0$, then the SMCs are on the left of the zeros of Y_n and the SCCs are on the right. By Theorem 4.6, the FPI with starting value z_1 converges to a zero $z_n^{(1)}$ of Y_n.

Given a zero of Y_n, $z_n^{(k)} \in (z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$, the right subinterval $(z_n^{(k)}, z_{n+i}^{(j+1)})$ is an SCC, with length smaller than $\pi/2$, and therefore $z_n^{(k)} + \pi/2$ "jumps" outside this SCC and falls into the next SMC $(z_{n+i}^{(j+1)}, z_n^{(k+1)})$, which guarantees monotonic convergence to the next zero of Y_n, $z_n^{(k+1)}$. This shows that the step $\Delta z = \pi/2$ guarantees convergence to the successive zeros of Y_n.

Because the algorithm always evaluates successive values of z in increasing order, except, perhaps, when calculating the first zero, after $z_n^{(m)} + \pi/2 > z_2$ for some m, the iteration can be stopped since the next zero of Y_n would be larger than z_2.

2. The proof for $\eta_i > 0$ is very similar and it is omitted. □

6. Special cases. In order to circumvent the limitations present in Proposition 5.1 and to thus be able to build a general algorithm which will find all the real roots inside any real interval $[z_1, z_2]$, we must study the convergence of the method when

1. the starting value lies in an interval which is not contained in an interval $(z_{n+i}^{(1)}, z_{n+i}^{(M)})$.
2. η_i changes sign in an interval $(z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$. We will say that $(z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$ is a transition subinterval.

The potential problems will be avoided by choosing starting values which guarantee monotonic convergence.

It is easy to prove the following result regarding convergence in transition subintervals.

PROPOSITION 6.1 (transition subintervals). *Let $J_i \equiv (z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$, $z_{n+i}^{(j)}$ and $z_{n+i}^{(j+1)}$ being two consecutive zeros of Y_{n+i}, and let z_η be the zero of η_i in J_i. If $\lim_{z \to z_{n+i}^{(j)}} \eta_i > 0$ and $\lim_{z \to z_{n+i}^{(j+1)}} \eta_i < 0$, then the following hold:*

1. $\lim_{p \to \infty} T_i^{(p)}(z_\eta) = z_n$, *where z_n is the zero of Y_n inside J_i.*
2. *If $z_\eta \le z_n$, then $z_\eta - \pi/2 < z_{n+i}^{(j)}$ and $z_n + \pi/2 > z_{n+i}^{(j+1)}$.*
3. *If $z_\eta \ge z_n$, then $z_n - \pi/2 < z_{n+i}^{(j)}$ and $z_\eta + \pi/2 > z_{n+i}^{(j+1)}$.*

Proof. 1. η_i is negative between z_η and z_n if $z_\eta < z_n$, while η_i will be positive between z_η and z_n if $z_\eta > z_n$; in both cases, by Lemma 4.2, there is monotonic convergence for all starting values between z_η and z_n (including both values).

2. $\eta_i < 0$ for $z \in (z_n, z_{n+i}^{(j+1)})$; thus, z_n has an SCC on the right and hence $z_n + \pi/2 > z_{n+i}^{(j+1)}$ (Corollary 4.4). On the other hand, $\eta_i > 0$ in $(z_{n+i}^{(j)}, z_\eta)$, where H_i is negative (such as for an SCC); by using similar arguments to those in Corollary 4.4 (applying (4.3)) one can show that $z_\eta - \pi/2 < z_{n+i}^{(j)}$.

3. The proof is analogous to the previous case. □

This result will be applied when η_i is a decreasing function with a zero z_η (recall that η_i is monotonic and cannot have more than one zero when $A_n - A_{n+i} \neq 0\ \forall x$). The algorithm is said to be expansive, because the zeros are evaluated in increasing order on the right of z_η and in decreasing order on the left. This proposition gives a recipe for finding the zero in the transition subinterval and for starting the forward and backward sweeps outside this subinterval.

Regarding convergence in intervals that are not contained in intervals $(z_{n+i}^{(1)}, z_{n+i}^{(M)})$, the most important case is for intervals of the type $[z_1, z_{n+i})$ or $(z_{n+i}, z_2]$. We have the following result.

PROPOSITION 6.2 (first and last subintervals). *Let J_i be an interval where there is no zero of Y_{n+i}.*

1. *If $J_i = [z_1, z_{n+i})$ and $\eta_i < 0$ in J_i, then we have the following:*
   (a) *If $H_i(z_1) > 0$, then there is no zero of Y_n in J_i and $z_1 + \pi/2 > z_{n+i}$.*
   (b) *If $H_i(z_1) < 0$, then there is a zero, z_n, of Y_n in J_i; $\lim_{p\to\infty} T^{(p)}(z_1) = z_n$ (monotonically) and $z_n + \pi/2 > z_{n+i}$.*
2. *If $J_i = (z_{n+i}, z_2]$ and $\eta_i > 0$ in J_i, then we have the following:*
   (a) *If $H_i(z_1) < 0$, then there is no zero of Y_n in J_i and $z_2 - \pi/2 < z_{n+i}$.*
   (b) *If $H_i(z_1) > 0$, then there is a zero, z_n, of Y_n in J_i; $\lim_{p\to\infty} T^{(p)}(z_2) = z_n$ (monotonically) and $z_n - \pi/2 < z_{n+i}$.*

Proof. 1(a) There can be no zero of Y_n in J_i because H_i is increasing at the zeros. Equation (4.4) can be used to prove that $z_1 + \pi/2 > z_{n+i}$.

1(b) There is a zero in J_i as a consequence of Lemma 4.1. Convergence is guaranteed by Lemma 4.2, and Corollary 4.4 proves that $z_n^{(1)} + \pi/2 > z_{n+i}$. (There is an SCC on the right of $z_n^{(1)}$.)

Cases 2(a) and 2(b) can be proved in the same way. □

This result will be useful when evaluating the smallest (largest) zero when $\eta_i < 0$ ($\eta_i > 0$) inside an interval $[z_1, z_2]$. Monotonic convergence is guaranteed.

Propositions 6.1 and 6.2 provide convergence results for the zeros of the first and last subintervals and transition subintervals, as well as steps to abandon such subintervals. It is important to note that such steps ($\Delta z = \pi/2$) guarantee monotonic convergence once such subintervals are abandoned.

7. Algorithms. We will build explicit algorithms to compute with certainty all the zeros inside an interval. These algorithms implement Propositions 5.1, 6.1, and 6.2.

We will first consider evaluation of the zeros inside an interval $[z_1, z_2]$, where η_i does not change sign. The algorithm computes zeros in increasing order if $\eta_i < 0$ and in decreasing order in the other case (similarly to Proposition 5.1). Both forward ($\eta_i < 0$) and backward ($\eta_i > 0$) sweeps can be summarized in a single algorithm as follows. We use FORTRAN-like syntax.

ALGORITHM 1. FORWARD AND BACKWARD SWEEPS. *Let $[z_1, z_2]$ be an interval where η_i does not change sign. Let $H(z) \equiv H_i(z)$. The zeros of Y_n in this interval can be found with certainty using the following algorithm:*

Input: $j = -\text{sign}(\eta_i)$ *(+1 forward sweep; −1 backward); z_1;z_2;$\epsilon \equiv$ relative precision*
Output: *i (number of zeros); $z(1), \ldots, z(i)$: zeros in the interval*

```
(0)   SUBROUTINE SWEEP(j,z1,z2,ϵ,i,z(i))
(1)   NOTERM = 1
(2)   z̄1 = (z1 + z2)/2 + j((z1 − z2)/2)
(3)   z = z̄1
(4)   z̄2 = (z1 + z2)/2 − j((z1 − z2)/2)
(5)   IF (jH(z) > 0) THEN z = z + jπ/2
(6)   i = 0
(7)   DO WHILE (j(z − z̄2) < 0)
(8)        CALL FIXEDPOINT(z,z̄2,ϵ,zn,NOTERM)
(9)        IF(NOTERM = 1) THEN
(10)            z(j i) = zn
(11)            i = i + 1
(12)            z = zn + jπ/2
```

(13) *ELSE*
(14) $z = \bar{z}_2 + j$
(15) *ENDIF*
(16) *END WHILE*
(17) *END*

($0'$) *SUBROUTINE FIXEDPOINT*(z,$\bar{z}_2$,ϵ,z_n,*NOTERM*)
($1'$) $Err = 1 + \epsilon$
($2'$) *DO WHILE* (*NOTERM* $= 1$) AND ($Err > \epsilon$)
($3'$) $z_p = z$
($4'$) $z = z - \arctan(H(z))$
($5'$) $Err = |1 - z/z_p|$
($6'$) *IF* ($j(z - \bar{z}_2) > 0$) *THEN NOTERM* $= 0$
($7'$) *END WHILE*
($8'$) $z_n = z$
($9'$) *END*

and the zeros generated in a forward sweep are stored in the positive positions of the array $z(i)$ $(z(1), z(2), \ldots)$, *while those generated in a backward sweep are stored in the negative positions* $(z(-1),\ z(-2), \ldots)$.

Lines ($0'$)–($9'$) implement the fixed point iteration T_i. Lines (0)–(17) implement Proposition 5.1, complemented with Proposition 6.2 (line (5)).

Let us, for instance, discuss how the algorithm works in the case of a forward sweep ($j = +1$) in an interval $[z_1, z_2]$; for a backward sweep, the description is very similar. The first 4 lines are used to set the initial and final values for the sweep. In the case we are considering (forward), the starting value for the process is $\bar{z}_1 = z_1$ and the algorithm stops when, at any stage of the process, a value of z is reached which is greater than $\bar{z}_2 = z_2$ (lines ($6'$) and (7) test whether this condition is reached).

Line (5) tests whether there is a zero of Y_n between z_1 and the first zero of Y_{n+i} in $[z_1, z_2]$; if there is no such zero ($jH(z) < 0$), the instruction $z = z + j\pi/2$ sets a new starting value z to compute the smallest zero of Y_n in $[z_1, z_2]$ (see Proposition 6.2).

Lines (7)–(17) implement Proposition 5.1.

Note that all successive values of z generated in the process form an increasing sequence because convergence is always monotonic. This explains why the process can be stopped when a value of z outside this interval is reached. When this happens, we can be sure that all the zeros in the interval have been found.

The previous algorithm always can be applied when η_i does not change sign inside the interval under consideration; however, when η_i changes sign in I, we have to combine the forward and backward sweeps. If η_i is negative on the left of the interval and positive on the right, then we can use an algorithm which approaches the TP from both sides of the interval.

ALGORITHM 2. CONTRACTIVE SWEEP. *Let* $[z_1, z_2]$ *be an interval where* η_i *changes sign in such a way that* $\eta_i < 0$ *as* $z \to z_1$. *Let* z_η *be such that* $\eta_i(z_\eta) = 0$. *The zeros of* Y_n *in such an interval can be found with certainty using the following two successive calls*:
CALL SWEEP($+1$,z_1,z_η,ϵ,i,$z(i)$)
CALL SWEEP(-1,z_η,z_2,ϵ,i,$z(i)$)

If η_i is negative on the right of the interval and positive on the left, then the sweep starts from the TP and propagates away from it (expansive sweep). Convergence

starting from any value z in an interval $(z_{n+i}^{(j)}, z_{n+i}^{(j+1)})$ with a TP is not guaranteed because there is no SMC in such an interval. However, given the way η_i changes sign, Proposition 6.1 tells us that the starting value z_η ensures convergence to the zero of Y_n near the TP. In addition, the same proposition gives the recipe for abandoning the region of convergence to this zero. These are the main characteristics of the following algorithm.

ALGORITHM 3. EXPANSIVE SWEEP. *Let $[z_1, z_2]$ be a real interval where η_i changes sign in such a way that $\eta_i > 0$ as $z \to z_1$. Let z_η be such that $\eta_i(z_\eta) = 0$. The zeros of Y_n in such an interval can be found with certainty using the following algorithm*:

(0) *SUBROUTINE EXPANSIVE*(z_1,z_2,z_η,ϵ,$z(i)$)
(1) $z = z_\eta$
(2) $Err = 1 + \epsilon$
(3) *DO WHILE* $(Err > \epsilon)$ *AND* $(NOTERM = 1)$
(4) $\quad z_p = z$
(5) $\quad z = z - \arctan(H(z))$
(6) $\quad$ *IF* $(z > z_2)$ *OR* $(z < z_1)$ $NOTERM = 0$
(7) $\quad Err = |1 - z/z_p|$
(8) *END WHILE*
(9) $z_{tp} = z$
(10) *IF NOTERM*$= 1$ *THEN*
(11) $\quad z(0) = z$
(12) $\quad$ *IF* $(z_{tp} > z_\eta)$ *THEN*
(13) $\quad\quad$ *CALL SWEEP*($+1$,$z(0) + \pi/2$,z_2,ϵ,i,$z(i)$)
(14) $\quad\quad$ *CALL SWEEP*(-1,z_1,$z_\eta - \pi/2$,ϵ,i,$z(i)$)
(15) $\quad$ *ELSE*
(16) $\quad\quad$ *CALL SWEEP*($+1$,$z_\eta + \pi/2$,z_2,ϵ,i,$z(i)$)
(17) $\quad\quad$ *CALL SWEEP*(-1,z_1,$z(0) - \pi/2$,ϵ,i,$z(i)$)
(18) $\quad\quad$ *ENDIF*
(19) *ENDIF*
(20) *END*

$z(0)$ *is the zero of Y_n in the transition subinterval.*

The three algorithms described are sufficient to find all the zeros of any solution of a second order ODE $y_n''(x) + B(x)y_n'(x) + A_n(x)y_n(x) = 0$, $A_n(x) - A_{n-1}(x) \neq 0\,\forall x$ satisfying general DDEs (satisfied by two fundamental sets) with continuous coefficients. The forward and backward (or expansive and contractive) sweeps are most often simultaneously available given that we will normally have two iterations to choose, one with $\eta_{-1} = -(a_n - b_n)/2d_n$ and another with $\eta_{+1} = (a_{n+1} - b_{n+1})/2d_{n+1}$.

The condition $A_n(x) - A_{n-1}(x) \neq 0\,\forall x$ was used to show that η_i can change sign only once. However, if this condition is not met, strategies combining contractive and expansive sweeps can be developed. In any case, as we will see, this condition is met with great generality.

8. Applications. Considering that we started from basic ingredients, it is not surprising that the method developed is valid for a wide spectrum of special functions. The method can be applied to any combination of fundamental solutions of the ODE. In contrast, matrix eigenvalue methods apply only to orthogonal polynomials and to minimal solutions of three term recurrence relations

$$y_{n+1} + \alpha_n y_n + \beta_n y_{n-1} = 0. \tag{8.1}$$

The matrix methods for minimal solutions were developed by Grad and Zakrajsek [8], enhanced by Ikebe et al. (see [10, 11, 16] and references therein), and later revisited by Ball [2].

The fixed point method described here becomes especially simple to implement for minimal solutions of the recurrence (8.1) related to (2.3). The reason is that Pincherles's theorem [4, 18] guarantees the existence of continued fractions for the ratios of minimal solutions y_n/y_{n-1} and y_n/y_{n+1}, which are the only evaluations needed in our method. In this case, the method is based solely on the coefficients of the recurrence relations, as happens with matrix methods [10], with the advantage that the fixed point method is not affected by truncation errors, which need to be estimated. The fixed point method is not restricted to minimal solutions; for any dominant solution, the method still can be applied by using an alternative method for calculating $y_n/y_{n\pm1}$.

We are now explicitly showing how to implement these algorithms for general solutions of the Bessel, Coulomb, and Legendre equations (with emphasis on conical functions and Legendre polynomials) as well as for general solutions of second order ODEs satisfied by some classical orthogonal polynomials (Hermite, Laguerre). Computational details are given in [7].

8.1. Bessel equation (cylinder functions). The Bessel equation is

$$y_n''(x) + B(x)y_n'(x) + A_n(x)y_n(x) = 0, \tag{8.2}$$

with $B(x) = 1/x$, $A_n(x) = 1 - n^2/x^2$, and $A_n(x) - A_{n-1}(x) = \dfrac{2n-1}{x^2}$ which is only zero for $n = 1/2$. This case is trivial: $y_n(x) \propto \cos(x + \phi)$. The coefficients of the DDEs are

$$a_n = -n/x\,; \quad b_n = (n-1)/x\,; \quad d_n = -e_n = 1, \tag{8.3}$$

and then $\eta_i = -i\frac{n_i - 1/2}{x}$, with $i = \pm1$, $n_{+1} = n + 1$, and $n_{-1} = n$. There are no TPs ($x = 0$ is a regular singular point and our analysis is only for intervals where all the solutions are continuous). No change of variable $z(x)$ is required. The DDEs are satisfied by general cylinder functions; therefore, the algorithm can be applied to combinations of first and second kind Bessel functions, including Airy functions.

8.2. Coulomb equation (Coulomb wave functions). The coefficient $A_n(x)$ for the Coulomb equation (which is an equation in normal form) is $A_n(x) = 1 - 2\gamma/x - n(n+1)/x^2$; we use n instead of the more standard notation L, γ instead of η, and x instead of ρ (see [1, 25]). n and γ are parameters.

We can take $n > -1$ because $A_{-n-1} = A_n$. Observe that $A_n(x) - A_{n-1}(x) = -2n/x$ which equals zero only for $n = 0$. $n = 0$ is a special case only in the sense that the iteration with $i = -1$ has no meaning because y_0 and y_{-1} are not independent solutions (but the iteration for $i = +1$ always can be used).

The coefficients for the DDEs (see [1, eqs. (14.2.1) and (14.2.2)]) are

$$a_n = -b_n = -\frac{1}{n}\left(\frac{n^2}{x} + \gamma\right); \qquad d_n = -e_n = \frac{\sqrt{n^2 + \gamma^2}}{n}\ . \tag{8.4}$$

We need only consider positive zeros because the negative ones are positive zeros of the functions with the sign of γ reversed. The change of variables can be chosen as $z = \int d_{n_i}(x)dx = d_{n_i}x$ (again, $n_i = n$ for $i = -1$ and $n_i = n + 1$ for $i = +1$).

We have that $\eta_i = -i(n_i^2 + \gamma x)/(x\sqrt{n_i^2 + \gamma^2})$. Thus, there is a TP at $z_\eta = -d_n n^2/\gamma$ which is positive for $n \in (-1, 0)$ and $\gamma > 0$ as well as for $n > 0$ and $\gamma < 0$.

If there is no positive TP, then the forward sweep ($i = +1$) always can be used and the backward sweep ($i = -1$) can be used except for $n = 0$. When there is a TP, we can use both contractive and expansive sweeps (except for the case $n = 0$, for which only the expansive sweep can be used).

The DDEs are satisfied by both the regular and irregular Coulomb wave functions; the algorithm applies to any combination of them (general solutions of the differential equation).

8.3. Legendre equation (conical functions). Conical functions are chosen to illustrate the application of the fixed point method to Legendre functions P_m^n, Q_m^n, which are solutions of a differential equation $y_n''(x) + B(x)y_n'(x) + A_n(x)y_n(x) = 0$ with coefficients

$$B(x) = 2x/(x^2 - 1);\ A_n(x) = -[m(m+1) + n^2/(x^2 - 1)](x^2 - 1). \tag{8.5}$$

We have reversed the use of m and n with respect to the standard notation [1] in order to be consistent with our own notation. We will consider the case $x > 1$.

We have $A_n - A_{n-1} = (-2n + 1)/(x^2 - 1)^2$ which vanishes (identically) only for $n = 1/2$; this is, as for Bessel functions with $n = 1/2$, a trivial case.

Conical functions are defined by $m = -1/2 + i\tau$ with real τ. The corresponding coefficients of the difference-differential relations are

$$a_n = \frac{-nx}{x^2 - 1}, \quad b_n = \frac{(n-1)x}{x^2 - 1}, \quad d_n = -\frac{\lambda_n^2}{\sqrt{x^2 - 1}}, \quad e_n = \frac{1}{\sqrt{x^2 - 1}}, \tag{8.6}$$

with

$$\lambda_n = \sqrt{(n - 1/2)^2 + \tau^2}\ . \tag{8.7}$$

In order to set $d_n = -e_n$ we need to renormalize the solutions. We then use new functions $\bar{y}$: $y_p = K_p \bar{y}_p$ with $K_n = \sqrt{(n - 1/2)^2 + \tau^2} K_{n-1}$. With this, the new coefficients d_n and e_n are $d_n = -e_n = -\lambda_n/\sqrt{x^2 - 1}$ and the change of variables $z(x)$ can be taken as $z(x) = \int |d_{n_i}(x)| dx = \lambda_{n_i} \ln(x + \sqrt{x^2 - 1})$, which means that $x(z) = \cosh(z/\lambda_{n_i})$.

The iterations read $T_i(z) = z - \arctan(H_i(z))$ with

$$\begin{aligned} H_{-1}(z) &= \mathrm{sign}(d_n)\frac{\bar{y}_n(x(z))}{\bar{y}_{n-1}(x(z))} = -\frac{1}{\lambda_n}\frac{y_n(x(z))}{y_{n-1}(x(z))}, \\ H_{+1}(z) &= -\mathrm{sign}(d_{n+1})\frac{\bar{y}_n(x(z))}{\bar{y}_{n+1}(x(z))} = \lambda_{n+1}\frac{y_n(x(z))}{y_{n+1}(x(z))}. \end{aligned} \tag{8.8}$$

From the coefficients we get $\eta_i = i(-n_i + 1/2)x/(\lambda_{n_i}\sqrt{x^2 - 1})$, with n_i as defined before. There are no TP for $x > 1$. However, η_i vanishes identically for $n_i = 1/2$. This is a trivial case because $\dot{H}_i = 1 + H_i^2$.

Forward and backward sweeps are possible.

The same DDEs are satisfied by first kind ($P^n_{-1/2+i\tau}$) and second kind ($Q^n_{-1/2+i\tau}$) conical functions and therefore the algorithm also applies to any linear combination.

8.4. Legendre equation (Legendre polynomials). Differential equation: $(1-x^2)P_n(x) - 2xP_n'(x) + n(n+1)P_n(x) = 0$, $|x| < 1$.

Coefficients in the DDEs: $a_n = -b_n = -\frac{nx}{1-x^2}$, $d_n = -e_n = \frac{n}{1-x^2}$.

Change of variable: $z(x) = \frac{n_i}{2}\ln\frac{1+x}{1-x}$; $x(z) = \tanh(z/n_i)$. Range of z: $(-\infty, +\infty)$.

TPs: $\eta_i = -ix$; TP at $x_\eta = 0 = z(x_\eta) \equiv z_\eta$.

Applicable algorithms: expansive from $z = 0$; contractive is not appropriate given the range of z. Only the positive z-roots of Legendre polynomials need to be calculated because they are symmetric around $x = 0$ and $z(-x) = -z(x)$; thus, a forward sweep from $z = 0$ is sufficient to find all the zeros.

The algorithms apply to any solution of the differential equation, satisfying the same DDEs. Then we can use them for $\cos\alpha P_n - \sin\alpha Q_n$ for any α.

8.5. Hermite equation (Hermite polynomials). Differential equation: $y_n'' - 2xy_n' + 2ny_n = 0$.

Coefficients in the DDEs: $a_n = 0$, $b_n = 2x$, $d_n = 2n$, $e_n = -1$.

Normalization: We must renormalize Hermite polynomials by $y_p = k_p\bar{y}_p$ with $k_n = \sqrt{2n}k_{n-1}$ $(n > 0)$.

Renormalized coefficients: same as before except that $d_n = -e_n = \sqrt{2n}$.

Change of variable: $z(x) = \sqrt{2n_i}x$. Range of z: $(-\infty, +\infty)$.

TPs: $\eta_i = -ix(z)/\sqrt{2n_i}$; TP at $x_\eta = 0 = z(x_\eta) \equiv z_\eta$.

Applicable algorithms: same as for Legendre polynomials.

The functions $e^{-x^2/2}y_n(x)$ satisfy an equation in normal form and two DDEs with $A_n(x) = 1 - x^2 + 2n$, $a_n = -b_n = -x$, $d_n = 2n$, $e_n = -1$ which satisfy all the conditions of Lemma 2.2. Thus, the DDEs are satisfied by general solutions (with a convenient normalization). The algorithms apply to arbitrary solutions of the ODE.

8.6. Laguerre equation (generalized Laguerre polynomials). Differential equation: $xy_n'' + (\alpha + 1 - x)y_n' + ny_n = 0$, $x > 0$.

Coefficients: $a_n = n/x$, $b_n = 1 - (n+\alpha)/x$, $d_n = -(n+\alpha)/x$, $e_n = n/x$.

Normalization: We must renormalize Laguerre polynomials by $y_p = k_p\bar{y}_p$ with $k_n = \sqrt{(n+\alpha)/n}k_{n-1}$.

Renormalized coefficients: same as before except that $d_n = -e_n = -\frac{1}{x}\sqrt{(n+\alpha)n}$.

Change of variable: $z(x) = \sqrt{(n_i+\alpha)n_i}\ln x$. Range of z: $(-\infty, +\infty)$.

TPs: $\eta_i = i(n_i + \alpha/2 - x/2)/\sqrt{n_i(n_i+\alpha)}$; TP at $x_\eta = 2n_i + \alpha$, $z_\eta = z(x_\eta) = \sqrt{n_i(n_i+\alpha)}\ln(2n_i+\alpha)$.

Applicable algorithms: expansive or contractive from z_η. Given the range of z, expansive is more appropriate.

Writing the differential equation in normal form and using Lemma 2.2 we conclude that the algorithms apply to arbitrary solutions of the ODE.

9. Conclusions and perspectives. The method described here can be applied to any oscillating special function when DDEs (with continuous coefficients) satisfied by two sets of fundamental solutions are available. We also require that A_n and A_{n+i} are never equal (which is a condition generally met).

The method has several advantages with respect to matrix methods: not only can the zeros of minimal solutions be obtained, but also the zeros of general solutions. The zeros in any interval (x_1, x_2) can be found, and not necessarily the first n zeros. The method can be used to investigate the existence of zeros in a given interval (where perhaps there is no zero at all). In addition, although fast (see [22] for a similar method for Bessel functions), the method admits improvements when there is

additional information on the zeros of a function (like asymptotic [24] or Chebyshev approximations for the zeros of Bessel functions), in contrast to matrix methods. Finally, the fixed point method is not affected by matrix truncation errors, which are function dependent [16, 2]. With respect to more general purpose methods [12], the fixed point method has the advantage of being much more simple to implement. Besides, it proves to be faster: the global Newton method for Bessel functions in [22] compared favorably with respect to [28] and the speed of convergence of this Newton method [21] is improved by the implementation of the fixed point method described in this paper. More computational details are given in [7].

The main difficulty when evaluating zeros of a given function is usually the bracketing of the roots, given that the distribution of zeros can change drastically from one set of parameters to another. The fixed point method here presented does not require bracketing. A good example of this variability are Coulomb functions, which present very rapid oscillations for $\gamma < 0$, large $|\gamma|$, and not too large x and oscillate slowly for large and positive n and/or γ and not too large x. It is observed that the number of iterations needed by the algorithm to converge to the different zeros is quite stable with varying parameters [7]. For instance, for Coulomb functions we have experienced that around 10 iterations are enough to get a precision of 10^{-8} for moderate values of n and γ and that as n and γ increase this number is quite stable, reaching 20–30 for values as large as 500 (in spite of the fact that the spacing between the first zeros increases considerably).

Not only can the spacing between zeros vary for different parameter selection but also, for a fixed function, the spacing between zeros can vary drastically for different ranges of x. The change of variable $z(x)$ tends to smooth this variation; for instance, this is observed for conical functions, which oscillate very fast for x close to 1 and slow down for large x. The change of variables $z(x)$ (see the previous section) automatically sets a much more uniform distribution of zeros with respect to z, which explains why the fixed point method has a fairly uniform behavior regarding convergence to all the zeros.

Both the method and the bonus information regarding the distances between zeros have the property of being valid for a broad family of special functions, which is a rather singular event. In future papers, we plan to investigate how broad this class of functions is and to provide a comprehensive lexicon for its application.

Also, the zeros of the derivatives of this class of functions can be investigated using similar techniques. Given a family of functions $\{y_n\}$ satisfying a second order linear ODE in normal form, after a change of the dependent variable one can write a second order linear ODE for $z_n \propto y_n'$, together with first order DDEs; the main difficulty consists of finding analytic expressions for the change of variable $z(x)$ and its inverse. An alternative consists of realizing that for positive A_n the zeros of y_n and y_n' are interlaced. This suggests that fixed point methods based on the logarithmic derivative of y_n are more appropriate, as discussed in [7].

Acknowledgments. The author thanks A. Gil, W. Koepf, and N.M. Temme for carefully reading the manuscript. The author thanks A. Gil for testing the algorithms.

REFERENCES

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* Dover, New York, 1972.

[2] J.S. Ball, *Automatic computation of zeros of Bessel functions and other special functions,* SIAM J. Sci. Comput., 21 (2000), pp. 1458–1464.

[3] C. Belingeri and P.E. Ricci, *On asymptotic formulas for the first zero of the Bessel function* $J_\nu(x)$, J. Inform. Optim. Sci., 17 (1996), pp. 267–274.
[4] C. Brezinski, ed., *Continued Fractions and Padé Approximants,* North-Holland, Amsterdam, 1990.
[5] Á Elbert and A. Laforgia, *An upper bound for the zeros of the cylinder function* $\mathcal{C}_\nu(x)$, Math. Inequal. Appl., 1 (1998), pp. 105–111.
[6] B.R. Fabijonas and F.W.J. Olver, *On the reversion of an asymptotic expansion and the zeros of the airy functions,* SIAM Rev., 41 (1999), pp. 762–773.
[7] A. Gil and J. Segura, *Computing zeros and turning points of special functions from fixed point methods,* submitted.
[8] J. Grad and E. Zakrajsek, *Method for evaluation of zeros of Bessel functions,* J. Inst. Math. Appl., 11 (1973), pp. 57–72.
[9] E.K. Ifantis and P.D. Siafarikas, *A differential inequality for the positive zeros of Bessel functions,* J. Comput. Appl. Math., 44 (1992), pp. 115–120.
[10] Y. Ikebe, *The zeros of regular Coulomb wave functions and of their derivatives,* Math. Comput., 29 (1975), pp. 878–887.
[11] Y. Ikebe, Y. Kikuchi, and I. Fujishiro, *Computing zeros and orders of Bessel functions,* J. Comput. Appl. Math., 38 (1991), pp. 169–184.
[12] D.J. Kavvadias and M.N. Vrahatis, *Locating and computing all the simple roots and extrema of a function,* SIAM J. Sci. Comput., 17 (1996), pp. 1232–1248.
[13] K.S. Kölbig, C345: *Zeros of Bessel functions J and Y,* CERNLIB Program Library, European Laboratory for Particle Physics, 1994. Available online at http://wwwinfo.cern.ch/asdoc/shortwrupsdir/c345/top.html.
[14] I. Marx, *On the structure of recurrence relations,* Michigan Math. J., 2 (1953), pp. 45–50.
[15] I. Marx, *On the structure of recurrence relations* II, Michigan Math. J., 2 (1953), pp. 99–103.
[16] Y. Miyazaki, Y. Kikuchi, D. Cai, and Y. Ikebe, *Error analysis for the computation of zeros of regular Coulomb wave function and its first derivative,* Math. Comp., 70 (2001), pp. 1195–1209.
[17] M.E. Muldoon, *Convexity properties of special functions and their zeros,* in Recent Progress in Inequalities, Kluwer Academics Publishers, Dordrecht, Boston, London, 1998, pp. 309–323.
[18] W.H. Press, S.A. Teukoslky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in Fortran,* Cambridge University Press, Cambridge, UK, 1992.
[19] R. Piessens, *Chebyshev series approximations for the zeros of the Bessel functions,* J. Comput. Phys., 53 (1984), pp. 188–192.
[20] R. Piessens, *On the computation of zeros and turning points of Bessel functions,* Bull. Greek Math. Soc., 31 (1990), pp. 117–122.
[21] J. Segura, *A global Newton method for the zeros of cylinder functions*, Numer. Algorithms, 18 (1999), pp. 259–276.
[22] J. Segura and A. Gil, *ELF and GNOME: Two tiny codes to evaluate the real zeros of the Bessel functions of the first kind for real orders,* Comput. Phys. Comm., 117 (1999), pp. 250–262.
[23] J. Segura, *Bounds on differences of adjacent zeros of Bessel functions and iterative relations between consecutive zeros*, Math. Comp., 70 (2001), pp. 1205–1220.
[24] N.M. Temme, *An algorithm with ALGOL* 60 *program for the computation of the zeros of ordinary Bessel functions and those of their derivatives,* J. Comput. Phys., 32 (1979), pp. 270–279.
[25] N.M. Temme, *Special Functions. An Introduction to the Classical Functions of Mathematical Physics,* John Wiley, New York, 1996.
[26] N.M. Vrahatis, O. Ragos, F.A. Zafiropoulos, and T.N. Grapsa, *Locating and computing zeros of Airy functions,* Z. Angew. Math. Mech., 76 (1996), pp. 419–422.
[27] M.N. Vrahatis, T.N. Grapsa, O. Ragos, and F.A. Zafiropoulos, *On the location and computation of zeros of Bessel functions,* Z. Angew. Math. Mech., 77 (1997), pp. 467–475.
[28] N.M. Vrahatis, O. Ragos, T. Skiniotis, F.A. Zafiropoulos, and T.N. Grapsa, *RFSFNS: A portable package for the numerical determination of the number and the calculation of roots of Bessel functions,* Comput. Phys. Comm., 92 (1995), pp. 252–266.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 134–158

WEAK ILL-POSEDNESS OF SPATIAL DISCRETIZATIONS OF ABSORBING BOUNDARY CONDITIONS FOR SCHRÖDINGER-TYPE EQUATIONS*

ISAÍAS ALONSO-MALLO† AND NURIA REGUERA‡

Abstract. When we wish to solve numerically a differential problem defined on an infinite domain, it is necessary to consider a finite subdomain and to use artificial boundary conditions in such a way that the solutions in the finite subdomain approximate the original solution. These boundary conditions are called absorbing when small reflections to the interior domain are allowed. In this paper, we develop a general class of absorbing boundary conditions for Schrödinger-type equations by using rational approximations to the transparent boundary conditions. With this approach, previous absorbing boundary conditions in the literature are included in this class. We use the method of lines for the discretization of the initial boundary value problems obtained this way. We show that the ordinary differential systems that arise after the spatial discretization are weakly ill-posed, explaining a previous conjecture of Fevens and Jiang. The time discretization is carried out with A-stable Runge–Kutta methods, where the high order ones may be used to compensate for the possible troubles present in the problems semidiscretized in space.

Key words. Schrödinger equation, absorbing boundary conditions, initial boundary value problems, method of lines

AMS subject classifications. 65M12, 65M20, 65M99

PII. S0036142900374433

1. Introduction.

The numerical solution of initial value problems on unbounded domains arises in a wide variety of differential equations. In order to obtain a manageable problem, it is necessary to consider a finite subdomain $[x_l, x_r]$ and to impose artificial boundary conditions. When the solution of this new problem is equal to the restriction to the subdomain of the original solution, we say that the boundary condition is *transparent* (TBC). However, when the transparent boundary conditions are nonlocal it is convenient to use local *absorbing* boundary conditions (ABCs), permitting that the computed solution presents some reflections.

The TBCs and ABCs and their discretizations are widely studied (see [5, 9, 23] for hyperbolic problems, [4, 8] for parabolic ones). In this paper, we are interested in the study of ABCs for the equation

$$\partial_t u = -\frac{i}{c}(\partial_x^2 u + V(x,t)u), \quad x \in \mathbf{R},\ t > 0, \tag{1.1}$$

where c is a real constant and V is the potential. (In this paper, we suppose without loss of generality that $c > 0$. The case $c < 0$ is analogous.) This equation arises in two well-known cases: the one-dimensional time dependent Schrödinger equation for

*Received by the editors June 26, 2000; accepted for publication (in revised form) October 5, 2001; published electronically April 12, 2002.

http://www.siam.org/journals/sinum/40-1/37443.html

†Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, C/ Doctor Mergelina s.n., 47014 Valladolid, Spain (isaias@mac.cie.uva.es). The research of this author was supported by MCYT BFM 2001-2013 and JCYL VA025/01.

‡Departamento de Matemáticas y Computación, Universidad de Burgos, Avda. Cantabria s.n., 09006 Burgos, Spain (nreguera@ubu.es). The research of this author was supported by JCYL VA025/01.

a particle with mass m,

$$i\hbar\partial_t\Psi = -\frac{\hbar^2}{2m}\partial_x^2\Psi + V(x,t)\Psi,$$

and the Fresnel equation for the evolution of a paraxial electrical field E along the z-direction in a Cartesian coordinate system [17, 24],

$$2in_0k_0\partial_z E = \partial_x^2 E + (n^2(x) - n_0^2)k_0^2 E, \tag{1.2}$$

where $n(x)$ is the refractive index ($n = 1$ if the solution propagates in a vacuum) and n_0 the reference index. We suppose in this paper that the potential V is constant or almost constant. With these hypotheses, it is possible to obtain an expression for the TBCs. Equation (1.1) along with these TBCs give rise to a well-posed problem when the support of the initial condition is contained in the domain $[x_l, x_r]$. But these TBCs are nonlocal and thus their practical interest is limited. Baskakov and Popov [2] considered these TBCs, using for their implementation linear approximations of the solution in the time intervals. However, such an approach may cause instabilities, as is proved in [15].

As an alternative, Schmidt and Yevick [19] proposed a TBC for the Fresnel equation discretized in time with the θ-method and another one for the fully discretized problem. As they are transparent, we can expect very small reflections at the boundary. Nevertheless, these boundary conditions have some troubles: they are nonlocal and thus the computational cost is high, and they are specific to the method used for the discretization in time (θ-method) and also in space in the case of the fully discretized ones. In this way, we cannot use other different numerical schemes in the interior domain. Previously, Schmidt and Deuflhard [18] had already obtained these boundary conditions for the Fresnel equation discretized in time with variable coefficients.

Several works have been done to develop local ABCs for the Schrödinger equation. Di Menza [3] considered the Schrödinger equation without potential in several dimensions. In order to get local ABCs from the transparent ones, he obtains the expression analogous to (2.3) below for the TBCs in Fourier variables, and he uses interpolating rational functions whose numerator and denominator are both polynomials of the same degree. The interpolatory nodes are chosen in order to obtain least squares approximations to the Fourier symbol of the TBCs in a given interval.

In a recent work, Fevens and Jiang [6] have developed the following ABCs for the Schrödinger equation:

$$\prod_{l=1}^{p}\left(i\partial_x - \frac{a_l c}{2}\right)u = 0. \tag{1.3}$$

This boundary condition is developed in order to absorb completely the components of the wave solutions that travel with group velocities a_l, $l = 1, \ldots, p$. They showed how Shibata's boundary conditions [20] is equivalent to (1.3) with $p = 2$ and so is Kuska's [13] for $p = 3$ and $a_1 = a_2 = a_3$.

In this paper, we define a general class of ABCs considering interpolating rational functions, as Di Menza did, but here the degree of the numerator and denominator is not necessarily the same, and we do not fix the interpolatory nodes in any way. However, we show what is the optimal choice in order to absorb the solution when it is a plane wave. These ABCs are denoted ABC(j_1, j_2), where j_1, j_2 are the degrees of the

polynomials in the numerator and denominator of the rational function, respectively. We will call order of absorption the value $j_1 + j_2 + 1$.

We prove that the ABCs obtained are a generalization of ABCs in [6]. However, a well-posedness study in a similar way to the one carried out in [6] shows that the initial boundary value problem that arises by using these ABCs may be well-posed only for the ABCs obtained in [6].

We study the discretization of the ABCs by using the method of lines. First, we discretize in space, obtaining a system of ordinary differential equations. We use the same standard finite differences discretization used in [6, 19] for the interior domain, together with several implementations at the boundary, depending on the ABCs that we use.

With our analysis, we can prove that the semidiscrete problems that arise after the spatial discretization are weakly ill-posed. In fact, these problems are dissipative when we use an adequate weighted norm. However, the bound depends on the parameter of the spatial discretization and this may lead to a growth of this bound when the spatial grid is refined. We prove this fact for the case of ABC(1,0), showing that the ill-posedness is very weak. Therefore, the ABC(1,0) may be useful in practical applications, although they are not very absorbing. Moreover, we show numerically that the semidiscrete problems associated with higher order ABCs are worse posed. This explains why in some cases we can get better results with ABCs of smaller order of absorption, as is conjectured in [6]. On the other hand, when the solution is a wave traveling with a concrete group velocity, the worse behavior of ABCs of higher order can be canceled because of the greater absorption at the boundary.

Second, we use a Runge–Kutta method for the time integration, although other time integration methods can be considered. Since the problem discretized in space is weakly ill-posed, it is very important to use a high order method for the time integration to compensate for the troubles in space, allowing a moderate time stepsize. Therefore, it is crucial to use an algorithm obtained by using the method of lines.

The organization of the paper is the following. The ABCs are studied in section 2, together with the relations with other ABCs previously proposed in the literature. Section 3 is devoted to the spatial discretization of the case ABC(1,0). Some theorems of linear algebra, which are necessary to show the weak ill-posedness of the semidiscrete problems, are proved in section 4. In section 5 we study the time discretization by using Runge–Kutta methods. In section 6 we analyze the discretization of ABCs of higher order, showing numerically the weak ill-posedness. The higher-dimensional case is studied in section 7. Finally, section 8 presents some numerical experiments showing results previously obtained.

2. Absorbing boundary conditions. Let us suppose $x_r = 0$ and consider the right exterior domain $x > 0$. In order to obtain an expression for the TBCs, we use an argument similar to the one in [19]. Let us denote by $\hat{u}(x,\omega)$ the Fourier–Laplace transform of $u(x,t)$, where $\omega = \Re(\omega) + i\Im(\omega)$ with $\Im(\omega) < 0$. From (1.1), we get

$$\partial_x^2 \hat{u} - \lambda^2 \hat{u} = 0, \tag{2.1}$$

where $\lambda^2 = -(V + c\omega)$, and since we are interested in solutions that go to 0 as $x \to \infty$, it should be

$$\hat{u}(x,\omega) = \hat{u}(0,\omega) e^{-\lambda x} \tag{2.2}$$

with $\lambda = \sqrt{-(V + c\omega)}$ ($\sqrt{\ }$ is the square root with positive real part). If $\hat{U}(p,\omega)$ denotes the Laplace transform of $\hat{u}(x,\omega)$, in view of (2.2), it is $\hat{U}(p,\omega) = \hat{u}(0,\omega)/(p +$

λ), and then $\hat{U}(\lambda,\omega) < \infty$. On the other hand, if we take Laplace transform in (2.1), we obtain

$$\hat{U}(p,\omega) = \frac{p\hat{u}(0,\omega) + \partial_x \hat{u}(0,\omega)}{(p^2 - \lambda^2)}.$$

So in order for $\hat{U}(p,\omega)$ to be finite at $p = \lambda$, the numerator in the previous expression should be 0 for this value of p. We have developed the expression for the TBCs:

$$i\sqrt{V + c\omega}\,\hat{u}(x_r,\omega) + \partial_x \hat{u}(x_r,\omega) = 0. \tag{2.3}$$

An analogous expression can be obtained for the left boundary. Our purpose is to approximate these nonlocal boundary conditions by local ones. We are going to consider approximations

$$\sqrt{V + c\omega} \approx q(V + c\omega), \tag{2.4}$$

where $q(s)$ is a rational function that interpolates $\sqrt{s}$. This is a generalization of [1] where we considered as q the Taylor expansion of first order. We use the notation ABC(j_1, j_2) when $q(s) = p_1(s)/p_2(s)$, where p_1 and p_2 are relatively prime polynomials with degrees j_1 and j_2, respectively. We recall that $j_1 + j_2 + 1$ is the order of absorption. Notice that the ABC also depends on the interpolatory nodes, but this dependence is not displayed in our notation.

We will not consider the cases ABC(0,j) for $j \geq 0$, since with a study similar to the one made later in this section, we see that ABC(0,1) and ABC(0,2) cannot give rise to a well-posed problem, and although this does not happen for ABC(0,0), its order of absorption is too small. The simplest cases are the following.

ABC(1,0). Let us consider in (2.4) the polynomial $q(s)$ that interpolates $\sqrt{s}$ at the points s_1^2 and s_2^2 (with $s_1, s_2 > 0$),

$$q(s) = \frac{s_1 s_2}{s_1 + s_2} + \frac{s}{s_1 + s_2}. \tag{2.5}$$

Then the differential operator associated with the symbol obtained in (2.3) by considering the approximation (2.4) is

$$\beta_0 \partial_t u(x_r,t) + \beta_1 u(x_r,t) + \partial_x u(x_r,t) = 0, \tag{2.6}$$

where $\beta_0 = c/(s_1 + s_2)$, $\beta_1 = i(s_1 s_2 + V)/(s_1 + s_2)$. In the particular case when $s_1^2 = s_2^2 = b$, the approximation (2.4) reduces to the Taylor expansion of first order of $\sqrt{V + c\omega}$ at $\omega = \omega^*$, with $\omega^* = (b - V)/c$.

ABC(1,1). We can also replace $\sqrt{V + c\omega}$ by $q(V + c\omega)$, where $q(s) = (\alpha_0 + \alpha_1 s)/(1 + \alpha_2 s)$ is a rational function that interpolates $\sqrt{s}$ at s_1^2, s_2^2, and s_3^2. This gives rise to

$$\beta_0 u(x_r,t) + \beta_1 \partial_t u(x_r,t) + \beta_2 \partial_x u(x_r,t) + \beta_3 \partial_{tx} u(x_r,t) = 0 \tag{2.7}$$

for certain coefficients β_j depending on s_1, s_2, and s_3. In particular, when $s_1^2 = s_2^2 = s_3^2 = b$, the approximation considered reduces to the Padé(1,1) expansion of $\sqrt{V + c\omega}$ at $\omega = \omega^*$.

ABC(2,1). Let us now consider $q(s) = (\alpha_0 + \alpha_1 s + \alpha_2 s^2)/(1 + \alpha_3 s)$ so that it interpolates $\sqrt{s}$ at the points s_1^2, s_2^2, s_3^2, and s_4^2. In this way, we get

$$\beta_0 u(x_r,t) + \beta_1 \partial_t u(x_r,t) + \beta_2 \partial_t^2 u(x_r,t) + \beta_3 \partial_x u(x_r,t) + \beta_4 \partial_{tx} u(x_r,t) = 0 \tag{2.8}$$

for certain coefficients β_j that depend on s_1, s_2, s_3, and s_4. The special case $s_1^2 = s_2^2 = s_3^2 = s_4^2 = b$ corresponds to a Padé(2,1) approximation of $\sqrt{V + c\omega}$ at $\omega = \omega^*$.

ABC(3,2). Finally, let $q(s) = p_1(s)/q_1(s)$ with $p_1(s)$ and $q_1(s)$ of degree 3, 2, respectively, interpolate $\sqrt{s}$ at the points s_1^2, s_2^2, s_3^2, s_4^2, and s_5^2. With a similar reasoning as in previous cases, we obtain

$$0 = \beta_0 u(x_r,t) + \beta_1 \partial_t u(x_r,t) + \beta_2 \partial_t^2 u(x_r,t) + \beta_3 \partial_t^3 u(x_r,t) + \beta_4 \partial_x u(x_r,t) + \beta_5 \partial_{tx} u(x_r,t) + \beta_6 \partial_{ttx} u(x_r,t). \tag{2.9}$$

ABC(2,0). We can also consider the second order polynomial $q(s)$ that interpolates $\sqrt{s}$ at three points. Nevertheless, the ABC obtained is useless for practice as we will see later on.

Let us now see which should be the value for the interpolatory nodes. Consider the equation in the interior domain along with a solution of kind $u(x,t) = \exp i(kx - \omega(k)t)$, where $\omega(k) = (V - k^2)/c$. When we use the ABC obtained by considering the approximation (2.4) in (2.3), we need $\sqrt{V + c\omega} - q(V + c\omega)$ to be small when $\omega = -\omega(k)$, that is, when $V + c\omega = k^2$. This way, the approximation should be good when at least one of the nodes of interpolation is $s_1^2 = k^2$.

Let us now study the relation between our ABCs and the ones proposed in [6]. Because of the derivatives of the solution appearing in the ABCs, only ABC($j+1$,j) and ABC(j,j) for a natural number j could coincide with (1.3). If we take $s_j = a_j c/2$, $j = 1, 2$ in ABC(1,0), we obtain the ABCs proposed in [6] for $p = 2$. Similarly, ABC(1,1) coincides with the one in [6] for $p = 3$ and $s_j = a_j c/2$, $j = 1, 2, 3$. In the same way, ABC(2,1) is that of [6] for $p = 4$ and $s_j = a_j c/2$, $j = 1, 2, 3, 4$.

The study in a rigorous way of the well-posedness of the initial boundary value problems obtained with these ABC is not the purpose of this paper. However, the theory of well-posedness for hyperbolic problems of Kreiss [12, 21] provides necessary conditions for the study of well-posedness for these problems (cf. [10]). In this way, in [6] it is checked whether there exist solutions $\Psi(s) = \exp(st + \eta x)$ of the equation and the ABCs with $\Re(s) \geq 0$, concluding that the only solution of this kind satisfies $\Re(s) = 0$. For our ABC(1,0) (that coincides with the one of [6] for $p = 2$), these solutions should satisfy

$$ics - V = \eta^2, \qquad \beta_0 s + \beta_1 + \eta = 0,$$

and then the only possibilities for s are $s = i(s_j^2 - V)/c$, for $j = 1, 2$, both imaginary. The group velocities associated with these solutions are $2s_j/c > 0$ and as we are considering the right boundary, this will not produce a reflected wave. We have obtained similar conclusions for ABC(1,1), ABC(2,1), ABC(2,2), and ABC(3,2). For ABC(2,0) and ABC(1,2) there exists a value of s that although it is imaginary, it gives rise to a solution with negative group velocity. For ABC(3,0) and ABC(1,3) there exists s with $\Re(s) > 0$ such that $\exp(st + \eta x)$ is a solution of the equation and the ABCs, giving rise to an exponentially unstable problem. We see that among the ABCs previously mentioned, the only ones that could give rise to a well-posed problem are those of the form ABC($j+1$,j) and ABC(j,j), that is, particular cases of [6].

3. Spatial discretization of ABC(1,0). Let $\{x^j : 0 \leq j \leq N\}$ be a uniform mesh of the interval $[x_l, x_r]$, where $x^j = x_l + jh$, $0 \leq j \leq N$ with $h = L/N$ and $L = x_r - x_l$. We will denote by $u^j(t)$ an approximation of $u(x^j, t)$. We have considered,

as in [6, 19], finite differences for the discretization in space of the equation in the interior domain $[x_l, x_r]$:

$$\frac{d}{dt}u^j(t) = \widetilde{m}_1(h)u^{j-1}(t) + \widetilde{m}_2(h)u^j(t) + \widetilde{m}_1(h)u^{j+1}(t), \quad 1 \le j \le N-1, \tag{3.1}$$

with $\widetilde{m}_1(h) = -i/ch^2$, $\widetilde{m}_2(h) = i(2 - Vh^2)/ch^2$.

Let us study now how to discretize the equation at the right boundary when we consider ABC(1,0) given by (2.6). An analogous reasoning is valid for the left boundary. Let us consider, as an approximation to $\partial_x^2 u$ at the right boundary x^N,

$$\partial_x^2 u(x^N, t) \approx \frac{2}{h^2}(u^{N-1}(t) - u^N(t) + h\partial_x u(x^N, t)) \tag{3.2}$$

so that, along with (2.6), we can develop the following discretization of the equation at x^N:

$$\begin{aligned}\frac{d}{dt}u^N(t) &\approx \frac{-i}{c}(\partial_x^2 u(x^N, t) + Vu^N(t)) \\ &\approx \frac{-2i}{ch^2}\left(u^{N-1}(t) - u^N(t) - h\beta_0\frac{d}{dt}u^N(t) - h\beta_1 u^N(t)\right) - \frac{iV}{c}u^N(t).\end{aligned}$$

We have obtained $du^N(t)/dt \approx \widetilde{\alpha}(h)u^N(t) + \widetilde{\beta}(h)u^{N-1}(t)$ this way, where

$$\widetilde{\alpha}(h) = \frac{i(2 - h^2V)(s_1 + s_2) - 2h(s_1s_2 + V)}{ch(-2i + hs_1 + hs_2)}, \quad \widetilde{\beta}(h) = \frac{-2i}{ch^2\left(1 - \frac{2i}{h(s_1+s_2)}\right)}.$$

For the left boundary, we have a similar expression. With this implementation of ABC(1,0), we have obtained a first order ordinary differential system

$$U'(t) = M(h)U(t) \tag{3.3}$$

with $U(t) = \left[u^0(t), u^1(t), \ldots, u^{N-1}(t), u^N(t)\right]^T$ and

$$M(h) = \begin{bmatrix} \widetilde{\alpha} & \widetilde{\beta} & 0 & 0 & \cdots & 0 \\ \widetilde{m}_1 & \widetilde{m}_2 & \widetilde{m}_1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \widetilde{m}_1 & \widetilde{m}_2 & \widetilde{m}_1 \\ 0 & \cdots & 0 & 0 & \widetilde{\beta} & \widetilde{\alpha} \end{bmatrix} \in \mathcal{M}_{(N+1)\times(N+1)}, \tag{3.4}$$

where we omit the dependence on h in the notation of the elements of $M(h)$. The solution of (3.3) is

$$U(t) = \exp(M(h)t)U(0).$$

Our following objective is to study the well-posedness in the Euclidean norm of (3.3), with the corresponding scalar product denoted by $\langle\cdot,\cdot\rangle$. We have

$$\frac{d}{dt}\|U\|^2 = \frac{d}{dt}\langle U, U\rangle = 2\Re\langle U, U'\rangle = 2\Re\langle U, M(h)U\rangle.$$

TABLE 3.1
Maximum of the real part of the ϵ-pseudoeigenvalues.

ϵ	ABC(1,0)	ABC(1,1)	ABC(1,1) 2nd implementation	ABC(2,1)	ABC(3,2)
1.0d-1	1.2004d-2	2.3979	2.8647d-1	4.3121d-1	7.3975d+2
1.0d-3	-5.4649d-4	-8.5165d-2	-6.8467d-4	-1.1292d-3	3.3262
1.0d-6	-5.6102d-4	-9.0410d-2	-6.6837d-4	-1.1210d-3	-3.6804d-3
1.0d-9	-5.6107d-4	-9.0412d-2	-6.6838d-4	-1.1210d-3	-3.8212d-3
1.0d-12	-5.6107d-4	-9.0412d-2	-6.6838d-4	-1.1210d-3	-3.8213d-3

Let $\mu_2(M(h))$ be the logarithmic norm of $M(h)$, i.e., the largest eigenvalue of $(M(h) + M(h)^*)/2$. From the estimate [7]

$$\Re\langle y, M(h)y\rangle \leq \mu_2(M(h))\|y\|^2, \tag{3.5}$$

the solution decays in the Euclidean norm (i.e., (3.3) is dissipative) when $\mu_2(M(h)) \leq 0$. But a straightforward calculation shows that $\mu_2(M(h)) > 0$ and $\mu_2(M(h)) = 1/(2ch^2) + O(h^{-1})$, providing a bound that permits an exponential instability when h goes to 0.

Since the numerical experiments do not show this catastrophic behavior, and $\mu_2(M(h))$ is the smallest number satisfying (3.5) [7], some kind of weak well-posedness must be present. In fact, the following result is proved in section 4.

THEOREM 3.1. *For every $h > 0$, all the eigenvalues of the matrix $M(h)$ have negative real part.*

A first consequence of Theorem 3.1 is that the solution $U(t)$ of (3.3) goes to zero as $t \to \infty$. However, the matrix $M(h)$ is nonnormal and it is known that in this case the eigenanalysis is not always an efficient means to determine the behavior of an exponential matrix [16, 22]. In fact, for small values of t it is possible that $U(t)$ is unbounded when h goes to zero. To examine this growth, we first make use of the analysis of the ϵ-pseudospectrum, defined in [22] by

$$\Lambda_\epsilon(M(h)) = \{\mu_\epsilon \in \mathbf{C} : \mu_\epsilon \text{ is an eigenvalue of } M(h) + E \text{ for some } E \text{ with } \|E\| \leq \epsilon\}.$$

From Theorem 5.1 in [16], when the estimate

$$\|\exp(tM(h))\| \leq C\exp(\omega t)$$

holds for all $t \geq 0$, we have the following bound for the real parts of the ϵ-pseudoeigenvalues μ_ϵ of $M(h)$:

$$\Re\mu_\epsilon \leq \omega + C\epsilon$$

for all $\epsilon \geq 0$. We have computed these ϵ-pseudoeigenvalues for some typical values of ϵ for an example of section 8 with a moderate size of h ($h = 1/80$). The maximum of the real parts are displayed in the second column of Table 3.1. Notice that all these values are negative and vary slowly for $\epsilon = 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}$, showing that the nonnormality of $M(h)$ is mild for the h considered.

However, it is necessary to estimate the initial growth of the solution of (3.3) when h goes to zero. Suppose that $M(h)$ is diagonalizable for $h > 0$. Let $P(h)$ be

an invertible matrix such that $M(h) = P(h)D(h)P(h)^{-1}$, where $D(h)$ is the diagonal matrix of eigenvalues. Therefore,

$$\begin{aligned}\|\exp(M(h)t)\| &= \left\|P(h)\exp(D(h)t)P(h)^{-1}\right\| \\ &\le \|P(h)\|\,\|\exp(D(h)t)\|\,\left\|P(h)^{-1}\right\| \le \|P(h)\|\,\left\|P(h)^{-1}\right\| = \kappa_h,\end{aligned}$$

where κ_h is the condition number of $P(h)$.

If $M(h)$ is not diagonalizable, the analysis of the behavior of $U(t)$ may be more complicated (see, for example, [11]). However, the diagonalizable case can be considered as generic. Moreover, since $\widetilde{\alpha}(h) = O(h^{-1})$ and $\widetilde{\beta}(h) = O(h^{-1})$ while the rest of the elements of $M(h)$ are $O(h^{-2})$, $M(h)$ can be considered as a perturbation of the diagonalizable matrix M_0, defined as the tridiagonal matrix such that $M_0(j,j) = 2i/ch^2$, $M_0(j,j+1) = M_0(j,j-1) = -i/ch^2$, and the first and last rows vanish. For this matrix, it is straightforward to prove that $M_0 = P_0 D_0 P_0^{-1}$, where D_0 is the diagonal matrix of eigenvalues, with $\sigma(M_0) = \left\{(2i/ch^2)(1-\cos(j\pi/N)) : j = 1,\dots,N-1\right\}\cup\{0\}$. Therefore,

$$\|\exp(M_0 t)\| = \left\|P_0\exp(D_0 t)P_0^{-1}\right\| \le \kappa_0,$$

where $\kappa_0 = \|P_0\|\,\left\|P_0^{-1}\right\|$ is the condition number of the matrix P_0 whose columns are the eigenvectors of M_0. Therefore,

$$P_0 = \begin{bmatrix} \frac{1}{\sqrt{N+1}} & \mathbf{0} & 0 \\ \frac{\mathbf{e}}{\sqrt{N+1}} & C_0 & \frac{\mathbf{d}}{s_N} \\ \frac{1}{\sqrt{N+1}} & \mathbf{0} & \frac{N}{s_N} \end{bmatrix}, \quad P_0^{-1} = \begin{bmatrix} \sqrt{N+1} & \mathbf{0} & 0 \\ \mathbf{a} & C_0 & \mathbf{b} \\ -\frac{s_N}{N} & \mathbf{0} & \frac{s_N}{N} \end{bmatrix},$$

where C_0 is the unitary matrix with elements $C_0(l,j) = \sqrt{2/N}\sin(lj\pi/N)$, $1 \le l,j \le N-1$, $s_N = \sqrt{N(N+1)(2N+1)/6}$, $\mathbf{e} = [1,\dots,1]^T$, $\mathbf{d} = [1,2,\dots,N-1]^T$, $\mathbf{a} = -\frac{1}{N}C_0[N-1,N-2,\dots,1]^T$, and $\mathbf{b} = -\frac{1}{N}C_0\mathbf{d}$.

Now, it is possible to prove that $\|P_0\| = O(1)$ and $\left\|P_0^{-1}\right\| = O(N^{1/2})$, and we deduce that $\kappa_0 = O(N^{1/2})$. By considering κ_0 as an approximation of κ_h, it is reasonable to suppose that $\kappa_h = O(h^{-1/2})$. We have numerically checked this value with an example used in section 8 (see Figure 3.1), obtaining this behavior for κ_h. The matrices $M(h)$ of these numerical tests were always diagonalizable because the eigenvalues were distinct.

4. Results on the eigenvalues of a certain class of matrices. Let us consider a matrix M_{N+1} of dimension $(N+1)\times(N+1)$ with the structure of (3.4) with $\widetilde{\alpha} = \alpha$, $\widetilde{\beta} = \beta$, $\widetilde{m}_j = m_j$, $j = 1,2$, where α, β, m_1, and m_2 are, at first, arbitrary nonzero complex numbers. If λ is an eigenvalue of this matrix with eigenvector $\mathbf{x} = [x_0,\dots,x_N]^T$, the following equation must be satisfied:

$$m_1 x_{j-1} + (m_2 - \lambda)x_j + m_1 x_{j+1} = 0, \quad 1 \le j \le N-1.$$

The solutions of this equation are $x_j = Ay_1^j + By_2^j$ for certain constants A, B, where y_1, y_2 are the roots of

$$y^2 + \frac{m_2 - \lambda}{m_1}y + 1 = 0, \tag{4.1}$$

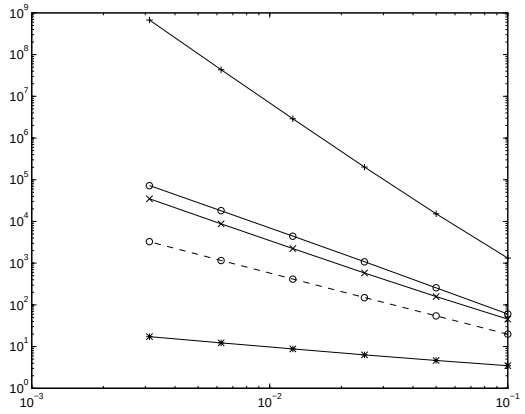


FIG. 3.1. *Condition number as a function of h:* $-*$ $ABC(1,0)$, $-\circ$ $ABC(1,1)$, $-\,-\circ$ $ABC(1,1)$ *(second implementation),* $-\times$ $ABC(2,1)$, $-+$ $ABC(3,2)$.

and then $\lambda = m_2 + m_1(y_1 + y_2)$. The following system also has to be satisfied:

$$\begin{aligned}(\alpha-\lambda)(A+B)+\beta(Ay_1+By_2)=0,\\ \beta(Ay_1^{N-1}+By_2^{N-1})+(\alpha-\lambda)(Ay_1^N+By_2^N)=0.\end{aligned} \tag{4.2}$$

In order that this system has nontrivial solutions A, B, the determinant of its coefficients should vanish. In this way, taking into account that $y_1y_2 = 1$, we obtain

$$(\alpha-\lambda)^2(y_1^N-y_2^N)+2\beta(\alpha-\lambda)(y_1^{N-1}-y_2^{N-1})+\beta^2(y_1^{N-2}-y_2^{N-2})=0. \tag{4.3}$$

Let A_{N-1} be the tridiagonal matrix obtained from M_{N+1} by eliminating the first and last rows and columns. It is a well-known fact that $\sigma(A_{N-1}) = \{m_2+2m_1\cos(\pi j/N) : j = 1,\dots,N-1\}$. Notice that $\sigma(A_{N-1}) \cap \sigma(A_N) = \emptyset$, $N > 0$. Let us denote $R_{N-1}(\lambda) = \det(A_{N-1} - \lambda I_{N-1})$ (where I_{N-1} is the identity matrix of dimension $N-1$). It is clear that

$$\begin{aligned}\det(M_{N+1}-\lambda I_{N+1}) &= (\alpha-\lambda)^2R_{N-1}(\lambda)\\ &\quad -2m_1(\alpha-\lambda)\beta R_{N-2}(\lambda)+\beta^2m_1^2R_{N-3}(\lambda).\end{aligned} \tag{4.4}$$

Notice that $R_N(\lambda) = m_1^N P_N((m_2-\lambda)/m_1)$, where $P_N(x)$ is the monic polynomial of degree N:

$$P_N(x)=\det\begin{bmatrix} x & 1 & 0 & \cdots & 0\\ 1 & x & 1 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & \cdots & 0 & 1 & x\end{bmatrix}.$$

From this expression we deduce that the following recurrence formula is satisfied:

$$P_n(x) = xP_{n-1}(x) - P_{n-2}(x), \quad n = 1,2,\dots, \tag{4.5}$$

with $P_0(x) = 1$ and $P_{-1}(x) = 0$. $R_{N-1}(\lambda)$ and $R_N(\lambda)$ have no common roots, so the same is true for the polynomials $P_N(x)$. Finally, (4.4) reduces to

$$\det(M_{N+1}-\lambda I_{N+1}) = m_1^{N-1}Q_{N+1}\left(\frac{m_2-\lambda}{m_1}\right),$$

where

$$Q_{N+1}(x) = (\alpha - m_2 + m_1 x)^2 P_{N-1}(x) - 2\beta(\alpha - m_2 + m_1 x)P_{N-2}(x) + \beta^2 P_{N-3}(x). \tag{4.6}$$

So we can conclude that λ is an eigenvalue of M_{N+1} iff

$$Q_{N+1}\left(\frac{m_2 - \lambda}{m_1}\right) = 0.$$

In the rest of this section, we will assume that the coefficients m_1 and m_2 of M_{N+1} are $m_1 = -i$, $m_2 = i(2-\delta)$.

LEMMA 4.1. *Let us consider the matrix M_{N+1} previously defined, with α, $\beta \in \mathbf{C}$ and $\delta \in \mathbf{R}$. We shall use the notation $\alpha_r = \Re(\alpha)$, $\alpha_i = \Im(\alpha)$, $\beta_r = \Re(\beta) \neq 0$, $\beta_i = \Im(\beta) \neq 0$, where these coefficients satisfy one of the following properties:*

$$|\beta| \leq |\alpha_r|, \tag{4.7}$$

$$|\beta| > |\alpha_r| \quad \textit{and} \quad \sqrt{|\beta|^2 - \alpha_r^2} < \delta + \alpha_i < 4 - \sqrt{|\beta|^2 - \alpha_r^2}. \tag{4.8}$$

We will also assume the following hypotheses:

$$\delta + \alpha_i < \frac{\alpha_r \beta_i}{\beta_r}, \tag{4.9}$$

$$(\delta + \alpha_i)(2 + \beta_i) < -\alpha_r\beta_r - 2\beta_i. \tag{4.10}$$

Then the matrix M_{N+1} has no imaginary eigenvalues.

Proof. Let us suppose there exists an imaginary eigenvalue λ of M, $\lambda = i\lambda_0$ with λ_0 real. The Gersgorin disks are $D_1 = \{x \in \mathbf{C} : |x - i(2-\delta)| \leq 2\}$ and $D_2 = \{x \in \mathbf{C} : |x - \alpha| \leq |\beta|\}$. If (4.7) is satisfied, D_2 has no imaginary eigenvalues except maybe $i\alpha_i$. But if this point was an eigenvalue of M_{N+1}, it should be also in the boundary of D_1 (M_{N+1} is irreducible; see the theorem of Taussky [14]). Therefore, $-\delta \leq \lambda_0 \leq 4 - \delta$. Otherwise, if $|\beta| > |\alpha_r|$, D_2 intersects the imaginary axis at the points $i(\alpha_i \pm \sqrt{|\beta|^2 - \alpha_r^2})$ that are in D_1 because of (4.8), then we obtain this way the same condition for λ_0, $-\delta \leq \lambda_0 \leq 4 - \delta$.

On the other hand, as we have already seen, if λ is an eigenvalue of M_{N+1}, $Q_{N+1}(X_0) = 0$, where $X_0 = -2 + \delta + \lambda_0 \in \mathbf{R}$, so X_0 should satisfy $-2 \leq X_0 \leq 2$. Making use of the recurrence formula (4.5) in (4.6),

$$\begin{aligned} Q_{N+1}(x) = {} & (-\beta^2 + (\alpha_r - i\gamma)^2 - 2i(\alpha_r - i\gamma)x - x^2)P_{N-1}(x) \\ & - (2\beta(\alpha_r - i\gamma) - (2i\beta + \beta^2)x)P_{N-2}(x), \end{aligned}$$

where $\gamma = 2 - \delta - \alpha_i$. We can write $Q_{N+1}(x) = Q^r_{N+1}(x) + iQ^i_{N+1}(x)$ where $Q^r_{N+1}(x)$ and $Q^i_{N+1}(x)$ are real polynomials. Since X_0 is a real root of $Q_{N+1}(x)$, it will be a root of $Q^r_{N+1}(x)$ and $Q^i_{N+1}(x)$:

$$\begin{aligned} 0 = Q^r_{N+1}(X_0) = {} & (-\beta_r^2 + \beta_i^2 + \alpha_r^2 - \gamma^2 - 2\gamma X_0 - X_0^2)P_{N-1}(X_0) \\ & - (2\alpha_r\beta_r + 2\gamma\beta_i - (-2\beta_i + \beta_r^2 - \beta_i^2)X_0)P_{N-2}(X_0), \end{aligned} \tag{4.11}$$

$$\begin{aligned} 0 = Q^i_{N+1}(X_0) = {} & 2(-\beta_r\beta_i - \alpha_r\gamma - \alpha_r X_0)P_{N-1}(X_0) \\ & - 2(\alpha_r\beta_i - \gamma\beta_r - (\beta_r + \beta_r\beta_i)X_0)P_{N-2}(X_0). \end{aligned} \tag{4.12}$$

As we have previously mentioned, $P_{N-1}(x)$ and $P_{N-2}(x)$ have no common roots. If we consider (4.11)–(4.12) as a system with variables $P_{N-1}(X_0)$ and $P_{N-2}(X_0)$, in order that it has a nontrivial solution, it should be

$$\begin{aligned}(4.13)\quad 0 &= (-\beta_r^2+\beta_i^2+\alpha_r^2-\gamma^2-2\gamma X_0-X_0^2)(\alpha_r\beta_i-\gamma\beta_r-(\beta_r+\beta_r\beta_i)X_0)\\ &\quad-(2\alpha_r\beta_r+2\gamma\beta_i-(-2\beta_i+\beta_r^2-\beta_i^2)X_0)(-\beta_r\beta_i-\alpha_r\gamma-\alpha_r X_0).\end{aligned}$$

Let us suppose first that $\beta_i\neq -1$. The roots of this equation are

$$X_0^a=-\gamma-\frac{\alpha_r\beta_i}{\beta_r},\quad X_0^{b,c}=\frac{A\pm\sqrt{A^2-B}}{D}$$

with $A=\alpha_r\beta_r-\gamma(2+\beta_i)$, $B=4(1+\beta_i)(\alpha_r^2+\beta_i^2+\beta_r^2+\gamma^2)$, $D=2(1+\beta_i)$.

Because of (4.9), $X_0^a<-2$, which is not possible. If $A^2-B<0$, $X_0^{b,c}\notin\mathbf{R}$, concluding the proof. Let us suppose then that $A^2-B\geq 0$. We consider first the case $\beta_i+1>0$. Let us prove $X_0^b<-2$, which is equivalent to $A+2D<-\sqrt{A^2-B}$. Making use of (4.10), we see that $A+2D<0$, and then

$$A+2D<-\sqrt{A^2-B}\quad\text{iff}\quad A^2+4AD+4D^2=|A+2D|^2>A^2-B.$$

It remains only to prove that

$$4AD+4D^2+B=4(1+\beta_i)\phi_1(\delta)>0,$$

where $\phi_1(\delta)=(\alpha_i+\beta_i)^2+(\alpha_r+\beta_r)^2+2\delta(\alpha_i+\beta_i)+\delta^2$. Notice that $\phi_1(\delta)$ goes to ∞ as $\delta\to\pm\infty$. The roots of $\phi_1(\delta)$ are $-\alpha_i-\beta_i\pm\sqrt{-(\alpha_r+\beta_r)^2}$. If $|\beta|\leq|\alpha_r|$, since $\beta_i\neq 0$, it should be $\alpha_r+\beta_r\neq 0$, and then $\phi_1(\delta)>0$ for all $\delta\in\mathbf{R}$. Otherwise, if (4.8) is satisfied, we see that the only case when we cannot guarantee that $\phi_1(\delta)>0$ is for $\delta=-\alpha_i-\beta_i$ when $\alpha_r+\beta_r=0$. But then $\sqrt{|\beta|^2-\alpha_r^2}=|\beta_i|\not<\delta+\alpha_i=-\beta_i$, which goes against hypothesis (4.8). So we obtain that $\phi_1(\delta)>0$, and then $X_0^b<-2$, which is not possible. Similarly, $X_0^c<-2$ iff $A+2D<\sqrt{A^2-B}$ which is true since $A+2D<0$ from (4.10).

Let us now consider the case $\beta_i+1<0$. It can be proved that $X_0^b<-2$ with a reasoning analogous to that for $\beta_i+1>0$. Let us see $X_0^c>2$, which is equivalent to $A-2D<\sqrt{A^2-B}$. If $A-2D<0$, this is trivially true. Otherwise,

$$A-2D<\sqrt{A^2-B}\quad\text{iff}\quad -4AD+4D^2+B=4(1+\beta_i)\phi_2(\delta)<0,$$

where $\phi_2(\delta)=(4-\alpha_i)^2+(\alpha_r-\beta_r)^2+8\beta_i-2\alpha_i\beta_i+\beta_i^2+2(\alpha_i-4-\beta_i)\delta+\delta^2$. It can be proved that $\phi_2(\delta)>0$ in a manner similar to that for ϕ_1, concluding $X_0^c>2$.

Finally, if $\beta_i=-1$, (4.13) reduces to

$$\begin{aligned}(4.14)\qquad 0 &= (-\beta_r^2+1+\alpha_r^2-\gamma^2-2\gamma X_0-X_0^2)(-\alpha_r-\gamma\beta_r)\\ &\quad-(2\alpha_r\beta_r-2\gamma-(1+\beta_r^2)X_0)(\beta_r-\alpha_r\gamma-\alpha_r X_0),\end{aligned}$$

whose roots are

$$X_0^d=\frac{\alpha_r}{\beta_r}-\gamma,\qquad X_0^e=\frac{1+\alpha_r^2+\beta_r^2+\gamma^2}{\alpha_r\beta_r-\gamma}.$$

From hypothesis (4.9) for $\beta_i=-1$, we deduce that $X_0^d<-2$. Let us see that $X_0^e<-2$. Taking into account (4.10), we have that $\alpha_r\beta_r-\gamma<0$, so $X_0^e<-2$ iff

$$1+\alpha_r^2+\beta_r^2+\gamma^2+2\alpha_r\beta_r-2\gamma=(\alpha_r+\beta_r)^2+(1-\gamma)^2>0,$$

which is true unless $\alpha_r = -\beta_r$ and $\delta + \alpha_i = 1$. It can be seen that this cannot happen if (4.7) or (4.8) is satisfied. Therefore, (4.14) cannot be satisfied, and then M_{N+1} cannot have any imaginary eigenvalues. ∎

LEMMA 4.2. *Let $M_{N+1}(h)$ be the matrix M_{N+1} where the coefficients $\alpha(h)$, $\beta(h)$, $\delta(h)$ are functions regular enough in a neighborhood $\mathcal{V}$ of zero and the dimension of the matrix is independent of h. Let us suppose that there exist $a_1, b_1, d_1 \in \mathbf{R}$ such that*

$$\alpha(h) = h\alpha_0 + h^2\alpha_1(h), \quad \textit{where} \quad \alpha_0 < 0, \quad \alpha_1(0) = ia_1, \tag{4.15}$$

$$\beta(h) = -h\alpha_0 + h^2\beta_1(h), \quad \textit{where} \quad \beta_1(0) = ib_1, \tag{4.16}$$

$$\delta(h) = h^2\delta_1(h) \in \mathbf{R}, \quad \textit{where} \quad \delta_1(0) = d_1, \tag{4.17}$$

$$0 > a_1 + b_1 + d_1, \tag{4.18}$$

$$0 > \alpha_0(1-N)(a_1 + b_1 + d_1) + 2\Re(\alpha_1'(0) + \beta_1'(0)). \tag{4.19}$$

Then there exists $h_0 > 0$ such that for all $h \in (0, h_0)$, the eigenvalues of $M_{N+1}(h)$ have negative real part.

Proof. Let λ be an eigenvalue of $M_{N+1}(h)$ with eigenvector $\mathbf{x} = [x_0, \dots, x_N]^T$. As we have remarked previously, $x_j = Ay_1^j + By_2^j$ where y_1 and y_2 are the roots of (4.1) and $\lambda = i(2 - \delta(h) - y_1 - y_2)$. We observe this way that

$$\Re(\lambda) = \Im(y_1)\left(1 - \frac{1}{|y_1|^2}\right) < 0 \quad \text{iff} \quad \begin{cases} \Im(y_1) < 0 & \text{and} \quad |y_1| > 1 \quad \text{or} \\ \Im(y_1) > 0 & \text{and} \quad |y_1| < 1, \end{cases} \tag{4.20}$$

where we have used that $y_1 y_2 = 1$. Equation (4.3) also has to be satisfied (where the coefficients now depend on h). If we multiply this expression by y_1^{N+2}, we obtain $T(y_1, h) = 0$, with

$$\begin{aligned} T(y,h) = {} & \alpha(h)^2(y^{2N+2} - y^2) - 2i\alpha(h)((2-\delta(h))y - y^2 - 1)(y^{2N+1} - y) \\ & - ((2-\delta(h))y - y^2 - 1)^2(y^{2N} - 1) + 2\alpha(h)\beta(h)(y^{2N+1} - y^3) \\ & - 2i\beta(h)((2-\delta(h))y - y^2 - 1)(y^{2N} - y^2) + \beta(h)^2(y^{2N} - y^4). \end{aligned}$$

Our purpose is to study whether the roots of $T(y,h)$ satisfy (4.20) for h small enough. Let us notice that $T(y,0) = -(y-1)^4(y^{2N} - 1)$, and therefore, the roots of $T(y,0)$ are $y = 1$ and $\{\exp(i\pi j/N) : j = 0, \dots, 2N-1\}$. Taking into account that $\alpha(0) = \beta(0) = \delta(0) = 0$, we get

$$\partial_y T(y,0) = -2N(y-1)^4 y^{2N-1} + 4(y-1)^3(1 - y^{2N}),$$

so for $j \in \{1, \dots, 2N-1\}$, $T(\exp(i\pi j/N), 0) = 0$ and $\partial_y T(\exp(i\pi j/N), 0) \neq 0$. Then we can apply the implicit function theorem: there exist neighborhoods $\mathcal{A}_j \subset \mathcal{V}$ of 0 and $\mathcal{B}_j$ of $\exp(i\pi j/N)$ and a unique regular function $y_j : \mathcal{A}_j \to \mathcal{B}_j$ such that

$$T(y_j(h), h) = 0 \quad \text{for all} \quad h \in \mathcal{A}_j \quad \text{and} \quad y_j(0) = \exp(i\pi j/N).$$

In a neighborhood of 0, $y_j(h) = \exp(i\pi j/N) + hy_j'(0) + O(h^2)$. Therefore,

$$\begin{aligned} |y_j(h)|^2 &= |\exp(i\pi j/N) + hy_j'(0) + O(h^2)|^2 \\ &= 1 + 2h(\Re(y_j'(0))\cos(\pi j/N) + \Im(y_j'(0))\sin(\pi j/N)) + O(h^2). \end{aligned}$$

Let us now study $y_j'(0)$. Taking into account that $\beta'(0) = -\alpha'(0)$, $\delta'(0) = 0$, $y_j(0)^{2N} = 1$, and $y_j(0)^{-1} = \bar{y}_j(0)$,

$$y_j'(0) = \frac{-\partial_h T(y_j(0),0)}{\partial_y T(y_j(0),0)} = \frac{-2i\alpha'(0)(y_j(0)+y_j(0)^{2N})}{-2N(y_j(0)-1)y_j(0)^{2N-1}+4(1-y_j(0)^{2N})}$$

$$= \frac{i\alpha'(0)(y_j(0)+1)}{N(y_j(0)-1)y_j(0)^{-1}} = \frac{i\alpha_0(y_j(0)+1)}{N(1-\bar{y}_j(0))} = \frac{i\alpha_0(1-y_j(0)^2)}{N|1-\bar{y}_j(0)|^2},$$

and then

$$\Re(y_j'(0)) = \frac{2\alpha_0\cos(\pi j/N)\sin(\pi j/N)}{N|1-\bar{y}_j(0)|^2}, \quad \Im(y_j'(0)) = \frac{2\alpha_0\sin^2(\pi j/N)}{N|1-\bar{y}_j(0)|^2}.$$

Since $\alpha_0 < 0$ from (4.15), we deduce that $\Im(y_j'(0)) < 0$ for $0 < j < 2N$, $j \neq N$.

If $0 < j < \lfloor \frac{N}{2} \rfloor$, $\cos(\pi j/N)\sin(\pi j/N) > 0$ and then $\Re(y_j'(0)) < 0$. We have $\Re(y_j'(0))\cos(\pi j/N) + \Im(y_j'(0))\sin(\pi j/N) < 0$ this way, and therefore, $|y_j(h)|^2 < 1$ and $\Im(y_j(h)) > 0$ for h small enough. Reasoning this way for the rest of the cases, we can conclude that there exists $h_0 > 0$ such that, for all $h \in (0, h_0)$,

(i) if $0 < j < N$, $\Im(y_j(h)) > 0$ and $|y_j(h)| < 1$;
(ii) if $N < j < 2N$, $\Im(y_j(h)) < 0$ and $|y_j(h)| > 1$.

For $j = N$, $y_j(0) = -1$. It can be checked that $T(-1,h) = 0$, so $y_N \equiv -1$ because of the uniqueness of y_N. Nevertheless, it does not correspond to an eigenvalue of $M_{N+1}(h)$. Otherwise, the associated eigenvalue $\lambda = i(2-\delta(h)) + 2i = 4i - i\delta(h)$ would have the eigenvector $\mathbf{x}$ with $x_j = (-1)^j(A+B)$, and from (4.2),

$$(\alpha(h) - \lambda)(A+B) - \beta(h)(A+B) = 0,$$

so $A + B = 0$, since $\alpha(h) - \lambda - \beta(h) = -4i + 2h\alpha_0 + h^2(\alpha_1(h) + i\delta_1(h) - \beta_1(h)) \neq 0$ for h small enough.

Finally, let us consider the zero $y = 1$ of $T(y,0)$. We cannot use the implicit function theorem directly because $\partial_y T(1,0) = 0$. Let us define r such that $y = 1 + rh$, obtaining in this way the function $t(r,h)$ defined by

$$\begin{aligned} t(r,h) = \frac{1}{h^4}T(1+rh,h) &= (r^2 + (1+rh)(i\alpha_0 r - i\alpha_1(h) - i\beta_1(h)(1+rh) + \delta_1(h)))^2 \\ &\quad - (1+hr)^{2N}(-i\alpha_0 r + r^2 - i\beta_1(h) + (1+rh)(-i\alpha_1(h) + \delta_1(h)))^2 \\ &= 4i\alpha_0 r(a_1 + b_1 + d_1 + r^2) + h(-2nr(a_1 + b_1 + d_1 - i\alpha_0 r + r^2)^2 \\ &\quad + 2r(2b_1 + i\alpha_0 r)(a_1 + b_1 + d_1 + r^2) + 2i\alpha_0 r(2a_1 r + 2b_1 r + 2d_1 r + i\alpha_0 r^2) \\ &\quad + 2i\alpha_0 r(-2i\alpha_1'(0) - 2i\beta_1'(0) + 2\delta_1(0))) + O(h^2). \end{aligned} \tag{4.21}$$

We are going to use the implicit function theorem for $t(r,h)$. The roots of $t(r,0)$ are $r_0 = 0$, $r_1 = \sqrt{-a_1 - b_1 - d_1}$, and $r_2 = -\sqrt{-a_1 - b_1 - d_1}$. Making use of (4.18),

$$\partial_r t(r_0,0) = 4\alpha_0(\alpha_1(0) + \beta_1(0) + i\delta_1(0)) = 4i\alpha_0(a_1 + b_1 + d_1) \neq 0,$$

$$\partial_r t(r_1,0) = \partial_r t(r_2,0) = -8i\alpha_0(a_1 + b_1 + d_1) \neq 0.$$

Therefore, for $j = 0, 1, 2$, there exist neighborhoods $\mathcal{U}_j \subset \mathcal{V}$ of zero and $\mathcal{W}_j$ of r_j and a unique regular function $r_j : \mathcal{U}_j \to \mathcal{W}_j$ such that

$$t(r_j(h), h) = 0 \quad \text{for all} \quad h \in \mathcal{U}_j, \quad \text{and} \quad r_j(0) = r_j.$$

Let us define $\phi_j(h) = 1 + hr_j(h)$ for $j = 0, 1, 2$. For these functions,

$$T(\phi_j(h), h) = T(1 + hr_j(h), h) = h^4 t(r_j(h), h) = 0 \quad \text{and} \quad \phi_j(0) = 1.$$

For $j = 0, 1, 2$, $\phi_j(h) = 1 + h\phi_j'(0) + (h^2/2)\phi_j''(0) + O(h^3)$. Let us study each of the cases:

1. $\phi_1(h) = 1 + hr_1(h)$. Taking into account (4.21), (4.18), (4.19), and $\delta_1(h) \in \mathbf{R}$,

$$\phi_1'(0) = r_1(0) = \sqrt{-a_1 - b_1 - d_1} > 0,$$

$$\Im(\phi_1''(0)) = \Im(2r_1'(0)) = -\Im\left(\frac{2\partial_h t(r_1, 0)}{\partial_r t(r_1, 0)}\right) = \frac{-\Re(\partial_h t(r_1, 0))}{4\alpha_0(a_1 + b_1 + d_1)}$$

$$= \frac{\alpha_0(1 - N)(a_1 + b_1 + d_1) + 2\Re(\alpha_1'(0) + \beta_1'(0))}{2\sqrt{-a_1 - b_1 - d_1}} < 0.$$

Therefore, for $h > 0$ small enough, $\Im(\phi_1(h)) < 0$ and

$$|\phi_1(h)|^2 = \left|1 + h\phi_1'(0) + \frac{h^2}{2}\phi_1''(0) + O(h^3)\right|^2 = 1 + 2h\phi_1'(0) + O(h^2) > 1.$$

2. Similarly for $\phi_2(h) = 1 + hr_2(h)$, we obtain $\Im(\phi_2(h)) > 0$ and $|\phi_2(h)|^2 < 1$ for $h > 0$ small enough.

3. Finally, notice that $t(0, h) = 0$ for all h, so $r_0 \equiv 0$ and $\phi_0 \equiv 1$. Nevertheless, the root $y = 1$ of $T(y, h)$ does not give rise to any eigenvalue of $M_{N+1}(h)$. This can be proved similarly as we have done previously for the root $y = -1$. ∎

THEOREM 4.3. *With the notation of Lemma* 4.2 *let us consider the matrix* $M_{N+1}(h)$ *whose coefficients satisfy* (4.15)–(4.19)*, and let us suppose that the hypotheses of Lemma* 4.1 *are satisfied for all* $h > 0$ *(where now the coefficients are functions of* h*). Then, for every* $h > 0$*, all the eigenvalues of* $M_{N+1}(h)$ *have negative real part.*

Proof. Lemma 4.2 asserts that there exists h_0 such that for $h \in (0, h_0)$, all the eigenvalues of $M_{N+1}(h)$ have negative real part. Lemma 4.1 asserts that for $h > 0$, $M_{N+1}(h)$ has not imaginary eigenvalues. On the other hand, the eigenvalues of $M_{N+1}(h)$ are continuous functions of h, so if there were $h^* \geq h_0$ such that $M_{N+1}(h^*)$ had an eigenvalue with real part greater or equal to zero, there should exist $h \in [h_0, h^*]$ such that $M_{N+1}(h)$ had an imaginary eigenvalue, which is not possible. ∎

Proof of Theorem 3.1. It suffices to prove that all the eigenvalues of $ch^2 M(h)$ have negative real part. Notice that $ch^2 M(h)$ is the matrix $M_{N+1}(h)$ previously considered with $\alpha(h) = ch^2\widetilde{\alpha}(h)$, $\beta(h) = ch^2\widetilde{\beta}(h)$, and $\delta(h) = h^2 V$. These coefficients satisfy (4.15)–(4.17) for

$$\alpha_0 = -(s_1 + s_2) < 0,\ a_1 = \frac{s_1^2 + s_2^2}{2} - V \in \mathbf{R},\ b_1 = \frac{-(s_1 + s_2)^2}{2} \in \mathbf{R},\ d_1 = V \in \mathbf{R}.$$

We also have

$$\alpha_1'(0) = \frac{1}{4}(s_1^3 + s_1^2 s_2 + s_1 s_2^2 + s_2^3), \quad \beta_1'(0) = -\frac{1}{4}(s_1 + s_2)^3.$$

Then, hypotheses (4.18) and (4.19) are satisfied,

$$a_1 + b_1 + d_1 = -s_1 s_2 < 0,$$

$$\alpha_0(1 - N)(a_1 + b_1 + d_1) + 2\Re(\alpha_1'(0) + \beta_1'(0)) = -N s_1 s_2 (s_1 + s_2) < 0.$$

Let us see now that the hypotheses of Lemma 4.1 are satisfied. With the previous notation,

$$\alpha_r(h) = \frac{-2h(s_1+s_2)(2+s_1s_2h^2)}{4+h^2(s_1+s_2)^2}, \quad \alpha_i(h) = \frac{h^2(2(s_1^2+s_2^2)-4V-h^2V(s_1+s_2)^2)}{4+h^2(s_1+s_2)^2},$$

$$\beta_r(h) = \frac{4}{h(s_1+s_2)+\frac{4}{h(s_1+s_2)}}, \quad \beta_i(h) = \frac{-2}{1+\frac{4}{h^2(s_1+s_2)^2}}.$$

For these coefficients we have, omitting the dependence on h in the notation,

$$|\beta|^2 - |\alpha_r|^2 = \frac{4h^4(s_1+s_2)^2((s_1-s_2)^2-h^2s_1^2s_2^2)}{(4+h^2(s_1+s_2)^2)^2},$$

so $|\beta| \le |\alpha_r|$ for $h \ge |s_1-s_2|/s_1s_2$ and (4.7) is satisfied. Let us see (4.8) is satisfied if $0 < h < |s_1-s_2|/s_1s_2$. It can be checked that $\delta - 4 + \alpha_i < 0$; therefore, $\delta - 4 + \alpha_i < -\sqrt{|\beta|^2-\alpha_r^2}$ iff $(\delta-4+\alpha_i)^2 > |\beta|^2 - \alpha_r^2$, which is satisfied:

$$(\delta-4+\alpha_i)^2 - |\beta|^2 + \alpha_r^2 = \frac{4(4+s_1s_2h^2)^2}{4+h^2(s_1+s_2)^2} > 0.$$

Similarly, $-\alpha_i + \sqrt{|\beta|^2-\alpha_r^2} < \delta$ iff $(\delta+\alpha_i)^2 > |\beta|^2 - \alpha_r^2$, which is true:

$$(\delta+\alpha_i)^2 - |\beta|^2 + \alpha_r^2 = \frac{4s_1^2s_2^2h^4}{4+h^2(s_1+s_2)^2} > 0.$$

Therefore, (4.8) is satisfied.

The following relations are true for every h:

$$\delta + \alpha_i - \frac{\alpha_r\beta_i}{\beta_r} = -h^2s_1s_2 < 0,$$

$$(\delta+\alpha_i)(2+\beta_i) + \alpha_r\beta_r + 2\beta_i = \frac{-4h^2(s_1^2+4s_1s_2+s_2^2)}{4+h^2(s_1+s_2)^2} < 0.$$

Therefore, hypotheses (4.9) and (4.10) are satisfied.

Therefore, Theorem 4.3 asserts that for all $h > 0$ the eigenvalues of $M(h)$ (where $h = L/N$) have negative real part. □

5. Time discretization. We now study the full discrete method obtained by using the method of lines. For this, it is necessary to apply a time integration method to solve numerically the ordinary differential system (3.3). Although there are many available integration schemes, notice that (3.3) is stiff and the stiffness grows when h, the parameter of the spatial discretization, goes to zero. Moreover, there exist several eigenvalues of the matrix $M(h)$ with real part nearly 0 but with large imaginary parts. Therefore, it is convenient to use A-stable methods. As a consequence, the widely used backward differentiation formulae methods are not suited to solve (3.3), and we consider only implicit Runge–Kutta methods in this paper.

Let us take a Runge–Kutta method of order p and denote by $r(z)$ its stability function. We suppose that the Runge–Kutta method is A-stable, i.e., $|r(z)| \le 1$ for $\Re(z) \le 0$. Moreover, since the order is p, $r(z)$ is a rational approximation to $\exp(z)$ of order p, i.e.,

$$\exp(z) - r(z) = Cz^{p+1} + O(z^{p+2}) \quad \text{as} \quad z \to 0,$$

where $C \neq 0$ is the error constant. We remark that this constant may be dependent on the parameter k, causing the well-known order reduction phenomenon [7]. Therefore, the order observed in practice is smaller than the classical order p when the time stepsize is not small enough. We do not study in detail this problem here; nevertheless, we remark that the semidiscrete problems arising when we use higher order ABCs are usually stiffer.

By applying the Runge–Kutta method with stepsize k to (3.3), we obtain approximations $U_n = r^n(kM(h))U(0)$ to the values $U(nk) = U(t_n) = \exp(t_n M(h))U(0)$, $0 \leq n \leq N$, with U being the solution of (3.3). Our main objective is to study the troubles caused by the weak ill-posedness of the semidiscrete problems (3.3) to the convergence of these approximations. Notice that we do not take into account the error due to the spatial approximation and the ABCs.

We denote by $e_n = U(t_n) - r^n(kM(h))U(0)$ the global errors and by $\rho_n = U(t_n) - r(kM(h))U(t_{n-1})$ the local errors. Then

$$\begin{aligned} e_{n+1} &= U(t_{n+1}) - r(kM(h))U(t_n) + r(kM(h))U(t_n) - r^{n+1}(kM(h))U(0) \\ &= \rho_{n+1} + r(kM(h))e_n, \end{aligned}$$

and, assuming that $M(h)$ is diagonalizable and by using a recursion argument, we deduce that

$$\begin{aligned} e_n &= \sum_{j=1}^{n} r^{n-j}(kM(h))\rho_j = \sum_{j=1}^{n} r^{n-j}(kM(h))(U(t_j) - r(kM(h))U(t_{j-1})) \\ &= \sum_{j=1}^{n} r^{n-j}(kM(h))(\exp(kM(h)) - r(kM(h)))\exp(t_{j-1}M(h))U(0) \\ &= \sum_{j=1}^{n} P(h)r^{n-j}(kD(h))(\exp(kD(h)) - r(kD(h)))\exp(t_{j-1}D(h))P(h)^{-1}U(0). \end{aligned}$$

Taking the Euclidean norm, we deduce

$$\begin{aligned} \|e_n\| &\leq \kappa_h \sum_{j=1}^{n} \|r^{n-j}(kD(h))(\exp(kD(h)) - r(kD(h)))\exp(t_{j-1}D(h))\|\|U(0)\| \\ &\leq \kappa_h \sum_{j=1}^{n} \sup_{\Re z \leq 0} \{|r^{n-j}(z)||\exp(t_{j-1}z)|\} \sup_{z\in\sigma(D(h))} |\exp(kz) - r(kz)|\|U(0)\| \\ &\leq \kappa_h \sum_{j=1}^{n} \sup_{z\in\sigma(D(h))} \{|\exp(kz) - r(kz)|\}\|U(0)\| = O(\kappa_h k^p). \end{aligned}$$

From this estimate, we conclude that the global error may grow when h goes to zero and k is fixed because of the growth of κ_h. Nevertheless, this possible growth is made up for the presence of the factor k^p, which goes to zero when the stepsize k goes to zero. Therefore, it is very convenient to use Runge–Kutta methods of high order when the factor κ_h is ill-behaved. Otherwise, the stepsize k must be very small.

6. Discretization of other ABC. We are going to propose spatial approximations of some ABC defined in section 2. Notice that it is possible to consider several spatial approximations for each one. Therefore, the properties of the corresponding ordinary differential system analogous to (3.3) depend on both the ABCs and the spatial discretization. As an example, we propose two different discretizations of ABC(1,1).

ABC(1,1). First, let us define the new function $v(x,t) = \partial_t u(x,t)$ and denote by $v^N(t)$ an approximation of $v(x^N,t)$. Taking ABC(1,1) and (3.2) into account,

$$\begin{aligned} v^N &\approx \frac{d}{dt}u^N \approx \frac{-i}{c}(\partial_x^2 u(x^N,t) + Vu^N) \\ &\approx \frac{-2i}{ch^2}\left(u^{N-1} - u^N - \frac{h}{\beta_2}(\beta_0 u^N + \beta_1 v^N + \beta_3 \partial_x v(x^N,t))\right) - \frac{iV}{c}u^N, \end{aligned}$$

where we have omitted the dependence on t in the notation. This allows us to obtain $\partial_x v(x^N,t) \approx \gamma_0 u^{N-1} + \gamma_1 u^N + \gamma_2 v^N$. Finally, making use of this expression,

$$\begin{aligned} \frac{d}{dt}v^N &\approx \frac{-i}{c}(\partial_x^2 v(x^N,t) + Vv^N) \approx \frac{-2i}{ch^2}\left(\frac{d}{dt}u^{N-1} - v^N + h\partial_x v(x^N,t)\right) - \frac{iV}{c}v^N \\ &\approx \frac{-2i}{ch^2}(\tilde{m}_1(u^{N-2}+u^N) + \tilde{m}_2 u^{N-1} - v^N + h(\gamma_0 u^{N-1} + \gamma_1 u^N + \gamma_2 v^N)) \\ &\quad - \frac{iV}{c}v^N = \delta_0 v^N + \delta_1 u^N + \delta_2 u^{N-1} + \delta_3 u^{N-2} \end{aligned}$$

for certain coefficients δ_j. We have obtained in this way a system $U' = MU$ with $U = [v^0, u^0, \ldots, u^N, v^N]^T$ and

$$M = \begin{bmatrix} \delta_0 & \delta_1 & \delta_2 & \delta_3 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \tilde{m}_1 & \tilde{m}_2 & \tilde{m}_1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & \tilde{m}_1 & \tilde{m}_2 & \tilde{m}_1 & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 1 \\ 0 & \cdots & 0 & \delta_3 & \delta_2 & \delta_1 & \delta_0 \end{bmatrix}. \tag{6.1}$$

Let us implement ABC(1,1) in a different way. As for the previous implementation, we consider (3.2) so that

$$\frac{d}{dt}u^N \approx \frac{-2i}{ch^2}\left(u^{N-1} - u^N - \frac{h}{\beta_2}\left(\beta_0 u^N + \beta_1 \frac{d}{dt}u^N + \beta_0 \partial_{xt} u(x^N,t)\right)\right) - \frac{iV}{c}u^N.$$

Using the approximation

$$\partial_{xt} u(x^N,t) \approx \frac{3\frac{d}{dt}u^N - 4\frac{d}{dt}u^{N-1} + \frac{d}{dt}u^{N-2}}{2h}$$

and taking (3.1) into account, we obtain

$$\frac{d}{dt}u^N = \gamma_0 u^N + \gamma_1 u^{N-1} + \gamma_2 u^{N-2} + \gamma_3 u^{N-3}$$

for certain coefficients γ_j. Therefore, in this case, we have a system with matrix

$$M = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & 0 & \cdots & 0 \\ \tilde{m}_1 & \tilde{m}_2 & \tilde{m}_1 & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & 0 & \tilde{m}_1 & \tilde{m}_2 & \tilde{m}_1 \\ 0 & \cdots & 0 & \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}. \tag{6.2}$$

A study similar to the one done previously for ABC(1,0) could be done for ABC(1,1). We have checked numerically that matrices (6.1) and (6.2) have their eigenvalues with negative real part and the ϵ-pseudoeigenvalues are close to the eigenvalues for some typical values of ϵ. These numerical results are displayed in Table 3.1. Nevertheless, the condition number κ_h of the corresponding matrix P_h behaves with h worse than it did for matrix (3.4). In Figure 3.1, we can observe that for matrix (6.1) κ_h is approximately $O(h^{-2})$ and for (6.2) it is $O(h^{-3/2})$.

ABC(2,1). Continuing this approach and using the notation $v(x,t) = \partial_t u(x,t)$, for the ABC(2,1) we have

$$v^N \approx \frac{-i}{c}(\partial_x^2 u(x^N,t) + Vu^N) \approx \frac{-2i}{ch^2}\Bigg(u^{N-1} - u^N - \frac{h}{\beta_3}\left(\beta_0 u^N + \beta_1 v^N + \beta_2 \frac{d}{dt}v^N + \beta_4 \partial_x v(x^N,t)\right)\Bigg) - \frac{iV}{c}u^N.$$

Solving last equation for $\partial_x v^N$, we get

$$\partial_x v(x^N,t) \approx \gamma_0 u^{N-1} + \gamma_1 u^N + \gamma_2 v^N + \gamma_3 \frac{d}{dt}v^N$$

for certain coefficients γ_j. This expression allows us to obtain

$$\frac{d}{dt}v^N \approx \frac{-i}{c}(\partial_x^2 v(x^N,t) + Vv^N) \approx \frac{-2i}{ch^2}\left(\frac{d}{dt}u^{N-1} - v^N + h\partial_x v(x^N,t)\right) - \frac{iV}{c}v^N$$

$$\approx \frac{-2i}{ch^2}\Bigg(\tilde{m}_1(u^{N-2} + u^N) + \tilde{m}_2 u^{N-1} - v^N + h\left(\gamma_0 u^{N-1} + \gamma_1 u^N + \gamma_2 v^N + \gamma_3 \frac{d}{dt}v^N\right)\Bigg) - \frac{iV}{c}v^N,$$

where an approximation similar to (3.2) has been used for $\partial_x^2 v(x^N,t)$. Finally, we obtain

$$\frac{d}{dt}v^N = \delta_0 v^N + \delta_1 u^N + \delta_2 u^{N-1} + \delta_3 u^{N-2}.$$

Similar to previous cases, after discretization in space, we have to solve a system whose matrix is of the form (6.1) (with different coefficients). We have checked numerically (see Table 3.1) that all the ϵ-pseudoeigenvalues of this matrix have negative real part for some typical values of ϵ smaller or equal to 1.0d-3, and in Figure 3.1 we can observe that $\kappa_h = O(h^{-2})$.

For the discretization in time of these ordinary differential systems, we may use the same analysis of section 5. We remember that it is convenient to use higher order Runge–Kutta methods for the more ill-behaved problems semidiscretized in space. This fact will be clear in section 8.

7. Higher dimensions. We extend the ABCs obtained in section 2 to the two-dimensional case (the three-dimensional case is analogous). Consider the equation

$$\partial_t u(x,y,t) = \frac{-i}{c}(\partial_x^2 u(x,y,t) + \partial_y^2 u(x,y,t) + Vu(x,y,t)), \quad (x,y) \in \mathbf{R}^2,\ t>0. \tag{7.1}$$

Let us take the finite subdomain $[x_l, x_r] \times [y_l, y_r]$, where we will impose artificial boundary conditions. We consider the boundary $x = x_r$. Let us call $\hat{u}(x,\xi,\omega)$ the Fourier–Laplace transform in time and Fourier transform in variable y of $u(x,y,t)$. From (7.1) we obtain

$$\partial_x^2 \hat{u}(x,\xi,\omega) - \lambda^2 \hat{u}(x,\xi,\omega) = 0,$$

which is analogous to (2.1) with $\lambda^2 = -V - c\omega + \xi^2$. With an argument similar to that of the one-dimensional case, we obtain the TBC for the right boundary,

$$i\sqrt{V + c\omega - \xi^2}\,\hat{u}(x_r,\xi,\omega) + \partial_x \hat{u}(x_r,\xi,\omega) = 0. \tag{7.2}$$

In order to approximate this nonlocal boundary condition, we consider the approximation $\sqrt{V + c\omega - \xi^2} \approx q(V + c\omega - \xi^2)$, where, as in section 2, $q(s)$ is a rational interpolatory function of $\sqrt{s}$. In this way, we may consider the following simplest cases (other ABC of higher order can be obtained with a similar argument).

ABC(1,0). We take the rational function given by (2.5), and we deduce the ABC

$$\beta_0 \partial_t u(x_r,y,t) + \beta_1 u(x_r,y,t) + \beta_2 \partial_y^2 u(x_r,y,t) + \partial_x u(x_r,y,t) = 0, \tag{7.3}$$

where $\beta_0 = c/(s_1+s_2)$, $\beta_1 = i(s_1 s_2 + V)/(s_1+s_2)$, and $\beta_2 = i/(s_1+s_2)$. In a similar way, we obtain the following ABC for the boundary $y = y_r$:

$$\beta_0 \partial_t u(x,y_r,t) + \beta_1 u(x,y_r,t) + \beta_2 \partial_x^2 u(x,y_r,t) + \partial_y u(x,y_r,t) = 0. \tag{7.4}$$

ABC(1,1). We now approximate $\sqrt{s}$ by $q(s) = (\alpha_0 + \alpha_1 s)/(1+\alpha_2 s)$. In this case, the ABC is of the form

$$\begin{aligned} 0 = {} & \beta_0 u(x_r,y,t) + \beta_1 \partial_t u(x_r,y,t) + \beta_2 \partial_y^2 u(x_r,y,t) + \beta_3 \partial_x u(x_r,y,t) \\ & + \beta_4 \partial_{xt} u(x_r,y,t) + \beta_5 \partial_{xyy} u(x_r,y,t) \end{aligned} \tag{7.5}$$

for certain coefficients β_j. In fact, when $q(s)$ is the Padé(1,1) approximation to $\sqrt{s}$, (7.5) coincides with the ABC given by Kuska in [13] and by Fevens and Jiang in [6].

Let us consider the spatial discretization of (7.1) with ABC(1,0). Without loss of generality, we suppose that $x_r - x_l = y_r - y_l$. We consider a uniform grid of $[x_l, x_r] \times [y_l, y_r]$ given by $\{(x^j, y^m) : 0 \le j,m \le N\}$, where $x^j = x_l + jh$, $y^m = y_l + mh$ with $h = (x_r - x_l)/N = (y_r - y_l)/N$, and we denote by $u^{j,m}(t)$ the numerical approximation of $u(x^j, y^m, t)$. For the discretization of (7.1) in the interior domain, we consider the finite difference scheme

$$\frac{d}{dt}u^{j,m} = \frac{-i}{c}\left(\frac{u^{j-1,m} - 2u^{j,m} + u^{j+1,m}}{h^2} + \frac{u^{j,m-1} - 2u^{j,m} + u^{j,m+1}}{h^2} + Vu^{j,m}\right).$$

Now, we consider the spatial discretization at the boundary $x = x_r$. We take the node (x^N, y^m) with $0 < m < N$. First, we consider the approximation

$$\frac{d}{dt}u^{N,m} \approx \frac{-i}{c}\left(\frac{2}{h^2}(u^{N-1,m} - u^{N,m} + h\partial_x u(x^N,y^m,t)) + \partial_y^2 u(x^N,y^m,t) + Vu^{N,m}\right),$$

and taking into account the ABC(1,0) given by (7.3),

$$\frac{d}{dt}u^{N,m} \approx \frac{i}{c}\left(\frac{-2}{h^2}u^{N-1,m} - \left(V - \frac{2}{h^2} - \frac{2\beta_1}{h}\right)u^{N,m} - \left(1 - \frac{2\beta_2}{h}\right)\partial_y^2 u(x^N, y^m, t) + \frac{2\beta_0}{h}\frac{d}{dt}u^{N,m}\right).$$

Finally, using the approximation $\partial_y^2 u(x^N, y^m, t) \approx (u^{N,m-1} - 2u^{N,m} + u^{N,m+1})/h^2$,

$$\frac{d}{dt}u^{N,m} = \alpha_0 u^{N-1,m} + \alpha_1 u^{N,m-1} + \alpha_1 u^{N,m+1} + \alpha_2 u^{N,m}$$

for certain coefficients α_j. Similar expressions are obtained for $du^{j,N}/dt$, $du^{0,m}/dt$, and $du^{j,0}/dt$, $1 \le j, m \le N-1$.

Let us see now the spatial discretization for the corner (x^N, y^N). We consider the approximation

$$\partial_x^2 u(x^N, y^N, t) \approx \frac{2}{h^2}\left(u^{N-1,N} + u^{N,N} - h\left(\beta_0\frac{d}{dt}u^{N,N} + \beta_1 u^{N,N} + \beta_2\partial_y^2 u(x^N, y^N, t)\right)\right), \tag{7.6}$$

where we have used ABC(1,0) given by (7.3). Similarly, taking into account (7.4),

$$\partial_y^2 u(x^N, y^N, t) \approx \frac{2}{h^2}\left(u^{N,N-1} + u^{N,N} - h\left(\beta_0\frac{d}{dt}u^{N,N} + \beta_1 u^{N,N} + \beta_2\partial_x^2 u(x^N, y^N, t)\right)\right). \tag{7.7}$$

Adding expressions (7.6) and (7.7), we obtain

$$\partial_x^2 u(x^N, y^N, t) + \partial_y^2 u(x^N, y^N, t) \approx \epsilon_0 u^{N-1,N} + \epsilon_0 u^{N,N-1} + \epsilon_1 u^{N,N} + \epsilon_2\frac{d}{dt}u^{N,N}$$

for certain coefficients ϵ_j. Finally, making use of this expression in

$$\frac{d}{dt}u^{N,N} \approx \frac{-i}{c}(\partial_x^2 u(x^N, y^N, t) + \partial_y^2 u(x^N, y^N, t) + V u^{N,N}),$$

we obtain

$$\frac{d}{dt}u^{N,N} \approx \delta_0 u^{N,N} + \delta_1 u^{N,N-1} + \delta_1 u^{N-1,N}$$

for certain coefficients δ_j. Analogous expressions are obtained for the spatial discretization of $du^{N,0}/dt$, $du^{0,N}/dt$, and $du^{0,0}/dt$.

We have obtained a first order ordinary differential system

$$U'(t) = M(h)U(t), \tag{7.8}$$

where $U(t) = [u^{0,0}(t), u^{0,1}(t), \ldots, u^{0,N}(t), u^{1,0}(t), \ldots, u^{N,N}(t)]^T$ and $M(h)$ is a matrix of dimension $(N+1)^2 \times (N+1)^2$ with coefficients depending on h. This matrix

TABLE 7.1
Maximum of the real part of the ϵ-pseudoeigenvalues for ABC$(1,0)$ 2-dimensional.

ϵ =1.0d-1	ϵ =1.0d-3	ϵ =1.0d-6	ϵ =1.0d-9	ϵ =1.0d-12
-2.3324d-01	-2.2152d-01	-2.2172d-01	-2.2172d-01	-2.2172d-01

is nonnormal as in the one-dimensional case. In order to study the behavior of the solutions of (7.8), we have carried out an analysis of the ϵ-pseudospectrum of $M(h)$. In Table 7.1 we show an example of the maximum real part of the ϵ-eigenvalues for some values of ϵ. As in section 3, we deduce from this table that the nonnormality of $M(h)$ is mild. We have also computed for some examples the condition number κ_h of the matrix $P(h)$ such that $M(h) = P(h)D(h)P(h)^{-1}$ with $D(h)$ diagonal. The growth of κ_h observed is approximately $O(h^{-1})$.

8. Numerical experiments. Like Schmidt and Yevick in [19], we will consider for the numerical experiments the Fresnel equation (1.2) with $n = 1$, $\beta = 21.8°$, $n_0 = \cos(\beta)$, $\lambda = 0.832$, $k_0 = 2\pi/\lambda$. This case is a generalization of the two test cases of [25] associated with optical beam propagation in the Fresnel approximation. In order to obtain a solution describing an angle α with respect to the t-axis, we are going to consider the initial condition

$$u_0(x) = \exp\left(-((x - L/2)/\sigma)^2\right) \exp(i\eta(x - L/2)), \quad x \in [0, L], \tag{8.1}$$

with $\eta = -\cos(\beta)k_0 \tan(\alpha)$.

We will compare the results, in terms of reflection, obtained when we consider ABC(1,0), ABC(1,1), and ABC(2,1) using the spatial discretizations previously discussed. In the case of ABC(1,1) we will use the discretization in space that gives rise to a system with matrix (6.1). We will also consider ABC(3,2) using a spatial discretization similar to the one of ABC(2,1). In every case, the interpolatory nodes s_j^2 are considered equal to a unique positive number b. As we have already said, with the initial condition (8.1) we will hope to obtain optimal results if $b = \eta^2$.

Let us consider first an initial condition (8.1) with $\alpha = 10°$, $\sigma = 10$, and $L = 200$ so that it is 0 at the boundary. (This is very important, as we will discuss later.) In Figure 8.1(a) we have represented the reflection (the discrete L^2-norm of the solution remaining inside the computational window) as a function of time when we consider the optimal value $b = \eta^2$ and $h = 0.025$. The integration in time has been carried out with the implicit midpoint rule using a stepsize $k = 0.4$. It is clear that the behavior observed in Figure 8.1(a) is due to the order of absorption of the ABCs. Nevertheless, this is not the only thing we should take into account to compare the different ABC. As we have already remarked, the semidiscrete problems associated with these ABC are weakly ill-posed and are worse posed for higher order absorbing boundary conditions. This bad behavior is not visible in Figure 8.1(a) because the initial condition is 0 at the boundary and as the solution is a wave traveling with a concrete group velocity, this bad behavior of high order ABCs is canceled because of the greater absorption at the boundary.

Let us see the influence of taking an initial condition (8.1) which is not zero at the boundary, although this should not be done in practice since the interior domain must be long enough to contain the support of the initial condition. Let us consider (8.1) with $\sigma = 10$, $\alpha = 20°$, and $L = 40$. We will take the optimal value $b = \eta^2$ and the same discretization and stepsizes as in the previous experiment. Notice that

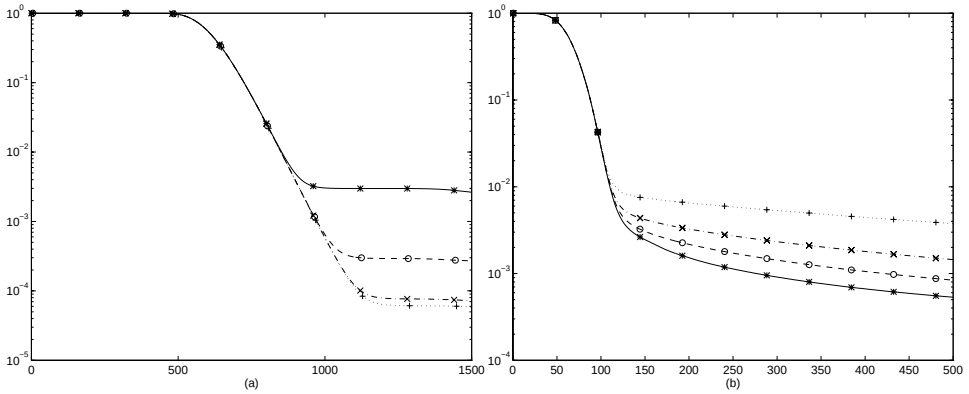


FIG. 8.1. *Reflection as a function of time:* $-*$ $ABC(1,0)$, $--\circ$ $ABC(1,1)$, $-\cdot-\times$ $ABC(2,1)$, $\cdots+$ $ABC(3,2)$. (a) *Initial value vanishing at the boundary.* (b) *Initial value nonvanishing at the boundary.*

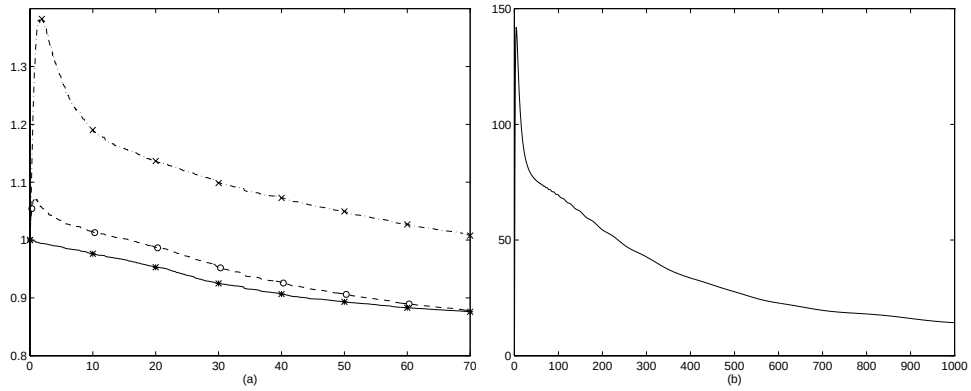


FIG. 8.2. *Reflection as a function of time. Random initial value nonvanishing at the boundary.* (a) $-*$ $ABC(1,0)$, $--\circ$ $ABC(1,1)$, $-\cdot-\times$ $ABC(2,1)$. (b) $ABC(3,2)$

$u_0(x)$ is approximately 0.018 at the boundary. This way, in Figure 8.1(b) the worse behavior of high ABCs due to its weak ill-posedness (see Figure 3.1) is present. We observe that ABCs with high order of absorption produce more reflection than the less absorbing ones. We remark that a similar behavior is present in the numerical experiments in [6]. The authors take as initial condition a single Gaussian distribution $w_0(x) = \exp\left(-(x-\xi)^2/2\sigma_0^2\right)\exp(iK_0x)$, $x \in [0,L]$, with $L = 10$, $\xi = 3L/4$, and $\sigma_0 = L/10$. They also consider a narrow Gaussian pulse distribution $z_0(x)$ taking $\xi = 3L/4$, $\sigma_0 = L/100$, and $L = 10$, concluding that although for the initial condition $w_0(x)$ the ABC given by (1.3) with $p = 4$ behaves worse that the one with $p = 3$; for $z_0(x)$ the higher order ABC is more effective. Taking into account our previous analysis, this behavior is due to the fact that $w_0(x)$ is not 0 at the boundary.

In the previous examples we have considered only regular solutions. In order to see that the weak ill-posedness does not produce unbounded growths when the solution is not regular, we consider now a random initial condition. The discretization of the equation is done as in the previous experiment with $h = 0.025$ and $k = 0.1$. We observe in Figure 8.2 that although for all ABC the norm of the solution goes to 0

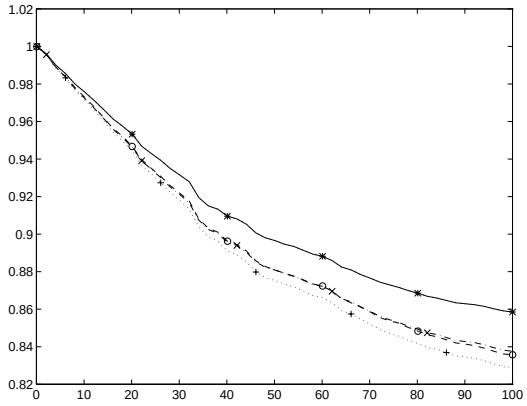


FIG. 8.3. *Reflection as a function of time. Random initial value vanishing at the boundary:* $-*$ $ABC(1,0)$, $--\circ$ $ABC(1,1)$, $-\cdot-\times$ $ABC(2,1)$, $\cdots+$ $ABC(3,2)$.

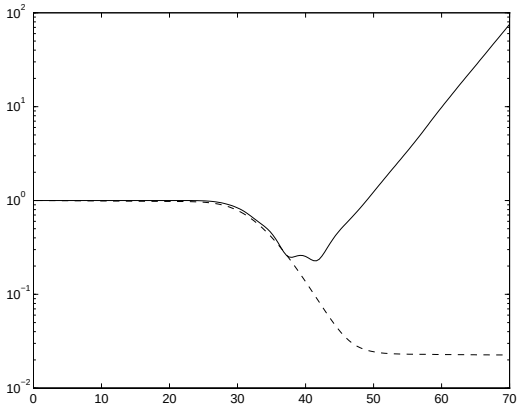


FIG. 8.4. *Reflection as a function of time. Nonoptimal value of* b *and very small spatial stepsize:* $ABC(3,2)$, *– IMPR, – – DIRK order* 3.

as $t \to \infty$, for ABC(1,1), ABC(2,1), and ABC(3,2) there is an initial growth due to the condition number of the semidiscrete systems in space associated with these ABC (see Figure 3.1). In the case of ABC(3,2), this condition number is $O(h^{-4})$ and the norm of the solutions grows initially to 140. This behavior is more accentuated the smaller h is.

Let us see how important it is to consider an initial condition which is 0 at the boundary. In Figure 8.3 we have considered the same random initial condition but with vanishing values for the four first and last nodes. For the same values of h and k as in Figure 8.2, the worse behavior of high order ABCs is not visible.

The fact that we have used the method of lines is very important since the problem discretized in space is weakly ill-posed. We can consider this way a high order method for the time integration that compensates for the troubles in space. This can be seen in Figure 8.4. We have considered ABC(3,2), taking as initial condition (8.1) with

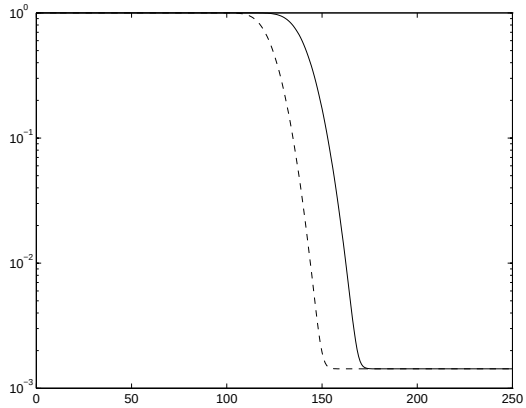


FIG. 8.5. *Reflection as a function of time. Comparison of the velocity in the interior domain:* $ABC(1,0)$ *– IMPR,* – – *2-stage Gauss.*

$\alpha = 30°$, $\sigma = 3$, and $L = 36$ (notice that it is 0 at the boundary). We have not chosen the optimal value of b for this solution but rather the optimal value for a wave traveling with a velocity $\tan(10°)$, so that the weak ill-posedness of the semidiscrete problem is not canceled because the great absorption of this ABC. We have integrated the semidiscrete problem with two different methods: the implicit midpoint rule and a diagonally implicit Runge–Kutta of order 3. In both cases we have considered the very small value $h = 36/76800$ and the time stepsize $k = 0.4$. In Figure 8.4 we can observe that with the implicit midpoint rule the instability in space makes the norm of the solution grow, while with the method of order 3 this behavior is not present.

The possibility of using different integrators in time is also important for the integration of the solution in the interior domain. Let us see the result of taking as initial condition (8.1) with $\alpha = 40°$, $\sigma = 10$, and $L = 200$. We have considered ABC(1,0) using for the integration in time first the implicit midpoint rule, and second the 2-stage Gauss method. In both cases we have taken $h = 0.025$ and $k = 0.4$. In Figure 8.5 we observe that with the implicit midpoint rule the numerical solution does not travel with the right velocity (describing an angle $\alpha = 40°$ with respect to the t-axis) because k is not small enough, while this does not happen for the 2-stage Gauss method. The possibility of considering high order methods for the time integration allows us to use bigger stepsizes.

REFERENCES

[1] I. ALONSO-MALLO AND N. REGUERA, *Condiciones frontera transparentes y absorbentes para la ecuación de Schrödinger en una dimensión*, in Proceedings of the 15th Congress on Differential Equations and Applications and 6th Congress on Applied Mathematics, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain, 1999, pp. 467–473.

[2] V. A. BASKAKOV AND A. V. POPOV, *Implementation of transparent boundaries for numerical solution of the Schrödinger equation*, Wave Motion, 14 (1991), pp. 123–128.

[3] L. DI MENZA, *Transparent and absorbing boundary conditions for the Schrödinger equation in a bounded domain*, Numer. Funct. Anal. Optim., 18 (1997), pp. 759–775.

[4] E. DUBACH, *Artificial boundary conditions for diffusion equations: Numerical study*, J. Comput. Appl. Math., 70 (1996), pp. 127–144.

[5] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.

[6] T. Fevens and H. Jiang, *Absorbing boundary conditions for the Schrödinger equation*, SIAM J. Sci. Comput., 21 (1999), pp. 255–282.
[7] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, Springer-Verlag, Berlin, 1991.
[8] L. Halpern and J. Rauch, *Absorbing boundary conditions for diffusion equations*, Numer. Math., 71 (1995), pp. 185–224.
[9] R. L. Higdon, *Absorbing boundary conditions for difference aproximations to the multi-dimensional wave equation*, Math. Comp., 47 (1986), pp. 437–459.
[10] L. H. Howell and L. N. Trefethen, *Ill-posedness of absorbing boundary conditions for migration*, Geophysics, 53 (1988), pp. 593–603.
[11] B. Kagström, *Bounds and perturbation bounds for the matrix exponential*, BIT, 17 (1977), pp. 39–57.
[12] H. O. Kreiss, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–298.
[13] J.-P. Kuska, *Absorbing boundary conditions for the Schrödinger equation on finite intervals*, Phys. Rev. B, 46 (1992), pp. 5000–5003.
[14] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.
[15] B. Mayfield, *Nonlocal Boundary Condition for the Schrödinger Equation*, Ph.D. thesis, University of Rhode Island, Providence, RI, 1989.
[16] S. C. Reddy and L. N. Trefethen, *Stability of the method of lines*, Numer. Math., 62 (1992), pp. 235–267.
[17] F. Schmidt, *An adaptive approach to the numerical solution of Fresnel's wave equation*, J. Lightwave Technology, 11 (1993), pp. 1425–1434.
[18] F. Schmidt and P. Deuflhard, *Discrete transparent boundary conditions for numerical solutions of Fresnel's equation*, Comput. Math. Appl., 29 (1995), pp. 53–76.
[19] F. Schmidt and D. Yevick, *Discrete transparent boundary conditions for Schrödinger-type equations*, J. Comput. Phys., 134 (1997), pp. 96–107.
[20] T. Shibata, *Absorbing boundary conditions for the finite difference time-domain calculation of the one-dimensional Schrödinger equation*, Phys. Rev. B, 43 (1991), pp. 6760–6763.
[21] L. N. Trefethen, *Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundström*, J. Comput. Phys., 49 (1983), pp. 199–217.
[22] L. N. Trefethen, *Pseudoespectra of matrices*, in Proceedings of the 14th Dundee Biennial Conference on Numerical Analysis, Pitman Res. Notes Math. Ser. 260, D. F. Griffiths and G. A. Watson, eds., Longman, New York, 1992, pp. 234–266.
[23] L. N. Trefethen and L. Halpern, *Well-posedness of one-way wave equations and absorbing boundary conditions*, Math. Comp., 47 (1986), pp. 421–435.
[24] T. B. Koch, R. März, and J. B. Davies, *Beam propagation method using z-transient variational principle*, in Proceedings 16th European Conference on Optical Communication (ECOC) 1, Amsterdam, The Netherlands, 1990, pp. 163–166.
[25] D. Yevick, J. Yu, and Y. Yayon, *Optimal absorbing boundary conditions*, J. Opt. Soc. Amer. A, 12 (1995), pp. 107–110.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 159–179

DUAL-PRIMAL FETI METHODS FOR THREE-DIMENSIONAL ELLIPTIC PROBLEMS WITH HETEROGENEOUS COEFFICIENTS*

AXEL KLAWONN†, OLOF B. WIDLUND‡, AND MAKSYMILIAN DRYJA§

Abstract. In this paper, certain iterative substructuring methods with Lagrange multipliers are considered for elliptic problems in three dimensions. The algorithms belong to the family of dual-primal finite element tearing and interconnecting (FETI) methods which recently have been introduced and analyzed successfully for elliptic problems in the plane. The family of algorithms for three dimensions is extended and a full analysis is provided for the new algorithms. Particular attention is paid to finding algorithms with a small primal subspace since that subspace represents the only global part of the dual-primal preconditioner. It is shown that the condition numbers of several of the dual-primal FETI methods can be bounded polylogarithmically as a function of the dimension of the individual subregion problems and that the bounds are otherwise independent of the number of subdomains, the mesh size, and jumps in the coefficients. These results closely parallel those of other successful iterative substructuring methods of primal as well as dual type.

Key words. domain decomposition, Lagrange multipliers, FETI, dual-primal methods, preconditioners, elliptic equations, finite elements, heterogeneous coefficients

AMS subject classifications. 65F10, 65N30, 65N55

PII. S0036142901388081

1. Introduction.

The finite element tearing and interconnecting (FETI) methods are domain decomposition methods of iterative substructuring type. They are thus a special type of preconditioned conjugate gradient methods which have been developed for solving the often huge algebraic systems of equations which arise in finite element computations. The dual-primal FETI (FETI-DP) methods were introduced recently by Farhat et al. [9]. Their work was followed by a significant contribution to the theory of two-dimensional second and fourth order problems by Mandel and Tezaur [17]; by Farhat, Lesoinne, and Pierson [10], who specifically address an algorithm for three-dimensional problems; and by Pierson [19], who has also recently used his codes to solve very difficult and huge problems. The algorithm presented in [10, 19] uses constraints on the averages over edges and faces in a way similar to that of the algorithms considered in this paper. Our contribution is to both the extension of the family of algorithms for problems in three dimensions and the analysis. We also show that good convergence bounds can be maintained even for quite general coefficients such as those that model highly heterogeneous materials. Our work has

*Received by the editors April 17, 2001; accepted for publication (in revised form) November 26, 2001; published electronically April 12, 2002.
http://www.siam.org/journals/sinum/40-1/38808.html

†Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 Sankt Augustin, Germany (klawonn@scai.fhg.de, http://www.scai.fhg.de/people/klawonn.html). The research of this author was supported in part by the National Science Foundation under grant NSF-CCR-9732208.

‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (widlund@cs.nyu.edu, http://www.cs.nyu.edu/cs/faculty/widlund). The research of this author was supported in part by the National Science Foundation under grant NSF-CCR-9732208 and in part by the U.S. Department of Energy under contract DE-FG02-92ER25127.

§Department of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland (dryja@mimuw.edu.pl). The research of this author was supported in part by the National Science Foundation under grant NSF-CCR-9732208 and in part by the Polish Science Foundation under grant 2P03A 021 16.

been inspired by that of Mandel and Tezaur and is also based on our own earlier work, in particular, [5, 6], and [13].

It is well known that domain decomposition algorithms cannot be scalable, i.e., have a rate of convergence which is independent of the number of subregions, unless a coarse space component is included. We note that the underlying coarse spaces for three-dimensional problems are often more complicated than the quite simple constructions that work well for problems in the plane; see [24] for a discussion. We will construct several of our FETI-DP methods using relatively exotic coarse spaces. Thus, our Algorithms B and C in section 4 are closely related to certain interpolation operators and coarse spaces known from earlier work on primal iterative substructuring methods; see [5, 6]. Both of these methods have relatively large global, primal subspaces.

The term dual-primal refers to the idea of enforcing some continuity constraints across the interface between the subregions, throughout the iteration, as in a primal method, while all other constraints are enforced by using dual variables, i.e., Lagrange multipliers, as in a dual method. We will see that the FETI-DP methods differ in several important respects from the strictly dual FETI methods, in particular, the one-level FETI method which is described in section 3. In fact, from both an algorithmic and an analytic point of view, the FETI-DP methods are closer to the primal iterative substructuring methods than the FETI methods previously developed. While the global part of the preconditioner for a strictly dual FETI method is directly associated with the dual variables, that of a FETI-DP method is not.

We note that primal iterative substructuring methods have been studied quite extensively (see, e.g., [6, 8], and [5]) well before a similarly complete, and quite challenging, mathematical theory was developed for the FETI methods (see [16, 21], and [13]; FETI algorithms using inexact subdomain solvers also have been developed and analyzed by two of the authors in [12]). We note that primal iterative substructuring methods have been developed extensively even for elliptic systems, e.g., in [18], and that we believe we have many of the tools necessary to extend our current results and algorithms to the systems of linear elasticity; cf. also [12]. We also note that, algorithmically, some of the FETI-DP methods that we consider have certain features in common with very early work on iterative substructuring methods for problems with many substructures; cf. the studies on Neumann–Dirichlet algorithms by Dryja, Proskurowski, and Widlund [4] and the contributions of Dryja [3] and Widlund [23] to the first international symposium on domain decomposition. We note, in particular, that the Neumann subsystems of these early algorithms are nonsingular; there are no floating subregions because of a device very similar to that used in the FETI-DP methods. The use of Lagrange multipliers, in a special context, was also suggested in [23].

The remainder of this paper is organized as follows. In section 2, we introduce our scalar elliptic equation, which can have very different coefficients in different subregions and which has served as a standard, nontrivial model problem in many studies of iterative substructuring methods. We also introduce a simple finite element space, the decomposition of our region, and our variational problem. In section 3, we give a brief description of a one-level FETI method to provide a necessary background. In section 4, we introduce our four FETI-DP methods; we note that we have recently analyzed still another FETI-DP algorithm in [14]. In section 5, we provide, with a few proofs, some auxiliary results, many of which have been previously developed for the analysis of primal iterative substructuring methods. In section 6, we prove

almost optimal bounds on the condition number of three of the methods. They are independent of the number of substructures and grow only polylogarithmically with the number of degrees of freedom associated with the individual substructures.

2. Elliptic model problem, finite elements, and geometry. Let $\Omega \subset \mathbf{R}^3$ be a bounded, polyhedral region; let $\partial\Omega_D \subset \partial\Omega$ be a closed set of positive measure, and let $\partial\Omega_N := \partial\Omega \setminus \partial\Omega_D$ be its complement. We impose homogeneous Dirichlet and general Neumann boundary conditions, respectively, on these two subsets and introduce the Sobolev space $H^1_0(\Omega, \partial\Omega_D) := \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega_D\}$.

For simplicity, we will consider only a piecewise linear, conforming finite element approximation of the following scalar, second order model problem: Find $u \in H^1_0(\Omega, \partial\Omega_D)$, such that

$$a(u, v) = f(v) \quad \forall v \in H^1_0(\Omega, \partial\Omega_D), \tag{2.1}$$

where

$$a(u, v) := \int_\Omega \rho(x)\nabla u \cdot \nabla v dx, \qquad f(v) := \int_\Omega f v dx \; + \; \int_{\partial\Omega_N} g_N v ds, \tag{2.2}$$

and where g_N is the Neumann boundary data defined on $\partial\Omega_N$; it provides, together with the volume load f, the contributions to the load vector of the finite element problem. The coefficient $\rho(x) > 0$ for $x \in \Omega$.

We decompose Ω into nonoverlapping subdomains $\Omega_i, i = 1, \ldots, N$, also known as substructures, and each of which is the union of shape-regular elements with the finite element nodes on the boundaries of neighboring subdomains matching across the interface $\Gamma := \overline{(\bigcup_{i=1}^N \partial\Omega_i) \setminus \partial\Omega}$. The interface Γ is composed of subdomain faces, regarded as open sets, which are shared by two subregions; of edges which are shared by more than two subregions; and of vertices which are endpoints of edges. If Γ intersects $\partial\Omega_N$ along an edge common to the boundaries of only two subdomains, we will regard it as part of the face common to this pair of subdomains. We denote the faces of Ω_i by $\mathcal{F}^{ij}$, its edges by $\mathcal{E}^{ik}$, and its vertices by $\mathcal{V}^{i\ell}$.

We denote the standard finite element space of continuous, piecewise linear functions on Ω_i by $W^h(\Omega_i)$; we always assume that these functions vanish on $\partial\Omega_D$. For simplicity, we assume that the triangulation of each subdomain is quasi-uniform. The diameter of Ω_i is H_i or, generically, H. We denote the corresponding finite element trace spaces by $W_i := W^h(\partial\Omega_i \cap \Gamma), i = 1, \ldots, N$, and the associated product space by $W := \prod_{i=1}^N W_i$. We will often consider elements of W which are discontinuous across the interface.

The finite element approximation of the elliptic problem is continuous across Γ and we denote the corresponding subspace of W by $\widehat{W}$. We note that while the stiffness matrix K and Schur complement S, which correspond to the product space W, generally are singular, those of $\widehat{W}$ are not.

We also will use additional, intermediate subspaces $\widetilde{W}$ of W for which only a relatively small number of continuity constraints is enforced across the interface. One of the benefits of working in $\widetilde{W}$, rather than in W, will be that certain related Schur complements, $\widetilde{S}$ and S_Δ, are strictly positive definite; see sections 3 and 4.

We assume that possible jumps of $\rho(x)$ are aligned with the subdomain boundaries and, for simplicity, that on each subregion Ω_i, $\rho(x)$ has the constant value $\rho_i > 0$. Our bilinear form and load vector can then be written, in terms of contributions from

individual subregions, as

$$a(u,v) = \sum_{i=1}^{N} \rho_i \int_{\Omega_i} \nabla u \cdot \nabla v dx, \qquad f(v) = \sum_{i=1}^{N} \left(\int_{\Omega_i} f v dx \; + \; \int_{\partial\Omega_i \cap \partial\Omega_N} g_N v ds \right). \tag{2.3}$$

In our theoretical analysis, we assume that each subregion Ω_i is the union of a number of shape-regular tetrahedral coarse elements and that the number of such tetrahedra is uniformly bounded for each subdomain. Thus, the subregions are not very thin and we also can easily show that the diameters of any pair of neighboring subdomains are comparable. We also assume that if a face of a subdomain intersects $\partial\Omega_D$, then the measure of this set is comparable to that of the face. Similarly, if an edge of a subdomain intersects $\partial\Omega_D$, we assume that the length of this intersection is bounded from below in terms of the length of the edge as a whole. The sets of nodes on $\partial\Omega, \partial\Omega_i$, and Γ are denoted by $\partial\Omega_h, \partial\Omega_{i,h}$, and Γ_h, respectively.

As in previous work on Neumann–Neumann and FETI algorithms, a crucial role is played by *the weighted counting functions* μ_i, which are associated with the individual subdomain boundaries $\partial\Omega_i$; cf. [5, 8, 15, 20]. In this paper they will be used in the definition of certain diagonal scaling matrices. These functions are defined for $\gamma \in [1/2, \infty)$ and for $x \in \Gamma_h \cup \partial\Omega_h$ by a sum of contributions from Ω_i and its relevant neighbors

$$\mu_i(x) = \begin{cases} \sum\limits_{j \in \mathcal{N}_x} \rho_j^{\gamma}(x) & x \in \partial\Omega_{i,h} \cap \partial\Omega_{j,h}, \\ \rho_i^{\gamma}(x) & x \in \partial\Omega_{i,h} \cap (\partial\Omega_h \setminus \Gamma_h), \\ 0 & x \in (\Gamma_h \cup \partial\Omega_h) \setminus \partial\Omega_{i,h}. \end{cases} \tag{2.4}$$

Here, $\mathcal{N}_x$ is the set of indices of the subregions which have x on their boundaries. We note that any node of Γ_h either belongs to two faces, belongs to at least three edges, or is a vertex of several substructures. The μ_i are continuous, piecewise discrete harmonic functions; for a definition see section 3. The pseudoinverses $\mu_i^{\dagger}$, which belong to the same class of functions, are defined for $x \in \Gamma_h \cup \partial\Omega_h$ by

$$\mu_i^{\dagger}(x) = \begin{cases} \mu_i^{-1}(x) & \text{if } \mu_i(x) \neq 0, \\ 0 & \text{if } \mu_i(x) = 0. \end{cases} \tag{2.5}$$

We note that these functions provide a partition of unity:

$$\sum_i \rho_i^{\gamma}(x) \mu_i^{\dagger}(x) \equiv 1 \quad \forall x \in \Gamma_h \cup \partial\Omega_h. \tag{2.6}$$

3. One-level FETI methods. In this section, we will introduce some notation and certain other aspects of the older one-level FETI methods which we will use in the rest of this paper. We begin by defining a stiffness matrix K for the entire product space $\prod_{i=1}^{N} W^h(\Omega_i)$. K is a direct sum of local stiffness matrices $K^{(i)}$ which correspond to the subdomains $\Omega_i, i = 1, \ldots, N$, and to the appropriate terms in the first formula of (2.3). The local load vectors are obtained similarly; see the second formula of (2.3).

Any nodal variable not associated with Γ_h is called interior and belongs only to one substructure; the nodal values on $\partial\Omega_N \backslash \Gamma$ also belong to this set. The interior variables of any subdomain can be eliminated by block Gaussian elimination in work which

clearly can be parallelized across the subdomains. The resulting reduced matrices are the Schur complements

$$S^{(i)} = K^{(i)}_{\Gamma\Gamma} - K^{(i)}_{\Gamma I}(K^{(i)}_{II})^{-1}K^{(i)}_{I\Gamma}, \qquad i = 1, \ldots, N.$$

Here Γ and I represent the interface and interior, respectively. We note that the $S^{(i)}$, and their inverses or pseudoinverses, are needed only in terms of matrix-vector products and that their elements therefore need not be explicitly computed. We also obtain a reduced load vector for each subdomain. The one originating in Ω_i is denoted by f_i, and the local vector of interface nodal values, which can be regarded as a component of an element of the product space W, is denoted by u_i.

The elimination of the interior variables of a substructure also can be viewed in terms of an orthogonal projection, with respect to the bilinear form $\langle K^{(i)}\cdot, \cdot\rangle$, onto the subspace of vectors with components that vanish at all the nodes of $\overline{\partial\Omega_i \setminus \partial\Omega_N}$. Here $\langle\cdot,\cdot\rangle$ denotes the ℓ_2-inner product. We note that these vectors represent elements of $W^h(\Omega_i) \cap H^1_0(\Omega_i, \overline{\partial\Omega_i \setminus \partial\Omega_N})$. These local subspaces are orthogonal, in this energy inner product, to the space of discrete harmonic vectors which represent discrete harmonic finite element functions with v_Γ and w_Γ vectors of interface values, such a vector $w = (w_I, w_\Gamma)$ is defined by

$$\langle K^{(i)}w, v\rangle = 0 \quad \forall v \ \text{ such that } \ v_\Gamma = 0 \tag{3.1}$$

on the subdomain Ω_i, or, equivalently, by

$$K^{(i)}_{II}w_I + K^{(i)}_{I\Gamma}w_\Gamma = 0. \tag{3.2}$$

We can regard w_Γ as a vector of Dirichlet data given on $\partial\Omega_{i,h} \cap \Gamma_h$ and note that a piecewise discrete harmonic function is completely defined by its values on the interface.

The Schur complement $S^{(i)}$ satisfies the following minimum property: For all $w \in W_i$,

$$\langle S^{(i)}w, w\rangle = \min\langle K^{(i)}v, v\rangle, \tag{3.3}$$

where the minimum is taken over all $v = (v_I, v_\Gamma) \in W^h(\Omega_i)$ such that $v_\Gamma = w$.

We note that we can view the Schur complement $S^{(i)}$ as the restriction of the stiffness matrix $K^{(i)}$ to the space of discrete harmonic functions. In what follows, we will work almost exclusively with functions in the trace spaces W_i and, whenever convenient, consider such an element as representing a discrete harmonic function in Ω_i. We also note that it is this piecewise discrete harmonic part of the solution, representing an element of $\widehat{W}$, that is determined by any iterative substructuring method; the other, interior, parts of the solution are computed locally as indicated above.

We now briefly review a part of the derivation of the traditional FETI methods prior to showing in the next section how matters change in the FETI-DP case. We begin by reformulating the finite element problem, reduced to the interface Γ, as a minimization problem with constraints given by the requirement of continuity across Γ: find $u \in W$, such that

$$\left.\begin{array}{c} J(u) := \frac{1}{2}\langle Su, u\rangle - \langle f, u\rangle \rightarrow \ \min \\ Bu = 0 \end{array}\right\}, \tag{3.4}$$

where $u = [u_1, \ldots, u_N]^t$, $f = [f_1, \ldots, f_N]^t$, and $S = \mathrm{diag}_i(S^{(i)})$.

The matrix $B = [B^{(1)}, \ldots, B^{(N)}]$ is constructed from $\{0, 1, -1\}$ such that the values of the solution u, associated with more than one subdomain, coincide when $Bu = 0$. Here, as in [13, sections 5 and 4], we can work with either fully redundant or nonredundant constraints, i.e., with either all possible or the smallest possible number of constraints for each node of Γ_h. The local Schur complement matrices $S^{(i)}$ are positive semidefinite and, in fact, in many cases, there are floating subdomains, i.e., subregions for which the $S^{(i)}$ are singular. Problem (3.4) is uniquely solvable if and only if $\ker(S) \cap \ker(B) = \{0\}$, i.e., S is invertible on the null space of B. This condition holds since the original finite element model is elliptic.

In a standard one-level FETI method, a vector of Lagrange multipliers λ is introduced to enforce all the constraints $Bu = 0$, and we obtain a saddle point formulation of (3.4): find $(u, \lambda) \in W \times U$ such that

$$\left.\begin{array}{rcrcl} Su & + & B^t\lambda & = & f \\ Bu & & & = & 0 \end{array}\right\}. \tag{3.5}$$

In this paper, we will work exclusively with fully redundant sets of Lagrange multipliers. The matrix B^t then has a null space and, to ensure uniqueness, it is appropriate to restrict the choice of Lagrange multipliers to $\mathrm{range}(B)$. In fact, in the one-level FETI methods the space of Lagrange multipliers is chosen as a subspace of $\mathrm{range}(B)$, since further constraints on the Lagrange multipliers must be introduced in order to ensure the solvability of the first equation of (3.5); see, e.g., [11, 16, 13].

We also will use a full column rank matrix R built from all of the null space elements of S; these elements are associated with individual subdomains (the rigid body motions in the case of elasticity) and are continued by zero outside the subregion in question. Thus, $\mathrm{range}(R) = \ker(S)$. We note that no subdomain with a boundary which intersects $\partial\Omega_D$ contributes to R. We also note that the case of linear elasticity is somewhat more complicated. There also can be contributions to R from subdomains with boundaries intersecting $\partial\Omega$ for which there are not enough essential boundary conditions to fully eliminate the entire space of rigid body motions.

The solution of the first equation in (3.5) exists if and only if $f - B^t\lambda \in \mathrm{range}(S)$; this constraint leads to the introduction of an orthogonal projection P from U onto $\ker(G^t)$ with $G := BR$. We note that we do not need any such projection in the FETI-DP methods defined in the next section.

Eliminating the primal variables from (3.5) and considering the component orthogonal to $\mathrm{range}(G)$, we obtain

$$\left.\begin{array}{rcl} P^tF\lambda & = & P^td \\ G^t\lambda & = & e \end{array}\right\} \tag{3.6}$$

with $F := BS^\dagger B^t$, $d := BS^\dagger f$, $S^\dagger$ a pseudoinverse of S, and $e := R^t f$; this last constraint ensures that the first equation of (3.5) is solvable.

The original FETI method is a conjugate gradient method applied to

$$P^tF\lambda = P^td, \qquad \lambda \in \lambda_0 + \mathrm{range}(P), \tag{3.7}$$

with an initial approximation λ_0 chosen such that $G^t\lambda_0 = e$; this constraint guarantees that the first equation of (3.5) is consistent.

We will not describe the preconditioners used in the solution of this dual problem but will postpone this topic to the next section; there are no essential differences

between the two cases as far as preconditioners are concerned. For a more detailed description and analysis of a number of one-level FETI algorithms, see Klawonn and Widlund [13].

4. FETI-DP methods. In previous studies of FETI-DP methods for problems in two dimensions (see Farhat et al. [9] and Mandel and Tezaur [17]), the constraints on the degrees of freedom associated with the vertices of the substructures are enforced in each iteration; i.e., the corresponding degrees of freedom belong to the primal set of variables, while all the constraints associated with the edge nodes are fully enforced only at the convergence of the iterative method. A linear system of algebraic equations is solved exactly in each step of the iteration. All unknowns except those of the subdomain vertices can be eliminated at a modest expense, and in parallel across the subdomains, resulting in a Schur complement for the vertex variables. In this first step, we can take full advantage of a high quality sparse matrix Cholesky solver when solving the individual subdomain problems, which in fact are Neumann problems on the individual subregions except for a Dirichlet condition at the subdomain vertices. The order of the Schur complement equals the number of subdomain vertices which do not belong to $\partial\Omega_D$. It is sparse since it can be shown quite easily that no nonzero off-diagonal elements exist in the reduced system matrix except those that correspond to pairs of vertices which belong to the same substructure.

In their recent paper, Mandel and Tezaur [17] established a condition number bound of the form $C(1+\log(H/h))^2$ for the resulting FETI-DP method, in two dimensions, and if it is equipped with a Dirichlet preconditioner similar to those used for some of the older FETI methods; cf. Farhat, Mandel, and Roux [11]. This preconditioner is built from local solvers on the subregions with zero Dirichlet conditions at the vertices of the subregions. This algorithm is scalable in the sense that the constant C is independent of the number of subregions, the subdomain diameters, as well as the mesh size h of the finite element model. Mandel and Tezaur also established a corresponding result for a fourth order elliptic problem in the plane. Their proof in [17] for the second order equation uses linear algebra arguments and a lemma from a classical paper by Bramble, Pasciak, and Schatz [2, Lem. 3.5].

The same algorithm, our Algorithm A, also can be defined for the three-dimensional case but it does not perform well; see Farhat et al. [9, sect. 5]. This is undoubtedly related to the poor performance of *vertex-based* iterative substructuring methods; see [6, sect. 6.1]. A condition number estimate for this algorithm is given in Remark 2 at the end of the paper.

In the present study, as well as in others of FETI-DP methods, we work with subspaces $\widetilde{W} \subset W$ for which sufficiently many constraints are enforced so that the resulting leading diagonal block of the saddle point problem, though no longer block diagonal, is strictly positive definite. We also introduce two subspaces, $\widehat{W}_\Pi \subset \widehat{W}$ and $\widetilde{W}_\Delta$, corresponding to a primal and a dual part of the space $\widetilde{W}$. These subspaces will play an important role in the description and analysis of our iterative method. The direct sum of these spaces equals $\widetilde{W}$, i.e.,

$$\widetilde{W} = \widehat{W}_\Pi \oplus \widetilde{W}_\Delta. \tag{4.1}$$

The second subspace, $\widetilde{W}_\Delta$, is the direct sum of local subspaces $\widetilde{W}_{\Delta,i}$ of $\widetilde{W}$, where each subdomain Ω_i contributes a subspace $\widetilde{W}_{\Delta,i}$; only its ith component in the sense of the product space $\widetilde{W}$ is nontrivial.

In the descriptions of our algorithms, we will need certain standard finite element cutoff functions $\theta_{\mathcal{E}^{ik}}$, $\theta_{\mathcal{F}^{ij}}$, and $\theta_{\mathcal{V}^{i\ell}}$. The first two are the discrete harmonic functions

which equal 1 on $\mathcal{E}_h^{ik}$ and $\mathcal{F}_h^{ij}$, respectively, and which vanish elsewhere on Γ_h; $\theta_{\mathcal{V}^{i\ell}}$ denotes the piecewise discrete harmonic extension of the standard nodal basis function associated with the vertex $\mathcal{V}^{i\ell}$. These cutoff functions also will be used in the analysis of the methods; see sections 5 and 6.

We are now ready to define our algorithms in terms of pairs of subspaces.

ALGORITHM A. *The primal subspace, $\widehat{W}_\Pi$, is spanned by the nodal finite element basis functions $\theta_{\mathcal{V}^{i\ell}}$. The local subspace $\widetilde{W}_{\Delta,i}$ is defined as the subspace of W_i of elements which vanish at the subdomain vertices, i.e., by*

$$\widetilde{W}_{\Delta,i} := \{u \in W_i : u(\mathcal{V}^{i\ell}) = 0 \ \forall \mathcal{V}^{i\ell} \in \partial\Omega_i\}.$$

Hence, $\widetilde{W} = \widetilde{W}_A$ is the subspace of W of functions that are continuous at the subdomain vertices.

ALGORITHM B. *The primal subspace, $\widehat{W}_\Pi$, is spanned by the vertex nodal finite element basis functions $\theta_{\mathcal{V}^{i\ell}}$ and the cutoff functions $\theta_{\mathcal{E}^{ik}}$ and $\theta_{\mathcal{F}^{ij}}$ associated with all the edges and faces, respectively, of the interface. The local subspace $\widetilde{W}_{\Delta,i}$ is defined as the subspace of W_i, where the values at the subdomain vertices vanish, together with the averages $\overline{u}_{\mathcal{E}^{ik}}$ and $\overline{u}_{\mathcal{F}^{ij}}$, i.e., by*

$$\widetilde{W}_{\Delta,i} := \{u \in W_i : u(\mathcal{V}^{i\ell}) = 0, \overline{u}_{\mathcal{E}^{ik}} = 0, \overline{u}_{\mathcal{F}^{ij}} = 0 \ \forall \mathcal{V}^{i\ell}, \mathcal{E}^{ik}, \mathcal{F}^{ij} \subset \partial\Omega_i\}.$$

Here,

$$\overline{u}_{\mathcal{E}^{ik}} = \frac{\int_{\mathcal{E}^{ik}} u ds}{\int_{\mathcal{E}^{ik}} 1 ds} \quad \text{and} \quad \overline{u}_{\mathcal{F}^{ij}} = \frac{\int_{\mathcal{F}^{ij}} u dx}{\int_{\mathcal{F}^{ij}} 1 dx}. \tag{4.2}$$

Hence, $\widetilde{W} = \widetilde{W}_B$ is the subspace of W of functions that are continuous at the subdomain vertices and that have the same values of $\overline{u}_{\mathcal{E}^{ik}}$ and $\overline{u}_{\mathcal{F}^{ij}}$. These functions are independent of the component of $u \in \widetilde{W}_B$ that is used in the evaluation of these averages.

ALGORITHM C. *The primal subspace, $\widehat{W}_\Pi$, is spanned by the vertex nodal finite element basis functions $\theta_{\mathcal{V}^{i\ell}}$ and the cutoff functions $\theta_{\mathcal{E}^{ik}}$ of all the edges of Γ. The local subspace $\widetilde{W}_{\Delta,i}$ is defined as the subspace of W_i, where the values at the subdomain vertices vanish together with the averages $\overline{u}_{\mathcal{E}^{ik}}$, i.e., by*

$$\widetilde{W}_{\Delta,i} := \{u \in W_i : u(\mathcal{V}^{i\ell}) = 0, \overline{u}_{\mathcal{E}^{ik}} = 0 \ \forall \mathcal{V}^{i\ell}, \mathcal{E}^{ik} \subset \partial\Omega_i\}.$$

Hence, $\widetilde{W} = \widetilde{W}_C$ is the subspace of W of functions that are continuous at the subdomain vertices and have common averages $\overline{u}_{\mathcal{E}^{ik}}$ for all the edges. The number of degrees of freedom of the corresponding primal subspace $\widehat{W}_\Pi$ is therefore equal to the sum of the number of vertices and the number of edges; this $\widehat{W}_\Pi$ will be of lower dimension than the primal space of Algorithm B.

The number of constraints enforced in all the iterations of Algorithms B and C is substantially larger than when only the vertex constraints are satisfied, as in Algorithm A, but we are still able to work with a uniformly bounded number of such constraints for each substructure. In order to put this in perspective, we consider Algorithms B and C in the very regular case of cubic substructures. There are then seven global variables for each interior substructure in the case of Algorithm B since there are eight vertices, each shared by eight cubes; twelve edges, each shared by four cubes; and six faces, each shared by a pair of substructures. The count for Algorithm

C is four. We note that the counts would be different, relative to the number of substructures, in the case of tetrahedral subregions.

It is useful to distinguish between the continuity constraints at the vertices and the other constraints. The latter are sometimes called optional constraints since they are not needed to guarantee solvability of the subproblems if there are enough vertex constraints. The optional constraints could be handled as the vertex variables after a change of basis. Another possibility, which we advocate, is to introduce an additional set of Lagrange multipliers which are computed exactly in each iteration to enforce the required optional constraints of the primal subspace; see Farhat, Lesoinne, and Pierson [10], where this approach is used; for a more detailed description, see section 4.2, especially formulae (24)–(28), of that paper.

We are able to show as strong a result for Algorithm C as for Algorithm B. It is therefore natural to attempt to drop additional constraints, i.e., further decrease the primal subspace $\widehat{W}_\Pi$ while preserving the fast convergence of the FETI-DP method. This leads to the introduction of our final algorithm.

ALGORITHM D. *The primal subspace, $\widehat{W}_\Pi$, is defined in terms of constraints associated with a subset of the edges and vertices of the interface. Our recipe for selecting such primal edges and vertices is relatively complicated and can only be fully understood by carefully reading the proof of Lemma* 10 *in section* 6.

We first describe the requirements on a minimal set of primal constraints which we have found necessary to give a complete proof of a good bound for Algorithm D. For each face, we should have at least one designated, primal edge. Additionally, for all pairs of substructures Ω_i, Ω_j, which have an edge in common, we must have an acceptable *edge path* between the two subdomains. An acceptable edge path is a path from Ω_i to Ω_j, possibly via several other subdomains, Ω_k, which have the edge $\mathcal{E}^{ij}$ in common and such that their coefficients satisfy $TOL*\rho_k \geq \min(\rho_i, \rho_j)$ for some chosen tolerance TOL. The path can only pass from one subdomain to another through an edge designated as primal. Finally, we consider all pairs of substructures which have in common a vertex $\mathcal{V}^{i\ell}$ but not a face or an edge. Then, we assume either that $\mathcal{V}^{i\ell}$ is a primal vertex or that we have an acceptable edge path of the same nature as above, except that we can be more lenient and insist only on $TOL*\rho_k \geq (h_k/H_k)\min(\rho_i, \rho_j)$. We also note that we could allow our edge paths to stray somewhat further away from the edge $\mathcal{E}^{ij}$, or the vertex $\mathcal{V}^{i\ell}$, and that in fact a careful examination of the proof of Lemma 10 would reveal that alternative, more liberal rules concerning the paths could be adopted.

We now give a description of a possible way of selecting the set of primal constraints. We start by choosing enough edges so that for each face of the interface there is at least one designated, primal edge which is part of the boundary of the face. In addition, we can exercise an option of designating some of the vertices of the substructures as primal; this is not strictly necessary but if constraints are enforced at enough vertices throughout the computation, then the related Schur complement can be made invertible even without any edge constraints. As pointed out above, this can be an advantage in the implementation of the method.

After this initial phase, which in the case of hexagonal substructures can involve as few as three edge constraints per subdomain, and hence a very small primal space, we consider the effects of the possibly very large variation of the coefficients ρ_i; if there are no great variations in the coefficients, we need do nothing more. We examine one by one each edge $\mathcal{E}^{ij}$ not previously designated as primal. We consider all pairs of subdomains that have this edge in common and try to find an acceptable edge path

between the two subdomains Ω_i and Ω_j. If no such path can be found, we add the edge $\mathcal{E}^{ij}$ to the set of designated edges; a trivial, acceptable edge path is then created. We also note that since two subdomains that share a face always have at least one designated edge in common, we need not consider any such pairs of subdomains in this step.

Finally, we consider, one by one, all vertices which so far have not been designated as primal. We consider pairs of substructures that have such a vertex $\mathcal{V}^{i\ell}$ in common but which do not have a face or edge in common. For each vertex inspected, we try to find an acceptable edge path subject only to the more lenient condition on the coefficients. If we fail to find such a path, we mark the vertex $\mathcal{V}^{i\ell}$ as primal, i.e., a vertex where the constraints should be exactly satisfied throughout the FETI iteration.

We note that we are free to add any other vertex, edge, or face constraints to our definition of the primal space; the bounds on the condition numbers will only improve. If all edges and vertices are primal, we are back to Algorithm C.

We can now formulate our FETI-DP algorithms. Each is expressed in terms of a Schur complement $\widetilde{S}$ related to the dual space $\widetilde{W}_\Delta$. We can arrive at this reduced problem by eliminating the primal variables associated with the interior nodes, the vertex nodes designated as primal, as well as the Lagrange multipliers related to the optional constraints. This Schur complement $\widetilde{S}$ can be equally well defined by a variational problem: for all $w_\Delta \in \widetilde{W}_\Delta$,

$$\langle \widetilde{S} w_\Delta, w_\Delta \rangle = \min \langle Sw, w \rangle, \tag{4.3}$$

where we take the minimum over all $w \in \widetilde{W}$ of the form $w = w_\Pi + w_\Delta, w_\Pi \in \widehat{W}_\Pi$. We note that any Schur complement of a positive definite, symmetric matrix is always associated with such a variational problem. We also obtain, analogously, a reduced right-hand side $\tilde{f}_\Delta$ from the load vectors associated with the individual subdomains.

We now reformulate the original finite element problem, reduced to the degrees of freedom of the second subspace $\widetilde{W}_\Delta$, as a minimization problem with constraints given by the requirement of continuity across all of Γ_h: find $u_\Delta \in \widetilde{W}_\Delta$, such that

$$\left.\begin{array}{c} J(u_\Delta) := \frac{1}{2}\langle \widetilde{S} u_\Delta, u_\Delta \rangle - \langle \tilde{f}_\Delta, u_\Delta \rangle \rightarrow \min \\ B_\Delta u_\Delta = 0 \end{array}\right\}. \tag{4.4}$$

The matrix B_Δ is constructed from $\{0, 1, -1\}$ in a way very similar to the matrix B discussed in section 3 and in such a way that the values of the solution u_Δ, associated with more than one subdomain, coincide when $B_\Delta u_\Delta = 0$. Again, these constraints are very simple and just express that the nodal values coincide across the interface; in comparison with the FETI method described in the previous section, we can drop some of the constraints, in particular, those associated with the vertex nodes of the primal space. However, we will otherwise use all possible constraints and thus work with a fully redundant set of Lagrange multipliers as in [13, sect. 5].

By introducing a set of Lagrange multipliers $\lambda \in V := \text{range}\,(B_\Delta)$ to enforce the constraints $B_\Delta u_\Delta = 0$, we obtain a saddle point formulation of (4.4), as in (3.5). Since $\widetilde{S}$ is invertible, we can eliminate the subvector u_Δ, and we obtain the following system for the dual variables:

$$F\lambda = d := B_\Delta \widetilde{S}^{-1} \tilde{f}_\Delta, \tag{4.5}$$

where

$$F := B_\Delta \widetilde{S}^{-1} B_\Delta^t.$$

Algorithmically, the matrix $\widetilde{S}$ is needed only in terms of $\widetilde{S}^{-1}$ times a vector, and such an operation can be computed relatively inexpensively. While it is natural to describe a Schur complement in terms of a second set of variables and resulting from the elimination of a first set, the action of its inverse on a vector can often be obtained advantageously by solving the entire linear system from which it originates after augmenting the given right-hand side with zeros. Full advantage can then be taken of algorithms that symmetrically reorder the larger matrix so as to preserve sparsity. In the case at hand, it is thus advantageous to group together all the interior and dual variables of each subdomain and to factor the resulting blocks in parallel across the subdomains using a good ordering algorithm. The contributions to the remaining Schur complement, of the primal variables, can also be computed locally prior to subassembly and factorization of this final, global part of the linear system of equations.

The operator F will obviously depend on the choice of the subspaces $\widehat{W}_\Pi$ and $\widetilde{W}_\Delta$, and we denote the operators of the resulting linear systems by F_A, F_B, F_C, and F_D, respectively. To define the FETI-DP Dirichlet preconditioner, we need to introduce an additional set of local Schur complement matrices, $S_\Delta^{(i)}$, which is obtained by restricting $S^{(i)}$ to the space $\widetilde{W}_{\Delta,i}$; in the case of Algorithm A, we simply remove the rows and columns corresponding to the subdomain vertices from $S^{(i)}$. The associated block-diagonal matrix is given by

$$S_\Delta := \operatorname{diag}_{i=1}^N(S_\Delta^{(i)}).$$

We can compute S_Δ times a vector $w_\Delta \in \widetilde{W}_\Delta$ by solving a local Dirichlet problem with a solution in $\widetilde{W}_{\Delta,i},\ i = 1, \ldots, N$, and then multiplying it by the stiffness matrix of its subdomains. Such solutions are constrained to vanish at designated subdomain vertices and to have zero edge and face averages, as required by the algorithm in question.

We also introduce diagonal scaling matrices $D_\Delta^{(i)}$ that operate on the Lagrange multiplier spaces. Each of their diagonal elements corresponds to a Lagrange multiplier which enforces continuity between the nodal values of some $w_i \in \widetilde{W}_i$ and $w_j \in \widetilde{W}_j$ at some point $x \in \Gamma_h$; it is given by $\rho_j^\gamma(x)\mu_j^\dagger(x)$. Finally, we define a scaled jump operator by

$$B_{D,\Delta} := \left[D_\Delta^{(1)} B_\Delta^{(1)}, \ldots, D_\Delta^{(N)} B_\Delta^{(N)}\right].$$

As in Klawonn and Widlund [13, sect. 5], we solve the dual system (4.5) using the preconditioned conjugate gradient algorithm with the preconditioner

$$M^{-1} := B_{D,\Delta} S_\Delta B_{D,\Delta}^t. \tag{4.6}$$

The FETI-DP method is the standard preconditioned conjugate gradient algorithm for solving the preconditioned system

$$M^{-1} F\lambda = M^{-1} d.$$

This definition of M clearly depends on the choice of the subspaces $\widehat{W}_\Pi$ and $\widetilde{W}_\Delta$ for the different algorithms. The resulting preconditioners are denoted by $M_A^{-1}, M_B^{-1}, M_C^{-1}$, and M_D^{-1}, respectively.

5. Some auxiliary lemmas. The purpose of this section is to provide, in most cases without proofs, the few auxiliary results that are required for a complete proof of Lemmas 9 and 10, which provide the core of the proofs of our main results. Some of these results are borrowed from [6, 8, 7]; see also [25] for similar material. Here, we formulate them using trace spaces on the subdomain boundaries, i.e., $H^{1/2}(\partial\Omega_i)$ instead of the spaces $H^1(\Omega_i)$ and discrete harmonic extensions; given the well-known equivalence of the norms, nothing essentially new needs to be proven. In our proofs, we will work with the S-norm defined by $|u|_S^2 = \sum_{i=1}^N |u_i|_{S^{(i)}}^2$ and $|u_i|_{S^{(i)}}^2 = \langle S^{(i)}u_i, u_i\rangle$. A proof of the equivalence of $S^{(i)}$- and $H^{1/2}(\partial\Omega_i)$-seminorms of elements of W_i can be found in [1] for the case of piecewise linear elements and two dimensions. The tools necessary to extend this result to more general finite elements are provided in [22]; in our case, we of course have to multiply $|u_i|_{H^{1/2}(\partial\Omega_i)}^2$ by the factor ρ_i.

We also recall that we can define the $H_{00}^{1/2}(\tilde\Gamma)$-norm, $\tilde\Gamma \subset \partial\Omega_i$, of an element of W_i which is supported in $\tilde\Gamma$, as the $H^{1/2}(\partial\Omega_i)$-norm of the function extended by zero onto $\partial\Omega_i \setminus \tilde\Gamma$.

The first lemma can, essentially, be found in Dryja, Smith, and Widlund [6, Lem. 4.4].

LEMMA 1. *Let $\theta_{\mathcal{F}^{ij}}$ be the finite element function that is equal to* 1 *at the nodal points on the face $\mathcal{F}^{ij}$, which is common to two subregions Ω_i and Ω_j, and that vanishes on $(\partial\Omega_{i,h} \cup \partial\Omega_{j,h}) \setminus \mathcal{F}_h^{ij}$. Then,*

$$|\theta_{\mathcal{F}^{ij}}|_{H^{1/2}(\partial\Omega_i)}^2 \leq C(1+\log(H_i/h_i))H_i.$$

The same bounds also hold for the other subregion Ω_j.

The following result can, essentially, be found in Dryja, Smith, and Widlund [6, Lem. 4.5] or in Dryja [3, Lem. 3].

LEMMA 2. *Let $\theta_{\mathcal{F}^{ij}}$ be the function introduced in Lemma* 1 *and let I^h denote the interpolation operator onto the finite element space $W^h(\Omega_i)$. Then, for all $u \in W_i$,*

$$\|I^h(\theta_{\mathcal{F}^{ij}}u)\|_{H_{00}^{1/2}(\mathcal{F}^{ij})}^2 \leq C(1+\log(H_i/h_i))^2\left(|u|_{H^{1/2}(\mathcal{F}^{ij})}^2 + \frac{1}{H_i}\|u\|_{L_2(\mathcal{F}^{ij})}^2\right).$$

We will also need two additional results which are used to estimate the contributions to our bounds from the edges of Ω_i. For the next lemma, see Dryja, Smith, and Widlund [6, Lem. 4.7].

LEMMA 3. *Let $\theta_{\mathcal{E}^{ik}}$ be the cutoff function associated with the edge $\mathcal{E}^{ik}$. Then, for all $u \in W_i$,*

$$|I^h(\theta_{\mathcal{E}^{ik}}u)|_{H^{1/2}(\partial\Omega_i)}^2 \leq C\|u\|_{L_2(\mathcal{E}^{ik})}^2.$$

This result follows by an elementary estimate of the energy norm of the zero extension of the boundary values and by noting that the harmonic extension has a smaller energy.

We will also need a Sobolev-type inequality for finite element functions; see Dryja and Widlund [7, Lem. 3.3] or Dryja [3, Lem. 1].

LEMMA 4. *Let $\mathcal{E}^{ik}$ be any edge of Ω_i which forms part of the boundary of a face $\mathcal{F}^{ij} \subset \partial\Omega_i$. Then, for all $u \in W_i$,*

$$\|u\|_{L_2(\mathcal{E}^{ik})}^2 \leq C(1+\log(H_i/h_i))\left(|u|_{H^{1/2}(\mathcal{F}^{ij})}^2 + \frac{1}{H_i}\|u\|_{L_2(\mathcal{F}^{ij})}^2\right).$$

We also state a nonstandard version of Friedrichs' inequality that is given in a somewhat different form in [8, Lem. 6].

LEMMA 5. *Let $\mathcal{E}^{ik}$ be an edge of $\mathcal{F}^{ij}$. Then, for all $u \in W_i$ that vanish on $\mathcal{E}^{ik}$,*

$$\|u\|^2_{L_2(\mathcal{F}^{ij})} \leq CH_i(1+\log(H_i/h_i))|u|^2_{H^{1/2}(\mathcal{F}^{ij})}.$$

The proof of the main results in Mandel and Tezaur [17] is based on a bound for a certain interpolation operator. In our proofs, we also could use a different interpolation operator for each of our algorithms. Although these operators now play no direct role in the proofs of our main results, they are nevertheless of independent interest. They also illustrate how, in the cases of Algorithms B and C, we can approximate an arbitrary element in $\widetilde{W}_B$ and $\widetilde{W}_C$, respectively, by a continuous interpolant which is almost uniformly stable in the energy norm; concerning $\widetilde{W}_D$, see Remark 1.

The first interpolation operator, I^h_A, is given by the continuous piecewise linear interpolant on the coarse triangulation of Γ used in the definition of the Ω_i.

Our second interpolation operator I^h_B is defined, for all $u \in \widetilde{W}_B$, by sums over all the vertices, edges, and faces of Γ:

$$I^h_B u(x) = \sum_{\mathcal{V}^{i\ell}\in\Gamma} u(\mathcal{V}^{i\ell})\theta_{\mathcal{V}^{i\ell}}(x) + \sum_{\mathcal{E}^{ik}\subset\Gamma} \overline{u}_{\mathcal{E}^{ik}}\theta_{\mathcal{E}^{ik}}(x) + \sum_{\mathcal{F}^{ij}\subset\Gamma} \overline{u}_{\mathcal{F}^{ij}}\theta_{\mathcal{F}^{ij}}(x). \tag{5.1}$$

The operator I^h_B, a modification of an operator introduced in [6, p. 1690], has almost optimal stability properties. We note that the values of $I^h_B u(x)$ on $\partial\Omega_i$ depend only on the W_i component of u.

We also introduce a third interpolation operator, I^h_C, which provides an alternative to I^h_B:

$$I^h_C u(x) = \sum_{\mathcal{V}^{i\ell}\in\Gamma} u(\mathcal{V}^{i\ell})\theta_{\mathcal{V}^{i\ell}}(x) + \sum_{\mathcal{E}^{ik}\subset\Gamma} \overline{u}_{\mathcal{E}^{ik}}\theta_{\mathcal{E}^{ik}}(x) + \sum_{\mathcal{F}^{ij}\subset\Gamma} \overline{u}_{\partial\mathcal{F}^{ij}}\theta_{\mathcal{F}^{ij}}(x). \tag{5.2}$$

Here the average $\overline{u}_{\mathcal{E}^{ik}}$ is defined as in (4.2), and $\overline{u}_{\partial\mathcal{F}^{ij}}$ is given by

$$\overline{u}_{\partial\mathcal{F}^{ij}} = \frac{\int_{\partial\mathcal{F}^{ij}} u ds}{\int_{\partial\mathcal{F}^{ij}} 1 ds}.$$

This average is a convex combination of the values of the $\overline{u}_{\mathcal{E}^{ik}}$ of the face in question. This interpolant is well defined for any element $u \in \widetilde{W}_C$.

The next lemma provides L_2- and $H^{1/2}$-estimates for the vertex-based interpolation operator I^h_A. This is essentially Lemma 4.1 of Dryja, Smith, and Widlund [6]. The proof follows directly from Poincaré's inequality and a standard discrete Sobolev inequality; see also [6, sect. 4].

LEMMA 6. *The vertex-based interpolation operator I^h_A satisfies*

$$|I^h_A u|^2_{H^{1/2}(\mathcal{F}^{ij})} \leq C\,(H_i/h_i)\,|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i$$

and

$$\|u - I^h_A u\|^2_{L_2(\mathcal{F}^{ij})} \leq C\,(H_i/h_i)\,H_i\,|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i.$$

Here the constant C is independent of the diameter H_i of Ω_i and the mesh size h_i.

We have better results for the interpolation operators I^h_B and I^h_C, introduced in (5.1) and (5.2), respectively. A bound for I^h_B can be found in a somewhat different

form in Dryja, Smith, and Widlund [6, pp. 1689–1690]. We note that our L_2-estimate is now improved in comparison to [6, p. 1690] since our estimate of the interpolation error contains no logarithmic factor.

LEMMA 7. *The interpolation operators I_B^h and I_C^h, defined in* (5.1) *and* (5.2), *respectively, satisfy*

$$|I_B^h u|^2_{H^{1/2}(\mathcal{F}^{ij})} \leq C(1+\log(H_i/h_i))|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i,$$
$$|I_C^h u|^2_{H^{1/2}(\mathcal{F}^{ij})} \leq C(1+\log(H_i/h_i))|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i$$

and

$$\|u - I_B^h u\|^2_{L_2(\mathcal{F}^{ij})} \leq CH_i|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i,$$
$$\|u - I_C^h u\|^2_{L_2(\mathcal{F}^{ij})} \leq CH_i(1+\log(H_i/h_i))|u|^2_{H^{1/2}(\mathcal{F}^{ij})} \quad \forall u \in W_i.$$

Here the constant C is independent of the diameter H_i of Ω_i and the mesh size h_i.

6. Convergence analysis. Our analysis borrows ideas from the recent paper by Mandel and Tezaur [17] and from our own paper [13]. In the latter, fast one-level FETI algorithms and a theory for the elliptic problem of the class defined by (2.3) were developed for an arbitrary choice of the ρ_i.

As in [17], the two different Schur complements, $\widetilde{S}$ and S_Δ, introduced in section 4 play an important role in the analysis of the dual-primal iterative algorithm. Both operate on the second subspace $\widetilde{W}_\Delta$, and we also recall that $\widetilde{S}$ represents a global problem while S_Δ does not.

Let $V := \text{range}\,(B_\Delta)$ be the space of Lagrange multipliers. As in [13, sect. 5], we introduce a projection

$$P_\Delta := B_{D,\Delta}^t B_\Delta.$$

A simple computation shows (see [13, Lem. 4.2]) that P_Δ preserves the jump of any function $u_\Delta \in \widetilde{W}_\Delta$, i.e., $B_\Delta P_\Delta u_\Delta = B_\Delta u_\Delta$, and we also have $P_\Delta u = 0$ for all $u \in \widehat{W}$.

Analogous to [13, Lem. 5.2], we have the following.

LEMMA 8. *For any $\mu \in V$, there exists a $w_\Delta \in \text{range}\,(P_\Delta)$ such that $\mu = B_\Delta w_\Delta$.*

Proof. We note that for any $\mu \in V = \text{range}\,(B_\Delta)$, there exists a w'_Δ such that $\mu = B_\Delta w'_\Delta$. Choosing $w_\Delta := P_\Delta w'_\Delta$, we have $B_\Delta w_\Delta = B_\Delta w'_\Delta = \mu$. ∎

Let $x \in \Gamma_h$ and let $w_\Delta \in \widetilde{W}_\Delta$. We borrow the following formula from [13]:

$$P_\Delta w_\Delta(x) \;=\; \sum_{j \in \mathcal{N}_{\Delta,x}} \rho_j^\gamma \mu_j^\dagger (w_{\Delta,i}(x) - w_{\Delta,j}(x)), x \in \partial\Omega_{i,h} \cap \Gamma_h. \tag{6.1}$$

Here, $\mathcal{N}_{\Delta,x}$ is the set of indices of the subregions which have the node x on their boundaries. We note that the coefficients in this expression are constant on the set of the nodal points of each face and each edge of $\partial\Omega_i$ and that this formula is independent of the particular choice of B_Δ.

We first analyze Algorithm B and begin by proving the following core estimate.

LEMMA 9 (Algorithm B). *For all $w_\Delta \in \widetilde{W}_{\Delta,B}$, we have*

$$|P_\Delta w_\Delta|^2_{S_\Delta} \leq C\,(1+\log(H/h))^2 |w_\Delta|^2_{\widetilde{S}},$$

where $C > 0$ is independent of h, H, γ, and the ρ_i.

Proof. We consider an arbitrary $w_\Delta \in \widetilde{W}_{\Delta,B}$. In order to compute its $\widetilde{S}$-norm (cf. (4.3)), we determine the element $w = w_\Pi + w_\Delta \in \widetilde{W}_B, w_\Pi \in \widehat{W}_{\Pi,B}$, with the correct minimal property. Then, by the definition of $\widetilde{S}$, $|w_\Delta|_{\widetilde{S}} = |w|_S$. We next note that we can subtract any continuous function from w_Δ without changing the values of $P_\Delta w_\Delta$; thus, $P_\Delta w = P_\Delta w_\Delta$. It is also easy to see, by carrying out a simple computation and by using formula (6.1), that $P_\Delta w_\Delta \in \widetilde{W}_{\Delta,B}$. We also recall that the S_Δ-norm of any element of $\widetilde{W}_\Delta$ equals its S-norm.

We model our proof on [13, Lem. 4.7 and 5.4] but note that the arguments need to be modified to some extent. We also note that we have contributions only from faces and edges since all elements in $\widetilde{W}_B$ are continuous at the vertices. Here, in contrast to the proof in [13], we do not need to assume that there is no subdomain, with boundaries, that intersects $\partial\Omega_D$ only in isolated points.

We introduce the notation $(v_i)_{i=1,\dots,N} := P_\Delta w$. Then, we have to estimate

$$|P_\Delta w|_S^2 = \sum_{i=1}^N |v_i|_{S^{(i)}}^2.$$

We can therefore focus on the estimate of the contribution from a single subdomain Ω_i. We first assume that its boundary and the boundaries of its relevant neighbors do not intersect $\partial\Omega_D$.

We cut the function v_i, using the functions $\theta_{\mathcal{F}^{ij}}$ and $\theta_{\mathcal{E}^{ik}}$, and write it as a sum of terms which vanish at all the interface nodes outside individual faces and edges; cf., e.g., [6, 8, 7]. We then have, since v_i vanishes at the subdomain vertices,

$$v_i = \sum_{\mathcal{F}^{ij}\subset\partial\Omega_i} I^h(\theta_{\mathcal{F}^{ij}} v_i) + \sum_{\mathcal{E}^{ik}\subset\partial\Omega_i} I^h(\theta_{\mathcal{E}^{ik}} v_i).$$

We find that the face $\mathcal{F}^{ij}$ contributes

$$I^h(\theta_{\mathcal{F}^{ij}} \rho_j^\gamma \mu_j^\dagger (w_i - w_j))$$

and we have to estimate its $H_{00}^{1/2}(\mathcal{F}^{ij})$-norm; this formula follows from (6.1).

With $\gamma \geq 1/2$, we can easily prove that

$$\rho_i(\rho_j^\gamma \mu_j^\dagger)^2 \leq \min(\rho_i, \rho_j). \tag{6.2}$$

We note that $\rho_j^\gamma \mu_j^\dagger$ is constant on $\mathcal{F}_h^{ij}$ and that w has common face averages, i.e., $\overline{w}_{i,\mathcal{F}^{ij}} = \overline{w}_{j,\mathcal{F}^{ij}}$. Using inequality (6.2), these observations, and Lemma 2, we obtain

(6.3)

$$\begin{aligned}
&\rho_i \|I^h(\theta_{\mathcal{F}^{ij}} \rho_j^\gamma \mu_j^\dagger (w_i - w_j))\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})} \\
&= \rho_i \|I^h(\theta_{\mathcal{F}^{ij}} \rho_j^\gamma \mu_j^\dagger ((w_i - \overline{w}_{i,\mathcal{F}^{ij}}) - (w_j - \overline{w}_{j,\mathcal{F}^{ij}})))\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})} \\
&\leq C\,(1 + \log(H_i/h_i))^2 \min(\rho_i, \rho_j) \Big(|w_i - w_j|^2_{H^{1/2}(\mathcal{F}^{ij})} \\
&\qquad\qquad + \frac{1}{H_i} \|(w_i - \overline{w}_{i,\mathcal{F}^{ij}}) - (w_j - \overline{w}_{j,\mathcal{F}^{ij}})\|^2_{L_2(\mathcal{F}^{ij})} \Big).
\end{aligned}$$

We can estimate this expression by

$$C\,(1 + \log(H_i/h_i))^2 \left(\rho_i |w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + \rho_j |w_j|^2_{H^{1/2}(\mathcal{F}^{ij})} \right),$$

as desired, by applying a Poincaré inequality. We note that, by assumption, H_j and H_i are comparable and so are h_j and h_i, since the triangulations of Ω_i and Ω_j are quasi-uniform.

By using Lemma 3, we can estimate the contributions of the edges of Ω_i to the energy of v_i in terms of L_2-norms over the edges. These L_2-terms are then estimated by using Lemma 4. If four subdomains, e.g., $\Omega_i, \Omega_j, \Omega_k$, and Ω_ℓ, have an edge $\mathcal{E}^{ik}$ in common, then, according to (6.1), there are three contributions to the estimate of the contribution of Ω_i to $|P_\Delta w|_S^2$, namely,

$$\begin{aligned}\rho_i \, \|I^h(\rho_j^\gamma \, \mu_j^\dagger \, \theta_{\mathcal{E}^{ik}}(w_i - w_j))\|^2_{L_2(\mathcal{E}^{ik})} \; &+ \; \rho_i \, \|I^h(\rho_k^\gamma \, \mu_k^\dagger \, \theta_{\mathcal{E}^{ik}}(w_i - w_k))\|^2_{L_2(\mathcal{E}^{ik})} \\ &+ \; \rho_i \, \|I^h(\rho_\ell^\gamma \, \mu_\ell^\dagger \, \theta_{\mathcal{E}^{ik}}(w_i - w_\ell))\|^2_{L_2(\mathcal{E}^{ik})}.\end{aligned} \tag{6.4}$$

We first consider the second term in detail, assuming that Ω_i shares a face with each of Ω_j and Ω_ℓ but only an edge with Ω_k. In the next estimate, we use $|\overline{w}_{i,\mathcal{E}^{ik}}|^2 \leq 1/H_i \|w_i\|^2_{L_2(\mathcal{E}^{ik})}$ and $\|\theta_{\mathcal{E}^{ik}}\|^2_{L_2(\mathcal{E}^{ik})} \leq C\, H_i$. Using formula (6.2), Lemma 4, and the fact that w has common edge averages, i.e., $\overline{w}_{i,\mathcal{E}^{ik}} = \overline{w}_{k,\mathcal{E}^{ik}}$, we obtain

$$\begin{aligned}
& \rho_i \, \|I^h(\rho_k^\gamma \mu_k^\dagger \theta_{\mathcal{E}^{ik}}(w_i - w_k))\|^2_{L_2(\mathcal{E}^{ik})} \\
= \; & \rho_i \, \|I^h(\rho_k^\gamma \mu_k^\dagger (\theta_{\mathcal{E}^{ik}}(w_i - \overline{w}_{i,\mathcal{E}^{ik}}) - \theta_{\mathcal{E}^{ik}}(w_k - \overline{w}_{k,\mathcal{E}^{ik}})))\|^2_{L_2(\mathcal{E}^{ik})} \\
\leq \; & 2 \left(\rho_i \|I^h(\theta_{\mathcal{E}^{ik}}(w_i - \overline{w}_{i,\mathcal{E}^{ik}}))\|^2_{L_2(\mathcal{E}^{ik})} + \rho_k \|I^h(\theta_{\mathcal{E}^{ik}}(w_k - \overline{w}_{k,\mathcal{E}^{ik}}))\|^2_{L_2(\mathcal{E}^{ik})} \right) \\
\leq \; & C \left(\rho_i \|w_i\|^2_{L_2(\mathcal{E}^{ik})} + \rho_k \|w_k\|^2_{L_2(\mathcal{E}^{ik})} \right) \\
\leq \; & C(1 + \log(H/h)) \left(\rho_i \left(|w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + \tfrac{1}{H_i} \|w_i\|^2_{L_2(\mathcal{F}^{ij})} \right) \right. \\
& \qquad \left. + \rho_k \left(|w_k|^2_{H^{1/2}(\mathcal{F}^{kj})} + \tfrac{1}{H_k} \|w_k\|^2_{L_2(\mathcal{F}^{kj})} \right) \right) \\
\leq \; & C(1 + \log(H/h)) \left(\rho_i |w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + \rho_k |w_k|^2_{H^{1/2}(\mathcal{F}^{kj})} \right),
\end{aligned} \tag{6.5}$$

with $\mathcal{F}^{ij}$ a face of Ω_i and $\mathcal{F}^{kj}$ a face of Ω_k, both of which have the edge $\mathcal{E}^{ik}$ in common. The last inequality follows from the shift invariance of the expressions on the third line; i.e., we can add constants to w_i and w_k without changing the value of the expressions and then use Poincaré's inequality.

Since Ω_i and Ω_j, as well as Ω_i and Ω_ℓ, have a face in common, the argument given above could be simplified for the first and third edge contributions; they can be reduced to estimates for face terms directly.

Finally, we have to consider boundary subregions which have a nonempty intersection with $\partial\Omega_D$ and show that we can obtain bounds of the same quality. We then need different arguments to eliminate the $L_2(\mathcal{F}^{ij})$ terms. In case this intersection is a face or an edge, we can use exactly the same arguments as in [13, p. 71] which include using Lemma 5. If the boundary of a substructure intersects $\partial\Omega_D$ in just one or a few single points, the shifting can be done exactly as above for the face and edge terms of an interior subregion. ∎

We now prove our condition number estimate for Algorithm B, which only depends polylogarithmically on the dimension of the subproblems.

THEOREM 1 (Algorithm B). *The condition number satisfies*

$$\kappa(M_B^{-1} F_B) \leq C \, (1 + \log(H/h))^2.$$

Here, C is independent of h, H, γ, and the values of the ρ_i.

Proof. We have to estimate the smallest eigenvalue $\lambda_{min}(M_B^{-1}F_B)$ from below and the largest eigenvalue $\lambda_{max}(M_B^{-1}F_B)$ from above. We will show that

$$\langle M_B\lambda,\lambda\rangle \leq \langle F_B\lambda,\lambda\rangle \leq C\,(1+\log(H/h))^2\langle M_B\lambda,\lambda\rangle \quad \forall \lambda\in V. \tag{6.6}$$

Lower bound. This bound is derived using purely algebraic arguments. As in the analysis of the one-level FETI methods, we can use the following formula (see Mandel and Tezaur [16] or Klawonn and Widlund [13, p. 73]):

$$\langle F_B\lambda,\lambda\rangle = \sup_{0\neq v_\Delta\in\widetilde{W}_\Delta} \frac{\langle\lambda, B_\Delta v_\Delta\rangle^2}{|v_\Delta|^2_{\widetilde{S}}}.$$

Let $\mu\in V$ be arbitrary. It then follows from Lemma 8 that there exists a $w_\Delta\in$ range (P_Δ) with $\mu = B_\Delta w_\Delta$. Since $w_\Delta = P_\Delta w_\Delta$ and $|u_\Delta|_{\widetilde{S}}\leq |u_\Delta|_{S_\Delta}$ for all $u_\Delta\in\widetilde{W}_\Delta$, we obtain

$$\langle F_B\lambda,\lambda\rangle \geq \frac{\langle\lambda,B_\Delta w_\Delta\rangle^2}{|w_\Delta|^2_{\widetilde{S}}} \geq \frac{\langle\lambda,B_\Delta w_\Delta\rangle^2}{|w_\Delta|^2_{S_\Delta}} = \frac{\langle\lambda,\mu\rangle^2}{|B^t_{D,\Delta}\mu|^2_{S_\Delta}} = \frac{\langle\lambda,\mu\rangle^2}{\langle M_B^{-1}\mu,\mu\rangle}.$$

The left inequality of (6.6) follows by choosing $\mu := M_B\lambda$.

Upper bound. Using Lemma 9, we obtain, for all $\lambda\in V$,

$$\begin{aligned}
\langle F_B\lambda,\lambda\rangle &= \sup_{0\neq w_\Delta\in\widetilde{W}_\Delta} \frac{\langle\lambda,B_\Delta w_\Delta\rangle^2}{|w_\Delta|^2_{\widetilde{S}}} \\
&\leq C\,(1+\log(H/h))^2 \sup_{w_\Delta\neq 0} \frac{\langle\lambda,B_\Delta w_\Delta\rangle^2}{|P_\Delta w_\Delta|^2_{S_\Delta}} \\
&= C\,(1+\log(H/h))^2 \sup_{w_\Delta\neq 0} \frac{\langle\lambda,B_\Delta w_\Delta\rangle^2}{\langle M_B^{-1}B_\Delta w_\Delta, B_\Delta w_\Delta\rangle} \\
&= C\,(1+\log(H/h))^2 \sup_{\mu\in V} \frac{\langle\lambda,\mu\rangle^2}{\langle M_B^{-1}\mu,\mu\rangle} = C\,(1+\log(H/h))^2\langle M_B\lambda,\lambda\rangle. \qquad \square
\end{aligned}$$

We now turn to the analysis of Algorithms C and D.

LEMMA 10 (Algorithms C and D). *For all $w_\Delta\in\widetilde{W}_{\Delta,C}$, we have*

$$|P_\Delta w_\Delta|^2_{S_\Delta} \leq C\,(1+\log(H/h))^2|w_\Delta|^2_{\widetilde{S}}.$$

For all $w_\Delta\in\widetilde{W}_{\Delta,D}$, we have

$$|P_\Delta w_\Delta|^2_{S_\Delta} \leq C\,\max(1,TOL)\,(1+\log(H/h))^2|w_\Delta|^2_{\widetilde{S}}.$$

In both cases, $C>0$ is independent of h, H, γ, and the ρ_i.

Proof. We can proceed as in the proof of Lemma 9; we will use the same notation and discuss only details that are technically different. We note that in Algorithm D all vertices are not necessarily constrained and that therefore we have to estimate terms of $P_\Delta w(x)$ related to the vertices which are not primal.

We cut the function v_i using the functions $\theta_{\mathcal{F}^{ij}}$, $\theta_{\mathcal{E}^{ik}}$, and $\theta_{\mathcal{V}^{il}}$ and write it as a sum of terms which vanish at all the interface nodes outside individual faces, edges, and vertices, respectively; cf., e.g., [6, 8, 7]. We then have

$$v_i = \sum_{\mathcal{F}^{ij}\subset\partial\Omega_i} I^h(\theta_{\mathcal{F}^{ij}}v_i) + \sum_{\mathcal{E}^{ik}\subset\partial\Omega_i} I^h(\theta_{\mathcal{E}^{ik}}v_i) + \sum_{\mathcal{V}^{il}\in\partial\Omega_i} \theta_{\mathcal{V}^{il}}v_i(\mathcal{V}^{il}).$$

As in [13] and the proof of Lemma 9, we find that the face $\mathcal{F}^{ij}$ contributes

$$I^h(\theta_{\mathcal{F}^{ij}}\rho_j^\gamma\mu_j^\dagger(w_i-w_j))$$

and we have to estimate its $H_{00}^{1/2}(\mathcal{F}^{ij})$-norm. Using inequality (6.2) and the fact that $\rho_j^\gamma\mu_j^\dagger$ is constant on $\mathcal{F}_h^{ij}$, we obtain

$$\begin{aligned}
&\rho_i\|I^h(\theta_{\mathcal{F}^{ij}}\rho_j^\gamma\mu_j^\dagger(w_i-w_j))\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}\\
=\;&\rho_i\|I^h(\theta_{\mathcal{F}^{ij}}\rho_j^\gamma\mu_j^\dagger((w_i-\overline{w}_{i,\mathcal{F}^{ij}})-(w_j-\overline{w}_{j,\mathcal{F}^{ij}})\\
&\qquad+(\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{F}^{ij}})))\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}\\
\le\;&2\min(\rho_i,\rho_j)\left(\|I^h(\theta_{\mathcal{F}^{ij}}((w_i-\overline{w}_{i,\mathcal{F}^{ij}})-(w_j-\overline{w}_{j,\mathcal{F}^{ij}})))\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}\right.\\
&\qquad\left.+\|(\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{F}^{ij}})\theta_{\mathcal{F}^{ij}}\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}\right).
\end{aligned}\tag{6.7}$$

The first term can be estimated as in (6.3) by

$$C\,(1+\log(H_i/h_i))^2\left(\rho_i|w_i|^2_{H^{1/2}(\mathcal{F}^{ij})}+\rho_j|w_j|^2_{H^{1/2}(\mathcal{F}^{ij})}\right),$$

as desired, by applying a Poincaré inequality. There remains to estimate $\|((\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{F}^{ij}})\theta_{\mathcal{F}^{ij}}\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}$. Let $\mathcal{E}^{ik}\subset\partial\mathcal{F}^{ij}$ be a designated, primal edge. Then, we have

$$|\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{F}^{ij}}|^2\le 2\left(|\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{i,\mathcal{E}^{ik}}|^2+|\overline{w}_{j,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{E}^{ik}}|^2\right).$$

It is sufficient to consider the first term on the right-hand side. The shift invariance allows us to assume that $\overline{w}_{i,\mathcal{F}^{ij}}=0$. Using $|\overline{w}_{\mathcal{E}^{ik}}|^2\le C/H_i\|w_i\|^2_{L_2(\mathcal{E}^{ik})}$ and Lemmas 1 and 4, we obtain

$$\|(\overline{w}_{i,\mathcal{F}^{ij}}-\overline{w}_{j,\mathcal{F}^{ij}})\theta_{\mathcal{F}^{ij}}\|^2_{H_{00}^{1/2}(\mathcal{F}^{ij})}\le C\,(1+\log(H/h))^2\left(|w_i|^2_{H^{1/2}(\mathcal{F}^{ij})}+|w_j|^2_{H^{1/2}(\mathcal{F}^{ij})}\right).$$

The remainder of the proof of the result for Algorithm C can be carried out as in the proof of Lemma 9. However, for Algorithm D, we need to do some further work.

Proceeding as in the proof of Lemma 9, we can estimate the contributions of the edges of Ω_i to the energy of v_i in terms of L_2-norms over the edges. We first consider the second term of (6.4) in detail, again assuming that Ω_i shares a face with each of Ω_j and Ω_ℓ but only an edge with Ω_k. If we have a trivial, acceptable edge path, i.e., the common edge is designated as primal, we can proceed exactly as in (6.5). Otherwise assume that we have a nontrivial, acceptable edge path through the subdomain Ω_j via the edges $\mathcal{E}^{ij}$ and $\mathcal{E}^{jk}$; in general the acceptable edge path could be more complicated but such a case could be analyzed similarly. We obtain

$$\begin{aligned}
&\rho_i\,\|\rho_k^\gamma\mu_k^\dagger I^h(\theta_{\mathcal{E}^{ik}}(w_i-w_k))\|^2_{L_2(\mathcal{E}^{ik})}\\
=\;&\rho_i\,\|\rho_k^\gamma\mu_k^\dagger\left(I^h(\theta_{\mathcal{E}^{ik}}(w_i-\overline{w}_{i,\mathcal{E}^{ij}}))+\theta_{\mathcal{E}^{ik}}(\overline{w}_{j,\mathcal{E}^{ij}}-\overline{w}_{j,\mathcal{E}^{jk}})\right.\\
&\qquad\left.-I^h(\theta_{\mathcal{E}^{ik}}(w_k-\overline{w}_{k,\mathcal{E}^{jk}}))\right)\|^2_{L_2(\mathcal{E}^{ik})}\\
\le\;&C\min(\rho_i,\rho_k)\left(\|I^h(\theta_{\mathcal{E}^{ik}}(w_i-\overline{w}_{i,\mathcal{E}^{ij}}))\|^2_{L_2(\mathcal{E}^{ik})}+H_j\,|\overline{w}_{j,\mathcal{E}^{ij}}-\overline{w}_{j,\mathcal{E}^{jk}}|^2\right.\\
&\qquad\left.+\|I^h(\theta_{\mathcal{E}^{ik}}(w_k-\overline{w}_{k,\mathcal{E}^{jk}}))\|^2_{L_2(\mathcal{E}^{ik})}\right).
\end{aligned}\tag{6.8}$$

The terms of the last expression can be estimated as in (6.5). The only difference is that, additionally, we have to use $TOL * \rho_j \geq \min(\rho_i, \rho_k)$. We obtain

$$\rho_i \, \|\rho_k^\gamma \mu_k^\dagger I^h(\theta_{\mathcal{E}^{ik}}(w_i - w_k))\|^2_{L_2(\mathcal{E}^{ik})} \leq C(1+\log(H/h)) \left(\rho_i |w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + \rho_k |w_k|^2_{H^{1/2}(\mathcal{F}^{ik})} + TOL * \rho_j \left(|w_j|^2_{H^{1/2}(\mathcal{F}^{ij})} + |w_j|^2_{H^{1/2}(\mathcal{F}^{jk})}\right)\right).$$

Since Ω_i and Ω_j, as well as Ω_i and Ω_ℓ, have a face in common, the argument given above could be simplified for the first and third edge contributions (see (6.4)); they can be reduced to estimates of face terms.

Finally, we consider the terms resulting from the vertices. We have, according to (6.1),

$$\begin{aligned} &\rho_i |\theta_{\mathcal{V}^{i\ell}} v_i(\mathcal{V}^{i\ell})|^2_{H^{1/2}(\partial\Omega_i)} \\ &\leq C \sum_{j \in \mathcal{N}_{\Delta,\mathcal{V}^{i\ell}}} \rho_i (\rho_j^\gamma \mu_j^\dagger)^2 |\theta_{\mathcal{V}^{i\ell}}|^2_{H^{1/2}(\partial\Omega_i)} |w_i(\mathcal{V}^{i\ell}) - w_j(\mathcal{V}^{i\ell})|^2 \\ &\leq C \sum_{j \in \mathcal{N}_{\Delta,\mathcal{V}^{i\ell}}} \min(\rho_i, \rho_j) h_i |w_i(\mathcal{V}^{i\ell}) - w_j(\mathcal{V}^{i\ell})|^2. \end{aligned}$$

We now consider each pair of substructures separately. Let Ω_i, Ω_l be such a pair and assume that we have an acceptable edge path through Ω_j via the edges $\mathcal{E}^{ij}$ and $\mathcal{E}^{jl}$ with the condition

$$TOL * \rho_j \geq \frac{h_j}{H_j} \min(\rho_i, \rho_l). \tag{6.9}$$

We can proceed as in the analysis of the edge terms and obtain

$$\begin{aligned} &\min(\rho_i, \rho_l) h_i |w_i(\mathcal{V}^{il}) - w_l(\mathcal{V}^{il})|^2 \\ &\leq 3 \, \min(\rho_i, \rho_l) h_i \left(|w_i(\mathcal{V}^{il}) - \overline{w}_{i,\mathcal{E}^{ij}}|^2 + |\overline{w}_{j,\mathcal{E}^{ij}} - \overline{w}_{j,\mathcal{E}^{jl}}|^2 + |w_l(\mathcal{V}^{il}) - \overline{w}_{l,\mathcal{E}^{jl}}|^2\right). \end{aligned}$$

It is sufficient to estimate the first term on the last line; the third term can be treated in exactly the same way, and the second term can be estimated as above with the only difference of an additional factor h_j/H_j which is accounted for in (6.9). Using $h_i |w_i(\mathcal{V}^{il})|^2 \leq C \|w_i\|^2_{L_2(\mathcal{E}^{ij})}$ and Lemma 4, and estimating $|\overline{w}_{i,\mathcal{E}^{ij}}|$ as before, we obtain

$$\begin{aligned} |w_i(\mathcal{V}^{il}) - \overline{w}_{i,\mathcal{E}^{ij}}|^2 &\leq 2 \left(|w_i(\mathcal{V}^{il})|^2 + |\overline{w}_{i,\mathcal{E}^{ij}}|^2\right) \\ &\leq C(1+\log(H_i/h_i)) h_i^{-1} \left(|w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + 1/H_i \|w_i\|^2_{L_2(\mathcal{F}^{ij})}\right) \\ &\leq C(1+\log(H_i/h_i)) h_i^{-1} |w_i|^2_{H^{1/2}(\mathcal{F}^{ij})}. \end{aligned}$$

Here, the last line follows again from the shift invariance of the first expression. Using (6.9), we finally obtain

$$\begin{aligned} &\min(\rho_i, \rho_l) h_i |w_i(\mathcal{V}^{il}) - w_l(\mathcal{V}^{il})|^2 \\ &\leq C \, (1+\log(H/h)) \left(\rho_i |w_i|^2_{H^{1/2}(\mathcal{F}^{ij})} + \rho_\ell |w_\ell|^2_{H^{1/2}(\mathcal{F}^{j\ell})} + TOL * \rho_j \left(|w_j|^2_{H^{1/2}(\mathcal{F}^{ij})} + |w_j|^2_{H^{1/2}(\mathcal{F}^{j\ell})}\right)\right). \end{aligned}$$

The boundary subregions can again be treated as in the proof of Lemma 9. ∎

We can now prove our condition number estimates for Algorithms C and D, which are as strong as those in Theorem 1. The proof can be carried out exactly as for Theorem 1, using Lemma 10 instead of Lemma 9.

THEOREM 2 (Algorithms C and D). *The condition numbers satisfy*

$$\kappa(M_C^{-1}F_C) \leq C\,(1+\log(H/h))^2$$

and

$$\kappa(M_D^{-1}F_D) \leq C\,\max(1,TOL)\,(1+\log(H/h))^2.$$

Here, C is independent of h, H, γ, and the values of the ρ_i.

Remark 1. It is possible to define a fourth interpolation operator I_D^h, based on the weights ρ_i, the pseudoinverses $\mu_i^\dagger$, and the averages over the subdomain boundaries, by

$$I_D^h u(x) = \sum_i \overline{u}_{\partial\Omega_i}\rho_i^\gamma(x)\mu_i^\dagger(x). \tag{6.10}$$

Here the average $\overline{u}_{\partial\Omega_i}$ is defined by

$$\overline{u}_{\partial\Omega_i} = \frac{\int_{\partial\Omega_i} u ds}{\int_{\partial\Omega_i} 1 ds},$$

where we use the component in W_i when computing this average. This operator naturally appears in studies of Neumann–Neumann algorithms. We can establish the same type of bounds as for I_C^h in Lemma 7, provided that we introduce the same constraints as for Algorithm D.

Remark 2. It is already known from the numerical results in [9, 10] that Algorithm A is not competitive. We can prove that the condition number of Algorithm A satisfies the weaker bound,

$$\kappa(M_A^{-1}F_A) \leq C\,(H/h)\,(1+\log(H/h))^2,$$

in the same way as Theorem 1, using Lemma 6 and a variant of Lemma 10. Here, C is independent of h, H, γ, and the values of the ρ_i.

REFERENCES

[1] P. E. BJØRSTAD AND O. B. WIDLUND, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures,* SIAM J. Numer. Anal., 23 (1986), pp. 1097–1120.

[2] J. H. BRAMBLE, J. E. PASCIAK, AND A. H. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring,* I, Math. Comp., 47 (1986), pp. 103–134.

[3] M. DRYJA, *A method of domain decomposition for three-dimensional finite element problems,* in Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski et al., eds., SIAM, Philadelphia, 1988, pp. 43–61.

[4] M. DRYJA, W. PROSKUROWSKI, AND O. WIDLUND, *A method of domain decomposition with crosspoints for elliptic finite element problems,* in Optimal Algorithms, Bl. Sendov, ed., Bulgarian Academy of Sciences, Sofia, Bulgaria, 1986, pp. 97–111.

[5] M. DRYJA, M. V. SARKIS, AND O. B. WIDLUND, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions,* Numer. Math., 72 (1996), pp. 313–348.

[6] M. DRYJA, B. F. SMITH, AND O. B. WIDLUND, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions,* SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

[7] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap,* SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[8] M. Dryja and O. B. Widlund, *Schwarz methods of Neumann-Neumann type for three-dimensional elliptic finite element problems,* Comm. Pure Appl. Math., 48 (1995), pp. 121–155.

[9] C. Farhat, M. Lesoinne, P. Le Tallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method – part* I: *A faster alternative to the two-level FETI method,* Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[10] C. Farhat, M. Lesoinne, and K. Pierson, *A scalable dual-primal domain decomposition method,* Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.

[11] C. Farhat, J. Mandel, and F.-X. Roux, *Optimal convergence properties of the FETI domain decomposition method,* Comput. Methods Appl. Mech. Engrg., 115 (1994), pp. 367–388.

[12] A. Klawonn and O. B. Widlund, *A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity*, SIAM J. Sci. Comput., 22 (2000), pp. 1199–1219.

[13] A. Klawonn and O. B. Widlund, *FETI and Neumann–Neumann iterative substructuring methods: Connections and new results,* Comm. Pure Appl. Math., 54 (2001), pp. 57–90.

[14] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods with face constraints,* in Recent Developments in Domain Decomposition Methods, Lecture Notes in Comput. Sci. Engrg. 23, L. F. Pavarino and A. Toselli, eds., Springer-Verlag, New York, 2002, to appear.

[15] J. Mandel and M. Brezina, *Balancing domain decomposition for problems with large jumps in coefficients,* Math. Comp., 65 (1996), pp. 1387–1401.

[16] J. Mandel and R. Tezaur, *Convergence of a substructuring method with Lagrange multipliers,* Numer. Math., 73 (1996), pp. 473–487.

[17] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method,* Numer. Math., 88 (2001), pp. 543–558.

[18] L. F. Pavarino and O. B. Widlund, *Iterative substructuring methods for spectral element discretizations of elliptic systems* I: *Compressible linear elasticity,* SIAM J. Numer. Anal., 37 (1999), pp. 353–374.

[19] K. H. Pierson, *A Family of Domain Decomposition Methods for the Massively Parallel Solution of Computational Mechanics Problems*, Ph.D. thesis, Aerospace Engineering, University of Colorado at Boulder, Boulder, CO, 2000.

[20] M. V. Sarkis, *Schwarz Preconditioners for Elliptic Problems with Discontinuous Coefficients Using Conforming and Non-Conforming Elements*, Ph.D. thesis, Courant Institute, New York University, New York, NY, 1994.

[21] R. Tezaur, *Analysis of Lagrange Multiplier Based Domain Decomposition*, Ph.D. thesis, Department of Mathematics, University of Colorado at Denver, Denver, CO, 1998. Available online at http://www-math.cudenver.edu/graduate/thesis/rtezaur.ps.gz.

[22] O. B. Widlund, *An extension theorem for finite element spaces with three applications,* in Numerical Techniques in Continuum Mechanics, Proceedings of the Second GAMM-Seminar, (Kiel, January 1986), Notes Numer. Fluid Mech. 16, W. Hackbusch and K. Witsch, eds., Friedr. Vieweg und Sohn, Braunschweig/Wiesbaden, 1987, pages 110–122.

[23] O. B. Widlund, *Iterative substructuring methods: Algorithms and theory for elliptic problems in the plane,* in Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski et al., eds., SIAM, Philadelphia, 1988, pp. 113–128.

[24] O. B. Widlund, *Exotic coarse spaces for Schwarz methods for lower order and spectral finite elements,* in Proceedings of Seventh International Conference of Domain Decomposition Methods on Scientific and Engineering Computing (University Park, PA, October 27–30, 1993), Contemp. Math. 180, D. E. Keyes and J. Xu, eds., AMS, Providence, RI, 1994, pages 131–136.

[25] J. Xu and J. Zou, *Some nonoverlapping domain decomposition methods,* SIAM Rev., 40 (1998), pp. 857–914.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 180–197

A STABLE AFFINE-APPROXIMATE FINITE ELEMENT METHOD*

K. ARUNAKIRINATHAR† AND B. D. REDDY‡

Abstract. The notion of the affine figure closest to a given quadrilateral can be given a precise mathematical definition. The resulting figure is referred to as the equivalent parallelogram associated with a quadrilateral. Equipped with such a concept, it is then feasible to consider finite element approximations in which the original quadrilateral elements are replaced by their equivalent parallelograms. The idea is appealing, not least because of the resulting economy arising from computations performed on an element generated by an affine map. Furthermore, numerical experiments reported recently indicate that highly efficient and accurate schemes result when such a concept is combined with the enhanced strain method or the method of incompatible modes. The purpose of this work is to analyze finite element schemes resulting from approximation of quadrilaterals by their equivalent parallelograms. The focus is on low-order (bilinear) elements, and the analysis is carried out in the context of linear elasticity for standard approximations as well as for those which use enhanced strains. The affine approximation applies only to the element map, and the primary unknown (the displacement vector in the context of elasticity) is approximated by conventional piecewise bilinear functions. The analysis confirms convergence at the optimal rate, provided that the deviations of the quadrilaterals from their equivalent parallelograms are at most $O(h)$.

Key words. equivalent parallelogram, affine approximations, four-noded quadrilateral, error estimates

AMS subject classifications. 65N30, 65N15

PII. S0036142900382442

1. Introduction.

The use of low-order elements in finite element analyses of complex problems carries with it significant advantages. Most particularly, such schemes are highly economical and for this reason are attractive.

Finite element analyses based on four-noded quadrilaterals in two dimensions, and on eight-noded hexahedral elements in three, are widely used. Unfortunately, they are not without their drawbacks. In problems of solid mechanics in which bending deformations dominate, analyses based on these elements exhibit poor accuracy, at least when coarse meshes are used. In addition, in the incompressible limit, or when the compressibility is small, locking behavior is experienced.

There is a vast literature that is devoted to the construction of methods which are intended to overcome the problems referred to, while retaining the advantages of using low-order elements. One commonly used set of remedies is that based on a combination of underintegration plus stabilization (see, for example, the work of [4] and [8]). The great advantage of this approach is its efficiency, in that only a single integration point is used. However, the eigenvalues of the stiffness matrix are required in the process, and it is not possible to evaluate these without a relatively high degree of effort, for nonaffine elements.

Another popular approach is that associated with the enrichment or enhancement of the strain by the addition of suitably chosen basis functions. The key work dealing with enhanced strain formulations is [14], which in turn contains as a special case

*Received by the editors December 13, 2000; accepted for publication (in revised form) October 9, 2001; published electronically April 12, 2002.
http://www.siam.org/journals/sinum/40-1/38244.html
†Department of Mathematics, University of Swaziland, PO Luyengo M205, Kwaluseni, Swaziland.
‡Faculty of Science, University of Cape Town, 7701 Rondebosch, South Africa (bdr@science.uct.ac.za). The work of this author was supported by the National Research Foundation of South Africa.

the nonconforming method of incompatible modes due to Wilson, Taylor, Doherty, and Ghaboussi [17] for rectangular elements and extended by Taylor, Beresford, and Wilson [16] to incorporate arbitrary quadrilaterals. The method has been successfully extended to nonlinear problems (see, for example, [15]). Reddy and Simo [11] have shown, for linear problems and for affine elements, that the enhanced strain method is stable and convergent, while Arunakirinathar and Reddy [3] have extended that work to include the case of arbitrary quadrilaterals.

The enhanced strain method is still not without its drawbacks. For example, the quality of approximations for arbitrary elements declines with an increase in distortion of the elements. Consideration of this shortcoming, together with a desire to improve the efficiency of computations associated with arbitrary quadrilaterals or hexahedra, leads naturally to the notion of replacing the arbitrary quadrilateral by the affine element that is closest to it, in a manner that can be made precise. Such an affine element is known as the equivalent parallelogram in two dimensions, and the equivalent parallelepiped in three. It is important to bear in mind that the affine approximation applies only to the element map and that the primary unknown (the displacement vector in the context of elasticity) is still approximated by piecewise bilinear functions in two dimensions and by piecewise trilinear functions in three.

It has been shown in [2] that the interpolation error obtained by using the equivalent parallelogram instead of the original quadrilateral is of the same order as that corresponding to the usual interpolation error. The element stiffness matrices associated with the equivalent elements are therefore admissible alternatives to the "exact" stiffness matrices of the original elements, while at the same time they are far easier to construct. This set of ideas has been proposed, and then tested numerically, first in the context of problems of linear elasticity in [9] and subsequently for problems involving nonlinearly elastic materials in [12, 13]. In all cases the numerical results are encouraging and suggest a significant improvement in efficiency and accuracy when this approach is used, particularly in circumstances in which element distortions are significant.

The purpose of this work is to carry out a detailed analysis of these affine-approximate finite element methods, for linear problems. The analysis is confined to plane problems but can be extended to three dimensions with little difficulty, though the details are messy. The analysis includes treatment of finite elements without and with the inclusion of enhanced strains. The key results, with respect to both classes of approximations, is that the method converges at the optimal rate provided that the element distortion is sufficiently small—more precisely, provided that deviation of the quadrilateral from the equivalent parallelogram is of the order of mesh size. This notion will be made precise in what follows.

The plan of the remainder of this work is as follows. In section 2 the problem is formulated. Finite element approximations are introduced in section 3, as is the notion of the equivalent parallelogram. The analysis of the affine-approximate method is carried in section 4 for the problem without enhancement, while section 5 is devoted to an analysis of the problem with enhancement.

2. The boundary-value problem of elasticity. The model problem of relevance is the displacement boundary-value problem of linear elasticity. Suppose that a linear elastic body occupies a region $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$. The body has boundary Γ. Then the governing equations which are required to hold on Ω are the equation of equilibrium

$$-\operatorname{div} \boldsymbol{\sigma} = \boldsymbol{b}, \tag{1}$$

the strain-displacement relation

$$\boldsymbol{\epsilon}(\boldsymbol{u}) = \tfrac{1}{2}(\nabla \boldsymbol{u} + [\nabla \boldsymbol{u}]^T), \tag{2}$$

and the elastic constitutive equation

$$\boldsymbol{\sigma} = \mathbb{C}[\boldsymbol{\epsilon}]. \tag{3}$$

Here $\boldsymbol{\sigma}$ is the symmetric Cauchy stress tensor, $\boldsymbol{\epsilon}$ is the infinitesimal strain tensor, $\boldsymbol{u}$ is the displacement vector, $\mathbb{C}$ is the fourth-order elasticity tensor, and $\boldsymbol{b}$ is a prescribed body force vector.

For convenience we assume that the displacement satisfies the homogeneous Dirichlet boundary condition, that is,

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma. \tag{4}$$

The tensor $\mathbb{C}$ has the symmetries

$$\mathbb{C}_{ijkl} = \mathbb{C}_{jikl} = \mathbb{C}_{ijlk} = \mathbb{C}_{klij} \tag{5}$$

and is assumed to be pointwise stable, in the sense that there exists a constant $c_0 > 0$ such that

$$\mathbb{C}_{ijkl} M_{ij} M_{kl} \geq c_0 M_{ij} M_{ij} \tag{6}$$

for all symmetric matrices $\boldsymbol{M}$, the summation convention for repeated indices being invoked here and henceforth. In addition, the components of $\mathbb{C}$ are assumed to be bounded measurable functions; that is,

$$\mathbb{C}_{ijkl} \in L^\infty(\Omega) \tag{7}$$

for all indices i, j, k, l ranging over 1 to d, with

$$c_\infty := \max_{i,j,k,l} \operatorname{ess\,sup}\{\mathbb{C}_{ijkl}(\boldsymbol{x}) : \ \boldsymbol{x} \in \Omega \ \}. \tag{8}$$

For isotropic elastic materials the elasticity tensor takes the simple form

$$\mathbb{C}[\boldsymbol{\epsilon}] = \lambda \operatorname{tr} \boldsymbol{\epsilon} + 2\mu\, \boldsymbol{\epsilon},$$

in which λ and μ are the Lamé constants. Pointwise stability of $\mathbb{C}$ is equivalent to the condition that the Lamé constants satisfy the inequalities [10]

$$\mu > 0 \quad \text{and} \quad \lambda > -\tfrac{2}{3}\mu.$$

We will make use of the space $L^2(\Omega)$ of square-integrable functions defined on Ω. The inner product and norm on this space are denoted, respectively, by $(\cdot, \cdot)_0$ and $\| \cdot \|_0$. We recall also the definition of the Sobolev spaces $H^m(\Omega)$, where m is an integer; for nonnegative m, these are equivalence classes of functions which, together with their generalized derivatives up to and including those of order m, are members of $L^2(\Omega)$. The Sobolev spaces are Hilbert spaces with inner product and associated norm given by

$$(u, v)_m = \int_\Omega \sum_{|\alpha| \leq m} D^\alpha u(\boldsymbol{x}) D^\alpha v(\boldsymbol{x}) \, d\boldsymbol{x}, \quad \|u\|_m = (u, u)_m^{1/2}$$

for all $u, v \in H^m(\Omega)$. Here $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index whose components α_i are nonnegative integers, $|\alpha| = \alpha_1 + \cdots + \alpha_d$, and $D^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$. The seminorm $|\cdot|_m$ on $H^m(\Omega)$ is defined by

$$|u|_m = \int_\Omega \sum_{|\alpha|=m} D^\alpha u(\boldsymbol{x}) D^\alpha v(\boldsymbol{x})\ d\boldsymbol{x}.$$

We define the space $H_0^m(\Omega)$ of functions in $H^m(\Omega)$ which, together with their derivatives of order up to and including those of order $m-1$, vanish on the boundary in the sense of traces. The seminorm $|\cdot|_m$ is a norm on $H_0^m(\Omega)$, equivalent to the standard norm $\|\cdot\|_m$. Finally, the space $H^{-m}(\Omega)$, for m a nonnegative integer, may be identified with the topological dual space of $H_0^m(\Omega)$.

We are now in a position to define the standard variational problem in linear elasticity. For this purpose we denote by $V := [H_0^1(\Omega)]^d$ the space of admissible displacements and define the bilinear form $a(\cdot,\cdot)$ and linear functional $\ell(\cdot)$ by

$$a : V \times V \to \mathbb{R}, \qquad a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \mathbb{C}\boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\epsilon}(\boldsymbol{v})\ d\boldsymbol{x}, \tag{9}$$

$$\ell : V \to \mathbb{R}, \qquad \ell(\boldsymbol{v}) = \int_\Omega \boldsymbol{b} \cdot \boldsymbol{v}\ d\boldsymbol{x}. \tag{10}$$

The properties of $\mathbb{C}$ guarantee that $a(\cdot,\cdot)$ is symmetric, continuous, and V-elliptic; that is, $a(\boldsymbol{v}, \boldsymbol{u}) = a(\boldsymbol{u}, \boldsymbol{v})$, and there exist positive constants M and α such that

$$|a(\boldsymbol{u}, \boldsymbol{v})| \le M\|\boldsymbol{u}\|_V \|\boldsymbol{v}\|_V \quad \text{and} \quad a(\boldsymbol{v}, \boldsymbol{v}) \ge \alpha \|\boldsymbol{v}\|_V^2$$

for all $\boldsymbol{u}$, $\boldsymbol{v} \in V$.

The topological dual of V is denoted by V'.

The standard variational problem is as follows.

Problem S. Given $\boldsymbol{b} \in V'$, find $\boldsymbol{u} \in V$ which satisfies

$$a(\boldsymbol{u}, \boldsymbol{v}) = \ell(\boldsymbol{v}) \tag{11}$$

for all $\boldsymbol{v} \in V$.

It is well known (see, for example, [7]) that Problem S has a unique solution, which satisfies the bound

$$\|\boldsymbol{u}\|_V \le (1/\alpha)\, \|\ell\|_{V'}. \tag{12}$$

3. Finite element approximations. We confine attention to plane situations, so that $d = 2$. The domain Ω is assumed to be polygonal, and a finite element mesh $\mathcal{T}$ of quadrilateral elements is constructed on Ω in the usual manner. A typical element K in $\mathcal{T}$ is generated by an isoparametric map F from a reference element $\hat{K} \equiv (-1,1) \times (-1,1)$. The mesh parameter h is defined by

$$h = \max_{K \in \mathcal{T}} \sup\{|\boldsymbol{x} - \boldsymbol{y}| : \quad \boldsymbol{x}, \boldsymbol{y} \in K\}. \tag{13}$$

We define basis functions $\hat{N}_A \quad (A = 1, \ldots, 4)$ on $\hat{K}$ by

$$\hat{N}_A(\boldsymbol{\xi}) = \tfrac{1}{4}(1 + \xi\xi_A)(1 + \eta\eta_A),$$

where $\boldsymbol{\xi}_A \equiv (\xi_A, \eta_A)$ are the nodal coordinates on $\hat{K}$ with $(\xi_1 \cdots \xi_4) = (1\ -1\ -1\ 1)$ and $(\eta_1 \cdots \eta_4) = (1\ 1\ -1\ -1)$. Denote by Q_1 the space of bilinear functions spanned by $\hat{N}_A$. Then it is convenient to express the map F in the form

$$F : \hat{K} \to K, \qquad \boldsymbol{x} = F(\boldsymbol{\xi}) = \sum_{A=1}^{4} \boldsymbol{x}_A \hat{N}_A(\boldsymbol{\xi}), \tag{14}$$

in which $\boldsymbol{x}_A$ are the nodal points of K.

The Jacobian matrix $\boldsymbol{J}$ is defined to be the gradient of the map F and is the matrix with components

$$J_{ij} = \frac{\partial F_i(\boldsymbol{\xi})}{\partial \xi_j}, \quad i, j = 1, 2.$$

We also set

$$j = \det \boldsymbol{J}.$$

Next we define the space V^h by

$$V^h = \{\boldsymbol{v}_h \in V : \ (v_h)_i|_K \circ F \in Q_1\}. \tag{15}$$

In other words, if we define the function $\hat{\boldsymbol{v}}$ on $\hat{K}$ by

$$\hat{\boldsymbol{v}}(\boldsymbol{\xi}) = \boldsymbol{v}|_K(\boldsymbol{x}),$$

in which $\boldsymbol{\xi}$ and $\boldsymbol{x}$ are related through (14), then $\hat{\boldsymbol{v}} \in Q_1$.

The standard discrete variational problem takes the following form.

Problem $\mathrm{S^h}$. Given $\boldsymbol{b} \in V'$, find $\boldsymbol{u}_h \in V^h$ which satisfies

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h) = \ell(\boldsymbol{v}_h) \tag{16}$$

for all $\boldsymbol{v}_h \in V^h$.

It is well known (see [7]) that Problem $\mathrm{S^h}$ has a unique solution, and furthermore, provided that $\boldsymbol{u} \in [H^2(\Omega)]^2$, there exists a constant $C > 0$, depending on Ω and on u, but independent of h, such that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \leq Ch.$$

It is instructive, and relevant to the developments that follow, to note that the bilinear form and linear functional appearing in (16) are usually evaluated on the reference element. In order to do this it is necessary to carry out transformations of the functions appearing in $a(\cdot, \cdot)$ and $\ell(\cdot)$. Thus, if we define $\hat{\boldsymbol{v}}$ as above, the chain rule gives

$$\frac{\partial \hat{v}_i}{\partial \xi_j} = \sum_{k=1}^{2} \frac{\partial v_i}{\partial x_k} \frac{\partial F_k}{\partial \xi_j},$$

or if we define the tensors or matrices $\boldsymbol{L}$ and $\hat{\boldsymbol{L}}$ by

$$L_{ij} = \frac{\partial v_i}{\partial x_j}, \quad \hat{L}_{ij} = \frac{\partial \hat{v}_i}{\partial \xi_j},$$

then

$$\hat{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{J}.$$

The transformation of the strain tensor may now be easily carried out, and we have, using an obvious notation,

$$\begin{aligned} \boldsymbol{\epsilon}(\boldsymbol{v}) &= \tfrac{1}{2}(\boldsymbol{L} + \boldsymbol{L}^T) \\ &= \tfrac{1}{2}(\hat{\boldsymbol{L}}\boldsymbol{J}^{-1} + \boldsymbol{J}^{-T}\hat{\boldsymbol{L}}^T) := \hat{\boldsymbol{\epsilon}}(\hat{\boldsymbol{v}}). \end{aligned} \tag{17}$$

Hence the bilinear form may be evaluated according to

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_{K \in \mathcal{T}} a_K(\boldsymbol{u}_h, \boldsymbol{v}_h) \tag{18}$$

in which

$$\begin{aligned} a_K(\boldsymbol{u}_h, \boldsymbol{v}_h) &= \int_K \mathbb{C}[\boldsymbol{\epsilon}(\boldsymbol{u}_h)] : \boldsymbol{\epsilon}(\boldsymbol{v}_h)\ dxdy \\ &= \int_{\hat{K}} \hat{\mathbb{C}}[\hat{\boldsymbol{\epsilon}}(\hat{\boldsymbol{u}}_h)] : \hat{\boldsymbol{\epsilon}}(\hat{\boldsymbol{v}}_h)\ j\, d\xi d\eta \end{aligned} \tag{19}$$

and where $\hat{\mathbb{C}} = \mathbb{C} \circ F$.

3.1. The equivalent parallelogram. The notion of the equivalent parallelogram associated with a quadrilateral arises naturally when one seeks to define the parallelogram that is closest to the quadrilateral in a sense that can be made mathematically precise. Assuming that such a parallelogram can be constructed, the quadrilateral can then be viewed as a perturbation of the parallelogram.

This problem was considered by Arunakirinathar and Reddy [2], who showed that the equivalent parallelogram can be constructed as follows. Suppose that the map from the reference element $\hat{K}$ to an arbitrary quadrilateral K is given by (14); then the equivalent parallelogram $\tilde{K}$ associated with K is defined by the *affine* map $\tilde{F}$ obtained simply by discarding the bilinear terms in (14). That is, if we define the vector $\boldsymbol{k}$ by

$$\boldsymbol{k} = \tfrac{1}{4}(\boldsymbol{x}_1 - \boldsymbol{x}_2 + \boldsymbol{x}_3 - \boldsymbol{x}_4),$$

then the map $\tilde{F}$ may be expressed in the form

$$\begin{aligned} \tilde{F}(\boldsymbol{\xi}) &= F(\boldsymbol{\xi}) - \boldsymbol{k}\xi\eta \\ &= \sum_{A=1}^{4} N_A(\boldsymbol{\xi})\tilde{\boldsymbol{x}}_A, \end{aligned}$$

in which the nodal points $\tilde{\boldsymbol{x}}_A$ of the equivalent parallelogram are defined by

$$\tilde{\boldsymbol{x}}_A = \tfrac{3}{4}\boldsymbol{x}_A + \tfrac{1}{4}(\boldsymbol{x}_{A+1} - \boldsymbol{x}_{A+2} + \boldsymbol{x}_{A+3}), \quad A = 1, \ldots, 4 \text{ (modulo 4)}.$$

These notions are illustrated in Figure 1.

The equivalent parallelogram has some interesting properties [2]: for example, K and $\tilde{K}$ have the same areas, their sides intersect at midpoints, and the lengths of the corresponding diagonals are equal. These last two properties are evident in Figure 1.

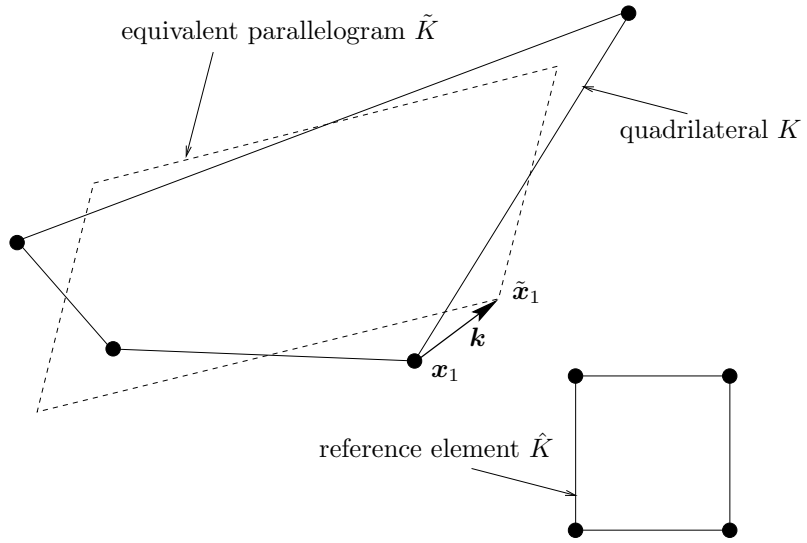


FIG. 1. *The equivalent parallelogram associated with a quadrilateral.*

It is necessary to characterize mathematically the relationship between the quadrilateral and its equivalent parallelogram; more particularly, we need to characterize the notion of closeness between K and $\tilde{K}$. Suppose that the affine map from $\hat{K}$ to $\tilde{K}$ takes the form

$$\tilde{\boldsymbol{x}} = \boldsymbol{C}\boldsymbol{\xi} + \boldsymbol{c}, \tag{20}$$

in which $\boldsymbol{C}$ and $\boldsymbol{c}$ are, respectively, a constant matrix and vector; then the *distortion parameter* τ_K for element K is defined by

$$\tau_K = |\boldsymbol{C}^{-1}\boldsymbol{k}|. \tag{21}$$

The distortion is thus measured by mapping the vector $\boldsymbol{k}$ back to the reference element $\hat{K}$, using $\boldsymbol{C}^{-1}$. From Figure 1 and the definition of $\boldsymbol{k}$ it is clear that $\boldsymbol{k} = \boldsymbol{0}$, and $\tau_K = 0$, if and only if K is a parallelogram.

The distortion parameter associated with a finite element mesh may now be defined by

$$\tau = \max_{K \in \mathcal{T}} |\tau_K|. \tag{22}$$

We will also require the notion of an *h-regular mesh*, which is defined to be a finite element mesh for which $\tau = O(h)$.

The role played by τ_K in characterizing the difference between K and $\tilde{K}$ may be seen more clearly by examining the properties of the map $G : \tilde{K} \to K$ from the parallelogram to the quadrilateral. If we set $\boldsymbol{J}' = DG$ and $j' = \det \boldsymbol{J}'$, then since

$$\boldsymbol{x} = G(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}} + \boldsymbol{k}\xi\eta,$$

it is straightforward to show, also using (20), that

$$\boldsymbol{J}' = \boldsymbol{I} + \boldsymbol{C}^{-1}\boldsymbol{k} \otimes \boldsymbol{l}$$

and

$$j' = 1 + \boldsymbol{C}^{-1}\boldsymbol{k} \cdot \boldsymbol{l},$$

in which $\boldsymbol{l} = (\eta, \xi)$. It follows from (21) that

$$\sup_{\boldsymbol{x} \in K} |j'(\boldsymbol{x})| := \|j'\|_\infty \leq 1 + \sqrt{2}\,\tau_K. \tag{23}$$

If $|\boldsymbol{A}| = (\sum_{i,j=1}^{2} A_{ij}A_{ij})^{1/2}$ for any matrix $\boldsymbol{A}$, then

$$\begin{aligned} |\boldsymbol{J}'| &\leq |\boldsymbol{I}| + |\boldsymbol{C}^{-1}\boldsymbol{k} \otimes \boldsymbol{l}| \\ &\leq |\boldsymbol{I}| + |\boldsymbol{C}^{-1}\boldsymbol{k}|\,|\boldsymbol{l}| \\ &\leq \sqrt{2}\,(1 + \tau_K). \end{aligned} \tag{24}$$

The inverse of the map $\boldsymbol{J}'$ is given by

$$(\boldsymbol{J}')^{-1} = \boldsymbol{I} - \boldsymbol{R}, \tag{25}$$

in which

$$\boldsymbol{R} = \frac{\boldsymbol{C}^{-1}\boldsymbol{k} \otimes \boldsymbol{l}}{1 + \boldsymbol{C}^{-1}\boldsymbol{k} \cdot \boldsymbol{l}}. \tag{26}$$

It follows from (21) that

$$|\boldsymbol{R}| \leq \frac{|\boldsymbol{C}^{-1}\boldsymbol{k}|\,|\boldsymbol{l}|}{|1 + \boldsymbol{C}^{-1}\boldsymbol{k} \cdot \boldsymbol{l}|} \leq \frac{\sqrt{2}\tau_K}{1 - \sqrt{2}\tau_K}. \tag{27}$$

Likewise, using the identity

$$\det(\boldsymbol{I} + \boldsymbol{B}) = 1 + \det \boldsymbol{B}(1 + \operatorname{tr} \boldsymbol{B}^{-1}) + \operatorname{tr} \boldsymbol{B}$$

and (25) and (26), we find that

$$\begin{aligned} (j')^{-1} &= \det((\boldsymbol{J}')^{-1}) \\ &= 1 - \frac{\boldsymbol{C}^{-1}\boldsymbol{k} \cdot \boldsymbol{l}}{1 + \boldsymbol{C}^{-1}\boldsymbol{k} \cdot \boldsymbol{l}} \\ &\leq \frac{1}{1 - \sqrt{2}\tau_K} \end{aligned}$$

so that

$$\|(j')^{-1}\|_\infty \leq \frac{1}{1 - \sqrt{2}\tau_K}. \tag{28}$$

Thus, roughly, it is seen that the Jacobians of $\tilde{K}$ and K differ by a term that is $O(\tau_K)$.

Finally, a transformation similar to that in (17) may be obtained by defining on $\tilde{K}$ a coordinate system $\tilde{\boldsymbol{x}} = (\tilde{x}, \tilde{y})$ and by defining the gradient $\tilde{\nabla}$ and strain $\tilde{\boldsymbol{\epsilon}}$ relative to this coordinate system by

$$(\tilde{\nabla}\tilde{\boldsymbol{v}})_{ij} = \frac{\partial \tilde{v}_i}{\partial \tilde{x}_j}, \quad \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}) = \frac{1}{2}(\tilde{\nabla}\tilde{\boldsymbol{v}} + [\tilde{\nabla}\tilde{\boldsymbol{v}}]^T). \tag{29}$$

For convenience we set

$$\tilde{\boldsymbol{L}}(\tilde{\boldsymbol{v}}) = \tilde{\nabla}\tilde{\boldsymbol{v}};$$

then we have

$$\tilde{\boldsymbol{L}}(\tilde{\boldsymbol{v}}) = \boldsymbol{L}(\boldsymbol{v})\tilde{\boldsymbol{J}}. \tag{30}$$

4. Affine finite element approximations. Since the equivalent parallelogram associated with a quadrilateral is "close" to that quadrilateral, it is natural to enquire whether approximate solutions to Problem S of the desired degree of accuracy might be obtained if Problem S^{h} were modified by replacing the integral in (18) with an integral over $\tilde{K}$ for each element K in the mesh. Such a procedure would have the advantage that, since the map from K to $\tilde{K}$ is affine, the associated Jacobian matrix and determinant are constant, and the integrals can be evaluated exactly, for homogeneous materials at least, for which case $\mathbb{C}$ is constant.

We now show that such a procedure does in fact lead to finite element approximations that converge at the usual rate. Numerical results presented by Küssner and Reddy [9] in the case of linear elasticity and by Reese and coworkers [12, 13] for problems involving nonlinear elasticity and finite deformations show in addition that, when this concept is applied to enhanced assumed strain formulations, the results represent in many cases an improvement over those obtained by the conventional approach.

From (25), (29), and (30), we have

$$\begin{aligned}\boldsymbol{\epsilon}(\boldsymbol{v}) &= \tfrac{1}{2}(\boldsymbol{L}(\boldsymbol{v}) + \boldsymbol{L}^T(\boldsymbol{v})) \\ &= \tfrac{1}{2}(\tilde{\boldsymbol{L}}(\boldsymbol{J}')^{-1} + (\boldsymbol{J}')^{-T}\tilde{\boldsymbol{L}}^T) \\ &= \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}) - \Delta(\tilde{\boldsymbol{v}}),\end{aligned} \tag{31}$$

in which

$$\Delta(\tilde{\boldsymbol{v}}) := \tfrac{1}{2}[\tilde{\boldsymbol{L}}\boldsymbol{R} + \boldsymbol{R}^T\tilde{\boldsymbol{L}}^T], \tag{32}$$

and $\boldsymbol{R}$ is defined by (26).

Next, for continuous functions $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$ on $\tilde{K}$ we define the bilinear form $a_{\tilde{K}}(\cdot,\cdot)$ and linear functional $\ell_{\tilde{K}}(\cdot)$ by

$$a_{\tilde{K}}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}) = \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{u}})] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}})\; d\tilde{x}d\tilde{y}, \quad \ell_{\tilde{K}}(\tilde{\boldsymbol{v}}) = \int_{\tilde{K}} \boldsymbol{b}\cdot\tilde{\boldsymbol{v}}\; d\tilde{x}d\tilde{y}. \tag{33}$$

Here, and henceforth, $\tilde{\mathbb{C}}(\tilde{\boldsymbol{x}}) := \mathbb{C}(G(\tilde{\boldsymbol{x}}))$, and we write $d\tilde{x}d\tilde{y}$ for dx_1dx_2 for convenience. Likewise, for functions $\boldsymbol{u}$, $\boldsymbol{v} \in V$ we set

$$\tilde{a}(\boldsymbol{u},\boldsymbol{v}) = \sum_{K\in\mathcal{T}} a_{\tilde{K}}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}), \quad \tilde{\ell}(\boldsymbol{v}) = \sum_{K\in\mathcal{T}} \ell_{\tilde{K}}(\tilde{\boldsymbol{v}}), \tag{34}$$

where it is to be understood that $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$ are the maps to $\tilde{K}$ of the restrictions $\boldsymbol{u}|_K$ and $\boldsymbol{v}|_K$. In particular, in what follows, for any function $\boldsymbol{v}_h \in V^h$ we set

$$\tilde{\boldsymbol{v}}_h := \boldsymbol{v}_h|_K \circ G^{-1}.$$

We are now in a position to define the affine-approximate problem.

Problem $\tilde{\mathrm{S}}^{\mathrm{h}}$. Given $\boldsymbol{b} \in V'$, find $\boldsymbol{w}_h \in V^h$ which satisfies

$$\tilde{a}(\boldsymbol{w}_h,\boldsymbol{v}_h) = \tilde{\ell}(\boldsymbol{v}_h) \tag{35}$$

for all $\boldsymbol{v}_h \in V^h$.

Remark. It is important to note that Problem $\tilde{\mathrm{S}}^{\mathrm{h}}$ is defined not by *transforming* the integrals in (18) onto $\tilde{K}$ but rather by *replacing* the integrals over K with those over $\tilde{K}$. Thus the relevant functions are mapped from K to $\tilde{K}$, after which the appropriate gradients are evaluated on $\tilde{K}$. So, for example, the integrals in the definitions of $a_{\tilde{K}}$ and $\ell_{\tilde{K}}$ contain no Jacobian determinants.

LEMMA 1. *The bilinear form $\tilde{a}$ is V^h-elliptic; that is, there exists a constant $\tilde{\alpha} > 0$, independent of τ, such that*

$$\tilde{a}(\boldsymbol{v}_h, \boldsymbol{v}_h) \geq \tilde{\alpha}(1 + \tau + \tau^2)\|\boldsymbol{v}_h\|_V^2 \quad \text{for all } \boldsymbol{v}_h \in V^h, \tag{36}$$

where τ is defined by (22).

Proof. From the V^h-ellipticity of $a(\cdot,\cdot)$ we have

$$\alpha \|\boldsymbol{v}_h\|_V^2 \leq a(\boldsymbol{v}_h, \boldsymbol{v}_h) = \sum_{K \in \mathcal{T}} a_K(\boldsymbol{v}_h, \boldsymbol{v}_h), \tag{37}$$

where $a_K(\cdot,\cdot)$ is given by

$$a_K(\boldsymbol{u}, \boldsymbol{v}) := \int_K \mathbb{C}[\boldsymbol{\epsilon}(\boldsymbol{u})] : \boldsymbol{\epsilon}(\boldsymbol{v})\, dxdy. \tag{38}$$

Using (31) we have

$$a_K(\boldsymbol{v}_h, \boldsymbol{v}_h) = \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) - \Delta(\tilde{\boldsymbol{v}}_h)] : (\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) - \Delta(\tilde{\boldsymbol{v}}_h))\, j'\, d\tilde{x}d\tilde{y}. \tag{39}$$

We now examine each of the terms on the right-hand side of (39) in turn.

We set

$$(\boldsymbol{E}, \boldsymbol{F})_{\mathbb{C},\tilde{K}} := \int_{\tilde{K}} \tilde{\mathbb{C}}[\boldsymbol{E}] : \boldsymbol{F}\, d\tilde{x}d\tilde{y}$$

for any $\boldsymbol{E}$, $\boldsymbol{F} \in \mathbb{L}^2(\tilde{K})$, and we note that this is an inner product on $\mathbb{L}^2(\tilde{K})$, equivalent to the standard inner product, as a result of the properties (6) and (7) of $\mathbb{C}$. We will also make use of the symmetry property $\mathbb{C}[\boldsymbol{E}] : \boldsymbol{F} = \mathbb{C}[\boldsymbol{F}] : \boldsymbol{E}$, which follows from (5).

Now, setting $\tilde{\boldsymbol{L}}\boldsymbol{R} := \boldsymbol{A}$ in the definition (32) of Δ, consider the expression

$$\begin{aligned}
\text{(I)} &:= \int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta] : \Delta\, j'\, d\tilde{x}d\tilde{y} \\
&\leq \|j'\|_\infty (\Delta, \Delta)_{\mathbb{C},\tilde{K}} \\
&= \tfrac{1}{4}\|j'\|_\infty (\boldsymbol{A} + \boldsymbol{A}^T, \boldsymbol{A} + \boldsymbol{A}^T)_{\mathbb{C},\tilde{K}} \\
&= \tfrac{1}{2}\|j'\|_\infty \left(\|\boldsymbol{A}\|_{\mathbb{C},\tilde{K}}^2 + (\boldsymbol{A}, \boldsymbol{A}^T)_{\mathbb{C},\tilde{K}} \right) \\
&\leq \|j'\|_\infty \|\boldsymbol{A}\|_{\mathbb{C},\tilde{K}}^2. \qquad (40)
\end{aligned}$$

Next, from the definition of $\boldsymbol{A}$ we have

$$\|\boldsymbol{A}\|_{\mathbb{C},\tilde{K}} \leq \|\tilde{\boldsymbol{L}}\|_{\mathbb{C},\tilde{K}} \|\boldsymbol{R}\|_{\mathbb{C},\tilde{K}}. \tag{41}$$

Furthermore, we have, using the definition of the norm $\|\cdot\|_{\mathbb{C},\tilde{K}}$, (8), and (27),

$$\begin{aligned}\|\boldsymbol{R}\|^2_{\mathbb{C},\tilde{K}} &\le c_\infty \mu_K \max_{\tilde{\boldsymbol{x}}\in\tilde{K}} |\boldsymbol{R}|^2(\tilde{\boldsymbol{x}}) \\ &\le \frac{2c_\infty \mu_K \tau_K^2}{(1-\sqrt{2}\,\tau_K)^2}, \end{aligned} \tag{42}$$

in which μ_K is the area of K (and of $\tilde{K}$).

We also have

$$\begin{aligned}\|\tilde{\boldsymbol{L}}\|^2_{\mathbb{C},\tilde{K}} &= \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{L}}] : \tilde{\boldsymbol{L}}\; d\tilde{x}d\tilde{y} \\ &= \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}] : \tilde{\boldsymbol{\epsilon}}\; d\tilde{x}d\tilde{y} \\ &= a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h). \end{aligned} \tag{43}$$

From (41)–(43), it follows that

$$\text{(I)} \le c_K^2\, a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h), \tag{44}$$

where the constant c_K is defined by

$$c_K = \tau_K(1-\sqrt{2}\,\tau_K)^{-1}\sqrt{2c_\infty\mu_K(1+\sqrt{2}\,\tau_K)}. \tag{45}$$

Next, consider the expression

$$\begin{aligned}\text{(II)} &:= \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) j'\; d\tilde{x}d\tilde{y} \\ &\le (1+\sqrt{2}\,\tau_K)\int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}] : \tilde{\boldsymbol{\epsilon}}\; d\tilde{x}d\tilde{y} \\ &= (1+\sqrt{2}\,\tau_K) a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h), \end{aligned} \tag{46}$$

where we have used the positivity of the integrand and (23).

Finally, consider the term

$$\begin{aligned}\text{(III)} &:= -\int_{\tilde{K}} \left[\tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{v}}_h) + \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)\right] j'\; d\tilde{x}d\tilde{y} \\ &= -2\int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{v}}_h) j'\; d\tilde{x}d\tilde{y} \\ &= -2(j'\tilde{\boldsymbol{\epsilon}}, \Delta)_{\mathbb{C},\tilde{K}} \\ &\le 2\|j'\|_\infty \left|(\tilde{\boldsymbol{\epsilon}}, \Delta)_{\mathbb{C},\tilde{K}}\right| \\ &= 2\|j'\|_\infty \left|(\tilde{\boldsymbol{\epsilon}}, \boldsymbol{A})_{\mathbb{C},\tilde{K}}\right| \\ &\le 2\|j'\|_\infty \|\tilde{\boldsymbol{\epsilon}}\|_{\mathbb{C},\tilde{K}} \|\boldsymbol{A}\|_{\mathbb{C},\tilde{K}}, \end{aligned}$$

where we have used the representation $\Delta = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T)$ and the Cauchy–Schwarz inequality. Now we know that $\|\tilde{\boldsymbol{\epsilon}}\|_{\mathbb{C},\tilde{K}} = \sqrt{a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h)}$ and, from (41)–(43),

$$\|\boldsymbol{A}\|^2_{\mathbb{C},\tilde{K}} \le c_K^2(1+\sqrt{2}\,\tau_K)^{-1} a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h).$$

So we have

$$\text{(III)} \le 2c_K\sqrt{1+\sqrt{2}\,\tau_K}\,a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h,\tilde{\boldsymbol{v}}_h). \tag{47}$$

Expanding (39) and substituting the estimates for (I), (II), and (III), we have

$$\begin{aligned}\alpha\|\boldsymbol{v}_h\|_V^2 \le a(\boldsymbol{v}_h,\boldsymbol{v}_h) &= \sum_{K\in\mathcal{T}} a_K(\boldsymbol{v}_h,\boldsymbol{v}_h)\\ &\le \sum_{K\in\mathcal{T}} \text{(I)}+\text{(II)}+\text{(III)}\\ &\le \sum_{K\in\mathcal{T}}\left[c_K+\sqrt{1+\sqrt{2}\,\tau_K}\right]^2 a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h,\tilde{\boldsymbol{v}}_h)\\ &\le \gamma\,\tilde{a}(\boldsymbol{v}_h,\boldsymbol{v}_h),\end{aligned}$$

where the constant γ is given by

$$\gamma = [1+\sqrt{2c_\infty}\,(1-\sqrt{2}\,\tau_K)^{-1}h\tau]^2(1+\sqrt{2}\,\tau).$$

It follows that $\tilde{a}(\cdot,\cdot)$ is V^h-elliptic. □

Since $\tilde{a}$ and $\tilde{\ell}$ are clearly continuous, it follows therefore that Problem $\tilde{\mathrm{S}}^h$ has a unique solution $\boldsymbol{w}_h$.

Remark. The consistency and convergence of the solution to Problem S^{h} depends in a fundamental way on the following lemma, which is analogous to the first Strang lemma (see [7, Theorem 4.4.1]) associated with errors induced by numerical quadrature in finite element approximations.

LEMMA 2. *There exists a positive constant C, independent of h, such that*

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{w}_h\|_V \le C\Bigg(&\inf_{\boldsymbol{v}_h\in V^h}\left\{\|\boldsymbol{u}-\boldsymbol{v}_h\|_V+\sup_{\boldsymbol{z}_h\in V^h}\frac{|a(\boldsymbol{v}_h,\boldsymbol{z}_h)-\tilde{a}(\boldsymbol{v}_h,\boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|_V}\right\}\\ &+\sup_{\boldsymbol{z}_h\in V^h}\frac{|\ell(\boldsymbol{z}_h)-\tilde{\ell}(\boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|_V}\Bigg).\end{aligned} \tag{48}$$

Proof. The proof follows that of the Strang lemma very closely, and we therefore merely sketch the details. Using the V^h-ellipticity of $\tilde{a}$, we have

$$\begin{aligned}\tilde{\alpha}\|\boldsymbol{w}_h-\boldsymbol{v}_h\|_V^2 &\le \tilde{a}(\boldsymbol{w}_h-\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)\\ &= a(\boldsymbol{u}-\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)+[a(\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)-\tilde{a}(\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)]\\ &\quad+[\tilde{\ell}(\boldsymbol{w}_h-\boldsymbol{v}_h)-\ell(\boldsymbol{w}_h-\boldsymbol{v}_h)].\end{aligned} \tag{49}$$

Here we have used (11) and (35). Now

$$\frac{a(\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)-\tilde{a}(\boldsymbol{v}_h,\boldsymbol{w}_h-\boldsymbol{v}_h)}{\|\boldsymbol{w}_h-\boldsymbol{v}_h\|_V} \le \sup_{\boldsymbol{z}_h\in V^h}\frac{|a(\boldsymbol{v}_h,\boldsymbol{z}_h)-\tilde{a}(\boldsymbol{v}_h,\boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|_V}, \tag{50}$$

and a similar inequality exists for $\ell-\tilde{\ell}$. The triangle inequality gives

$$\|\boldsymbol{u}-\boldsymbol{w}_h\|_V \le \|\boldsymbol{u}-\boldsymbol{v}_h\|_V+\|\boldsymbol{w}_h-\boldsymbol{v}_h\|_V. \tag{51}$$

We divide throughout in (49) by $\|\boldsymbol{w}_h-\boldsymbol{v}_h\|_V$ and make use of the continuity of a in the first term on the right-hand side of (49); next, we use (50) and the corresponding expression for $|\ell-\tilde{\ell}|$ and take the supremum of the terms in square brackets in

(49). Finally, (48) is obtained by using (51) and by taking the infimum over all $\boldsymbol{v}_h \in V^h$. ∎

Next, we address the task of estimating the expression

$$\sup_{\boldsymbol{z}_h \in V^h} \frac{|a(\boldsymbol{v}_h, \boldsymbol{z}_h) - \tilde{a}(\boldsymbol{v}_h, \boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|_V}.$$

From (33) and (39) we have

$$\begin{aligned}
&a_K(\boldsymbol{v}_h, \boldsymbol{z}_h) - a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{z}}_h) \\
&\quad = \int_K \mathbb{C}[\boldsymbol{\epsilon}(\boldsymbol{v}_h)] : \boldsymbol{\epsilon}(\boldsymbol{z}_h)\, dxdy - \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)\, d\tilde{x}d\tilde{y} \\
&\quad = \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h) j'\, d\tilde{x}d\tilde{y} - \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{z}}_h)\, j'\, d\tilde{x}d\tilde{y} \\
&\qquad - \int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)\, j'\, d\tilde{x}d\tilde{y} + \int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{z}}_h)\, j'\, d\tilde{x}d\tilde{y} \\
&\qquad - \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)\, d\tilde{x}d\tilde{y} \\
&\quad = \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)(j'-1)\, d\tilde{x}d\tilde{y} - \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{z}}_h) j'\, d\tilde{x}d\tilde{y} \\
&\qquad - \int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : \tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h) j'\, d\tilde{x}d\tilde{y} + \int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : \Delta(\tilde{\boldsymbol{z}}_h) j'\, d\tilde{x}d\tilde{y} \\
&\quad \le \sqrt{2}\,\tau_K\, a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{z}}_h) + \sqrt{a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{v}}_h) a_{\tilde{K}}(\tilde{\boldsymbol{z}}_h, \tilde{\boldsymbol{z}}_h)} \left[c_K^2 + 2c_K \sqrt{1 + \sqrt{2}\,\tau_K} \right]. \qquad (52)
\end{aligned}$$

Here we have used the estimates leading to (44) and (47).

Now

$$a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{z}}_h) \le M_{\tilde{K}}\, \|\tilde{\boldsymbol{v}}_h\|_{H^1(\tilde{K})} \|\tilde{\boldsymbol{z}}_h\|_{H^1(\tilde{K})}$$

for some positive constant $M_{\tilde{K}}$, independent of h_K and τ_K, and, using (24), (28), and (30), we have

$$\begin{aligned}
\|\tilde{\boldsymbol{v}}_h\|^2_{H^1(\tilde{K})} &= \int_{\tilde{K}} \left(|\tilde{\boldsymbol{v}}_h|^2 + |\tilde{\nabla} \tilde{\boldsymbol{v}}_h|^2 \right) d\tilde{x}d\tilde{y} \\
&= \int_K [|\boldsymbol{v}_h|^2 + |(\nabla \boldsymbol{v}_h)\boldsymbol{J}'|^2](j')^{-1}\, dxdy \\
&\le C_K \|\boldsymbol{v}_h\|^2_{H^1(K)},
\end{aligned}$$

where $C_K = 2(1+\tau_K)^2(1-\sqrt{2}\,\tau_K)^{-1}$.

Hence

$$\begin{aligned}
&\frac{|a(\boldsymbol{v}_h, \boldsymbol{z}_h) - \tilde{a}(\boldsymbol{v}_h, \boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|_V} \\
&\le \frac{\sum_{K \in \mathcal{T}} M_{\tilde{K}} C_K \left[\left(c_K + \sqrt{1 + \sqrt{2}\,\tau_K} \right)^2 - 1 \right] \|\boldsymbol{v}_h\|_{H^1(K)} \|\boldsymbol{z}_h\|_{H^1(K)}}{\|\boldsymbol{z}_h\|_V} \\
&\le C\tau \frac{\sum_{K \in \mathcal{T}} \|\boldsymbol{v}_h\|_{H^1(K)} \|\boldsymbol{z}_h\|_{H^1(K)}}{\|\boldsymbol{z}_h\|_V} \\
&\le C\tau \|\boldsymbol{v}_h\|_V, \qquad (53)
\end{aligned}$$

where the constant C is independent of h and τ.

Choosing $\boldsymbol{v}_h = \Pi_h \boldsymbol{u}$, where Π_h is the interpolation operator onto V^h, and noting that $\|\Pi_h \boldsymbol{u}\|_V \leq \|\boldsymbol{u}\|_V + \|\boldsymbol{u} - \Pi_h \boldsymbol{u}\|_V \leq ch|\boldsymbol{u}|_{H^2} + \|\boldsymbol{u}\|_V$, we finally obtain

$$\inf_{\boldsymbol{v}_h \in V^h} \sup_{\boldsymbol{z}_h \in V^h} \frac{|a(\boldsymbol{v}_h, \boldsymbol{z}_h) - \tilde{a}(\boldsymbol{v}_h, \boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|} \leq C\tau,$$

where the constant C depends on the geometry and on the solution $\boldsymbol{u}$. In exactly the same way we can derive the estimate

$$\inf_{\boldsymbol{v}_h \in V^h} \sup_{\boldsymbol{z}_h \in V^h} \frac{|\ell(\boldsymbol{z}_h) - \tilde{\ell}(\boldsymbol{z}_h)|}{\|\boldsymbol{z}_h\|} \leq C\tau,$$

where again C depends on the geometry and on $\boldsymbol{u}$. We therefore have the following result.

THEOREM 3. *Let $\mathcal{T}$ be an h-regular finite element mesh of quadrilaterals, with the maximum distortion of quadrilaterals being bounded in the sense that $\tau \leq ch$ for some constant c, independent of h, as $h \to 0$. Then problem $\tilde{\mathrm{S}}^h$ has a unique solution $\boldsymbol{w}_h$ which satisfies*

$$\|\boldsymbol{u} - \boldsymbol{w}_h\|_V \leq Ch,$$

the constant C depending on the geometry and the solution $\boldsymbol{u}$ to the continuous problem, but not on h.

5. The enhanced strain problem. In the context of the finite element method, the enhanced strain method refers to an approach proposed by Simo and Rifai [14], in which the discrete strain $\boldsymbol{\epsilon}_h$ takes the form

$$\boldsymbol{\epsilon}_h = \boldsymbol{\epsilon}(\boldsymbol{u}_h) + \boldsymbol{\eta}_h, \tag{54}$$

the first term on the right-hand side being evaluated as in (2), while the second term on the right-hand side is the enhanced strain, which is required to have the property $\boldsymbol{\eta}_h \to \mathbf{0}$ as $h \to 0$.

In order to formulate the problem in weak form it is necessary to add to the spaces already defined the space Γ^h of enhanced strains, which is defined by

$$\Gamma^h := \left\{ \boldsymbol{\gamma} = (\gamma_{ij}) : \ \gamma_{ij} \in L^2(\Omega), \ \gamma_{ji} = \gamma_{ij}, \ \int_K \mathbb{C}\boldsymbol{\gamma}|_K \, dxdy = \mathbf{0} \text{ for all } K \in \mathcal{T} \right\}. \tag{55}$$

In practice Γ^h will comprise functions of the form $\boldsymbol{\gamma} = j^{-1}\hat{\boldsymbol{\gamma}}$ on each element, in which the components of $\hat{\boldsymbol{\gamma}}$ are simple polynomials defined on the reference element $\hat{K}$. A consequence of this definition is that $\int_{\hat{K}} \mathbb{C}\hat{\boldsymbol{\gamma}} \, d\xi d\eta = \mathbf{0}$ [3]. Concrete examples of bases for Γ^h, together with applications, may be found in [14, 1].

We set $\boldsymbol{\phi}_h = (\boldsymbol{u}_h, \boldsymbol{\eta}_h)$ and $\boldsymbol{\psi}_h = (\boldsymbol{v}_h, \boldsymbol{\gamma}_h)$ for $\boldsymbol{u}_h$, $\boldsymbol{v}_h \in V^h$ and $\boldsymbol{\eta}_h$, $\boldsymbol{\gamma}_h \in \Gamma^h$. Also, we define the product space

$$\Psi^h := V^h \times \Gamma^h,$$

which is a Hilbert space with the natural norm

$$\|\boldsymbol{\psi}_h\|_\Psi := \left(\|\boldsymbol{v}_h\|_V^2 + \|\boldsymbol{\gamma}_h\|_{L^2}^2 \right)^{1/2}.$$

The bilinear form $A : \Psi^h \times \Psi^h \longrightarrow \mathbb{R}$ is defined by

$$A(\boldsymbol{\phi}_h, \boldsymbol{\psi}_h) = \int_\Omega \mathbb{C}(\boldsymbol{\epsilon}(\boldsymbol{u}_h) + \boldsymbol{\eta}_h) : (\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h)\, dxdy, \tag{56}$$

and we recall the definition (10) of the linear functional ℓ.

The weak formulation of the problem then takes the following form [11, 14].

Problem E^{h}. Find $(\boldsymbol{u}_h, \boldsymbol{\eta}_h) \in V^h \times \Gamma^h$ such that

$$A(\boldsymbol{\phi}_h, \boldsymbol{\psi}_h) = \ell(\boldsymbol{v}_h) \quad \text{for all } \boldsymbol{\psi}_h \in \Psi^h. \tag{57}$$

We have the following result, proved in [11] for affine-equivalent meshes (see also [5]) and in [3] for isoparametric meshes and stated here for the special case in which V^h is chosen as in (15).

THEOREM 4. *Let $\mathcal{T}$ be a regular mesh of quadrilaterals on a bounded polygonal domain $\Omega \in \mathbb{R}^2$. Let the space V^h be defined by* (15) *and the space Γ^h by* (55). *Assume, in addition, that*

(a) $\boldsymbol{\epsilon}(V^h) \cap \Gamma^h = \{\boldsymbol{0}\}$,

(b) *there exists a constant c_1 with $0 < c_1 < 1$ such that, for any $\boldsymbol{\gamma}_h \in \Gamma^h$, $\|P\boldsymbol{\gamma}_h\|_\Gamma \le c_1\|\boldsymbol{\gamma}_h\|$, where P is the L^2-orthogonal projection onto $\boldsymbol{\epsilon}(V^h)$.*

Then there exists a unique solution to Problem E^{h}. *Furthermore, if $\boldsymbol{u} \in [H^2(\Omega)]^2$, then there exists a constant $C > 0$, independent of h, such that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V + \|\boldsymbol{\eta}_h\|_\Gamma \le Ch|\boldsymbol{u}|_{H^2}.$$

We now consider the affine-approximate problem analogous to Problem $\tilde{\mathrm{S}}^h$.

Define

$$A_{\tilde{K}}(\tilde{\boldsymbol{\chi}}_h, \tilde{\boldsymbol{\psi}}_h) = \int_{\tilde{K}} \tilde{\mathbb{C}}(\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{w}}_h) + \tilde{\boldsymbol{\beta}}_h) : (\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) + \tilde{\boldsymbol{\gamma}}_h)\, d\tilde{x}d\tilde{y} \tag{58}$$

and

$$\ell_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h) = \int_{\tilde{K}} \boldsymbol{b} \cdot \tilde{\boldsymbol{v}}_h\, d\tilde{x}d\tilde{y}, \tag{59}$$

where $\boldsymbol{\chi}_h = (\boldsymbol{w}_h, \boldsymbol{\beta}_h)$ and superposed tildes have the same interpretation as previously. Set

$$\tilde{A}(\boldsymbol{\chi}_h, \boldsymbol{\psi}_h) = \sum_{K \in \mathcal{T}} A_{\tilde{K}}(\tilde{\boldsymbol{\chi}}_h, \tilde{\boldsymbol{\psi}}_h) \tag{60}$$

and

$$\tilde{\ell}(\boldsymbol{\psi}_h) = \sum_{K \in \mathcal{T}} \ell_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h). \tag{61}$$

Problem $\tilde{\mathrm{E}}^{\mathrm{h}}$. Given $\boldsymbol{b} \in V'$, find $\boldsymbol{\chi}_h := (\boldsymbol{w}_h, \boldsymbol{\beta}_h) \in V^h \times \Psi^h$ which satisfy

$$\tilde{A}(\boldsymbol{\chi}_h, \boldsymbol{\psi}_h) = \tilde{\ell}(\boldsymbol{\psi}_h) \tag{62}$$

for all $\boldsymbol{\psi}_h \in \Psi^h$.

We need to show that $\tilde{A}$ is Ψ^h-elliptic. We proceed as in the case of Lemma 1, beginning with the observation that $A(\cdot,\cdot)$ is Ψ^h-elliptic (see [11]), from which it follows that

$$\begin{aligned} \alpha_A \|\boldsymbol{\psi}_h\|_\Psi^2 &\le A(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h) \\ &= \sum_{K\in\mathcal{T}} A_K(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h). \end{aligned} \tag{63}$$

Now we have, from the positivity of the integrand of A_K and the symmetry of $\mathbb{C}$,

$$\begin{aligned} A_K(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h) &= \int_K \mathbb{C}(\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h) : (\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h)\ dxdy \\ &= \int_{\tilde{K}} \mathbb{C}(\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h) : (\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h)\ j'\, d\tilde{x}d\tilde{y} \\ &\le \|j'\|_\infty \int_{\tilde{K}} \mathbb{C}(\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h) : (\boldsymbol{\epsilon}(\boldsymbol{v}_h) + \boldsymbol{\gamma}_h)\ d\tilde{x}d\tilde{y} \\ &= \|j'\|_\infty \int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) - \Delta(\tilde{\boldsymbol{v}}_h) + \tilde{\boldsymbol{\gamma}}_h) : (\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) - \Delta(\tilde{\boldsymbol{v}}_h) + \tilde{\boldsymbol{\gamma}}_h)\ d\tilde{x}d\tilde{y} \\ &= \|j'\|_\infty \Bigg[A_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h, \tilde{\boldsymbol{\psi}}_h) \underbrace{-2\int_{\tilde{K}} \tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)] : [\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) - \tfrac{1}{2}\Delta(\tilde{\boldsymbol{v}}_h) + \tilde{\boldsymbol{\gamma}}_h]\ d\tilde{x}d\tilde{y}}_{\mathcal{F}} \\ &\qquad + 2\int_{\tilde{K}} \tilde{\mathbb{C}}[\tilde{\boldsymbol{\gamma}}_h] : (\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h) + \tfrac{1}{2}\tilde{\boldsymbol{\gamma}}_h)\ d\tilde{x}d\tilde{y} \Bigg] \\ &\le \|j'\|_\infty \Big[A_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h, \tilde{\boldsymbol{\psi}}_h) + 2(\tilde{\boldsymbol{\epsilon}}, \tilde{\boldsymbol{\gamma}})_{\mathbb{C},\tilde{K}} + (\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\gamma}})_{\mathbb{C},\tilde{K}} + \mathcal{F}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{\gamma}}_h, \tau_K) \Big]. \end{aligned}$$

Here the term $\mathcal{F}$ is of the form $\mathcal{F} = c_K \cdot [\text{terms depending on } (\tilde{\boldsymbol{v}}_h,\ \tilde{\boldsymbol{\gamma}}_h,\ \tau_K)]$, as can be deduced by a series of manipulations similar to those carried out in (44)–(46). From the definition of $A_{\tilde{K}}$ it follows therefore that

$$A_K(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h) \le (1 + \sqrt{2}\,\tau_K)[2A_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h, \tilde{\boldsymbol{\psi}}_h) + \mathcal{F}(\tilde{\boldsymbol{v}}_h, \tilde{\boldsymbol{\gamma}}_h, \tau_K)].$$

We therefore have the following result.

LEMMA 5. *The bilinear form $\tilde{A}$ is Ψ^h-elliptic for sufficiently small τ. Furthermore, Problem $\tilde{\mathrm{E}}^h$ has a unique solution $\boldsymbol{\chi}_h = (\boldsymbol{w}_h, \boldsymbol{\beta}_h)$ in Ψ^h.*

Next, we have the following counterpart to Lemma 2.

LEMMA 6. *Set $\boldsymbol{\phi} = (\boldsymbol{u}, \boldsymbol{0})$, where $\boldsymbol{u}$ is the solution to Problem* S*, and denote the solution to Problem $\tilde{\mathrm{E}}^{\mathrm{h}}$ by $\boldsymbol{\chi}_h = (\boldsymbol{w}_h, \boldsymbol{\beta}_h)$. Then there exists a constant C, independent of h, such that*

$$\|\boldsymbol{\phi} - \boldsymbol{\chi}_h\|_\Psi \le C \left[\inf_{\boldsymbol{\psi}_h \in \Psi^h} \left\{ \|\boldsymbol{\phi} - \boldsymbol{\psi}_h\|_\Psi + \sup_{\boldsymbol{\omega}_h = (\boldsymbol{z}_h, \boldsymbol{\rho}_h) \in \Psi^h} \frac{\diamond}{\|\boldsymbol{\omega}_h\|_\Psi} \right\} \right], \tag{64}$$

where $\diamond$ is given by

$$\diamond = |A(\boldsymbol{\psi}_h, \boldsymbol{\omega}_h) - \tilde{A}(\boldsymbol{\psi}_h, \boldsymbol{\omega}_h)| + |\ell(\boldsymbol{\psi}_h) - \tilde{\ell}(\boldsymbol{\psi}_h)| + \left| \int_\Omega \mathbb{C}\boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\gamma}_h\ dxdy \right|. \tag{65}$$

Proof. From the Ψ^h-ellipticity of $\tilde{A}$, we have

$$\begin{aligned}\tilde{\alpha}\|\boldsymbol{\chi}_h-\boldsymbol{\psi}_h\|_\Psi^2 &\le \tilde{A}(\boldsymbol{\chi}_h-\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)\\ &= A(\boldsymbol{\phi}-\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)+[A(\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)-\tilde{A}(\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)]\\ &\quad - A(\boldsymbol{\phi},\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)+\tilde{A}(\boldsymbol{\chi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)\\ &= A(\boldsymbol{\phi}-\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)+[A(\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)-\tilde{A}(\boldsymbol{\psi}_h,\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)]\\ &\quad +[\tilde{\ell}(\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)-\ell(\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)]+[\ell(\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)-A(\boldsymbol{\phi},\boldsymbol{\chi}_h-\boldsymbol{\psi}_h)].\end{aligned}$$

The rest of the proof proceeds in much the same way as the proof of Lemma 2. ∎

Finally, the expression $\diamond$ in (64) must be estimated. We have, from (58) and (64),

$$\begin{aligned}&A_K(\boldsymbol{\psi}_h,\boldsymbol{\omega}_h)-A_{\tilde{K}}(\tilde{\boldsymbol{\psi}}_h,\tilde{\boldsymbol{\omega}}_h)\\ &= a_K(\boldsymbol{v}_h,\boldsymbol{z}_h)-a_{\tilde{K}}(\tilde{\boldsymbol{v}}_h,\tilde{\boldsymbol{z}}_h)\quad\text{(i)}\\ &\quad+\int_K(\mathbb{C}[\boldsymbol{\epsilon}(\boldsymbol{v}_h)]:\boldsymbol{\rho}_h+\mathbb{C}[\boldsymbol{\epsilon}(\boldsymbol{z}_h)]:\boldsymbol{\gamma}_h)\,dxdy\\ &\quad-\int_{\tilde{K}}(\tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)]:\tilde{\boldsymbol{\rho}}_h-\tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)]:\tilde{\boldsymbol{\gamma}}_h)\,d\tilde{x}d\tilde{y}\quad\text{(ii)}\\ &\quad+\int_K\mathbb{C}[\boldsymbol{\gamma}_h]:\boldsymbol{\rho}_h\,dxdy-\int_{\tilde{K}}\tilde{\mathbb{C}}[\tilde{\boldsymbol{\gamma}}_h]:\tilde{\boldsymbol{\rho}}_h\,d\tilde{x}d\tilde{y}.\quad\text{(iii)}\end{aligned}\tag{66}$$

We now examine the expressions (i)–(iii) in (66) in turn. First, we see that (i) is estimated in (52). Next, we have

$$\begin{aligned}\text{(ii)} &\le \int_{\tilde{K}}(j'-1)\tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)]:\tilde{\boldsymbol{\rho}}_h\,d\tilde{x}d\tilde{y}+\int_{\tilde{K}}(j'-1)\tilde{\mathbb{C}}[\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)]:\tilde{\boldsymbol{\gamma}}_h\,d\tilde{x}d\tilde{y}\\ &\quad-\int_{\tilde{K}}\tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{v}}_h)]:\tilde{\boldsymbol{\rho}}_h\,\tilde{j}\,d\tilde{x}d\tilde{y}-\int_{\tilde{K}}\tilde{\mathbb{C}}[\Delta(\tilde{\boldsymbol{z}}_h)]:\tilde{\boldsymbol{\gamma}}_h\,\tilde{j}\,d\tilde{x}d\tilde{y}\\ &\le \left[\sqrt{2}\tau_K+\sqrt{1+\sqrt{2}\tau_K}\,c_K\right]\left\{\|\tilde{\boldsymbol{\rho}}_h\|_{\mathbb{C},\tilde{K}}\|\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{v}}_h)\|_{\mathbb{C},\tilde{K}}+\|\tilde{\boldsymbol{\gamma}}_h\|_{\mathbb{C},\tilde{K}}\|\tilde{\boldsymbol{\epsilon}}(\tilde{\boldsymbol{z}}_h)\|_{\mathbb{C},\tilde{K}}\right\}.\end{aligned}\tag{67}$$

Here we have used (23) and the manipulations leading to (44). Finally,

$$\text{(iii)}\le\sqrt{2}\tau_K\|\tilde{\boldsymbol{\rho}}_h\|_{\mathbb{C},\tilde{K}}\|\tilde{\boldsymbol{\gamma}}_h\|_{\mathbb{C},\tilde{K}}.\tag{68}$$

It is not difficult to see, from the arguments leading to (53) and from (67) and (68), that

$$\sup_{\boldsymbol{\omega}_h\in\Psi^h}\frac{A(\boldsymbol{\psi}_h,\boldsymbol{\omega}_h)-\tilde{A}(\tilde{\boldsymbol{\psi}}_h,\tilde{\boldsymbol{\omega}}_h)}{\|\boldsymbol{\omega}_h\|_\Psi}\le C\tau,\tag{69}$$

in which the constant C depends on the geometry and on the solution $\boldsymbol{u}$ but not on h nor on τ. In the same way, one may derive an estimate of the form (69) for the second term in the definition (65) of $\diamond$.

Finally, the last term on the right-hand side of (65) is shown in [11] to be bounded, up to a constant, by $h|\boldsymbol{u}|_1$.

THEOREM 7. *Let $\mathcal{T}$ be an h-regular finite element mesh of quadrilaterals, with the maximum distortion of quadrilaterals being bounded in the sense that $\tau\le ch$ for*

some constant c, independent of h, as $h \to 0$. Let $\boldsymbol{\phi} = (\boldsymbol{u}, \mathbf{0}) \in \Psi$, where $\boldsymbol{u}$ is the solution to Problem S. *Then Problem* $\tilde{\mathrm{E}}^{\mathrm{h}}$ *has a unique solution $\boldsymbol{\chi}_h$ which satisfies*

$$\|\boldsymbol{\phi} - \boldsymbol{\chi}_h\|_\Psi \leq Ch,$$

the constant C depending on the geometry, on the material tensor $\mathbb{C}$, and on $\boldsymbol{u}$, but not on h.

Remark. The analysis presented here has been carried out for the case of compressible materials, for which the components of $\mathbb{C}$ are bounded. A modified approach, such as that presented in [11, section 5] or in [6], is required for the limiting cases of incompressibility or near incompressibility. Such an analysis would combine the approaches presented here and in those works.

REFERENCES

[1] U. Andelfinger and E. Ramm, *EAS-elements for two-dimensional, three-dimensional, plate and shell structures and their equivalence to HR-elements*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 1311–1337.

[2] K. Arunakirinathar and B. D. Reddy, *Some geometrical results and estimates for quadrilateral finite elements*, Comput. Methods Appl. Mech. Engrg., 122 (1995), pp. 307–314.

[3] K. Arunakirinathar and B. D. Reddy, *Further results for enhanced strain methods with isoparametric elements*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 127–143.

[4] T. Belytschko and W. Bachrach, *Efficient implementation of quadrilaterals with high coarse-mesh accuracy*, Comput. Methods Appl. Mech. Engrg., 54 (1986), pp. 279–301.

[5] D. Braess, *Enhanced assumed strain elements and locking in membrane problems*, Comput. Methods Appl. Mech. Engrg., 165 (1998), pp. 155–174.

[6] D. Braess, C. Carstensen, and B. D. Reddy, *Uniform convergence and a posteriori estimators for the enhanced strain finite element method*, submitted.

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[8] D. P. Flanagan and T. Belytschko, *A uniform strain hexahedron and quadrilateral with orthogonal hourglass control*, Internat. J. Numer. Methods Engrg., 17 (1981), pp. 679–706.

[9] M. Küssner and B. D. Reddy, *The equivalent parallelogram and parallelipiped, and their application to stabilized finite elements in two and three dimensions*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 1967–1983.

[10] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[11] B. D. Reddy and J. C. Simo, *Stability and convergence of a class of enhanced strain methods*, SIAM J. Numer. Anal., 32 (1995), pp. 1705–1728.

[12] S. Reese, M. Küssner, and B. D. Reddy, *A new stabilization technique for finite elements in non-linear elasticity*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 1617–1652.

[13] S. Reese, P. Wriggers, and B. D. Reddy, *A new locking-free brick element technique for large deformation problems in elasticity*, Comput. & Structures, 75 (2000), pp. 291–304.

[14] J. C. Simo and S. Rifai, *A class of mixed assumed strain methods and the method of incompatible modes*, Internat. J. Numer. Methods Engrg., 29 (1990), pp. 1595–1638.

[15] J. C. Simo, F. Armero, and R. L.Taylor, *Improved version of assumed enhanced strain tri-linear elements for 3D finite deformation problems*, Comput. Methods Appl. Mech. Engrg., 110 (1993), pp. 359–366.

[16] R. L. Taylor, P. J. Beresford, and E. L. Wilson, *A non-conforming element for stress analysis*, Internat. J. Numer. Methods Engrg., 10 (1976), pp. 1211–1220.

[17] E. L. Wilson, R. L. Taylor, W. P. Doherty, and J. Ghaboussi, *Incompatible displacement models*, in Numerical and Computer Models in Structural Mechanics, Academic Press, New York, 1973, pp. 43–57.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 198–219

ERROR ANALYSIS OF A FINITE ELEMENT–INTEGRAL EQUATION SCHEME FOR APPROXIMATING THE TIME-HARMONIC MAXWELL SYSTEM*

G. C. HSIAO†, P. B. MONK†, AND N. NIGAM‡

Abstract. In 1996 Hazard and Lenoir suggested a variational formulation of Maxwell's equations using an overlapping integral equation and volume representation of the solution [*SIAM J. Math. Anal.,* 27 (1996), pp. 1597–1630]. They suggested a numerical scheme based on this approach, but no error analysis was provided. In this paper, we provide a convergence analysis of an edge finite element scheme for the method. The analysis uses the theory of collectively compact operators. Its novelty is that a perturbation argument is needed to obtain error estimates for the solution of the discrete problem that is best suited for implementation.

Key words. Maxwell system, edge finite element, integral representation, error estimate

AMS subject classifications. 65N15, 65N30, 78A45

PII. S003614290038131X

1. Introduction.

A key feature of scattering problems is that they are typically posed as exterior boundary value problems. When using finite element methods to compute approximate solutions to these problems, the truncation of the computational domain needs to be done carefully. The truncated problem should be chosen to provide a convenient and accurate approximation of the true problem. In [25], Hazard and Lenoir proposed a new variational approach to the time-harmonic scattering problem for Maxwell's equations that can be used as the basis of a finite element method. This was extended to layered media in [19]. Hazard and Lenoir suggested the use of standard continuous finite elements, which are known to require special care if the scatterer has corners [4, 17, 21]. In this paper, we propose the use of edge elements [31, 32] to approximate the problem. A direct application of this approach leads to unwieldy matrices. Thus we apply flux-recovery procedures [35, 5, 6, 7] in the discretization of the Hazard–Lenoir method resulting in a fully discrete problem that is better suited to implementation. We then provide the first error analysis of the discrete Hazard–Lenoir scheme for Maxwell's equations (see [18] for convergence studies of the method applied to the time-harmonic Helmholtz problem).

For simplicity we will not describe the general scattering problem discussed in [19]. Instead, we shall confine ourselves to time-harmonic scattering from a bounded perfect conductor. The finite element scheme applies in the more general case, but some estimates in the analysis still need to be performed. The plan for the paper is as follows. In the next subsection we describe the continuous problem and introduce a

*Received by the editors November 17, 2000; accepted for publication (in revised form) December 18, 2001; published electronically May 1, 2002. The research of the first and second authors was sponsored by the Air Force Office of Scientific Research, Air Force Materials Command, USAF, under grant number F49620-96-1-0039. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sinum/40-1/38131.html

†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (hsiao@math.udel.edu, monk@math.udel.edu).

‡IMA, University of Minnesota, Minneapolis, MN 55455 (nigam@ima.umn.edu).

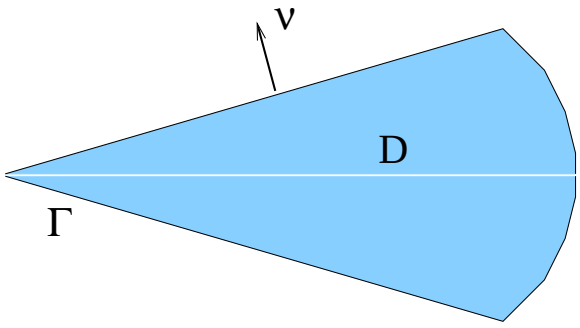


FIG. 1.1. *The perfectly conducting scatterer, D.*

truncated boundary value problem based on an integral representation of the scattered field. In section 2 we describe the finite elements to be used, and in section 3 we provide an error analysis of the method without discretizing the integral operator. In section 4 we provide an analysis of the fully discrete scheme and discuss briefly the solution of the discrete problem, and we draw some conclusions in section 5.

1.1. Problem description. We consider a perfectly conducting scatterer, which occupies a bounded, Lipschitz, polyhedral region D in $\mathbb{R}^3$. We assume that the boundary of the scatterer, Γ, is connected and we denote by $\boldsymbol{\nu}$ the unit outward normal (see Figure 1.1). For simplicity we shall also assume that both D and $\mathbb{R}^3 \setminus \bar{D}$ are simply connected. We wish to approximate the total electric field $\mathbf{E} = \mathbf{E}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^3 \setminus D$, where this field satisfies the time-harmonic Maxwell's equation,

$$\nabla \times \nabla \times \mathbf{E} - k^2 \mathbf{E} = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{D}, \tag{1.1}$$

and the boundary condition appropriate for a perfect conductor,

$$\boldsymbol{\nu} \times \mathbf{E} = 0 \quad \text{on } \Gamma. \tag{1.2}$$

The real parameter $k > 0$ is called the wave number of the time-harmonic field. The total field $\mathbf{E}$ is given by

$$\mathbf{E} = \mathbf{E}^i + \mathbf{E}^s \quad \text{in } \mathbb{R}^3 \setminus D, \tag{1.3}$$

where the given incident field $\mathbf{E}^i$ satisfies Maxwell's equation (1.1) in all of $\mathbb{R}^3$, and $\mathbf{E}^s$ is the unknown scattered field. A typical choice for the incident field is a plane wave, in which case

$$\mathbf{E}^i = \mathbf{p} \exp(ik\mathbf{d} \cdot \mathbf{x}),$$

where the real vectors $\mathbf{p}$ (polarization) and $\boldsymbol{d}$ satisfy $|\mathbf{p}| = |\mathbf{d}| = 1$ and $\mathbf{p} \cdot \mathbf{d} = 0$.

In order to uniquely determine the scattered field, we need to impose the Silver–Müller radiation condition,

$$\lim_{|x| \to \infty} (\nabla \times \mathbf{E}^s) \times \mathbf{x} - ik|\mathbf{x}|\mathbf{E}^s_T = 0, \tag{1.4}$$

uniformly in $\hat{\mathbf{x}} = \mathbf{x}/|\mathbf{x}|$, where

$$\mathbf{E}^s_T = (\hat{\mathbf{x}} \times \mathbf{E}^s) \times \hat{\mathbf{x}}.$$

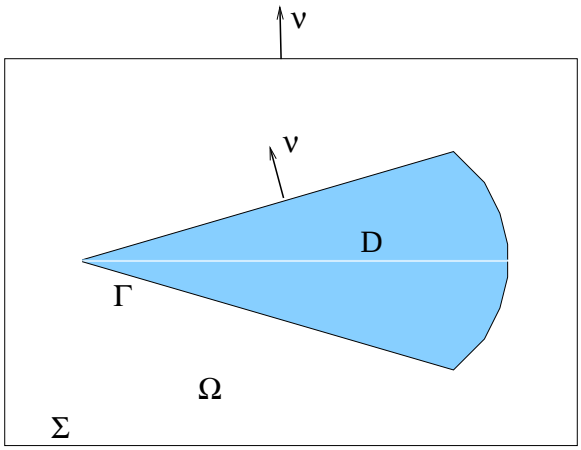


FIG. 1.2. *The truncated computational domain,* $\Omega := D_\Sigma \setminus D$.

More generally, for a sufficiently smooth vector function $\mathbf{u}$, we use the notation

$$\mathbf{u}_T := (\boldsymbol{\nu} \times \mathbf{u}|_S) \times \boldsymbol{\nu}$$

to denote the tangential component of $\mathbf{u}$ on a given surface S with normal $\boldsymbol{\nu}$. Equations (1.1)–(1.4) are uniquely solvable in appropriate function spaces which will be given shortly.

1.2. Truncation of the problem. As stated in the previous section, the scattering problem is posed on an infinite region. In order to apply a finite element method, we need to truncate the domain. Following Hazard and Lenoir, we introduce a simply connected, Lipschitz, polyhedral surface Σ, with interior D_Σ, such that $\overline{D} \subset D_\Sigma$. The outward unit normal on Σ is again denoted $\boldsymbol{\nu}$. For technical reasons associated with the proof of Lemma 3.3, we restrict Σ to be a right parallelepiped. We expect that a more general polyhedral surface also would be appropriate. A smooth surface such as a sphere might also be used, but the analysis of such a scheme would involve the use of curvilinear finite elements which is outside the scope of this paper. We define the truncated computational domain

$$\Omega := D_\Sigma \setminus \overline{D}.$$

Our assumptions imply that the boundary of Ω consists of two disjoint, connected components Σ and Γ, and the region Ω is simply connected (see Figure 1.2). The goal is to use finite elements on Ω to approximate $\mathbf{E}$, but we need a boundary condition on Σ. This is provided by recalling the Stratton–Chu formula that gives a representation of classical solutions of (1.1)–(1.4) away from Γ (see [15]). More precisely, let

$$\mathbb{G}(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}, \mathbf{y})\mathbb{I} + k^{-2}\mathrm{Hess}(\Phi)(\mathbf{x}, \mathbf{y}),$$

where $\mathbb{I}$ is the identity matrix, $\Phi(\mathbf{x}, \mathbf{y})$ is the fundamental solution of the Helmholtz equation in $\mathbb{R}^3$, given by

$$\Phi(\mathbf{x}, \mathbf{y}) := \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|},$$

and Hess$(\cdot)$ is the Hessian operator defined by

$$\mathrm{Hess}(\Phi)_{l,m} = \frac{\partial^2 \Phi}{\partial y_l \partial y_m}, \qquad 1 \le l,\, m \le 3.$$

For $\mathbf{x} \in \mathbb{R}^3 \setminus \overline{D}$, we have

$$\begin{aligned} \mathbf{E}^s(\mathbf{x}) &= \int_\Gamma \left\{ \mathbb{G}(\mathbf{x},\mathbf{y})^T \left(\boldsymbol{\nu}_y \times (\nabla \times \mathbf{E}^s(\mathbf{y}))\right) \right. \\ &\qquad \left. (\nabla_y \times \mathbb{G}(\mathbf{x},\mathbf{y}))^T (\boldsymbol{\nu}_y \times \mathbf{E}^s(\mathbf{y})) \right\} dA(\mathbf{y}) \\ &=: \mathcal{I}(\mathbf{E}^s), \end{aligned} \tag{1.5}$$

where $\nabla_y \times \mathbb{G}(\mathbf{x},\mathbf{y})$ is the columnwise curl of $\mathbb{G}$ with respect to $\mathbf{y}$. Using the fact that $\mathbf{E}^i$ is a regular solution of Maxwell's equations inside D, we have $\mathcal{I}(\mathbf{E}^i) = 0$ in Ω, and thus

$$\mathbf{E} = \mathbf{E}^i + \mathcal{I}(\mathbf{E}) \quad \text{in } \Omega,$$

provided $\mathbf{E}$ is regular enough for $\mathcal{I}(\mathbf{E})$, defined in (1.5), to be well defined.

1.3. A modified integral representation. Unfortunately, the regularity requirement implicit in (1.5) is not met by edge finite elements. We therefore need to extend the definition of $\mathcal{I}$ to allow for less regular arguments $\mathbf{E}^s$. In order to do this, we first recall that

$$H(\mathrm{curl};\Omega) = \left\{ \mathbf{u} \in (L^2(\Omega))^3 |\ \nabla \times \mathbf{u} \in (L^2(\Omega))^3 \right\}$$

and define the subspace

$$X := \left\{ \mathbf{u} \in H(\mathrm{curl};\Omega) |\ \boldsymbol{\nu} \times \mathbf{u} = 0 \,\mathrm{on}\, \Gamma, \text{ and } \boldsymbol{\nu} \times \mathbf{u}|_\Sigma \in (L^2(\Sigma))^3 \right\}.$$

The space X is equipped with the norm

$$\|\mathbf{u}\|_X^2 = \|\mathbf{u}\|_{H(\mathrm{curl};\Omega)}^2 + \|\mathbf{u}_T\|_{(L^2(\Sigma))^3}^2, \tag{1.6}$$

where $\|\cdot\|_{H(\mathrm{curl};\Omega)}$ is the standard norm on $H(\mathrm{curl};\Omega)$, and $\|\cdot\|_{(L^2(\Sigma))^3}$ is the $(L^2(\Sigma))^3$ norm.

Let $\chi \in C_0^\infty(D_\Sigma)$ denote a cutoff function such that $\chi = 1$ on Γ, and define $\widetilde{\mathbb{G}}(\mathbf{x},\cdot) \in H(\mathrm{curl};\Omega)$ by

$$\widetilde{\mathbb{G}}(\mathbf{x},\mathbf{y}) = \chi(\mathbf{y}) \mathbb{G}(\mathbf{x},\mathbf{y}). \tag{1.7}$$

We can now define the integral operator

$$\begin{aligned} \mathcal{I}^R(\mathbf{E}^s) := &\int_\Omega \left((\nabla_y \times \widetilde{\mathbb{G}})^T \nabla \times \mathbf{E}^s - k^2 \widetilde{\mathbb{G}}^T \mathbf{E}^s \right) dV(\mathbf{y}) \\ &+ \int_\Gamma (\nabla_y \times \mathbb{G})^T \boldsymbol{\nu}_y \times \mathbf{E}^s \, dA(\mathbf{y}), \end{aligned} \tag{1.8}$$

where the curl is taken with respect to $\mathbf{y}$ and the integral is evaluated for $\mathbf{x}$ outside the support of the cutoff function χ (in particular, for $\mathbf{x} \in \Sigma$). Using integration by parts we can verify that for a smooth solution $\mathbf{E}^s$ of (1.1)–(1.4),

$$\mathcal{I}^R(\mathbf{E}^s) = \mathcal{I}(\mathbf{E}^s),$$

and thus

$$\mathbf{E} = \mathbf{E}^i + \mathcal{I}^R(\mathbf{E}).$$

Note also that since $\mathcal{I}^R(\mathbf{E})$ is evaluated outside the support of χ (for example, on Σ), a further integration by parts, and the use of the perfect conducting boundary condition on Γ, show that

$$\mathcal{I}^R(\mathbf{E}) = \int_\Omega \left(\nabla \times (\nabla \times \widetilde{\mathbb{G}}) - k^2 \widetilde{\mathbb{G}}\right)^T \mathbf{E}\, dV(\mathbf{y}). \tag{1.9}$$

This is the form of $\mathcal{I}^R$ we shall use for the first part of the analysis.

Before stating the variational problem for Maxwell's equations, we define one further operator. For a sufficiently smooth field $\mathbf{u}$, we can define a tangential boundary operator on Σ as follows:

$$T(\mathbf{u}) := (\nabla \times \mathbf{u})|_\Sigma \times \boldsymbol{\nu} - ik\mathbf{u}_T, \tag{1.10}$$

where $\mathbf{u}_T$ is the tangential component of $\mathbf{u}$ on Σ. With this notation, the truncated version of problem (1.1)–(1.4) is to find $\mathbf{E} \in X$ such that

$$\begin{aligned} &\int_\Omega \nabla \times \mathbf{E} \cdot \nabla \times \bar{\boldsymbol{\phi}} - k^2 \mathbf{E} \cdot \bar{\boldsymbol{\phi}}\, dV \\ &- \int_\Sigma ik\mathbf{E}_T \cdot \bar{\boldsymbol{\phi}}_T\, dA - \int_\Sigma T(\mathcal{I}^R(\mathbf{E})) \cdot \bar{\boldsymbol{\phi}}_T\, dA = \int_\Sigma T(\mathbf{E}^i) \cdot \bar{\boldsymbol{\phi}}_T\, dA \quad \forall \boldsymbol{\phi} \in X. \end{aligned} \tag{1.11}$$

Hazard and Lenoir show [25] that problem (1.11) has a unique solution for every $k > 0$ and given incident field $\mathbf{E}^i$. It is also easy to see that the solution does not depend on the choice of χ. Indeed, (1.11) is identical to (4.12) of [25], allowing for differences in notation and with $\tau = t = \infty$, $\xi_0 = 1$, $\xi = 1$, $\lambda = ik$, and $\zeta^{-1} = 1$. The space X is H^E_τ from [25] but without the constraint on $\nabla \cdot \mathbf{E}$. As we shall see, this constraint is implied directly by the variational equation (1.11).

We shall use (1.11) as the basis of the finite element method we shall analyze. There are several advantages to this formulation compared to other finite element formulations. One advantage is that, unlike methods which use standard absorbing boundary conditions, the convergence of the formulation in this paper can be verified as the mesh size decreases; moreover, the method can be easily applied to a layered medium. Further, compared to a standard coupled finite element–boundary element scheme, the advantage of this method is that no singular integrals must be approximated (since $\mathbf{x} \neq \mathbf{y}$ in (1.5)). The main disadvantages relate to the matrices arising from the discrete problem. We shall discuss these issues more at the end of the paper, but mention here that the matrices have dense blocks.

2. The finite element method. We describe a method based on the tetrahedral edge elements of Nédélec [31] which we summarize next. Results for the hexahedral elements discussed in the same paper follow in the same way and will not be detailed here. In addition, smooth curved boundaries can be handled using the mapping scheme described in [22], but only for elements of the lowest order.

We suppose that Ω has been covered by a regular, quasi-uniform mesh τ_h consisting of tetrahedra of maximum diameter h. Let P_ℓ denote the set of all polynomials in x_1, x_2, and x_3 of maximum degree ℓ, and let $\widetilde{P}_\ell$ denote the set of homogeneous

polynomials of degree ℓ in x_1, x_2, and x_3. On each tetrahedron $K \in \tau_h$, the finite element functions are taken from the set

$$R_\ell = (P_{\ell-1})^3 \oplus S_\ell$$

for some $\ell = 1, 2, \ldots$, where

$$S_\ell = \{\mathbf{p} \in \widetilde{P}_\ell^3 \mid \mathbf{p}(\mathbf{x}) \cdot \mathbf{x} = 0 \quad \forall \mathbf{x}\}.$$

Using these basis functions, we can define

$$V_h = \{\mathbf{u}_h \in H(\mathrm{curl}; \Omega) \mid \mathbf{u}_h|_K \in R_\ell \quad \forall K \in \tau_h\}$$

and

$$X_h = \{\mathbf{u}_h \in V_h \mid \mathbf{u}_h \times \boldsymbol{\nu} = 0 \quad \text{on } \Gamma\}. \tag{2.1}$$

Following [2], we can define an interpolation operator. We proceed by recalling the standard degrees of freedom for these tetrahedral elements. Let K be an element and suppose that $\mathbf{u} \in (H^{\frac{1}{2}+\epsilon}(K))^3$, $\epsilon > 0$, and $\nabla \times \mathbf{u} \in (L^q(K))^3$, $q > 2$. Then the following degrees of freedom on K are well defined. Let e be any edge of K with unit tangent vector $\boldsymbol{\tau}$ and let $\boldsymbol{\nu}$ denote the outward normal to K for any face f. Then let

$$M_e(\mathbf{u}) = \left\{ \int_e \mathbf{u} \cdot \boldsymbol{\tau}\, q\, ds \quad \forall q \in P_{\ell-1}(e),\ \forall \text{ edges } e \text{ of } K \right\}, \tag{2.2a}$$

$$M_f(\mathbf{u}) = \left\{ \int_f \mathbf{u} \times \boldsymbol{\nu} \cdot \mathbf{q}\, dA \quad \forall \mathbf{q} \in (P_{\ell-2}(f))^2,\ \forall \text{ faces } f \text{ of } K \right\}, \tag{2.2b}$$

$$M_K(\mathbf{u}) = \left\{ \int_K \mathbf{u} \cdot \mathbf{q}\, dV \quad \forall \mathbf{q} \in (P_{\ell-3}(K))^3 \right\}. \tag{2.2c}$$

The set $M_\ell(\mathbf{u}) \cup M_f(\mathbf{u}) \cup M_K(\mathbf{u})$ is unisolvent for R_l and curl-conforming (see [31]). Thus, we can define an interpolant π_h elementwise by requiring that $\pi_h \mathbf{u}\big|_K \in R_\ell$ and

$$M_e(\mathbf{u} - \pi_h \mathbf{u}) = M_f(\mathbf{u} - \pi_h \mathbf{u}) = M_K(\mathbf{u} - \pi_h \mathbf{u}) = \{0\}.$$

Error estimates can be proved by scaling to a reference element. In [1] it is shown that

$$\|\mathbf{u} - \pi_h \mathbf{u}\|_{H(\mathrm{curl};\Omega)} \le C h^{\ell'} \left(\|\mathbf{u}\|_{(H^{\ell'}(\Omega))^3} + \|\nabla \times \mathbf{u}\|_{(H^{\ell'}(\Omega))^3} \right) \tag{2.3}$$

for any ℓ' with $\ell \ge \ell' > \frac{1}{2}$.

Now suppose that $\mathbf{u}$ is such that $\nabla \times \mathbf{u}|_K \in (P_\ell)^3$ for each element $K \in \tau_h$. Then using a standard scaling argument like the one proving the above estimate, and using the equivalence of norms for piecewise polynomials on the reference element as in the proof of equation (2.4) of [3], we obtain that for $0 < \epsilon \le \frac{1}{2}$,

$$\|\mathbf{u} - \pi_h \mathbf{u}\|_{(L^2(\Omega))^3} \le C \left(h^{\frac{1}{2}+\epsilon} \|\mathbf{u}\|_{(H^{\frac{1}{2}+\epsilon}(\Omega))^3} + h \|\nabla \times \mathbf{u}\|_{(L^2(\Omega))^2} \right). \tag{2.4}$$

For later use we need to discretize the operator $\mathcal{I}^R$ defined in (1.8). Let $\widetilde{\mathbb{G}}_h(\mathbf{x}, \cdot)$ denote the matrix function such that if $\tilde{\mathbf{g}}_{h,m}(\mathbf{x}, \cdot)$ is the mth column of $\widetilde{\mathbb{G}}_h(\mathbf{x}, \cdot)$ and $\tilde{\mathbf{g}}_m(\mathbf{x}, \cdot)$ is the mth column of $\widetilde{\mathbb{G}}(\mathbf{x}, \cdot)$, then (recalling that $\mathbf{x} \in \Sigma$)

1. $\tilde{\mathbf{g}}_{h,m}(\mathbf{x},\cdot) \in V_h$, $1 \le m \le 3$;
2. $(\tilde{\mathbf{g}}_{h,m}(\mathbf{x},\cdot))_T$ interpolates $(\mathbf{g}_m(\mathbf{x},\cdot))_T$ on Γ (using edge and face degrees of freedom (2.2a) and (2.2b));
3. $(\tilde{\mathbf{g}}_{h,i}(\mathbf{x},\cdot))_T = 0$ on all tetrahedra having a face or edge on Σ.

Obviously, this discretization of $\widetilde{\mathbb{G}}(\mathbf{x},\mathbf{y})$ is not uniquely determined. For computational convenience, we use $(\tilde{\mathbf{g}}_{h,m})$, $m = 1,2,3$, that decay to zero rapidly away from Γ. This minimizes the support of $\widetilde{\mathbb{G}}_h$ and is the reason for discretizing $\widetilde{\mathbb{G}}$.

We can then define the discretized version of the integral operator defined in (1.8), denoted by $\mathcal{I}_h(\mathbf{u})$, for $\mathbf{u} \in H(\operatorname{curl}\Omega)$ and $\mathbf{x}$ outside the support of $\widetilde{\mathbb{G}}_h$ (in particular, for $\mathbf{x} \in \Sigma$) by

$$\mathcal{I}_h(\mathbf{u})(\mathbf{x}) = \int_\Omega \Big((\nabla \times \widetilde{\mathbb{G}}_h(\mathbf{x},\mathbf{y}))^T \nabla \times \mathbf{u}(\mathbf{y}) - k^2 (\widetilde{\mathbb{G}}_h(\mathbf{x},\mathbf{y}))^T \mathbf{u}(\mathbf{y}) \Big)\, dV(\mathbf{y}). \tag{2.5}$$

As long as $\mathbf{x}$ is on Σ, $\mathcal{I}_h(\mathbf{u})$ is a smooth function of $\mathbf{x}$. Hence $T(\mathcal{I}_h(\mathbf{u}))$ is a well defined and smooth (tangential) vector field on each face on Σ.

The finite element analogue of (1.11) is to find $\mathbf{E}_h \in X_h$ such that

$$\begin{aligned} &\int_\Omega \nabla \times \mathbf{E}_h \cdot \nabla \times \overline{\boldsymbol{\phi}}_h - k^2 \mathbf{E}_h \cdot \overline{\boldsymbol{\phi}}_h \, dV - \int_\Sigma (ik \mathbf{E}_{h,T} + T(\mathcal{I}_h(\mathbf{E}_h))) \cdot \overline{\boldsymbol{\phi}}_{h,T} \, dA \\ &\quad = \int_\Sigma T(\mathbf{E}^i) \cdot \overline{\boldsymbol{\phi}}_h \, dA \quad \forall \boldsymbol{\phi}_h \in X_h. \end{aligned} \tag{2.6}$$

Unfortunately, we have been unable to prove directly that $\mathbf{E}_h$ converges to $\mathbf{E}$. Instead we first analyze the convergence of the solution of the following intermediate problem of finding $\tilde{\mathbf{E}}_h \in X_h$ such that

$$\begin{aligned} &\int_\Omega \nabla \times \tilde{\mathbf{E}}_h \cdot \nabla \times \overline{\boldsymbol{\phi}}_h - k^2 \tilde{\mathbf{E}}_h \cdot \overline{\boldsymbol{\phi}}_h \, dV - \int_\Sigma \Big(ik \tilde{\mathbf{E}}_{h,T} + T(\mathcal{I}^R(\tilde{\mathbf{E}}_h)) \Big) \cdot \overline{\boldsymbol{\phi}}_{h,T} \, dA \\ &\quad = \int_\Sigma T(\mathbf{E}^i) \cdot \overline{\boldsymbol{\phi}}_h \, dA \quad \forall \boldsymbol{\phi}_h \in X_h. \end{aligned} \tag{2.7}$$

Here the operator $\mathcal{I}^R$ is not discretized.

In the next section we shall show that $\tilde{\mathbf{E}}_h$ is well defined and converges to the true solution $\mathbf{E}$. In principle, we could implement (2.7) but the integral operator $\mathcal{I}^R$ would become increasingly more expensive to evaluate as the mesh size decreases, since a volume integral over a fixed volume must be evaluated. Hence we prefer to compute with (2.6) since $\mathcal{I}_h$ can be constructed to involve only a skin of tetrahedra that share an edge with Γ.

Another justification for the use of (2.6) is that the solution $\mathbf{E}_h$ is independent of the choice of $\widetilde{\mathbb{G}}_h$ (provided the conditions mentioned earlier in this section are satisfied). In order to show this, we state the following lemma, which also partially justifies our choice of $\widetilde{\mathbb{G}}_h$. The proof of this lemma is straightforward and is postponed until the appendix.

LEMMA 2.1. *Suppose* (2.6) *has a unique solution for each* $\widetilde{\mathbb{G}}_h$ *satisfying requirements* (1)–(3) *discussed earlier in this section. Then the solution is independent of the choice of* $\widetilde{\mathbb{G}}_h$.

From this lemma it suffices to prove existence and uniqueness for a particular choice of $\widetilde{\mathbb{G}}_h$ to then conclude the result for any $\widetilde{\mathbb{G}}_h$.

3. Analysis of the scheme. We will prove that as the mesh size h decreases, the solutions of the discrete problem (2.7) approach the exact solution of (1.11). In order to prove this, we first need to carefully describe the function spaces we will be working with. This is done in subsection 3.1. In subsection 3.2, we rewrite both the continuous and the discrete problems in a convenient operator form. In subsection 3.3, we show that the operator equations are of Fredholm type. We also demonstrate the convergence (in some suitable norm) of the discrete operators to their continuous analogues as the mesh size decreases.

Subsection 3.5 concerns a collective compactness result. We follow the general approach of [34] in that we verify convergence of some operation using the theory of collectively compact operators. The main property of edge elements relevant to this approach is the discrete compactness property (see [27, 28, 29, 34]). Subsection 3.6 combines all these results into our first theorem about the finite element scheme.

An alternative approach to proving the operator convergence via the theory of mixed methods has been employed by Boffi [9] and Boffi, Brezzi, and Gastaldi [10, 11]. So far this approach has been aimed at proving convergence in $(L^2(\Omega))^3$ which is appropriate for eigenvalue problems. In that case, Boffi proved the equivalence of the mixed method and the discrete compactness approaches [8] (see also [14]), so the choice of which method to use is immaterial.

For the remainder of the paper we shall assume that Γ and Σ are each connected, so $\partial\Omega = \Gamma \cup \Sigma$ and $\Gamma \cap \Sigma = \phi$. We could allow both boundaries to be disconnected if necessary at the cost of using more notation and complexity.

3.1. Some function spaces and estimates. In this subsection, we define useful function spaces, which shall be used in the remainder of the paper. Convenient decompositions and properties of these spaces are also listed. Let

$$S = \left\{ p \in H^1(\Omega) \mid p = 0 \text{ on } \Gamma \text{ and } p = \text{ constant on } \Sigma \right\}. \tag{3.1}$$

Then $\nabla S \subset X$ and is a closed subspace of X. Hence we may write, using the $(L^2(\Omega))^3$ inner product,

$$X = X_0 \oplus \nabla S \tag{3.2}$$

and

$$X_0 = \left\{ \mathbf{u} \in X \;\middle|\; \int_\Omega \mathbf{u} \cdot \nabla\xi \, dV = 0 \quad \forall \xi \in S \right\}. \tag{3.3}$$

Using Costabel's regularity result [16] we know that the injection $X_0 \to (L^2(\Omega))^3$ is compact. Furthermore, suppose $\mathbf{u} \in X_0$, $\nabla \times \mathbf{u} = 0$ in Ω and $\mathbf{u}_T = 0$ on Σ. Then since $\nabla \times \mathbf{u} = 0$, and Ω is simply connected, there is a scalar potential $p \in H^1(\Omega)$ such that

$$\mathbf{u} = \nabla p.$$

The tangential components of $\mathbf{u}$ vanish on Γ and Σ, and thus we can take $p \in S$. The fact that $\mathbf{u} \in \nabla S$ and $\mathbf{u} \in X_0$ then implies $\mathbf{u} = 0$. Thus, using the compactness result above, for $\mathbf{u} \in X_0$ there is a constant $C > 0$ such that

$$\|\mathbf{u}\|_X \le C \left(\|\nabla \times \mathbf{u}\|_{(L^2(\Omega))^3} + \|\mathbf{u}_T\|_{(L^2(\Sigma)^3)} \right). \tag{3.4}$$

We can also decompose X_h. It is well known (see [31]) that if

$$S_h = \{p_h \in S \mid p_h|_K \in P_\ell \quad \forall K \in \tau_h\}, \tag{3.5}$$

then

$$\nabla S_h \subset X_h, \tag{3.6}$$

and we may write (again using the $(L^2(\Omega))^3$ inner product)

$$X_h = X_{0,h} \oplus \nabla S_h, \tag{3.7}$$

where

$$X_{0,h} = \left\{ \mathbf{u}_h \in X_h \;\middle|\; \int_\Omega \mathbf{u}_h \cdot \nabla \xi_h \, dV = 0 \quad \forall \xi_h \in S_h \right\} \tag{3.8}$$

is the space of discrete divergence-free fields. The main difficulty with the analysis of the error is that $X_{0,h} \not\subset X_0$.

3.2. An operator equation. In order to use the Fredholm alternative in the analysis of the finite element formulation, we rewrite the continuous variational problem (1.11) and the discrete finite element problem (2.6) as operator equations. We introduce some convenient notation to be used in the remainder of this section. For $\mathbf{u}, \mathbf{v} \in X$ we denote

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \nabla \times \mathbf{u} \cdot \nabla \times \bar{\mathbf{v}} + k^2 \mathbf{u} \cdot \overline{\mathbf{v}} \, dV - ik \int_\Sigma \mathbf{u}_T \cdot \overline{\mathbf{v}}_T \, dA. \tag{3.9}$$

Note that $|a(\mathbf{u}, \mathbf{u})|$ is a norm equivalent to $\|\mathbf{u}\|_X$. Define the operator $A : (L^2(\Omega))^3 \to (L^2(\Omega))^3$ such that for all $\mathbf{f} \in (L^2(\Omega))^3$, $A\mathbf{f} \in X_0 \subset (L^2(\Omega))^3$ satisfies

$$a(A\mathbf{f}, \boldsymbol{\phi}) = -2k^2 \int_\Omega \mathbf{f} \cdot \overline{\boldsymbol{\phi}} \, dV - \int_\Sigma T(\mathcal{I}^R(\mathbf{f})) \cdot \overline{\boldsymbol{\phi}}_T dA \quad \forall \boldsymbol{\phi} \in X_0. \tag{3.10}$$

By the Lax–Milgram lemma this problem is well posed. In particular, using the expression for $\mathcal{I}^R$ in (1.9) shows that

$$\|T(\mathcal{I}^R(\mathbf{u}))\|_{(L^2(\Sigma))^3} \leq C \|\mathbf{u}\|_{(L^2(\Omega))^3}$$

which allows us to prove the continuity of the right-hand side of (3.10).

Similarly, we define $\mathbf{F} \in X_0$ by

$$a(\mathbf{F}, \boldsymbol{\phi}) = \int_\Sigma T(\mathbf{E}^i) \cdot \bar{\phi} dA \quad \forall \boldsymbol{\phi} \in X_0. \tag{3.11}$$

We proceed to show that the operator problem of finding $\mathbf{E} \in (L^2(\Omega))^3$ such that

$$\mathbf{E} + A\mathbf{E} = \mathbf{F} \tag{3.12}$$

is exactly equivalent to solving the Hazard–Lenoir equation (1.11). Any solution of (1.11) is divergence free, and thus if we pick a test function $\boldsymbol{\phi} \in X_0$, then (1.11) can be recast as follows: *Find* $\mathbf{E} \in X_0$ *such that*

$$a(\mathbf{E} + A\mathbf{E} - \mathbf{F}, \boldsymbol{\phi}) = 0 \quad \forall \boldsymbol{\phi} \in X_0.$$

Hence, $\mathbf{E} + A\mathbf{E} - \mathbf{F} = 0$ in X_0 and this certainly implies equality in $(L^2(\Omega))^3$. Conversely, if we have a solution $\mathbf{E} \in (L^2(\Omega))^3$ of

$$\mathbf{E} + A\mathbf{E} = \mathbf{F},$$

then since $\mathbf{E} = \mathbf{F} - A\mathbf{E}$, we know that $\mathbf{E} \in X_0$. Therefore $\mathbf{E}$ satisfies

$$a(\mathbf{E} + A\mathbf{E} - \mathbf{F}, \boldsymbol{\xi}) = 0 \quad \forall \boldsymbol{\xi} \in X,$$

which is the Hazard–Lenoir equation (1.11). This shows the equivalence of the operator equation (3.12) and the Hazard–Lenoir equation (1.11). The existence and uniqueness of solutions to (3.12) now follow from those of (1.11) (see [25]).

3.3. The Fredholm alternative. Hazard and Lenoir proved the compactness of A as an operator from X_0 to X_0. We need to perform the analysis in $(L^2(\Omega))^3$ since $X_{0,h} \not\subset X_0$. In fact, A is compact as a map from $(L^2(\Omega))^3$ to $(L^2(\Omega))^3$, as the next lemma shows.

LEMMA 3.1. *The map $A : (L^2(\Omega))^3 \to (L^2(\Omega))^3$ is compact.*

Proof. By the Lax–Milgram lemma, A is well defined and bounded as a map from $(L^2(\Omega))^3$ into X_0. The extension of Weber's compactness theorem due to Costabel [16] proves that X_0 is compactly embedded in $(L^2(\Omega))^3$. This proves the compactness of A. ∎

Using this lemma we can see that (3.12) is a Fredholm equation on $(L^2(\Omega))^3$ and hence, via Hazard and Lenoir's uniqueness result, (3.12) has a unique solution $\mathbf{E}$ in $(L^2(\Omega))^3$.

Now we write the discrete problem (2.7) as an operator equation. We define the operator $\tilde{A}_h : (L^2(\Omega))^3 \to (L^2(\Omega))^3$ as the straightforward discrete analogue of A. By this we mean that for a given $\mathbf{f} \in (L^2(\Omega))^3$, $\tilde{A}_h\mathbf{f} \in X_{0,h}$ satisfies

$$a(\tilde{A}_h\mathbf{f}, \boldsymbol{\xi}_h) = -2k^2 \int_\Omega \mathbf{f} \cdot \bar{\boldsymbol{\xi}}_h \, dV - \int_\Sigma T(\mathcal{I}^R(\mathbf{f})) \cdot \bar{\boldsymbol{\xi}}_{h,T} \, dA \quad \forall \boldsymbol{\xi}_h \in X_{0,h}. \tag{3.13}$$

We can also define $\mathbf{F}_h \in X_{0,h}$ by

$$a(\mathbf{F}_h, \boldsymbol{\xi}_h) = \int_\Sigma T(\mathbf{E}^i) \cdot \overline{\boldsymbol{\xi}}_{h,T} \, dA \quad \forall \boldsymbol{\xi}_h \in X_{0,h}. \tag{3.14}$$

The operator $\tilde{A}_h$ and vector $\mathbf{F}_h$ are well defined by the Lax–Milgram lemma.

We can then pose the problem of finding $\tilde{\mathbf{E}}_h \in (L^2(\Omega))^3$ such that

$$\tilde{\mathbf{E}}_h + \tilde{A}_h\tilde{\mathbf{E}}_h = \mathbf{F}_h. \tag{3.15}$$

Assuming such a solution can be found, we have

$$\tilde{\mathbf{E}}_h = \mathbf{F}_h - \tilde{A}_h\tilde{\mathbf{E}}_h \in X_{0,h}.$$

As a first step in our analysis of this problem, we now need to demonstrate that as the mesh size h decreases, the discrete operator $\tilde{A}_h$ converges to A. This is the content of the next lemma, the proof of which is rather classical (see [26]).

LEMMA 3.2. *For fixed $\mathbf{f} \in (L^2(\Omega))^3$, $\tilde{A}_h\mathbf{f} \to A\mathbf{f}$ in X as $h \to 0$.*

Proof. We rewrite the problem defining A, (3.10), as the mixed problem of finding $A\mathbf{f} \in X$ and $p \in S$ such that

$$\begin{aligned} a(A\mathbf{f}, \boldsymbol{\xi}) + \int_\Omega \boldsymbol{\xi} \cdot \nabla \bar{p} \, dV &= -2k^2 \int_\Omega \mathbf{f} \cdot \bar{\boldsymbol{\xi}} \, dV - \int_\Sigma T(\mathcal{I}_R(\mathbf{f})) \cdot \bar{\boldsymbol{\xi}}_T \, dA \quad \forall \boldsymbol{\xi} \in X, \\ \int_\Omega A\mathbf{f} \cdot \nabla \bar{\phi} \, dV &= 0 \quad \forall \phi \in S. \end{aligned}$$

Since $a(\cdot,\cdot)$ is coercive on X and $\nabla S \subset X$, we can easily verify the Babuška–Brezzi condition and conclude that this is a well-posed problem (taking $\boldsymbol{\xi} \in X_0$ shows it reduces to (3.10) in this case).

Similarly, the discrete problem (3.13) may be written as

$$a(\tilde{A}_h\mathbf{f},\boldsymbol{\xi}_h) + \int_\Omega \boldsymbol{\xi}_h \cdot \nabla \bar{p}_h \, dV = -2k^2 \int_\Omega \mathbf{f}\cdot\bar{\boldsymbol{\xi}}_h \, dV - \int_\Sigma T(\mathcal{I}_R(\mathbf{f}))\cdot\bar{\boldsymbol{\xi}}_{h,T}\, dA \quad \forall \boldsymbol{\xi}_h \in X_h,$$

$$\int_\Omega \tilde{A}_h\mathbf{f}\cdot\nabla\bar{\phi}_h \, dV = 0 \quad \forall \phi_h \in S_h. \tag{3.16}$$

The coercivity of $a(\cdot,\cdot)$ and the fact that $\nabla S_h \subset X_h$ allows us to verify the discrete Babuška–Brezzi condition, and we conclude that the following estimate holds (see [13]):

$$\begin{aligned}\|A\mathbf{f} - \tilde{A}_h\mathbf{f}\|_X + \|\nabla(p-p_h)\|_{(L^2(\Omega))^3} \\ \le c\left\{\inf_{\chi_h\in X_h}\|A\mathbf{f}-\chi_h\|_X + \inf_{\xi_h\in S_h}\|\nabla(p-\xi_h)\|_{(L^2(\Omega))^3}\right\}.\end{aligned} \tag{3.17}$$

The theorem follows from standard arguments using the density of S_h in S and X_h in X. ∎

The pointwise convergence of $\tilde{A}_h$ to A is not sufficient to conclude that the operator $(I+\tilde{A}_h)$ is invertible. Before proving this invertibility we need to prove a technical regularity result.

3.4. A regularity result. Before stating and proving our main compactness result we need the following regularity result. This result claims that if a vector field $\mathbf{u} \in X$ and a discrete field $\mathbf{u}_h \in X_h$ have curls which agree in Ω, and if the tangential components of the fields agree on the boundary, then $\mathbf{u}$ possesses some extra regularity. The proof of this lemma proceeds by considering a decomposition of $\mathbf{u}$ and establishing a regularity result for each component.

LEMMA 3.3. *Let $\mathbf{u}_h \in X_{0,h}$ and suppose $\mathbf{u} \in X_0$ satisfies*

$$\begin{aligned}\nabla\times\mathbf{u} &= \nabla\times\mathbf{u}_h \text{ in } \Omega,\\ \boldsymbol{\nu}\times\mathbf{u} &= \boldsymbol{\nu}\times\mathbf{u}_h \text{ on } \partial\Omega.\end{aligned}$$

Then there is an $\epsilon_{max} > 0$ such that $\mathbf{u} \in (H^s(\Omega))^3$, for $1/2 \le s < 1/2+\epsilon_{max}$, and

$$\|\mathbf{u}\|_{(H^s(\Omega))^3} \le C\left(\|\nabla\times\mathbf{u}\|_{(L^2(\Omega))^3} + \|\boldsymbol{\nu}\times\mathbf{u}\|_{(H^{s-1/2}(\Sigma))^3}\right).$$

Remark 1. (1) In [16] this result is proved for $S = 1/2$, and in [2] the result is proved when $\boldsymbol{\nu}\times\mathbf{u} = 0$ on $\partial\Omega$ (including Σ). The result here is possible because $\boldsymbol{\nu}\times\mathbf{u}_h$ is a piecewise polynomial on Γ and Σ and hence is smoother than just square integrable on Γ and Σ. The proof we shall give combines those in [2] and [16].

(2) It is in the proof of this theorem that we use the fact that Σ is a parallelepiped. This is a hypothesis for an extension result given in Lemma A.2. Note that extension results of this type are valid for arbitrary Lipschitz polygons in $\mathbb{R}^2$ (see [24, Thm. 1.5.2.3], for example). Dauge (in private communication) suggested that the same is true in $\mathbb{R}^3$. Assuming this is so, the proofs in this paper are valid for more general outer boundaries Σ.

Proof. In this proof we shall use the spaces $H^l(\Sigma)$ for $l > 0$. We define (on Σ here, but using obvious notation also on Γ)

$$H^l(\Sigma) = \left\{g = \xi|_\Sigma \mid \xi \in H^{l+1/2}(\Omega)\right\}$$

with the norm

$$\|g\|_{H^l(\Sigma)} = \inf_{\xi \in H^{l+1/2}(\Omega),\ \xi|_\Sigma = g} \|\xi\|_{H^{l+1/2}(\Omega)}.$$

This definition makes sense for all $l > 0$ and agrees with the trace space defined in terms of intrinsic norms for $l \leq 1$.

As in [16], let $\mathcal{O}$ denote a smooth, bounded, connected domain with simply connected boundary $\partial\mathcal{O}$ containing $\bar{\Omega}$ in its interior (see Figure 3.1). First we construct a vector potential $\mathbf{w} \in (H^1(\mathcal{O}))^3$ such that

$$\left.\begin{array}{rcl} \nabla \times \mathbf{w} & = & \nabla \times \mathbf{u} \\ \nabla \cdot \mathbf{w} & = & 0 \end{array}\right\} \text{ in } \Omega. \tag{3.18}$$

Let $\mathbf{z}$ be defined on $\mathcal{O}$ by

$$\mathbf{z} = \begin{cases} 0 & \text{in } D, \\ \nabla \times \mathbf{u} & \text{in } \Omega, \\ \nabla \xi & \text{in } \mathcal{O} \setminus \overline{(\Omega \cup D)}, \end{cases}$$

where $\xi \in H^1(\mathcal{O} \setminus \overline{(\Omega \cup D)})/\mathbb{R}$ solves the boundary value problem

$$\begin{aligned} &\Delta \xi = 0 \text{ in } \mathcal{O} \setminus \overline{(\Omega \cup D)}, \\ &\frac{\partial \xi}{\partial \nu} = \boldsymbol{\nu} \cdot \nabla \times \mathbf{u} \text{ on } \Sigma, \\ &\frac{\partial \xi}{\partial \nu} = 0 \text{ on } \partial\mathcal{O}. \end{aligned}$$

Note that $\boldsymbol{\nu} \cdot \nabla \times \mathbf{u} \in H^{-1/2}(\Sigma)$ since $\nabla \cdot (\nabla \times \mathbf{u}) = 0$ and that this also implies the necessary compatibility condition for solvability.

Of course, $\boldsymbol{\nu} \cdot \nabla \times \mathbf{u} = 0$ on Γ since the boundary condition is perfectly conducting. Thus $\mathbf{z}$ has a continuous normal component across Σ and Γ, $\nabla \cdot \mathbf{z} = 0$ in $\mathcal{O}$. Hence Lemma 3.5 of [2] ensures that there is a function $\mathbf{w} \in (H^1(\mathcal{O}))^3$ such that

$$\nabla \times \mathbf{w} = \mathbf{z} \text{ and } \nabla \cdot \mathbf{w} = 0 \text{ in } \mathcal{O}.$$

Hence $\mathbf{w}$ verifies the desired properties in (3.18).

As $\nabla \times (\mathbf{u} - \mathbf{w}) = 0$, in Ω there is a scalar potential $p \in H^1(\Omega)$ such that

$$\mathbf{u} - \mathbf{w} = \nabla p.$$

The fact that $\nabla \cdot (\mathbf{u} - \mathbf{w}) = 0$ then implies that

$$\Delta p = 0 \text{ in } \Omega.$$

We need to demonstrate that p possesses extra regularity in order to establish the smoothness of $\mathbf{u}$. This can be done following [2]. Inside D we have $\nabla \times \mathbf{w} = 0$, and since $\mathbf{w} \in (H^1(\mathcal{O}))^3$, there is a scalar potential $\eta \in H^2(D)$ such that $\mathbf{w} = \nabla \eta$ in D. On the boundary Γ, we have

$$(\boldsymbol{\nu} \times (\mathbf{u} - \mathbf{w})) \times \boldsymbol{\nu} = (\boldsymbol{\nu} \times \nabla p) \times \boldsymbol{\nu}.$$

We now use the boundary condition on Γ and the fact that $\mathbf{w} = \nabla \eta$ in D to obtain

$$\nabla_\Gamma p = -\nabla_\Gamma \eta,$$

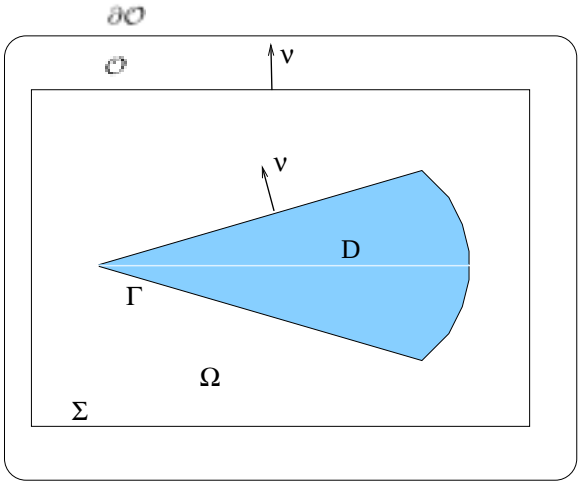


FIG. 3.1. *The configuration of the smooth enclosing domain O.*

where we have used the notation $\nabla_\Gamma p = (\boldsymbol{\nu} \times \nabla p) \times \boldsymbol{\nu}$ so that ∇_Γ is the surface gradient. Since $\eta \in H^{3/2}(\Gamma)$, we can conclude $p \in H^{3/2}(\Gamma)$.

For the outer boundary Σ, we follow [16]. Clearly $p|_\Sigma \in H^{1/2}(\Sigma)$ and $\mathbf{w}|_\Sigma \in (H^{1/2}(\Sigma))^3$. Furthermore, on Σ,

$$(\boldsymbol{\nu} \times (\mathbf{u} - \mathbf{w})) \times \boldsymbol{\nu} = (\boldsymbol{\nu} \times \nabla p) \times \boldsymbol{\nu}.$$

Thus, on Σ,

$$\mathbf{u}_T - \mathbf{w}_T = \nabla_\Sigma p.$$

Unlike the boundary condition on Γ, now $\mathbf{u}_T$ does not vanish on Σ, since

$$\mathbf{u}_T = \mathbf{u}_{h,T}.$$

The polyhedral nature of Σ implies that on each face F of the surface, we have that $\boldsymbol{\nu}$ is a constant vector, and $\mathbf{u}_h$ is a piecewise polynomial, so

$$\mathbf{u}_{h,T} = (\boldsymbol{\nu} \times \mathbf{u}_h) \times \boldsymbol{\nu} \in (H^\epsilon(F))^3$$

for $0 \le \epsilon < 1/2$. Thus $\nabla_\Sigma p \in (H^\epsilon(F))^3$ and hence $p \in H^{1+\epsilon}(F)$. Moreover, since $p \in H^{1/2}(\Sigma)$ it cannot have line discontinuities, so in fact p is continuous on Σ.

We are thus assured that p is continuous at each edge of Σ, and $p \in H^{1+\epsilon}(F)$ for each face F. Hence, via Lemma A.2 there is an extension denoted $\tilde{p} \in H^{3/2+\epsilon}(\Omega)$. Then using this extension and considering $p - \tilde{p}$, we can conclude, using Corollary 18.15 of Dauge [20], that there is an ϵ_{max} with $0 < \epsilon_{max} < 1/2$ such that $p \in H^{3/2+\epsilon}(\Omega)$ for $0 \le \epsilon < \epsilon_{max}$. Using the fact that $\mathbf{u} = \mathbf{w} + \nabla p$ completes the proof. ∎

3.5. A collective compactness result. Let Λ be a countable set of positive real numbers whose only accumulation point is at zero. We assume that the mesh size $h \in \Lambda$ and hence that there is a sequence of mesh sizes $h_n \to 0$ as $n \to \infty$.

LEMMA 3.4. *The set of operators* $\{\tilde{A}_h : (L^2(\Omega))^3 \to (L^2(\Omega))^3\}_{h\in\Lambda}$ *is collectively compact.*

Remark 2. This is essentially the discrete compactness property (for a full discussion of this property see [27, 28, 29, 9]).

Proof. Let $U \subset (L^2(\Omega))^3$ be a bounded set and define

$$\mathcal{A}(U) = \{\mathbf{w} \in (L^2(\Omega))^3 \mid \mathbf{w} = \tilde{A}_h(\mathbf{u}) \text{ for some } \mathbf{u} \in U \text{ and } h \in \Lambda\}.$$

To prove collective compactness, we need to show that $\mathcal{A}(U)$ is precompact in $(L^2(\Omega))^3$. Let $\{\mathbf{w}_n\} \subset \mathcal{A}(U)$ be a sequence. Then for each n there is an h_n and $\mathbf{u}_n \in U$ such that

$$\mathbf{w}_n = \tilde{A}_{h_n}(\mathbf{u}_n) \in X_{0,h_n}.$$

Without loss of generality, we can assume $h_n \to 0$ as $n \to \infty$ (otherwise we are in a finite-dimensional vector space, and a convergent subsequence of $\{\mathbf{w}_n\}$ is guaranteed).

Let $p^n \in S$ satisfy

$$\int_\Omega \nabla p^n \cdot \nabla \overline{\xi}\, dV = \int_\Omega \mathbf{w}_n \cdot \nabla \overline{\xi}\, dV \quad \forall \xi \in S.$$

We can decompose $\mathbf{w}_n \in (L^2(\Omega))^3$ by defining $\mathbf{w}^n$ as

$$\mathbf{w}^n = \mathbf{w}_n - \nabla p^n.$$

Clearly, $\mathbf{w}^n \in X_0$. Since $\{\mathbf{w}_n\}$ is a bounded subset of X, we can conclude $\{\mathbf{w}^n\}$ is a bounded subset in X_0. Thus, using the compactness result of [16], there is a subsequence, still denoted $\{\mathbf{w}^n\}$, and an element $\mathbf{w} \in X_0$, such that

$$\mathbf{w}^n \longrightarrow \mathbf{w} \text{ as } n \to \infty \begin{cases} \text{weakly} & \text{in} \quad X, \\ \text{strongly} & \text{in} \quad (L^2(\Omega))^3. \end{cases}$$

Moreover, the definition of $\mathbf{w}^n$ reveals that since $\mathbf{w}_n$ is a finite element function,

$$\nabla \times \mathbf{w}^n = \nabla \times \mathbf{w}_n \in (L^q(\Omega))^3,$$

and $\boldsymbol{\nu} \times \mathbf{w}^n = \boldsymbol{\nu} \times \mathbf{w}_n$ on Σ (and Γ).

Hence by Lemma 3.3, $\mathbf{w}^n \in (H^{\frac{1}{2}+\epsilon}(\Omega))^3$ for some $\epsilon > 0$. Since $\nabla \times \mathbf{w}^n \in (L^q(\Omega))^3$ for any $q \geq 2$, we can conclude that the interpolant $\pi_{h_n}\mathbf{w}^n$ is well defined. The interpolant of $\mathbf{w}_n$, and hence the interpolant of ∇p^n, is also well defined. Thus,

$$\pi_{h_n}\mathbf{w}^n = \pi_{h_n}(\mathbf{w}_n - \nabla p^n) = \mathbf{w}_n - \nabla p_n$$

for some $p_n \in S_{h_n}$. The relation

$$\pi_{h_n} \nabla p = \nabla p_n \quad \text{for some } p_n \in S_{h_n}$$

follows from properties of edge elements (see [23]). Hence, using the fact that $\mathbf{w} \in X_0$ and $\mathbf{w}_n \in X_{0,h_n}$, we have

$$\begin{aligned} \int_\Omega (\mathbf{w} - \mathbf{w}_n) \cdot \overline{(\mathbf{w}^n - \mathbf{w}_n)}\, dV &= \int_\Omega (\mathbf{w} - \mathbf{w}_n) \cdot \overline{(\mathbf{w}^n - \pi_{h_n}\mathbf{w}^n - \nabla p_n)}\, dV \\ &= \int_\Omega (\mathbf{w} - \mathbf{w}_n) \cdot \overline{(\mathbf{w}^n - \pi_{h_n}\mathbf{w}^n)}\, dV \end{aligned}$$

so that

$$\|\mathbf{w}-\mathbf{w}_n\|_{(L^2(\Omega))^3} \le \|\mathbf{w}-\mathbf{w}^n\|_{(L^2(\Omega))^3} + \|\mathbf{w}^n-\pi_{h_n}\mathbf{w}^n\|_{(L^2(\Omega))^3}.$$

Now we can use the estimate (2.4) to conclude (using the fact that $\nabla\times\mathbf{w}^n = \nabla\times\mathbf{w}_n$) that

$$\begin{aligned}\|\mathbf{w}-\mathbf{w}_n\|_{(L^2(\Omega))^3} \le{}& \|\mathbf{w}-\mathbf{w}^n\|_{(L^2(\Omega))^3}\\ &+C\left(h^{1/2+\epsilon}\|\mathbf{w}^n\|_{(H^{1/2+\epsilon}(\Omega))^3} + h\|\nabla\times\mathbf{w}_n\|_{(L^2(\Omega))^3}\right).\end{aligned}$$

Using this estimate, the a priori estimate in Lemma 3.3, and the following inverse estimate for fractional power Sobolev spaces norms of piecewise polynomial functions on quasi-uniform meshes proved in [12]:

$$\|\boldsymbol{\nu}\times\mathbf{w}_n\|_{(H^{\epsilon}(\Sigma))^3} \le Ch^{-\epsilon}\|\boldsymbol{\nu}\times\mathbf{w}_n\|_{(L^2(\Sigma))^3},$$

we have that

$$\|\mathbf{w}-\mathbf{w}_n\|_{(L^2(\Omega))^3} \le \|\mathbf{w}-\mathbf{w}^n\|_{(L^2(\Omega))^3} + Ch^{1/2}\|\mathbf{w}_n\|_X.$$

The first term on the right-hand side tends to zero by construction, and the second term tends to zero since $h_n \to 0$ as $n\to\infty$. Hence we have proved the desired result. ∎

3.6. Error estimates. We can now analyze the operator based problems (3.12) and (3.15) which are to find $\mathbf{E}\in(L^2(\Omega))^3$ and $\mathbf{E}_h\in(L^2(\Omega))^3$ such that

$$\begin{aligned}(I+A)\mathbf{E} &= \mathbf{F},\\ (I+\tilde{A}_h)\tilde{\mathbf{E}}_h &= \mathbf{F}_h,\end{aligned}$$

for $h\in\Lambda$. We have the following theorem.

THEOREM 3.5. *For $h\in\Lambda$ sufficiently small, $(I+\tilde{A}_h)^{-1}$ exists and is uniformly bounded as a map from $(L^2(\Omega))^3$ to $(L^2(\Omega))^3$. The error estimate*

$$\begin{aligned}\|\tilde{\mathbf{E}}_h-\mathbf{E}\|_{(L^2(\Omega))^3} \le c\Big(&\|\mathbf{F}-\mathbf{F}_h\|_{(L^2(\Omega))^3} + \|(A-\tilde{A}_h)\mathbf{F}\|_{(L^2(\Omega))^3}\\ &+\|(A-\tilde{A}_h)A\mathbf{E}\|_{(L^2(\Omega))^3}\Big)\end{aligned}$$

holds with c independent of h, $\mathbf{E}$, and $\mathbf{F}$.

Proof. We start with a slight modification of the proof of Theorem 10.9 of [30]. From that theorem we know that if $h\in\Lambda$ is small enough, then $(I+\tilde{A}_h)$ is invertible with a uniformly bounded inverse as a map from $(L^2(\Omega))^3$ to $(L^2(\Omega))^3$ (because $\{A_h\}_{h\in\Lambda}$ is a collectively compact and pointwise convergent sequence of operators). Thus $\tilde{\mathbf{E}}_h$ is well defined.

Then

$$\mathbf{E}-\tilde{\mathbf{E}}_h = (I+\tilde{A}_h)^{-1}(\mathbf{F}-\mathbf{F}_h) + \left((I+A)^{-1}-(I+\tilde{A}_h)^{-1}\right)\mathbf{F}$$

and the following error estimate follows from the bound on $((I+A)^{-1}-(I+\tilde{A}_h)^{-1})$ in [30]:

$$\begin{aligned}\|\tilde{\mathbf{E}}_h-\mathbf{E}\|_{(L^2(\Omega))^3} \le c\Big(&\|\mathbf{F}-\mathbf{F}_h\|_{(L^2(\Omega))^3} + \|(A-\tilde{A}_h)\mathbf{F}\|_{(L^2(\Omega))^3}\\ &+ \|(A-\tilde{A}_h)A\mathbf{E}\|_{(L^2(\Omega))^3}\Big).\end{aligned} \tag{3.19}$$

∎

THEOREM 3.6. *Provided $h \in \Lambda$ is small enough, the discrete variational problem* (2.7) *has a unique solution $\tilde{\mathbf{E}}_h \in X_h$. Furthermore,*

$$\|\tilde{\mathbf{E}}_h - \mathbf{E}\|_X \le C \left(\inf_{\boldsymbol{\chi}_h \in X_h} \|\mathbf{F} - \boldsymbol{\chi}_h\|_X + \inf_{\boldsymbol{\phi}_h \in X_h} \|A\mathbf{F} - \boldsymbol{\phi}_h\|_X + \inf_{\boldsymbol{\psi}_h \in X_h} \|A\mathbf{E} - \boldsymbol{\psi}_h\|_X \right).$$

In general, $\tilde{\mathbf{E}}_h \to \mathbf{E}$ in X as $h \to 0$.

Proof. From the previous theorem, $\mathbf{E}_h$ is proved to exist uniquely, and it remains to estimate the error in X. From (3.12) and (3.15),

$$\begin{aligned} \|\mathbf{E} - \tilde{\mathbf{E}}_h\|_X &= \|A\mathbf{E} - \tilde{A}_h\tilde{\mathbf{E}}_h + \mathbf{F} - \mathbf{F}_h\|_X \\ &\le \|(A - \tilde{A}_h)\mathbf{E}\|_X + \|\tilde{A}_h(\mathbf{E} - \tilde{\mathbf{E}}_h)\|_X + \|\mathbf{F} - \mathbf{F}_h\|_X. \end{aligned}$$

However, using the definition of $\tilde{A}_h$,

$$\|\tilde{A}_h(\mathbf{E} - \tilde{\mathbf{E}}_h)\|_X \le C\|\mathbf{E} - \tilde{\mathbf{E}}_h\|_{(L^2(\Omega))^3}.$$

Hence, via the previous theorem (and using the fact that the $(L^2(\Omega))^3$ norm is bounded by the X norm),

$$\begin{aligned} \|\mathbf{E} - \tilde{\mathbf{E}}_h\|_X \le C \Big(&\|(A - \tilde{A}_h)\mathbf{E}\|_X + \|\mathbf{F} - \mathbf{F}_h\|_X \\ &+ \|(A - \tilde{A}_h)\mathbf{F}\|_{(L^2(\Omega))^3} + \|(A - \tilde{A}_h)A\mathbf{E}\|_{(L^2(\Omega))^3} \Big). \end{aligned}$$

Since $A\mathbf{E} = \mathbf{F} - \mathbf{E}$, this can be rewritten

$$\|\mathbf{E} - \tilde{\mathbf{E}}_h\|_X \le C \left(\|(A - \tilde{A}_h)\mathbf{E}\|_X + \|\mathbf{F} - \mathbf{F}_h\|_X + \|(A - \tilde{A}_h)\mathbf{F}\|_{(L^2(\Omega))^3} \right).$$

Now we can estimate each term. Via the estimates for the mixed method used previously, we get

$$\|\mathbf{F} - \mathbf{F}_h\|_X \le C \inf_{\boldsymbol{\chi}_h \in X_h} \|\mathbf{F} - \boldsymbol{\chi}_h\|_X.$$

Since $\mathbf{E} \in X_0$, the potential p_h in (3.16) vanishes when $\mathbf{F} = \mathbf{E}$. Therefore,

$$\|(A - \tilde{A}_h)\mathbf{E}\|_X \le C \inf_{\boldsymbol{\phi}_h \in X_h} \|A\mathbf{E} - \boldsymbol{\phi}_h\|_X.$$

In the same way, since $\mathbf{F} \in X_0$,

$$\|(A - \tilde{A}_h)\mathbf{F}\|_X \le C \inf_{\boldsymbol{\xi}_h \in X_h} \|A\mathbf{F} - \boldsymbol{\xi}_h\|_X.$$

The convergence result now follows from a density argument. ∎

This result can be made more specific provided the solution is regular enough. Let

$$\begin{aligned} H^s(\text{curl};\, \Omega) = \big\{ &\mathbf{u} \in (H^s(\Omega))^3 \mid \nabla \times \mathbf{u} \in (H^s(\Omega))^3,\ \boldsymbol{\nu} \times \mathbf{u} \in (H^s(f))^3 \\ &\text{for each face } f \text{ of } \Sigma \big\} \end{aligned}$$

for some $s \ge 0$ with norm

$$\|\mathbf{u}\|^2_{H^s(\text{curl};\, \Omega)} := \|\mathbf{u}\|^2_{H^s(\Omega)} + \|\nabla \times \mathbf{u}\|^2_{H^s(\Omega)} + \sum_{f \in \Sigma} \|\boldsymbol{\nu} \times \mathbf{u}\|^2_{H^s(f)}.$$

Then, the error estimate of Theorem 3.6 can be written as shown below.

COROLLARY 3.7. *If* $\mathbf{F}, A\mathbf{F}, A\mathbf{E} \in H^s(\text{curl};\ \Omega)$ *for some* $s > \frac{1}{2}$, *then*

$$\|\mathbf{E} - \tilde{\mathbf{E}}_h\|_X \leq ch^{\min(s,\ell)},$$

where the norm on X *is given by* (1.6).

Remark 3. For a Lipschitz polyhedral domain, the best we can generally expect is that the above regularity requirements hold for some s with $\frac{1}{2} < s$ but possibly with s less than 1.

4. The fully discrete problem. The discretization we have considered up to this point is not optimal for implementation since $\mathcal{I}^R$ is expensive to compute. We prefer to use (2.6) in place of (2.7). Let us define $A_h : (L^2(\Omega))^3 \to (L^2(\Omega))^3$ such that if $\mathbf{f} \in (L^2(\Omega))^3$, then $A_h\mathbf{f} \in X_{0,h}$ satisfies

$$a(A_h\mathbf{f}, \boldsymbol{\xi}_h) = -2k^2 \int_\Omega \mathbf{f} \cdot \bar{\boldsymbol{\xi}}_h \, dV - \int_\Sigma T(\mathcal{I}_h(\mathbf{f})) \cdot \bar{\boldsymbol{\xi}}_{h,T} \, dA \quad \forall \boldsymbol{\xi}_h \in X_{0,h}. \tag{4.1}$$

Then (2.6) is equivalent to solving

$$(I + A_h)\mathbf{E}_h = \mathbf{F}_h. \tag{4.2}$$

In order to prove convergence we make a specific choice of $\widetilde{\mathbb{G}}_h$. We choose $\widetilde{\mathbb{G}}_h$ to interpolate $\widetilde{\mathbb{G}}$ on Ω (as a function of $\mathbf{y}$). Using this choice we can prove the following lemma.

LEMMA 4.1. *There is a constant* C *such that for any* $\mathbf{u} \in X$,

$$\|(A_h - \tilde{A}_h)\mathbf{u}\|_X \leq Ch^l \|\mathbf{u}\|_X.$$

Proof. By the definition of A_h and $\tilde{A}_h$,

$$a((A_h - \tilde{A}_h)\mathbf{u}, (A_h - \tilde{A}_h)\mathbf{u}) = -\int_\Sigma T(\mathcal{I}_h(\mathbf{u}) - \mathcal{I}^R(\mathbf{u})) \cdot \overline{(A_h - \tilde{A}_h)\mathbf{u}} \, dA.$$

Thus

$$\|(A_h - \tilde{A}_h)\mathbf{u}\|_X \leq C\|T(\mathcal{I}_h(\mathbf{u}) - \mathcal{I}^R(\mathbf{u}))\|_{(L^2(\Sigma))^3}.$$

However, for any derivative D_x with respect to $\mathbf{x}$, for any $\mathbf{x} \in \Sigma$,

$$\begin{aligned} \left|D_x(\mathcal{I}_h(\mathbf{u}) - \mathcal{I}^R(\mathbf{u}))\right| &= \left| \int_\Omega (\nabla \times D_x(\widetilde{\mathbb{G}}_h - \widetilde{\mathbb{G}}))^T \nabla \times \mathbf{u} - k^2 D_x(\widetilde{\mathbb{G}}_h - \widetilde{\mathbb{G}})^T \mathbf{u} \, dV \right| \\ &\leq C\|D_x(\widetilde{\mathbb{G}}_h - \widetilde{\mathbb{G}})\|_X \|\mathbf{u}\|_X. \end{aligned}$$

But since $D_x\mathbb{G}$ is smooth when $\mathbf{x} \neq \mathbf{y}$, and $\widetilde{\mathbb{G}}_h$ interpolates $\widetilde{\mathbb{G}}$, we may use the interpolation estimate (2.3) to show that

$$\|D_x(\widetilde{\mathbb{G}}_h - \widetilde{\mathbb{G}})\|_X \leq Ch^l,$$

and we are done. ∎

Next we verify that $(I + \tilde{A}_h)$ is invertible as a map from X to X.

LEMMA 4.2. *The operator* $(I + \tilde{A}_h)$ *is invertible with a uniformly bounded inverse as a map from* X *to* X.

Proof. We have already seen that this lemma holds with $(L^2(\Omega))^3$ in place of X. Now let $\mathbf{u} \in (L^2(\Omega))^3$ solve

$$\mathbf{u} + \tilde{A}_h \mathbf{u} = \mathbf{F}$$

for some $\mathbf{F} \in X$; then since $\mathbf{u} = \mathbf{F} - \tilde{A}_h \mathbf{u} \in X$,

$$\begin{aligned} \|\mathbf{u}\|_X &\le \|\mathbf{F}\|_X + \|\tilde{A}_h \mathbf{u}\|_X \\ &\le \|\mathbf{F}\|_X + C\|\mathbf{u}\|_{(L^2(\Omega))^3} \\ &\le \|\mathbf{F}\|_X + C\|\mathbf{F}\|_{(L^2(\Omega))^3}. \end{aligned}$$

Thus

$$\|(I + \tilde{A}_h)^{-1}\mathbf{F}\|_X \le C\|\mathbf{F}\|_X,$$

and we are done. ∎

Now we can prove that (4.2) has a unique solution that is close to the solution $\tilde{\mathbf{E}}_h$ of (2.7).

THEOREM 4.3. *Provided h is sufficiently small,* (4.2) (*or equivalently,* (2.6)) *has a unique solution $\mathbf{E}_h \in X_h$, and if $\tilde{\mathbf{E}}_h$ is the solution of* (2.7)*, then*

$$\|\mathbf{E}_h - \tilde{\mathbf{E}}_h\|_X \le Ch^l \|\mathbf{E}_h\|_X.$$

Remark 4. As a result of this theorem we can conclude that $\mathbf{E}_h$ satisfies the error estimates in Theorem 3.6 and Corollary 3.7.

Proof. We have already verified (Lemma 4.2) that $(I + \tilde{A}_h)$ is invertible as a map from X to X and that the inverse is uniformly bounded. Since

$$\mathbf{E}_h + \tilde{A}_h \mathbf{E}_h + (A_h - \tilde{A}_h)\mathbf{E}_h = \mathbf{F}_h,$$

we have

$$(I + C_h)\mathbf{E}_h = (I + \tilde{A}_h)^{-1}\mathbf{F}_h,$$

where $C_h = (I + \tilde{A}_h)^{-1}(A_h - \tilde{A}_h)$ and hence, using Lemma 4.1, $\|C_h\|_{L(X,X)} \le Ch^l < 1$ for h sufficiently small. This implies that $(I + C_h)$ is invertible with bounded inverse in X, and hence $\mathbf{E}_h$ exists.

Furthermore,

$$(I + \tilde{A}_h)(\mathbf{E}_h - \tilde{\mathbf{E}}_h) = (\tilde{A}_h - A_h)\mathbf{E}_h$$

so that by using Lemma 4.1 and the boundedness of $(I + \tilde{A}_h)^{-1}$ we have

$$\|\mathbf{E}_h - \tilde{\mathbf{E}}_h\|_X \le C\|(\tilde{A}_h - A_h)\mathbf{E}_h\|_X \le Ch^l \|\mathbf{E}_h\|_X.$$

Thus we can conclude that Theorem 3.6 holds for $\mathbf{E}_h$. ∎

Now we shall show why (2.6) helps in the discretization of this problem. Let $\{\boldsymbol{\xi}_i\}_{i=1}^{N_h}$ be a basis for X_h. Usually this basis would be constructed using the degrees of freedom (2.2), but other choices are possible [32]. Then we can express $\mathbf{E}_h \in X_h$ as

$$\mathbf{E}_h = \sum_{i=1}^{N_h} E_i \boldsymbol{\xi}_i,$$

and we may write the variational equation (2.6) as a matrix equation. Let $\vec{E} = (E_1, \ldots, E_{N_h})^T$, and let S and L be $N_h \times N_n$ matrices with

$$S_{i,j} = \int_\Omega \nabla \times \boldsymbol{\xi}_j \cdot \nabla \times \bar{\boldsymbol{\xi}}_i - k^2 \boldsymbol{\xi}_j \cdot \bar{\boldsymbol{\xi}}_i \, dV - ik \int_\Sigma \boldsymbol{\xi}_{j,T} \cdot \bar{\boldsymbol{\xi}}_{i,T} \, dA,$$
$$L_{ij} = -\int_\Sigma T(\mathcal{I}_h(\boldsymbol{\xi}_j)) \cdot \bar{\boldsymbol{\xi}}_i \, dA.$$

Let $\vec{F}$ be the vector with

$$F_j = \int_\Sigma T(\mathbf{E}^i) \cdot \bar{\boldsymbol{\xi}}_j \, dA.$$

Then

$$(S + L)\vec{E} = \vec{F}. \tag{4.3}$$

Our analysis guarantees that $S + L$ is invertible for h sufficiently small, but $S + L$ is not particularly well structured from the point of view of numerical linear algebra. It is nondefinite and nonsymmetric.

The matrix S is somewhat better behaved than L. It is sparse and symmetric (but not Hermitian). It corresponds to the standard discretization of an interior boundary value problem for Maxwell's equations and is also invertible for h sufficiently small. In general, S has $O(N_h)$ nonzero entries.

If we choose $\widetilde{\mathbb{G}}_h$ to interpolate zero away from Γ, then $\mathcal{I}_h(\boldsymbol{\xi}_j)$ vanishes when $\boldsymbol{\xi}_j$ is zero on all tetrahedra sharing an edge with Γ. Thus $L_{ij} \neq 0$ only if $\boldsymbol{\xi}_i$ is associated with an edge or face on Σ and if $\boldsymbol{\xi}_j$ is associated with a tetrahedron touching Γ. For a quasi-uniform mesh, we expect $O(N_h^{\frac{2}{3}})$ edges and faces on Σ, and $O(N_h^{\frac{2}{3}})$ tetrahedra to touch Γ. Hence L has $O(N_h^{\frac{4}{3}})$ nonzero entries which is far more than S. Thus L is very expensive to compute and store. This suggests that (4.3) should be solved by an iterative technique (Bi-CGSTAB has worked well for us when applying similar methods to the Helmholtz equation) and then only the action of L needs to be computed. We expect that this can be computed rapidly using the fast multipole method [33] to yield a fast overall solver.

5. Conclusion. We have proved convergence of the combined finite element–integral equation technique under fairly general conditions on the scatterer and the auxiliary boundary. When the domain is well behaved so that the exact solution is regular, we can even obtain optimal order estimates.

The scheme results in a large dense submatrix in the overall matrix for the discrete problem. This suggests the necessity of using the fast multipole scheme to evaluate the integral operator $\mathcal{I}_h(\mathbf{f})$. We are now programming the combined scheme and hope to report numerical results and algorithmic details in the near future.

Appendix. In this appendix, we provide a proof of Lemma 2.1 and an extension theorem for functions defined on a parallelepiped (Lemma A.2). Let $\widetilde{\mathbb{G}}_h(\mathbf{x}, \cdot)$ denote the matrix function such that if $\tilde{\mathbf{g}}_{h,l}(\mathbf{x}, \cdot)$ is the lth column of $\widetilde{\mathbb{G}}_h(\mathbf{x}, \cdot)$ and $\mathbf{g}_l(\mathbf{x}, \cdot)$ is the lth column of $\mathbb{G}(\mathbf{x}, \cdot)$, then

(a) $\tilde{\mathbf{g}}_{h,l}(\mathbf{x}, \cdot) \in V_h, \quad 1 \leq l \leq 3$;
(b) $(\tilde{\mathbf{g}}_{h,l}(\mathbf{x}, \cdot))_T$ interpolates $(\mathbf{g}_l(\mathbf{x}, \cdot))_T$ on Γ (using edge and face freedoms (2.2a) and (2.2b));
(c) $(\tilde{\mathbf{g}}_{h,l}(\mathbf{x}, \cdot))_T = 0$ on all tetrahedra sharing an edge or face with Σ.

LEMMA A.1. *Suppose* (2.6) *has a unique solution for each given* $\widetilde{\mathbb{G}}_h$ *satisfying the properties assumed in section* 2. *Then the solution is independent of the choice of* $\widetilde{\mathbb{G}}_h$.

Proof. Suppose $\mathbf{E}_h^{(i)}$ is the solution of (2.6) corresponding to $\widetilde{\mathbb{G}}_h^{(i)}$, $i = 1, 2$. Let $\mathcal{I}_h^{(i)}$, $i = 1, 2$, denote the operator in (2.5) using $\widetilde{\mathbb{G}}_h^{(i)}$. Since solutions of the finite element formulation (2.6) are unique, it suffices to show that

$$\mathcal{I}_h^{(2)}(\mathbf{E}_h^{(1)}) = \mathcal{I}_h^{(1)}(\mathbf{E}_h^{(1)}).$$

By definition,

$$\begin{aligned}&\left(\mathcal{I}_h^{(2)}(\mathbf{E}_h^{(1)}) - \mathcal{I}_h^{(1)}(\mathbf{E}_h^{(1)})\right)^T \\ &\quad = \int_\Omega \left\{(\nabla \times \mathbf{E}_h^{(1)})^T \nabla \times (\widetilde{\mathbb{G}}_h^{(2)} - \widetilde{\mathbb{G}}_h^{(1)}) - k^2 (\mathbf{E}_h^{(1)})^T (\widetilde{\mathbb{G}}_h^{(2)} - \widetilde{\mathbb{G}}_h^{(1)})\right\} dV.\end{aligned}$$

Now the lth column of $\widetilde{\mathbb{G}}_h^{(2)} - \widetilde{\mathbb{G}}_h^{(1)}$ is

$$(\widetilde{\mathbb{G}}_h^{(2)} - \widetilde{\mathbb{G}}_h^{(1)})_l = \tilde{\mathbf{g}}_{h,l}^{(2)} - \tilde{\mathbf{g}}_{h,l}^{(1)},$$

and since $\mathbf{g}_{h,l}^{(j)}$, $j = 1, 2$, interpolates $\mathbf{g}_l$ on Γ, the tangential component of the difference vanishes there. Hence $\tilde{\mathbf{g}}_{h,l}^{(2)} - \tilde{\mathbf{g}}_{h,l}^{(1)} \in X_h$. Since $(\mathbf{g}_{h,i}^{(2)} - \mathbf{g}_{h,i}^{(1)})_T = 0$ on Σ we have, from the definition of $\mathbf{E}_h^{(1)}$ in (2.6) and using the test function $\boldsymbol{\phi}_h = \overline{(\tilde{\mathbf{g}}_{h,l}^{(2)} - \tilde{\mathbf{g}}_{h,l}^{(1)})}$,

$$\int_\Omega \left\{(\nabla \times \mathbf{E}_h^{(1)}) \cdot \nabla \times (\tilde{\mathbf{g}}_{h,l}^{(2)} - \tilde{\mathbf{g}}_{h,l}^{(1)}) - k^2 \mathbf{E}_h^{(1)} \cdot (\tilde{\mathbf{g}}_{h,l}^{(2)} - \tilde{\mathbf{g}}_{h,l}^{(1)})\right\} dV = 0.$$

This implies $\mathcal{I}_h^{(2)}(\mathbf{E}_h^{(1)}) = \mathcal{I}_h^{(1)}(\mathbf{E}_h^{(1)})$. Thus $\mathbf{E}_h^{(1)}$ satisfies (2.6) with $\mathcal{I}_h = \mathcal{I}_h^{(2)}$. The assumed uniqueness of $\mathbf{E}_h^{(2)}$ then implies $\mathbf{E}_h^{(2)} = \mathbf{E}_h^{(1)}$. ∎

LEMMA A.2. *Suppose* Σ *is the surface of a right parallelepiped, that there is a function* $u \in H^1(\Sigma)$ *such that on each face* f *of* Σ, $u \in H^{1+\epsilon}(\Sigma)$, $0 \le \epsilon < 1/2$, *and that* u *is continuous at the edges of the face of* Σ. *Then there is an extension* $u \in H^{3/2+\epsilon}(P)$, *where* P *denotes the interior of the parallelepiped.*

Remark 5. The proof is similar to that of Theorem 1.5.2.4 of [24].

Proof. By a partition of unity we need only consider the problem in the neighborhood of a corner. Suppose the corner is at $(0, 0, 0)$ and that the planes $x = 0$, $y = 0$, and $z = 0$ meet there. Furthermore, suppose we wish to extend u to the octant $x > 0$, $y > 0$, and $z > 0$.

Let $u = g_1$ on the quarter plane $\{0\} \times \mathbb{R}_+ \times \mathbb{R}_+$, $u = g_2$ on the quarter plane $\mathbb{R}_+ \times \{0\} \times \mathbb{R}_+$, and $u = g_3$ on the quarter plane $\mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$. We can assume g_1, g_2, and g_3 vanish outside a sphere of radius R around $(0, 0, 0)$ (because of the partition of unity).

Using the standard extension theorem we can extend g_1 to a function $\tilde{g}_1 \in H^{1+\epsilon}(\{0\} \times \mathbb{R} \times \mathbb{R})$. This can then be extended to a function $G_1 \in H^{3/2+\epsilon}(\mathbb{R}^3)$ such that $G_1|_{x=0} = \tilde{g}_1$.

Now consider $v = u - G_1$. Clearly, $v = 0$ on $\{0\} \times \mathbb{R}_+ \times \mathbb{R}_+$, $v = g_2 - G_1$ on $\mathbb{R}_+ \times \{0\} \times \mathbb{R}_+$, and $v = g_3 - G_1$ on $\mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$. In particular, $v = 0$ on the line

$x = 0$, $z = 0$, $y > 0$. Define a function w by

$$w(x,y) = \begin{cases} v(x,y,0) & \text{if } x > 0, y > 0, \\ -v(-x,y,0) & \text{if } x < 0, y > 0, \\ v(x,-y,0) & \text{if } x > 0, y < 0, \\ -v(-x,-y,0) & \text{if } x < 0, y < 0. \end{cases}$$

Clearly, w is continuous on the whole plane $z = 0$ and is antisymmetric in x. Since it is continuous we can conclude $w \in H^{1+\epsilon}(\mathbb{R} \times \mathbb{R} \times \{0\})$. Using this as Dirichlet data for solving Laplace's equation (with zero data on the hemisphere of radius R in the upper half-space) shows that there is a function $G_2 \in H^{3/2+\epsilon}(\mathbb{R} \times \mathbb{R} \times \mathbb{R}_+)$ such that $v = G_2$ on the plane $z = 0$. The odd symmetry of the data is maintained by the solution and so G_2 vanishes on the plane $x = 0$. Hence if $p = v - G_2$, then we can conclude $p = 0$ on $\{0\} \times \mathbb{R}_+ \times \mathbb{R}_+$, $p = g_2 - G_1 - G_2$ on $\mathbb{R}_+ \times \{0\} \times \mathbb{R}_+$, and $p = 0$ on $\mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$.

Now we again extend p to the plane $y = 0$ by reflection. Let q be defined for $y = 0$ by

$$q(x,z) = \begin{cases} p(x,0,z) & \text{if } x > 0, z > 0, \\ -p(-x,0,z) & \text{if } x < 0, z > 0, \\ -p(x,0,-z) & \text{if } x > 0, z < 0, \\ p(-x,0,-z) & \text{if } x < 0, z < 0. \end{cases}$$

Now $q \in H^{1+\epsilon}(\mathbb{R} \times \{0\} \times \mathbb{R})$ and is antisymmetric about the lines $y = z = 0$ and $y = x = 0$. Thus, again using this as boundary data for the Dirichlet problem for Laplace's equation (again with zero Dirichlet data on the hemisphere in the appropriate half-space), we can conclude that there is a function G_3 such that $p = G_3|_{y=0}$ and that G_3 vanishes on $x = 0$ and $z = 0$. Thus the function

$$u = G_1 + G_2 + G_3$$

is the required extension with the required smoothness. □

REFERENCES

[1] A. Alonso and A. Valli, *An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations*, Math. Comp., 68 (1999), pp. 607–631.

[2] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[3] D. Arnold, R. Falk, and R. Winthur, *Multigrid in h(div) and h(curl)*, Numer. Math., 85 (2000), pp. 197–217.

[4] F. Assous, P. Ciarlet Jr., and E. Sonnendrücker, *Resolution of the Maxwell equations in a domain with reentrant corners*, RAIRO Model. Math. Anal. Numér., 32 (1998), pp. 359–389.

[5] I. Babuška and A. Miller, *The post-processing approach in the finite element method – Part 1: Calculation of displacements, stresses and other higher derivatives of the displacements*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 1085–1109.

[6] I. Babuška and A. Miller, *The post-processing approach in the finite element method – Part 2: The calculation of stress intensity factors*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 1110–1129.

[7] I. Babuška and A. Miller, *The post-processing approach in the finite element method – Part 3: A posteriori error estimates and adaptive mesh selection*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 2311–2324.

[8] D. Boffi, *Fortin operators and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[9] D. Boffi, *A note on the de Rham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.
[10] D. Boffi, F. Brezzi, and L. Gastaldi, *On the convergence of eigenvalues for mixed formulations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 131–154.
[11] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (2000), pp. 121–140.
[12] J. Bramble, J. Pasciak, and J. Xu, *The analysis of multigrid algorithms with non-nested spaces or non-inherited quadratic forms*, Math. Comp., 56 (1991), pp. 1–34.
[13] S. Brenner and L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.
[14] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM. J. Numer. Anal., 38 (2000), pp. 580–607.
[15] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, New York, 1998.
[16] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.
[17] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276. Tech. report available online at http://www.maths.univ-rennes1.fr/ costabel/.
[18] J. Coyle and P. Monk, *The finite element approximation of scattering in a layered medium*, in Analytical and Computational Methods in Scattering and Applied Mathematics, F. Santosa and I. Stakgold, eds., Res. Notes Math. 417, Chapman & Hall/CRC, London, 2000, pp. 67–84.
[19] P.-M. Cutzach and C. Hazard, *Existence, uniqueness and analyticity properties for electromagnetic scattering in a two-layered medium*, Math. Methods Appl. Sci., 21 (1998), pp. 433–461.
[20] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.
[21] M. Dauge, M. Costabel, and D. Martin, *Numerical Investigation of a Boundary Penalization Method for Maxwell Equations*, Tech. report, University of Rennes, Rennes, France, 1999. Available online at http://www.maths.univ-rennes1.fr/˜dauge/core/index.html.
[22] F. Dubois, *Discrete vector potential representation of a divergence-free vector field in three-dimensional domains: Numerical analysis of a model problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1103–1141.
[23] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.
[24] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.
[25] C. Hazard and M. Lenoir, *On the solution of time-harmonic scattering problems for Maxwell's equations*, SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.
[26] P. Joly, C. Poirier, J.-E. Roberts, and P. Trouve, *A new nonconforming finite element method for the computation of electromagnetic guided waves.* I. *Mathematical analysis,* SIAM J. Numer. Anal., 33 (1996), pp. 1494–1525.
[27] F. Kikuchi, *An isomorphic property of two Hilbert spaces appearing in electromagnetism: Analysis by the mixed formulation*, Japan J. Appl. Math., 3 (1986), pp. 53–58.
[28] F. Kikuchi, *Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism*, Comput. Methods Appl. Mech. Engrg., 64 (1987), pp. 509–521.
[29] F. Kikuchi, *On a discrete compactness property for the Nedelec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. 1A Math., 36 (1989), pp. 479–490.
[30] R. Kress, *Linear Integral Equations*, 2nd ed., Springer-Verlag, Berlin, 1999.
[31] J. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[32] J. Nédélec, *Eléments finis mixtes incompressibles pour l'équation de Stokes dans* $\mathbb{R}^3$, Numer. Math., 39 (1982), pp. 97–112.
[33] V. Rokhlin, *Rapid solution of integral equations of scattering theory in two dimensions*, J. Comput. Phys., 86 (1990), pp. 414–439.
[34] L. Vardapetyan and L. Demkowicz, *hp-adaptive finite elements in electromagnetics*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 331–344.
[35] J. Wheeler, *Permafrost thermal design for the trans-Alaska pipeline*, in Moving Boundary Problems, D. Wilson, A. Solomon, and P. Boggs, eds., Academic Press, New York, 1978, pp. 267–284.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 220–232

EFFICIENT COMPUTATION OF SENSITIVITIES FOR ORDINARY DIFFERENTIAL EQUATION BOUNDARY VALUE PROBLEMS*

RADU SERBAN† AND LINDA R. PETZOLD‡

Abstract. For models described by ordinary differential equation boundary value problems (ODE BVPs), we derive adjoint equations for sensitivity analysis, giving explicit forms for the boundary conditions of the adjoint boundary value problem. The solutions of the adjoint equations are used to efficiently compute gradients of both integral-form and pointwise constraints. Existence and stability results are given for the adjoint system and its numerical solution. The use of the method is demonstrated for a simple example, where it is seen that the method is particularly advantageous for problems with more than a few parameters.

Key words. sensitivity analysis, ODE boundary value problem, adjoint method

AMS subject classifications. 65L10, 65L99

PII. S0036142900376870

1. Introduction.

Sensitivity analysis generates essential information for design optimization, parameter estimation, optimal control, data assimilation, process sensitivity, and experimental design. Virtually any scientific or engineering problem can take advantage of sensitivity analysis, for example, problems in chemical engineering applications, multibody mechanical systems, structural engineering, materials science, electric and electronic circuit simulation, and weather prediction models.

There is a large body of work on methods and software for forward sensitivity analysis of initial value problems (IVPs) for ordinary differential equation (ODE) systems [16] and differential-algebraic equation (DAE) systems [11, 22]. Recent research [21, 25] has demonstrated that forward sensitivities can be computed reliably and efficiently via automatic differentiation [4] in combination with ODE/DAE/PDE solution techniques designed to exploit the structure of the sensitivity system.

Forward sensitivity analysis has been shown to be very efficient for problems in which the sensitivities of a (potentially) very large number of quantities, with respect to relatively few parameters, are needed. However, for problems where the number of uncertain parameters is large, the forward sensitivity method becomes computationally intractable. The adjoint (reverse) method is advantageous in the complementary situation, where the sensitivities of a few quantities, with respect to a large number of parameters, are needed. Adjoint sensitivity analysis is particularly attractive for boundary value problems (BVPs). In contrast to the situation for IVPs, where the adjoint method requires considerable memory resources above what is required for the solution of the original problem, the solution values required by

*Received by the editors August 16, 2000; accepted for publication (in revised form) December 19, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/37687.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 (radu@llnl.gov). The research of this author was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under contract W-7405-Eng-48.

‡Department of Computer Science, University of California, Santa Barbara, CA 93106 (petzold@engineering.ucsb.edu). The research of this author was partially supported by NSF ACI-0086061, DOE DE-FG03-00ER25430, NSF KDI ATM-9873133, NSF-ARPA PC 239415, and EPRI WO-8333-06.

the adjoint method for BVPs are naturally available from the solution of the original problem.

Adjoint sensitivity analysis raises a set of entirely new issues ranging from existence of adjoint operators [3, 5, 6, 20] to construction of adjoint models [10, 17, 18], derivation of boundary conditions for the adjoint states [5], and algorithm implementation [13]. Adjoint sensitivity analysis for BVPs has focused mainly on models described by partial differential equations (PDEs). We cite here the work of Cacuci [5, 6] on general sensitivity theory for nonlinear systems, that of Ghione and Filicori on sensitivity of semiconductor devices [12], the work of Giles and Pierce [14] on adjoint equations in computational fluid dynamics, and that of Machiels, Maclay, and Patera [23, 24] on the use of adjoint methods to obtain a posteriori finite element output bounds.

In [5, 6, 26], adjoint operators are constructed for general nonlinear systems, and results are given for solvability of the original and adjoint systems. However, because of the generality of this setting, boundary conditions for the adjoint problem cannot be explicitly constructed. Instead, for each particular example, proper boundary conditions are obtained by imposing the condition that the Lagrange identity is satisfied. For the adjoint equations for inviscid and viscous compressible flow, Giles and Pierce [14] constructed correct boundary conditions for adjoint problems used in evaluating integral-form quantities. In computational fluid dynamics, most quantities of interest are in integral form. However, in other engineering areas, point quantities such as maximum stresses and/or deformations in structural analysis are of major concern. Being able to efficiently compute gradients of such quantities is thus of high interest.

In the present work we derive adjoint equations for sensitivity analysis of models described by ODE BVPs. For a general form ODE BVP, which is assumed to be well conditioned and to have a unique solution, we derive in section 2 adjoint systems for which we explicitly construct proper boundary conditions. Our goals are to demonstrate that the adjoint method offers an efficient means of computing ODE BVP sensitivities, particularly if there are many parameters, and to show how the adjoint method is formulated for ODE BVPs for different classes of derived functions. Thus we derive adjoint equations to efficiently evaluate not only gradients of integral-form quantities, but also (using the Leibnitz integral rule) gradients of pointwise constraints. In section 3 we establish that, for the problems considered here, the adjoint problems are well-posed and inherit the stability of the original system. We show that numerical stability of the midpoint method for the original system implies numerical stability for the adjoint system. In section 4 we illustrate the computation of sensitivities via the adjoint method on a simple example.

2. Derivation of the adjoint BVP. Consider a state vector $\mathbf{x} \in R^N$ that satisfies the BVP depending on parameters $\mathbf{p} \in R^{N_p}$,

$$\begin{aligned} \mathbf{F}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{p}, t) &= \mathbf{0}, \\ \mathbf{h}(\mathbf{x}(a), \mathbf{x}(b), \mathbf{p}) &= \mathbf{0}, \end{aligned} \tag{1}$$

and the function $g(\mathbf{x}, \mathbf{p}, t)$ whose gradient with respect to $\mathbf{p}$, $dg/d\mathbf{p}$ is to be evaluated at some time $\tau \in [a, b]$. We assume that the Jacobian of $\mathbf{F}$ with respect to the vector $\dot{\mathbf{x}}$ is nonsingular (meaning that (1) represents a system of ODEs and not DAEs) and that $\mathbf{h}$ represents a set of N independent equations. Note that if g also depends on the time derivatives $\dot{\mathbf{x}}$, then the first set of relations (1) can be used to express g as a function of only $\mathbf{x}$, $\mathbf{p}$, and t. We assume also that (1) is well conditioned and has a unique solution.

In section 2.1 we derive the gradient $(dg/d\mathbf{p})$ at $\tau \in (a,b)$. Using the resulting expressions, the particular case of computing $(dg/d\mathbf{p})$ at $\tau = b$ is analyzed in section 2.2.

2.1. Gradients of g between the integration bounds. We start by deriving the gradient $(dg/d\mathbf{p})$ at some $\tau \in (a,b)$. The derivation closely follows the IVP case [7], with differences arising from the definition of proper boundary conditions for the adjoint equations.

First, define the function

$$G^\tau(\mathbf{p}) = \int_a^\tau g(\mathbf{x},\mathbf{p},t)dt.$$

The gradient of G^τ with respect to $\mathbf{p}$ is then simply

$$\frac{dG^\tau}{d\mathbf{p}} = \int_a^\tau \frac{dg}{d\mathbf{p}}(\mathbf{x},\mathbf{p},t)dt = \int_a^\tau \left(g_\mathbf{p} + g_\mathbf{x}\mathbf{x}_\mathbf{p}\right) dt, \tag{2}$$

where subscripts represent partial differentiation. Applying the Leibnitz integral rule we obtain

$$\frac{d}{d\tau}\frac{dG^\tau}{d\mathbf{p}} = \left.\frac{dg}{d\mathbf{p}}\right|_\tau.$$

Thus $(dg/d\mathbf{p})|_\tau$ can be computed as

$$\left.\frac{dg}{d\mathbf{p}}\right|_\tau = \frac{d}{d\tau}\left(\int_a^\tau \left(g_\mathbf{p} + g_\mathbf{x}\mathbf{x}_\mathbf{p}\right) dt\right).$$

The challenge of adjoint sensitivity analysis is now to compute the above quantity without solving for the sensitivities $\mathbf{x}_\mathbf{p}$. To do this, we first form the linear sensitivity problem from the BVP (1),

$$\begin{aligned} \mathbf{F}_{\dot{\mathbf{x}}}\dot{\mathbf{x}}_\mathbf{p} + \mathbf{F}_\mathbf{x}\mathbf{x}_\mathbf{p} + \mathbf{F}_\mathbf{p} &= \mathbf{0}, \\ \mathbf{A}\mathbf{x}_\mathbf{p}(a) + \mathbf{B}\mathbf{x}_\mathbf{p}(b) + \mathbf{h}_\mathbf{p} &= \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{A} = \mathbf{h}_{\mathbf{x}_0}(\mathbf{x}(a),\mathbf{x}(b),\mathbf{p})$ and $\mathbf{B} = \mathbf{h}_{\mathbf{x}_1}(\mathbf{x}(a),\mathbf{x}(b),\mathbf{p})$. Then, for arbitrary $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in R^N$, the following holds:

$$0 \equiv \int_a^\tau \boldsymbol{\lambda}_1^* \left(\mathbf{F}_{\dot{\mathbf{x}}}\dot{\mathbf{x}}_\mathbf{p} + \mathbf{F}_\mathbf{x}\mathbf{x}_\mathbf{p} + \mathbf{F}_\mathbf{p}\right) dt + \int_\tau^b \boldsymbol{\lambda}_2^* \left(\mathbf{F}_{\dot{\mathbf{x}}}\dot{\mathbf{x}}_\mathbf{p} + \mathbf{F}_\mathbf{x}\mathbf{x}_\mathbf{p} + \mathbf{F}_\mathbf{p}\right) dt, \tag{4}$$

where * indicates the transposed conjugate. Integrating by parts, the first term in the first integral in the above relation becomes

$$\int_a^\tau \boldsymbol{\lambda}_1^*\mathbf{F}_{\dot{\mathbf{x}}}\dot{\mathbf{x}}_\mathbf{p}dt = \left(\boldsymbol{\lambda}_1^*\mathbf{F}_{\dot{\mathbf{x}}}\mathbf{x}_\mathbf{p}\right)|_a^\tau - \int_a^\tau \left(\dot{\boldsymbol{\lambda}}_1^*\mathbf{F}_{\dot{\mathbf{x}}} + \boldsymbol{\lambda}_1^*\frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right)\mathbf{x}_\mathbf{p}dt,$$

where

$$\boldsymbol{\lambda}_1^*\frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt} = \left[\left(\frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right)^*\boldsymbol{\lambda}_1\right]^* = \left(\mathbf{F}_{\dot{\mathbf{x}}}^*\bar{\boldsymbol{\lambda}}_1\right)_t^* + \left[\left(\mathbf{F}_{\dot{\mathbf{x}}}^*\bar{\boldsymbol{\lambda}}_1\right)_\mathbf{x}\dot{\mathbf{x}}\right]^* + \left[\left(\mathbf{F}_{\dot{\mathbf{x}}}^*\bar{\boldsymbol{\lambda}}_1\right)_{\dot{\mathbf{x}}}\ddot{\mathbf{x}}\right]^*. \tag{5}$$

A bar over a variable indicates that the variable is held fixed for the purpose of the current differentiation. Without loss of generality, we can assume that $\mathbf{F}$ depends

linearly on $\dot{\mathbf{x}}$ and that therefore, the last term in (5) is zero. Indeed, any other case can be reduced to this one by introducing the additional variables $\mathbf{y} = \dot{\mathbf{x}}$. So from now on, we calculate $\boldsymbol{\lambda}_1^* \left(d\mathbf{F}_{\dot{\mathbf{x}}}/dt\right)$ by

$$\boldsymbol{\lambda}_1^* \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt} = \left(\mathbf{F}_{\dot{\mathbf{x}}}^* \bar{\boldsymbol{\lambda}}_1\right)_t^* + \left[\left(\mathbf{F}_{\dot{\mathbf{x}}}^* \bar{\boldsymbol{\lambda}}_1\right)_{\mathbf{x}} \dot{\mathbf{x}}\right]^*.$$

Thus we have from (4) that

$$\begin{aligned} 0 \equiv\ & \left(\boldsymbol{\lambda}_1^* \mathbf{F}_{\dot{\mathbf{x}}} \mathbf{x}_{\mathbf{p}}\right)\big|_a^\tau + \left(\boldsymbol{\lambda}_2^* \mathbf{F}_{\dot{\mathbf{x}}} \mathbf{x}_{\mathbf{p}}\right)\big|_\tau^b - \int_a^\tau \left(\dot{\boldsymbol{\lambda}}_1^* \mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\lambda}_1^* \mathbf{F}_{\mathbf{x}} + \boldsymbol{\lambda}_1^* \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) \mathbf{x}_{\mathbf{p}} dt \\ & - \int_\tau^b \left(\dot{\boldsymbol{\lambda}}_2^* \mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\lambda}_2^* \mathbf{F}_{\mathbf{x}} + \boldsymbol{\lambda}_2^* \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) \mathbf{x}_{\mathbf{p}} dt + \int_a^\tau \boldsymbol{\lambda}_1^* \mathbf{F}_{\mathbf{p}} dt + \int_\tau^b \boldsymbol{\lambda}_2^* \mathbf{F}_{\mathbf{p}} dt. \end{aligned} \tag{6}$$

A suitable choice for $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ to compute $dG^\tau/d\mathbf{p}$ is given by the following.

PROPOSITION 1. *If $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ satisfy*

$$\begin{gathered} \dot{\boldsymbol{\lambda}}_1^* \mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\lambda}_1^* \left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) = g_{\mathbf{x}}, \\ \dot{\boldsymbol{\lambda}}_2^* \mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\lambda}_2^* \left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) = \mathbf{0}, \\ \bar{\mathbf{A}} \mathbf{F}_{\dot{\mathbf{x}}}^*(a) \boldsymbol{\lambda}_1(a) + \bar{\mathbf{B}} \mathbf{F}_{\dot{\mathbf{x}}}^*(b) \boldsymbol{\lambda}_2(b) = \mathbf{0}, \\ \boldsymbol{\lambda}_1(\tau) - \boldsymbol{\lambda}_2(\tau) = \mathbf{0}, \end{gathered} \tag{7}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are such that

$$\textit{span} \begin{bmatrix} \bar{\mathbf{A}}^T \\ \bar{\mathbf{B}}^T \end{bmatrix} = \textit{null}\left[-\mathbf{A}|\mathbf{B}\right], \tag{8}$$

that is, the rows of $[\bar{\mathbf{A}}|\bar{\mathbf{B}}]$ span the null space of $[-\mathbf{A}|\mathbf{B}]$, then

$$\frac{dG^\tau}{d\mathbf{p}} = -\boldsymbol{\alpha}^* \mathbf{h}_{\mathbf{p}} + \int_a^\tau \left(g_{\mathbf{p}} + \boldsymbol{\lambda}_1^* \mathbf{F}_{\mathbf{p}}\right) dt + \int_\tau^b \boldsymbol{\lambda}_2^* \mathbf{F}_{\mathbf{p}} dt, \tag{9}$$

where $\boldsymbol{\alpha} = \left(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*\right)^{-1} \left(-\mathbf{A}(\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_1)(a) + \mathbf{B}(\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_2)(b)\right)$.

Proof. First note that the requirement that the boundary conditions in (1) represent N linearly independent equations is equivalent to $[\mathbf{A}|\mathbf{B}]$ (as well as $[-\mathbf{A}|\mathbf{B}]$) having full row rank. As a consequence, the matrix $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ is invertible. The definition (8) of $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ implies that the rows of $[-\mathbf{A}|\mathbf{B}]$ span the null space of $[\bar{\mathbf{A}}|\bar{\mathbf{B}}]$. On the other hand, the third relation in (7) imposes that the vector $[(\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_1)(a), (\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_2)(b)]$ is in the null space of $[\bar{\mathbf{A}}|\bar{\mathbf{B}}]$. Therefore, there exists a vector $\boldsymbol{\alpha} \in R^N$ such that

$$\begin{aligned} (\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_1)(a) &= -\mathbf{A}^* \boldsymbol{\alpha}, \\ (\mathbf{F}_{\dot{\mathbf{x}}}^* \boldsymbol{\lambda}_2)(b) &= \mathbf{B}^* \boldsymbol{\alpha}, \end{aligned} \tag{10}$$

and thus

$$\begin{gathered} \left(\boldsymbol{\lambda}_1^* \mathbf{F}_{\dot{\mathbf{x}}} \mathbf{x}_{\mathbf{p}}\right)\big|_a^\tau + \left(\boldsymbol{\lambda}_2^* \mathbf{F}_{\dot{\mathbf{x}}} \mathbf{x}_{\mathbf{p}}\right)\big|_\tau^b = \left(\boldsymbol{\lambda}_1(\tau) - \boldsymbol{\lambda}_2(\tau)\right)^* \left(\mathbf{F}_{\dot{\mathbf{x}}} \mathbf{x}_{\mathbf{p}}\right)(\tau) \\ - \left(\boldsymbol{\lambda}_1^* \mathbf{F}_{\dot{\mathbf{x}}}\right)(a) \mathbf{x}_{\mathbf{p}}(a) + \left(\boldsymbol{\lambda}_2^* \mathbf{F}_{\dot{\mathbf{x}}}\right)(b) \mathbf{x}_{\mathbf{p}}(b) = \boldsymbol{\alpha}^* \left(\mathbf{A}\mathbf{x}_{\mathbf{p}}(a) + \mathbf{B}\mathbf{x}_{\mathbf{p}}(b)\right) = -\boldsymbol{\alpha}^* \mathbf{h}_{\mathbf{p}}, \end{gathered}$$

where the second relation in (3) and the last relation in (7) have been used. Since $[\mathbf{A}|\mathbf{B}]$ has full row rank, the $N \times N$ matrix $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ is nonsingular, and $\boldsymbol{\alpha}$ can be computed from (10) as

$$\boldsymbol{\alpha} = (\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)^{-1} \left(-\mathbf{A}(\mathbf{F}^*_{\dot{\mathbf{x}}}\boldsymbol{\lambda}_1)(a) + \mathbf{B}(\mathbf{F}^*_{\dot{\mathbf{x}}}\boldsymbol{\lambda}_2)(b)\right).$$

Substituting this result together with the first two relations of (7) into (6), we have

$$0 = -\boldsymbol{\alpha}^*\mathbf{h}_{\mathbf{p}} - \int_a^\tau g_{\mathbf{x}}\mathbf{x}_{\mathbf{p}}dt + \int_a^\tau \boldsymbol{\lambda}_1^*\mathbf{F}_{\mathbf{p}}dt + \int_\tau^b \boldsymbol{\lambda}_2^*\mathbf{F}_{\mathbf{p}}dt$$

and therefore

$$\frac{dG^\tau}{d\mathbf{p}} = \int_a^\tau \left(g_{\mathbf{p}} + g_{\mathbf{x}}\mathbf{x}_{\mathbf{p}}\right)dt = -\boldsymbol{\alpha}^*\mathbf{h}_{\mathbf{p}} + \int_a^\tau \left(g_{\mathbf{p}} + \boldsymbol{\lambda}_1^*\mathbf{F}_{\mathbf{p}}\right)dt + \int_\tau^b \boldsymbol{\lambda}_2^*\mathbf{F}_{\mathbf{p}}dt. \qquad \square$$

Returning to the problem of computing $dg/d\mathbf{p}$ at τ, by taking the total derivative with respect to τ in (9) we obtain

$$\frac{d}{d\tau}\frac{dG^\tau}{d\mathbf{p}} = \frac{d}{d\tau}\left(-\boldsymbol{\alpha}^*\mathbf{h}_{\mathbf{p}} + \int_a^\tau \left(g_{\mathbf{p}} + \boldsymbol{\lambda}_1^*\mathbf{F}_{\mathbf{p}}\right)dt + \int_\tau^b \boldsymbol{\lambda}_2^*\mathbf{F}_{\mathbf{p}}dt\right)$$

and therefore

$$\left.\frac{dg}{d\mathbf{p}}\right|_\tau = -\boldsymbol{\alpha}_\tau^T\mathbf{h}_{\mathbf{p}} + g_{\mathbf{p}}(\tau) + \int_a^\tau \boldsymbol{\mu}_1^*\mathbf{F}_{\mathbf{p}}dt + \int_\tau^b \boldsymbol{\mu}_2^*\mathbf{F}_{\mathbf{p}}dt, \tag{11}$$

where we have used $\boldsymbol{\lambda}_1(\tau) = \boldsymbol{\lambda}_2(\tau)$. The quantities $\boldsymbol{\mu}_1 = (\boldsymbol{\lambda}_1)_\tau$ and $\boldsymbol{\mu}_2 = (\boldsymbol{\lambda}_2)_\tau$ are the solution of the following sensitivity system, obtained by direct differentiation of (7):

$$\begin{aligned}
\dot{\boldsymbol{\mu}}_1^*\mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\mu}_1^*\left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) &= \mathbf{0},\\
\dot{\boldsymbol{\mu}}_2^*\mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\mu}_2^*\left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) &= \mathbf{0},\\
\bar{\mathbf{A}}\mathbf{F}^*_{\dot{\mathbf{x}}}(a)\boldsymbol{\mu}_1(a) + \bar{\mathbf{B}}\mathbf{F}^*_{\dot{\mathbf{x}}}(b)\boldsymbol{\mu}_2(b) &= \mathbf{0},\\
\boldsymbol{\mu}_1(\tau) + \dot{\boldsymbol{\lambda}}_1(\tau) - \boldsymbol{\mu}_2(\tau) - \dot{\boldsymbol{\lambda}}_2(\tau) &= \mathbf{0}.
\end{aligned}$$

The last boundary condition is obtained by taking the total derivative with respect to τ of the boundary condition $\boldsymbol{\lambda}_1(\tau) - \boldsymbol{\lambda}_2(\tau) = 0$ and taking into account all dependencies on τ. These can be better seen if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are considered as functions of two arguments, $\boldsymbol{\lambda}_1(t,\tau)$ and $\boldsymbol{\lambda}_2(t,\tau)$. In this case, direct differentiation of

$$\boldsymbol{\lambda}(t,\tau)|_{t=\tau} - \boldsymbol{\lambda}(t,\tau)|_{t=\tau} = \mathbf{0}$$

yields

$$\boldsymbol{\lambda}_{1t}(\tau,\tau) + \boldsymbol{\lambda}_{1\tau}(\tau,\tau) - \boldsymbol{\lambda}_{2t}(\tau,\tau) - \boldsymbol{\lambda}_{2\tau}(\tau,\tau) = 0,$$

that is,

$$\dot{\boldsymbol{\lambda}}_1(\tau) + \boldsymbol{\mu}_1(\tau) - \dot{\boldsymbol{\lambda}}_2(\tau) + \boldsymbol{\mu}_2(\tau) = 0.$$

Note that, upon substitution of $\dot{\boldsymbol{\lambda}}_1(\tau)$ and $\dot{\boldsymbol{\lambda}}_2(\tau)$, this boundary condition can be further simplified to

$$\boldsymbol{\mu}_1(\tau) - \boldsymbol{\mu}_2(\tau) + \left(g_{\mathbf{x}}\mathbf{F}_{\dot{\mathbf{x}}}^{-1}\right)^*(\tau) = \mathbf{0}.$$

The quantity $\boldsymbol{\alpha}_\tau$ is obtained by taking the total derivative of $\boldsymbol{\alpha}$ with respect to τ:

$$\boldsymbol{\alpha}_\tau = \left(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*\right)^{-1}\left(-\mathbf{A}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\mu}_1)(a) + \mathbf{B}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\mu}_2)(b)\right).$$

2.2. Gradients of g at the integration bounds. The gradient of $G^b = \int_a^b g(\mathbf{x},\mathbf{p},t)dt$ can be derived by applying a similar procedure, leading to

$$\frac{dG^b}{d\mathbf{p}} = -\boldsymbol{\alpha}^*\mathbf{h}_{\mathbf{p}} + \int_a^b \left(g_{\mathbf{p}} + \boldsymbol{\lambda}^*\mathbf{f}_{\mathbf{p}}\right)dt, \tag{12}$$

where $\boldsymbol{\lambda}$ is the solution of

$$\dot{\boldsymbol{\lambda}}^*\mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\lambda}^*\left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) = g_{\mathbf{x}},$$
$$\bar{\mathbf{A}}\mathbf{F}_{\dot{\mathbf{x}}}^*(a)\boldsymbol{\lambda}(a) + \bar{\mathbf{B}}\mathbf{F}_{\dot{\mathbf{x}}}^*(b)\boldsymbol{\lambda}(b) = \mathbf{0},$$

and $\boldsymbol{\alpha} = \left(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*\right)^{-1}\left(-\mathbf{A}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\lambda})(a) + \mathbf{B}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\lambda})(b)\right)$.

The gradient of g at $t = b$ could, in principle, be obtained as in the previous section by taking the total derivative of (12) with respect to b. Such an approach would be considerably complicated by the fact that g now depends on b implicitly through x. However, if we take

$$\left.\frac{dg}{d\mathbf{p}}\right|_b = \lim_{\tau\to b}\left.\frac{dg}{d\mathbf{p}}\right|_\tau,$$

then these difficulties can be circumvented. Indeed, if we specify $\tau = b$ in (11), we obtain

$$\left.\frac{dg}{d\mathbf{p}}\right|_b = -\boldsymbol{\alpha}_b^*\mathbf{h}_{\mathbf{p}} + g_{\mathbf{p}}(b) + \int_a^b \boldsymbol{\mu}^*\mathbf{F}_{\mathbf{p}}dt, \tag{13}$$

where $\boldsymbol{\mu} = \boldsymbol{\lambda}_b$ is the solution of

$$\dot{\boldsymbol{\mu}}^*\mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\mu}^*\left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) = \mathbf{0},$$
$$\bar{\mathbf{A}}\mathbf{F}_{\dot{\mathbf{x}}}^*(a)\boldsymbol{\mu}(a) + \bar{\mathbf{B}}\mathbf{F}_{\dot{\mathbf{x}}}^*(b)\left(\boldsymbol{\mu}(b) + \left(g_{\mathbf{x}}\mathbf{F}_{\dot{\mathbf{x}}}^{-1}\right)^*(b)\right) = \mathbf{0},$$

or, rearranging the boundary condition, is the solution of

$$\dot{\boldsymbol{\mu}}^*\mathbf{F}_{\dot{\mathbf{x}}} - \boldsymbol{\mu}^*\left(\mathbf{F}_{\mathbf{x}} - \frac{d\mathbf{F}_{\dot{\mathbf{x}}}}{dt}\right) = \mathbf{0},$$
$$\bar{\mathbf{A}}\mathbf{F}_{\dot{\mathbf{x}}}^*(a)\boldsymbol{\mu}(a) + \bar{\mathbf{B}}\mathbf{F}_{\dot{\mathbf{x}}}^*(b)\boldsymbol{\mu}(b) = -\bar{\mathbf{B}}g_{\dot{\mathbf{x}}}^*(b).$$

The expression of $\boldsymbol{\alpha}_b$ in (13) can be derived as

$$\boldsymbol{\alpha}_b = \left(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*\right)^{-1}\left(-\mathbf{A}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\mu})(a) + \mathbf{B}(\mathbf{F}_{\dot{\mathbf{x}}}^*\boldsymbol{\mu} + g_{\dot{\mathbf{x}}}^*)(b)\right).$$

3. On existence and stability of the adjoint solution. Consider a linear implicit ODE BVP of the form (3), written here as

$$\begin{aligned} \mathbf{M}(t)\dot{\mathbf{x}} + \mathbf{K}(t)\mathbf{x} + \mathbf{q}(t) &= \mathbf{0}, \\ \mathbf{A}\mathbf{x}(a) + \mathbf{B}x(b) + \mathbf{c} &= \mathbf{0}, \end{aligned} \tag{14}$$

whose adjoint BVP can be written as

$$\begin{aligned} \frac{d}{dt}\left(\mathbf{M}^*(t)\boldsymbol{\lambda}\right) - \mathbf{K}^*(t)\boldsymbol{\lambda} + \mathbf{r}(t) &= \mathbf{0}, \\ \bar{\mathbf{A}}\mathbf{M}^*(a)\boldsymbol{\lambda}(a) + \bar{\mathbf{B}}\mathbf{M}^*(b)\boldsymbol{\lambda}(b) &= \mathbf{0}, \end{aligned} \tag{15}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are defined as before.

In this section we investigate the stability of the adjoint system. More precisely, if the original system is stable, will the adjoint system also be stable? If we consider the adjoint system (15), the answer may be negative. Indeed, consider the following IVP example [7]:

$$e^t\dot{x} + \frac{1}{2}e^t x = 0, \tag{16}$$

with some initial condition at $t = a$. This system is equivalent to

$$\dot{x} + \frac{1}{2}x = 0,$$

so it is stable to integration from the left. However, the adjoint system (15) for (16) is

$$e^t\dot{\lambda} - \frac{1}{2}e^t\lambda + e^t\lambda = 0 \quad \Rightarrow \quad \dot{\lambda} + \frac{1}{2}\lambda = 0. \tag{17}$$

Note that the adjoint system must be solved in a backward direction. Thus the adjoint system (17) is unstable.

Denoting $\bar{\boldsymbol{\lambda}}(t) = \mathbf{M}^*(t)\boldsymbol{\lambda}(t)$, we can form the *augmented adjoint system* for (15),

$$\begin{aligned} \dot{\bar{\boldsymbol{\lambda}}} - \mathbf{K}^*(t)\boldsymbol{\lambda} + \mathbf{r}(t) &= \mathbf{0}, \\ \bar{\boldsymbol{\lambda}} - \mathbf{M}^*(t)\boldsymbol{\lambda} &= \mathbf{0}, \\ \bar{\mathbf{A}}\mathbf{M}^*(a)\boldsymbol{\lambda}(a) + \bar{\mathbf{B}}\mathbf{M}^*(b)\boldsymbol{\lambda}(b) &= \mathbf{0}. \end{aligned} \tag{18}$$

If, instead of (17), we solve the augmented adjoint system (18), then $\bar{\boldsymbol{\lambda}}$ satisfies

$$\dot{\bar{\boldsymbol{\lambda}}} - \frac{1}{2}\bar{\boldsymbol{\lambda}} = 0,$$

which is stable to integration from the right. We will show that, in general, if the original system (14) is stable, then the augmented adjoint system (18) for $\bar{\boldsymbol{\lambda}}$ is stable. First, note that since $\mathbf{M}$ is nonsingular, $\bar{\boldsymbol{\lambda}}$ satisfies

$$\begin{aligned} \dot{\bar{\boldsymbol{\lambda}}} - \mathbf{K}^*(t)\left(\mathbf{M}^*(t)\right)^{-1}\bar{\boldsymbol{\lambda}} + \mathbf{r}(t) &= \mathbf{0}, \\ \bar{\mathbf{A}}\bar{\boldsymbol{\lambda}}(a) + \bar{\mathbf{B}}\bar{\boldsymbol{\lambda}}(b) &= \mathbf{0}. \end{aligned}$$

In other words, for the implicit ODE BVP, the augmented adjoint system for $\bar{\boldsymbol{\lambda}}$ is the same as the adjoint system of the explicit ODE BVP equivalent to the original

system (14). Therefore it is sufficient to investigate stability of the adjoint system for the explicit ODE BVP,

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{C}(t)\mathbf{x} + \mathbf{q}(t), \\ \mathbf{A}\mathbf{x}(a) + \mathbf{B}x(b) &= \mathbf{c}, \end{aligned} \tag{19}$$

whose adjoint BVP can be written as

$$\begin{aligned} \dot{\boldsymbol{\lambda}} &= -\mathbf{C}^*(t)\boldsymbol{\lambda} + \mathbf{r}(t), \\ \bar{\mathbf{A}}\boldsymbol{\lambda}(a) + \bar{\mathbf{B}}\boldsymbol{\lambda}(b) &= \mathbf{0}. \end{aligned} \tag{20}$$

We begin by deriving the relationship between fundamental solutions of these two problems. This is given by the following.

LEMMA 1. *Let* $\mathbf{X}$ *and* $\boldsymbol{\Lambda}$ *be any fundamental solutions of* (19) *and* (20), *respectively. Then, for any* $t, s \in [a, b]$, $\boldsymbol{\Lambda}^*(t)\mathbf{X}(t) = \boldsymbol{\Lambda}^*(s)\mathbf{X}(s)$.

Proof. Consider $\mathbf{Z}(t) = \boldsymbol{\Lambda}^*(t)\mathbf{X}(t)$. Then

$$\dot{\mathbf{Z}} = \dot{\boldsymbol{\Lambda}}^*(t)\mathbf{X}(t) + \boldsymbol{\Lambda}^*(t)\dot{\mathbf{X}}(t) = \left(-\mathbf{C}^*(t)\boldsymbol{\Lambda}(t)\right)^* \mathbf{X}(t) + \boldsymbol{\Lambda}^*(t)\left(\mathbf{C}(t)\mathbf{X}(t)\right) = \mathbf{0}.$$

Therefore $\boldsymbol{\Lambda}^*(t)\mathbf{X}(t) = \boldsymbol{\Lambda}^*(s)\mathbf{X}(s)$ for all $t, s \in [a, b]$; in particular, $\boldsymbol{\Lambda}^*(t)\mathbf{X}(t) = \boldsymbol{\Lambda}^*(a)\mathbf{X}(a) = \boldsymbol{\Lambda}^*(b)\mathbf{X}(b)$ for any $t \in [a, b]$. ∎

As a direct consequence of Lemma 1, $\|\boldsymbol{\Lambda}(t)\boldsymbol{\Lambda}^{-1}(s)\| = \|\mathbf{X}(s)\mathbf{X}^{-1}(t)\|$ for all $s \geq t$. This proves the following.

THEOREM 1. *The adjoint system* (20) *of an (asymptotically) stable linear ODE IVP* (19) *is (asymptotically) stable.*

We now concentrate on the ODE BVP. We first show the following.

THEOREM 2. *If the BVP* (19) *has a unique solution, then a solution for the adjoint BVP* (20) *exists and is unique.*

Proof. Consider the fundamental solution $\mathbf{X}(t)$ of the homogeneous equivalent of (19) which satisfies $\mathbf{X}(a) = \mathbf{I}$. Then the matrix $\mathbf{Q} = \mathbf{A} + \mathbf{B}\mathbf{X}(b)$ is nonsingular [2]. Similarly, let $\boldsymbol{\Lambda}(t)$ be the fundamental solution of the homogeneous equivalent of (20), which satisfies $\boldsymbol{\Lambda}(b) = \mathbf{I}$, and construct the matrix $\bar{\mathbf{Q}} = \bar{\mathbf{A}}\boldsymbol{\lambda}(a) + \bar{\mathbf{B}}$. From Lemma 1 we have that $\boldsymbol{\Lambda}(a) = \mathbf{X}^*(b)$. Then

$$\mathbf{Q}\bar{\mathbf{A}}^* = \mathbf{A}\bar{\mathbf{A}}^* + \mathbf{B}\mathbf{X}(b)\bar{\mathbf{A}}^* = \mathbf{B}\bar{\mathbf{B}}^* + \mathbf{B}\boldsymbol{\Lambda}^*(a)\bar{\mathbf{A}}^* = \mathbf{B}\bar{\mathbf{Q}}^*$$

and

$$\begin{aligned} \mathbf{Q}\mathbf{X}^{-1}(b)\bar{\mathbf{B}}^* &= \mathbf{A}\mathbf{X}^{-1}(b)\bar{\mathbf{B}}^* + \mathbf{B}\bar{\mathbf{B}}^* = \mathbf{A}\left(\boldsymbol{\Lambda}^*(a)\right)^{-1}\bar{\mathbf{B}}^* + \mathbf{A}\bar{\mathbf{A}}^* \\ &= \mathbf{A}\left(\boldsymbol{\Lambda}^*(a)\right)^{-1}\bar{\mathbf{Q}}^* = \mathbf{A}\mathbf{X}^{-1}(b)\bar{\mathbf{Q}}^*, \end{aligned}$$

where we have used $\mathbf{A}\bar{\mathbf{A}}^* = \mathbf{B}\bar{\mathbf{B}}^*$. Since $\mathbf{Q}$ is invertible, we can write

$$\begin{aligned} \bar{\mathbf{A}}^* &= \mathbf{Q}^{-1}\mathbf{B}\bar{\mathbf{Q}}^*, \\ \bar{\mathbf{B}}^* &= \mathbf{X}(b)\mathbf{Q}^{-1}\mathbf{A}\mathbf{X}^{-1}(b)\bar{\mathbf{Q}}^*. \end{aligned}$$

Thus

$$\begin{bmatrix} \bar{\mathbf{A}}^* \\ \bar{\mathbf{B}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{Q}^{-1}\mathbf{B} \\ \mathbf{X}(b)\mathbf{Q}^{-1}\mathbf{A}\mathbf{X}^{-1}(b) \end{bmatrix} \bar{\mathbf{Q}}^*.$$

Since $[\bar{\mathbf{A}}|\bar{\mathbf{B}}]$ has full row rank, it follows that $\bar{\mathbf{Q}}^*$ has full rank. Therefore, $\bar{\mathbf{Q}}$ is invertible and (20) has a unique solution. ∎

Stability results for the adjoint problem are given by the following [9].

THEOREM 3. *If the BVP* (19) *is well conditioned, then its adjoint BVP* (20) *is well conditioned.*

We now consider numerical stability for the adjoint system. As a consequence of Theorem 3, zero-stability (i.e., stability as the stepsize $h \to 0$ and the number of steps $n \to \infty$) for the adjoint BVP (20) is inherited from zero-stability of the original BVP (19). Here we are concerned with the question of whether a numerical method which is stable for the original system (19) on the mesh

$$\pi : a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$$

is also stable for the adjoint system (20). We consider the midpoint method for which we show the following.

THEOREM 4. *Numerical stability of the midpoint method for the original system on some mesh π implies numerical stability for the adjoint system on the same mesh.*

Proof. Discretizing the original system (19) with the midpoint rule, we obtain

$$\begin{aligned} \frac{\mathbf{x}_n - \mathbf{x}_{n-1}}{h_n} &= \mathbf{C}(t_{n-1/2})\frac{\mathbf{x}_n + \mathbf{x}_{n-1}}{2} + \mathbf{q}(t_{n-1/2}), \qquad n = 1, \ldots, N, \\ \mathbf{A}\mathbf{x}_0 + \mathbf{B}\mathbf{x}_N &= -\mathbf{c}, \end{aligned} \tag{21}$$

where $h_n = t_n - t_{n-1}$. The first N relations in (21) can be written as

$$-\mathbf{S}_n\mathbf{x}_{n-1} + \mathbf{R}_n\mathbf{x}_n = \mathbf{q}(t_{n-1/2}), \qquad n = 1, \ldots, N,$$

where

$$\begin{aligned} \mathbf{S}_n &= \frac{1}{h_n}\mathbf{I} + \frac{1}{2}\mathbf{C}(t_{n-1/2}), \\ \mathbf{R}_n &= \frac{1}{h_n}\mathbf{I} - \frac{1}{2}\mathbf{C}(t_{n-1/2}). \end{aligned}$$

Thus we have that $[\mathbf{x}_0^*; \mathbf{x}_N^*]$ is the solution of a linear system of the form

$$\begin{bmatrix} \mathbf{P} & -\mathbf{I} \\ \mathbf{A} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{q}} \\ -\mathbf{c} \end{bmatrix} \tag{22}$$

for some right-hand side $\hat{\mathbf{q}}$, where $\mathbf{P} = \mathbf{R}_N^{-1}\mathbf{S}_N\mathbf{R}_{N-1}^{-1}\mathbf{S}_{N-1}\cdots\mathbf{R}_1^{-1}\mathbf{S}_1$.

The midpoint method applied to the adjoint problem (20) yields the linear equations

$$\mathbf{S}_n^*\boldsymbol{\lambda}_n - \mathbf{R}_n^*\boldsymbol{\lambda}_{n-1} = \mathbf{r}(t_{n-1/2}), \qquad n = 1, \ldots, N.$$

Since $\mathbf{S}_n$ and $\mathbf{R}_n^{-1}$ commute, it follows that $[\boldsymbol{\lambda}_N^*; \boldsymbol{\lambda}_0^*]$ is the solution of a linear system

$$\begin{bmatrix} \mathbf{P}^* & -\mathbf{I} \\ \bar{\mathbf{B}} & \bar{\mathbf{A}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_N \\ \boldsymbol{\lambda}_0 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{r}} \\ \mathbf{0} \end{bmatrix} \tag{23}$$

for some right-hand side $\hat{\mathbf{r}}$. We show next that with $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ defined by (8), solving linear system (23) is equivalent to solving linear system (22). Indeed, system (22) can be solved as

$$\begin{aligned}(\mathbf{A}+\mathbf{BP})\mathbf{x}_0 &= \mathbf{B}\hat{\mathbf{q}}-\mathbf{c},\\ \mathbf{x}_N &= \mathbf{P}\mathbf{x}_0-\hat{\mathbf{q}}.\end{aligned}$$

On the other hand, by construction, there exists a vector $\boldsymbol{\alpha}$ defined as in (10). Thus system (23) can be solved as

$$\begin{aligned}(\mathbf{P}^*\mathbf{B}^*+\mathbf{A}^*)\boldsymbol{\alpha} &= \hat{\mathbf{r}},\\ \boldsymbol{\lambda}_0 &= -\mathbf{A}^*\boldsymbol{\alpha},\\ \boldsymbol{\lambda}_N &= \mathbf{B}^*\boldsymbol{\alpha}.\end{aligned}$$

Noting that $\|(\mathbf{P}^*\mathbf{B}^*+\mathbf{A}^*)^{-1}\| = \|(\mathbf{A}+\mathbf{BP})^{-1}\|$, this concludes the proof. ∎

4. Numerical example. As an example we consider the following ODE BVP:

$$\begin{aligned}(J+ml^2)\ddot{\theta} &= u(t)-mgl\cos(\theta),\\ \theta(a) &= \theta_0,\\ \theta(b) &= \theta_1,\end{aligned} \tag{24}$$

which describes the motion of a 2-D pendulum of length $2l$, mass m, and inertia J under the action of gravity (g) and a time varying applied torque $u(t)$. The position of the pendulum is imposed at the initial and final times. Considering the torque $u(t)$ parameterized by $p \in R^{N_p}$, we wish to estimate the sensitivity with respect to p of the energy $g(\theta,\dot{\theta},p,t) = \frac{1}{2}(J+ml^2)\dot{\theta}^2 + mgl\sin(\theta)$ at some time $t \in (a,b)$, as well as the sensitivity of the total energy $G^\tau(p) = \int_a^\tau g(\theta,\dot{\theta},p,t)dt$ over the interval $[a,\tau]$.

As an alternative to using adjoint sensitivity analysis for the solution of these problems, one could generate the sensitivity ODE BVP systems (3) by the following forward method: For each of the parameters p_i, compute the sensitivities of the trajectories $(\theta(t),\dot{\theta}(t))$ and then, using the chain rule of differentiation, evaluate the gradients of g and G^τ. However, such an approach is computationally expensive, especially if the dimension N_p of the parameterization of $u(t)$ is very large.

We first transform (24) into a first order ODE BVP:

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= \frac{1}{J+ml^2}\left(u(t)-mgl\cos(x_1)\right),\\ x_1(a) &= \theta_0,\\ x_1(b) &= \theta_1,\end{aligned} \tag{25}$$

in which case

$$g(x,p,t) = \frac{1}{2}(J+ml^2)x_2^2 + mgl\sin(x_1) \tag{26}$$

and

$$G^\tau(p) = \int_a^\tau \left(\frac{1}{2}(J+ml^2)x_2^2 + mgl\sin(x_1)\right)dt. \tag{27}$$

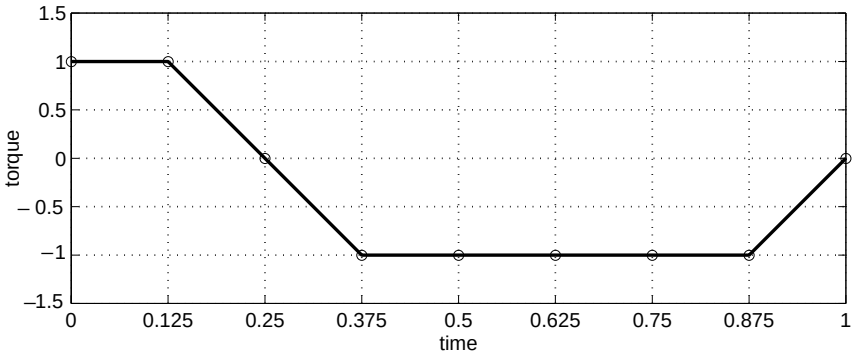


FIG. 1. *Torque $u(t)$ for the pendulum example ($N = 8$, $N_p = 9$).*

Consider a piecewise linear approximation of u given by

$$u(t) = -p_k \frac{t - k\Delta t}{\Delta t} + p_{k+1} \frac{t - (k-1)\Delta t}{\Delta t}; \tag{28}$$
$$(k-1)\Delta t \le t \le k\Delta t, \qquad k = 1, 2, \ldots, N,$$

where $\Delta t = (b-a)/N$. This gives $N_p = N + 1$ parameters p. Let $N = 8$ and let u be as in Figure 1. For $a = 0$, $b = 1$, $\theta_0 = 0$, $\theta_1 = 0$, $\tau = 0.25$, and $m = g = l = J = 1$, we compare gradients of g and G^τ at both $\tau \in (a, b)$ and $\tau = b$ obtained by the adjoint sensitivity analysis presented in the previous sections with gradients computed through forward sensitivity analysis. Differences in gradients computed with the two methods are summarized in Table 1.

TABLE 1
Differences in gradients computed with adjoint (a) and forward methods (f).

i	$\left[\frac{dG^\tau}{dp}\right]_{(a)} - \left[\frac{dG^\tau}{dp}\right]_{(f)}$	$\left[\frac{dg(\tau)}{dp}\right]_{(a)} - \left[\frac{dg(\tau)}{dp}\right]_{(f)}$
1	$4.068486 \cdot 10^{-06}$	$-5.219818 \cdot 10^{-09}$
2	$8.042969 \cdot 10^{-06}$	$-9.258004 \cdot 10^{-12}$
3	$7.934972 \cdot 10^{-06}$	$-3.032900 \cdot 10^{-11}$
4	$5.331125 \cdot 10^{-06}$	$1.268874 \cdot 10^{-06}$
5	$9.902579 \cdot 10^{-08}$	$3.172192 \cdot 10^{-06}$
6	$-4.355139 \cdot 10^{-08}$	$5.555800 \cdot 10^{-11}$
7	$-2.891162 \cdot 10^{-08}$	$2.859599 \cdot 10^{-11}$
8	$-1.443129 \cdot 10^{-08}$	$8.619001 \cdot 10^{-12}$
9	$-3.178720 \cdot 10^{-10}$	$5.191848 \cdot 10^{-09}$
i	$\left[\frac{dG^b}{dp}\right]_{(a)} - \left[\frac{dG^b}{dp}\right]_{(f)}$	$\left[\frac{dg(b)}{dp}\right]_{(a)} - \left[\frac{dg(b)}{dp}\right]_{(f)}$
1	$1.013121 \cdot 10^{-05}$	$-4.789156 \cdot 10^{-09}$
2	$1.990550 \cdot 10^{-05}$	$-8.493997 \cdot 10^{-12}$
3	$1.943275 \cdot 10^{-05}$	$-2.782700 \cdot 10^{-11}$
4	$1.906737 \cdot 10^{-05}$	$-5.279502 \cdot 10^{-11}$
5	$1.883816 \cdot 10^{-05}$	$-7.485801 \cdot 10^{-11}$
6	$1.875241 \cdot 10^{-05}$	$-8.563900 \cdot 10^{-11}$
7	$1.879674 \cdot 10^{-05}$	$-7.838902 \cdot 10^{-11}$
8	$1.893979 \cdot 10^{-05}$	$-4.928702 \cdot 10^{-11}$
9	$9.530949 \cdot 10^{-06}$	$-1.530841 \cdot 10^{-06}$

All ODE BVPs involved in both adjoint and forward computations were numerically solved with COLSYS [1]. We note that the version of COLSYS that we used has a limit of 20 on the number of differential equations, thus limiting the number of parameters that we could include for the forward sensitivity system to $N_p = 9$. Of course, when using the adjoint approach this is not an issue, as only one additional BVP of the same size as the original problem must be solved to evaluate gradients with respect to an array of parameters of any size. The other obvious advantage of using adjoint sensitivity versus forward sensitivity is, of course, computational efficiency. Indeed, solution of the BVP (original + adjoint) required by the adjoint approach was 15 times faster than solution of the BVP (original + N_p forward sensitivities) required by the forward approach. In all fairness, we must note that a careful implementation of forward sensitivity analysis (which takes advantage of the special structure of the sensitivity systems and the fact that they share the same Jacobian matrices with the original BVP) will lead to a speedup of only about $(1 + N_p)/(1 + 1) = 5$. Since COLSYS does not provide a sensitivity analysis capability, the overhead computations (especially in the linear algebra) explain the much higher speedup obtained.

REFERENCES

[1] U.M. Ascher, J. Christiansen, and R.D. Russel, *A collocation solver for mixed order systems of boundary value problems*, Math. Comp., 33 (1979), pp. 659–679.

[2] U.M. Ascher, R.M.M. Mattheij, and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM, Philadelphia, 1995.

[3] K. Balla and R. Marz, *Linear differential algebraic equations of index* 1 *and their adjoint equations*, Results Math., 37 (2000), pp. 13–35.

[4] C. Bischof, A. Carle, G. Corliss, A. Griewank, and P. Hovland, *ADIFOR - Generating derivative codes from Fortran programs*, Scientific Programming, 1 (1992), pp. 11–29.

[5] D.G. Cacuci, *Sensitivity theory for nonlinear systems.* I. *Nonlinear functional analysis approach*, J. Math. Phys., 22 (1981), pp. 2794–2802.

[6] D.G. Cacuci, *Sensitivity theory for nonlinear systems.* II. *Extension to additional classes of responses*, J. Math. Phys., 22 (1981), pp. 2803–2812.

[7] Y. Cao, S. Li, L.R. Petzold, and R. Serban, *Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution*, submitted.

[8] M. Caracotsios and W.E. Stewart, *Sensitivity analysis of initial value problems with mixed ODEs and algebraic equations*, Computers and Chemical Engineering, 9 (1985), pp. 359–365.

[9] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[10] R.M. Errico, *What is an adjoint model*? Bull. Amer. Meteorological Society, 78 (1997), pp. 2577–2591.

[11] W.F. Feehery and P.I. Barton, *Efficient sensitivity analysis of large-scale differential-algebraic systems*, Appl. Numer. Math., 25 (1997), pp. 41–54.

[12] G. Ghione and F. Filicori, *A computationally efficient unified approach to the analysis of the sensitivity and noise of semiconductor devices*, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, 12 (1993), pp. 425–438.

[13] R. Giering and T. Kaminski, *Recipes for adjoint code construction*, ACM Trans. Math. Software, 24 (1998), pp. 437–474.

[14] M.B. Giles and N.A. Pierce, *Adjoint Equations in CFD: Duality, Boundary Conditions and Solution Behaviour*, AIAA Paper 97-1850, American Institute of Aeronautics and Astronautics, New York, 1997.

[15] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[16] E. Hairer, S.P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.

[17] E.J. Haug and P.E. Ehle, *Second-order design sensitivity analysis of mechanical system dynamics*, Internat. J. Numer. Methods Engrg., 18 (1982), pp. 1699–1717.

[18] E.J. Haug, R. Wehage, and N.C. Barman, *Design sensitivity analysis of planar mechanism and machine dynamics*, Trans. ASME, 103 (1981).
[19] H.B. Keller, *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdel, Waltham, MA, 1968.
[20] R.M. Lewis, *Numerical Computation of Sensitivities and the Adjoint Approach*, Tech. report 97-61, ICASE, NASA Langley Research Center, Hampton, VA, 1997.
[21] S. Li and L.R. Petzold, *Design of New DASPK for Sensitivity Analysis*, Tech. report, Department of Computer Science, University of California, Santa Barbara, CA, 1999.
[22] S. Li and L.R. Petzold, *Software and algorithms for sensitivity analysis of large-scale differential-algebraic systems*, J. Comput. Appl. Math., 125 (2000), pp. 131–145.
[23] L. Machiels, *A Posteriori Finite Element Output Bounds for Discontinuous Galerkin Discretizations of Parabolic Problems*, Tech. report UCRL-JC-136614, Lawrence Livermore National Laboratory, Livermore, CA, 1999.
[24] L. Machiels, Y. Maday, and A.T. Patera, *Output Bounds for Reduced-Order Approximations of Elliptic Partial Differential Equations*, Tech. report FML 99-5-1, Massachusetts Institute of Technology, Cambridge, MA, 1999.
[25] T. Maly and L.R. Petzold, *Numerical methods and software for sensitivity analysis of differential-algebraic systems*, Appl. Numer. Math., 20 (1997), pp. 57–79.
[26] G.I. Marchuk, V.I. Agoshkov, and V.P. Shutyaev, *Adjoint Equations and Perturbation Algorithms*, CRC Press, Boca Raton, FL, 1996.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 233–253

AUGMENTED LAGRANGE-SQP METHODS WITH LIPSCHITZ-CONTINUOUS LAGRANGE MULTIPLIER UPDATES*

E. SACHS† AND S. VOLKWEIN‡

Abstract. Augmented Lagrangian-SQP methods using Lipschitz-continuous Lagrange multiplier updates are analyzed. Kantorovich-style convergence results are proved and applied to the discretization of optimal control problems. The existence of stationary points for the discretized problems is also discussed.

Key words. nonlinear programming, multiplier methods, Lipschitz-continuous multiplier methods, Burgers equation

AMS subject classifications. 49M15, 49M37, 65K05, 90C30, 90C55

PII. S0036142998348212

1. Introduction.

This paper is concerned with an optimization problem of the following type:

$$\text{minimize } J(x) \text{ subject to } e(x) = 0, \tag{1.1}$$

where $J : X \to \mathbb{R}$ and $e : X \to Y$ are sufficiently smooth functions and X, Y are real Hilbert spaces. These types of problems occur, for example, in the optimal control of systems described by partial differential equations. In many applications (1.1) is solved by variants of SQP methods. The principal idea of these algorithms is to replace J and e by a quadratic approximation of the Lagrangian and a linearization of the constraint. The resulting augmented Lagrangian-SQP method is as follows [GMW81, pp. 225–233].

ALGORITHM 1.

(a) *Choose* $(x^0, \lambda^0) \in X \times Y$, $c \geq 0$, *and set* $n = 0$.

(b) *Solve the following quadratic minimization problem for* s:

$$\text{(QP)} \quad \min L_c'(x^n, \lambda^n)s + \frac{1}{2} L_c''(x^n, \lambda^n)(s, s) \text{ subject to } e'(x^n)s + e(x^n) = 0,$$

where L_c *denotes the augmented Lagrangian*

$$L_c(x, \lambda) = J(x) + \langle e(x), \lambda \rangle_Y + \frac{c}{2} \|e(x)\|_Y^2 \quad \text{for } c \geq 0. \tag{1.2}$$

(c) *Set* $x^{n+1} = x^n + s$, $\lambda^{n+1} = \Lambda(x^n, \lambda^n)$, *and go back to step* (b).

The parameter c is called the augmentation or penalty parameter. Versions of the augmented Lagrangian-SQP methods differ in the choice of the multiplier update $\Lambda(x, \lambda)$. We refer the reader to [FST87] for a review of multiplier updates in finite

*Received by the editors November 20, 1998; accepted for publication (in revised form) November 14, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/34821.html

†Fachbereich IV-Mathematik, Universität Trier, D-54286 Trier, Germany (sachs@uni-trier.de) and Department of Mathematics, ICAM, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (sachs@icam.vt.edu).

‡Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstraße 36, A-8010 Graz, Austria (stefan.volkwein@uni-graz.at). This author's research was supported in part by the Graduiertenkolleg Mathematische Optimierung at the University of Trier.

dimensions. It is well known that different multiplier updates lead to different rates of convergence. If Λ is the Newton multiplier update, i.e.,

$$\Lambda_N(x,\lambda) = \left(e'(x)L_c''(x,\lambda)^{-1}e'(x)^\star\right)^{-1}\left(e(x) - e'(x)L_c''(x,\lambda)^{-1}L_c'(x,\lambda)\right), \tag{1.3}$$

then Algorithm 1 coincides with the augmented Lagrangian Newton-SQP method. This method was analyzed by Ito and Kunisch in [IK96]. In this paper we consider Lipschitz-continuous update formulas for the Lagrange multiplier that depend only on the x-variable. We assume that $\lambda^n = \Lambda(x^n)$, where $\lambda : X \to Y$ satisfies

$$\|\Lambda(x) - \Lambda(\bar{x})\|_Y \le \gamma_\Lambda \|x - \bar{x}\|_X \tag{1.4}$$

for all $x, \bar{x}$ in an appropriate neighborhood of the starting point and a constant $\gamma_\Lambda \ge 0$ such that as an example, consider the least-squares multipliers [Kle97], i.e., the solution of

$$\text{minimize } \|J'(x^{n+1}) + e'(x^{n+1})^\star\lambda\|_X \text{ over } \lambda \in Y.$$

The least-squares solution is given by

$$\lambda^{n+1} = -\left(e'(x^{n+1})e'(x^{n+1})^\star\right)^{-1} e'(x^{n+1})J'(x^{n+1}), \tag{1.5}$$

where $e'(x^{n+1})^\star$ denotes the adjoint of the operator $e'(x^{n+1})$. Another example is given by the update of Kunisch and Sachs in [KS92]. Note that, in contrast, the Newton multiplier update (1.3) does not generally satisfy the Lipschitz continuity condition (1.4).

In this paper we prove convergence results using the Kantorovich theory. As is well known, the advantage of this theory is that the existence of a stationary point is not required. This is particularly useful in the case of discretized infinite-dimensional optimization problems arising in optimal control. In addition to proving that the iterates converge, we show that stationary points for the discretized problem exist. In [KV97] and [Vol00a] the convergence of the augmented Lagrangian-SQP method is considered in terms of the pair (x, λ). In this paper we sharpen the convergence statement by considering only the variable x. This requires the definition of a new fixed point map in order to prove contraction properties only for x. This approach is presented for Lipschitz-continuous multiplier updates. We keep the update rule as general as possible and impose additional assumptions such as $\Lambda(x^*) = \lambda^*$ at a later stage.

The paper is organized as follows. In section 2 an example of a constrained optimal control problem is presented. In the next section the augmented Lagrangian-SQP method is formulated with a Lipschitz-continuous update. We prove under certain conditions that the first-order necessary optimality conditions for (1.1) have a solution x^*. Therefore, we need not assume that $\Lambda(x^*) = \lambda^*$, where λ^* is the Lagrange multiplier associated with x^*. In section 4 these results are applied to discretized problems. If the infinite-dimensional first-order necessary optimality conditions have a solution, it can be shown that the discretized optimality conditions also have a solution. Finally, a numerical example is presented in section 5.

We introduce some notation that will be used throughout the paper. Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed linear spaces. The set $B(v; r)$ denotes an open ball of radius $r > 0$ centered at the point $x \in X$. By $\mathcal{L}(V, W)$ we denote the normed linear space of all bounded linear operators from V into W and write $\mathcal{L}(V) = \mathcal{L}(V, V)$. By $\langle \cdot, \cdot \rangle_{V',V}$

we denote the dual pair associated with V and its dual. For $\mathcal{A} \in \mathcal{L}(V,W)$ the set $N(\mathcal{A}) = \{v \in V : \mathcal{A}(v) = 0\}$ denotes the null space of $\mathcal{A}$. Let X and Y be two real Hilbert spaces with inner products $\langle \cdot\,, \cdot \rangle_X$ and $\langle \cdot\,, \cdot \rangle_Y$, respectively, and $\mathcal{A} \in \mathcal{L}(X,Y)$. The (Hilbert space) adjoint $\mathcal{A}^\star \in \mathcal{L}(Y,X)$ of $\mathcal{A}$ is defined by

$$\langle \mathcal{A}^\star(\lambda), x \rangle_X = \langle \lambda, \mathcal{A}(x) \rangle_Y \quad \text{for all } x \in X, \lambda \in Y.$$

2. Formulation of the problem. Let us consider the following constrained optimal control problem:

$$\text{(P)} \qquad \text{minimize } J(x) \text{ subject to } e(x) = 0,$$

where $J : X \to \mathbb{R}$, $e : X \to Y$, and X, Y are real Hilbert spaces. Note that we do not distinguish between a functional in the dual space and its Riesz representation in the Hilbert space. The Hilbert space $X \times Y$ is endowed with the Hilbert space product topology. The Fréchet-derivatives with respect to the variable x are denoted by primes. We start with an example that motivates the material of later sections. We also refer to [KV97] and [Vol00a], where further examples were presented.

Example 2.1. Let $\Omega = (0,1) \subset \mathbb{R}$, $\Omega_\circ \subseteq \Omega$ with positive measure, and $f \in H^{-1}(\Omega)$. For a control $u \in L^2(\Omega_\circ)$ the state $y \in H_0^1(\Omega)$ is given by the variational solution of the steady-state Burgers equation

$$-\nu y'' + yy' = f + \mathcal{E}(u) \quad \text{in } H^{-1}(\Omega), \tag{2.1}$$

where $\mathcal{E} \in \mathcal{L}(L^2(\Omega_\circ), H^{-1}(\Omega))$ is the extension operator

$$\mathcal{E}(u) = \langle u, \cdot \rangle_{L^2(\Omega_\circ)}. \tag{2.2}$$

Equation (2.1) is a second-order approximation to the one-dimensional steady-state Navier–Stokes equations [Lig56]. In the general case $f \in H^{-1}(\Omega)$ would be a force density, y a velocity field, and $\nu > 0$ a viscosity parameter. Note that for $u \in L^2(\Omega_\circ)$, $\mathcal{E}(u)$ belongs to $L^2(\Omega_\circ)'$, which can be identified with $L^2(\Omega_\circ)$. We equip $H_0^1(\Omega)$ with the inner product

$$\langle \varphi, \psi \rangle_{H_0^1} = \int_\Omega \varphi' \psi' \, dx \quad \text{for } \varphi, \psi \in H_0^1(\Omega).$$

It is proved in [Vol00b] that for all $u \in L^2(\Omega_\circ)$ there exists at least one $y \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \nu y' \varphi' + yy' \varphi \, dx = \langle f + \mathcal{E}(u), \varphi \rangle_{H^{-1}, H_0^1} \quad \text{for all } \varphi \in H_0^1(\Omega),$$

i.e., (2.1) has at least one variational solution. The cost function is of the tracking type:

$$J(y,u) = \frac{1}{2} \int_\Omega |y - z|^2 \, dt + \frac{\alpha}{2} \int_{\Omega_\circ} |u|^2 \, dt,$$

where $z \in L^2(\Omega)$ denotes a desired state and $\alpha > 0$ is fixed. We introduce the nonlinear operator $e = H_0^1(\Omega) \times L^2(\Omega_\circ) \to H_0^1(\Omega)$ such that

$$e(y,u) = (-\Delta)^{-1}\big(-\nu y'' + yy' - f - \mathcal{E}(u) \big),$$

where Δ is the Laplace operator from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$; i.e., for $\tilde{f} \in H^{-1}(\Omega)$ $v = (-\Delta)^{-1}(\tilde{f})$ solves

$$\int_\Omega v'\varphi'\,dt = \langle \tilde{f}, \varphi \rangle_{H^{-1},H_0^1} \quad \text{for all } \varphi \in H_0^1(\Omega).$$

The resulting optimal control problem is of the form (P) with $X = H_0^1(\Omega) \times L^2(\Omega_\circ)$ and $Y = H_0^1(\Omega)$.

For SQP methods in Hilbert spaces the following smoothness assumptions are imposed on J and e.

Assumption 1. J and e are twice continuously Fréchet-differentiable, and the mappings J'' and e'' are Lipschitz-continuous in a neighborhood $U(x^0)$ of some given point $x^0 \in X$.

Remark 2.2. In Example 2.1 both J and e are twice continuously Fréchet-differentiable in $X = H_0^1(\Omega) \times L^2(\Omega_\circ)$. In particular, for $(y,u) \in X$ we have

$$\begin{aligned}
J'(y,u) &= \begin{pmatrix} (-\Delta)^{-1}\big(\langle y - z, \cdot\rangle_{L^2}\big) \\ \alpha u \end{pmatrix} \in X, \\
J''(y,u)(v,q) &= \begin{pmatrix} (-\Delta)^{-1}\big(\langle v, \cdot\rangle_{L^2}\big) \\ \alpha q \end{pmatrix} \in X, \\
e'(y,u)(v,q) &= (-\Delta)^{-1}\big(-\nu v'' + (yv)' - \mathcal{E}(q)\big) \in Y, \\
e''(y,u)(v,q)(w,p) &= (-\Delta)^{-1}\big(\langle (vw)', \cdot\rangle_{L^2}\big) \in Y
\end{aligned} \tag{2.3}$$

for $(v,q),(w,p) \in X$. Since the second Fréchet-derivatives of J and e do not depend on (y,u), J'' and e'' are Lipschitz-continuous in the whole space X.

The Lagrange multiplier rule holds for (1.5) if the following regularity condition is true.

Assumption 2. The Fréchet derivative $e'(x^0)$ is surjective.

Remark 2.3. In [Vol00b] it was shown that for Example 2.1 the operator $e'(y,u)$ is surjective for all $(y,u) \in X$. In particular, Assumption 2 is satisfied.

Assumption 3. Let $\Lambda : X \to Y$ denote a continuous mapping with a uniform Lipschitz-constant $\gamma_\Lambda \geq 0$ on the set $U(x^0)$, i.e.,

$$\|\Lambda(x) - \Lambda(\bar{x})\|_Y \leq \gamma_\Lambda \|x - \bar{x}\|_X \quad \text{for all } x, \bar{x} \in U(x^0). \tag{2.4}$$

The function Λ denotes the Lipschitz-continuous update for the Lagrange multiplier.

Remark 2.4. In case of the least-squares update (1.5) we have

$$\Lambda(x) = -\left(e'(x)e'(x)^\star\right)^{-1} e'(x)J'(x). \tag{2.5}$$

Under Assumptions 1 and 2 we conclude that Λ is Lipschitz-continuous in a neighborhood of x^0. In fact, Assumption 2 implies that $e'(x)^\star$ is injective in a neighborhood of x^0. It follows from the closed range theorem [Bre87, p. 29] that there exists a $\beta > 0$ such that

$$\langle e'(x)e'(x)^\star\lambda, \lambda\rangle_X = \|e'(x)^*\lambda\|_X^2 \geq \beta\,\|\lambda\|_Y^2 \quad \text{for all } \lambda \in Y.$$

Hence, $e'(x)e'(x)^\star$ is a positive operator and thus invertible.

Remark 2.5. We specify (2.5) for Example 2.1. Let $(y,u) \in X$ and let v be the solution of the Poisson problem

$$-v'' = y - z \text{ in } H^{-1}(\Omega), \quad v(0) = v(1) = 0.$$

Using (2.3) we set

$$q = (-\Delta)^{-1}\left(e'(y,u)J'(y,u)\right) = -\nu v'' + (yv)' - \alpha\mathcal{E}(u) \in H^{-1}(\Omega). \tag{2.6}$$

Note that the adjoint of $(-\Delta)^{-1}$ is given by its inverse operator $-\Delta$. It follows that for $\lambda \in H_0^1(\Omega)$,

$$\begin{aligned}(-\Delta)^{-1}e'(y,u)e'(y,u)^\star\lambda &= (-\Delta)^{-1}e'(y,u)\left((-\Delta)^{-1}(-\nu\lambda'' - y\lambda'), \quad \mathcal{E}^\star(-\Delta)(\lambda)\right)\\ &= -\nu\tilde{\lambda}'' + (y\tilde{\lambda})' - \mathcal{E}\mathcal{E}^\star(-\Delta)(\lambda),\end{aligned}$$

where $\tilde{\lambda} = (-\Delta)^{-1}(-\nu\lambda'' - y\lambda') \in H_0^1(\Omega)$. The adjoint $\mathcal{E}^\star$ of $\mathcal{E}$ is given by $\mathcal{E}^\star(f) = [(-\Delta)^{-1}(f)]\big|_{\Omega_\circ} \in L^2(\Omega_\circ)$ for any $f \in H^{-1}(\Omega)$. Then $\lambda = \Lambda(y,u)$ is the unique solution $(\lambda,\tilde{\lambda}) \in H_0^1(\Omega)\times H_0^1(\Omega)$ of the linear system

$$\left\{\begin{array}{lcll} -\nu\tilde{\lambda}'' + (y\tilde{\lambda})' + \mathcal{E}\mathcal{E}^\star(\lambda'') &=& -q & \text{in } H^{-1}(\Omega),\\ -\tilde{\lambda}'' + \nu\lambda'' + y\lambda' &=& 0 & \text{in } H^{-1}(\Omega).\end{array}\right. \tag{2.7}$$

We give a sufficient condition that the first-order necessary optimality conditions for problem (P),

$$L_c'(x,\lambda) = 0, \quad e(x) = 0 \quad \text{for } c \geq 0, \tag{2.8}$$

have a solution. Note that L_c' is the derivative of the augmented Lagrangian (1.2) with respect to x.

Define the vector

$$R_c(x) = \begin{pmatrix} L_c'(x,\Lambda(x)) \\ e(x)\end{pmatrix}$$

and matrix

$$M_c(x) = \nabla_{(x,\lambda)}\begin{pmatrix} L_c'(x,\lambda) \\ e(x)\end{pmatrix}\Big|_{\lambda=\Lambda(x)} = \begin{pmatrix} L_c''(x,\Lambda(x)) & e'(x)^\star \\ e'(x) & 0\end{pmatrix} \tag{2.9}$$

for all $x \in U(x^0)$ and for $c \geq 0$. Since Λ, e, L_c', e', and L_c'' are Lipschitz-continuous in $U(x^0)$, the maps R_c and M_c are also Lipschitz-continuous with Lipschitz-constants $\gamma_R > 0$ and $\gamma_M > 0$. For convenience, we assume that γ_M is also the Lipschitz-constant for

$$\nabla_{(x,\lambda)}\begin{pmatrix} L_c'(x,\lambda) \\ e(x)\end{pmatrix} = \begin{pmatrix} L_c''(x,\lambda) & e'(x)^\star \\ e'(x) & 0\end{pmatrix}$$

in a neighborhood of $(x^0,\Lambda(x^0))$. To guarantee the invertibility of the matrix M_c at x^0 we need Assumption 2 and the following condition.

Assumption 4. The operator $L_0''(x^0,\Lambda(x^0))$ is coercive on the null space of $e'(x^0)$, i.e., there exists a constant $\kappa > 0$ such that

$$L_0''(x^0,\Lambda(x^0))(\phi,\phi) \geq \kappa\,\|\phi\|_X^2 \quad \text{for all } \phi \in N(e'(x^0)).$$

Remark 2.6. In case $x^0 = x^*$ and $\lambda^* = \Lambda(x^*)$, this assumption includes the usual second-order sufficient optimality condition. However, other choices are also possible. For example, in the case $\Lambda(x^0) = 0$ Assumption 4 means $J''(x^0)$ must be positive definite on the null space of $e'(x^0)$.

Remark 2.7. We discuss Assumption 4 for the least-squares update applied to Example 2.1. Let $x^0 = (y^0, u^0) \in X$. For arbitrary $\phi = (v, q) \in N(e'(x^0))$

$$\begin{aligned} L_0''(x^0, \Lambda(x^0))(\phi, \phi) &= \|v\|_{L^2}^2 + \alpha \|q\|_{L^2(\Omega_\circ)}^2 + 2 \int_\Omega v v' \Lambda(x^0)\, dx \\ &\geq \alpha \|q\|_{L^2(\Omega_\circ)}^2 - 2 \int_\Omega v v' \left(e'(x^0) e'(x^0)^\star \right)^{-1} e'(x^0) J'(x^0)\, dx. \end{aligned}$$

Note that there exists a constant $C_1 > 0$ such that $\|(e'(x^0)e'(x^0)^\star)^{-1}\|_{\mathcal{L}(H_0^1)} \leq C_1$. This bound and the Hölder inequality imply that

$$L_0''(x^0, \Lambda(x^0))(\phi, \phi) \geq \alpha \|q\|_{L^2(\Omega_\circ)}^2 - 2C_1 \|v\|_{L^\infty} \|v'\|_{L^2} \|e'(x^0)J'(x^0)\|_{H_0^1}. \tag{2.10}$$

Here we also use the estimate $\|v\|_{L^\infty} \leq \|v\|_{H_0^1}$ which holds for all $v \in H_0^1(\Omega)$. It may be inferred from (2.6) that

$$\|e'(x^0)J'(x^0)\|_{H_0^1} = \| - \nu w'' + (wy^0)' - \alpha \mathcal{E}(u^0)\|_{H^{-1}},$$

where we put $w = (-\Delta)^{-1}(\langle y^0 - z, \cdot \rangle_{L^2})$. Since $\phi = (v, q) \in N(e'(x^0))$, we obtain that $-\nu v'' + (y^0 v)' - \mathcal{E}(q) = 0$ in $H^{-1}(\Omega)$, which leads to

$$\int_\Omega \nu v' \varphi' + (y^0 v)' \varphi \, dx = \int_{\Omega_\circ} q \varphi \, dx \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Using Remark B.4 of the appendix, there exists a constant $C_2 > 0$ with

$$\|v\|_{H_0^1}^2 \leq C_2 \|q\|_{L^2(\Omega_\circ)}^2.$$

Thus, we have

$$L_0''(x^0, \Lambda(x^0))(\phi, \phi) \geq \frac{\alpha}{2} \|q\|_{L^2(\Omega_\circ)}^2 + \left(\frac{\alpha}{2C_2} - 2C_1 \|e'(x^0)J'(x^0)\|_{H_0^1} \right) \|v\|_{H_0^1}^2. \tag{2.11}$$

If the inequality

$$\| - \nu w'' + (wy^0)' - \alpha \mathcal{E}(u^0)\|_{H^{-1}} < \frac{\alpha}{4C_1 C_2} \tag{2.12}$$

holds, it follows that

$$\kappa = \min \left(\frac{\alpha}{2}, \frac{\alpha}{2C_2} - 2C_1 \|e'(x^0)J'(x^0)\|_{H_0^1} \right) > 0$$

and from (2.11) we obtain

$$L_0''(x^0, \Lambda(x^0))(\phi, \phi) \geq \kappa \|\phi\|_X^2 \quad \text{for all } \phi \in N(e'(x^0)).$$

Condition (2.12) can be deduced from (2.10) if the residual $\|(y^0 - z, \alpha u^0)\|_{L^2 \times L^2(\Omega_\circ)}$ is sufficiently small.

Assumption 4 implies the existence of a constant $\eta_c > 0$ depending on c, such that

$$\|M_c(x^0)^{-1}\|_{\mathcal{L}(X \times Y)} \leq \eta_c. \tag{2.13}$$

Moreover, there exists a neighborhood $V(x^0) \subseteq U(x^0)$ such that, for all $x \in V(x^0)$,

(a) $J(x)$ and $e(x)$ are twice Fréchet-differentiable and their second Fréchet-derivatives are Lipschitz-continuous in $\overline{V(x^0)}$;
(b) $e'(x)$ is surjective; and
(c) $L_0''(x, \Lambda(x))$ is coercive on the null space of $e'(x)$.

3. The augmented Lagrangian-SQP method with Lipschitz-continuous Lagrange multiplier update. If the Newton multiplier update (1.3) is used in Algorithm 2, then the rate of convergence results hold for the pair of iterates (x, λ) [IK96, KS97]. If one uses a Lipschitz-continuous multiplier update, the dependence on λ^n is eliminated and we derive convergence estimates that depend only on the variable x. We consider the map $x^n \mapsto x^{n+1}$ and omit λ. Note that the solution to (QP) is given by the first component of $M_c(x^n)^{-1}R_c(x^n) \in X \times Y$ [Kle97]. Therefore, we introduce the linear and bounded projection $\mathcal{P} : X \times Y \to X$ defined by $\mathcal{P}(x, \lambda) = x$ for all $(x, \lambda) \in X \times Y$. We formulate Algorithm 2 using a Lipschitz-continuous multiplier update.

ALGORITHM 2.

(a) *Choose* $x^0 \in V(x^0)$, $c \geq 0$, *and set* $n = 0$.

(b) *Compute*

$$x^{n+1} = x^n - \mathcal{P}\left(M_c(x^n)^{-1}R_c(x^n)\right). \tag{3.1}$$

(c) *Set* $n = n + 1$ *and go back to* (b).

Define the nonlinear operator $\Phi_c : X \to X$ such that

$$\Phi_c(x) = x - \mathcal{P}\left(M_c(x)^{-1}R_c(x)\right) \quad \text{for } c \geq 0.$$

The computation (3.1) can be written as a fixed-point iteration $x^{n+1} = \Phi_c(x^n)$ for $n = 0, 1, \ldots$.

In (1.3) we introduced the Newton multiplier update Λ_N. On the neighborhood $V(x^0)$ of x^0 the element $\Lambda_N(x, \Lambda(x))$ is well defined and there exists a constant $\gamma_N > 0$ satisfying

$$\|\Lambda_N(x, \Lambda(x)) - \Lambda_N(\bar{x}, \Lambda(\bar{x}))\|_Y \leq \gamma_N \|x - \bar{x}\|_X \quad \text{for all } x, \bar{x} \in V(x^0).$$

In particular, $\Lambda_N(x, \Lambda(x))$ satisfies

$$M_c(x)\begin{pmatrix} \Phi_c(x) - x \\ \Lambda_N(x, \Lambda(x)) \end{pmatrix} = -R_c(x). \tag{3.2}$$

Remark 3.1. From (3.2) and (2.13) we infer that

$$\|x^1 - x^0\|_X = \|\Phi_c(x^0) - x^0\|_X = \|\mathcal{P}(M_c(x^0)^{-1}R_c(x^0))\|_X \leq \eta_c \|R_c(x^0)\|_{X\times Y}$$

holds. Thus we conclude that $\|x^1 - x^0\|_X$ is small if $\|R_c(x^0)\|_{X\times Y}$ is sufficiently small.

The following theorem is of the Kantorovich type since it makes assumptions only on the initial iterate and some problem data rather than on an unknown solution of the problem.

THEOREM 3.2. *Let Assumptions* 1–4 *at some* $x^0 \in X$ *be valid. Assume that*

$$\|\Lambda_N(x^0, \Lambda(x^0))\|_Y \leq \frac{1}{4\gamma_M \eta_c} \tag{3.3}$$

and

$$\|\Phi_c(x^0) - x^0\|_X \leq \frac{1}{16p_1} \tag{3.4}$$

with $p_1 = \eta_c\gamma_M((1+\gamma_\Lambda^2)^2 + 2\gamma_\Lambda)/2$. *Then the iterates defined by Algorithm* 2 *lie in*

$$D = V(x^0) \cap \left\{x \in X : \|x - x^0\|_X < \varrho\right\},$$

where $\varrho = \min\left((\gamma_M\eta_c)^{-1}, (2\gamma_N\gamma_M\eta_c)^{-1}, p_1^{-1}\right)/2$, *and converge to a fixed point* $x^* \in D$ *satisfying*

$$x^* = \Phi_c(x^*), \tag{3.5}$$

i.e., x^* *solves* $\mathcal{P}(M_c(x^*)^{-1}R_c(x^*)) = 0$.

Proof. To prove the claim we want to apply Lemma A.2; see Appendix A. Therefore, we have to estimate $\|\Phi_c(\Phi_c(x)) - \Phi_c(x)\|_X$ for $x, \Phi_c(x) \in D$. First we show that Φ_c is well defined on D. We have for all $x \in D$

$$\|M_c(x) - M_c(x^0)\|_{\mathcal{L}(X\times Y)} \le \gamma_M \, \|x - x^0\|_X < \gamma_M\varrho < \frac{1}{2\eta_c} < \frac{1}{\eta_c}. \tag{3.6}$$

It follows from (2.13) and (3.6) that

$$\|M_c(x^0)^{-1}\|_{\mathcal{L}(X\times Y)} \, \|M_c(x) - M_c(x^0)\|_{\mathcal{L}(X\times Y)} < 1.$$

Now the Banach lemma [OR70, p. 45] gives

$$\|M_c(x)^{-1}\|_{\mathcal{L}(X\times Y)} \le \frac{\eta_c}{1 - \eta_c\gamma_M \, \|x - x^0\|_X} \le 2\eta_c \quad \text{for all } x \in D. \tag{3.7}$$

This implies that Φ_c is well defined on D. Now we proceed by deriving an estimate of the form (A.1). For that purpose let $x, \Phi_c(x) \in D$. Then we have

(3.8)

$$\begin{aligned}
&\|\Phi_c(\Phi_c(x)) - \Phi_c(x)\|_X \\
&= \|\mathcal{P}\big[M_c(\Phi_c(x))^{-1}R_c(\Phi_c(x))\big]\|_X \\
&= \Bigg\|\mathcal{P}\Bigg[M_c(\Phi_c(x))^{-1}\Bigg\{R_c(\Phi_c(x)) - R_c(x) - M_c(x)\begin{pmatrix} \Phi_c(x) - x \\ \Lambda(\Phi_c(x)) - \Lambda(x) \end{pmatrix} \\
&\qquad + (M_c(x) - M_c(\Phi_c(x)))\begin{pmatrix} 0 \\ \Lambda(\Phi_c(x)) - \Lambda(x) - \Lambda_N(x, \Lambda(x)) \end{pmatrix} \\
&\qquad + M_c(\Phi_c(x))\begin{pmatrix} 0 \\ \Lambda(\Phi_c(x)) - \Lambda(x) - \Lambda_N(x, \Lambda(x)) \end{pmatrix}\Bigg\}\Bigg]\Bigg\|_X.
\end{aligned}$$

From the expression

$$\begin{aligned}
&R_c(\Phi(x)) - R_c(x) - M_c(x)\begin{pmatrix} \Phi_c(x) - x \\ \Lambda(\Phi_c(x)) - \Lambda(x) \end{pmatrix} \\
&\quad = \begin{pmatrix} L_c'(\Phi_c(x), \Lambda(\Phi_c(x))) \\ e(\Phi_c(x)) \end{pmatrix} - \begin{pmatrix} L_c'(x, \Lambda(x)) \\ e(x) \end{pmatrix} \\
&\qquad - \nabla_{(x,\lambda)}\begin{pmatrix} L_c'(x,\lambda) \\ e(x) \end{pmatrix}\Bigg|_{\lambda=\Lambda(x)}\begin{pmatrix} \Phi_c(x) - x \\ \Lambda(\Phi_c(x)) - \Lambda(x) \end{pmatrix}
\end{aligned}$$

and Lemma 3.2.12 in [OR70, p. 73] we obtain that

$$\begin{aligned}
&\left\|R_c(\Phi(x)) - R_c(x) - M_c(x)\begin{pmatrix} \Phi_c(x) - x \\ \Lambda(\Phi_c(x)) - \Lambda(x) \end{pmatrix}\right\|_{X\times Y} \\
&\quad \le \frac{\gamma_M}{2} \, \|(\Phi_c(x) - x, \Lambda(\Phi_c(x)) - \Lambda(x))\|_{X\times Y}^2 \\
&\quad \le \frac{\gamma_M}{2}\,(1 + \gamma_\Lambda^2) \, \|\Phi_c(x) - x\|_X^2.
\end{aligned} \tag{3.9}$$

Furthermore,

$$\mathcal{P}\left[M_c(\Phi_c(x))^{-1}M_c(\Phi_c(x))\begin{pmatrix} 0 \\ \Lambda(\Phi_c(x)) - \Lambda(x) - \Lambda_N(x,\Lambda(x)) \end{pmatrix}\right] = 0. \tag{3.10}$$

Since $\mathcal{P}$ is linear, we derive from (3.8), (3.9), and (3.10) that

$$\begin{aligned} &\|\Phi_c(\Phi_c(x)) - \Phi_c(x)\|_X \\ &\quad \le \|M_c(\Phi_c(x))^{-1}\|_{\mathcal{L}(X\times Y)}\left(\frac{\gamma_M}{2}(1+\gamma_\Lambda^2) + \gamma_\Lambda\gamma_M\right)\|\Phi_c(x) - x\|_X^2 \\ &\qquad + \|M_c(\Phi_c(x))^{-1}\|_{\mathcal{L}(X\times Y)}\gamma_M\|\Lambda_N(x,\Lambda(x))\|_Y\|\Phi_c(x) - x\|_X. \end{aligned} \tag{3.11}$$

Using (3.3) the term $\|\Lambda_N(x,\Lambda(x))\|_Y$ can be estimated as follows:

$$\begin{aligned} \|\Lambda_N(x,\Lambda(x))\|_Y &\le \|\Lambda_N(x^0,\Lambda(x^0))\|_Y + \|\Lambda_N(x,\Lambda(x)) - \Lambda_N(x^0,\Lambda(x^0))\|_Y \\ &\le \frac{1}{4\gamma_M\eta_c} + \gamma_N\ \|x - x^0\|_X < \frac{1}{4\gamma_M\eta_c} + \gamma_N\varrho \le \frac{1}{2\gamma_M\eta_c}. \end{aligned}$$

Define $p_2 = 1/2$ and $p_3 = 2p_1$. Then, we have $p_3 \ge \gamma_M\eta_c$ such that we conclude from (3.7) and (3.11) that

$$\|\Phi_c(\Phi_c(x)) - \Phi_c(x)\|_X \le \frac{p_1\ \|\Phi_c(x) - x\|_X^2 + p_2\ \|\Phi_c(x) - x\|_X}{1 - p_3\ \|\Phi_c(x) - x^0\|_X} \tag{3.12}$$

for all $x, \Phi_c(x) \in D$. Note that $\|\Phi_c(x) - x^0\|_X < \varrho \le 1/p_3$. Thus (A.1) holds with

$$\varphi : [0,\infty) \times \left[0, \frac{1}{p_3}\right) \to \mathbb{R}, \quad (t,s) \mapsto \frac{p_1t^2 + p_2t}{1 - p_3s}.$$

Obviously, φ is monotonically increasing in both variables. Note that $p_1 > 0$, $p_2 \ge 0$, and $2p_1 = p_3$. Hence p_1, p_2, and p_3 satisfy the hypotheses of Lemma A.3. Using Lemmas A.2 and A.3 we have proved the theorem. ∎

Remark 3.3. From Theorem 3.2, the iterates $x^{n+1} = \Phi_c(x^n)$, $n \le 1$, lie in D. Applying estimate (3.12) for $x^{n+1} \in D$ leads to

$$\|x^{n+1} - x^n\|_X \le \frac{p_1\ \|x^n - x^{n-1}\|_X^2 + p_2\ \|x^n - x^{n-1}\|_X}{1 - p_3\ \|x^n - x^0\|_X} \quad \text{for } n \ge 1.$$

If $\|x^n - x^0\|_X \le 1/(2p_3)$, we obtain

$$\|x^{n+1} - x^n\|_X \le \frac{p_1}{2}\ \|x^n - x^{n-1}\|_X^2 + \frac{p_2}{2}\ \|x^n - x^{n-1}\|_X \quad \text{for } n \ge 1.$$

This is an estimate for the iterates of the sequence $\{x^n\}_{n\in\mathbb{N}}$. To prove the quadratic convergence of the fixed-point iteration we need, in addition, that $\lambda^* = \Lambda(x^*)$ is satisfied; see Theorem 3.6 below.

Theorem 3.2 yields not only the existence of a fixed point x^* but also the existence of a pair $(x^*,\lambda^*) \in D \times Y$ that solves the first-order necessary optimality conditions (2.8) for problem (P).

COROLLARY 3.4. *Under the hypotheses of Theorem* 3.2 *the multiplier* $\lambda^* = \Lambda_N(x^*,\Lambda(x^*)) + \Lambda(x^*) \in Y$ *satisfies together with* x^* *the first-order necessary optimality conditions for problem* (P).

Proof. Equations (2.9), (3.2), and (3.5) give

$$e'(x^*)^\star \Lambda_N(x^*, \Lambda(x^*)) = -L_c'(x^*, \Lambda(x^*)), \tag{3.13}$$

$$0 = -e(x^*). \tag{3.14}$$

From (3.14) we have $e(x^*) = 0$. Equation (3.13) implies that

$$L_c'(x^*, \Lambda_N(x^*, \Lambda(x^*)) + \Lambda(x^*)) = 0 \quad \text{for } c \geq 0.$$

Setting $\lambda^* = \Lambda_N(x^*, \Lambda(x^*)) + \Lambda(x^*) \in Y$ we conclude that $L_c'(x^*, \lambda^*) = 0$. ∎

COROLLARY 3.5. *Let all hypotheses of Theorem* 3.2 *hold. In case of the least-squares update we have* $\lambda^* = \Lambda(x^*)$. *Furthermore,* x^* *is a solution to* (P).

Proof. From (3.13) and $e(x^*) = 0$ we obtain that

$$\begin{aligned} e'(x^*)^\star \Lambda_N(x^*, \Lambda(x^*)) &= -J'(x^*) - e'(x^*)^\star \Lambda(x^*) \\ &= -J'(x^*) + e'(x^*)^\star (e'(x^*)e'(x^*)^\star)^{-1} e'(x^*) J'(x^*). \end{aligned}$$

Since $e'(x^*)$ is surjective, we have $(e'(x^*)e'(x^*)^\star)^{-1} \in \mathcal{L}(Y)$. Hence, $\Lambda_N(x^*, \Lambda(x^*)) = 0$ and $\lambda^* = \Lambda_N(x^*, \Lambda(x^*)) + \Lambda(x^*) = \Lambda(x^*)$. The rest follows directly from the second-order sufficient optimality condition (Remark 2.6). ∎

If $\lambda^* = \Lambda(x^*)$ for a Lipschitz-continuous multiplier update is used in Algorithm 2, then it is possible to recover the q-quadratic rate of convergence in the iterates x^n. For the proof we refer to [KS97].

THEOREM 3.6. *Let all hypotheses of Theorem* 3.2 *and* $\lambda^* = \Lambda(x^*)$ *hold. Then there exists* $\varepsilon > 0$ *such that for all* $x^0 \in B(x^*, \varepsilon)$ *the sequence* $x^{n+1} = \Phi_c(x^n)$ *converges to* x^* *and*

$$\|x^{n+1} - x^*\|_X \leq C \|x^n - x^*\|_X^2 \quad \textit{for } n = 0, 1, \ldots \tag{3.15}$$

and

$$\|\lambda^{n+1} - \lambda^*\|_X \leq \gamma_\Lambda C \|x^n - x^*\|_X^2 \quad \textit{for } n = 0, 1, \ldots$$

for a constant $C > 0$, *which is independent of* n.

4. Application for discretization methods. Let X and Y be infinite-dimensional Hilbert spaces. The goal of this section is to prove convergence of the discretized version of Algorithm 3 to a solution of the discrete first-order necessary optimality conditions. The existence of such a solution can also be demonstrated by Theorem 3.2.

We assume the following: the constrained minimization problem (**P**) has a local solution x^*; J and e are twice continuously Fréchet-differentiable; the mappings J'' and e'' are Lipschitz-continuous; $e'(x)$ is surjective in the neighborhood $U(x^*)$ of x^*; and $L_0''(x, \Lambda(x))$ is coercive on the null space of $e'(x)$ in $U(x^*)$ of x^* (see the Assumptions 1–4).

Since $e'(x^*)$ is surjective, there exists an element $\lambda^* \in Y$ satisfying the first-order necessary optimality conditions

$$L_c'(x^*, \lambda^*) = L'(x^*, \lambda^*) = 0 \text{ and } e(x^*) = 0 \quad \text{for all } c \geq 0. \tag{4.1}$$

In addition to (2.4) we assume that the multiplier update Λ satisfies

$$\Lambda(x^*) = \lambda^*. \tag{4.2}$$

Remark 4.1. In Remark 2.4 we introduced the least-squares update

$$\Lambda(x) = -\left(e'(x)e'(x)^\star\right)^{-1} e'(x)J'(x).$$

Using (4.1) we get $J'(x^*) = -e'(x^*)^\star\lambda^*$. Thus, (4.2) holds for the least-squares update.

Since X and Y are infinite-dimensional real Hilbert spaces, they have to be discretized for a numerical solution of (P). For this purpose we suppose that we are given finite-dimensional spaces $\{X_h\}_h$ and $\{Y_h\}_h$ approximating X and Y, respectively.

Let $p_h^X \in \mathcal{L}(X_h, X)$ and $p_h^Y \in \mathcal{L}(Y_h, Y)$ be given injective prolongations. In addition, we introduce surjective restrictions $r_h^X \in \mathcal{L}(X, X_h)$ and $r_h^Y \in \mathcal{L}(Y, Y_h)$. For brevity we set $p_h = (p_h^X, p_h^Y)$ and $r_h = (r_h^X, r_h^Y)$.

In many applications it turns out that the solution (x^*, λ^*) of (4.1) as well as the iterates x^n of Algorithm 2 have "better smoothness" than the elements of $X \times Y$ and X, respectively. This is a motivation for the following assumption.

Assumption 5. *There are bounded subsets $V^* \subset X$ such that*

$$x^*,\, x^n,\, x^n - x^*,\, x^{n+1} - x^n \in V^* \tag{4.3}$$

hold for all $n = 0, 1, \ldots$.

Let e_h, L_h', L_h'', and e_h' denote Lipschitz-continuous discretizations of the operators e, L_0', L_0'', and e', respectively. We refer the reader to [Vol00a], where Lipschitz-continuous discretizations are introduced as internal approximations of the infinite-dimensional operators, for instance, $e_h = (p_h^Y)^\star \circ e \circ p_h^X$. Further, we define $\Lambda_h = r_h^Y \circ \Lambda \circ p_h^X$.

We make the following assumptions on the discretization method.

Assumption 6. The discretization methods are described by a family of quadruples

$$\{R_{c,h}, M_{c,h}, p_h, r_h\} \text{ for } h > 0, \tag{4.4}$$

where the operators

$$R_{c,h} : U_h \subseteq X_h \to X_h \times Y_h, \quad M_{c,h} : U_h \subseteq X_h \to \mathcal{L}(X_h \times Y_h) \quad \text{for } h > 0$$

are given by

$$R_{c,h}(x_h) = \begin{pmatrix} L_0'(x_h, \Lambda_h(x_h)) \\ e_h(x_h) \end{pmatrix}, \quad M_{c,h}(x_h) = \begin{pmatrix} L_0''(x_h, \Lambda_h(x_h)) & e_h'(x_h)^\star \\ e_h'(x_h) & 0 \end{pmatrix},$$

and

$$r_h^X(U(x^*) \cap V^*) \subseteq U_h \quad \text{for } h > 0. \tag{4.5}$$

The discretization (4.4) is Lipschitz-uniform with respect to h, i.e., there exist constants $r > 0$, $\Gamma_R > 0$, and $\Gamma_M > 0$ independent of h such that

$$B_h = \{x_h \in X_h : \|x_h - r_h^X x^*\|_X \le r\} \subseteq U_h \quad \text{for } h > 0 \tag{4.6}$$

and

$$\begin{aligned} \|R_{c,h}(x_h) - R_{c,h}(\bar{x}_h)\|_{X\times Y} &\le \Gamma_R \|x_h - \bar{x}_h\|_X, \\ \|M_{c,h}(x_h) - M_{c,h}(\bar{x}_h)\|_{\mathcal{L}(X_h\times Y_h)} &\le \Gamma_M \|x_h - x_h\|_X \end{aligned} \tag{4.7}$$

for all $x_h, \bar{x}_h \in B_h$. Moreover, the discretization family $\{R_{c,h}, M_{c,h}, p_h, r_h\}_{h>0}$ is

(1) uniformly bounded, i.e., there exists a constant $\mu > 0$ independent of h such that

$$\|p_h\|_{\mathcal{L}(X_h \times Y_h, X \times Y)} \leq \mu \text{ and } \|r_h\|_{\mathcal{L}(X \times Y, X_h \times Y_h)} \leq \mu; \tag{4.8}$$

(2) stable in U_h, i.e.,

$$\|M_{c,h}(x_h)^{-1}\|_{\mathcal{L}(X_h \times Y_h)} \leq \sigma_c \quad \text{for all } x_h \in U_h,$$

where the constant $\eta_c > 0$ does not depend on h; and

(3) consistent of order $s > 0$, i.e., there is a constant $\bar{C} > 0$ independent of h such that

$$\|R_{c,h}(r_h^X x) - (p_h)^\star R_c(x)\|_{X \times Y} \leq \bar{C} h^s$$

for all $x \in U(x^*) \cap V^*$.

We derive from (2.4) and (4.8) that

$$\|\Lambda_h(r_h^X x) - \Lambda_h(r_h^X \bar{x})\|_Y = \|r_h^Y \left(\Lambda(p_h^X r_h^X x) - \Lambda(p_h^X r_h^X \bar{x})\right)\|_Y \leq \mu^2 \gamma_\Lambda \|r_h^X x - r_h^X \bar{x}\|_X$$

for all $x, \bar{x} \in U(x^*)$. Hence, $\Gamma_\Lambda = \mu^2 \gamma_\Lambda$ is a uniform Lipschitz-constant for Λ_h.

To formulate the discretization of Algorithm 2 we need the following linear and bounded projection $\mathcal{P}_h : X_h \times Y_h \to X_h$ defined by $\mathcal{P}_h(x_h, \lambda_h) = x_h$ for all $(x_h, \lambda_h) \in X_h \times Y_h$.

ALGORITHM 3.

(a) *Choose* $c \geq 0, x_h^0 \in U_h$, *and* $n = 0$.

(b) *Compute*

$$x_h^{n+1} = \Phi_{c,h}(x_h^n) = x_h^n - \mathcal{P}_h \left(M_{c,h}(x_h^n)^{-1} R_{c,h}(x_h^n)\right). \tag{4.9}$$

(c) *Set* $n = n + 1$ *and go back to* (b).

THEOREM 4.2. *Let Assumptions* 1–6 *hold. Then there exists an* $\bar{h} \in [0, 1]$ *such that for all* $h \in (0, \bar{h}]$ *the operators* $\Phi_{c,h}$ *possess fixed points* x_h^* *satisfying*

$$\|x_h^* - r_h^X x^*\|_X = O(h^s). \tag{4.10}$$

If in addition $\|x^0 - x^*\|_X$ *is sufficiently small, then there exists an* $\tilde{h} \in (0, \bar{h}]$ *such that for all* $h \in (0, \tilde{h}]$ *the sequence* $x_h^{n+1} = \Phi_{c,h}(x_h^n)$ *with the starting value* $x_h^0 = r_h^X x^*$ *converges to* x_h^*.

Proof. We apply Theorem 3.2 to prove this claim. In the constants p_1 and ϱ of Theorem 3.2 we replace γ_M and γ_Λ by the Lipschitz-constants of $M_{c,h}$ and Λ_h, i.e.,

$$p_1 = \frac{1}{2} \sigma_c \Gamma_M \left((1 + \Gamma_\Lambda^2)^2 + 2\Gamma_\Lambda\right).$$

Further, define the constant

$$\rho = \min\left(\frac{1}{8\sigma_c^2 \Gamma_R \Gamma_M}, \frac{1}{2p_1}\right)$$

and the set

$$D_h = \left\{x \in B_h : \|x_h - r_h^X x^*\|_{X_h} < \rho\right\}.$$

We prove that the claim holds for

$$\bar{h} = \min\left(\left(\frac{1}{8\sigma_c^2\bar{C}\Gamma_M}\right)^{1/s}, \left(\frac{1}{16p_1\sigma_c\bar{C}}\right)^{1/s}\right). \tag{4.11}$$

By applying Assumption 6 we find that

$$\|M_{c,h}(x_h^0)^{-1}\|_{\mathcal{L}(X_h\times Y_h)} \le \sigma_c$$

for a constant $\eta_c > 0$ independent of h. This is the estimate (2.13). Condition (3.3) has the form

$$\|\Lambda_{N,h}(x_h^0, \Lambda_h(x^0))\|_Y < \frac{1}{4\Gamma_M\sigma_c}, \tag{4.12}$$

where $\Lambda_{N,h}(x_h^0, \Lambda_h(x^0)) \in Y_h$ satisfies

$$M_{c,h}(x_h^0)\begin{pmatrix} \Phi_{c,h}(x_h^0) - x_h^0 \\ \Lambda_{N,h}(x_h^0, \Lambda_h(x^0)) \end{pmatrix} = -R_{c,h}(x_h^0).$$

We find that

$$\|\Lambda_{N,h}(x_h^0, \Lambda_h(x^0))\|_Y \le \sigma_c \, \|R_{c,h}(x_h^0)\|_{X\times Y}. \tag{4.13}$$

From $R_c(x^*) = 0$ and Assumption 6 we deduce that for $x_h \in D_h$

$$\begin{aligned}\|R_{c,h}(x_h)\|_{X\times Y} &= \|R_{c,h}(x_h) - R_{c,h}(r_h^X x^*)\|_{X\times Y} \\ &\qquad + \|R_{c,h}(r_h^X x^*) - (p_h)^\star R_c(x^*)\|_{X\times Y} \\ &\le \Gamma_R \, \|x_h - r_h^X x^*\|_X + \bar{C}h^s \\ &< \frac{1}{8\sigma_c^2\Gamma_M} + \bar{C}h^s.\end{aligned}$$

Thus (4.13) and (4.11) yield (4.12):

$$\|\Lambda_{N,h}(x_h^0, \Lambda_h(x^0))\|_Y \le \frac{1}{8\sigma_c\Gamma_M} + \sigma_c\bar{C}\bar{h}^s \le \frac{1}{4\Gamma_M\sigma_c}.$$

From Assumption 6, $R_c(x^*) = 0$, and $x_h^0 = r_h^X x^0$ we conclude that

$$\begin{aligned}&\|\Phi_{c,h}(x_h^0) - x_h^0\|_X \\ &\quad \le \|M_{c,h}(x_h^0)^{-1}\|_{\mathcal{L}(X_h\times Y_h)} \, \|R_{c,h}(r_h^X x^*) - (p_h)^\star R_c(x^*)\|_{X\times Y} \\ &\quad \le \sigma_c\bar{C}\bar{h}^s \le \frac{1}{16p_1}.\end{aligned} \tag{4.14}$$

This is the estimate (3.4). Applying Theorem 3.2 leads to the existence of (x_h^*, λ_h^*) which solve the first-order necessary optimality condition of (P_h).

We now verify (4.10). Let $p_2 = 1/2$ and $p_3 = 2p_1$. From Lemmas A.2 and A.3 and (4.14) it follows that

$$\begin{aligned}\|x_h^* - r_h^X x^*\|_X = t^* &\le \frac{1}{p_3}\left(1 - p_2 - \sqrt{(1-p_2)^2 - 4p_1\,\|\Phi_{c,h}(x_h^0) - x_h^0\|_X}\right) \\ &= \frac{4p_1\,\|\Phi_{c,h}(x_h^0) - x_h^0\|_X}{p_3\left(1 - p_2 + \sqrt{(1-p_2)^2 - 4p_1\,\|\Phi_{c,h}(x_h^0) - x_h^0\|_X}\right)} \\ &\le \frac{4p_1\,\|\Phi_{c,h}(x_h^0) - x_h^0\|_X}{p_3(1-p_2)} \le \frac{4p_1\sigma_c\bar{C}}{p_3(1-p_2)}\,h^s,\end{aligned}$$

which gives (4.10). ∎

Remark 4.3. From Theorem 4.2 and Corollary 3.4 there exists $\lambda_h^* \in Y_h$ such that (x_h^*, λ_h^*) solves the first-order necessary optimality conditions for (P_h). If $\lambda_h^* = \Lambda_h(x_h^*)$ holds, the q-quadratic convergence follows for the iterates $x_h^{n+1} = \Phi_{c,h}(x_h^n)$ analogous to the proof of Theorem 3.6.

5. Application to Example 2.1. This section is devoted to the discussion of the assumptions of Theorem 4.2 for Example 2.1.

It was proved in [Vol00b] that (P) has a solution, and that $e'(y^*, u^*)$ is surjective, if $f \in L^2(\Omega)$ holds. Moreover, $L_0''(y^*, u^*, \lambda^*)$ is coercive on the null space of $e'(y^*, u^*)$ if $\|y^* - z\|_{L^2}$ is sufficiently small.

To approximate the Hilbert spaces X and Y in case of Example 2.1 we set $h = \frac{1}{m+1}$, $x_j = jh$ for $j = 0, \dots, m+1$ and introduce piecewise linear functions $\varphi_1, \dots, \varphi_m \in H_0^1(\Omega)$ satisfying $\varphi_i(x_j) = \delta_{ij}$. We restrict our discussion to the case where the set $\Omega_\circ = (a, b) \subseteq \Omega$ is an open nonempty interval. The numbers $i_a, i_b \in \{1, \dots, m\}$ are defined by

$$0 \le a < x_{i_a} < x_{i_a+1} < \cdots < x_{i_b} < b \le 1.$$

For the approximation of X and Y we set

$$X_h = \mathrm{Span}\,\{\varphi_1, \dots, \varphi_m\} \times \mathrm{Span}\,\{\varphi_{i_a}, \dots, \varphi_{i_b}\}, \quad Y_h = \mathrm{Span}\,\{\varphi_1, \dots, \varphi_m\}.$$

To define the prolongations and restrictions we set

$$p_h(y_h, u_h, \lambda_h) = (y_h, u_h, \lambda_h) \quad \text{for } (y_h, u_h) \in X_h \text{ and } \lambda_h \in Y_h$$

and

$$r_h(y, u, \lambda) = \left(\sum_{i=0}^{m} y(x_i)\varphi_i, \frac{1}{h} \sum_{i=i_a}^{i_b} \int_{x_i - \frac{h}{2}}^{x_i + \frac{h}{2}} u\, dx\, \varphi_i, \left(\sum_{i=1}^{m-1} \lambda(x_i)\varphi_i, \mu \right) \right)$$

for $(y, u) \in X$ and $(\lambda, \mu) \in Y$. The Hilbert spaces X_h and Y_h are endowed with the inner products in X and Y, respectively. It follows directly that p_h is a linear injective operator satisfying $\|p_h\|_{\mathcal{L}(X_h \times Y_h, X \times Y)} = 1$ and that r_h is linear and surjective.

We discuss Assumption 5 in the case of Example 2.1. We prove that there exists a bounded subset V^* in $H^2(\Omega) \times H^1(\Omega_\circ)$ such that (4.3) holds. By Theorem 3.6 there exists $\varepsilon > 0$ such that

$$\lim_{n \to \infty} \|(y^n, u^n) - (y^*, u^*)\|_X = 0$$

for all initial values $(y^0, u^0) \in B((y^*, u^*); \varepsilon)$. Hence,

$$\|(y^n, u^n)\|_X \le C_3 \quad \text{for all } n, \tag{5.1}$$

where $C_3 > 0$ does not depend on n. This implies that

$$\|\Lambda(y^n, u^n)\|_{H_0^1} \le \|\Lambda(y^n, u^n) - \Lambda(y^*, u^*)\|_{H_0^1} + \|\lambda^*\|_{H_0^1} \le C_4 \quad \text{for all } n, \tag{5.2}$$

where $C_4 = \gamma_g C_3 + \|(y^*, u^*)\|_X + \|\lambda^*\|_Y$ is independent of n. Let $y^0 \in H^2(\Omega)$ and $u^0 \in H^1(\Omega_\circ)$. Then both (5.1) and (5.2) hold with $n = 0$ and $n = 1$. If $(\bar{y}, \bar{u}, \bar{\lambda})$ is the solution of the linear system

$$M_c(y^0, u^0) \begin{pmatrix} \bar{y} \\ \bar{u} \\ \bar{\lambda} \end{pmatrix} = -R_c(y^0, u^0) + M_c(y^0, u^0) \begin{pmatrix} y^0 \\ u^0 \\ \Lambda(y^0, u^0) \end{pmatrix},$$

then the new iterate $(y^1, u^1) = \Phi_c(y^0, u^0)$ is given by $(y^1, u^1) = (\bar{y}, \bar{u})$. This implies that

$$-\nu\bar{\lambda}'' - y^n\bar{\lambda}' = -\bar{y} + \Lambda(y^n, u^n)'\bar{y} + z - y^n\Lambda(y^n, u^n)', \tag{5.3}$$

$$\alpha\bar{u} = \bar{\lambda}, \tag{5.4}$$

$$-\nu\bar{y}'' = -(y^n\bar{y})' + \mathcal{E}(\bar{u}) + f + y^n(y^n)'. \tag{5.5}$$

Since $y^{n+1} \in L^\infty(\Omega)$, we have $-\nu\lambda'' - y^n\lambda' \in L^2(\Omega)$. As (5.1) holds for all n, we have $\|\bar{y}\|_{H_0^1} \le C_3$. To show that $\bar{\lambda}'' \in L^2(\Omega)$ we estimate the right-hand side of (5.3):

$$\begin{aligned} &\| -\bar{y} + \Lambda(y^0, u^0)'\bar{y} + z - y^0\Lambda(y^0, u^0)'\|_{L^2} \\ &\quad \le \|\bar{y}\|_{L^2} + \|\Lambda(y^0, u^0)\|_{H_0^1}\|\bar{y}\|_{L^\infty} + \|z\|_{L^2} + \|y^0\|_{L^\infty}\|\Lambda(y^0, u^0)\|_{H_0^1} \\ &\quad \le C_3 + C_4C_3 + \|z\|_{L^2} + C_3C_4 =: C_5. \end{aligned} \tag{5.6}$$

Hence $\bar{\lambda}'' \in L^2(\Omega)$ and we can apply the Sturm–Liouville theory in Appendix B. We obtain from Remark B.2, (5.1), and (5.6) that

$$\begin{aligned} \|\bar{\lambda}\|_{H_0^1} &\le \frac{1}{\nu}\exp\left(\frac{2}{\nu}\|y^0\|_{L^1}\right)\| -\bar{y} + \Lambda(y^0, u^0)'\bar{y} + z - y^0\Lambda(y^0, u^0)'\|_{L^2} \\ &\le \frac{1}{\nu}\exp\left(\frac{2C_3}{\nu}\right)C_5 =: C_6. \end{aligned} \tag{5.7}$$

Thus, we infer from (5.4) that

$$\|\bar{u}'\|_{L^2(\Omega_\circ)} = \frac{1}{\alpha}\|\bar{\lambda}'\|_{L^2(\Omega_\circ)} \le \frac{1}{\alpha}\|\bar{\lambda}\|_{H_0^1} \le \frac{C_6}{\alpha}. \tag{5.8}$$

By (5.5) and (5.8) we find that

$$\begin{aligned} \|\bar{y}''\|_{L^2} &\le \frac{1}{\nu}\left(2\|y^0\|_{H_0^1}\|\bar{y}\|_{H_0^1} + \|\bar{u}\|_{L^2(\Omega_\circ)} + \|f\|_{L^2} + \|y^0\|_{H_0^1}^2\right) \\ &\le \frac{1}{\nu}\left(2C_3^2 + C_3 + \|f\|_{L^2} + C_3^2\right) =: C_7. \end{aligned} \tag{5.9}$$

Hence, $(y^1, u^1) \in H^2(\Omega) \times H^1(\Omega_\circ)$ and for $C_8 = \max(C_6/\alpha, C_7)$ it follows that

$$\|(y^1)''\|_{L^2} \le C_8 \text{ and } \|(u^1)'\|_{L^2(\Omega_\circ)} \le C_8. \tag{5.10}$$

Now let (y^n, u^n) be given for $n \ge 0$. Then (5.1) and (5.2) hold. As in the case $n = 0$ we derive that

$$\|(y^{n+1})''\|_{L^2} \le C_8 \text{ and } \|(u^{n+1})'\|_{L^2(\Omega_\circ)} \le C_8$$

hold, where C_8 is the same constant as in (5.10). The first-order necessary optimality conditions for Example 2.1 are given by

$$\begin{aligned} y^* - z - \nu(\lambda^*)'' - y^*(\lambda^*)' = 0 \quad &\text{in } \Omega \quad \text{(the adjoint equation)}, \\ \alpha u^* - \lambda^* = 0 \quad &\text{in } \Omega_\circ \text{ (the optimality condition)}, \\ -\nu(y^*)'' + y^*(y^*)' - f - \mathcal{E}(u^*) = 0 \quad &\text{in } \Omega \quad \text{(the state equation)}. \end{aligned}$$

By using Remark B.2, the adjoint equation, and $\|y^*\|_{H_0^1} \le C_3$ it can be shown that

$$\|\lambda^*\|_{H_0^1} \le \frac{1}{\nu}\exp\left(\frac{2}{\nu}\|y^*\|_{L^2}\right)\|y^* - z\|_{L^2} \le C_6,$$

where the constant $C_6 > 0$ is given by estimate (5.7). Thus, from the optimality condition we infer that $\|(u^*)'\|_{L^2(\Omega_\circ)} \le C_6/\alpha$, which coincides with estimate (5.8). The state equation gives

$$\|(y^*)''\|_{L^2} \le \frac{1}{\nu}\left(2\,\|y^*\|_{L^\infty}\,\|y^*\|_{H_0^1} + \|u^*\|_{L^2(\Omega_\circ)} + \|f\|_{L^2}\right) \le C_7,$$

where the constant $C_7 > 0$ is given by estimate (5.9). We set

$$V^* = \left\{(v,q) \in H^2(\Omega) \times H^1(\Omega_\circ) : \|v''\|_{L^2} \le 2C_8 \text{ and } \|q'\|_{L^2(\Omega_\circ)} \le 2C_8\right\}. \tag{5.11}$$

Thus, we obtain that

$$(y^*,u^*), (y^n,u^n), (y^n,u^n) - (y^*,u^*), (y^{n+1},u^{n+1}) - (y^n,u^n) \in V^*.$$

Now let us discuss Assumption 6 in case of Example 2.1.

(a) We have $\|p_h\|_{\mathcal{L}(X_h \times Y_h, X \times Y)} = 1$. It was proved in [Aub72, p. 38] that

$$\lim_{h\to\infty} \|((y,u),\lambda) - r_h((y,u),\lambda)\|_{X\times Y} = 0 \quad \text{for all } ((y,u),\lambda) \in X \times Y.$$

Hence $\|r_h((y,u),\lambda)\|_{X\times Y}$ is bounded for all $((y,u),\lambda) \in X \times Y$. According to the principle of uniform boundedness [Wou79, p. 112], there exists a constant $C_9 > 0$ such that $\|r_h\|_{\mathcal{L}(X\times Y, X_h \times Y_h)} \le C_9$. Thus, the discretization family is uniformly bounded.

(b) If we define the operator $R_{c,h}$ by the internal approximation of R_c, i.e., $R_{c,h} = (p_h)^\star \circ R_c \circ p_h^X$, then it follows from $\|p_h^\star\|_{\mathcal{L}(X\times Y, X_h \times Y_h)} = 1$ that

$$\|R_{c,h}(y_h,u_h) - R_{c,h}(\bar y_h,\bar u_h)\|_{X_h\times Y_h} \le \gamma_R\,\|(y_h,u_h) - (\bar y_h,\bar u_h)\|_{X\times Y}.$$

Analogously, if we define $M_{c,h}(\cdot) = (p_h)^\star \circ M_c(p_h^X(\cdot)) \circ p_h^X$, we find that

$$\|M_{c,h}(y_h,u_h) - M_{c,h}(\bar y_h,\bar u_h)\|_{\mathcal{L}(X_h\times Y_h)} \le \gamma_M\,\|(y_h,u_h) - (\bar y_h,\bar u_h)\|_{X\times Y}.$$

Thus, (4.7) follows with $\Gamma_R = \gamma_R$ and $\Gamma_M = \gamma_M$.

(c) The discretization family is stable in U_h if the uniform sufficient optimality condition

$$L_h''(y_h,u_h,\Lambda_h(y_h,u_h))(\phi_h,\phi_h) \ge \kappa^*\,\|\phi_h\|_X^2 \quad \text{for all } \phi_h \in N(e_h'(y_h,u_h)) \tag{5.12}$$

and the uniform Babuška–Brezzi condition

$$\inf_{\lambda_h \in Y_h\setminus\{0\}} \sup_{\phi_h \in X_h\setminus\{0\}} \frac{\langle e_h'(y_h,u_h)^\star\lambda_h, \phi_h\rangle_X}{\|\phi_h\|_X\,\|\lambda_h\|_Y} \ge \beta^* \tag{5.13}$$

hold for all $(y_h,u_h) \in U_h$ and constants $\kappa^* \ge 0$ and $\beta^* > 0$ that do not depend on h; see [GR86, p. 114]. If we define $L_h'' = (p_h^X)^\star \circ L_0''(p_h(\cdot)) \circ p_h^X$ and use (a), then (5.12) leads to

$$L''(y_h,u_h,\Lambda_h(y_h,u_h))(\phi_h,\phi_h) \ge \kappa^*\,\|\phi_h\|_X^2 \quad \text{for all } \phi_h \in N(e_h'(y_h,u_h)).$$

Hence, (5.12) follows if L_0'' is coercive in the whole space X. It is proved in [Vol00b] that (5.13) holds if the mesh-size is sufficiently small and if (y_h,u_h) is sufficiently close to $r_h^X(x^*,u^*)$.

TABLE 1
$R_{c,h}(y_h^n, u_h^n)$ for the Newton (Λ_N) and the least-squares update (Λ).

n	$R_{c,h}(y_h^n, u_h^n)$ for Λ_N	$R_{c,h}(y_h^n, u_h^n)$ for Λ
1	1.56e-01	1.62e-01
2	5.98e-02	4.33e-02
3	7.30e-03	9.15e-03
4	1.31e-03	1.29e-03
5	1.04e-04	5.26e-06
6	1.56e-09	1.37e-09
7	6.42e-14	5.87e-14

(d) In (5.11) we introduced the set V^*. Standard estimates for finite elements (see, for instance, [Hac92]) lead to

$$\|(y,u) - r_h^X(y,u)\|_X \le C_{10} h \quad \text{for all } (y,u) \in V^*.$$

Thus,

$$\begin{aligned}\|R_{c,h}(r_h^X(y,u)) - (p_h)^\star R_c(y,u)\|_{X\times Y} &\le \gamma_R \|p_h^X r_h^X(y,u) - (\bar y, \bar u)\|_{X\times Y} \\ &\le \gamma_R C_{10} h = C_{11} h,\end{aligned}$$

with $C_{11} = \gamma_R C_{10}$, so that the discretization family is consistent of order $s = 1$.

6. Numerical example. Let us consider the optimal control problem

$$\text{minimize } J(y,u) = \frac{1}{2}\,\|y - z\|_{L^2}^2 + \frac{\alpha}{2}\,\|u\|_{L^2(\Omega_\circ)}^2$$

with $z(x) = \sin(2\pi x)$, $\Omega = (0,1)$, $\Omega_\circ = (0, 0.5)$, and $\alpha = 0.001$

$$\text{subject to } -\nu y'' + yy' = \mathcal{E}(u) \quad \text{in } \Omega \text{ and } y(0) = 0, y(1) = 0,$$

where $\nu = 0.01$. The codes are written in MATLAB 5.3 and executed on a 550 MHz PC Pentium III. We apply the augmented Lagrangian-SQP method with $c = 1$ and compare the Newton update Λ_N with the least-squares update Λ given in (2.5). The two algorithms are stopped if

$$\|R_c(y_h^n, u_h^n)\|_{X\times Y} < 10^{-10}.$$

As initial iterates we choose $y_h^0 = z$ and $u_h^0 = 0$, and the step-size h is set equal to $\frac{1}{150}$. It turns out that both updates behave similarly; see Table 1.

Further, the second-order convergence rate in the variables (y_h^n, u_h^n) can be observed empirically in both cases. For this purpose we define

$$e_h(n) = \frac{\|(y_h^n, u_h^n) - (y_h^*, u_h^*)\|_{H^1\times L^2}}{\|(y_h^{n-1}, u_h^{n-1}) - (y_h^*, u_h^*)\|_{H^1\times L^2}^2} \quad \text{for } n \ge 2.$$

Then we observe in Table 2 that the term $e_h(n)$ is bounded, which yields the quadratic convergence rate of convergence.

An important feature of iterative approximation schemes for infinite-dimensional problems is mesh-independence. It asserts that the number or iterations to reach a

TABLE 2
$e_h(n)$ *for the Newton* (Λ_N) *and the least-squares update* (Λ).

n	$e_h(n)$ for Λ_N	$e_h(n)$ for Λ
1	1.33e − 01	1.33e − 01
2	3.63e − 01	2.80e − 01
3	3.98e + 00	5.90e + 00
4	1.66e − 00	3.43e − 01
5	1.44e − 00	5.89e + 00
6	4.29e − 01	1.29e + 00

TABLE 3
$n_h(\varepsilon)$ *for different* h.

$m = 1/h$	150	200	300	500	700	900	1300
$\varepsilon = 10^{-0}$	1	1	1	1	1	1	1
$\varepsilon = 10^{-1}$	2	2	2	2	2	1	1
$\varepsilon = 10^{-2}$	3	3	3	3	3	2	2
$\varepsilon = 10^{-3}$	5	4	4	4	4	4	4
$\varepsilon = 10^{-4}$	5	5	5	5	5	5	5
$\varepsilon = 10^{-5}$	5	5	5	5	5	5	5
$\varepsilon = 10^{-6}$	6	6	5	5	5	5	5
$\varepsilon = 10^{-7}$	6	6	6	6	6	6	6
$\varepsilon = 10^{-8}$	6	6	6	6	6	6	6
$\varepsilon = 10^{-9}$	7	7	6	6	6	6	6

certain approximation quality $\varepsilon > 0$ is independent of the mesh-size. Let us introduce the following notation:

$$n_h(\varepsilon) = \min \left\{ n_\circ | n \geq n_\circ : \|R_{c,h}(x_h^n)\|_{X \times Y} < \varepsilon \right\}.$$

From Table 3 one can observe empirically that Algorithm 3 has a mesh-independent behavior. For a proof we refer the reader to [Vol01].

7. Conclusions. In this paper we prove convergence results of the Kantorovich type for Lagrange-SQP methods. The theory is extended to the case of Lipschitz-continuous multiplier updates and is applied to obtain convergence rates for the discretized optimization problems. In addition, we obtain the existence of stationary points for the discretized problems. The assumptions are verified for an optimal control problem for the steady-state Burgers equation, and a numerical example illustrates the theoretical results. Numerical examples for the augmented Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates exhibit mesh-independence.

Appendix A. Nonlinear majorants. For more details and for the proofs we refer the reader to [OR70, Part V].

DEFINITION A.1. *Let* $\{x^n\}_{n \in \mathbb{N}}$ *be any sequence in a real Hilbert space* X. *Then a sequence* $\{t^n\}_{n \in \mathbb{N}} \subset [0, \infty)$ *for which*

$$\|x^{n+1} - x^n\|_X \leq t^{n+1} - t^n, \qquad n = 0, 1, \dots,$$

holds is a majorizing sequence *for* $\{x^n\}_{n \in \mathbb{N}}$.

Note that any majorizing sequence is necessarily monotonically increasing. In the proof of Theorem 3.2 a majorizing sequence occurs as the solution of a certain nonlinear difference equation. It is based on an estimate for $\Phi(\Phi(x)) - \Phi(x)$. The idea is presented in the following lemma.

LEMMA A.2. *Let X be a real Hilbert space, $\Phi : D \subseteq X \to X$, and $\varphi : J_1 \times J_2 \subset \mathbb{R}^2 \to [0,\infty)$, where each J_i is an interval of the form $[0,\alpha]$, $[0,\alpha)$, or $[0,\infty)$ and φ is monotonically increasing in each variable. Suppose that there is an element $x^0 \in D$ such that*

$$\|\Phi(\Phi(x)) - \Phi(x)\|_X \le \varphi\left(\|\Phi(x) - x\|_X, \|\Phi(x) - x^0\|_X\right) \tag{A.1}$$

holds whenever $x, \Phi(x) \in D$, and that with $t^0 = 0$, $t^1 = \|\Phi(x^0) - x^0\|_X$ the solution of the difference equation

$$t^{n+1} - t^n = \varphi(t^n - t^{n-1}, t^n), \qquad n = 1, 2, \dots, \tag{A.2}$$

exists and converges to $t^* < \infty$. Finally, assume that $B(x^0, t^*) \subset D$. Then the iterates $x^{n+1} = \Phi(x^n)$, $n = 0, 1, \dots$, are well defined, lie in $\overline{B(x^0, t^*)}$, converge to some $x^* \in \overline{B(x^0, t^*)}$, and satisfy

$$\|x^* - x^n\|_X \le t^* - t^n \quad \text{for all } n = 0, 1, \dots .$$

If $x^* \in D$ and Φ is continuous at x^*, then $x^* = \Phi(x^*)$.

Using (3.12) the difference equation (A.2) has the form

$$t^{n+1} - t^n = \frac{p_1\ (t^n - t^{n-1})^2 + p_2\ (t^n - t^{n-1})}{1 - p_3\ t^n} \quad \text{for all } n = 0, 1, \dots . \tag{A.3}$$

The following lemma gives sufficient conditions for the convergence of the sequence satisfying (A.3).

LEMMA A.3. *Assume that $p_1 > 0$, $1 > p_2 \ge 0$, $2p_1 = p_3$, and $0 \le \|\Phi(x^0) - x^0\|_X \le (1 - p_2)^2 (4p_1)^{-1}$. Then the sequence $\{t^n\}_{n\in\mathbb{N}}$ of (A.3) with initial values $t^0 = 0$ and $t^1 = \|\Phi(x^0) - x^0\|_X$ is strictly increasing unless $\|\Phi(x^0) - x^0\|_X = 0$, and*

$$\lim_{n\to\infty} t^n = t^* = \frac{1}{p_3}\left(1 - p_2 - \sqrt{(1 - p_2)^2 - 4p_1\ \|\Phi(x^0) - x^0\|_X}\right).$$

Appendix B. Linear boundary value problems of second order. Let $f \in L^2(\Omega)$, $y \in C(\overline{\Omega})$, and $\nu > 0$. We consider the linear boundary value problem

$$v'' + \frac{y}{\nu}\, v' = f \quad \text{in } \Omega, \quad v(0) = v(1) = 0. \tag{B.1}$$

Let

$$\varphi_0(x) = -\int_0^x \exp\left(-\frac{1}{\nu}\int_0^t y(s)\, ds\right) dt, \quad \varphi_1(x) = \int_x^1 \exp\left(-\frac{1}{\nu}\int_0^t y(s)\, ds\right) dt.$$

Observe that φ_0 and φ_1 are two linear independent fundamental solutions of the homogeneous problem $v'' + \frac{y}{\nu}\, v' = 0$ with $\varphi_0(0) = 0$ and $\varphi_1(1) = 0$. The Green's function associated with problem (B.2) is given by

$$G(x,t) = \begin{cases} \dfrac{\varphi_0(x)\varphi_1(t)}{\varphi_1(0)}, & 0 \le x \le t \le 1, \\[2ex] \dfrac{\varphi_0(t)\varphi_1(x)}{\varphi_1(0)}, & 0 \le t \le x \le 1. \end{cases} \tag{B.2}$$

For the proof of the next result we refer the reader to [Con78, p. 51].

PROPOSITION B.1. *Let $f \in L^2(\Omega)$ and let G denote the Green's function given by* (B.2). *Then there exists a unique solution $v \in H^2(\Omega) \cap H_0^1(\Omega)$ of* (B.1), *with the integral representation*

$$v(x) = \int_\Omega G(x,t) f(t)\, dt \quad \text{for } x \in \overline{\Omega}. \tag{B.3}$$

Remark B.2. Proposition B.1 allows us to give an H_0^1-estimate for the solution of (B.1):

$$\|v\|_{H_0^1} \le \frac{1}{|\varphi_1(1)|} \|\varphi_0\, \varphi_1\|_{L^\infty} \|f\|_{L^2} \le \exp\left(\frac{2}{\nu} \|y\|_{L^1}\right) \|f\|_{L^2}.$$

For $y \in C^1(\overline{\Omega})$ we consider the linear boundary value problem

$$v'' - \frac{1}{\nu}(yv)' = f \quad \text{in } \Omega, \quad v(0) = v(1) = 0. \tag{B.4}$$

Let

$$\varphi_0(x) = -\exp\left(\int_0^x \frac{y(s)}{\nu}\, ds\right) \int_0^x \exp\left(\int_t^0 \frac{y(s)}{\nu}\, ds\right) dt,$$

$$\varphi_1(x) = \exp\left(\int_0^x \frac{y(s)}{\nu}\, ds\right) \int_0^1 \exp\left(\int_t^1 \frac{y(s)}{\nu}\, ds\right) dt - \exp\left(\int_0^1 \frac{y(s)}{\nu}\, ds\right)$$
$$\cdot \exp\left(\int_0^x \frac{y(s)}{\nu}\, ds\right) \int_0^x \exp\left(-\int_0^t \frac{y(s)}{\nu}\, ds\right) dt.$$

Note that φ_0 and φ_1 are linear independent fundamental solutions of the homogeneous problem $v'' \frac{1}{\nu}(yv)' = 0$ with $\varphi_0(0) = 0$ and $\varphi_1(1) = 0$. Let the Green's function associated with problem (B.4) be given by

$$G(x,t) = \begin{cases} \dfrac{\varphi_0(x)\varphi_1(t)}{\varphi_1(1)}, & 0 \le x \le t \le 1, \\ \dfrac{\varphi_0(t)\varphi_1(x)}{\varphi_1(1)}, & 0 \le t \le x \le 1. \end{cases} \tag{B.5}$$

PROPOSITION B.3. *Let $f \in L^2(\Omega)$ and let G denote the Green's function given by* (B.5). *Then there exists a unique solution $v \in H^2(\Omega \cap H_0^1(\Omega)$ of* (B.4) *with the integral representation*

$$v(x) = \int_\Omega G(x,t) f(t)\, dt \quad \text{for } x \in \overline{\Omega}.$$

Remark B.4. Proposition B.3 allows us to give an estimate for the solution of (B.4) in the H_0^1-norm:

$$\|v\|_{H_0^1} \le \max_{(x,t)\in[0,1]^2} |G_x(x,t)|\, \|f\|_{L^2} \le c_\circ \|f\|_{L^2},$$

where

$$c_\circ = \exp\left(\frac{3}{\nu} \|y\|_{L^1}\right) \left(1 + \frac{\|y\|_{L^\infty}}{\nu}\right)^2 \left[1 + \frac{1}{|\varphi_1(1)|} \exp\left(\frac{2}{\nu} \|y\|_{L^1}\right)\right].$$

Acknowledgments. The authors thank both referees for helpful comments on the paper.

REFERENCES

[Aub72] J.-P. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Pure Appl. Math. 26, Wiley-Interscience, New York, 1972.

[Bre87] H. Brezis, *Analyse fonctionnelle. Théorie et Applications*, Masson, Paris, 1987.

[Con78] J. B. Conway, *A Course in Functional Analysis*, Graduate Texts in Math. 96, Springer-Verlag, New York, 1990.

[FST87] R. Fontecilla, T. Steihaug, and R. A. Tapia, *A convergence theory for a class of quasi-Newton methods for constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1133–1151.

[GR86] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[GMW81] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, New York, Toronto, Sydney, San Francisco, 1981.

[Hac92] W. Hackbusch, *Elliptic Differential Equations. Theory and Numerical Treatment*, Springer-Verlag, Berlin, 1992.

[IK96] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP-methods in Hilbert spaces and application to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96–125.

[Kle97] D. Kleis, *Augmented Lagrange SQP Methods and Application to the Sterilization of Prepackaged Food*, Ph.D. thesis, Fachbereich IV, Universität Trier, Trier, Germany, 1997.

[KS92] K. Kunisch and E. W. Sachs, *Reduced SQP methods for parameter identification problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1793–1820.

[KS97] D. Kleis and E. W. Sachs, *Convergence rate of the augmented Lagrangian SQP-method*, J. Optim. Theory Appl., 95 (1997), pp. 49–74.

[KV97] K. Kunisch and S. Volkwein, *Augmented Lagrangian-SQP techniques and their approximations*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, S. Cox and I. Lasiecka, eds., AMS, Providence, RI, 1997, pp. 147–159.

[Lig56] M. J. Lighthill, *Viscosity effects in sound waves of finite amplitude*, in Surveys in Mechanics, Cambridge University Press, Cambridge, UK, 1956, pp. 250–351.

[OR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Computer Science and Applied Mathematics, New York, London, Academic Press, 1970.

[Vol00a] S. Volkwein, *Mesh-independence for an augmented Lagrangian-SQP method in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 767–785.

[Vol00b] S. Volkwein, *Augmented Lagrangian-SQP techniques and optimal control problems for the stationary Burgers equation*, Comput. Optim. Appl., 16 (2000), pp. 57–81.

[Vol01] S. Volkwein, *Mesh-independence of Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates*, Optim. Methods Softw., to appear.

[Wou79] A. Wouk, *A Course of Applied Functional Analysis*, Wiley-Interscience, New York, 1979.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 254–281

AN ANALYSIS OF SMOOTHING EFFECTS OF UPWINDING STRATEGIES FOR THE CONVECTION-DIFFUSION EQUATION*

HOWARD C. ELMAN† AND ALISON RAMAGE‡

Abstract. Using a technique for constructing analytic expressions for discrete solutions to the convection-diffusion equation, we examine and characterize the effects of upwinding strategies on solution quality. In particular, for grid-aligned flow and discretization based on bilinear finite elements with streamline upwinding, we show precisely how the amount of upwinding included in the discrete operator affects solution oscillations and accuracy when different types of boundary layers are present. This analysis provides a basis for choosing a streamline upwinding parameter which also gives accurate solutions for problems with non-grid-aligned and variable speed flows. In addition, we show that the same analytic techniques provide insight into other discretizations, such as a finite difference method that incorporates streamline diffusion and the isotropic artificial diffusion method.

Key words. convection-diffusion equation, oscillations, Galerkin finite element method, streamline diffusion

AMS subject classifications. 65N22, 65N30, 65Q05, 35J25

PII. S0036142901374877

1. Introduction.

There are many discretization strategies available for the linear convection-diffusion equation

$$\begin{aligned} -\epsilon\nabla^2 u(x,y) + \mathbf{w}\cdot\nabla u(x,y) &= f(x,y) \qquad \text{in} \quad \Omega, \\ u(x,y) &= g(x,y) \qquad \text{on} \quad \delta\Omega, \end{aligned} \tag{1.1}$$

where the small parameter ϵ and divergence-free convective velocity field $\mathbf{w} = (w_1(x,y), w_2(x,y))$ are given. In this paper, we analyze some well-known methods which involve the addition of upwinding to stabilize the discretization for problems involving boundary layers. In particular, we focus on characterizing exactly how this upwinding affects the resulting discrete solutions.

A standard discretization technique is the Galerkin finite element method (see, for example, [5], [9], [10], [11], [13]). This is based on seeking a solution u of the weak form of (1.1),

$$\epsilon(\nabla u, \nabla v) + (\mathbf{w}.\nabla u, v) = (f, v) \qquad \forall\, v \in V,$$

where the test functions v are in the Sobolev space $V = \mathcal{H}_0^1(\Omega)$. Restricting this to a finite-dimensional subspace V_h of V gives

$$\epsilon(\nabla u_h, \nabla v) + (\mathbf{w}.\nabla u_h, v) = (f_h, v) \qquad \forall\, v \in V_h, \tag{1.2}$$

where f_h is the $L^2(\Omega)$ orthogonal projection of f into V_h and h is a discretization parameter. Choosing the test functions equal to a set of basis functions for V_h (usually

*Received by the editors June 18, 2001; accepted for publication (in revised form) January 10, 2002; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/37487.html

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (elman@cs.umd.edu). The work of this author was supported by National Science Foundation grant DMS9972490.

‡Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland (ar@maths.strath.ac.uk). The work of this author was supported by the Leverhulme Trust.

continuous piecewise polynomials with local support) leads to a sparse linear system whose solution can be used to recover the discrete solution u_h.

One quantity which has an important effect on the quality of the resulting discrete solution is the mesh Péclet number

$$P_e^{el} = \frac{h^{el}|\mathbf{w}|}{2\epsilon},$$

where h^{el} is a measure of element size and $|\mathbf{w}|$ represents the strength of the convective field within an element. In particular, if the mesh Péclet number is greater than one, then the discrete solution obtained from the Galerkin method may exhibit nonphysical oscillations. For the one-dimensional analogue of (2.1), this is well understood (see, for example, [10, p. 14]); for an analysis of the Galerkin discretization of the two-dimensional case, see [2]. An approach for minimizing the deleterious effects of these oscillations, especially in areas of the domain away from boundary layers, is to stabilize the discrete problem by using an upwind discretization. A particularly effective implementation of this idea is via the streamline diffusion method (see, e.g., [8], [9, sect. 9.7]). For linear or bilinear elements, the weak form (1.2) is replaced by

$$\epsilon(\nabla u_h, \nabla v) + (\mathbf{w}.\nabla u_h, v) + \sum \alpha^{el}(\mathbf{w}\cdot\nabla u_h, \mathbf{w}\cdot\nabla v)_{el} = (f_h, v) + \sum \alpha^{el}(f_h, \mathbf{w}\cdot\nabla v)_{el} \quad \forall v \in V_h, \tag{1.3}$$

where the sums are taken over all elements in the discretization. The stabilization parameters α^{el} are given by

$$\alpha^{el} = \frac{\delta^{el} h^{el}}{|\mathbf{w}|}, \tag{1.4}$$

where $\delta^{el} \geq 0$ are parameters to be chosen. Note that setting $\delta^{el} = 0$ on each element reduces (1.3) to the standard Galerkin case (1.2): this is the usual practice when $P_e^{el} < 1$. Formulation (1.3) has additional coercivity in the local flow direction, resulting in improved stability. More on the motivation behind this method can be found in [6, p. 289]. However, the best way of choosing δ^{el} for a general convection-diffusion problem is not known: for a discussion of this difficulty, see, for example, [13, Remark 3.34, p. 234].

In [2], we developed an analytic technique for characterizing the nature of oscillations in discrete solutions arising from the Galerkin discretization (1.2). More specifically, for the case of grid-aligned flow, we presented an analytic representation of the discrete solution, enabling isolation of any oscillatory behavior in the direction of the flow. Using this framework, we studied the dependence of solution behavior on the mesh Péclet number in some detail.

In this paper, we apply the tools developed in [2] to various upwinding strategies for discretizing (1.1). For the most part, we focus on the streamline diffusion method (1.3), examining the effect of stabilization on the quality of the resulting discrete solutions. In section 2, we summarize the Fourier analysis presented in [2] and derive an explicit formula for the discrete streamline diffusion solution for a model problem with constant grid-aligned flow. Section 3 contains the details of this process in the case of bilinear finite elements. The resulting formulae allow us to investigate various issues which influence the choice of stabilization parameters. In section 4, we characterize the effect of stabilization on oscillations in the discrete solution in the flow direction for three test problems whose solutions exhibit different types of

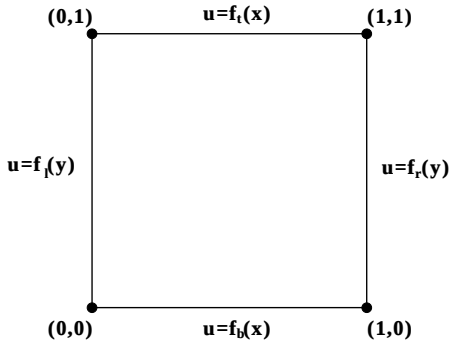


FIG. 1. *Boundary conditions.*

boundary layers. The implications of this analysis for solution accuracy are examined in section 5. In section 6, we discuss the relevance of our results for problems with non-grid-aligned and variable flow and present our recommended choice for the streamline diffusion parameters. Finally, in section 7, we illustrate how the same approach can be used to understand other discretization methods. We analyze an analogous streamline diffusion (upwind) discretization for a finite difference stencil and explain the comparative lack of effectiveness of isotropic artificial diffusion.

2. Summary of Fourier analysis. In this section, we summarize the Fourier techniques used in [2] to construct an analytic expression for the entries in the discrete solution vector $\mathbf{u}$.

Setting $\mathbf{w} = (0, 1)$ and $f = 0$ in (1.1), we obtain the "vertical wind" model problem

$$-\epsilon\nabla^2 u + \frac{\partial u}{\partial y} = 0 \qquad \text{in } \Omega = (0,1) \times (0,1), \tag{2.1}$$

with Dirichlet boundary conditions as shown in Figure 1. Using a natural ordering of the unknowns on a uniform grid of square bilinear elements with $N = 1/h$ elements in each dimension, both (1.2) and (1.3) give rise to a linear system

$$A\mathbf{u} = \mathbf{f}, \tag{2.2}$$

where the coefficient matrix A is of order $(N-1)^2$. Denoting the coefficients of the computational molecule by

$$\begin{array}{ccccc} m_4 & & m_3 & & m_4 \\ & \nwarrow & \uparrow & \nearrow & \\ m_2 & \leftarrow & m_1 & \rightarrow & m_2 \\ & \swarrow & \downarrow & \searrow & \\ m_6 & & m_5 & & m_6 \end{array}, \tag{2.3}$$

the matrix A can be written as

$$A = \begin{bmatrix} M_1 & M_2 & & & 0 \\ M_3 & M_1 & M_2 & & \\ & \ddots & \ddots & \ddots & \\ & & M_3 & M_1 & M_2 \\ 0 & & & M_3 & M_1 \end{bmatrix}, \tag{2.4}$$

where $M_1 = \text{tridiag}(m_2, m_1, m_2)$, $M_2 = \text{tridiag}(m_4, m_3, m_4)$, and $M_3 = \text{tridiag}(m_6, m_5, m_6)$ are all tridiagonal matrices of order $N-1$. Given that the eigenvalues and eigenvectors of the blocks of A satisfy

$$\begin{aligned} M_1\mathbf{v}_j &= \lambda_j \mathbf{v}_j, \qquad \lambda_j = m_1 + 2m_2 \cos\tfrac{j\pi}{N}, \\ M_2\mathbf{v}_j &= \sigma_j \mathbf{v}_j, \qquad \sigma_j = m_3 + 2m_4 \cos\tfrac{j\pi}{N}, \\ M_3\mathbf{v}_j &= \gamma_j \mathbf{v}_j, \qquad \gamma_j = m_5 + 2m_6 \cos\tfrac{j\pi}{N} \end{aligned} \tag{2.5}$$

for $j = 1, \ldots, N-1$, where the eigenvectors are

$$\mathbf{v}_j = \sqrt{\frac{2}{N}} \left[\sin\frac{j\pi}{N}, \quad \sin\frac{2j\pi}{N}, \quad \ldots, \sin\frac{(N-1)j\pi}{N} \right]^T, \tag{2.6}$$

we may obtain the decomposition

$$A = (\mathcal{V}P)T(\mathcal{V}P)^T, \tag{2.7}$$

where $\mathcal{V} = \text{diag}(V, V, \ldots, V)$ is a block diagonal matrix with each block V having the $N-1$ eigenvectors (2.6) as its columns, and P is a permutation matrix of order $(N-1)^2$. The matrix T is also block diagonal, with diagonal blocks $T_i = \text{tridiag}(\gamma_i, \lambda_i, \sigma_i)$, $i = 1, \ldots, N-1$. Using this decomposition and observing that P and V are both orthogonal, (2.2) implies

$$\mathbf{u} = \mathcal{V}P\mathbf{y}, \tag{2.8}$$

where the vector $\mathbf{y}$ is the solution to the linear system

$$T\mathbf{y} = P^T\mathcal{V}^T\mathbf{f} \equiv \hat{\mathbf{f}}. \tag{2.9}$$

As T is block diagonal, this system can be partitioned into $N-1$ independent systems of the form

$$T_i\mathbf{y}_i = \hat{\mathbf{f}}_i, \tag{2.10}$$

where T_i is defined above and $\mathbf{y}$ and $\hat{\mathbf{f}}$ are partitioned in the obvious way. Because T_i is a Toeplitz matrix, each of these systems can be considered as a three-term recurrence relation which can be solved analytically to give an expression for each entry y_{ik} of $\mathbf{y}_i$, $k = 1, \ldots, N-1$, in (2.10). Finally, to obtain an explicit formula for the entries of $\mathbf{u}$, we permute and transform these entries via (2.8) to get

$$u_{jk} = \sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin\frac{ij\pi}{N} y_{ik} \tag{2.11}$$

for $j, k = 1, \ldots, N-1$.

To obtain an expression for the entries y_{ik} in (2.11), we must consider the vectors $\hat{\mathbf{f}}_i$. As $f = 0$ in (2.1), the only nonzero entries in the original right-hand side vector $\mathbf{f}$ in (2.2) involve sums of certain matrix coefficients times boundary values, which are transformed and permuted to obtain $\hat{\mathbf{f}}$ in (2.9). The details of this process can be found in [2]. Here we simply state that each right-hand side vector $\hat{\mathbf{f}}_i$, $i = 1, \ldots, N-1$, in (2.10) can be written as

$$\hat{\mathbf{f}}_i = \begin{bmatrix} \bar{b}_i + \bar{s}_i \\ \bar{s}_i \\ \vdots \\ \bar{s}_i \\ \bar{t}_i + \bar{s}_i \end{bmatrix}_{N-1},$$

where $\bar{b}_i$ involves data from the bottom boundary values, $\bar{t}_i$ involves data from the top boundary values, and $\bar{s}_i$ combines information from the left and right boundary values. We will make the same assumption as in [2] that the functions $f_l(y)$ and $f_r(y)$ on the left and right boundaries are constant. This simplifies the presentation of the analysis.

The solution of each system (2.10) is now the solution of a three-term recurrence relation with constant coefficients whose auxiliary equation has roots

$$\mu_1(i) = \frac{-\lambda_i + \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}, \qquad \mu_2(i) = \frac{-\lambda_i - \sqrt{\lambda_i^2 - 4\sigma_i\gamma_i}}{2\sigma_i}. \tag{2.12}$$

The solution of this recurrence relation can be written as

$$y_{ik} = F_3(i) + [F_1(i) - F_3(i)]\, G_1(i,k) + [F_2(i) - F_3(i)]\, G_2(i,k), \tag{2.13}$$

where

$$G_1(i,k) = \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N},$$
$$G_2(i,k) = (1 - \mu_1^k) - (1 - \mu_1^N)\left[\frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N}\right],$$

and the functions

$$F_1(i) = -\frac{\bar{t}_i}{\sigma_i}, \qquad F_2(i) = \frac{\bar{s}_i}{\sigma_i + \lambda_i + \gamma_i}, \qquad F_3(i) = -\frac{\bar{b}_i}{\gamma_i}$$

involve the coefficient matrix entries and boundary condition information (see [2] for details).

We emphasize that the functions $F_m(i)$, $m = 1, 2, 3$, in (2.13) are independent of the vertical grid index k: for fixed i, the behavior of $\mathbf{y}$ in the streamline (vertical) direction depends only on the functions $G_1(i,k)$ and $G_2(i,k)$. In addition, as $F_1(i)$ is related to the top boundary values, $F_2(i)$ is related to the sum of the left and right boundary values (which have been assumed to be constant for this analysis), and $F_3(i)$ is related to the bottom boundary values, (2.13) shows that different boundary conditions will dictate how the functions $G_1(i,k)$ and $G_2(i,k)$ combine to produce different two-dimensional recurrence relation solutions y_{ik}. In the next section, we analyze the behavior of these solutions in some detail for the streamline diffusion finite element discretization (1.3) with bilinear elements.

3. Streamline diffusion discretization. In [2], an explicit expression for (2.13) for the Galerkin finite element method with bilinear elements was derived and analyzed. Here we present the equivalent analysis for the streamline diffusion finite element discretization (1.3) with a view to precisely characterizing the effect of the extra diffusion on the oscillations that occur with the Galerkin method when $P_e^{el} > 1$. We again use bilinear elements. Note that for a uniform grid and constant grid-aligned flow, $\delta = \delta^{el}$ is constant over all elements.

3.1. The recurrence relation solution. The coefficients in stencil (2.3) for a streamline diffusion discretization (1.3) using bilinear finite elements are given by

$$m_1 = \tfrac{4}{3}(\delta h + 2\epsilon), \qquad m_2 = \tfrac{1}{3}(\delta h - \epsilon), \qquad m_3 = -\tfrac{1}{3}[(2\delta - 1)h + \epsilon],$$

$$m_4 = -\tfrac{1}{12}[(2\delta - 1)h + 4\epsilon], \quad m_5 = -\tfrac{1}{3}[(2\delta + 1)h + \epsilon], \quad m_6 = -\tfrac{1}{12}[(2\delta + 1)h + 4\epsilon].$$

For convenience, we introduce the notation

$$C_i = \cos\frac{i\pi}{N}$$

and write the eigenvalues (2.5) as

$$\begin{aligned}
\gamma_i &= \frac{1}{6}\{-2[\delta h(2+C_i)+\epsilon(1+2C_i)]-h(2+C_i)\},\\
\lambda_i &= \frac{2}{3}\{[\delta h(2+C_i)+\epsilon(1+2C_i)]+3\epsilon(1-C_i)\},\\
\sigma_i &= \frac{1}{6}\{-2[\delta h(2+C_i)+\epsilon(1+2C_i)]+h(2+C_i)\},
\end{aligned}$$

$i = 1, \ldots, N-1$. Substituting these into (2.12) gives the expressions

$$\mu_{1,2} = \frac{-2\delta - \left[\dfrac{4-C_i}{2+C_i}\right]\dfrac{1}{P_e} \pm \sqrt{1+\dfrac{12\delta(1-C_i)}{(2+C_i)}\dfrac{1}{P_e}+\dfrac{3(5+C_i)(1-C_i)}{(2+C_i)^2}\dfrac{1}{P_e^2}}}{-2\delta+1-\left[\dfrac{1+2C_i}{2+C_i}\right]\dfrac{1}{P_e}} \tag{3.1}$$

for the auxiliary equation roots in (2.13).

3.2. Oscillations in the recurrence relation solution. We know from [2, Thm 5.1] that if $P_e > 1$, then the recurrence relation solution $\mathbf{y}$ and the related discrete solution $\mathbf{u}$ to the pure Galerkin problem (1.2) usually exhibit oscillations. In this section we address the question of how the streamline diffusion parameter δ can be chosen to eliminate oscillations in the recurrence relation solution $\mathbf{y}$. The issue of how this affects the resulting $\mathbf{u}$ will be discussed in section 3.3.

THEOREM 3.1. *If $P_e > 1$, then for any value of $i \in S_N \equiv \{1, \ldots, N-1\}$ there exists a parameter*

$$\delta_i^c = \frac{1}{2}\left(1 - \left[\frac{1+2C_i}{2+C_i}\right]\frac{1}{P_e}\right) \tag{3.2}$$

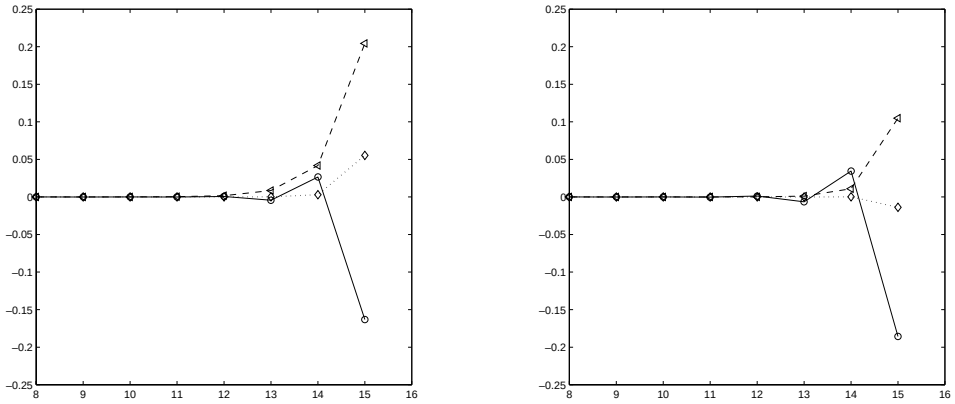
such that $\delta > \delta_i^c$ implies that $G_1(i,k)$ and $G_2(i,k)$ in (2.13) are nonoscillatory functions of k.

Proof. We have

$$G_1(i,k) = \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N} = \left[\frac{\left(\dfrac{\mu_1}{\mu_2}\right)^k - 1}{\left(\dfrac{\mu_1}{\mu_2}\right)^N - 1}\right]\mu_2^{k-N} = \Theta(i,k)\,\mu_2^{k-N}.$$

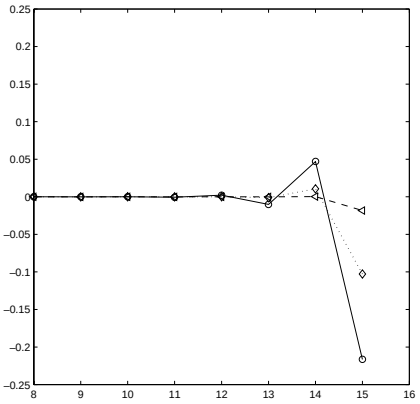
As $|\mu_1/\mu_2| < 1$, $\Theta(i,k)$ is always positive. Hence if μ_2 is negative, $G_1(i,k)$ alternates in sign as k goes from 1 to $N-1$, that is, $G_1(i,k)$ is oscillatory for fixed $i \in S_N$. From (3.1), the numerator of μ_2 is always negative so, for δ_i^c given by (3.2), we have the conditions

$$\begin{cases}
\delta > \delta_i^c \;\Rightarrow\; \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory,}\\[2ex]
\delta < \delta_i^c \;\Rightarrow\; \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory.}
\end{cases}$$



(a) $i = 1$.

(b) $i = N/2$.



(c) $i = N - 1$.

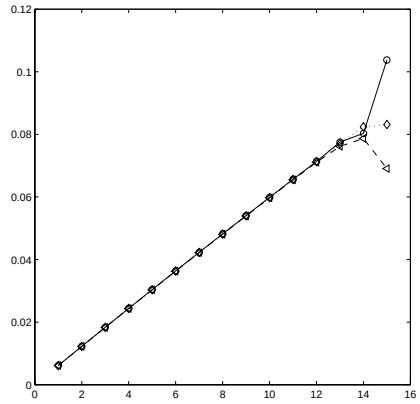
FIG. 2. *Plots of $G_1(i,k)$ against k for fixed i with $\delta = 0.2$ (solid, o), $\delta = 0.4$ (dotted, $\diamondsuit$), and $\delta = 0.6$ (dashed, $\triangle$).*

In addition, it can be shown that $0 < \mu_1 < 1$ so that if $G_1(i,k)$ is nonoscillatory, then $G_2(i,k) = (1-\mu_1^k) - (1-\mu_1^N)G_1(i,k)$ must also be nonoscillatory. □

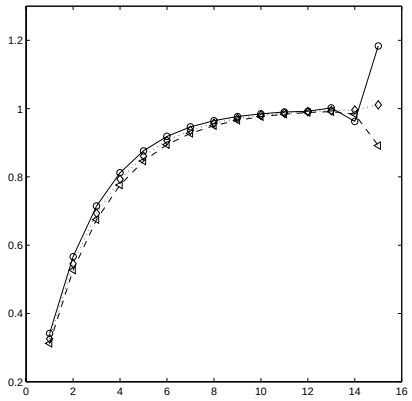
Sample plots of $G_1(i,k)$ for various values of $i \in S_N$ when $N = 16$ and $P_e = 3.125$ are given in Figure 2. Only the right half of the range of k has been plotted in each case to magnify the area of interest. Each subplot shows the behavior for three distinct values of δ, namely $\delta = 0.2$ (solid line, o), $\delta = 0.4$ (dotted line, $\diamondsuit$), and $\delta = 0.6$ (dashed line, $\triangle$). Given the relevant critical values $\delta_1^c \simeq 0.34$, $\delta_{N/2}^c \simeq 0.42$, and $\delta_{N-1}^c \simeq 0.65$ for this problem, the dependence of oscillations on the value of δ is clear. For $\delta = 0.2$ (that is, $\delta < \delta_i^c$ for all $i \in S_N$), all functions $G_1(i,k)$ are oscillatory; for $\delta = 0.4$, $G_1(1,k)$ is nonoscillatory (as $\delta > \delta_1^c$) and $G_1(N/2,k)$ is only very mildly oscillatory; for $\delta = 0.6$, only $G_1(N-1,k)$ is oscillatory (as $\delta > \delta_i^c$ for $i = 1, N/2$). Analogous behavior is seen in Figure 3 for $G_2(i,k)$ with the same parameter values, although the oscillations here occur about the function $1 - \mu_1^k$ rather than zero.

We now define

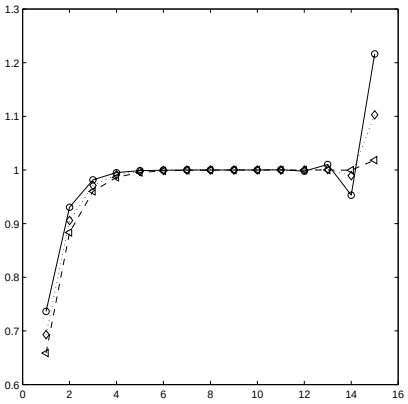
$$\delta_* = \frac{1}{2}\left(1 - \frac{1}{P_e}\right), \qquad \delta^* = \frac{1}{2}\left(1 + \frac{1}{P_e}\right) \tag{3.3}$$



(a) $i = 1$.



(b) $i = N/2$.



(c) $i = N - 1$.

FIG. 3. *Plots of $G_2(i,k)$ against k for fixed i with $\delta = 0.2$ (solid, o), $\delta = 0.4$ (dotted, $\diamondsuit$), and $\delta = 0.6$ (dashed, $\triangle$).*

(as in [3]) so that

$$\delta_* < \delta_i^c < \delta^* \tag{3.4}$$

for all values of $i \in S_N$. If $\delta \geq \delta^*$, then $\delta > \delta_i^c$ for each $i \in S_N$ and all of the functions $G_1(i,k)$ and $G_2(i,k)$ will be nonoscillatory in terms of k. We therefore have the following corollary to Theorem 3.1.

COROLLARY 3.2. *For any value of δ such that $\delta \geq \delta^*$, the functions $G_1(i,k)$ and $G_2(i,k)$ in* (2.13) *are nonoscillatory functions of k for every $i \in S_N$. Hence the recurrence relation solution $\mathbf{y}$ is a sum of smooth functions and will not exhibit oscillations in the streamline direction.*

The case $\delta = \delta_i^c$ requires special attention. With this value, $\sigma_i = 0$ in (2.5) and the resulting matrix T_i in (2.10) is bidiagonal. This leads to a two-term recurrence relation with auxiliary equation root

$$\rho = \frac{1}{1 + \dfrac{3(1 - C_i)}{2 + C_i} \dfrac{1}{P_e}}$$

and solution

$$y_{ik} = F_3(i)\rho^k + F_2(i)(1-\rho^k). \tag{3.5}$$

As $0 < \rho < 1$ for any $i \in S_N$, y_{ik} is nonoscillatory in the streamline direction. In addition, $\rho \to 1$ as $P_e \to \infty$, giving the solution $y_{ik} = F_3(i)$. Looking ahead to section 3.3, applying transformation (2.11) gives $u_{jk} = f_b(x_j)$ (see (3.8)). This is the solution to the reduced problem (obtained by setting $\epsilon = 0$ in (2.1)) where the bottom boundary values are simply transported in the direction of the flow without any diffusion present. That is, with the choice $\delta = \delta_i^c$ for each i, the discrete solution is exact at every interior node in the limit as $P_e \to \infty$.

3.3. Oscillations in the discrete solution. In this section we consider the impact of transformation (2.11) on the recurrence relation solution $\mathbf{y}$, with a view to choosing δ to obtain an oscillation-free discrete solution $\mathbf{u}$. We begin by considering the functions $F_m(i)$, $m = 1, 2, 3$, in (2.13). Following the analysis of [2, sect. 4.4 and appendix] we can derive the following expressions

$$\begin{aligned} F_1(i) &= \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_t(x_s) \sin \frac{si\pi}{N}, \\ F_2(i) &= f_l \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N}, \\ F_3(i) &= \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin \frac{si\pi}{N} \end{aligned} \tag{3.6}$$

for the streamline diffusion weight functions in the special case where the constant left and right boundary values f_l and f_r are equal. From (2.13), we therefore have

$$\begin{aligned} y_{ik} = &\sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin \frac{si\pi}{N} + \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} [f_t(x_s) - f_b(x_s)] \sin \frac{si\pi}{N} G_1(i,k) \\ &+ \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} [f_l - f_b(x_s)] \sin \frac{si\pi}{N} G_2(i,k) \end{aligned} \tag{3.7}$$

[2, Thm 4.2]. Note that the expressions in (3.6) hold for any stencil of the form (2.3) whose entries sum to zero. In particular, this implies that the functions in (3.6) are the same for discretizations (1.2) and (1.3).

We now apply transformation (2.11) to (3.7) to obtain an expression for the entries of the discrete solution vector $\mathbf{u}$. As in [2], for the first term we have

$$\sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin \frac{ij\pi}{N} \left\{ \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} f_b(x_s) \sin \frac{si\pi}{N} \right\} = f_b(x_j), \tag{3.8}$$

where $f_b(x)$ is the bottom boundary function in Figure 1. Applying (2.11) to the full expression (3.7) therefore gives

$$u_{jk} = f_b(x_j) + \frac{2}{N} \sum_{i=1}^{N-1} \left[a_{ij} G_1(i,k) + b_{ij} G_2(i,k) \right], \tag{3.9}$$

where

$$a_{ij} = \sin\frac{ij\pi}{N}\sum_{s=1}^{N-1}[f_t(x_s) - f_b(x_s)]\sin\frac{si\pi}{N},$$

(3.10)

$$b_{ij} = \sin\frac{ij\pi}{N}\sum_{s=1}^{N-1}[f_l - f_b(x_s)]\sin\frac{si\pi}{N}.$$

That is, along a streamline (j fixed), $\mathbf{u}$ consists of the bottom boundary value on that line plus a linear combination of the functions $G_1(i,k)$ and $G_2(i,k)$ for $i \in S_N$. Note that $a_{i(N-j)} = a_{ij}$ and $b_{i(N-j)} = b_{ij}$, so that if $f_b(x)$ is symmetric about the center vertical line of the grid, then so is $\mathbf{u}$.

We can use the representation (3.9) to obtain insight into the effect of δ on the quality of the solution in the streamline direction. Recall from section 3.2 that if $\delta \geq \delta_i^c$ in (3.2), then the functions $G_1(i,k)$ and $G_2(i,k)$ are nonoscillatory in the streamline direction for that particular $i \in S_N$. It follows from Corollary 3.2 that if $\delta \geq \delta^*$ in (3.3), then (3.9) is a sum of smooth functions. We have therefore established a sufficient condition for the discrete solution to be nonoscillatory.

THEOREM 3.3. *For a streamline diffusion discretization of* (2.1) *with bilinear finite elements, the discrete solution* $\mathbf{u}$ *does not exhibit oscillations in the streamline direction when* $\delta \geq \delta^*$.

4. Analysis of boundary layer effects. In practice, it turns out that the restriction on δ given by Theorem 3.3 is too harsh, and better solutions can be obtained using values of δ smaller than δ^* due to the "smoothing" nature of transformation (2.11). The precise effect of this transformation in the context of the behavior of the Galerkin finite element solution for different mesh Péclet numbers was studied in [2]. Here we present a discussion of the effects of varying δ in the streamline diffusion method. We illustrate the ideas with three examples containing different types of boundary layers. The first two examples contain an exponential layer at the outflow and parabolic layers along the characteristic (vertical) boundaries, respectively. The third example has a Neumann boundary condition at the outflow, and we show that the analysis generalizes to this case.

Throughout this section we will use notation based on considering u_{jk} in (3.9) as a sum of smooth and oscillatory parts. That is, letting i^* be the lowest value of $i \in S_N$ such that $\delta < \delta_i^c$, we write

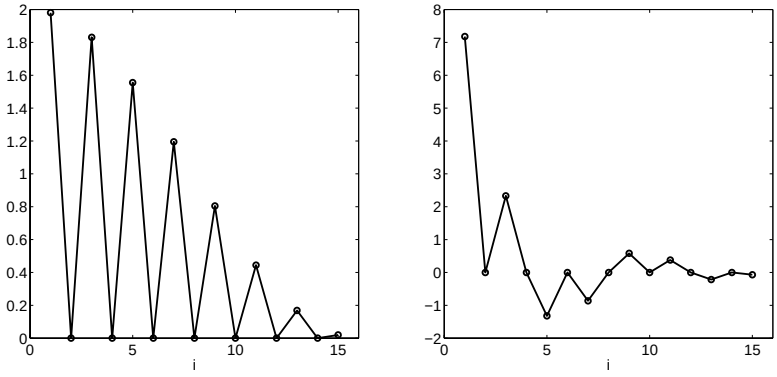
(4.1)

$$\begin{aligned} u_{jk} &= f_b(x_j) + \frac{2}{N}\left(\sum_{i=1}^{i^*-1}[a_{ij}G_1(i,k) + b_{ij}G_2(i,k)] + \sum_{i=i^*}^{N-1}[a_{ij}G_1(i,k) + b_{ij}G_2(i,k)]\right) \\ &= f_b(x_j) + S_{\text{smooth}} + S_{\text{osc}}. \end{aligned}$$

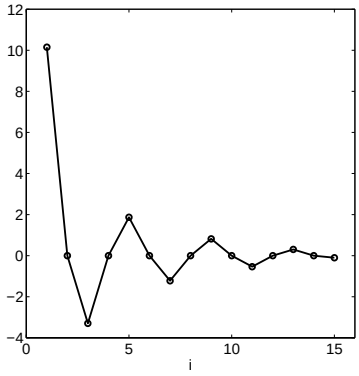
Note that the preceding analysis implies $S_{\text{smooth}} = 0$ when $\delta \leq \delta_*$ and $S_{\text{osc}} = 0$ when $\delta \geq \delta^*$. As δ increases from δ_*, i^* will increase so that S_{smooth} contains more and more of the terms, with the overall smoothness of $\mathbf{u}$ dependent on the relative size of the two sums S_{smooth} and S_{osc}.

Problem I. In this example we apply the Dirichlet boundary conditions

$$f_t(x) = 1, \qquad f_b(x) = f_l(y) = f_r(y) = 0,$$



(a) $j = 1$. (b) $j = N/4$.



(c) $j = N/2$.

FIG. 4. *Plots of coefficients a_{ij} against i for $N = 16$.*

as per Figure 1, so that the solution has an exponential boundary layer of width ϵ along the top boundary. For this problem, (3.7) implies

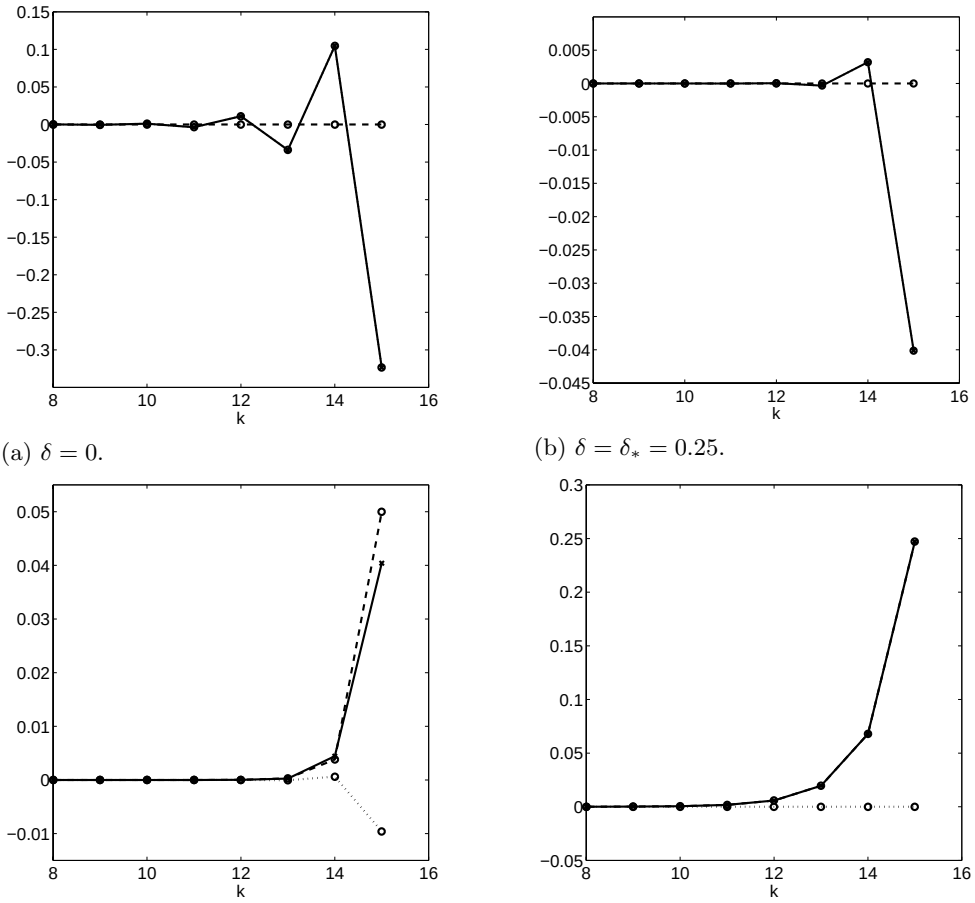
$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} G_1(i,k) \tag{4.2}$$

so the coefficients in (3.10) simplify to

$$a_{ij} = \sin \frac{ij\pi}{N} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N}, \qquad b_{ij} = 0, \tag{4.3}$$

with the magnitude of each a_{ij} decreasing rapidly as i goes from 1 to $N-1$ as shown in Figure 4 (taken from [2]). This means that the contributions to u_{jk} from the functions $G_1(i,k)$ are much larger for small indices i, so that the smoothness of $G_1(i,k)$ for small i plays a much more important role. In particular, it is not necessary for $G_1(i,k)$ to be nonoscillatory for all $i \in S_N$ in order for $|S_{\text{smooth}}|$ to dominate $|S_{\text{osc}}|$ and the resulting function $\mathbf{u}$ to be smooth.

We illustrate these ideas in Figures 5 and 6 for this example problem with $N = 16$ and $P_e = 2$. The first figure shows u_{1k} (or, equivalently, $u_{(N-1)k}$) plotted against k. This is the vertical cross-section of the solution obtained by fixing $j = 1$, which is



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

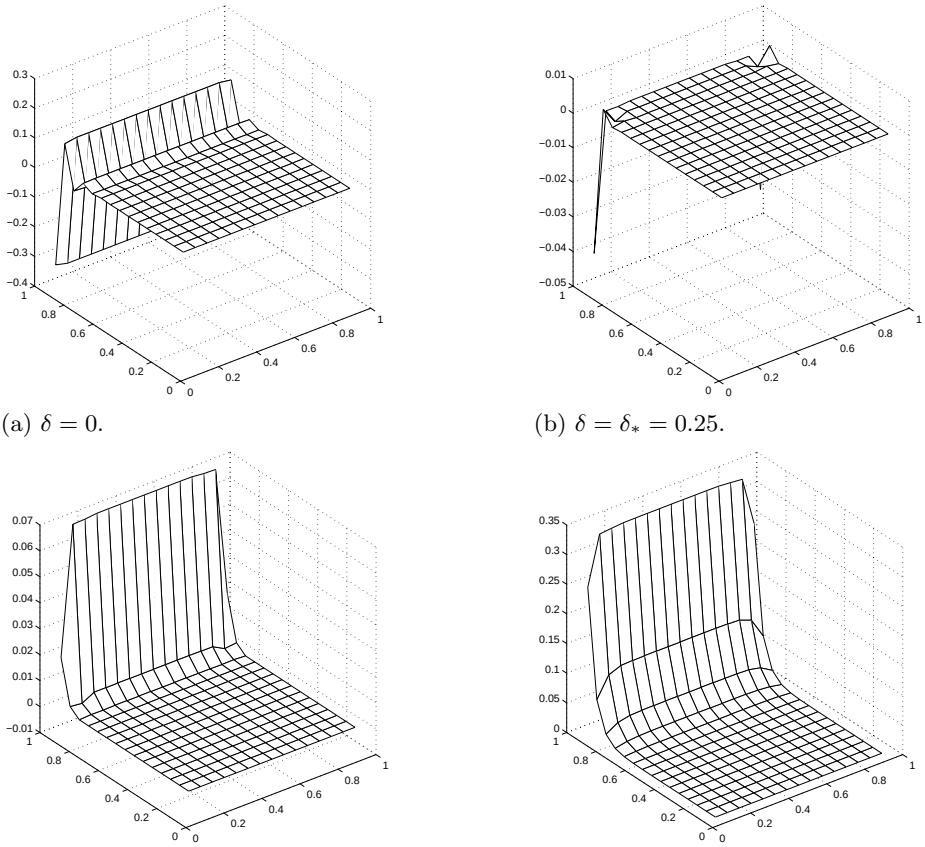
(c) $\delta = \delta_{\rm s} = 0.354$.

(d) $\delta = \delta^* = 0.75$.

FIG. 5. *Comparison of* $S_{\rm smooth}$ *(dashed line, o) and* $S_{\rm osc}$ *(dotted line, o) with* u_{1k} *(solid line, x) for Problem* I.

the most oscillatory of the vertical cross-sections for this problem. Each plot shows a comparison of $S_{\rm smooth}$ (dotted line, o) and $S_{\rm osc}$ (dashed line, o) with u_{1k} (solid line, x) for a different value of δ, where again only the right half of the range of k has been plotted to magnify the area of interest. For this example, $\delta_* = 0.25$ and $\delta^* = 0.75$. Plot (a) shows the Galerkin case ($\delta = 0$) where all of the functions $G_1(i,k)$ are oscillatory and $S_{\rm smooth}$ is zero. This is still true in plot (b), where $\delta = \delta_*$, but the magnitude and extent of the oscillations has been reduced considerably. The result of choosing $\delta = \delta^*$ according to Theorem 3.3 to guarantee an oscillation-free discrete solution by ensuring a nonoscillatory $\mathbf{y}$ is shown in plot (d). Here too much extra diffusion has been added. Plot (c) shows u_{1k} for $\delta = \delta_{\rm s} = 0.354$, which lies in the interval (δ_7^c, δ_8^c), that is, $i^* = 8$. This is the lowest value of i^* such that $S_{\rm smooth}$ dominates (3.9) for this problem and u_{1k} is nonoscillatory.

The corresponding full two-dimensional solutions $\mathbf{u}$ are shown in Figure 6, where the boundary values have been omitted so that the fine detail of each solution is visible. The overall behavior corresponds to that seen from the cross-sections: the severe oscillations present when $\delta = 0$ are almost eliminated by choosing $\delta = \delta_*$, and



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

(c) $\delta = \delta_s = 0.354$.

(d) $\delta = \delta^* = 0.75$.

FIG. 6. *Discrete solution at interior node points for Problem* I *with* $N = 16$, $P_e = 2$.

setting $\delta = \delta^*$ gives a smooth but overly diffuse solution. For $\delta = \delta_s$, the oscillations along the lines u_{1k} and $u_{(15)k}$ have just been eliminated to give a completely smooth solution in the flow direction.

Problem II. Next we consider the Dirichlet boundary conditions

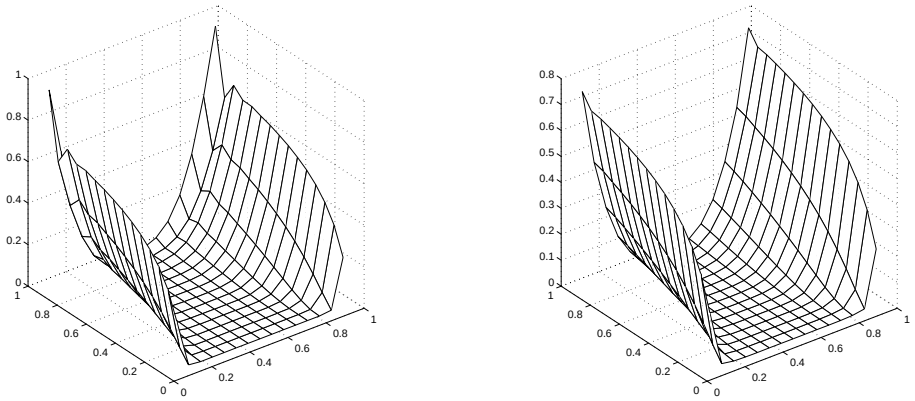
$$f_b(x) = f_t(x) = 0, \qquad f_l(y) = f_r(y) = 1,$$

which result in a solution which has parabolic layers on both vertical sides of the domain. The recurrence relation solution is

$$y_{ik} = \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} G_2(i,k), \tag{4.4}$$

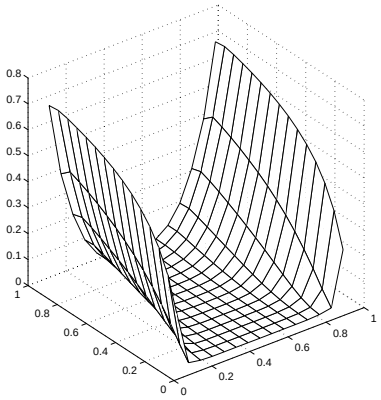
which is the same as for Problem I, except with G_2 in place of G_1 (see (4.2)). In addition, the coefficients in the full solution (3.10) are identical to those in Problem I as given by (4.3). The analysis for this problem is therefore very similar. In particular, as observed in section 3.2, G_2 is oscillatory if and only if G_1 is oscillatory, so exactly the same argument applies as to the effect of δ on solution quality.

Sample solutions for $N = 16$ with $P_e = 2$ are shown in Figure 7. These plots

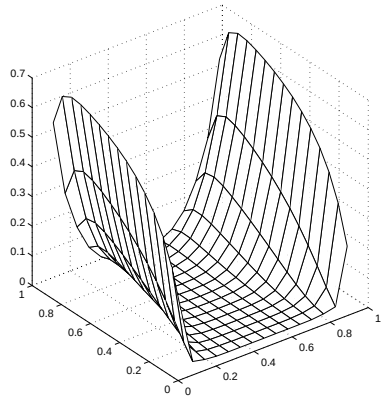


(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.



(c) $\delta = \delta_s = 0.354$.



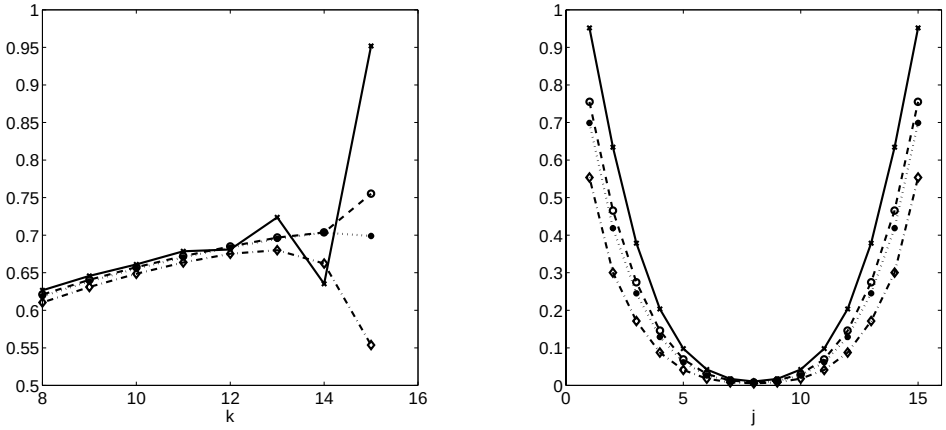
(d) $\delta = \delta^* = 0.75$.

FIG. 7. *Discrete solution at interior node points for Problem* II *with* $N = 16$, $P_e = 2$.

show the effect of increasing δ on the solution in the streamline direction: again, the solutions with $\delta = 0$ and $\delta = \delta_*$ exhibit oscillations while the solution with $\delta = \delta^*$ is overly diffuse. The value δ_s is the first for which the smooth part dominates to give a smooth solution. Figure 8 shows cross-sections of these plots for fixed values $j = 1$ on the left and $k = 15$ on the right.

It is known that parabolic layers such as those exhibited by the solution of this problem are wider than the exponential layers of the previous example (the widths are proportional to $\sqrt{\epsilon}$ and ϵ, respectively [13]). Oscillations transverse to the flow caused by inadequate resolution of parabolic layers will occur, but only for mesh Péclet numbers much larger than in the examples shown. However, the results given here demonstrate that *streamwise* effects also cause difficulties for problems with parabolic layers. The analysis shows that these are manifested in Problem II by the presence of G_2 in the solution and that streamline upwinding ameliorates these difficulties by making $G_2(i, \cdot)$ smoother for enough indices i. The right-hand plot in Figure 8 also shows that excessive diffusivity in the streamline direction gives the appearance of smearing of the characteristic layers.

Problem III. For this example, we replace the Dirichlet boundary condition $u = f_t(x)$ on the top boundary in Problem I by the Neumann boundary condition $\frac{\partial u}{\partial n} = 1$.



(a) $j = 1$. (b) $k = N - 1$.

FIG. 8. *Cross-sections of solutions to Problem* II *for* $N = 16$*;* $P_e = 2$ *for* $\delta = 0$ *(solid line,* $\times$*),* $\delta = \delta_*$ *(dashed line,* $\circ$*),* $\delta = \delta_s$ *(dotted line,* $*$*), and* $\delta = \delta^*$ *(dot-dash line,* $\diamondsuit$*).*

The other Dirichlet boundary conditions remain the same. The analysis of section 2 needs to be modified slightly to handle this case. There are now $N(N-1)$ unknowns, and the coefficient matrix A in (2.4) is replaced by

$$A^\star = \begin{bmatrix} M_1 & M_2 & & & 0 \\ M_3 & M_1 & M_2 & & \\ & \ddots & \ddots & \ddots & \\ & & M_3 & M_1 & M_2 \\ 0 & & & M_3 & M_1^\star \end{bmatrix},$$

where there are N rows of $(N-1) \times (N-1)$ blocks. For bilinear finite elements on a square mesh, $M_1^\star = \text{tridiag}(m_2^\star, m_1^\star, m_2^\star)$ with entries

$$m_1^\star = \frac{1}{3}[(2\delta + 1)h + 4\epsilon], \qquad m_2^\star = -\frac{1}{12}[(2\delta - 1)h + 2\epsilon].$$

As the vectors $\mathbf{v}_j$ in (2.6) are eigenvectors of $M_1^\star$, we may construct a matrix $\mathcal{V}^\star$ with N copies of V on its diagonal and a permutation matrix $P^\star$ of order $N(N-1)$ such that a decomposition of type (2.7) exists. The associated block tridiagonal matrix $T^\star$ has $N-1$ diagonal blocks, each one of the form

$$T_i^\star = \begin{bmatrix} \lambda_i & \sigma_i & & & 0 \\ \gamma_i & \lambda_i & \sigma_i & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_i & \lambda_i & \sigma_i \\ 0 & & & \gamma_i & \lambda_i^\star \end{bmatrix}_{N \times N},$$

where

$$\lambda_i^\star = m_1^\star + 2m_2^\star \cos\frac{i\pi}{N}, \qquad i = 1, \ldots, N-1,$$

are the eigenvalues of $M_1^\star$. Similarly, the transformed right-hand side vector $\hat{\mathbf{f}}^\star$ can be partitioned into $N-1$ vectors of length N to give $N-1$ independent systems

$$T_i^\star \mathbf{y}_i = \hat{\mathbf{f}}_i^\star. \tag{4.5}$$

For this specific example, the vectors $\hat{\mathbf{f}}_i^\star$ are given by

$$\hat{\mathbf{f}}_i^\star = \epsilon h \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} \end{bmatrix}_N .$$

The solution of each system (4.5) is therefore the solution of the same constant-coefficient recurrence relation as in the Dirichlet case, but with the right-hand boundary condition now of Neumann type. The roots of the auxiliary equation are given by (2.12), and the recurrence relation solution is

$$y_{ik}^\star = \epsilon h \sqrt{\frac{2}{N}} \sum_{s=1}^{N-1} \sin \frac{si\pi}{N} G_1^\star(i,k), \tag{4.6}$$

where

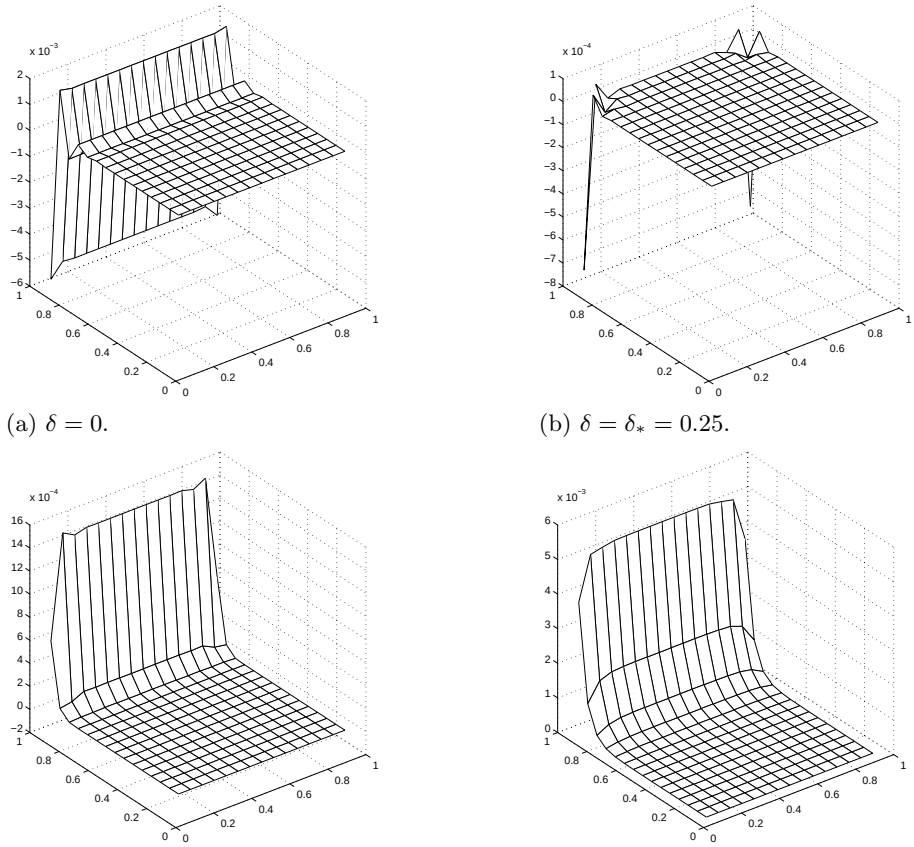
$$G_1^\star(i,k) = \frac{\mu_1^k - \mu_2^k}{(\gamma_i + \lambda_i^\star \mu_1)\mu_1^{N-1} - (\gamma_i + \lambda_i^\star \mu_2)\mu_2^{N-1}}.$$

This expression compares with (4.2) in the Dirichlet case. The most significant difference is the factor of ϵh in front of the Neumann solution: this means that for this problem the oscillations will be much smaller than those in the Dirichlet case. Because of the nature of $G_1^\star$ and G_1, however, the effect of changing δ will be very similar in both cases. This is borne out by the plots of the Neumann solution shown in Figure 9 (for $N = 16$ and $P_e = 2$ so that $\epsilon h = 9.8 \times 10^{-4}$). As predicted by the analysis, these solutions are almost identical in shape to those obtained for the Dirichlet problem (see Figure 6), but any oscillations are much smaller in magnitude.

5. Solution accuracy. We have now characterized the effect of δ on oscillations in the flow direction. One important question which remains is how the choice of δ affects the overall accuracy of the discrete solution. To investigate this, we begin with the example problems of the previous section. In each case, we compare solutions on a 16×16 grid with $\epsilon = 1/64$ (so $P_e = 2$) with a reference solution for the same value of ϵ on a 256×256 grid. On this fine grid, we use the Galerkin method ($\delta = 0$) as $P_e = 0.125 \ll 1$ and there are no oscillations. In what follows, we will denote the fine grid nodal solution vector by $\mathbf{u}_{256}$ and its associated finite element solution by u_{256}, likewise for the coarse grid solutions $\mathbf{u}_{16}^\delta$ and u_{16}^δ.

Figure 10 shows the variation with δ of the error for our test problems measured in two different norms. In all cases the norm of the error is plotted against δ for $0 \le \delta \le 1$ with the values of δ_* (o), δ_s ($\diamondsuit$), and δ^* (x) highlighted. For $P_e = 6.25$ ($\epsilon = 1/200$), $\delta_* = 0.42$, $\delta_s = 0.468$, and $\delta^* = 0.58$. The solid line represents the discrete $L_\infty[0,1]$ norm defined by

$$\|\mathbf{u}_{256} - \mathbf{u}_{16}^\delta\|_\infty = \max_{i,j} |\mathbf{u}_{256}(x_i, y_j) - \mathbf{u}_{16}^\delta(x_i, y_j)|,$$



(a) $\delta = 0$.

(b) $\delta = \delta_* = 0.25$.

(c) $\delta = \delta_s = 0.354$.

(d) $\delta = \delta^* = 0.75$.

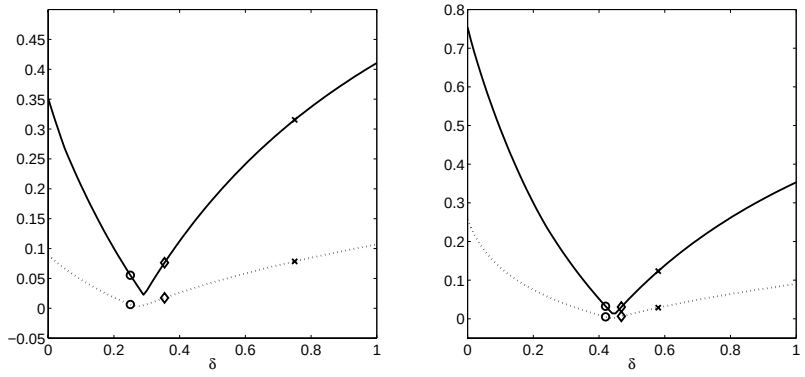
FIG. 9. *Discrete solution at interior node points for Problem* III *with* $N = 16$, $P_e = 2$.

where the points $(x_i, y_j) = (ih, jh)$ are the nodes of the 16×16 grid. When using the finite element method, it may be more natural to work with the L_2 norm

$$\|u_{256} - u_{16}^{\delta}\|_2 = \left\{ \int_{\Omega} \left(u_{256} - u_{16}^{\delta} \right)^2 \right\}^{\frac{1}{2}}. \tag{5.1}$$

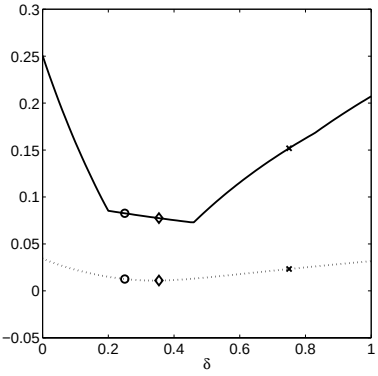
However, this measure leads to misleading results for certain singular perturbation problems of this type where the overall error is heavily dominated by the error in the boundary layer, which we cannot hope to resolve on a 16×16 uniform grid using low order elements. For Problems I and III, a more meaningful measure of the error for our purposes is obtained using the L_2 norm of the error away from the boundary layer; that is, in these cases, we omit the top row of coarse grid elements from the region of integration in (5.1) and integrate over $(0,1) \times (0, 0.9375)$ instead of $\Omega = (0,1) \times (0,1)$. This norm is represented by a dotted line in the error plots. We note in passing that in all of the examples, this curve is very similar to that obtained for the discrete L_2 norm defined by

$$\|\mathbf{u}_{256} - \mathbf{u}_{16}^{\delta}\|_2 = \left\{ \sum_{i,j=0}^{N} \left(u_{256}(x_i, y_j) - u_{16}^{\delta}(x_i, y_j) \right)^2 \right\}^{\frac{1}{2}},$$

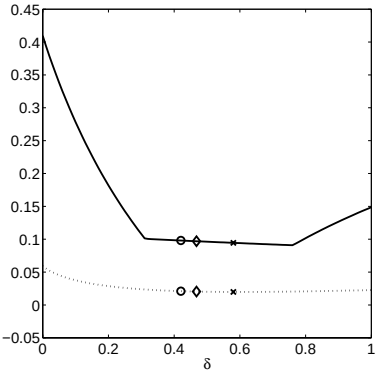


(a) Problem I: $P_e = 2$.

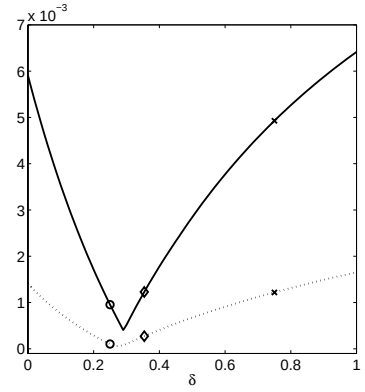
(d) Problem I: $P_e = 6.25$.



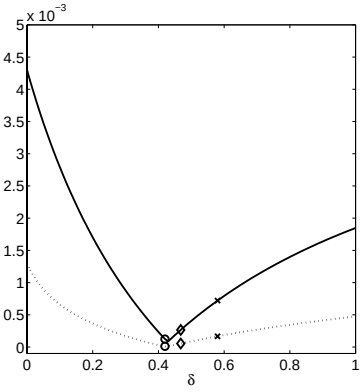
(b) Problem II: $P_e = 2$.



(e) Problem II: $P_e = 6.25$.



(c) Problem III: $P_e = 2$.



(f) Problem III: $P_e = 6.25$.

FIG. 10. *Error variation with δ in the discrete L_∞ norm (solid) and L_2 norm (dotted) for $N = 16$.*

where (x_i, y_j) is again a node of the coarse grid.

From Figure 10, we see that the optimal choice of δ in terms of solution accuracy depends on the norm in which the error is measured, although in most cases both δ_* and δ_s are closer than δ^* to the optimal choice. Note that setting $\delta = \delta_s$ to produce a completely oscillation-free discrete solution $\mathbf{u}$ does not result in the most accurate solution.

6. Guidelines for choosing the streamline diffusion parameter in practice. In sections 2–4, we presented model problem analysis which enabled us to characterize the behavior of the discrete finite element solutions. Three highlighted values of δ play important roles in this analysis: δ_*, where the solution is oscillatory but the oscillations are extremely small; δ_s, which is the smallest value of δ such that the solution is found by numerical experiment to be oscillation free; and δ^*, where the solution is guaranteed to be oscillation free via Theorem 3.3. The analysis, based on Fourier techniques, is restricted to grid-aligned flow. (This is needed for the tridiagonal matrices M_1, M_2, and M_3 of (2.4) to be symmetric and have a common set of eigenvectors.) In this section, we consider several more complex problems and make some observations about choosing δ in practice.

First, we observe that although with $\delta = \delta^*$ we have a way of guaranteeing that there are no oscillations, the resulting discrete solutions are overly diffuse and inaccurate: both δ_* and $\delta_{\rm s}$ are in general much better values to use. The choice $\delta = \delta_{\rm s}$ produces a completely oscillation-free solution but $\delta_{\rm s}$ is not readily determined even for the model problems considered above. However, we know that δ_s lies between δ_* and δ^*, and the empirical results for Problems I–III suggest that the computable expression

$$\delta_\bullet = \frac{1}{2}\left(1 - \frac{0.8}{P_e}\right) \tag{6.1}$$

is a good approximation to it. Note that in the limit as $P_e \to \infty$, both δ_* and $\delta_\bullet$ tend to 0.5.

We now introduce three new test problems with non-grid-aligned or variable winds. For these problems, we use a stabilization strategy which fixes δ^{el} locally on each element by using the local element mesh Péclet number

$$P_e^{el} = \frac{h^{el}\|\mathbf{w}^{el}\|_2}{2\epsilon}$$

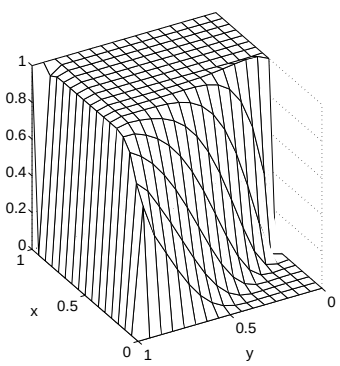
in formulae (3.3) and (6.1). This is calculated using the discrete L_2 norm of the wind value at the element center $\mathbf{w}^{el}$, with the local grid size value h^{el} taken as the distance across the element measured in the direction of the wind. In what follows, these element-based values of δ will be denoted using the superscript el. In all cases, the value of the stabilization parameter used is $\max(\delta^{el}, 0)$ on each element.

Problem IV. Here we impose the Dirichlet boundary conditions

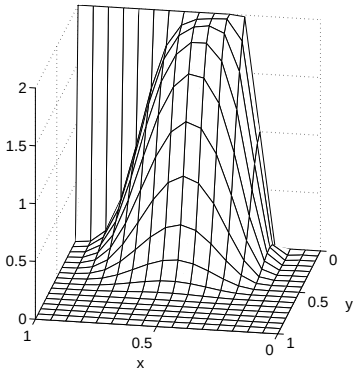
$$f_b(x) = \begin{cases} 0, & 0 < x \le \frac{1}{2}, \\ 1, & \frac{1}{2} < x < 1, \end{cases} \qquad f_t(x) = f_l(y) = 0, \quad f_r(y) = 1$$

on the domain in Figure 1 and apply the wind $\mathbf{w} = (\cos 115^\circ, \sin 115^\circ)$ which has constant magnitude and direction but is not aligned with the grid. This problem has an exponential boundary layer on a portion of the outflow boundary and an internal layer along the characteristic caused by the discontinuity on the inflow boundary. A sample solution with $N = 16$, $\epsilon = 1/200$, and $\delta = \delta_*^{el}$ is shown in Figure 11 (a).

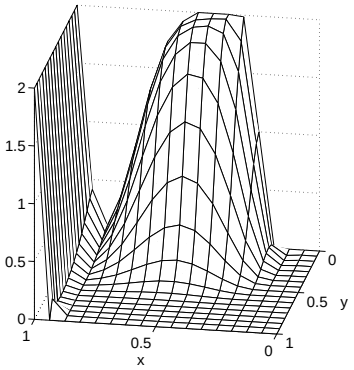
Error calculations carried out as described in the previous section lead to the plots in Figure 12 (a) and (d), where the values δ_*^{el} ($\circ$), $\delta^{el,*}$ ($\times$), and $\delta_\bullet^{el}$ ($\diamondsuit$) have been highlighted. When $\delta = 0$, the error is dominated by difficulties associated with the exponential layers at the outflow. As δ is increased so that these layers begin to be resolved, the error is then dominated by the effect of the discontinuity in the



(a) Problem IV.



(b) Problem V.



(c) Problem VI.

FIG. 11. *Sample solutions with* $N = 16$, $\epsilon = 1/200$, *and* $\delta = \delta_*^{el}$.

inflow boundary condition which is relatively insensitive to the value of δ, causing the middle of the plots to look fairly flat.

Problem V. Our two variable wind test problems are variants of the "IAHR/CEGB" test problem proposed in [14]. In this first case, we solve (1.1) on the unit square with

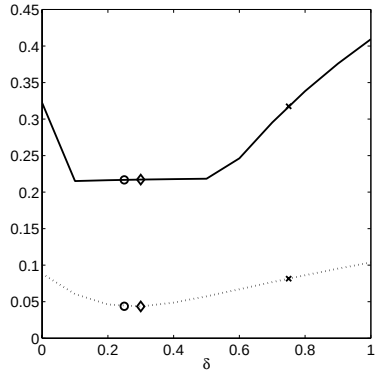
$$\mathbf{w} = (2y(1-(2x-1)^2), -2(2x-1)(1-y^2)). \tag{6.2}$$

The Dirichlet boundary conditions are given by

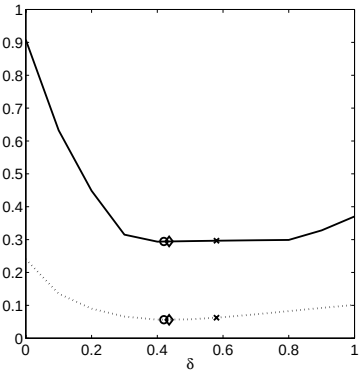
$$u(x,0) = 1 + \tanh[10 + 20(2x-1)] \tag{6.3}$$

on the inflow boundary (the interval $0 \le x \le 0.5$, $y = 0$) and $u(x,0) = 2$ on the outflow boundary (the interval $0.5 < x \le 1$, $y = 0$). On the remaining boundaries, we impose $f_t(x) = f_l(y) = f_r(y) = 0$. The Dirichlet boundary conditions at the bottom $y = 0$ are continuous but there is an exponential layer at the outflow portion, i.e., where $x \ge 1/2$. A sample solution for $N = 16$ and $\epsilon = 1/200$ is shown in Figure 11 (b).

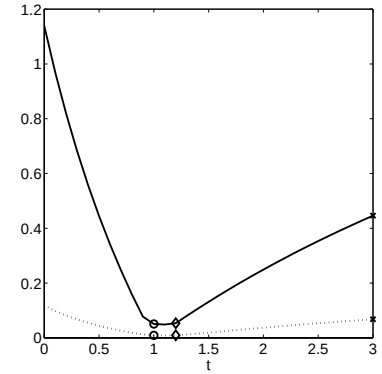
As the wind now varies in magnitude and direction from element to element, we cannot identify a single parameter δ which can be varied for the purposes of comparing errors as in the previous examples. However, we can compare various strategies for choosing δ^{el} locally within elements by considering the parameterized version of δ^{el}



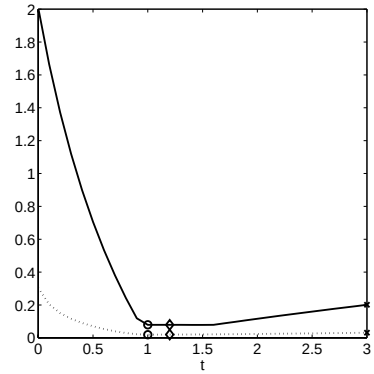
(a) Problem IV: $\epsilon = 1/64$.



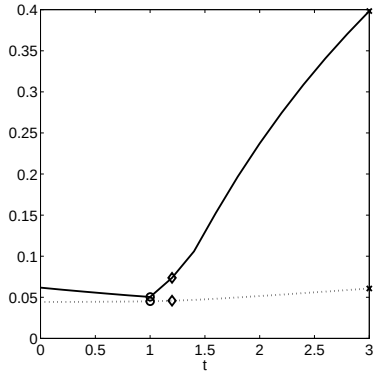
(d) Problem IV: $\epsilon = 1/200$.



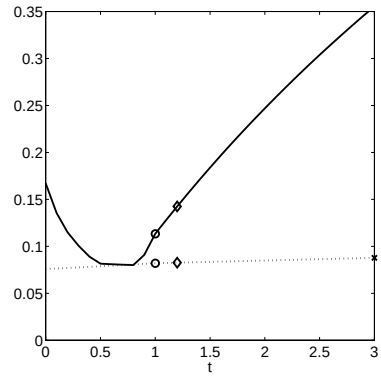
(b) Problem V: $\epsilon = 1/64$.



(e) Problem V: $\epsilon = 1/200$.



(c) Problem VI: $\epsilon = 1/64$.



(f) Problem VI: $\epsilon = 1/200$.

FIG. 12. *Error variation with δ in the discrete L_∞ norm (solid) and L_2 norm (dotted) for $N = 16$.*

given by

$$\delta^{el} = \begin{cases} \dfrac{t}{2}\left(1 - \dfrac{1}{P_e^{el}}\right), & 0 \le t \le 1, \\ \dfrac{1}{2}\left(1 + (t-2)\dfrac{1}{P_e^{el}}\right), & 1 < t \le 3. \end{cases} \tag{6.4}$$

As t varies from 0 to 3, the value of δ^{el} on each element first increases linearly from 0 to δ_*^{el} (at $t = 1$) and then varies linearly between δ_*^{el} and $\delta^{el,*}$. The variation with t of the error for this problem for $N = 16$ with two different values of ϵ is shown in Figure 12 (b) and (e). The errors are again calculated as described in section 5. The values δ_*^{el} ($\circ$), $\delta^{el,*}$ ($\times$), and $\delta_\bullet^{el}$ ($\diamondsuit$) are highlighted. The error is dominated by problems caused by the exponential layer along the outflow boundary in a similar way to Problem I.

Problem VI. Our final test problem also has a variable wind given by (6.2) but the boundary conditions are now of mixed type. We again impose the Dirichlet condition (6.3) on the inflow boundary but now the condition imposed on the outflow boundary is a homogeneous Neumann one. The Dirichlet boundary conditions on the remaining boundaries are $f_t(x) = f_l(y) = 0$, $f_r(y) = 2$. This results in the formation of a characteristic boundary layer along the right-hand wall. A sample solution for $N = 16$ and $\epsilon = 1/200$ is shown in Figure 11 (c). Error plots for this problem with δ parameterized by t as in (6.4) are shown in Figure 12 (c) and (f). This problem features a characteristic layer, so we expect the effects of changing δ to be less pronounced, as for Problem II. This is supported by the error plots: increasing δ helps to resolve the characteristic layer until the error becomes dominated by the effects of boundary discontinuities.

The results in these experiments are essentially the same as those for the model problems. We have not displayed oscillations here, but in all of the examples, the solutions for $\delta = \delta_*$ contain slight oscillations near layers, and the choice $\delta = \delta_\bullet$ reduces but does not eliminate them in these examples. There is little difference between these values in terms of solution quality obtained, and both choices are generally better than δ^*, which adds too much diffusion. Although it is tempting to use the interpolated value $\delta_\bullet$ to produce a qualitatively smoother solution, in our view δ_* is a better choice. The oscillations it produces indicate that in fact the layers are *not* fully resolved and that mesh refinement is needed where they occur; the smoothing of these effects will be misleading (see, e.g., [4]). Streamline diffusion alone cannot completely resolve this issue, and the choice δ_* adds the right amount of diffusion to keep the errors small in most of the domain. Note that this value has previously been recommended as a good choice in [1] and was shown to be good for efficient solution of the resulting linear system by the GMRES iterative method [3]. We also remark that although the analysis of sections 2–4 does not apply to linear elements, we have performed a few experiments which indicate that δ_* yields more accurate solutions than $\delta_\bullet$ in the linear case and that the latter choice adds excessive diffusion in this setting.

7. Application to other discretizations. To conclude, we emphasize that analysis of this type can be applied to any discretization whose stencil is of the form (2.3). We comment on two particular cases of interest here.

7.1. Finite differences with streamline diffusion. The usual central finite difference discretization of (1.1) can also be stabilized using streamline diffusion; see, for example, [12, p. 1465]. Specifically, we apply the finite difference method to the differential equation

$$-(\epsilon\nabla^2 + \nabla \cdot D\nabla)u(x,y) + \mathbf{w} \cdot \nabla u(x,y) = f(x,y),$$

where diffusion in the streamline direction is added using

$$D = \alpha \begin{bmatrix} c^2 & cs \\ cs & s^2 \end{bmatrix}$$

with

$$c = \frac{w_1}{\|\mathbf{w}\|_2}, \qquad s = \frac{w_2}{\|\mathbf{w}\|_2},$$

and α as in (1.4). Assuming for convenience that $\|\mathbf{w}\|_2 = 1$, the full computational molecule is given by

$$\begin{array}{ccc} \frac{w_1 w_2 \delta}{2h} & -\frac{\epsilon}{h^2} + \frac{w_2}{2h} - \frac{w_2^2 \delta}{h} & -\frac{w_1 w_2 \delta}{2h} \\ \nwarrow & \uparrow & \nearrow \\ -\frac{\epsilon}{h^2} - \frac{w_1}{2h} - \frac{w_1^2 \delta}{h} \leftarrow & \frac{4\epsilon}{h^2} + \frac{2\delta}{h} & \rightarrow -\frac{\epsilon}{h^2} + \frac{w_1}{2h} - \frac{w_1^2 \delta}{h} \\ \swarrow & \downarrow & \searrow \\ -\frac{w_1 w_2 \delta}{2h} & -\frac{\epsilon}{h^2} - \frac{w_2}{2h} - \frac{w_2^2 \delta}{h} & \frac{w_1 w_2 \delta}{2h} \end{array}.$$

This simplifies to a stencil of standard five-point type for our model problem (2.1) with grid-aligned flow. Using the notation of (2.3), the stencil entries are

$$m_1 = \frac{4\epsilon}{h^2} + \frac{2\delta}{h}, \qquad m_2 = -\frac{\epsilon}{h^2}, \qquad m_3 = -\frac{\epsilon}{h^2} + \frac{1}{2h} - \frac{\delta}{h},$$

$$m_4 = 0, \qquad m_5 = -\frac{\epsilon}{h^2} - \frac{1}{2h} - \frac{\delta}{h}, \qquad m_6 = 0$$

with related eigenvalues

$$\gamma_i = \frac{1}{h^2}\left[-(\epsilon + \delta h) - \frac{h}{2}\right], \qquad \lambda_i = \frac{1}{h^2}\left[2(\epsilon + \delta h) + 2\epsilon(1 - C_i)\right],$$

$$\sigma_i = \frac{1}{h^2}\left[-(\epsilon + \delta h) + \frac{h}{2}\right].$$

This results in the expressions

$$\mu_{1,2} = \frac{-2\delta - [2 - C_i]\frac{1}{P_e} \pm \sqrt{1 + 4\delta(1 - C_i)\frac{1}{P_e} + (1 - C_i)(3 - C_i)\frac{1}{P_e^2}}}{-2\delta + 1 - \frac{1}{P_e}}$$

for the roots of the recurrence relation which appear in (2.13).

Here the sign of μ_2 (and hence the nature of the corresponding functions $G_1(i,k)$ and $G_2(i,k)$, $i \in S_N$) is independent of i: as the numerator of μ_2 is always negative, we simply have the conditions

$$\begin{cases} \delta > \delta_* & \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory,} \\ \delta < \delta_* & \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory,} \end{cases}$$

where δ_* is given by (3.3). Hence the result equivalent to Theorem 3.1 is given by the following theorem.

THEOREM 7.1. *For a streamline diffusion finite difference discretization with $P_e > 1$, $\delta > \delta_*$ implies that $G_1(i,k)$ and $G_2(i,k)$ in* (2.13) *are nonoscillatory functions of k for any value of $i \in S_N$.*

The special case $\delta = \delta_*$ leads to the two-term recurrence with auxiliary equation root

$$\rho = \frac{1}{1 + \dfrac{(1 - C_i)}{P_e}}$$

and solution (3.5). Because $\rho < 1$, this solution is nonoscillatory in the streamline direction for all $i \in S_N$ and, as in the finite element case, tends to the nodally exact solution in the limit as $P_e \to \infty$.

The fact that there is one critical parameter (independent of i) here means that there is no issue about selecting a global parameter δ as we had in the finite element case. Furthermore, the analysis of the effect of transforming from $\mathbf{y}$ to $\mathbf{u}$ (cf. section 3.3) is greatly simplified. In particular, for the same specific example problem with $f_t = 1$ and $f_b = f_l = f_r = 0$ studied in section 3.3, the equivalent expression to (4.2) using finite differences has $S_{\text{smooth}} = 0$ when $\delta < \delta_*$ and $S_{\text{osc}} = 0$ when $\delta > \delta_*$. Thus we immediately have the following theorem (cf. Theorem 3.3).

THEOREM 7.2. *For a streamline diffusion finite difference discretization of* (2.1), *the discrete solution* $\mathbf{u}$ *does not exhibit oscillations in the streamline direction when $\delta \geq \delta_*$.*

That is, in contrast to the finite element case, there is no "smoothing" introduced by the Fourier transformation: the same single parameter governs the presence of oscillations in both the recurrence relation solution $\mathbf{y}$ and the discrete two-dimensional solution $\mathbf{u}$.

7.2. Artificial diffusion. So far we have focused on adding smoothing in the streamline direction only, which is just one of the many stabilization methods available. In this section we analyze the artificial diffusion method (see, for example, [7, pp. 218–219]) with a view to comparing its smoothing effect with that of streamline diffusion. The artificial diffusion method works by adding diffusion in an isotropic way which does not take account of flow direction, and it is well known that this can result in smearing of internal layers. We can use the analytical techniques presented in this paper to confirm that the streamline diffusion method avoids this problem.

We again consider a vertical wind model problem using bilinear finite elements on a uniform grid. The idea of the artificial diffusion method is to replace equation (2.1) with

$$-(\epsilon + \delta h)\nabla^2 u + \frac{\partial u}{\partial y} = 0 \qquad \text{in } \Omega = (0,1) \times (0,1), \tag{7.1}$$

with δ once again a stabilization parameter to be chosen. When $P_e < 1$, we set $\delta = 0$ as before. Galerkin discretization using bilinear finite elements results in a matrix of the form (2.4), which is therefore covered by our analysis. The stencil entries in this

case are given by

$$m_1 = \frac{8}{3}(\delta h + \epsilon), \qquad m_2 = -\frac{1}{3}(\delta h + \epsilon), \qquad m_3 = -\frac{1}{3}[(\delta - 1)h + \epsilon],$$

$$m_4 = -\frac{1}{12}[(4\delta - 1)h + 4\epsilon], \quad m_5 = -\frac{1}{3}[(\delta + 1)h + \epsilon], \quad m_6 = -\frac{1}{12}[(4\delta + 1)h + 4\epsilon],$$

so the roots (2.12) of the corresponding recurrence relation are given by

$$\mu_{1,2} = \frac{-\left(2\delta + \frac{1}{P_e}\right)\left[\frac{4 - C_i}{2 + C_i}\right] \pm \sqrt{1 + \frac{3(1 - C_i)(5 + C_i)}{(2 + C_i)^2}\left(2\delta + \frac{1}{P_e}\right)^2}}{1 - \left(2\delta + \frac{1}{P_e}\right)\left[\frac{1 + 2C_i}{2 + C_i}\right]}. \tag{7.2}$$

First we briefly consider the issue of oscillations in the streamline direction. Here, as in section 3.2, the sign of μ_2 (and hence the presence of oscillations in the recurrence relation solution) depends on the value of $i \in S_N$. Defining the new critical value

$$\tilde{\delta}_i^c = \frac{1}{2}\left(\left[\frac{2 + C_i}{1 + 2C_i}\right] - \frac{1}{P_e}\right),$$

we have different conditions for two sets of i values, namely

$$1 \le i \le \tfrac{2}{3}N: \quad \begin{cases} \delta > \tilde{\delta}_i^c & \Rightarrow \quad \mu_2 > 0,\ G_1(i,k) \text{ is nonoscillatory,} \\ \delta < \tilde{\delta}_i^c & \Rightarrow \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory,} \end{cases}$$

$$\tfrac{2}{3}N < i \le N - 1: \quad \mu_2 < 0,\ G_1(i,k) \text{ is oscillatory.}$$

Notice that this is different from the streamline diffusion case (cf. Theorem 3.1) in that there is no choice of δ which will make the recurrence relation solution oscillation free, as some of the contributing functions $G_1(i,k)$ are always oscillatory. However, it can be seen using an argument of the type presented in section 3.3 that the transformed solution is again dominated by contributions from functions pertaining to lower values of i. Hence, despite the fact that $G_1(i,k)$ is always oscillatory for large i, it is still possible to obtain a nonoscillatory discrete solution $\mathbf{u}$. Note that inequality (3.4) is satisfied with δ_c^i replaced by $\tilde{\delta}_c^i$. For the particular (i-independent) choice $\delta = \delta_*$ from (3.3), equation (7.1) (and hence the artificial diffusion solution) is independent of ϵ.

To gain insight into the main difference between this method and the streamline diffusion technique, we must examine solution behavior in the "crosswind" direction, that is, perpendicular to the direction of the flow. To fix ideas, we will use the discontinuous boundary conditions

$$f_b(x) = \begin{cases} 0, & x < 0.5, \\ 1, & x \ge 0.5, \end{cases} \qquad f_r(y) = 1, \quad f_t(x) = f_l(y) = 0$$

so that the solution has an internal layer along $x = 0.5$ as well as a boundary layer along the right half of the top boundary. The internal layer derives from propagation of the bottom boundary condition through the domain and, as $\epsilon \to 0$, the width of this layer tends to zero. Ideally, this phenomenon should be reproduced by a discretization method, that is, we would like to obtain a set of discrete solutions $\mathbf{u}$ in

this limit whose variation from the bottom boundary function is independent of j for fixed k. We now show that while the streamline diffusion method has this property, the artificial diffusion method does not.

Consider the recurrence relation solution vector $\mathbf{y}$ for this problem. From (2.13), its entries are given by

$$y_{ik} = F_3(i)\,(1 - G_1(i,k)) + [F_2(i) - F_3(i)]\,G_2(i,k) \tag{7.3}$$

with

$$F_2(i) = \sqrt{\frac{2}{N}}\left[\frac{(-1)^{i+1}\sin\dfrac{i\pi}{N}}{2\left(1-\cos\dfrac{i\pi}{N}\right)}\right]$$

[2, appendix] and $F_3(i)$ as in (3.6). As the functions $F_2(i)$ and $F_3(i)$ are the same for both discretizations, any difference in solution behavior must come from a difference in the behavior of the functions $G_1(i,k)$ and $G_2(i,k)$ associated with the two methods. We therefore now focus on how these functions vary with $i \in S_N$ as $\epsilon \to 0$ ($P_e \to \infty$) for $k \in S_N$ fixed. To simplify the presentation of this analysis, we will assume that δ is fixed independent of P_e, with $\delta \neq 0, 0.5$.

With the streamline diffusion discretization, neglecting terms of $O(P_e^{-1})$ and higher in (3.1) gives the approximations

$$\mu_1 \simeq 1, \qquad \mu_2 \simeq \frac{2\delta+1}{2\delta-1} \equiv \beta$$

so that

$$\begin{aligned} G_1(i,k) &= \frac{\mu_1^k - \mu_2^k}{\mu_1^N - \mu_2^N} \simeq \frac{1-\beta^k}{1-\beta^N} \equiv G_1^a(k), \\ G_2(i,k) &= (1-\mu_1^k) - (1-\mu_1^N)G_1(i,k) \simeq 0. \end{aligned}$$

Thus, in the limit as $P_e \to \infty$, both functions are independent of i. We then have

$$y_{ik} \simeq F_3(i)(1 - G_1^a(k));$$

hence, using (2.8),

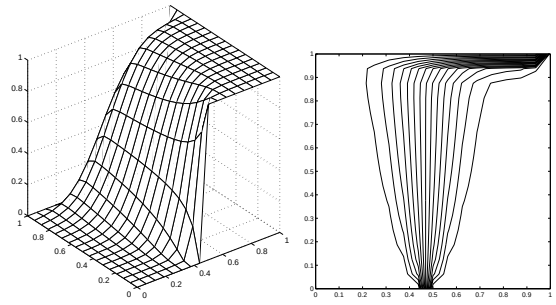
$$u_{jk} \simeq f_b(x_j)(1 - G_1^a(k)).$$

That is, the variation of u_{jk} from the bottom boundary function is independent of j in this limit. For the artificial diffusion discretization, however, neglecting terms of $O(P_e^{-1})$ and higher in (7.2) gives

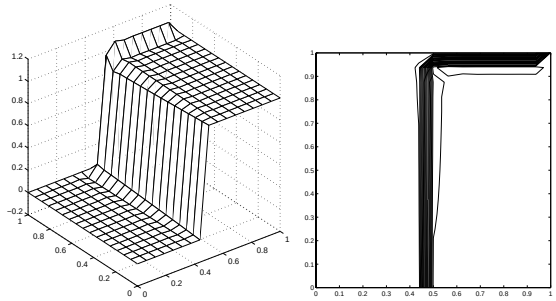
$$\mu_{1,2} \simeq \frac{-2\delta(4-C_i) \pm \sqrt{4(1+15\delta^2) + 4(1-12\delta^2)C_i + (1-12\delta^2)C_i^2}}{2(1-\delta) + (1-4\delta)C_i},$$

leading to approximations for $G_1(i,k)$ and $G_2(i,k)$ which depend on i through C_i. From (7.3) the solution is therefore

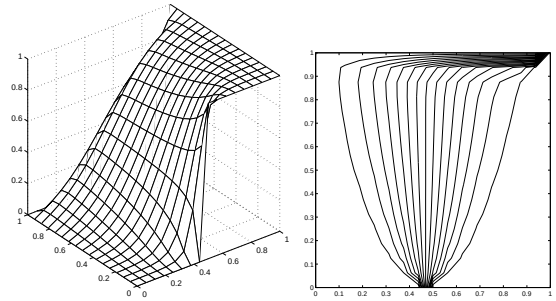
$$u_{jk} \simeq f_b(x_j) - \sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} \sin\frac{ij\pi}{N}\left(F_3(i)G_1(i,k) - [F_2(i) - F_3(i)]\,G_2(i,k)\right).$$



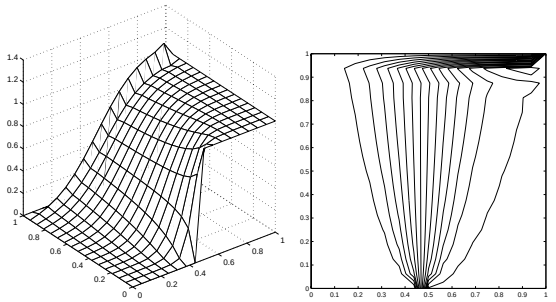
(a) Streamline diffusion: $P_e = 2$.



(b) Streamline diffusion: $P_e = 200$.



(c) Artificial diffusion: $P_e = 2$.



(d) Artificial diffusion: $P_e = 200$.

FIG. 13. *Solutions and contour plots for* $\delta = 0.4$ *and* $N = 16$.

This has a j-dependence which the continuous solution in this limit does not.

This fundamental difference between the discretizations is demonstrated pictorially in Figure 13, which shows streamline and artificial diffusion approximations (and associated contour plots) for this example problem with two values of ϵ, $\delta = 0.4$,

and $N = 16$. Plots (a) and (b) show that the streamline diffusion method captures the narrowing of the internal layer exhibited by the continuous solution as $\epsilon \to 0$ ($P_e \to \infty$). The equivalent artificial diffusion approximation does not, as shown in plots (c) and (d).

8. Summary. In this study, we have performed a Fourier analysis of model problems with grid-aligned flow that identifies the effects of upwinding in discretizations of the convection-diffusion equation. Our emphasis is on streamline-diffusion discretization with bilinear elements, where we show how the choice of streamline diffusion parameter affects the qualitative behavior of the solution with respect to oscillations. This analysis gives theoretical justification for the choice

$$\delta = \delta_* = \frac{1}{2}\left(1 - \frac{1}{P_e^{el}}\right).$$

Our analysis also shows that δ_* is the optimal choice for finite difference discretizations, provides insight into the method of isotropic artificial diffusion, and yields qualitatively good solutions in a variety of computational experiments.

REFERENCES

[1] A. Brooks and T. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[2] H.C. Elman and A. Ramage, *A characterisation of oscillations in the discrete two-dimensional convection-diffusion equation,* Math. Comp., to appear.

[3] B. Fischer, A. Ramage, D.J. Silvester, and A. J. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems,* Comput. Methods Appl. Mech. Engrg., 179 (1999), pp. 179–195.

[4] P.M. Gresho and R.L. Lee, *Don't suppress the wiggles—they're telling you something,* in Finite Element Methods for Convection Dominated Flows, AMD 34, T.J.R. Hughes, ed., ASME, New York, 1979, pp. 37–61.

[5] P.M. Gresho and R.L. Sani, *Incompressible Flow and the Finite Element Method*, John Wiley and Sons, Chichester, UK, 1999.

[6] M.D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Comput. Sci. Sci. Comput., Academic Press, New York, 1989.

[7] W. Hackbusch, *Multi-grid Methods and Applications*, Springer-Verlag, New York, 1980.

[8] T.J.R. Hughes and A. Brooks, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows, AMD 34, T.J.R. Hughes, ed., ASME, New York, 1979, pp. 120–131.

[9] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[10] K.W. Morton, *Numerical Solution of Convection-Diffusion Problems,* Chapman and Hall, London, 1996.

[11] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations.* Springer-Verlag, New York, 1994.

[12] H.-G. Roos, *Necessary convergence conditions for upwind schemes in the two-dimensional case,* Internat. J. Numer. Methods Engrg., 21 (1985), pp. 1459–1469.

[13] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[14] R.M. Smith and A.G. Hutton, *The numerical treatment of advection—a performance comparison of current methods,* Numer. Heat Trans., 5 (1982), pp. 439–461.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 282–306

ERROR ESTIMATES FOR SEMIDISCRETE FINITE ELEMENT APPROXIMATIONS OF LINEAR AND SEMILINEAR PARABOLIC EQUATIONS UNDER MINIMAL REGULARITY ASSUMPTIONS*

K. CHRYSAFINOS† AND L. S. HOU†

Abstract. Semidiscrete finite element error estimates for linear parabolic equations are derived under minimal regularity with the help of L^2 projectors. Then, analogous minimal regularity semidiscrete error estimates for semilinear parabolic equations are derived.

Key words. parabolic equations, finite element methods, semidiscrete error estimates, boundary value problems

AMS subject classifications. 65M60, 65M12

PII. S0036142900377991

1. Introduction.

Semidiscrete finite element approximations of linear parabolic equations have been studied extensively. [38] offers an excellent review of the main results and mathematical techniques of this subject and contains a comprehensive list of references. Here, we provide a brief summary along with some representative (but certainly not exhaustive) citations for each type of semidiscrete error estimate available in the literature.

- $L^2(\Omega)$ error estimates—either $L^\infty(0,T;L^2(\Omega))$ or $L^2(0,T;L^2(\Omega))$—[7], [18], [43], [22, Theorems 4.1–4.3], and [38, Theorems 1.2, 2.3, 2.5, 2.6];
- $H^1(\Omega)$ error estimates—either $L^\infty(0,T;H^1(\Omega))$ or $L^2(0,T;H^1(\Omega))$—[7], [18], [22, Theorems 4.1–4.3], and [38, Theorems 1.4, 2.4];
- $L^\infty(\Omega)$ error estimates—[33], [29], [30], [34], [40], [12], [16], [17], [10], [42], [41], [7], and [38, Theorems 5.2–5.5];
- Negative spatial norm error estimates for $u-u^h$ and $\partial_t u-\partial_t u^h$ and superconvergence results—[7], [37], [19], and [38, Theorems 6.2–6.6];
- Semidiscrete error estimates with nonsmooth initial data—[36], [37], [31], [27], [15], [25], and [38, Theorems 3.1–3.6].

There is also an abundant literature on fully discrete approximations; see, e.g., [4], [3], [18], [6], [8], [20], [32], [44], and [38]. In this paper, we will concern ourselves only with semidiscrete approximations.

Notwithstanding the vast number of papers and books devoted to semidiscrete approximations of solutions of parabolic equations, the literature seems to lack $L^2(\Omega)$ and $H^1(\Omega)$ convergence results for parabolic problems under minimal regularity assumptions. The goal of this work is to establish the convergence and derive error estimates in the norm of the solution space under minimal regularity assumptions on the solution. Such results will be established for both linear and semilinear parabolic problems.

The linear problem we consider is the parabolic initial boundary value problem

$$\partial_t u - \operatorname{div}\left[A(\mathbf{x})\nabla u\right] = f \qquad \text{in } [0,T]\times\Omega\,, \tag{1.1}$$

*Received by the editors September 12, 2000; accepted for publication (in revised form) November 6, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/37799.html
†Department of Mathematics, Iowa State University, Ames, IA 50011 (kchrysaf@iastate.edu, hou@math.iastate.edu).

$$u\big|_{(0,T)\times\partial\Omega} = 0\,, \tag{1.2}$$

and

$$u(0) = u_0 \qquad \text{in } \Omega\,, \tag{1.3}$$

where $\partial_t u = \partial u/\partial t$, Ω is a bounded spatial domain in $\mathbb{R}^d$ ($d = 2$ or 3), and the matrix function $A \in \mathbf{L}^\infty(\mathbb{R}^d; \mathbb{R}^{d\times d})$ is uniformly positive definite. In the standard weak formulation of (1.1)–(1.3) (see, e.g., [21]) the solution u is sought in the space

$$X \equiv L^2(0,T; H_0^1(\Omega)) \cap H^1(0,T; H^{-1}(\Omega)).$$

Of course, the minimal regularity of u for showing the X-norm convergence of semidiscrete finite element approximations is that $u \in X$. The assumption $u \in X$ is equivalent to the requirements that the forcing term $f \in L^2(0,T; H^{-1}(\Omega))$ and the initial condition $u_0 \in L^2(\Omega)$; see, e.g., [21, p. 356, Theorem 3]. The X-norm convergence results in the literature generally assume higher regularity than the minimal regularity $u \in X$. For instance, the estimates in [38] typically require $\partial_t u \in L^2(0,T; L^2(\Omega))$ and those in [43] require $\nabla u \in L^\infty(0,T; L^\infty(\Omega))$.

To derive $\mathcal{O}(h^m)$ ($m \geq 0$) error estimates in the X-norm, the minimal regularity assumption on u should be $u \in L^2(0,T; H^{m+1}(\Omega)) \cap H^1(0,T; H^{m-1}(\Omega))$. We indeed will prove $\mathcal{O}(h^m)$ error estimates under such minimal regularity assumptions. $\mathcal{O}(h^m)$ error estimates in the literature generally require higher regularity than what this paper assumes. We note that the assumption $u \in L^2(0,T; H^{m+1}(\Omega)) \cap H^1(0,T; H^{m-1}(\Omega))$ is essentially equivalent to the assumptions $u_0 \in H^m(\Omega)$ and $f \in L^2(0,T; H^{m-1}(\Omega))$ plus certain compatibility conditions; see, e.g., [28, p. 287, Theorem 2.10] (for all m) and [21, p. 365] (for odd m).

In addition to showing the convergence and deriving the error estimates under the minimal regularity assumptions explained above, we will also establish a parabolic Cea's lemma and a parabolic Aubin–Nitsche's lemma. The collection of these results comprises a semidiscrete finite element theory for parabolic problems that parallels classical finite element theories for elliptic problems.

A further linear result of this paper is the derivation of some pointwise-in-time error estimates under slightly weaker regularity requirements compared to similar results in the literature (e.g., [38, Theorem 2.4]).

It should be pointed out that a key element of our proof is the use of the $L^2(\Omega)$- and generalized $L^2(\Omega)$-projectors instead of the usual elliptic projector. Though some properties of the L^2-projections are well known (see, e.g., [38], [5], and [14] for interior L^2-projections and [24] for boundary L^2-projections), we include in this paper some detailed discussions of the approximation properties of $L^2(\Omega)$- and generalized $L^2(\Omega)$-projectors.

It also should be noted that the nonminimal regularity assumptions in the literature for showing convergence and error estimates are partly tied to the nonlinearities appearing in the equations. For instance, if the equation contains a term $\operatorname{div}[A(u)\nabla u]$, then for $u(t) \in H^1(\Omega)$ and $v \in H^1(\Omega)$ it is not guaranteed that $\int_\Omega A(u)(\nabla u)\cdot(\nabla v) < \infty$. In this case, an assumption such as $u(t) \in L^\infty(\Omega)$ is needed. For this reason the results of this paper do not apply to the approximations of nonlinear equations involving $\operatorname{div}[A(u)\nabla u]$. It can be checked that many nonminimal regularity assumptions in the literature could be weakened when the underlying equations are linear, e.g., the results of [43] and [22]. However, error estimates for

nonlinear equations under nonminimal regularity do not trivially lead to the minimal regularity error estimates of this paper when the nonlinearity is dropped. For instance, one may find in [38] a number of linear results under nonminimal regularity assumptions.

The semilinear problem we consider consists of the partial differential equation

$$\partial_t u - \operatorname{div}[A(\mathbf{x})\nabla u] + \mathbf{b}(t,\mathbf{x})\cdot\nabla u + \phi(u) = f \qquad \text{in } [0,T]\times\Omega\,, \tag{1.4}$$

with boundary-initial conditions (1.2)–(1.3), where $\mathbf{b} \in L^\infty(0,T;L^\gamma(\Omega))$ ($\gamma > d$, $d = 2$ or 3 is the space dimension) and ϕ satisfies certain power growth conditions. We extend the linear results to the semilinear case. Namely, we prove the X-norm convergence of semidiscrete finite element solutions if $u \in X$ and we derive X-norm error estimates under minimal regularity assumptions on u. Compared to the results of [38] and [39] for semilinear parabolic equations, our results not only assume merely minimal regularity but also allow for a slightly larger exponent in the growth conditions for $\phi(u)$.

In our study of the semilinear problems the linear theory of this paper and the approximation theory of Brezzi–Rappaz–Raviart both play a crucial role. Other worthmentioning technicalities for our semilinear theory involve (i) the proof of an embedding theorem for X and (ii) the choice of a merely continuously embedded subspace (as opposed to the choice of a compactly embedded subspace that is typical of the applications of Brezzi–Rappaz–Raviart theory).

The plan of the paper is as follows. In section 2, we introduce some notations and discuss the approximation properties of $L^2(\Omega)$- and generalized $L^2(\Omega)$-projectors. In section 3, we derive semidiscrete error estimates for finite element approximations of linear parabolic equations under minimal regularity assumptions on the solutions. Finally, in section 4, we derive error estimates for semidiscrete approximations of semilinear parabolic equations, again under minimal regularity assumptions on the solutions.

2. Notations and some properties of L^2-projections onto finite element spaces. We assume that Ω is a convex polygon in $\mathbb{R}^2$ or a convex polyhedron in $\mathbb{R}^3$. Throughout, we use standard notations (see, e.g., [1]) for the Sobolev spaces $H^r(\Omega)$ and $H^s(\Gamma)$ for all real r and s, with norms denoted by $\|\cdot\|_r$ and $\|\cdot\|_{s,\Gamma}$, respectively. We use $H_0^r(\Omega)$ to denote the completion of $C_0^\infty(\Omega)$ in the $H^r(\Omega)$-norm. We denote the inner products on $H^r(\Omega)$ and $H^s(\Gamma)$ by $(\cdot,\cdot)_r$ and $(\cdot,\cdot)_{s,\Gamma}$, respectively. Note that $H^0(\Omega) = L^2(\Omega)$ and we omit the subscript 0 for the $L^2(\Omega)$-norm and $L^2(\Omega)$ inner product, i.e., $(\cdot,\cdot)_0 = (\cdot,\cdot)$ and $\|\cdot\|_0 = \|\cdot\|$. The duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$ is denoted by $\langle\cdot,\cdot\rangle$, which we assume is generated from the $L^2(\Omega)$ inner product, i.e.,

$$\langle v, w\rangle = (v, w) \qquad \forall\, v, w \in L^2(\Omega).$$

Also, C denotes a generic constant whose value changes with context.

Let V^h be a family of finite element subspaces of $H_0^1(\Omega)$ with the following approximation properties:

$$\inf_{v^h\in V^h} \|v - v^h\|_s \to 0 \quad \text{as } h \to 0 \quad \forall\, v \in H^s(\Omega)\,,\ s = -1, 0, 1. \tag{2.1}$$

In order to properly state some additional approximation properties for functions in V^h having zero boundary values, we introduce, for all real r, the spaces $\Phi_0^r(\Omega) =$

$H_0^{\min\{1,r\}}(\Omega)$ equipped with the $H^{\min\{1,r\}}(\Omega)$-norm, i.e.,

$$\Phi_0^r(\Omega) = \begin{cases} H_0^1(\Omega) & \text{if } r \geq 1\,, \\ H_0^r(\Omega) & \text{if } 1/2 < r < 1\,, \\ H^r(\Omega) & \text{if } r \leq 1/2\,. \end{cases}$$

Note that we have used the result that $H^r(\Omega) = H_0^r(\Omega)$ for $r \leq 1/2$; see [26, p. 55]. Also, when $r > 1$, r is no longer the differentiability index in $\Phi_0^r(\Omega)$. The space $\Phi_0^r(\Omega)$ simply unifies the notation used below for stating relevant approximation properties.

We assume that V^h has the following additional approximation properties:

$$\inf_{v^h \in V^h} \|v - v^h\|_{-1} \leq Ch^{m+2}\|v\|_{m+1} \quad \forall\, v \in H^{m+1}(\Omega) \cap \Phi_0^{m+1}(\Omega),\ -2 \leq m \leq k\,, \tag{2.2}$$

$$\inf_{v^h \in V^h} \|v - v^h\| \leq Ch^{m+1}\|v\|_{m+1} \quad \forall\, v \in H^{m+1}(\Omega) \cap \Phi_0^{m+1}(\Omega), \quad -1 \leq m \leq k\,, \tag{2.3}$$

and

$$\inf_{v^h \in V^h} \|v - v^h\|_1 \leq Ch^{m}\|v\|_{m+1} \quad \forall\, v \in H^{m+1}(\Omega) \cap H_0^1(\Omega), \quad 0 \leq m \leq k\,. \tag{2.4}$$

In (2.2)–(2.4), C is independent of v and h and k is a positive integer that is usually determined by the order of the piecewise polynomials used to define V^h.

We also assume the following inverse inequalities:

$$\|v^h\|_1 \leq \frac{C}{h}\|v^h\|_0 \qquad \forall\, v^h \in V^h. \tag{2.5}$$

For a thorough discussion of the properties (2.1)–(2.5) and the construction of finite element spaces having these properties, see, e.g., [2] and [11].

We denote by P^h the $L^2(\Omega)$-projection from $L^2(\Omega)$ onto V^h, R^h the $H^1(\Omega)$-projection from $H_0^1(\Omega)$ onto V^h, and S^h the $H^{-1}(\Omega)$-projection from $H^{-1}(\Omega)$ onto V^h. Namely, for each $v \in L^2(\Omega)$,

$$(P^h v, w^h) = (v, w^h) \qquad \forall\, w^h \in V^h\,; \tag{2.6}$$

for each $v \in H_0^1(\Omega)$,

$$(R^h v, w^h)_1 = (v, w^h)_1 \qquad \forall\, w^h \in V^h\,; \tag{2.7}$$

and for each $v \in H^{-1}(\Omega)$,

$$(S^h v, w^h)_{-1} = (v, w^h)_{-1} \qquad \forall\, w^h \in V^h\,. \tag{2.8}$$

We also define the *generalized $L^2(\Omega)$-projection* operator $Q^h : H^{-1}(\Omega) \to V^h$ as follows: for each $v \in H^{-1}(\Omega)$,

$$\langle Q^h v, w^h \rangle = \langle v, w^h \rangle \qquad \forall\, w^h \in V^h\,. \tag{2.9}$$

Remark. One can easily show that Q^h is well defined by introducing a basis $\{e_i\}_{i=1}^M$ for V^h and examining the solvability of a $v^h = \sum_{j=1}^M c_j e_j$ from the equations $\langle \sum c_j e_j, e_i \rangle = \langle v, e_i \rangle$ for $i = 1, \ldots, M$. Also, it is evident that $Q^h v = P^h v$ whenever $v \in L^2(\Omega)$; thus, $Q^h : H^{-1}(\Omega) \to V^h$ can be thought of as the generalization of the operator $P^h : L^2(\Omega) \to V^h$ into an operator from $H^{-1}(\Omega)$ to V^h. ∎

It is trivial that $P^h v$ is the best approximation in the $L^2(\Omega)$-norm to $v \in L^2(\Omega)$. We will show that $P^h v$ is *a quasi-best* approximation to v in the $H^1(\Omega)$-norm and $Q^h v$ is *a quasi-best* approximation to v in the $H^{-1}(\Omega)$-norm in the sense of the following two propositions.

PROPOSITION 2.1. *Let $P^h : L^2(\Omega) \to V^h$ be defined by* (2.6) *and $R^h : H_0^1(\Omega) \to V^h$ be defined by* (2.7)*. Then there exists a constant C, independent of v and h, such that*

$$\|v - P^h v\|_1 \le C\|v - R^h v\|_1 \qquad \forall\, v \in H_0^1(\Omega)\,. \tag{2.10}$$

Proof. Let $v \in H_0^1(\Omega)$ be given. The inverse inequality (2.5) yields

$$\|R^h v - P^h v\|_1 \le \frac{C}{h}\|R^h v - P^h v\|\,.$$

The best approximation property of the projection operators gives

$$\|v - P^h v\| \le \|v - R^h v\|\,.$$

As Ω is assumed to be convex, the Aubin–Nitsche lemma (see, e.g., [11, p. 137]) implies

$$\|R^h v - v\| \le Ch\|R^h v - v\|_1\,.$$

Using triangle inequalities and the last three relations we obtain

$$\begin{aligned}\|v - P^h v\|_1 &\le \|v - R^h v\|_1 + \|R^h v - P^h v\|_1 \le \|v - R^h v\|_1 + \frac{C}{h}\|R^h v - P^h v\| \\ &\le \|v - R^h v\|_1 + \frac{C}{h}\Big(\|R^h v - v\| + \|v - P^h v\|\Big) \le \|v - R^h v\|_1 + \frac{2C}{h}\|R^h v - v\| \\ &\le C\|v - R^h v\|_1\,. \qquad \square\end{aligned}$$

Remark. The key elements of the proof of Proposition 2.1 can be found in, among others, [38]. We include this proposition and its proof here for completeness. $\square$

PROPOSITION 2.2. *Let $S^h : H^{-1}(\Omega) \to V^h$ and $Q^h : H^{-1}(\Omega) \to V^h$ be defined by* (2.8) *and* (2.9)*, respectively. Then, there exists a constant C, independent of v and h, such that*

$$\|v - Q^h v\|_{-1} \le C\|v - S^h v\|_{-1} \qquad \forall\, v \in H^{-1}(\Omega)\,. \tag{2.11}$$

Proof. Let $v \in H^{-1}(\Omega)$ be given. Then

$$\begin{aligned}\|v - Q^h v\|_{-1} &= \sup_{w \in H_0^1(\Omega), \|w\|_1 = 1} \langle v - Q^h v, w\rangle \\ &= \sup_{w \in H_0^1(\Omega), \|w\|_1 = 1} \langle v - Q^h v, w - P^h w\rangle = \sup_{w \in H_0^1(\Omega), \|w\|_1 = 1} \langle v, w - P^h w\rangle \\ &= \sup_{w \in H_0^1(\Omega), \|w\|_1 = 1} \langle v - S^h v, w - P^h w\rangle \le \|v - S^h v\|_{-1} \sup_{w \in H_0^1(\Omega), \|w\|_1 = 1} \|w - P^h w\|_1\,.\end{aligned}$$

From Proposition 2.1, we see that

$$\|w - P^h w\|_1 \le C\|w - R^h w\|_1 \le C\|w\|_1\,.$$

Thus, (2.11) is obtained trivially by combining the last two relations. $\square$

Remark. As a consequence of Propositions 2.1 and 2.2, approximation properties (2.1) with $s = -1$ and (2.2) can be proved from the inverse inequality (2.5) and approximation properties (2.1) with $s = 0, 1$ and (2.3)–(2.4). Indeed, if $v \in H^{k+1}(\Omega) \cap \Phi_0^{k+1}(\Omega)$,

$$\|v - S^h v\|_{-1} \le \|v - P^h v\|_{-1} = \sup_{\phi \in H_0^1(\Omega)} \frac{\langle v - P^h v, \phi \rangle}{\|\phi\|_1}$$

$$= \sup_{\phi \in H_0^1(\Omega)} \frac{\langle v - P^h v, \phi - P^h \phi \rangle}{\|\phi\|_1}$$

$$\le \|v - P^h v\| \cdot \sup_{\phi \in H_0^1(\Omega)} \frac{\|\phi - P^h \phi\|}{\|\phi\|_1} \le C h^{k+1} \|v\|_{k+1} \cdot h.$$

For $v \in H^{-1}(\Omega) \cap \Phi_0^{-1}(\Omega) = H^{-1}(\Omega)$,

$$\|v - S^h v\|_{-1} \le 2\|v\|_{-1}.$$

Thus (2.1) with $s = -1$ and (2.2) follows from these estimates. □

Next we examine approximation properties on time-space function spaces. If B is a spatial function space such as $H^r(\Omega)$ for a real r, then we denote by $L^2(0, T; B)$ and $H^1(0, T; B)$ the time-space function spaces such that

$$\|v\|^2_{L^2(0,T;B)} \equiv \int_0^T \|v(t)\|_B^2 \, dt < \infty \qquad \forall\, v \in L^2(0, T; B)$$

and

$$\|v\|^2_{H^1(0,T;B)} \equiv \int_0^T \left(\|v(t)\|_B^2 + \|\partial_t v(t)\|_B^2 \right) dt < \infty \qquad \forall\, v \in H^1(0, T; B),$$

respectively. If B^h is a finite dimensional subspace of B, the norm of $L^2(0, T; B^h)$ and $H^1(0, T; B^h)$ is taken to be that of $L^2(0, T; B)$ and $H^1(0, T; B)$, respectively.

Based on the approximation properties of V^h, we may derive the following approximation properties of the semidiscrete function spaces $L^2(0, T; V^h)$ and $H^1(0, T; V^h)$.

PROPOSITION 2.3. *Let V^h be a family of finite element subspaces of $H_0^1(\Omega)$ satisfying* (2.1)–(2.5). *Then, the following convergence properties hold as $h \to 0$:*

$$\inf_{v^h \in L^2(0,T;V^h)} \|v - v^h\|_{L^2(0,T;H^s(\Omega))} \to 0 \quad \forall\, v \in L^2(0, T; H^s(\Omega)),\ s = -1, 0, 1. \tag{2.12}$$

Moreover, the following approximation properties hold:

$$\inf_{v^h \in L^2(0,T;V^h)} \|v - v^h\|_{L^2(0,T;H^s(\Omega))} \le C h^{m+1-s} \|v\|_{L^2(0,T;H^{m+1}(\Omega))}$$
$$\forall\, v \in L^2(0, T; H^{m+1}(\Omega) \cap \Phi_0^{m+1}(\Omega)),\ -2 \le m \le k\,,\ s = -1, 0, 1. \tag{2.13}$$

Proof. We will prove (2.12) and (2.13) for the case $s = -1$ only; the cases when $s = 0$ or 1 can be proved in a similar manner.

To prove (2.13) when $s = -1$, we need only to verify it for the cases $m = k$ and $m = -2$ thanks to interpolation theorems. Let $v \in L^2(0, T; H^{k+1}(\Omega) \cap \Phi_0^{k+1}(\Omega))$ be given and let $S^h v(t)$ be defined by (2.8) for almost every $t \in (0, T)$, i.e.,

$$(S^h v(t), w^h)_{-1} = (v(t), w^h)_{-1} \qquad \forall\, w^h \in V^h,\ \text{a.e. } t\,.$$

It follows from $v \in L^2(0,T;H^{k+1}(\Omega)\cap\Phi_0^{k+1}(\Omega))$ that $S^h v \in L^2(0,T;V^h)$. Using (2.2), we see that for almost every $t\in(0,T)$,

$$\|v(t)-S^h v(t)\|_{-1} \le Ch^{k+2}\|v(t)\|_{k+1} \qquad \text{a.e. } t$$

so that integration in t yields

$$\|v-S^h v\|_{L^2(0,T;H^{-1}(\Omega))} \le Ch^{k+2}\|v(t)\|_{L^2(0,T;H^{k+1}(\Omega))}\,. \tag{2.14}$$

Let $v\in L^2(0,T;H^{-1}(\Omega)) = L^2(0,T;H^{-1}(\Omega)\cap\Phi_0^{-1}(\Omega))$ be given and let $S^h v(t)$ be defined by (2.8) again for almost every t. The best approximation property of projection operators implies

$$\|v(t)-S^h v(t)\|_{-1} \le C\|v(t)\|_{-1} \qquad \text{a.e. } t$$

so that integration in t leads to

$$\|v-S^h v\|_{L^2(0,T;H^{-1}(\Omega))} \le C\|v(t)\|_{L^2(0,T;H^{-1}(\Omega))}\,. \tag{2.15}$$

Interpolation of (2.14) and (2.15) yields (2.13).

We now proceed to prove (2.12) when $s=-1$. Let $v\in L^2(0,T;H^{-1}(\Omega))$ be given. For each $\epsilon>0$, we may choose a $v_\epsilon\in C([0,T];H^{-1}(\Omega))$ such that

$$\|v-v_\epsilon\|_{L^2(0,T;H^{-1}(\Omega))} < \epsilon/3\,.$$

Then, using triangle inequalities and the fact that $\|S^h w\|_{-1}\le\|w\|_{-1}$, we have that

$$\begin{aligned}
&\|v-S^h v\|_{L^2(0,T;H^{-1}(\Omega))}\\
&\le \|v-v_\epsilon\|_{L^2(0,T;H^{-1}(\Omega))} + \|v_\epsilon - S^h v_\epsilon\|_{L^2(0,T;H^{-1}(\Omega))} + \|S^h v_\epsilon - S^h v\|_{L^2(0,T;H^{-1}(\Omega))}\\
&\le 2\|v-v_\epsilon\|_{L^2(0,T;H^{-1}(\Omega))} + \sqrt{T}\max_{t\in[0,T]}\|v_\epsilon(t)-S^h v_\epsilon(t)\|_{-1}\\
&< 2\epsilon/3 + \sqrt{T}\,\|v_\epsilon(t_0)-S^h v_\epsilon(t_0)\|_{H^{-1}(\Omega)}
\end{aligned}$$

for some $t_0\in[0,T]$. The approximation property (2.1) implies that there exists an $h_0>0$ such that, for all $h\in(0,h_0)$,

$$\|v_\epsilon(t_0)-S^h v_\epsilon(t_0)\|_{-1} \le \epsilon/(3\sqrt{T})\,.$$

Combining the last two relations, we have that, for all $h\in(0,h_0)$,

$$\|v-S^h v\|_{L^2(0,T;H^{-1}(\Omega))} < \epsilon\,.$$

This completes the proof of (2.12) when $s=-1$. ∎

Similar to Propositions 2.1 and 2.2, we can establish the *best approximation* properties of the $L^2(\Omega)$-projector P^h in $L^2(0,T;H_0^1(\Omega))$ and of the generalized $L^2(\Omega)$-projector Q^h in $L^2(0,T;H^{-1}(\Omega))$.

PROPOSITION 2.4. *Let V^h be a family of finite element subspaces of $H_0^1(\Omega)$ satisfying* (2.1)–(2.5). *Then, the following approximation properties hold:*

$$\|v-P^h v\|_{L^2(0,T;H^1(\Omega))}\to 0 \quad \text{as } h\to 0 \quad \forall\, v\in L^2(0,T;H_0^1(\Omega))\,, \tag{2.16}$$

$$\|v-Q^h v\|_{L^2(0,T;H^{-1}(\Omega))}\to 0 \quad \text{as } h\to 0 \quad \forall\, v\in L^2(0,T;H^{-1}(\Omega))\,, \tag{2.17}$$

$$\|v - P^h v\|_{L^2(0,T;H^1(\Omega))} \le Ch^m \|v\|_{L^2(0,T;H^{m+1}(\Omega))} \quad \forall\, v \in L^2(0,T;H^{m+1}(\Omega)\cap H_0^1(\Omega)),\ 0 \le m \le k\,, \tag{2.18}$$

and

$$\|v - Q^h v\|_{L^2(0,T;H^{-1}(\Omega))} \le Ch^{m+2} \|v\|_{L^2(0,T;H^{m+1}(\Omega))} \quad \forall\, v \in L^2(0,T;H^{m+1}(\Omega)\cap \Phi_0^{m+1}(\Omega)),\ -1 \le m \le k\,. \tag{2.19}$$

Moreover, the following inequalities hold:

$$\|v - P^h v\|_{L^2(0,T;H^1(\Omega))} \le C\|v - v^h\|_{L^2(0,T;H^1(\Omega))} \quad \forall\, v \in L^2(0,T;H_0^1(\Omega))\ \forall\, v^h \in L^2(0,T;V^h) \tag{2.20}$$

and

$$\|v - Q^h v\|_{L^2(0,T;H^{-1}(\Omega))} \le C\|v - v^h\|_{L^2(0,T;H^{-1}(\Omega))} \quad \forall\, v \in L^2(0,T;H^{-1}(\Omega))\ \forall\, v^h \in L^2(0,T;V^h)\,. \tag{2.21}$$

Proof. Let $v \in L^2(0,T;H_0^1(\Omega))$ be given. From (2.10) we see that

$$\|v(t) - P^h v(t)\|_1 \le C\|v(t) - R^h v(t)\|_1 \qquad \text{a.e. } t\,.$$

Squaring both sides and integrating in t we obtain

$$\|v - P^h v\|_{L^2(0,T;H^1(\Omega))} \le C\|v - R^h v\|_{L^2(0,T;H^1(\Omega))}\,. \tag{2.22}$$

Also, it is evident from the best approximation properties of R^h in the $\|\cdot\|_1$-norm that

$$\|v(t) - R^h v(t)\|_{L^2(0,T;H^1(\Omega))} = \inf_{v^h \in V^h} \|v - v^h\|_{L^2(0,T;H^1(\Omega))}\,.$$

Thus, (2.16) and (2.18) follow from (2.12) (with $s = 1$), (2.13) (with $s = 1$), and the last two relations.

Let $v \in L^2(0,T;H^{-1}(\Omega))$ be given. From (2.11) we see that

$$\|v(t) - Q^h v(t)\|_{-1} \le C\|v(t) - S^h v(t)\|_{-1} \qquad \text{a.e. } t\,.$$

Squaring both sides and integrating in t we obtain

$$\|v - Q^h v\|_{L^2(0,T;H^{-1}(\Omega))} \le C\|v - S^h v\|_{L^2(0,T;H^{-1}(\Omega))}\,. \tag{2.23}$$

Also, it is evident from the best approximation properties of S^h in the $\|\cdot\|_{-1}$-norm that

$$\|v - S^h v\|_{L^2(0,T;H^{-1}(\Omega))} = \inf_{v^h \in V^h} \|v - v^h\|_{L^2(0,T;H^{-1}(\Omega))}\,.$$

Thus, (2.17) and (2.19) follow from (2.12), (2.13), and the last two relations.

Inequalities (2.20) and (2.21) are trivial consequences of (2.22) and (2.23), respectively. ∎

Remark. Note that using (2.6) and (2.8), we can prove that P^h is the L^2-projection from $L^2(0,T;L^2(\Omega))$ onto $L^2(0,T;V^h)$. ∎

3. Semidiscrete error estimates for linear parabolic equations. We assume in what follows that the data

$$f \in L^2(0,T;H^{-1}(\Omega)) \quad \text{and} \quad u_0 \in L^2(\Omega)$$

or, equivalently, the solution $u \in X$, where X is as introduced in section 1, is the space

$$X = L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$$

equipped with the norm

$$\|v\|_X^2 = \|v\|_{L^2(0,T;H^1(\Omega))}^2 + \|\partial_t v\|_{L^2(0,T;H^{-1}(\Omega))}^2 .$$

Whenever further regularity assumptions on the solution u (or equivalently on the data f and u_0) are used, they will be explicitly stated. In this section we consider the approximations of the linear parabolic problem

$$\partial_t u - \operatorname{div}[A(\mathbf{x})\nabla u] = f \qquad \text{in } (0,T)\times\Omega\,, \tag{3.1}$$

$$u\big|_{(0,T)\times\partial\Omega} = 0\,, \tag{3.2}$$

and

$$u(0) = u_0 \qquad \text{in } \Omega\,. \tag{3.3}$$

We introduce a bilinear form

$$a(v,w) = \int_\Omega [A(\mathbf{x})\nabla v(\mathbf{x})]\cdot[\nabla w(\mathbf{x})]\,d\mathbf{x} \qquad \forall\, v,w \in H^1(\Omega).$$

We assume that $A : \Omega \to \mathbb{R}^{d\times d}$ is uniformly positive definite and essentially bounded so that

$$a(v,v) \geq C_0\|v\|_1^2 \qquad \forall\, v \in H_0^1(\Omega) \tag{3.4}$$

and

$$a(v,w) \leq C_1\|v\|_1\|w\|_1 \qquad \forall\, v,w \in H^1(\Omega). \tag{3.5}$$

We define a weak formulation of (3.1)–(3.3) as follows: find

$$u \in X \tag{3.6}$$

such that

$$\langle \partial_t u(t), v\rangle + a(u(t),v) = \langle f(t), v\rangle \qquad \forall\, v \in H_0^1(\Omega),\ \text{a.e. } t, \tag{3.7}$$

and

$$(u(0), z) = (u_0, z) \qquad \forall\, z \in L^2(\Omega)\,. \tag{3.8}$$

Here, a.e. t means "for almost every $t \in (0,T)$."

Throughout this paper, V^h denotes a family of finite element subspaces of $H_0^1(\Omega)$ satisfying (2.1)–(2.5). The semidiscrete finite element approximation of the weak formulation (3.6)–(3.8) is defined by

$$u^h \in H^1(0,T;V^h)\,, \tag{3.9}$$

$$(\partial_t u^h(t), v^h) + a(u^h(t), v^h) = \langle f(t), v^h\rangle \qquad \forall\, v^h \in V^h,\ \text{a.e. } t\,, \tag{3.10}$$

and

$$u^h(0) = u_0^h\,, \tag{3.11}$$

where $u_0^h \in V^h$ is a suitable approximation of u_0.

The initial conditions (3.8) and (3.11) make sense because of part (i) of the following embedding results.

LEMMA 3.1. *The following two embedding results hold for X:*

(i) *X is continuously embedded into $C([0,T];L^2(\Omega))$. Furthermore, for $v \in X$, the mapping $t \mapsto \|v(t)\|$ is absolutely continuous on $[0,T]$ with*

$$\frac{1}{2}\frac{d}{dt}\|v(t)\|^2 = \langle v'(t), v(t)\rangle \qquad \text{a.e. } t \in [0,T]\,.$$

(ii) *X is compactly embedded into $L^2(0,T;L^2(\Omega))$.*

Proof. See, e.g., [21, p. 287] for part (i) and [35, p. 271, p. 274] for part (ii). ∎

Parallel to standard finite element theories for elliptic problems, we will derive error estimates in the norm of the solution space for the parabolic weak formulation and establish parabolic Cea's lemma and Aubin–Nitsche's lemma. (For the elliptic version of Cea's lemma and Aubin–Nitsche's lemma, see, e.g., [11].) We will also prove some additional estimates.

3.1. Semidiscrete error estimates in the norm of the solution space. We will estimate the errors in the X-norm by estimating $u-u^h$ and $\partial_t u - \partial_t u^h$ separately. We have the following error estimates for $u - u^h$.

THEOREM 3.2. *Let $u \in X$ be the solution of* (3.7)–(3.8) *and $u^h \in H^1(0,T;V^h)$ be the solution of* (3.10)–(3.11). *Assume that*

$$\|u_0 - u_0^h\| \to 0 \quad \text{as } h \to 0\,. \tag{3.12}$$

Then, for every $t \in [0,T]$,

$$\|u(t) - u^h(t)\|^2 + \int_0^T \|u(s) - u^h(s)\|_1^2\, ds \to 0 \quad \text{as } h \to 0\,. \tag{3.13}$$

If, in addition,

$$u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega)) \qquad \text{for some } m \in [0,k] \tag{3.14}$$

and $u_0^h = P^h u_0$, then

$$\begin{aligned}\|u(t) - u^h(t)\|^2 &+ \int_0^T \|u(s) - u^h(s)\|_1^2\, ds \\ &\le Ch^{2m}\Big(\|u\|^2_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|^2_{L^2(0,T;H^{m-1}(\Omega))}\Big).\end{aligned} \tag{3.15}$$

Proof. By subtracting (3.10) from (3.7) we obtain the "orthogonality" condition

$$\langle \partial_t u(t) - \partial_t u^h(t), v^h\rangle + a(u(t)-u^h(t), v^h) = 0 \qquad \forall\, v^h \in V^h,\ \text{a.e. } t\,. \tag{3.16}$$

This leads to the relation

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|u(t)-u^h(t)\|^2 + a(u(t)-u^h(t), u(t)-u^h(t)) \\ &\quad = \langle \partial_t u(t) - \partial_t u^h(t), u(t)-u^h(t)\rangle + a(u(t)-u^h(t), u(t)-u^h(t)) \\ &\quad = \langle \partial_t u(t) - \partial_t u^h(t), u(t)-v^h(t)\rangle + a(u(t)-u^h(t), u(t)-v^h(t)) \end{aligned} \tag{3.17}$$

for every $v^h(t) \in H^1(0,T;V^h)$ and almost every $t \in (0,T)$. Let $P^h u(t)$ be defined through (2.6) for almost every t. Then, we see that $P^h u$ has the same regularity as u, i.e., $P^h u \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$. Using repeatedly the fact that

$$(u(t) - P^h u(t), w^h) = 0 \qquad \forall\, w^h \in V^h,\ \text{a.e. } t\,,$$

we obtain

$$\begin{aligned} &\langle \partial_t u(t) - \partial_t u^h(t), u(t) - P^h u(t)\rangle = \langle \partial_t u(t), u(t) - P^h u(t)\rangle \\ &= \langle \partial_t u(t) - \partial_t P^h u(t), u(t) - P^h u(t)\rangle = \frac{1}{2}\frac{d}{dt}\|u(t) - P^h u(t)\|^2 \qquad \text{a.e. } t\,. \end{aligned} \tag{3.18}$$

Thus, by setting $v^h(t) = P^h u(t)$ in (3.17) and with the help of (3.18), (3.4), and (3.5), we are led to

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|u(t)-u^h(t)\|^2 + C_0\|u(t)-u^h(t)\|_1^2 \\ &\le \frac{1}{2}\frac{d}{dt}\|u(t)-u^h(t)\|^2 + a(u(t)-u^h(t), u(t)-u^h(t)) \\ &= \frac{1}{2}\frac{d}{dt}\|u(t)-P^h u(t)\|^2 + a(u(t)-u^h(t), u(t)-P^h u(t)) \\ &\le \frac{1}{2}\frac{d}{dt}\|u(t)-P^h u(t)\|^2 + \frac{C_0}{2}\|u(t)-u^h(t)\|_1^2 + C\|u(t)-P^h u(t)\|_1^2 \end{aligned} \tag{3.19}$$

for almost every t. Integrating (3.19) in t and eliminating the factor $(1/2)$ we obtain

$$\begin{aligned} &\|u(t)-u^h(t)\|^2 + C_0\int_0^t \|u(s)-u^h(s)\|_1^2\, ds \\ &\quad \le \|u_0 - u^h(0)\|^2 + \|u(t)-P^h u(t)\|^2 - \|u(0)-P^h u(0)\|^2 \\ &\qquad + C\int_0^t \|u(s)-P^h u(s)\|_1^2\, ds \\ &\quad \le \|u_0 - u_0^h\|^2 + \|u(t_1)-P^h u(t_1)\|^2 - \|u(0)-P^h u(0)\|^2 \\ &\qquad + C\int_0^t \|u(s)-P^h u(s)\|_1^2\, ds \qquad \forall\, t \in [0,T]\,, \end{aligned} \tag{3.20}$$

where $t_1 \in [0,T]$ satisfies $\|u(t_1)-P^h u(t_1)\|^2 = \max_{t\in[0,T]}\|u(t)-P^h u(t)\|^2$. Note that (3.20) makes sense for all t because of Lemma 3.1. Also, we note that (3.14) implies $u \in C([0,T];H^m(\Omega))$ so that $u_0 \in H^m(\Omega)$. Thus, (3.13) follows from (3.20), (3.12), (2.1) with $s=0$, and (2.16).

If $u_0^h = P^h u_0$, then (3.20) simplifies to

$$\begin{aligned}(3.21)\qquad &\|u(t)-u^h(t)\|^2 + C_0\int_0^t \|u(s)-u^h(s)\|_1^2\,ds \\ &\qquad \le \|u(t_1)-P^h u(t_1)\|^2 + C\int_0^T \|u(s)-P^h u(s)\|_1^2\,ds \qquad \forall\, t\in[0,T]\end{aligned}$$

so that under the assumption $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega))$, (3.15) follows from (3.21), (2.3), Lemma 3.1, and (2.18). □

Remark. Note that (3.14) implies that $u \in C([0,T];H^m(\Omega))$ so that u_0 necessarily satisfies $u_0 \in H^m(\Omega)$. Using the differential equation we also deduce that $f \in L^2(0,T;H^{m-1}(\Omega))$. On the other hand, if we assume $f \in L^2(0,T;H^{m-1}(\Omega))$ and $u_0 \in H^m(\Omega)$, we need certain compatibility conditions and some further conditions on f in order to deduce that $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega))$; see, e.g., [21, p. 365] and [28, pp. 386–387]. In Theorem 3.2 and the remaining theorems in this paper concerning the order of error estimates, we simply assume (3.14) holds without stating precisely the compatibility conditions and the conditions on f. Based on [21, p. 365] and [28, pp. 386–387], we see that regularity assumption (3.14) on u not only is nonvacuous, but holds quite generally. □

Remark. The use of the L^2-projection $P^h u(t)$ played an important role in the treatment of the $\partial_t u - \partial_t u^h$ term. [18, Theorem 3.1] derived an estimate seemingly similar to the first inequality in (3.20). However, that estimate contained an extra term, $\|\partial_t u - \partial_t u^h\|^2_{L^2(0,T;L^2(\Omega))}$, that cannot be estimated under the minimal regularity $\partial_t u \in L^2(0,T;H^{-1}(\Omega))$. [22, Theorem 4.1] derived an estimate involving the term $\|\partial_t u - \partial_t v^h\|^2_{L^2(0,T;H^{-1}(\Omega))}$ which is in the minimal regularity norm. However, that estimate by itself does not directly yield the convergence results or error estimates of this paper under minimal regularity; see [22, Theorem 4.3]. □

Now we turn to the convergence proof and the estimate for the error $\partial_t u - \partial_t u^h$ in $L^2(0,T;H^{-1}(\Omega))$.

THEOREM 3.3. *Let $u \in X$ be the solution of* (3.7)–(3.8) *and $u^h \in H^1(0,T;V^h)$ be the solution of* (3.10)–(3.11). *Assume that* (3.12) *holds. Then,*

$$(3.22)\qquad \|\partial_t u - \partial_t u^h\|_{L^2(0,T;H^{-1}(\Omega))} \to 0 \quad \text{as } h\to 0\,.$$

If, in addition, (3.14) *holds and $u_0^h = P^h u_0$, then*

$$\begin{aligned}(3.23)\qquad &\|\partial_t u - \partial_t u^h\|_{L^2(0,T;H^{-1}(\Omega))} \\ &\qquad \le Ch^m\Big(\|u\|_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|_{H^1(0,T;H^{m-1}(\Omega))}\Big).\end{aligned}$$

Proof. Using the orthogonality condition (3.16), we have

$$\begin{aligned}\langle \partial_t u(t) - \partial_t u^h(t), v\rangle &= \langle \partial_t u(t) - \partial_t u^h(t), Q^h v\rangle + \langle \partial_t u(t) - \partial_t u^h(t), v - Q^h v\rangle \\ &= -a(u(t)-u^h(t), Q^h v) + \langle \partial_t u(t) - \partial_t u^h(t), v - Q^h v\rangle\,.\end{aligned}$$

Note that $\langle \partial_t u^h(t), v - Q^h v\rangle = 0 = \langle Q^h \partial_t u(t), v - Q^h v\rangle$ so that the previous equality can be rewritten as

$$\begin{aligned}\langle \partial_t u(t) - \partial_t u^h(t), v\rangle &= -a(u(t)-u^h(t), Q^h v) + \langle \partial_t u(t) - Q^h\partial_t u(t), v - Q^h v\rangle \\ &\le C\|u(t)-u^h(t)\|_1\,\|Q^h v\|_1 + \|\partial_t u(t) - Q^h\partial_t u(t)\|_{-1}\|v - Q^h v\|_1 \qquad \text{a.e. } t\,.\end{aligned}$$

Taking the supremum over $v \in H_0^1(\Omega)$ with $\|v\|_1 = 1$ and noting that $\|P^h v - v\|_1 \le C\|R^h v - v\|_1 \le C\|v\|_1$, $\|P^h v\|_1 \le \|P^h v - v\|_1 + \|v\|_1 \le C\|R^h v - v\|_1 + \|v\|_1 \le C\|v\|_1$, and $P^h v = Q^h v$, we obtain

$$\begin{aligned} &\|\partial_t u(t) - \partial_t u^h(t)\|_{-1} \\ &\qquad \le C\|u(t) - u^h(t)\|_1 + C\|\partial_t u(t) - Q^h \partial_t u(t)\|_{-1} \qquad \text{a.e. } t\,. \end{aligned} \tag{3.24}$$

Integration in t in (3.24) yields

$$\begin{aligned} &\|\partial_t u - \partial_t u^h\|_{L^2(0,T;H^{-1}(\Omega))} \\ &\qquad \le C\|u - u^h\|_{L^2(0,T;H_0^1(\Omega))} + C\|\partial_t u - Q^h \partial_t u\|_{L^2(0,T;H^{-1}(\Omega))}\,. \end{aligned} \tag{3.25}$$

Thus, (3.22) follows from (3.25), (3.13), and (2.17), and (3.23) follows from (3.25), (3.15), and (2.19). ∎

3.2. Parabolic Cea's lemma and Aubin–Nitsche's lemma. As an immediate consequence of (2.20), (2.21), (3.21), and (3.25), we obtain the following parabolic version of Cea's lemma.

THEOREM 3.4. *Let $u \in X$ be the solution of* (3.7)–(3.8) *and $u^h \in H^1(0,T;V^h)$ be the solution of* (3.10)–(3.11). *Assume $u_0^h = P^h u_0$. Then*

$$\|u - u^h\|_X \le C\|u - v^h\|_X \qquad \forall\, v^h \in H^1(0,T;V^h),$$

i.e.,

$$\begin{aligned} &\|u(t) - u^h(t)\|_{L^2(\Omega)} + \|u - u^h\|_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t u - \partial_t u^h\|_{L^2(0,T;H^{-1}(\Omega))} \\ &\le C\Big(\|u(t) - v^h(t)\|_{L^2(\Omega)} + \|u - v^h\|_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t u - \partial_t v^h\|_{L^2(0,T;H^{-1}(\Omega))}\Big) \end{aligned}$$

for every $v^h \in H^1(0,T;V^h)$. Furthermore,

$$\begin{aligned} &\|u(t) - u^h(t)\|_{L^\infty(0,T;L^2(\Omega))} + \|u - u^h\|_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t u - \partial_t u^h\|_{L^2(0,T;H^{-1}(\Omega))} \\ &\le C\Big(\|u - v^h\|_{L^\infty(0,T;L^2(\Omega))} + \|u - w^h\|_{L^2(0,T;H_0^1(\Omega))} + \|\partial_t u - \partial_t z^h\|_{L^2(0,T;H^{-1}(\Omega))}\Big) \end{aligned}$$

for every $v^h, w^h, z^h \in H^1(0,T;V^h)$. ∎

Remark. Thanks to Lemma 3.1, we may freely add or delete the term

$$\|v\|_{L^\infty(0,T;L^2(\Omega))}$$

in the definition of the norm for X. ∎

Next we prove a parabolic Aubin–Nitsche lemma.

THEOREM 3.5. *Let $u \in X$ be the solution of* (3.7)–(3.8) *and $u^h \in H^1(0,T;V^h)$ be the solution of* (3.10)–(3.11). *Assume that $u_0^h = P^h u_0$. Then*

$$\|u - u^h\|_{L^2(0,T;L^2(\Omega))} \le Ch\|u - u^h\|_{L^2(0,T;H^1(\Omega))}. \tag{3.26}$$

Proof. Using the embedding

$$L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \hookrightarrow C([0,T];L^2(\Omega))$$

we have that $u \in C([0,T];L^2(\Omega))$ and $u^h \in C([0,T];L^2(\Omega))$. Thus for every $t \in [0,T]$ we may define $z(t) \in H_0^1(\Omega)$ and $z^h(t) \in V^h$ as the solutions of the following continuous and discrete elliptic problems, respectively:

$$a(z(t), v) = (u(t) - u^h(t), v) \qquad \forall\, v \in H_0^1(\Omega) \tag{3.27}$$

and

$$a(z^h(t), v^h) = (u(t) - u^h(t), v^h) \qquad \forall\, v^h \in V^h. \tag{3.28}$$

Setting $v = u(t) - u^h(t)$ in (3.27) we obtain

$$\begin{aligned}\|u(t) - u^h(t)\|_0^2 &= a(z(t), u(t) - u^h(t)) \\ &= a(z(t) - z^h(t), u(t) - u^h(t)) + a(z^h(t), u(t) - u^h(t)),\end{aligned}$$

which, upon an application of the "orthogonality condition" (3.16), can be rewritten as

$$\|u(t) - u^h(t)\|_0^2 = a(z(t) - z^h(t), u(t) - u^h(t)) - \langle \partial_t u(t) - \partial_t u^h(t), z^h(t)\rangle. \tag{3.29}$$

Differentiating (3.28) and then setting $v^h = z^h(t)$ yield

$$\frac{1}{2}\frac{d}{dt} a(z^h(t), z^h(t)) = a(\partial_t z^h(t), z^h(t)) = \langle \partial_t u(t) - \partial_t u^h(t), z^h(t)\rangle \quad \text{a.e. } t. \tag{3.30}$$

Substitution of (3.30) into (3.29) and integration over $t \in [0, T]$ lead us to

$$\begin{aligned}\|u - u^h\|^2_{L^2(0,T;L^2(\Omega))} \le C\|z - z^h\|_{L^2(0,T;H^1(\Omega))}\, \|u - u^h\|_{L^2(0,T;H^1(\Omega))} \\ - \frac{1}{2} a(z^h(T), z^h(T)) + \frac{1}{2} a(z^h(0), z^h(0)).\end{aligned} \tag{3.31}$$

From (3.28) at $t = 0$ we see that

$$a(z^h(0), z^h(0)) = (u(0) - u^h(0), z^h(0)) = (u(0) - P^h u(0), z^h(0)) = 0.$$

Also obviously,

$$a(z^h(T), z^h(T)) \ge 0.$$

Elliptic regularity on convex domains implies that $z(t) \in H^2(\Omega)$ with the estimate $\|z(t)\|_2 \le C\|u(t) - u^h(t)\|_0$. Standard error estimates for the finite element approximations of elliptic problems yield

$$\|z(t) - z^h(t)\|_1 \le Ch\, \|z(t)\|_2 \le Ch\, \|u(t) - u^h(t)\|_0. \tag{3.32}$$

By combining (3.31)–(3.32) we derive

$$\|u - u^h\|^2_{L^2(0,T;L^2(\Omega))} \le Ch\|u - u^h\|_{L^2(0,T;L^2(\Omega))}\, \|u - u^h\|_{L^2(0,T;H^1(\Omega))},$$

which yields (3.26) upon cancelling the common factor $\|u - u^h\|_{L^2(0,T;L^2(\Omega))}$. ∎

3.3. Pointwise-in-time error estimates. If we make stronger regularity assumptions on u, i.e., $u \in L^2(0, T; H^3(\Omega)) \cap H^1(0, T; H^1(\Omega))$, then we may obtain pointwise-in-time error estimates for $\|\partial_t u(t) - \partial_t u^h(t)\|_{-1}$ and $\|u(t) - u^h(t)\|_1$. We may also obtain an error estimate for $\|\partial_t u - \partial_t u^h\|_{L^2(0,T;L^2(\Omega))}$.

THEOREM 3.6. *Let $u \in X$ be the solution of* (3.7)–(3.8) *and $u^h \in H^1(0, T; V^h)$ be the solution of* (3.10)–(3.11). *Assume that $k \ge 2$,*

$$u \in L^2(0, T; H^{m+1}(\Omega)) \cap H^1(0, T; H^{m-1}(\Omega)) \qquad \textit{for some } m \in [2, k]\,, \tag{3.33}$$

and $u_0^h = P^h u_0$. *Then*

$$\|\partial_t u - \partial_t u^h\|_{L^2(0,T;L^2(\Omega))} \to 0 \quad \text{as } h \to 0\,, \tag{3.34}$$

$$\|u(t) - u^h(t)\|_1 \to 0 \quad \text{as } h \to 0 \quad \forall\, t \in [0,T]\,, \tag{3.35}$$

$$\|\partial_t u(t) - \partial_t u^h(t)\|_{-1} \to 0 \quad \text{as } h \to 0 \quad \forall\, t \in [0,T]\,, \tag{3.36}$$

$$\begin{aligned} &\|\partial_t u - \partial_t u^h\|_{L^2(0,T;L^2(\Omega))} \\ &\qquad \le C h^{m-1}\Big(\|u\|_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|_{H^1(0,T;H^{m-1}(\Omega))}\Big)\,, \end{aligned} \tag{3.37}$$

$$\begin{aligned} &\|u(t) - u^h(t)\|_1 \\ &\qquad \le C h^{m-1}\Big(\|u\|_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|_{H^1(0,T;H^{m-1}(\Omega))}\Big) \qquad \forall\, t \in [0,T]\,, \end{aligned} \tag{3.38}$$

and

$$\begin{aligned} &\|\partial_t u(t) - \partial_t u^h(t)\|_{-1} \\ &\qquad \le C h^{m-1}\Big(\|u\|_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|_{H^1(0,T;H^{m-1}(\Omega))}\Big) \qquad \forall\, t \in [0,T]\,. \end{aligned} \tag{3.39}$$

Proof. From (3.33) we deduce that $u_0 \in H^m(\Omega)$ and that $u \in L^2(0,T;H^3(\Omega)) \cap H^1(0,T;H^1(\Omega))$ so that $\partial_t u \in L^2(0,T;H^1(\Omega))$. Using the orthogonality condition (3.16) we obtain

$$(\partial_t u(t) - \partial_t u^h(t), P^h \partial_t u(t)) + a(u(t) - u^h(t), P^h \partial_t u(t)) = 0 \qquad \text{a.e. } t \tag{3.40}$$

and

$$(\partial_t u(t) - \partial_t u^h(t), \partial_t u^h(t)) + a(u(t) - u^h(t), \partial_t u^h(t)) = 0 \qquad \text{a.e. } t\,. \tag{3.41}$$

With the help of (3.40)–(3.41) we obtain

$$\begin{aligned} &\|\partial_t u(t) - \partial_t u^h(t)\|^2 + \frac{1}{2}\frac{d}{dt} a(u(t) - u^h(t), u(t) - u^h(t)) \\ &\quad = (\partial_t u(t) - \partial_t u^h(t), \partial_t u(t) - \partial_t u^h(t)) + a(u(t) - u^h(t), \partial_t u(t) - \partial_t u^h(t)) \\ &\quad = (\partial_t u(t) - \partial_t u^h(t), \partial_t u(t)) + a(u(t) - u^h(t), \partial_t u(t)) \\ &\quad = (\partial_t u(t) - \partial_t u^h(t), \partial_t u(t) - P^h \partial_t u(t)) + a(u(t) - u^h(t), \partial_t u(t) - P^h \partial_t u(t)) \end{aligned}$$

for almost every $t \in [0,T]$. Then, from the defining equation of P^h, i.e., (2.6), we see that

$$(\partial_t u^h(t), \partial_t u(t) - P^h \partial_t u(t)) = 0 = (P^h \partial_t u(t), \partial_t u(t) - P^h \partial_t u(t)) \qquad \text{a.e. } t\,.$$

Combining the last two equations we arrive at

$$\begin{aligned} &\|\partial_t u(t) - \partial_t u^h(t)\|^2 + \frac{1}{2}\frac{d}{dt} a(u(t) - u^h(t), u(t) - u^h(t)) \\ &= (\partial_t u(t) - P^h \partial_t u(t), \partial_t u(t) - P^h \partial_t u(t)) + a(u(t) - u^h(t), \partial_t u(t) - P^h \partial_t u(t)) \\ &\le \|\partial_t u(t) - P^h \partial_t u(t)\|^2 + C\|u(t) - u^h(t)\|_1^2 + C\|\partial_t u(t) - P^h \partial_t u(t)\|_1^2\,. \end{aligned}$$

Integration in t yields

$$\begin{aligned}
&\int_0^t \|\partial_t u(s) - \partial_t u^h(s)\|^2\,ds + \frac{1}{2}a(u(t)-u^h(t),u(t)-u^h(t)) \\
&\quad\le \frac{1}{2}a(u(0)-u^h(0),u(0)-u^h(0)) + \int_0^T \|\partial_t u(t) - P^h\partial_t u(t)\|^2\,dt \\
&\qquad + C\int_0^T \|u(t)-u^h(t)\|_1^2\,dt + C\int_0^T \|\partial_t u(t) - P^h\partial_t u(t)\|_1^2\,dt \\
&\quad\le \frac{1}{2}\|u_0 - P^h u_0\|_1^2 + \int_0^T \|\partial_t u(t) - P^h\partial_t u(t)\|^2\,dt \\
&\qquad + C\int_0^T \|u(t)-u^h(t)\|_1^2\,dt + C\int_0^T \|\partial_t u(t) - P^h\partial_t u(t)\|_1^2\,dt\,.
\end{aligned} \tag{3.42}$$

Thus, (3.34) and (3.35) follow from (3.42), (2.10), (2.1) with $s=1$, (2.12) with $s=0$, (3.13), and (2.16). Relations (3.37) and (3.38) follow from (3.42), (2.10), (2.4), (2.13) with $s=0$, (3.15), and (2.18). Also, (3.36) follows from (3.24), (3.35), (2.11), and (2.1). Finally, (3.39) follows from (3.24), (3.38), (2.11), and (2.2). ∎

Remark. The $H^1(\Omega)$-norm of $P^h\partial_t u(t)$ was used in the proof of Theorem 3.4. For this reason we need $k \ge m \ge 2$. ∎

Remark. We used the fact that $P^h\partial_t u = \partial_t P^h u$, which can be easily verified. ∎

Remark. (3.42) seems to suggest that, for pointwise-in-time estimate, one should assume $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^m(\Omega))$ instead of $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega))$. Under the former regularity together with the assumption $u_0 \in H^{m+1}(\Omega)$, (3.42) yields $\mathcal{O}(h^m)$ estimates instead of $\mathcal{O}(h^{m-1})$ estimates. ∎

4. Semidiscrete error estimates for semilinear parabolic equations. In this section, we derive error estimates for the semidiscrete finite element approximations of the semilinear parabolic problem

$$\partial_t u - \operatorname{div}[A(\mathbf{x})\nabla u] + \mathbf{b}(t,\mathbf{x})\cdot(\nabla u) + \phi(u) = f \qquad \text{in } (0,T)\times\Omega\,, \tag{4.1}$$

$$u\big|_{(0,T)\times\partial\Omega} = 0\,, \tag{4.2}$$

and

$$u(0) = u_0 \qquad \text{in } \Omega \tag{4.3}$$

under minimal regularity assumptions on u. Throughout this section we assume that $A(\mathbf{x})$ possesses sufficient smoothness that guarantees the validity of the regularity theorem [21, p. 360, Theorem 5] for the linear parabolic equation.

The error estimates to be derived in section 4.2 make use of results of [9] (see also [23] and [13]) concerning the approximation of a class of nonlinear problems. These results imply that, under certain hypotheses, the error of approximation of solutions of certain nonlinear problems is basically the same as the error of approximation of solutions of related linear problems. Here, for the sake of completeness, we state the relevant results, specialized to our needs.

4.1. Quotation of results concerning the approximation of a class of nonlinear problems. The nonlinear problems considered in [9] and [23] are of the following type. We seek a $\psi \in \mathcal{X}$ such that

$$\psi + \mathcal{T}\mathcal{G}(\psi) = 0\,, \tag{4.4}$$

where $\mathcal{T} \in \mathcal{L}(\mathcal{Y};\mathcal{X})$, $\mathcal{G}$ is a C^2 mapping from $\mathcal{X}$ into $\mathcal{Y}$, and $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces. We say that ψ is a *regular solution* if $\psi + T\mathcal{G}_\psi(\psi)$ is an isomorphism from $\mathcal{X}$ into $\mathcal{X}$. Here, $\mathcal{G}_\psi(\cdot)$ (or $\mathcal{G}'$ or $D\mathcal{G}$) denotes the Fréchet derivative of $\mathcal{G}(\cdot)$. We assume that there exists another Banach space $\mathcal{Z}$, contained in $\mathcal{Y}$, with continuous embedding, such that

$$\mathcal{G}_\psi(\psi) \in \mathcal{L}(\mathcal{X};\mathcal{Z}) \quad \forall\, \psi \in \mathcal{X}\,. \tag{4.5}$$

Approximations are defined by introducing a subspace $\mathcal{X}^h \subset \mathcal{X}$ and an approximating operator $\mathcal{T}^h \in \mathcal{L}(\mathcal{Y};\mathcal{X}^h)$. We seek a $\psi^h \in \mathcal{X}^h$ such that

$$\psi^h + \mathcal{T}^h\mathcal{G}(\psi^h) = 0\,. \tag{4.6}$$

Concerning the linear operator $\mathcal{T}^h$, we assume the approximation properties

$$\lim_{h\to 0} \|(\mathcal{T}^h - \mathcal{T})\omega\|_{\mathcal{X}} = 0 \quad \forall\, \omega \in \mathcal{Y} \tag{4.7}$$

and

$$\lim_{h\to 0} \|\mathcal{T}^h - \mathcal{T}\|_{\mathcal{L}(\mathcal{Z};\mathcal{X})} = 0\,. \tag{4.8}$$

Note that whenever the imbedding $\mathcal{Z} \subset \mathcal{Y}$ is compact, (4.8) follows from (4.7) and, moreover, (4.5) implies that the operator $\mathcal{T}\mathcal{G}_\psi(\psi) \in \mathcal{L}(\mathcal{X};\mathcal{X})$ is compact.

We can now state the result of [9] (see also [23]) that will be used in section 4.2. In the statement of the theorem, $D^2\mathcal{G}$ represents the second Fréchet derivatives of $\mathcal{G}$.

THEOREM 4.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. Assume that $\mathcal{G}$ is a C^2 mapping from $\mathcal{X}$ into $\mathcal{Y}$ and that $D^2\mathcal{G}$ is bounded on all bounded sets of $\mathcal{X}$. Assume that* (4.5), (4.7), *and* (4.8) *hold and that ψ is a regular solutions of* (4.4). *Then there exists a neighborhood $\mathcal{O}$ of the origin in $\mathcal{X}$ and, for $h \le h_0$ small enough, a unique $\psi^h \in \mathcal{X}^h$ such that ψ^h is a regular solution of* (4.6) *and $\psi^h - \psi \in \mathcal{O}$. Moreover, there exists a constant $C > 0$, independent of h, such that*

$$\|\psi^h - \psi\|_{\mathcal{X}} \le C\|(\mathcal{T}^h - \mathcal{T})\mathcal{G}(\psi)\|_{\mathcal{X}}\,. \qquad \square \tag{4.9}$$

4.2. Recasting the semilinear problem into the Brezzi–Rappaz–Raviart framework. We define the following weak form for the semilinear problem (4.1)–(4.3): find

$$u \in X \tag{4.10}$$

such that

$$\begin{aligned} \langle \partial_t u(t), v\rangle + a(u(t),v) + (\mathbf{b}(t)\cdot\nabla u(t), v) + \big(\phi(u(t)), v\big) \\ = \langle f(t), v\rangle \qquad \forall\, v \in H_0^1(\Omega),\ \text{a.e. } t \end{aligned} \tag{4.11}$$

and

$$(u(0), z) = (u_0, z) \qquad \forall\, z \in L^2(\Omega)\,. \tag{4.12}$$

Certain conditions must be imposed on ϕ in order to guarantee the existence of a solution for (4.10)–(4.12) in X. In this paper, we limit ourselves to finite element analysis only. Thus, we will simply assume the existence of a solution for (4.10)–(4.12) and then try to derive error estimates for its semidiscrete finite element solution. Of course, there are many choices of ϕ (e.g., $\phi(u) = u^{11/5}$) that guarantee a solution to (4.10)–(4.12) so that our error estimation is not vacuous.

Let V^h be a family of finite element subspaces of $H_0^1(\Omega)$ satisfying (2.1)–(2.5). We define a semidiscrete finite element approximation of the weak formulation (4.10)–(4.12) by

$$u^h \in H^1(0,T;V^h)\,, \tag{4.13}$$

$$\begin{aligned}(\partial_t u^h(t), v^h) + a(u^h(t), v^h) + (\mathbf{b}(t)\cdot\nabla u^h(t), v^h) + (\phi(u^h(t), v^h)\\ = \langle f(t), v^h\rangle \quad \forall\, v^h \in V^h, \text{ a.e. } t\,,\end{aligned} \tag{4.14}$$

and

$$(u^h(0), z^h) = (u_0, z^h) \qquad \forall\, z^h \in V^h\,. \tag{4.15}$$

Error estimates for such approximations were derived in [38, Chapter 14] and [39] under certain smoothness assumptions on the solution. We will derive error estimates under minimal regularity in the sense we described in section 1, and we determine the largest growth exponent for the derivative of the semilinear term that guarantees a solution with the minimal regularity. For the same order error estimates as those of [38, Chapter 14] and [39], we assume less smoothness on the solution than what was required in [38] or [39]. Also, we obtain an estimate for $\partial_t u - \partial_t u^h$ which was absent from [38] or [39].

To derive error estimates for the semidiscrete approximations of the semilinear problem, we first fit the problem into the Brezzi–Rappaz–Raviart framework described in section 4.1. Then by verifying all assumptions of the Brezzi–Rappaz–Raviart theory we obtain the desired error estimates.

We set

$$\mathcal{X} = X \equiv L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$$

with the norm $\|v\|_{\mathcal{X}}^2 = \|v\|_{L^2(0,T;H_0^1(\Omega))}^2 + \|\partial_t v\|_{L^2(0,T;H^{-1}(\Omega))}^2$ for all $v \in \mathcal{X}$ and

$$\mathcal{Y} = L^2(0,T;H^{-1}(\Omega)) \times L^2(\Omega)$$

with the norm $\|(y_1,y_2)\|_{\mathcal{Y}}^2 = \|y_1\|_{L^2(0,T;H^{-1}(\Omega))}^2 + \|y_2\|^2$ for all $y = (y_1,y_2) \in \mathcal{Y}$. We introduce the linear operator $\mathcal{T} : \mathcal{Y} \to \mathcal{X}$ to be the solution operator for the linear parabolic problem, i.e., $\mathcal{T}(\widetilde{f},\widetilde{u}_0) = \widetilde{u}$ for $(\widetilde{f},\widetilde{u}_0) \in \mathcal{Y}$ and $\widetilde{u} \in \mathcal{X}$ if and only if

$$\langle \partial_t \widetilde{u}(t), v\rangle + a(\widetilde{u}(t), v) = \langle \widetilde{f}(t), v\rangle \qquad \forall\, v \in H_0^1(\Omega), \text{ a.e. } t,$$

and

$$(\widetilde{u}(0), z) = (\widetilde{u}_0, z) \qquad \forall\, z \in L^2(\Omega)\,.$$

We define $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ by $\mathcal{G}(v) = (-f + \mathbf{b}\cdot\nabla v + \phi(v), -u_0)$ for all $v \in \mathcal{X}$.

Let $\mathcal{X}^h = H^1(0,T;V^h)$. We define the linear operator $\mathcal{T}^h : \mathcal{Y} \to \mathcal{X}^h$ to be the semidiscrete solution operator for the linear parabolic problem, i.e., $\mathcal{T}^h(\widetilde{f},\widetilde{u}_0) = \widetilde{u}^h$ for $(\widetilde{f},\widetilde{u}_0) \in \mathcal{Y}$ and $\widetilde{u}^h \in \mathcal{X}^h$ if and only if

$$(\partial_t \widetilde{u}^h(t), v^h) + a(\widetilde{u}^h(t), v^h) = \langle \widetilde{f}(t), v^h \rangle \qquad \forall\, v^h \in V^h, \text{ a.e. } t,$$

and

$$(\widetilde{u}^h(0), z^h) = (\widetilde{u}_0, z^h) \qquad \forall\, z^h \in V^h\,.$$

Clearly, (4.11)–(4.12) is equivalent to

$$u + \mathcal{T}\mathcal{G}(u) = 0,$$

and (4.14)–(4.15) is equivalent to

$$u^h + \mathcal{T}^h\mathcal{G}(u^h) = 0\,.$$

In other words, we have recast the semilinear problem and its approximation into the form of (4.4) and (4.6).

4.3. Semidiscrete error estimates for the approximation of semilinear parabolic equations. Our goal is to obtain convergence and error estimates in the X-norm for finite element approximations of the semilinear problem. We first examine conditions on ϕ that will guarantee the meaningfulness of the term $\phi(u)$ when $u \in X$. We consider the three-dimensional case only ($\Omega \subset \mathbb{R}^3$). Note that $L^2(0,T;H_0^1(\Omega)) \subset L^2(0,T;L^6(\Omega))$. If $\phi(s) = O(|s|^\alpha)$, then in order for the term $\phi(u) = O(|u|^\alpha)$ to make sense in the function space $L^2(0,T;H^{-1}(\Omega))$, we need $|u|^\alpha \in L^2(0,T;L^{6/5}(\Omega))$, which is equivalent to $u \in L^{2\alpha}(0,T;L^{6\alpha/5}(\Omega))$. The next lemma determines the largest value of allowable α when $\Omega \subset \mathbb{R}^3$.

LEMMA 4.2. *Assume $\Omega \subset \mathbb{R}^3$ and let $\overline{\alpha} \equiv 7/3$. Then the embedding*

$$L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \subset L^{2\alpha_1}(0,T;L^{6\alpha_2/5}(\Omega)) \tag{4.16}$$

is continuous for all $\alpha_1, \alpha_2 \in [1,\overline{\alpha}]$ and compact for all $\alpha_1, \alpha_2 \in [1,\overline{\alpha})$.

Proof. To show (4.16) is a continuous embedding for all $\alpha_1, \alpha_2 \in [1,\overline{\alpha}]$, it suffices to show that the embedding

$$\begin{aligned} &L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \\ &\qquad \subset L^{2\overline{\alpha}}(0,T;L^{\frac{6\overline{\alpha}}{5}}(\Omega)) = L^{\frac{14}{3}}(0,T;L^{\frac{14}{5}}(\Omega)) \end{aligned} \tag{4.17}$$

is continuous. It is well known that the embeddings

$$L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \subset L^2(0,T;L^6(\Omega))$$

and

$$L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \subset L^\infty(0,T;L^2(\Omega))$$

are continuous. For each $y \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$ the interpolation of L^6 and L^2 yields

$$\|y(t)\|_{L^q(\Omega)} \le C\|y(t)\|_{L^6(\Omega)}^{1-\theta}\,\|y(t)\|_{L^2(\Omega)}^{\theta}$$

for all $\theta \in [0,1]$ with $q \in [2,6]$ determined by

$$\frac{1}{q} = \frac{1-\theta}{6} + \frac{\theta}{2}\,. \tag{4.18}$$

This leads to

$$\|y(t)\|^{\frac{2}{1-\theta}}_{L^q(\Omega)} \le C\|y(t)\|^2_{L^6(\Omega)}\,\|y(t)\|^{\frac{2\theta}{1-\theta}}_{L^2(\Omega)}$$

so that

$$\|y\|_{L^{\frac{2}{1-\theta}}(0,T;L^q(\Omega))} \le C\|y\|^{\theta}_{L^\infty(0,T;L^2(\Omega))}\,\|y\|^{1-\theta}_{L^2(0,T;L^6(\Omega))}\,.$$

Thus, the embedding

$$L^2(0,T;H^1_0(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)) \subset L^{\frac{2}{1-\theta}}(0,T;L^q(\Omega)) \tag{4.19}$$

is continuous for all $\theta \in [0,1]$ with q determined by (4.18). By choosing $\theta = 4/7$ in (4.19), we see that the embedding (4.17) is continuous.

Next, we show the embedding (4.16) is compact for all $\alpha_1, \alpha_2 \in [1,\overline{\alpha})$. Let $\alpha_1, \alpha_2 \in [1,\overline{\alpha})$ be given. For each $y \in L^2(0,T;H^1_0(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$ the interpolation of $L^{2\overline{\alpha}}(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))$ and $L^2(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))$ yields

$$\begin{aligned}&\left(\int_0^T \|y(t)\|^{2\alpha_1}_{L^{\frac{6\alpha_2}{5}}(\Omega)}\,dt\right)^{\frac{1}{2\alpha_1}}\\ &\quad\le C\left(\int_0^T \|y(t)\|^{2\overline{\alpha}}_{L^{\frac{6\alpha_2}{5}}(\Omega)}\,dt\right)^{\frac{1-\theta_1}{2\overline{\alpha}}}\left(\int_0^T \|y(t)\|^2_{L^{\frac{6\alpha_2}{5}}(\Omega)}\,dt\right)^{\frac{\theta_1}{2}},\end{aligned} \tag{4.20}$$

where θ_1 satisfies

$$\frac{1}{2\alpha_1} = \frac{1-\theta_1}{2\overline{\alpha}} + \frac{\theta_1}{2}\,.$$

The interpolation of $L^{\frac{6\overline{\alpha}}{5}}(\Omega)$ and $L^2(\Omega)$ and Holder's inequality imply

$$\begin{aligned}\int_0^T \|y(t)\|^2_{L^{\frac{6\alpha_2}{5}}(\Omega)}\,dt &\le C\int_0^T \|y(t)\|^{2(1-\theta_2)}_{L^{\frac{6\overline{\alpha}}{5}}(\Omega)}\,\|y(t)\|^{2\theta_2}_{L^2(\Omega)}\,dt\\ &\le C\left(\int_0^T \|y(t)\|^2_{L^{\frac{6\overline{\alpha}}{5}}(\Omega)}\,dt\right)^{1-\theta_2}\left(\int_0^T \|y(t)\|^2_{L^2(\Omega)}\,dt\right)^{\theta_2},\end{aligned} \tag{4.21}$$

where θ_2 satisfies

$$\frac{5}{6\alpha_2} = \frac{5(1-\theta_2)}{6\overline{\alpha}} + \frac{\theta_2}{2}.$$

Substituting (4.21) into (4.20), we obtain

$$\begin{aligned}&\|y\|_{L^{2\alpha_1}(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))}\\ &\le C\left|\int_0^T \|y(t)\|^{2\overline{\alpha}}_{L^{\frac{6\alpha_2}{5}}(\Omega)}\,dt\right|^{\frac{1-\theta_1}{2\overline{\alpha}}}\left|\int_0^T \|y(t)\|^2_{L^{\frac{6\overline{\alpha}}{5}}(\Omega)}\,dt\right|^{\frac{\theta_1(1-\theta_2)}{2}}\left|\int_0^T \|y(t)\|^2_{L^2(\Omega)}\,dt\right|^{\frac{\theta_1\theta_2}{2}}\\ &= C\|y\|^{1-\theta_1}_{L^{2\overline{\alpha}}(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))}\|y\|^{\theta_1(1-\theta_2)}_{L^2(0,T;L^{\frac{6\overline{\alpha}}{5}}(\Omega))}\|y\|^{\theta_1\theta_2}_{L^2(0,T;L^2(\Omega))}\,.\end{aligned} \tag{4.22}$$

Let $\{y_n\}$ be a weakly convergent sequence in $L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$. Then $\{y_n\}$ is bounded in $L^{2\overline{\alpha}}(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))$ and in $L^2(0,T;L^{\frac{6\overline{\alpha}}{5}}(\Omega))$ (because of the continuous imbedding (4.17)), and $\{y_n\}$ converges strongly in $L^2(0,T;L^2(\Omega))$ (see Lemma 3.1, part (ii)). Thus, (4.22) implies that $\{y_n\}$ converges strongly in $L^{2\alpha_1}(0,T;L^{\frac{6\alpha_2}{5}}(\Omega))$. ∎

Based on the recast problems in section 4.2, we now verify all assumptions of Theorem 4.1 and derive the following error estimates for the semidiscrete solution u^h of the semilinear problem.

THEOREM 4.3. *Let $u \in X$ be a regular solution of (4.10)–(4.12). Assume that $\mathbf{b} \in L^\infty(0,T;\mathbf{L}^\gamma(\Omega))$ for some $\gamma > d$ ($d = 2$ or 3 being the space dimension), $\phi \in C^2(\mathbb{R};\mathbb{R})$,*

$$|\phi'(s)| \le C|s|^{\alpha-1} \quad \text{and} \quad |\phi''(s)| \le C|s|^{\alpha-2} \qquad \forall\, s \in \mathbb{R} \tag{4.23}$$

for an $\alpha \in [2, 7/3)$ when $d = 3$ or for an $\alpha \in [2,\infty)$ when $d = 2$. Then there exists a sufficiently small $h_0 > 0$ such that for all $h \in (0, h_0)$, (4.13)–(4.15) has a unique regular solution $u^h \in H^1(0,T;V^h)$ satisfying

$$\begin{aligned}\|u(t) - u^h(t)\|^2 + \|u - u^h\|^2_{L^2(0,T;H^1(\Omega))} + \|\partial_t u - \partial_t u^h\|^2_{L^2(0,T;H^{-1}(\Omega))} \\ \to 0 \quad \text{as } h \to 0\ \forall\, t \in [0,T]\,.\end{aligned} \tag{4.24}$$

If, in addition, $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega))$ for some $m \in [0,k]$, then

$$\begin{aligned}&\|u(t) - u^h(t)\|^2 + \|u - u^h\|^2_{L^2(0,T;H^1(\Omega))} + \|\partial_t u - \partial_t u^h\|^2_{L^2(0,T;H^{-1}(\Omega))} \\ &\quad \le Ch^{2m}\Big(\|u\|^2_{L^2([0,T;H^{m+1}(\Omega))} + \|\partial_t u\|^2_{L^2([0,T;H^{m-1}(\Omega))}\Big) \qquad \forall\, t \in [0,T]\,.\end{aligned} \tag{4.25}$$

Proof. We will treat the three-dimensional case only; the two-dimensional case can be handled similarly and more easily.

Let $\mathcal{X}$ and $\mathcal{Y}$ be the spaces defined in section 4.2 and let $\mathcal{T}$, $\mathcal{T}^h$, and $\mathcal{G}$ be the operators defined in section 4.2. Since $\alpha < 7/3 = \overline{\alpha}$, $\gamma > 3$, $\beta_\epsilon \equiv (6-\epsilon)/(5-\epsilon) \to 6/5$ as $\epsilon \to 0$, and $(12-2\epsilon)/(4-\epsilon) \to 3$ as $\epsilon \to 0$, we may fix a sufficiently small $\epsilon > 0$ such that $\beta_\epsilon \alpha < 6\overline{\alpha}/5$ and $(12-2\epsilon)/(4-\epsilon) < \gamma$. For this fixed ϵ, we set $\epsilon_1 = \epsilon/(12-2\epsilon)$ and $\mathcal{Z} = L^2(0,T;H^{-1+\epsilon_1}(\Omega)) \times H_0^{\epsilon_1}(\Omega)$. Note that ϵ_1 is so chosen to guarantee the continuous embedding $H^{1-\epsilon_1}(\Omega) \subset L^{6-\epsilon}(\Omega)$.

Theorems 3.2 and 3.3 imply that

$$\|(\mathcal{T} - \mathcal{T}^h)(\widetilde{f}, \widetilde{u}_0)\|_{\mathcal{X}} \to 0 \quad \text{as } h \to 0$$

for all $(\widetilde{f}, \widetilde{u}_0) \in \mathcal{Y}$.

For any $(\widetilde{f}, \widetilde{u}_0) \in \mathcal{Z} = L^2(0,T;H^{-1+\epsilon_1}(\Omega)) \times H_0^{\epsilon_1}(\Omega)$, a regularity theorem (see [21, p. 360, Theorem 5]) for the solution of the heat equation and interpolation theorems imply that $\widetilde{u} = T(\widetilde{f}, \widetilde{u}_0) \in L^2(0,T,H^{1+\epsilon_1}(\Omega)) \cap H^1(0,T;H^{-1+\epsilon_1}(\Omega))$ and

$$\|\widetilde{u}\|_{L^2(0,T,H^{1+\epsilon_1}(\Omega))} + \|\partial_t \widetilde{u}\|_{L^2(0,T,H^{-1+\epsilon_1}(\Omega))} \le C\Big(\|\widetilde{f}\|_{L^2(0,T,H^{-1+\epsilon_1}(\Omega))} + \|\widetilde{u}_0\|_{H^{\epsilon_1}(\Omega)}\Big)\,.$$

Thus, Theorems 3.2 and 3.3 with $m = \epsilon_1$ give

$$\begin{aligned}\|(\mathcal{T} - \mathcal{T}^h)(\widetilde{f}, \widetilde{u}_0)\|_{\mathcal{X}} &= \|\widetilde{u} - \widetilde{u}^h\|_{\mathcal{X}} \\ &\le Ch^{\epsilon_1}\Big(\|\widetilde{u}\|_{L^2(0,T,H^{1+\epsilon_1}(\Omega))} + \|\partial_t \widetilde{u}\|_{L^2(0,T,H^{-1+\epsilon_1}(\Omega))}\Big) \\ &\le Ch^{\epsilon_1}\Big(\|\widetilde{f}\|_{L^2(0,T,H^{-1+\epsilon_1}(\Omega))} + \|\widetilde{u}_0\|_{H^{\epsilon_1}(\Omega)}\Big) = Ch^{\epsilon_1}\|(\widetilde{f}, \widetilde{u}_0)\|_{\mathcal{Z}}\end{aligned}$$

so that by taking the supremum over $(\widetilde{f},\widetilde{u}_0)\in\mathcal{Z}$, we obtain

$$\|\mathcal{T}-\mathcal{T}^h\|_{\mathcal{L}(\mathcal{Z},\mathcal{X})}\le Ch^{\epsilon_1}\to 0\quad\text{as } h\to 0\,.$$

Next we verify (4.5). Let $v,w\in\mathcal{X}$ be given. A simple calculation reveals

$$[D\mathcal{G}(v)]w=(\mathbf{b}\cdot\nabla w+\phi'(v)w,0).$$

We recall the generalized Holder's inequality

$$\Big|\int_E f_1\,f_2\,f_3\,dE\Big|\le\|f_1\|_{L^{r_1}(E)}\,\|f_2\|_{L^{r_2}(E)}\,\|f_3\|_{L^{r_3}(E)}\,, \tag{4.26}$$

where E is a measurable set of any dimension and $(1/r_1)+(1/r_2)+(1/r_3)=1$. Using (4.26) with $r_1=\alpha\beta_\epsilon/(\alpha-1)$, $r_2=\alpha\beta_\epsilon$, and $r_3=6-\epsilon$, where $\beta_\epsilon=(6-\epsilon)/(5-\epsilon)$ (recall that ϵ was chosen such that $\beta_\epsilon\alpha<\overline{\alpha}$), we obtain

$$\begin{aligned}&\int_0^T\int_\Omega|v|^{\alpha-1}|w|\,|y|\,d\mathbf{x}\,dt\\&\le\int_0^T\|v(t)\|^{\alpha-1}_{L^{\alpha\beta_\epsilon}(\Omega)}\,\|w(t)\|_{L^{\alpha\beta_\epsilon}(\Omega)}\,\|y(t)\|_{L^{6-\epsilon}(\Omega)}\,dt\;\forall\,y\in L^2(0,T;H^{1-\epsilon_1}(\Omega))\,.\end{aligned}\tag{4.27}$$

Applying to (4.27) the inequality (4.26) with $r_1=2\alpha/(\alpha-1)$, $r_2=2\alpha$, and $r_3=2$ we have

$$\begin{aligned}&\int_0^T\int_\Omega|v|^{\alpha-1}|w|\,|y|\,d\mathbf{x}\,dt\\&\quad\le C\|v\|^{\alpha-1}_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}\,\|w\|_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}\,\|y\|_{L^2(0,T;L^{6-\epsilon}(\Omega))}\\&\quad\le C\|v\|^{\alpha-1}_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}\,\|w\|_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}\,\|y\|_{L^2(0,T;H^{1-\epsilon_1}(\Omega))}\end{aligned}$$

for every $y\in L^2(0,T;H^{1-\epsilon_1}(\Omega))$. Taking the supremum in the last estimate over all $y\in L^2(0,T;H^{1-\epsilon_1}(\Omega))$ we see that

$$\big\||v|^{\alpha-1}w\big\|_{L^2(0,T;H^{-1+\epsilon_1}(\Omega))}\le C\|v\|^{\alpha-1}_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}\,\|w\|_{L^{2\alpha}(0,T;L^{\alpha\beta_\epsilon}(\Omega))}.\tag{4.28}$$

We can also estimate the term $\mathbf{b}\cdot\nabla w$ as follows:

$$\begin{aligned}&\int_0^T\int_\Omega|\mathbf{b}||\nabla w|\,|y|\,d\mathbf{x}\,dt\\&\quad\le\int_0^T\|\mathbf{b}\|_{\mathbf{L}^{(12-2\epsilon)/(4-\epsilon)}(\Omega)}\,\|\nabla w(t)\|_{L^2(\Omega)}\,\|y(t)\|_{L^{6-\epsilon}(\Omega)}\,dt\\&\quad\le\|\mathbf{b}\|_{L^\infty(0,T;\mathbf{L}^\gamma(\Omega))}\|w\|_{L^2(0,T;H^1(\Omega))}\|y\|_{L^2(0,T;H^{1-\epsilon_1}(\Omega))}\quad\forall\,y\in L^2(0,T;H^{1-\epsilon_1}(\Omega))\end{aligned}$$

so that

$$\big\|\mathbf{b}\cdot\nabla w\big\|_{L^2(0,T;H^{-1+\epsilon_1}(\Omega))}\le C\|\mathbf{b}\|_{L^\infty(0,T;\mathbf{L}^\gamma(\Omega))}\|w\|_{L^2(0,T;H^1(\Omega))}.\tag{4.29}$$

Thus, (4.28), Lemma 4.2, and (4.29) imply that $D\mathcal{G}(v)\in\mathcal{L}(\mathcal{X},\mathcal{Z})$ for every $v\in\mathcal{X}$.

Similarly, we may use the growth condition for ϕ'' to prove that $D^2\mathcal{G}$ is locally bounded.

Hence all the assumptions in Theorem 4.1 are verified, and we conclude from that theorem that

$$\|u - u^h\|_{\mathcal{X}} \leq \|(\mathcal{T} - \mathcal{T}^h)\mathcal{G}(u)\|_{\mathcal{X}} \to 0 \quad \text{as } h \to 0,$$

and if, in addition, $u \in L^2(0,T;H^{m+1}(\Omega)) \cap H^1(0,T;H^{m-1}(\Omega))$,

$$\|u - u^h\|_{\mathcal{X}} \leq Ch^m \Big(\|u\|_{L^2(0,T;H^{m+1}(\Omega))} + \|\partial_t u\|_{L^2(0,T;H^{m-1}(\Omega))} \Big).$$

Thus, the desired estimates (4.24) and (4.25) follow from the definition of $\|\cdot\|_{\mathcal{X}}$, Lemma 3.1, and the last two relations. ∎

Remark. The proof of Theorem 4.3 used a regularity theorem of [21, p. 360] which was proved with the help of an elliptic regularity theorem. Thus, the convexity assumption on the domain Ω played a role here. Also, since we need only the regularity $L^2(0,T;H^{1+\epsilon_1}(\Omega)) \cap H^1(0,T;H^{-1+\epsilon_1}(\Omega))$ for solutions of the linear equation, we expect that the convexity assumption on Ω can be weakened. ∎

Remark. If $\alpha \in [1,2)$, then we have to use a modified version of Theorem 4.1, namely to replace the assumptions $\mathcal{G} \in C^2$ and $D^2\mathcal{G}$ being locally bounded by the assumptions $\mathcal{G} \in C^1$ and $D\mathcal{G}$ being uniformly continuous. This modified version of Theorem 4.1 can be easily proved upon a routine examination of the proof of [23, p. 307, Theorem 3.3]. Also, (4.23) should be replaced by $\phi'(s) = O(|s|^{\alpha-1})$ and ϕ' is uniformly continuous. For example, for $\phi(s) = s^{3/2}$ we have that $\phi'(s) = (3/2)\sqrt{s}$. The function $\sqrt{s}$ is obviously uniformly continuous away from 0. But on any neighborhood of $s = 0$, $\sqrt{s}$ is also uniformly continuous so that ϕ' is uniformly continuous on its entire domain. ∎

Acknowledgment. The authors thank Oleg Emanouilov for suggesting the ideas for the proof of Lemma 4.2.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic, New York, 1975.
[2] I. Babuska and A. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, (1972), pp. 3–359.
[3] C. Baiocchi and F. Brezzi, *Optimal error estimates for linear parabolic problems under minimal regularity assumptions*, Calcolo, 20 (1983), pp. 143–176.
[4] G.A. Baker, J.H. Bramble, and V. Thomee, *Single step Galerkin approximations for parabolic problems*, Math. Comp., 31 (1977), pp. 818–847.
[5] J.H. Bramble, J. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 35 (1980), pp. 655–677.
[6] J.H. Bramble and P. Sammon, *Efficient higher order single step methods for parabolic problems: Part* 1, Math. Comp., 35 (1980), pp. 655–677.
[7] J.H. Bramble, A.H. Schatz, V. Thomée, and L.B. Wahlbin, *Some convergence estimates for semidiscrete Galerkin type approximations for parabolic equations*, SIAM J. Numer. Anal., 14 (1977), pp. 218–241.
[8] J.H. Bramble and V. Thomee, *Discrete time Galerkin methods for a parabolic boundary value problem*, Ann. Mat. Pura Appl., 101 (1974), pp. 115–152.
[9] F. Brezzi, J. Rappaz, and P. Raviart, *Finite-dimensional approximation of nonlinear problems. Part* I: *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1–25.
[10] H. Chen, *An L^2 and L^∞-Error Analysis for Parabolic Finite Elements Equations with Applications by Superconvergence and Error Expansion*, Ph.D. thesis, Universität Heidelberg, Heidelberg, Germany, 1993.
[11] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[12] M. Crouzeix, S. Larrson, and V. Thomee, *Resolvent estimates for elliptic finite element operators in one dimension*, Math. Comp., 63 (1994), pp. 121–140.
[13] M. Crouzeix and J. Rappaz, *On Numerical Approximation in Bifurcation Theory*, Masson, Paris, 1990.
[14] M. Crouzeix and V. Thomee, *The stability in L_p and W_p^1 of the L^2-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.
[15] M. Crouzeix, V. Thomee, and L.B. Wahlbin, *Error estimates for spatially discrete approximations of semilinear parabolic equations with initial data of low regularity*, Math. Comp., 53 (1989), pp. 25–41.
[16] M. Dobrowolski, *L^∞-convergence of linear finite element approximation to quasilinear initial boundary value problems*, RAIRO Anal. Numér., 12 (1978), pp. 247–266.
[17] M. Dobrowolski, *L^∞-convergence of linear finite element approximation to nonlinear parabolic problems*, SIAM J. Numer. Anal, 17 (1980), pp. 663–674.
[18] J. Douglas, Jr., and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.
[19] J. Douglas, Jr., T. Dupont, and M.F. Wheeler, *Some superconvergence results for an H^1-Galerkin procedure for the heat equation*, in Computing Methods in Applied Sciences and Engineering (Proc. Internat. Sympos., Versailles, 1973), Part 1, Lecture Notes in Comput. Sci. 10, Spinger-Verlag, Berlin, New York, 1974, pp. 288–311.
[20] K. Eriksson, C. Johnson, and V. Thomee, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Model. Math. Anal. Numér., 19 (1985), pp. 611–643.
[21] L.C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.
[22] G. Fairweather, *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, New York, 1978.
[23] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
[24] M.D. Gunzburger and S.L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[25] C. Johnson, S. Larsson, V. Thomee, and L.B. Wahlbin, *Error estimates for spatially discrete approximations of semilinear parabolic equations with nonsmooth initial data*, Math. Comp., 49 (1987), pp. 331–357.
[26] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, 1972.
[27] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.
[28] J. Mélek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman & Hall, London, UK, 1996.
[29] J.A. Nitsche, *L^∞-convergence of finite element approximations for parabolic problems*, RAIRO Anal. Numér, 13 (1979), pp. 31–54.
[30] J.A. Nitsche and M. Wheeler, *L^∞-convergence of the finite element Galerkin operator for parabolic problems*, Numer. Funct. Anal. Optim., 4 (1981-82), pp. 325–353.
[31] P.H. Sammon, *Convergence estimates for semidiscrete parabolic equations approximations*, SIAM J. Numer. Anal., 19 (1982), pp. 68–92.
[32] P.H. Sammon, *Fully discrete approximation methods for parabolic problems with nonsmooth initial data*, SIAM J. Numer. Anal., 20 (1983), pp. 437–470.
[33] A.H. Schatz, V. Thomee, and L.B. Wahlbin, *Maximum norm stability, and error estimates in parabolic finite element equations*, Comm. Pure Appl. Math., 33 (1980), pp. 265–304.
[34] A.H. Schatz, V. Thomee, and L.B. Wahlbin, *Stability, analyticity and almost best approximation in maximum-norm for parabolic finite element equations*, Comm. Pure Appl. Math., 51 (1998), pp. 1349–1385.
[35] R. Temam, *Navier-Stokes Equations—Theory and Numerical Analysis*, Elsevier Science, Amsterdam, 1984.
[36] V. Thomee, *Some convergence results for Galerkin methods for parabolic boundary value problems*, in Mathematical Aspects of Finite Elements for Parabolic Boundary Value Problems, C. de Boor, ed., Academic Press, New York, 1974, pp. 55–88.
[37] V. Thomee, *Negative norm estimates and superconvergence in Galerkin methods for parabolic problems*, Math. Comp., 34 (1980), pp. 93–113.
[38] V. Thomee, *Galerkin Finite Element Methods for Parabolic Equations*, Springer-Verlag, Berlin, 1997.

[39] V. Thomée and L. Wahlbin, *On Galerkin methods in semilinear parabolic problems*, SIAM J. Numer. Anal., 12 (1975), pp. 378–389.

[40] V. Thomee and L. Wahlbin, *Maximum norm stability and error estimates in Galerkin methods for parabolic equations in one space variable*, Numer. Math., 41 (1983), pp. 345–371.

[41] L.B. Wahlbin, *On maximum norm error estimates for Galerkin approximations to one-dimensional second order parabolic boundary value problems*, SIAM J. Numer. Anal., 12 (1975), pp. 177–182

[42] M.F. Wheeler, *L_∞ estimates of optimal orders for Galerkin methods for one-dimensional second order parabolic and hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 908–913.

[43] M.F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

[44] M. Zlamal, *Finite element multistep discretizations of parabolic boundary value problems* Math. Comp., 29 (1975), pp. 350–359.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 307–318

THE DISCRETE FIRST-ORDER SYSTEM LEAST SQUARES: THE SECOND-ORDER ELLIPTIC BOUNDARY VALUE PROBLEM*

ZHIQIANG CAI† AND BYEONG CHUN SHIN‡

Abstract. In [Z. Cai, T. Manteuffel, and S. F. McCormick, *SIAM J. Numer. Anal.*, 34 (1997), pp. 425–454], an L^2-norm version of first-order system least squares (FOSLS) was developed for scalar second-order elliptic partial differential equations. A limitation of this approach is the requirement of sufficient smoothness of the original problem, which is used for the equivalence of spaces between $(H^1)^d$ and $H(\text{div}\,) \cap H(\text{curl})$-type, where $d = 2$ or 3 is the dimension. By directly approximating $H(\text{div}\,) \cap H(\text{curl})$-type space based on the Helmholtz decomposition, this paper develops a discrete FOSLS approach in two dimensions. Under general assumptions, we establish error estimates in the L^2 and H^1 norms for the vector and scalar variables, respectively. Such error estimates are optimal with respect to the required regularity of the solution. A preconditioner for the algebraic system arising from this approach is also considered.

Key words. least-squares discretization, multigrid, preconditioner, second-order elliptic problems

AMS subject classifications. 65F10, 65F30

PII. S0036142900381886

1. Introduction.

Recently, there has been substantial interest in the use of least-squares principles for numerical approximations of elliptic partial differential equations and systems (see the recent review article [1] and references therein). In [5], Cai, Manteuffel, and McCormick developed an L^2-norm version of first-order system least squares (FOSLS) for scalar second-order elliptic partial differential equations in $d = 2$ or 3 dimensions. It was shown that the homogeneous FOSLS functional is equivalent to a $\mathcal{V} \times H^1(\Omega)$ norm with $\mathcal{V} = H(\text{div}\,;\Omega) \cap H(\text{curl}\,A;\Omega)$ under general assumptions, where A is the diffusion coefficient and Ω is the domain of the underlying problem. Moreover, such a norm was shown to be in fact an $H^1(\Omega)^{d+1}$ norm under the assumption that the original problem is H^2-regular. This product H^1 equivalence means that the minimization process amounts to solving a loosely coupled system of Poisson-like scalar equations. This in turn implies that standard finite element discretization and standard multigrid solution methods admit optimal H^1-like performance.

The limitation of this L^2-norm FOSLS is the requirement of sufficient smoothness of the underlying problem. Such smoothness guarantees the equivalence of norms between $\mathcal{V}$ and $H^1(\Omega)^d$ so that it can be approximated by standard continuous finite element space as in [5]. In general, when the domain Ω is not smooth or not convex or the coefficient A is not continuous, these two spaces are not equivalent. In fact, $\mathcal{V}$ is equal to $H^1(\Omega)^d$ plus a finite-dimensional space which consists of singular functions associated with corners of the boundary and interfaces. Therefore, standard continuous finite element spaces are not good approximations to $\mathcal{V}$ in general. In this paper,

*Received by the editors December 3, 2000; accepted for publication (in revised form) November 6, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/38188.html

†Department of Mathematics, Purdue University, 1395 Mathematical Sciences Building, West Lafayette, IN 47907-1395 (zcai@math.purdue.edu). This research was supported in part by the National Science Foundation.

‡Department of Mathematics, Seoul National University, Seoul 151-747, Korea (bcshin@math.snu.ac.kr). This research was supported in part by the Korea Science and Engineering Foundation.

we will construct an appropriate approximation space for $\mathcal{V}$ based on the Helmholtz decomposition. Since our approximation space is discontinuous and is not contained in $\mathcal{V}$, we then modify the FOSLS functional to accommodate such discontinuity and nonconformity of finite element spaces. An alternative for overcoming such a limitation is the inverse-norm version of FOSLS (see [2]), but at the expense of rather awkward norm evaluation requirements.

The paper is organized as follows. The second-order elliptic boundary value problem and the L^2-norm version of the FOSLS approach are introduced in section 2, along with some notations. The discrete FOSLS approach is developed in section 3, and its error estimate is established in section 4. In section 5, we discuss preconditioners for the resulting system of linear equations.

2. First-order system least squares (FOSLS). Let Ω be a bounded, open, and simply connected domain in $\Re^2$ with Lipschitz boundary $\partial\Omega$. We consider the following scalar second-order elliptic boundary value problem:

$$\left\{\begin{array}{rcll} -\nabla\cdot(A\nabla p)+\mathbf{b}\cdot\nabla p+cp &=& f & \text{in } \Omega,\\ p &=& 0 & \text{on } \Gamma_D,\\ \mathbf{n}\cdot(A\nabla p) &=& 0 & \text{on } \Gamma_N, \end{array}\right. \tag{2.1}$$

where the symbols $\nabla\cdot$ and ∇ stand for the divergence and gradient operators, respectively; A is a 2×2 symmetric matrix of functions in $L^\infty(\Omega)$; $\mathbf{b}$ and c are the respective vector and scalar of functions in $L^\infty(\Omega)$; $f\in L^2(\Omega)$ is a given scalar function; $\partial\Omega=\Gamma_D\cup\Gamma_N$ is the partition of the boundary of Ω; and $\mathbf{n}$ is the outward unit vector normal to the boundary. For simplicity, assume that both Γ_D and Γ_N are nonempty, with the obvious generalization to quotient spaces when one of them is empty in the subsequent sections. We assume that A is uniformly symmetric positive definite and scaled appropriately; that is, there exist positive constants

$$0<\lambda\le 1\le\Lambda$$

such that

$$\lambda\boldsymbol{\xi}^T\boldsymbol{\xi}\le\boldsymbol{\xi}^T A\boldsymbol{\xi}\le\Lambda\boldsymbol{\xi}^T\boldsymbol{\xi} \tag{2.2}$$

for all $\boldsymbol{\xi}\in\Re^2$ and almost all $x\in\bar{\Omega}$.

We use standard notation and definitions for the Sobolev spaces $H^s(\Omega)^2$, associated inner products $(\cdot,\cdot)_s$, and respective norms $\|\cdot\|_s$, $s\ge 0$. (We suppress the designation Ω on the inner products and norms because dependence on region is clear by context.) $H^0(\Omega)^2$ coincides with $L^2(\Omega)^2$, in which case the norm and inner product will be denoted by $\|\cdot\|$ and $(\cdot,\cdot)$, respectively. Define subspaces of $H^1(\Omega)$:

$$H^1_D(\Omega)=\{q\in H^1(\Omega):\ q=0 \text{ on } \Gamma_D\} \text{ and } H^1_N(\Omega)=\{q\in H^1(\Omega):\ q=0 \text{ on } \Gamma_N\}.$$

Let $H_D^{-1}(\Omega)$ denote the dual of $H^1_D(\Omega)$ with the norm defined by

$$\|\phi\|_{H_D^{-1}(\Omega)}=\sup_{0\ne\psi\in H^1_D(\Omega)}\frac{(\phi,\,\psi)}{\|\psi\|_1}.$$

Denote the curl operator in $\Re^2$ by

$$\nabla\times=(-\partial_2,\,\partial_1)$$

and its formal adjoint by

$$\nabla^{\perp} = \begin{pmatrix} \partial_2 \\ -\partial_1 \end{pmatrix}.$$

Let

$$H(\operatorname{div} A^{\frac{1}{2}};\Omega) = \{\mathbf{v} \in L^2(\Omega)^2 : \nabla \cdot (A^{\frac{1}{2}}\mathbf{v}) \in L^2(\Omega)\}$$

and

$$H(\operatorname{curl} A^{-\frac{1}{2}};\Omega) = \{\mathbf{v} \in L^2(\Omega)^2 : \nabla\times(A^{-\frac{1}{2}}\mathbf{v}) \in L^2(\Omega)\},$$

which are Hilbert spaces under norms

$$\|\mathbf{v}\|_{H(\operatorname{div} A^{\frac{1}{2}};\Omega)} = \left(\|\mathbf{v}\|^2 + \left\|\nabla \cdot \left(A^{\frac{1}{2}}\mathbf{v}\right)\right\|^2\right)^{\frac{1}{2}}$$

and

$$\|\mathbf{v}\|_{H(\operatorname{curl} A^{-\frac{1}{2}};\Omega)} = \left(\|\mathbf{v}\|^2 + \left\|\nabla\times \left(A^{-\frac{1}{2}}\mathbf{v}\right)\right\|^2\right)^{\frac{1}{2}},$$

respectively. When A is the identity matrix, we use the simpler notations $H(\operatorname{div};\Omega)$ and $H(\operatorname{curl};\Omega)$. Define the subspaces

$$H_0(\operatorname{div} A^{\frac{1}{2}};\Omega) = \{\mathbf{v} \in H(\operatorname{div} A^{\frac{1}{2}};\Omega) : \mathbf{n} \cdot (A^{\frac{1}{2}}\mathbf{v}) = 0 \text{ on } \Gamma_N\},$$

$$H_0(\operatorname{curl} A^{-\frac{1}{2}};\Omega) = \{\mathbf{v} \in H(\operatorname{curl} A^{-\frac{1}{2}};\Omega) : \boldsymbol{\tau} \cdot (A^{-\frac{1}{2}}\mathbf{v}) = 0 \text{ on } \Gamma_D\},$$

and denote

$$\mathcal{U} = H_0(\operatorname{div} A^{\frac{1}{2}};\Omega) \cap H_0(\operatorname{curl} A^{-\frac{1}{2}};\Omega),$$

where $\boldsymbol{\tau}$ represents the unit vector tangent to the boundary oriented counterclockwise.

Introducing an independent vector variable

$$\mathbf{u} = A^{\frac{1}{2}}\,\nabla p,$$

by using the homogeneous Dirichlet boundary condition on Γ_D we have that

$$\nabla\times(A^{-\frac{1}{2}}\mathbf{u}) = 0 \quad \text{in } \Omega \quad \text{and} \quad \boldsymbol{\tau} \cdot (A^{-\frac{1}{2}}\mathbf{u}) = 0 \quad \text{on } \Gamma_D.$$

Then an equivalent extended system for problem (2.1) is

$$\left\{\begin{array}{rcll} \mathbf{u} - A^{\frac{1}{2}}\nabla p &=& \mathbf{0} & \text{in } \Omega, \\ -\nabla \cdot (A^{\frac{1}{2}}\mathbf{u}) + \mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{u}) + cp &=& f & \text{in } \Omega, \\ \nabla \times (A^{-\frac{1}{2}}\mathbf{u}) &=& 0 & \text{in } \Omega, \\ p &=& 0 & \text{on } \Gamma_D, \\ \mathbf{n} \cdot (A^{\frac{1}{2}}\mathbf{u}) &=& 0 & \text{on } \Gamma_N, \\ \boldsymbol{\tau} \cdot (A^{-\frac{1}{2}}\mathbf{u}) &=& 0 & \text{on } \Gamma_D. \end{array}\right. \tag{2.3}$$

Define the FOSLS functional as follows (see [5] or [6] for Poisson's equations):

$$G(\mathbf{v},q;f) = \|\mathbf{v} - A^{\frac{1}{2}}\nabla q\|^2 + \|f + \nabla \cdot (A^{\frac{1}{2}}\mathbf{v}) - \mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{v}) - cq\|^2 + \|\nabla \times (A^{-\frac{1}{2}}\mathbf{v})\|^2$$

for $(\mathbf{v},q) \in \mathcal{U} \times H^1_D(\Omega)$. Then the FOSLS variational problem for (2.1) is to minimize the quadratic functional $G(\mathbf{v},q;f)$ over $\mathcal{U} \times H^1_D(\Omega)$: find $(\mathbf{u},q) \in \mathcal{U} \times H^1_D(\Omega)$ such that

$$G(\mathbf{u},p;f) = \inf_{(\mathbf{v},q)\in\mathcal{U}\times H^1_D(\Omega)} G(\mathbf{v},q;f). \tag{2.4}$$

3. Discrete FOSLS. The least-squares approach defined in the previous section was proposed and analyzed in [5]. In particular, it was shown in [5] that the homogeneous functional is elliptic in the $H^1(\Omega)^3$ norm under certain H^2 regularity assumptions. This implies optimal H^1-like performance for standard finite element discretization and standard multigrid solution methods. An unfortunate limitation of this FOSLS approach is that this product H^1 equivalence generally requires sufficient smoothness of the original problem. Such a requirement is needed for the equivalence between the spaces $H^1(\Omega)^2$ and $\mathcal{U} = H_0(\operatorname{div} A^{\frac{1}{2}};\Omega) \cap H_0(\operatorname{curl} A^{-\frac{1}{2}};\Omega)$ and the quasi-optimality of finite element approximations in the H^1 norm for each variable. To overcome such a difficulty, we use discontinuous approximation spaces for the vector variable and modify this FOSLS functional to accommodate such a discontinuity of finite element spaces. Extension of the approach proposed in this section to the least-squares functional studied in [4, 8] is straightforward.

Discontinuous approximation spaces that we will employ are motivated by the following Helmholtz decomposition, for any $\mathbf{u} \in \mathcal{U}$:

$$\mathbf{u} = A^{\frac{1}{2}}\nabla s + A^{-\frac{1}{2}}\nabla^{\perp} t, \tag{3.1}$$

where $s \in H^1_D(\Omega)$ is the unique solution of

$$\left\{\begin{array}{rcccl} \nabla\cdot(A\nabla s) & = & \nabla\cdot(A^{\frac{1}{2}}\mathbf{u}) & \text{in} & \Omega, \\ s & = & 0 & \text{on} & \Gamma_D, \\ \mathbf{n}\cdot(A\nabla s) & = & 0 & \text{on} & \Gamma_N \end{array}\right.$$

and $t \in H^1_N(\Omega)$ is the unique solution of

$$\left\{\begin{array}{rcccl} \nabla\times(A^{-1}\nabla^{\perp} t) & = & \nabla\times(A^{-\frac{1}{2}}\mathbf{u}) & \text{in} & \Omega, \\ \boldsymbol{\tau}\cdot(A^{-1}\nabla^{\perp} t) & = & 0 & \text{on} & \Gamma_D, \\ t & = & 0 & \text{on} & \Gamma_N. \end{array}\right.$$

It is then natural to approximate the scalar functions $s \in H^1_D(\Omega)$ and $t \in H^1_N(\Omega)$ by standard continuous piecewise polynomials.

Let $\mathcal{T}_h$ be a partition of the domain Ω into finite elements; i.e., $\Omega = \cup_{K\in\mathcal{T}_h} K$ with $h = \max\{h_K = \operatorname{diam}(K) : K \in \mathcal{T}_h\}$. Assume that the triangulation $\mathcal{T}_h$ is regular (see [7]). Let $\mathcal{P}^h_{m-1}$ be a finite-dimensional space consisting of continuous piecewise polynomials of degree at most $m-1$ with respect to the triangulation $\mathcal{T}_h$. Denote standard finite element spaces by

$$\mathcal{S}^h_D = H^1_D(\Omega) \cap \mathcal{P}^h_{m-1} \quad \text{and} \quad \mathcal{S}^h_N = H^1_N(\Omega) \cap \mathcal{P}^h_{m-1}$$

and define the approximation space for the vector variable by

$$\mathcal{U}^h = \big(A^{\frac{1}{2}}\nabla\mathcal{S}^h_D\big) \oplus \big(A^{-\frac{1}{2}}\nabla^{\perp}\mathcal{S}^h_N\big).$$

It is an immediate consequence of the integration by parts and homogeneous boundary conditions that two subspaces $A^{\frac{1}{2}}\nabla\mathcal{S}^h_D$ and $A^{-\frac{1}{2}}\nabla^{\perp}\mathcal{S}^h_N$ are orthogonal with respect to the L^2 inner product. That is,

$$(A^{\frac{1}{2}}\nabla s,\, A^{-\frac{1}{2}}\nabla^{\perp} t) = 0 \tag{3.2}$$

for any $s \in \mathcal{S}^h_D$ and any $t \in \mathcal{S}^h_N$.

Note that $\mathcal{U}^h$ is not contained in $\mathcal{U}$ and, hence, the FOSLS functional $G(\cdot\,;\cdot)$ defined in the previous section is not well defined on $\mathcal{U}^h \times \mathcal{S}_D^h$. Therefore, we need to replace the divergence and curl operators in the $G(\cdot\,;\cdot)$ by the corresponding discrete operators. To this end, define the discrete divergence operator, $\nabla_h \cdot : L^2(\Omega)^2 \longrightarrow \mathcal{S}_D^h$, for given $\mathbf{v} \in L^2(\Omega)^2$ by $\phi = \nabla_h \cdot \mathbf{v} \in \mathcal{S}_D^h$ satisfying

$$(\phi,\, q) = -(\mathbf{v},\, \nabla q) \quad \forall\, q \in \mathcal{S}_D^h$$

and the discrete curl operator, $\nabla_h \times : L^2(\Omega)^2 \longrightarrow \mathcal{S}_N^h$, for given $\mathbf{v} \in L^2(\Omega)^2$ by $\psi = \nabla_h \times \mathbf{v} \in \mathcal{S}_N^h$ satisfying

$$(\psi,\, q) = (\mathbf{v}, \nabla^{\perp} q) \quad \forall\, q \in \mathcal{S}_N^h.$$

Finally, we denote Q_h the L^2-projection operator onto $\mathcal{S}_D^h$.

Now, we are ready to define the discrete FOSLS functional:

$$\begin{aligned} G_h(\mathbf{v}, q; f) = {} & \|\mathbf{v} - A^{\frac{1}{2}} \nabla q\|^2 + \|f + \nabla_h \cdot (A^{\frac{1}{2}} \mathbf{v}) - Q_h\big(\mathbf{b} \cdot (A^{-\frac{1}{2}} \mathbf{v})\big) - cq\|^2 \\ & + \|\nabla_h \times (A^{-\frac{1}{2}} \mathbf{v})\|^2 \end{aligned}$$

for $(\mathbf{v}, q) \in \mathcal{U}^h \times \mathcal{S}_D^h$. Our discrete FOSLS finite element approximation for (2.1) is then to minimize the quadratic functional $G_h(\mathbf{v}, q; f)$ over $\mathcal{U}^h \times \mathcal{S}_D^h$: find $(\mathbf{u}_h,\, p_h) \in \mathcal{U}^h \times \mathcal{S}_D^h$ such that

$$G_h(\mathbf{u}_h,\, p_h; f) = \inf_{(\mathbf{v},q) \in \mathcal{U}^h \times \mathcal{S}_D^h} G_h(\mathbf{v},\, q; f). \tag{3.3}$$

Denote the norm over $\mathcal{U}^h \times \mathcal{S}_D^h$ by

$$|||(\mathbf{v}, q)||| = \Big(\|q\|_1^2 + \|\mathbf{v}\|^2 + \|\nabla_h \cdot (A^{\frac{1}{2}} \mathbf{v})\|^2 + \|\nabla_h \times (A^{-\frac{1}{2}} \mathbf{v})\|^2 \Big)^{\frac{1}{2}}.$$

THEOREM 3.1. *The homogeneous functional $G_h(\cdot\,; 0)$ is uniformly elliptic and continuous in $\mathcal{U}^h \times \mathcal{S}_D^h$; i.e., for any $(\mathbf{v}, q) \in \mathcal{U}^h \times \mathcal{S}_D^h$, there exists a positive constant C such that*

$$\frac{1}{C} |||(\mathbf{v}, q)|||^2 \le G_h(\mathbf{v}, q; 0) \le C |||(\mathbf{v}, q)|||^2. \tag{3.4}$$

Proof. The upper bound in (3.4) is an immediate consequence of the triangle inequality and the boundedness of coefficients A, $\mathbf{b}$, c and the L^2-projection operator Q_h. To show the validity of the lower bound in (3.4), we first establish the following inequality: there exists a positive constant C such that

$$\frac{1}{C} |||(\mathbf{v}, q)|||^2 \le \tilde{G}(\mathbf{v}, q) \quad \forall\, (\mathbf{v}, q) \in \mathcal{U}^h \times \mathcal{S}_D^h, \tag{3.5}$$

where

$$\begin{aligned} \tilde{G}(\mathbf{v}, q) &= \|\mathbf{v} - A^{\frac{1}{2}} \nabla q\|^2 + \|\nabla_h \cdot (A^{\frac{1}{2}} \mathbf{v}) - Q_h(\mathbf{b} \cdot \nabla q) - cq\|^2 + \|\nabla_h \times (A^{-\frac{1}{2}} \mathbf{v})\|^2 \\ &= \left\| \begin{pmatrix} I & -A^{\frac{1}{2}} \nabla \\ \nabla_h \cdot A^{\frac{1}{2}} & -Q_h \mathbf{b} \cdot \nabla - cI \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\|^2 + \|\nabla_h \times (A^{-\frac{1}{2}} \mathbf{v})\|^2. \end{aligned}$$

Then the lower bound in (3.4) follows from the fact that

$$G_h(\mathbf{v},q;0) = \left\| \begin{pmatrix} I & 0 \\ -Q_h\mathbf{b}\cdot A^{-\frac{1}{2}} & I \end{pmatrix} \begin{pmatrix} I & -A^{\frac{1}{2}}\nabla \\ \nabla_h\cdot A^{\frac{1}{2}} & -Q_h\mathbf{b}\cdot\nabla - cI \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\|^2 + \|\nabla_h \times (A^{-\frac{1}{2}}\mathbf{v})\|^2$$

and that the largest and smallest singular values of the transformation matrix

$$\begin{pmatrix} I & 0 \\ -Q_h\mathbf{b}\cdot A^{-\frac{1}{2}} & I \end{pmatrix}$$

are bounded.

To prove the validity of (3.5), let $Xq = Q_h(\mathbf{b}\cdot\nabla q) + cq$ for convenience. Since $\|Q_h\| = 1$, (2.2) and the triangle and Poincaré–Friedrichs inequalities yield

$$\|Xq\| \le C\|A^{\frac{1}{2}}\nabla q\|.$$

It now follows from the definition of the discrete divergence operator and the Cauchy–Schwarz inequality that

$$\begin{aligned} \|A^{\frac{1}{2}}\nabla q\|^2 &= (A^{\frac{1}{2}}\nabla q - \mathbf{v}, A^{\frac{1}{2}}\nabla q) + (A^{\frac{1}{2}}\mathbf{v}, \nabla q) \\ &= (A^{\frac{1}{2}}\nabla q - \mathbf{v}, A^{\frac{1}{2}}\nabla q) - \big(\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}), q\big) \\ &= (A^{\frac{1}{2}}\nabla q - \mathbf{v}, A^{\frac{1}{2}}\nabla q) - \big(\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}) - Xq, q\big) - (Xq, q) \\ &\le \|A^{\frac{1}{2}}\nabla q - \mathbf{v}\|\,\|A^{\frac{1}{2}}\nabla q\| + \|\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}) - Xq\|\,\|q\| + \|Xq\|\,\|q\|, \end{aligned}$$

which, together with the Poincaré–Friedrichs inequality, implies that

$$\|A^{\frac{1}{2}}\nabla q\| \le C\left(\|A^{\frac{1}{2}}\nabla q - \mathbf{v}\| + \|\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}) - Q_h(\mathbf{b}\cdot\nabla q) - cq\| + \|q\|\right). \tag{3.6}$$

The triangle and Poincaré–Friedrichs inequalities and (3.6) give that

$$|||(\mathbf{v},q)|||^2 \le C\left(\tilde{G}(\mathbf{v},q) + \|q\|^2\right).$$

Now, (3.5) is a consequence of the standard compactness argument. This completes the proof of the theorem. ∎

4. Error estimates. This section establishes error estimates in the L^2 norm for the vector variable and the H^1 norm for the scalar variable (see Theorem 4.1). Such error estimates are optimal with respect to the required regularity of the solution.

Let $(\mathbf{u}_h, p_h) \in \mathcal{U}^h \times \mathcal{S}_D^h$ be the solution of the discrete problem in (3.3). The corresponding variational form of (3.3) is to find $(\mathbf{u}_h, p_h) \in \mathcal{U}^h \times \mathcal{S}_D^h$ such that

$$b_h(\mathbf{u}_h, p_h; \mathbf{v}, q) = \big(f, -\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}) + Q_h(\mathbf{b}\cdot(A^{-\frac{1}{2}}\mathbf{v})) + cq\big) \quad \forall\,(\mathbf{v},q) \in \mathcal{U}^h \times \mathcal{S}_D^h, \tag{4.1}$$

where the bilinear form $b_h(\cdot,\cdot)$ is induced from the quadratic form $G_h(\cdot;0)$:

$$\begin{aligned} b_h(\mathbf{u}_h, p_h; \mathbf{v}, q) &= \big(\mathbf{u}_h - A^{\frac{1}{2}}\nabla p_h, \mathbf{v} - A^{\frac{1}{2}}\nabla q\big) + \big(\nabla_h\times(A^{-\frac{1}{2}}\mathbf{u}_h), \nabla_h\times(A^{-\frac{1}{2}}\mathbf{v})\big) \\ &+\big(\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u}_h) - Q_h(\mathbf{b}\cdot(A^{-\frac{1}{2}}\mathbf{u}_h)) - cp_h, \nabla_h\cdot(A^{\frac{1}{2}}\mathbf{v}) - Q_h(\mathbf{b}\cdot(A^{-\frac{1}{2}}\mathbf{v})) - cq\big). \end{aligned} \tag{4.2}$$

To deduce the error equation, we need the following lemma.

LEMMA 4.1. *Let $(\mathbf{u}, p) \in \mathcal{U} \times H_D^1(\Omega)$ be the solution of first-order system (2.3). Then it satisfies the following equations:*

$$\big(-\nabla_h \cdot (A^{\frac{1}{2}}\mathbf{u}) + Q_h(\mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{u})) + cp, q \big) = (f, q) \quad \forall\, q \in \mathcal{S}_D^h \tag{4.3}$$

and

$$\big(\nabla_h \times (A^{-\frac{1}{2}}\mathbf{u}), r\big) = 0 \quad \forall\, r \in \mathcal{S}_N^h. \tag{4.4}$$

Proof. It follows from the definitions of the discrete divergence and curl operators and the L^2-projection and integration by parts that, for any $q \in \mathcal{S}_D^h$,

$$\big(\nabla_h \cdot (A^{\frac{1}{2}}\mathbf{u}), q\big) = \big(\nabla \cdot (A^{\frac{1}{2}}\mathbf{u}), q\big) \quad \text{and} \quad \big(Q_h(\mathbf{b} \cdot (A^{\frac{1}{2}}\mathbf{u})), q\big) = \big(\mathbf{b} \cdot (A^{\frac{1}{2}}\mathbf{u}), q\big)$$

and that, for any $r \in \mathcal{S}_N^h$,

$$\big(\nabla_h \times (A^{-\frac{1}{2}}\mathbf{u}), r\big) = \big(\nabla \times (A^{-\frac{1}{2}}\mathbf{u}), r\big) = 0,$$

which, together with the second and third equations in (2.3), imply equalities (4.3) and (4.4). ∎

For any $(\mathbf{v}, q) \in \mathcal{U}^h \times \mathcal{S}_D^h$, by (4.3) and (4.4) it is easy to see that

$$\begin{aligned} b_h(\mathbf{u}, p; \mathbf{v}, q) &= \big(f, -\nabla_h \cdot (A^{\frac{1}{2}}\mathbf{v}) + Q_h(\mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{v})) + cq\big) \\ &\quad + \big(\nabla \cdot (A^{\frac{1}{2}}\mathbf{u}) - \nabla_h \cdot (A^{\frac{1}{2}}\mathbf{u}), cq\big) - \big((I - Q_h)(\mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{u})), cq\big). \end{aligned} \tag{4.5}$$

The difference of equations (4.5) and (4.2) gives the following error equation:

$$\begin{aligned} &b_h(\mathbf{u} - \mathbf{u}_h, p - p_h; \mathbf{v}, q) \\ &= \big(\nabla \cdot (A^{\frac{1}{2}}\mathbf{u}) - \nabla_h \cdot (A^{\frac{1}{2}}\mathbf{u}), cq\big) - \big((I - Q_h)(\mathbf{b} \cdot (A^{-\frac{1}{2}}\mathbf{u})), cq\big) \end{aligned} \tag{4.6}$$

for all $(\mathbf{v}, q) \in \mathcal{U}^h \times \mathcal{S}_D^h$.

LEMMA 4.2. *For any $q \in H^{\alpha-2}(\Omega)$ with $\alpha \geq 2$, let $m-1$, the degree of the finite element space defined in section 3, be the smallest integer greater than or equal to $\alpha - 1$; we then have that*

$$\|(I - Q_h)q\|_{H_D^{-1}(\Omega)} \leq C\, h^{\alpha-1} \|q\|_{\alpha-2}. \tag{4.7}$$

Proof. It follows from the definitions of the $H_D^{-1}(\Omega)$ norm and the L^2-projection, the Cauchy–Schwarz inequality, and the approximation property that

$$\begin{aligned} \|(I - Q_h)q\|_{H_D^{-1}(\Omega)} &= \sup_{r \in H_D^1(\Omega)} \frac{\big((I - Q_h)q, r\big)}{\|r\|_1} = \sup_{r \in H_D^1(\Omega)} \frac{\big((I - Q_h)q, (I - Q_h)r\big)}{\|r\|_1} \\ &\leq C\, h\, \|(I - Q_h)q\| \leq C\, h^{\alpha-1}\, \|q\|_{\alpha-2}. \end{aligned}$$

This completes the proof of the lemma. ∎

Now, we are ready to establish error estimates in the L^2 and H^1 norms for the vector and scalar variables, respectively, which are optimal with respect to the required regularity of the solution. Note that the norm for $\mathbf{u}$ in the error estimate in (4.8) is L^2 only but H^1 in [5]. This contributes to the less smoothness requirement of the original problem here than in [5].

THEOREM 4.1. *Assume that* $(\mathbf{u}, p)$ *is in* $H^{\alpha-1}(\Omega)^2 \times H^{\alpha}(\Omega)$ *with* $\alpha > 1$, *and let* $m-1$, *the degree of the finite element space defined in section* 3, *be the smallest integer greater than or equal to* $\alpha - 1$. *Then the following error estimate holds:*

$$\|\mathbf{u} - \mathbf{u}_h\| + \|p - p_h\|_1 \le C\, h^{\alpha-1}\big(\|p\|_\alpha + \|\mathbf{u}\|_{\alpha-1}\big). \tag{4.8}$$

Proof. Let p_I be an interpolant of p in $\mathcal{S}_D^h$; one then has

$$\|p_I - p\|_1 \le C\, h^{\alpha-1}\, \|p\|_\alpha. \tag{4.9}$$

To establish the error bound in (4.8), by the triangle inequality, it suffices to show that there exists a $\tilde{\mathbf{u}}_h \in \mathcal{U}^h$ such that

$$\|\mathbf{u} - \tilde{\mathbf{u}}_h\| \le C\, h^{\alpha-1}\, \|\mathbf{u}\|_{\alpha-1} \tag{4.10}$$

and that

$$\|\tilde{\mathbf{u}}_h - \mathbf{u}_h\| + \|p_I - p_h\|_1 \le C\, h^{\alpha-1}\, \big(\|p\|_\alpha + \|\mathbf{u}\|_{\alpha-1}\big). \tag{4.11}$$

Note that $\mathbf{u}$ has the decomposition of the form

$$\mathbf{u} = A^{\frac{1}{2}} \nabla s + A^{-\frac{1}{2}} \nabla^\perp t,$$

where $s \in H_D^1(\Omega)$ and $t \in H_N^1(\Omega)$ are the unique solutions of

$$\big(A \nabla s, \nabla q\big) = \big(A^{\frac{1}{2}} \mathbf{u}, \nabla q\big) \quad \forall\, q \in H_D^1(\Omega) \tag{4.12}$$

and

$$\big(A^{-1} \nabla^\perp t, \nabla^\perp r\big) = \big(A^{-\frac{1}{2}} \mathbf{u}, \nabla^\perp r\big) \quad \forall\, r \in H_N^1(\Omega), \tag{4.13}$$

respectively. Let $\tilde{s}_h \in \mathcal{S}_D^h$ and $\tilde{t}_h \in \mathcal{S}_N^h$ be the respective finite element approximations of s and t; i.e., they satisfy the following equations:

$$\big(A \nabla \tilde{s}_h, \nabla q\big) = \big(A^{\frac{1}{2}} \mathbf{u}, \nabla q\big) \quad \forall\, q \in \mathcal{S}_D^h \tag{4.14}$$

and

$$\big(A^{-1} \nabla^\perp \tilde{t}_h, \nabla^\perp r\big) = \big(A^{-\frac{1}{2}} \mathbf{u}, \nabla^\perp r\big) \quad \forall\, r \in \mathcal{S}_N^h, \tag{4.15}$$

respectively. Assume that Poisson equations (4.12) and (4.13) have the following regularity estimates:

$$\|s\|_\alpha \le C\, \|\nabla \cdot (A^{\frac{1}{2}} \mathbf{u})\|_{\alpha-2} \quad \text{and} \quad \|t\|_\alpha \le C\, \|\nabla \times (A^{-\frac{1}{2}} \mathbf{u})\|_{\alpha-2},$$

respectively. Then standard finite element error bounds give that

$$\|A^{\frac{1}{2}} \nabla (s - \tilde{s}_h)\| \le C\, h^{\alpha-1}\, \|s\|_\alpha \le C\, h^{\alpha-1}\, \|\nabla \cdot (A^{\frac{1}{2}} \mathbf{u})\|_{\alpha-2}$$

and

$$\|A^{-\frac{1}{2}} \nabla^\perp (t - \tilde{t}_h)\| \le C\, h^{\alpha-1}\, \|t\|_\alpha \le C\, h^{\alpha-1}\, \|\nabla \times (A^{-\frac{1}{2}} \mathbf{u})\|_{\alpha-2}.$$

Hence, choosing

$$\tilde{\mathbf{u}}_h = A^{\frac{1}{2}} \nabla \tilde{s}_h + A^{-\frac{1}{2}} \nabla^\perp \tilde{t}_h,$$

by orthogonality (3.2) we have that

$$\begin{aligned}\|\mathbf{u}-\tilde{\mathbf{u}}_h\| &= \left(\|A^{\frac{1}{2}}\nabla(s-\tilde{s}_h)\|^2+\|A^{-\frac{1}{2}}\nabla^{\perp}(t-\tilde{t}_h)\|^2\right)^{\frac{1}{2}}\\ &\le C\,h^{\alpha-1}\,\left(\|\nabla\cdot(A^{\frac{1}{2}}\mathbf{u})\|_{\alpha-2}+\|\nabla\times(A^{-\frac{1}{2}}\mathbf{u})\|_{\alpha-2}\right).\end{aligned}$$

This completes the proof of inequality (4.10).

To show the validity of inequality (4.11), by the definition of the discrete divergence and curl operators, notice first that (4.14) and (4.15) imply

$$\nabla_h\cdot(A\nabla\tilde{s}_h)=\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u})\quad\text{and}\quad\nabla_h\times(A^{-1}\nabla^{\perp}\tilde{t}_h)=\nabla_h\times(A^{-\frac{1}{2}}\mathbf{u}),$$

respectively. Since $\nabla_h\cdot\nabla^{\perp}\tilde{t}_h=\nabla_h\times\nabla\tilde{s}_h=0$, we then have that

$$\nabla_h\cdot(A^{\frac{1}{2}}\tilde{\mathbf{u}}_h)=\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u})\quad\text{and}\quad\nabla_h\times(A^{-\frac{1}{2}}\tilde{\mathbf{u}}_h)=\nabla_h\times(A^{-\frac{1}{2}}\mathbf{u}). \tag{4.16}$$

It follows from Theorem 3.1 and error equation (4.6) that

$$\begin{aligned}&\frac{1}{C}|||(\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h)|||^2\\ &\le G_h(\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h;0)=b_h(\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h;\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h)\\ &=b_h(\tilde{\mathbf{u}}_h-\mathbf{u},p_I-p;\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h)-\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u})-\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u}),c(p_I-p_h)\big)\\ &\quad+\big((I-Q_h)(\mathbf{b}\cdot(A^{-\frac{1}{2}}\mathbf{u})),c(p_I-p_h)\big).\end{aligned} \tag{4.17}$$

Now, we bound each term in the above inequality. First, by the definitions of the discrete divergence operator, the L^2-projection, and $H_D^{-1}(\Omega)$ norm, we have that

$$\begin{aligned}&\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u})-\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u}),c(p_I-p_h)\big)\\ &=\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u}),c(p_I-p_h)\big)-\big(\nabla_h\cdot(A^{\frac{1}{2}}\mathbf{u}),Q_hc(p_I-p_h)\big)\\ &=\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u}),c(p_I-p_h)\big)-\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u}),Q_hc(p_I-p_h)\big)\\ &=\big((I-Q_h)\nabla\cdot(A^{\frac{1}{2}}\mathbf{u}),c(p_I-p_h)\big)\\ &\le C\,\|(I-Q_h)\nabla\cdot(A^{\frac{1}{2}}\mathbf{u})\|_{H_D^{-1}(\Omega)}\|p_I-p_h\|_1.\end{aligned} \tag{4.18}$$

Second, it follows from the Cauchy–Schwarz and triangle inequalities, equalities in (4.16), and the boundedness of the L^2-projection and coefficients A, $\mathbf{b}$, and c that

$$\begin{aligned}&b_h(\tilde{\mathbf{u}}_h-\mathbf{u},p_I-p;\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h)\\ &\le\left(\|\tilde{\mathbf{u}}_h-\mathbf{u}\|+\|A^{\frac{1}{2}}\nabla(p_I-p)\|\right)\left(\|\tilde{\mathbf{u}}_h-\mathbf{u}_h\|+\|A^{\frac{1}{2}}\nabla(p_I-p_h)\|\right)\\ &\quad+\left(\|Q_h\big(\mathbf{b}\cdot A^{-\frac{1}{2}}(\tilde{\mathbf{u}}_h-\mathbf{u})\big)\|+\|c(p_I-p)\|\right)\cdot\\ &\qquad\left(\|\nabla_h\cdot(A^{\frac{1}{2}}(\tilde{\mathbf{u}}_h-\mathbf{u}_h))\|+\|Q_h\big(\mathbf{b}\cdot A^{-\frac{1}{2}}(\tilde{\mathbf{u}}_h-\mathbf{u}_h)\big)\|+\|c(p_I-p_h)\|\right)\\ &\le C\,\left(\|\tilde{\mathbf{u}}_h-\mathbf{u}\|+\|p_I-p\|_1\right)|||(\tilde{\mathbf{u}}_h-\mathbf{u}_h,p_I-p_h)|||.\end{aligned} \tag{4.19}$$

Substituting (4.18) and (4.19) into (4.17) implies that

$$\begin{aligned}\|\tilde{\mathbf{u}}_h-\mathbf{u}_h\|+\|p_I-p_h\|_1\le C\Big(&\|\tilde{\mathbf{u}}_h-\mathbf{u}\|+\|p_I-p\|_1+\|(I-Q_h)\big(\nabla\cdot(A^{\frac{1}{2}}\mathbf{u})\big)\|_{H_D^{-1}(\Omega)}\\ &+\|(I-Q_h)\big(\mathbf{b}\cdot(A^{-\frac{1}{2}}\mathbf{u})\big)\|_{H_D^{-1}(\Omega)}\Big).\end{aligned}$$

Now, (4.11) is an immediate consequence of (4.9), (4.10), and Lemma 4.2. This completes the proof of the theorem. ∎

5. Preconditioners. In this section, we discuss a spectrally equivalent preconditioner for the system of linear equations arising from the FOSLS discretization which is uniform in the mesh size.

The equivalence in Theorem 3.1 do not give us an immediate preconditioner since $|||(\mathbf{v},p)|||^2$ involves the (discrete) divergence and curl operators. Instead of working with $\mathbf{v}\in\mathcal{U}^h$, we explicitly make use of its representation:

$$\mathbf{v} = A^{\frac{1}{2}}\nabla s + A^{-\frac{1}{2}}\nabla^{\perp} t, \quad \text{where } s\in\mathcal{S}_D^h,\ t\in\mathcal{S}_N^h. \tag{5.1}$$

Now, $|||(\mathbf{v},p)|||^2 = |||(s,t,p)|||^2$ would be equivalent to some weighted Sobolev norm in terms of (s,t,q) which gives indications on how to construct preconditioners. To this end, by the definitions of the discrete divergence and curl operators, we first note that

$$\nabla_h\cdot(\nabla^{\perp}t) = 0 \quad \text{in } \Omega \quad \text{and} \quad \nabla_h\times(\nabla s) = 0 \quad \text{in } \Omega \tag{5.2}$$

for any $t\in\mathcal{S}_N^h$ and any $s\in\mathcal{S}_D^h$, respectively. We then introduce two discrete diffusion operators, $\Delta_{h,A}:\mathcal{S}_D^h\longrightarrow\mathcal{S}_D^h$ and $\hat{\Delta}_{h,A}:\mathcal{S}_N^h\longrightarrow\mathcal{S}_N^h$. For a given $s\in\mathcal{S}_D^h$, define $\Delta_{h,A}s\in\mathcal{S}_D^h$ to be the solution of

$$(\Delta_{h,A}s,\, q) = -(A\nabla s,\,\nabla q) \quad \forall\ q\in\mathcal{S}_D^h, \tag{5.3}$$

and for a given $t\in\mathcal{S}_D^h$, define $\hat{\Delta}_{h,A}t\in\mathcal{S}_N^h$ to be the solution of

$$(\hat{\Delta}_{h,A}t,\, q) = (A^{-1}\nabla^{\perp}t,\,\nabla^{\perp}q) \quad \forall\ q\in\mathcal{S}_N^h. \tag{5.4}$$

It is easy to see that

$$\Delta_{h,A} = \nabla_h\cdot A\nabla \quad \text{and} \quad \hat{\Delta}_{h,A} = \nabla_h\times A^{-1}\nabla^{\perp}.$$

By using (3.2), we then have that

$$|||(\mathbf{v},p)|||^2 = |||(s,t,p)|||^2 = \|p\|_1^2 + |||s|||^2 + |||t|||^2, \tag{5.5}$$

where

$$|||s|||^2 = \|s\|^2 + \|A^{\frac{1}{2}}\nabla s\|^2 + \|\Delta_{h,A}s\|^2 \quad \text{and} \quad |||t|||^2 = \|t\|^2 + \|A^{-\frac{1}{2}}\nabla^{\perp}t\|^2 + \|\hat{\Delta}_{h,A}t\|^2.$$

Before discussing the preconditioner based on $|||(s,t,q)|||^2$, we restate our discrete FOSLS approach in terms of functions (s,t,q). Our FOSLS functional is as follows:

$$\begin{aligned} G_h(s,t,q;f) =& \|A^{\frac{1}{2}}\nabla s + A^{-\frac{1}{2}}\nabla^{\perp}t - A^{\frac{1}{2}}\nabla q\|^2 \\ &+ \|f + \Delta_{h,A}s - Q_h\big(\mathbf{b}\cdot(\nabla s + A^{-1}\nabla^{\perp}t)\big) - cq\|^2 + \|\hat{\Delta}_{h,A}t\|^2, \end{aligned} \tag{5.6}$$

and the FOSLS minimization problem is to find $(\phi_h,\psi_h,p_h)\in\mathcal{S}_D^h\times\mathcal{S}_N^h\times\mathcal{S}_D^h$ such that

$$G_h(\phi_h,\psi_h,p_h;f) = \inf_{(s,t,q)\in\mathcal{S}_D^h\times\mathcal{S}_N^h\times\mathcal{S}_D^h} G_h(s,t,q;f) \tag{5.7}$$

with $\mathbf{u}_h = A^{\frac{1}{2}}\nabla\phi_h + A^{-\frac{1}{2}}\nabla^{\perp}\psi_h$. The corresponding variational problem is to find $(\phi_h,\psi_h,p_h)\in\mathcal{S}_D^h\times\mathcal{S}_N^h\times\mathcal{S}_D^h$ such that

$$b_h(\phi_h,\psi_h,p_h;s,t,q) = f_h(s,t,q) \quad \forall\ (s,t,q)\in\mathcal{S}_D^h\times\mathcal{S}_N^h\times\mathcal{S}_D^h, \tag{5.8}$$

where the bilinear and linear forms are given by

$$\begin{aligned}
&b_h(\phi_h,\psi_h,p_h;s,t,q)\\
&=\big(A^{\frac12}\nabla\phi_h+A^{-\frac12}\nabla^\perp\psi_h-A^{\frac12}\nabla p_h,\ A^{\frac12}\nabla s+A^{-\frac12}\nabla^\perp t-A^{\frac12}\nabla q\big)+\big(\hat\Delta_{h,A}\psi_h,\ \hat\Delta_{h,A}t\big)\\
&+\big(\Delta_{h,A}\phi_h-Q_h\big(\mathbf{b}\cdot(\nabla\phi_h+A^{-1}\nabla^\perp\psi_h)\big)-cp,\ \Delta_{h,A}s-Q_h\big(\mathbf{b}\cdot(\nabla s+A^{-1}\nabla^\perp t)\big)-cq\big)
\end{aligned}$$

and

$$f_h(s,t,q)=\big(f,\ -\Delta_{h,A}s+Q_h\big(\mathbf{b}\cdot(\nabla s+A^{-1}\nabla^\perp t)\big)+cq\big).$$

THEOREM 5.1. *For any $(s,t,q)\in\mathcal{S}_D^h\times\mathcal{S}_N^h\times\mathcal{S}_D^h$, there exists a positive constant C such that*

$$\frac1C\,|||(s,t,q)|||^2\le G_h(s,t,q;0)=b_h(s,t,q;s,t,q)\le C\,|||(s,t,q)|||^2. \tag{5.9}$$

Proof. It is a direct consequence of Theorem 3.1 and equality (5.5). ∎

Theorem 5.1 indicates that the quadratic form $b_h(s,t,q;s,t,q)$ can be preconditioned well by the diagonal quadratic form $|||(s,t,q)|||^2$ because they are spectrally equivalent uniformly in the mesh size (see (5.9)). We further replace these diagonal blocks of $|||(s,t,q)|||^2$ by some multigrid preconditioners. To this end, note first that $\|q\|_1^2$ is uniformly equivalent to

$$\|q\|^2+\|A^{\frac12}\nabla q\|^2=\big((I-\Delta_{h,A})q,\,q\big)$$

by using (2.2) and the definitions of the discrete divergence and diffusion operators. Similarly, $|||s|||^2$ and $|||t|||^2$ are uniformly equivalent to

$$\begin{aligned}
&\|s\|^2+2\|A^{\frac12}\nabla s\|^2+\|\Delta_{h,A}s\|^2=\big((I-\Delta_{h,A})^2s,\,s\big)\\
\text{and}\quad&\|t\|^2+2\|A^{-\frac12}\nabla^\perp s\|^2+\|\hat\Delta_{h,A}t\|^2=\big((I-\hat\Delta_{h,A})^2t,\,t\big),
\end{aligned}$$

respectively. Let P_1 be a preconditioner based on a symmetric multigrid V-cycle applied to the diffusion problem: find $v\in\mathcal{S}_D^h$ such that

$$(A\nabla v,\,\nabla\xi)+(v,\,\xi)=0\quad\forall\,\xi\in\mathcal{S}_D^h.$$

It is well known that P_1 is spectrally equivalent to $I-\Delta_{h,A}$ uniformly in the mesh size. Since the solution of

$$(I-\Delta_{h,A})^2s=g$$

for a given $g\in\mathcal{S}_D^h$ can be obtained successively by solving two discrete diffusion equations, i.e.,

$$(I-\Delta_{h,A})\hat s=g\quad\text{and}\quad(I-\Delta_{h,A})s=\hat s,$$

it is then natural to precondition $(I-\Delta_{h,A})^2$ by P_1^2. For further discussions and numerical experiments on P_1^2 as a preconditioner for $(I-\Delta_{h,A})^2$, see [3]. Similarly, we precondition $(I-\hat\Delta_{h,A})^2$ by P_2^2, where P_2 is a preconditioner based on a symmetric multigrid V-cycle applied to the diffusion problem: find $v\in\mathcal{S}_N^h$ such that

$$(A^{-1}\nabla^\perp v,\,\nabla^\perp\xi)+(v,\,\xi)=0\quad\forall\,\xi\in\mathcal{S}_N^h.$$

REFERENCES

[1] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[2] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order system*, Math. Comp., 66 (1997), pp. 935–955.

[3] J. H. Bramble and T. Sun, *A negative-norm least squares method for Reissner-Mindlin plates*, Math. Comp., 67 (1998), pp. 901–916.

[4] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal. 31 (1994), pp. 1785–1799.

[5] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[6] C. L. Chang, *Finite element approximation for grad-div type systems in the plane*, SIAM J. Numer. Anal., 29 (1992), pp. 452–461.

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.

[8] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 319–343

LOCAL DISCONTINUOUS GALERKIN METHODS FOR THE STOKES SYSTEM*

BERNARDO COCKBURN†, GUIDO KANSCHAT‡, DOMINIK SCHÖTZAU§, AND CHRISTOPH SCHWAB¶

Abstract. In this paper, we introduce and analyze local discontinuous Galerkin methods for the Stokes system. For a class of shape regular meshes with hanging nodes we derive a priori estimates for the L^2-norm of the errors in the velocities and the pressure. We show that *optimal*-order estimates are obtained when polynomials of degree k are used for each component of the velocity and polynomials of degree $k-1$ for the pressure, for any $k \geq 1$. We also consider the case in which *all* the unknowns are approximated with polynomials of degree k and show that, although the orders of convergence remain the same, the method is more efficient. Numerical experiments verifying these facts are displayed.

Key words. finite elements, discontinuous Galerkin methods, Stokes system

AMS subject classification. 65N30

PII. S0036142900380121

1. Introduction.

In this paper, we introduce and analyze local discontinuous Galerkin (LDG) methods for the Stokes system

$$\begin{aligned} -\Delta \boldsymbol{u} + \nabla p &= \boldsymbol{f} && \text{in } \Omega, \\ \nabla \cdot \boldsymbol{u} &= 0 && \text{in } \Omega, \\ \boldsymbol{u} &= \boldsymbol{g}_{\mathcal{D}} && \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded domain of $\mathbb{R}^d$ and the Dirichlet datum satisfies the usual compatibility condition $\int_{\partial\Omega} \boldsymbol{g}_{\mathcal{D}} \cdot \boldsymbol{n}\, ds = 0$, with $\boldsymbol{n}$ denoting the outward unit normal to $\partial\Omega$. We thus continue the study of LDG methods as applied to diffusion-dominated problems started by Castillo, Cockburn, Perugia, and Schötzau [8], who carried out the analysis of general LDG methods for the Laplacian on general triangulations, and by Cockburn, Kanschat, Perugia, and Schötzau [13], who obtained superconvergence results for Cartesian grids and a special LDG method. Our long-term goal is to study LDG methods for the *incompressible* Navier–Stokes equations; the analysis of the Stokes system is thus a necessary intermediate step.

*Received by the editors October 26, 2000; accepted for publication (in revised form) December 12, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sinum/40-1/38012.html

†School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn@math.umn.edu). The research of this author was supported in part by the National Science Foundation (grant DMS-9807491) and by the University of Minnesota Supercomputing Institute.

‡Institut für Angewandte Mathematik, Universität Heidelberg, INF 293/294, 69120 Heidelberg, Germany (guido.kanschat@na-net.ornl.gov). The research of this author was supported in part by the ARO DAAG55-98-1-0335, the University of Minnesota Supercomputing Institute, and the Deutsche Forschungsgemeinschaft (grant Ka 1304/1-1). It was carried out while the author was a Visiting Professor at the School of Mathematics, University of Minnesota.

§School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (schoetza@math.umn.edu). The research of this author was supported in part by the Swiss National Science Foundation (Schweizerischer Nationalfonds).

¶Seminar für Angewandte Mathematik (SAM), ETHZ, CH-8092 Zürich, Switzerland (schwab@sam.math.ethz.ch).

There are mainly two motivations for using LDG methods for the Navier–Stokes equations. The first one is that these methods can easily handle meshes with hanging nodes, elements of general shapes, and local spaces of different types; this makes them ideally suited for *hp*-adaptivity. The second one, of no less importance, is that with their carefully devised *numerical fluxes* inherited from the corresponding discontinuous Galerkin (DG) discretizations of nonlinear hyperbolic conservation laws—see the work by Cockburn and Shu [16, 17, 19], Cockburn, Hou, and Shu [12], and Cockburn, Lin, and Shu [15]—the LDG methods weakly enforce the conservation laws element-by-element and in a *conservative* way. This last property is highly appreciated by the practitioners of computational fluid dynamics, especially in situations where there are shocks, steep gradients, or boundary layers. In fact, it was for the convection-dominated *compressible* Navier–Stokes equations that the DG discretization techniques were applied for the first time by Bassi and Rebay in [4] with excellent results; the LDG method was then introduced by Cockburn and Shu in [18] as an extension of Bassi and Rebay's method to general convection-diffusion problems. To give the reader a flavor of the LDG methods proposed in this paper, we briefly compare them with other methods.

• **Interior penalty methods.** In the framework of the Stokes system, the main difficulty in obtaining numerical approximations is the enforcement of the incompressibility condition on the velocity. For continuous approximations of the velocity, it is well known that a pointwise enforcement could yield an overconstrained velocity and the only divergence-free function might turn out to be identically zero; this is the so-called *locking* phenomenon. However, in 1990, Baker, Jureidini, and Karakashian [2] showed how to enforce the incompressibility condition pointwise *inside* each element and still obtain optimal error estimates. They achieved this by using interior penalty (IP) methods, that is, methods that take the velocity approximation to be *discontinuous* and penalize the size of its discontinuity jumps across the element boundaries; see also the recent extension of this method to the incompressible Navier–Stokes equations by Karakashian and Katsaounis [28]. Arnold, Brezzi, Cockburn, and Marini [1] briefly review IP methods for purely elliptic problems and then relate and compare them to the LDG and other DG methods. A similar comparison can easily be developed for the Stokes system, but here we restrict ourselves to pointing out that, like the IP method of Baker, Jureidini, and Karakashian, the LDG methods use a discontinuous approximate velocity whose discontinuity jumps across the element boundaries are also penalized. However, unlike the IP method of Baker, Jureidini, and Karakashian, the LDG methods use discontinuous pressure approximations and (at least in this paper) do not try to impose the incompressibility condition pointwise inside the elements; instead, like in standard mixed methods, this condition is imposed weakly.

• **Standard mixed methods.** In his review of standard mixed methods for the Navier–Stokes equations, Fortin [21] points out that the use of discontinuous approximations for the pressure ensures a better conservation of mass in comparison with the use of continuous approximations and refers to the work of Pelletier, Fortin, and Camarero [30] for situations that illustrate this point. This is a property that these methods have in common with the LDG methods, not only because of the use of discontinuous approximations of the pressure, but also because the LDG methods ensure mass conservation. Indeed, to obtain the LDG methods, we first rewrite the Stokes system as the following collection of conservation laws:

$$\underline{\sigma} = \nabla \boldsymbol{u} \qquad \text{in } \Omega, \tag{1.2}$$

$$-\nabla \cdot \underline{\sigma} + \nabla p = \boldsymbol{f} \qquad \text{in } \Omega, \tag{1.3}$$

$$\nabla \cdot \boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{1.4}$$

$$\boldsymbol{u} = \boldsymbol{g}_{\mathcal{D}} \qquad \text{on } \partial\Omega. \tag{1.5}$$

Then we discretize them by using the DG technique, that is, element-by-element and in a conservative way; this is what ensures mass conservation. Note that to achieve this, we introduced the stress tensor $\underline{\sigma}$. This could be considered a disadvantage of the LDG methods with respect to the classical mixed methods, but this is not so because $\underline{\sigma}$ can be eliminated independently and in parallel on each grid cell, as we shall see.

Let us briefly digress to point out that the issue of the possible advantages of methods that, like the LDG methods, enforce the conservation laws locally and in a conservative way over finite element methods which cannot do that, and are typically based on continuous approximations, is the subject of an ongoing discussion which is far from being exhausted. Although it has been firmly established that this property is certainly desirable for convection-dominated problems, its possible advantages in other situations still remain to be thoroughly explored. About this very point, see the review of DG methods by Cockburn, Karniadakis, and Shu [14] and the paper by Hughes, Engel, Mazzei, and Larson [25] where a comparison of discontinuous and continuous Galerkin methods is carried out.

• **Stabilized mixed methods.** Finally, let us emphasize that for the LDG methods, the approximation spaces for the velocity and the pressure can be chosen almost arbitrarily; only a mild local condition has to be satisfied. This is so because the LDG methods can be considered to be *stabilized* mixed methods; for a review of stabilized mixed methods, see the article by Franca, Hughes, and Stenberg [22]. They are thus related to the Galerkin least squares (GLS) mixed methods introduced in 1986/1987 by Hughes, Franca, and Balestra [27] and Hughes and Franca [26], who used the jumps of the pressures across boundary elements and residuals inside the elements to render them stable. However, unlike these methods, LDG methods use discontinuous approximations to the velocity and employ stabilization terms which involve jumps across the element boundaries only. Variations of the LDG methods we study here could be easily constructed which are closely related to the "locally" stabilized methods introduced and numerically studied in 1989 by Silvester and Kechkar [31] and then analyzed in 1992 by Kechkar and Silvester [29]; however, this subject will not be considered in this paper. Finally, we must also point out that in the GLS methods, one has, in particular for velocities which are piecewise quadratic or of higher degree, and also for curvilinear mapped elements, to evaluate the GLS stabilization terms which are quite costly due to the appearance of, e.g., the Laplacian in the bilinear forms. The LDG methods achieve, as we prove here, the same stabilization effect but as a rule do this without recourse to domain integrals of second-order derivatives of finite element functions. Rather, only edge/face integrals of jumps are evaluated.

Now, let us briefly describe our results. We show that if we use polynomials of degree k to approximate the pressure p, the stresses $\underline{\sigma}$, and the velocity $\boldsymbol{u}$, the order of convergence of k is obtained for the L^2-norm of p and $\underline{\sigma}$, and of $k+1$ for the L^2-norm of the velocity. These orders of convergence are *sharp*, as they are observed in our numerical experiments. We also explore the situation in which polynomials of degree $k-1$ are used to approximate the pressure p and the stress tensor $\underline{\sigma}$. In this case, we prove that the above mentioned orders of convergence remain *invariant*; in

other words, in this case the error estimates are *optimal*. Our numerical experiments confirm this fact; moreover, they also show that this choice of approximating spaces gives rise to a method which is *less efficient* than the one obtained by using the same approximation spaces for all the variables. In Table 1.1, we summarize our theoretical results and compare them with the orders of convergence obtained for the IP method of Baker, Jureidini, and Karakashian [2] and the stabilized mixed methods of Hughes and Franca [26]. (See also Franca and Stenberg [23] for a unified error analysis.) Note that when the approximations are continuous, the jumps across elements are zero and the corresponding penalization term vanishes; we indicate this by writing "none."

TABLE 1.1
Theoretical orders of convergence for $k \geq 1$.

Method		Penalization of the jumps of velocity and pressure		$\|\boldsymbol{u} - \boldsymbol{u}_N\|_0$	$\|p - p_N\|_0$
LDG		$\mathcal{O}(h^{-1})$	$\mathcal{O}(h)$	$k+1$	k
IP	[2]	$\mathcal{O}(h^{-1})$	none	$k+1$	k
Stabilized mixed	[26, 23]	none	$\mathcal{O}(h)$	$k+1$	k

Finally, let us point out that the technique we use in our analysis is an extension of that used in [8] for the Laplacian. One of the contributions in this paper is that we make the technique work for local spaces that might be different for different unknowns. In fact, in *all* previous error analyses of LDG methods involving second-order operators (see [18, 7, 9, 11, 8, 13, 20]), the local spaces for both the auxiliary stresses and the main unknowns have been taken to be identical. The second contribution is that we show how to obtain the inf-sup condition, which is nonstandard given the discontinuous nature of our elements, in order to obtain error estimates for the pressure. Note that, unlike the analysis technique used by Hughes, Franca, and Balestra [27] and Hughes and Franca [26], who obtained error estimates of the pressure in certain mesh-dependent norms, we obtain an error of the pressure in the L^2-norm by using an inf-sup condition; in this respect, our technique is closer to that employed in 1991 by Franca and Stenberg [23].

The paper is organized as follows. In section 2, we introduce the method, show that it determines a unique approximate solution, and then state and discuss our main results. Finally, a brief overview of its proof is given which is then completed in full detail in section 3. Section 4 is devoted to numerical experiments devised to verify our theoretical results and to compare the effect that the use of different spaces has on the quality of the LDG approximate solution. We end in section 5 by describing extensions of our analysis and giving some concluding remarks.

2. The main results. In this section, we formulate the LDG method and show that it possesses a well-defined solution. We then state and discuss our main results, and finally, we present an abstract framework upon which our error analysis is based.

We assume throughout this section, in order to avoid unnecessary technicalities, that the exact solution $(\boldsymbol{u}, p)$ of (1.1) belongs at least to $H^2(\Omega)^d \times H^1(\Omega)$.

2.1. Definition of the LDG method. To define the LDG method, we consider the system of first-order conservation laws (1.2)–(1.5). We use the standard notation $(\nabla \boldsymbol{v})_{ij} = \partial_j v_i$ and $(\nabla \cdot \underline{\sigma})_i = \sum_{j=1}^d \partial_j \sigma_{ij}$. We also denote by $\boldsymbol{v} \otimes \boldsymbol{n}$ the matrix whose

ijth component is $v_i\, n_j$ and write

$$\underline{\sigma} : \underline{\tau} := \sum_{i,j=1}^{d} \sigma_{ij}\tau_{ij}, \qquad\qquad \boldsymbol{v} \cdot \underline{\sigma} \cdot \boldsymbol{n} := \sum_{i,j=1}^{d} v_i \sigma_{ij} n_j = \underline{\sigma} : (\boldsymbol{v} \otimes \boldsymbol{n}).$$

Multiplying (1.2), (1.3), and (1.4) by arbitrary, smooth test functions $\underline{\tau}$, $\boldsymbol{v}$, and q, respectively, and integrating by parts over an arbitrary subset K of the domain Ω, we obtain

$$\int_K \underline{\sigma} : \underline{\tau}\, d\boldsymbol{x} = -\int_K \boldsymbol{u} \cdot \nabla \cdot \underline{\tau}\, d\boldsymbol{x} + \int_{\partial K} \boldsymbol{u} \cdot \underline{\tau} \cdot \boldsymbol{n}_K\, ds, \tag{2.1}$$

$$\int_K \underline{\sigma} : \nabla \boldsymbol{v}\, d\boldsymbol{x} - \int_{\partial K} \underline{\sigma} : (\boldsymbol{v} \otimes \boldsymbol{n}_K) ds - \int_K p\, \nabla \cdot \boldsymbol{v}\, d\boldsymbol{x} + \int_{\partial K} p\, \boldsymbol{v} \cdot \boldsymbol{n}_K ds$$

$$= \int_K \boldsymbol{f} \cdot \boldsymbol{v}\, d\boldsymbol{x}, \tag{2.2}$$

$$-\int_K \boldsymbol{u} \cdot \nabla q\, d\boldsymbol{x} + \int_{\partial K} \boldsymbol{u} \cdot \boldsymbol{n}_K\, q\, ds = 0, \tag{2.3}$$

where $\boldsymbol{n}_K$ is the outward unit normal to ∂K. This is the weak form of the Stokes system that we shall use to define the LDG method. We enforce the above equations on each element K of a general triangulation $\mathcal{T}$ of Ω which can have hanging nodes and elements of various shapes. Thus, since the above equations are well defined for any functions $(\underline{\sigma}, \boldsymbol{u}, p)$ and $(\underline{\tau}, \boldsymbol{v}, q)$ in $\underline{\Sigma} \times \boldsymbol{V} \times Q$, where

$$\underline{\Sigma} := \{\underline{\sigma} \in L^2(\Omega)^{d^2} \ : \sigma_{ij}\big|_K \in H^1(K)\ \forall K \in \mathcal{T},\ 1 \le i,j \le d\},$$
$$\boldsymbol{V} := \{\boldsymbol{v} \in L^2(\Omega)^d \ : v_i\big|_K \in H^1(K)\ \forall K \in \mathcal{T},\ 1 \le i \le d\},$$
$$Q := \left\{ q \in L^2(\Omega) \ : \int_\Omega q\, d\boldsymbol{x} = 0,\ q\big|_K \in H^1(K)\ \forall K \in \mathcal{T} \right\},$$

we seek to approximate the exact solution $(\underline{\sigma}, \boldsymbol{u}, p)$ with functions $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N)$ in the finite element space $\underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N \subset \underline{\Sigma} \times \boldsymbol{V} \times Q$, where

$$\underline{\Sigma}_N := \{\underline{\sigma} \in L^2(\Omega)^{d^2} \ : \sigma_{ij}\big|_K \in \mathcal{S}(K)\ \forall K \in \mathcal{T},\ 1 \le i,j \le d\},$$
$$\boldsymbol{V}_N := \{\boldsymbol{v} \in L^2(\Omega)^d \ : v_i\big|_K \in \mathcal{V}(K)\ \forall K \in \mathcal{T},\ 1 \le i \le d\},$$
$$Q_N := \left\{ q \in L^2(\Omega) \ : \int_\Omega q\, d\boldsymbol{x} = 0,\ q\big|_K \in \mathcal{Q}(K)\ \forall K \in \mathcal{T} \right\},$$

and the *local* finite element spaces $\mathcal{S}(K)$, $\mathcal{V}(K)$, and $\mathcal{Q}(K)$ typically consist of polynomials.

The approximate solution $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N)$ is now defined by imposing that for all $K \in \mathcal{T}$, for all $(\underline{\tau}, \boldsymbol{v}, q) \in \mathcal{S}(K)^{d^2} \times \mathcal{V}(K)^d \times \mathcal{Q}(K)$,

$$\int_K \underline{\sigma}_N : \underline{\tau}\, d\boldsymbol{x} = -\int_K \boldsymbol{u}_N \cdot \nabla \cdot \underline{\tau}\, d\boldsymbol{x} + \int_{\partial K} \widehat{\boldsymbol{u}}_{N,\sigma} \cdot \underline{\tau} \cdot \boldsymbol{n}_K\, ds, \tag{2.4}$$

$$\int_K \underline{\sigma}_N : \nabla \boldsymbol{v}\, d\boldsymbol{x} - \int_{\partial K} \widehat{\underline{\sigma}}_N : (\boldsymbol{v} \otimes \boldsymbol{n}_K)\, ds - \int_K p_N\, \nabla \cdot \boldsymbol{v}\, d\boldsymbol{x} + \int_{\partial K} \widehat{p}_N\, \boldsymbol{v} \cdot \boldsymbol{n}_K\, ds$$

$$= \int_K \boldsymbol{f} \cdot \boldsymbol{v}\, d\boldsymbol{x}, \tag{2.5}$$

$$-\int_K \boldsymbol{u}_N \cdot \nabla q\, d\boldsymbol{x} + \int_{\partial K} \widehat{\boldsymbol{u}}_{N,p} \cdot \boldsymbol{n}_K\, q\, ds = 0. \tag{2.6}$$

Here, $\widehat{\boldsymbol{u}}_{N,\sigma}$, $\widehat{\underline{\sigma}}_N$, $\widehat{p}_N$, and $\widehat{\boldsymbol{u}}_{N,p}$ are the so-called *numerical fluxes*, which are discrete approximations to traces on the boundary of the elements. Note how the numerical fluxes $\widehat{\boldsymbol{u}}_{N,\sigma}$ and $\widehat{\boldsymbol{u}}_{N,p}$ arise naturally from the weak formulation; although both are approximations to the trace of the velocity $\boldsymbol{u}$, they are defined in very different ways since they are associated with different conservation laws.

To define these numerical fluxes, we need to introduce some notation associated with traces. Let K^+ and K^- be two adjacent elements of $\mathcal{T}$; let $\boldsymbol{x}$ be an arbitrary point of the set $e = \partial K^+ \cap \partial K^-$, which is assumed to have a nonzero $(d-1)$-dimensional measure; and let $\boldsymbol{n}^+$ and $\boldsymbol{n}^-$ be the corresponding outward unit normals at that point. Let $(\underline{\sigma}, \boldsymbol{u}, p)$ be a function smooth inside each element $K^\pm$ and let us denote by $(\underline{\sigma}^\pm, \boldsymbol{u}^\pm, p^\pm)$ the traces of $(\underline{\sigma}, \boldsymbol{u}, p)$ on e from the interior of $K^\pm$. Then we define the mean values $\{\!\{\cdot\}\!\}$ and jumps $[\![\cdot]\!]$ at $\boldsymbol{x} \in e$ as

$$\{\!\{p\}\!\} := (p^+ + p^-)/2, \qquad \{\!\{\boldsymbol{u}\}\!\} := (\boldsymbol{u}^+ + \boldsymbol{u}^-)/2, \qquad \{\!\{\underline{\sigma}\}\!\} := (\underline{\sigma}^+ + \underline{\sigma}^-)/2,$$
$$[\![p]\!] := p^+\,\boldsymbol{n}^+ + p^-\,\boldsymbol{n}^-, \qquad [\![\boldsymbol{u}]\!] := \boldsymbol{u}^+ \cdot \boldsymbol{n}^+ + \boldsymbol{u}^- \cdot \boldsymbol{n}^-, \qquad [\![\underline{\sigma}]\!] := \underline{\sigma}^+ \cdot \boldsymbol{n}^+ + \underline{\sigma}^- \cdot \boldsymbol{n}^-.$$

Note that the jumps $[\![p]\!]$ and $[\![\underline{\sigma}]\!]$ are both vectors whereas the jump $[\![\boldsymbol{u}]\!]$ is a scalar. We also need to define a jump of the velocity $\boldsymbol{u}$ which is a matrix, namely,

$$\underline{[\![\boldsymbol{u}]\!]} := \boldsymbol{u}^+ \otimes \boldsymbol{n}^+ + \boldsymbol{u}^- \otimes \boldsymbol{n}^-.$$

In components, we have $\underline{[\![\boldsymbol{u}]\!]}^2 = \sum_{i=1}^d (u_i^+ - u_i^-)^2$ and $(\boldsymbol{u}^\pm \otimes \boldsymbol{n}^\pm)^2 = \sum_{i=1}^d (u_i^\pm)^2$. Also, we remark that, since $[\![\boldsymbol{u}]\!] = \sum_{i=1}^d (u_i^+ - u_i^-)\, n_i^+$, we have $[\![\boldsymbol{u}]\!]^2 \le \underline{[\![\boldsymbol{u}]\!]}^2$, that is, the norm of the scalar-valued jump of the velocity can be controlled by the norm of the matrix-valued jump.

We are now ready to introduce the numerical fluxes. We begin by defining the numerical fluxes $\widehat{\underline{\sigma}}$ and $\widehat{\boldsymbol{u}}_\sigma$ associated with the Laplacian. We pick a *direct* extension of the choice of numerical fluxes for the Laplace operator considered in [8] and [13]. That is, on a face e inside the domain Ω, we take

$$\begin{bmatrix} \widehat{\underline{\sigma}} \\ \widehat{\boldsymbol{u}}_\sigma \end{bmatrix} := \begin{bmatrix} \{\!\{\underline{\sigma}\}\!\} \\ \{\!\{\boldsymbol{u}\}\!\} \end{bmatrix} - \begin{bmatrix} C_{11} \underline{[\![\boldsymbol{u}]\!]} + [\![\underline{\sigma}]\!] \otimes \boldsymbol{C}_{12} \\ -\underline{[\![\boldsymbol{u}]\!]} \cdot \boldsymbol{C}_{12} \end{bmatrix}, \tag{2.7}$$

and if e lies on the boundary, we take

$$\begin{bmatrix} \widehat{\underline{\sigma}} \\ \widehat{\boldsymbol{u}}_\sigma \end{bmatrix} := \begin{bmatrix} \underline{\sigma}^+ - C_{11}\,(\boldsymbol{u}^+ - \boldsymbol{g}_{\mathcal{D}}) \otimes \boldsymbol{n}^+ \\ \boldsymbol{g}_{\mathcal{D}} \end{bmatrix}. \tag{2.8}$$

The numerical fluxes associated with the incompressibility constraint, $\widehat{\boldsymbol{u}}_p$ and $\widehat{p}$, are defined by using an analogous recipe. If the face e is on the interior of Ω, we take

$$\begin{bmatrix} \widehat{\boldsymbol{u}}_p \\ \widehat{p} \end{bmatrix} := \begin{bmatrix} \{\!\{\boldsymbol{u}\}\!\} \\ \{\!\{p\}\!\} \end{bmatrix} + \begin{bmatrix} D_{11}\,[\![p]\!] + \boldsymbol{D}_{12}\,[\![\boldsymbol{u}]\!] \\ -\boldsymbol{D}_{12} \cdot [\![p]\!] \end{bmatrix}, \tag{2.9}$$

and if e lies on the boundary, we take

$$\begin{bmatrix} \widehat{\boldsymbol{u}}_p \\ \widehat{p} \end{bmatrix} := \begin{bmatrix} \boldsymbol{g}_{\mathcal{D}} \\ p^+ \end{bmatrix}. \tag{2.10}$$

The parameters C_{11}, $\boldsymbol{C}_{12}$ and D_{11}, $\boldsymbol{D}_{12}$ depend on $\boldsymbol{x} \in e$. This completes the definition of the LDG method for the Stokes system (1.1).

We would like to stress the following points about this method:

• Note that since the numerical flux $\widehat{\boldsymbol{u}}_\sigma$ is independent of the variable $\underline{\sigma}$, it is possible to use (2.4) to solve $\underline{\sigma}_N$ in terms of $\boldsymbol{u}_N$ only, element-by-element. This local solvability, which allows us to *eliminate* the stresses $\underline{\sigma}_N$ from the equations, gives the name to the LDG method. (See [8, 18] for more details.)

• The numerical fluxes are consistent in the sense that equations (2.4)–(2.6) coincide with (2.1)–(2.3) for the exact solution $(\underline{\sigma}, \boldsymbol{u}, p)$. Note also that the boundary condition is taken into consideration *only* through the numerical fluxes $\widehat{\boldsymbol{u}}_\sigma$ and $\widehat{\boldsymbol{u}}_p$ on the boundary.

• The purpose of the coefficients C_{11} and D_{11} is to ensure the stability of the method. They are thus referred to as the stabilization coefficients. As we shall see, they can also affect the accuracy of the method. The parameters $\boldsymbol{C}_{12}$ and $\boldsymbol{D}_{12}$ can be chosen so as to reduce the sparsity of the matrices and, in special cases, to enhance the accuracy of the method; see the case of the Laplacian treated in [13]. In this paper, we simply assume that they are of order one.

• Note that if we rewrite the conservation law (1.3) as

$$-\nabla \cdot (\underline{\sigma} - p\, \underline{I}) = \boldsymbol{f} \qquad \text{in } \Omega,$$

where $\underline{I}$ is the identity tensor, we see that we need to define a single numerical flux for $(\underline{\sigma} - p\ \underline{I})$ which, in fact, has been taken to be $\widehat{\underline{\sigma}} - \widehat{p}\ \underline{I}$. We could have taken the following more general ansatz for the numerical flux for the pressure $\widehat{p} = \{\!\{p\}\!\} - \boldsymbol{D}_{12} \cdot [\![p]\!] + D_{22}\, [\![\boldsymbol{u}]\!]$, but this would result in

$$\widehat{\underline{\sigma}} - \widehat{p}\, \underline{I} \;=\; \{\!\{\underline{\sigma}\}\!\} - \{\!\{p\}\!\}\, \underline{I} \;+\; \boldsymbol{D}_{12} \cdot [\![p]\!]\, \underline{I} - \Big(C_{11} \underline{[\![\boldsymbol{u}]\!]} + D_{22}\, [\![\boldsymbol{u}]\!]\, \underline{I} \Big).$$

Since, as we shall see, the role of the term $C_{11}\, [\![\boldsymbol{u}]\!]$ is to control *all* the discontinuity jumps of the velocity $\boldsymbol{u}$ but the term $D_{22}\, [\![\boldsymbol{u}]\!]\, \underline{I}$ can induce a control on the jumps of only the *normal* component of the velocity, it is clear that we can always take $D_{22} \equiv 0$.

2.2. The mixed setting. The study of the LDG method is greatly facilitated if we recast its formulation in a classical mixed finite element setting. To do that, we denote by $\mathcal{E}_i$ the *union* of all interior faces of the triangulation $\mathcal{T}$ and by $\mathcal{E}_\mathcal{D}$ the union of faces lying on $\partial\Omega$. By summing (2.4), (2.5), and (2.6) over all elements and after simple algebraic manipulations, the LDG method can be reformulated more compactly as follows. Find $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$ such that

$$\begin{aligned} a(\underline{\sigma}_N, \underline{\tau}) + b(\boldsymbol{u}_N, \underline{\tau}) &= f(\underline{\tau}), \\ -b(\boldsymbol{v}, \underline{\sigma}_N) + c(\boldsymbol{u}_N, \boldsymbol{v}) + d(\boldsymbol{v}, p_N) &= g(\boldsymbol{v}), \\ -d(\boldsymbol{u}_N, q) + e(p_N, q) &= h(q) \end{aligned} \tag{2.11}$$

for all $(\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$.

Here,

$$
\begin{aligned}
a(\underline{\sigma},\underline{\tau}) &:= \int_\Omega \underline{\sigma} : \underline{\tau}\, d\boldsymbol{x},\\
b(\boldsymbol{u},\underline{\tau}) &:= \sum_{K\in\mathcal{T}} \int_K \boldsymbol{u}\cdot\nabla\cdot\underline{\tau}\, d\boldsymbol{x} - \int_{\mathcal{E}_i} (\{\!\{\boldsymbol{u}\}\!\} + \underline{[\![\boldsymbol{u}]\!]}\cdot \boldsymbol{C}_{12})\cdot [\![\underline{\tau}]\!]\, ds,\\
c(\boldsymbol{u},\boldsymbol{v}) &:= \int_{\mathcal{E}_i} C_{11}\underline{[\![\boldsymbol{u}]\!]} : \underline{[\![\boldsymbol{v}]\!]}\, ds + \int_{\mathcal{E}_\mathcal{D}} C_{11}(\boldsymbol{u}\otimes\boldsymbol{n}) : (\boldsymbol{v}\otimes\boldsymbol{n})\, ds,\\
d(\boldsymbol{v},p) &:= -\sum_{K\in\mathcal{T}} \int_K p\nabla\cdot\boldsymbol{v}\, d\boldsymbol{x} + \int_{\mathcal{E}_i} (\{\!\{p\}\!\} - \boldsymbol{D}_{12}\cdot[\![p]\!])[\![\boldsymbol{v}]\!]\, ds + \int_{\mathcal{E}_\mathcal{D}} p\boldsymbol{v}\cdot\boldsymbol{n}\, ds,\\
e(p,q) &:= \int_{\mathcal{E}_i} D_{11}[\![p]\!]\cdot[\![q]\!]\, ds
\end{aligned}
$$

and

$$
\begin{aligned}
f(\underline{\tau}) &:= \int_{\mathcal{E}_\mathcal{D}} \boldsymbol{g}_\mathcal{D}\cdot\underline{\tau}\cdot\boldsymbol{n}\, ds,\\
g(\boldsymbol{v}) &:= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\, d\boldsymbol{x} + \int_{\mathcal{E}_\mathcal{D}} C_{11}(\boldsymbol{g}_\mathcal{D}\otimes\boldsymbol{n}) : (\boldsymbol{v}\otimes\boldsymbol{n})\, ds,\\
h(q) &:= -\int_{\mathcal{E}_\mathcal{D}} \boldsymbol{g}_\mathcal{D}\cdot\boldsymbol{n}\, q\, ds.
\end{aligned}
$$

Note that, by integration by parts, the forms b and d can also be expressed as

$$
\begin{aligned}
b(\boldsymbol{u},\underline{\tau}) &= -\sum_{K\in\mathcal{T}} \int_K \nabla\boldsymbol{u} : \underline{\tau}\, d\boldsymbol{x} + \int_{\mathcal{E}_i} (\{\!\{\underline{\tau}\}\!\} - [\![\underline{\tau}]\!]\otimes\boldsymbol{C}_{12}) : \underline{[\![\boldsymbol{u}]\!]}\, ds + \int_{\mathcal{E}_\mathcal{D}} \underline{\tau} : (\boldsymbol{u}\otimes\boldsymbol{n})\, ds,\\
d(\boldsymbol{v},p) &= \sum_{K\in\mathcal{T}} \int_K \boldsymbol{v}\cdot\nabla p\, d\boldsymbol{x} - \int_{\mathcal{E}_i} (\{\!\{\boldsymbol{v}\}\!\} + \boldsymbol{D}_{12}[\![\boldsymbol{v}]\!])\cdot[\![p]\!]\, ds.
\end{aligned}
$$

Finally, in order to analyze the method, we write the mixed system (2.11) in the following equivalent form: Find $(\underline{\sigma}_N,\boldsymbol{u}_N,p_N)\in\underline{\Sigma}_N\times\boldsymbol{V}_N\times Q_N$ such that

$$
\mathcal{A}(\underline{\sigma}_N,\boldsymbol{u}_N,p_N;\underline{\tau},\boldsymbol{v},q) = \mathcal{F}(\underline{\tau},\boldsymbol{v},q) \tag{2.12}
$$

for all $(\underline{\tau},\boldsymbol{v},q)\in\underline{\Sigma}_N\times\boldsymbol{V}_N\times Q_N$ by setting

$$
\begin{aligned}
\mathcal{A}(\underline{\sigma},\boldsymbol{u},p;\underline{\tau},\boldsymbol{v},q) &:= a(\underline{\sigma},\underline{\tau}) + b(\boldsymbol{u},\underline{\tau}) - b(\boldsymbol{v},\underline{\sigma}) + c(\boldsymbol{u},\boldsymbol{v}) + d(\boldsymbol{v},p) - d(\boldsymbol{u},q) + e(p,q),\\
\mathcal{F}(\underline{\tau},\boldsymbol{v},q) &:= f(\underline{\tau}) + g(\boldsymbol{v}) + h(q).
\end{aligned}
$$

2.3. Existence and uniqueness of LDG solutions. Next, we show that the LDG method defines a unique approximate solution provided that for each element $K\in\mathcal{T}$ the following mild conditions on the local spaces hold:

$$
u\in\mathcal{V}(K) : \int_K \nabla u\cdot\boldsymbol{v}\, d\boldsymbol{x} = 0 \quad \forall\boldsymbol{v}\in\mathcal{S}^d(K) \quad \text{implies} \quad \nabla u\equiv\boldsymbol{0} \text{ on } K, \tag{2.13}
$$

$$
q\in\mathcal{Q}(K) : \int_K \boldsymbol{v}\cdot\nabla q\, d\boldsymbol{x} = 0 \quad \forall\boldsymbol{v}\in\mathcal{V}^d(K) \quad \text{implies} \quad \nabla q\equiv\boldsymbol{0} \text{ on } K. \tag{2.14}
$$

See [8] for simple examples of local spaces not satisfying the above conditions.

PROPOSITION 2.1 (well-posedness of the LDG method). *Consider the* LDG *method defined by the weak formulation* (2.4)–(2.6) *and by the numerical fluxes given by* (2.7)–(2.10). *Suppose that the coefficients* C_{11} *and* D_{11} *are positive. Finally, assume that the conditions* (2.13) *and* (2.14) *on the local spaces are satisfied. Then the* LDG *method defines a unique approximate solution* $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$.

Proof. It is enough to show that the only possible solution to the system (2.11) with $\boldsymbol{f} = \boldsymbol{0}$ and $\boldsymbol{g}_{\mathcal{D}} = \boldsymbol{0}$ is $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N) = (\underline{0}, \boldsymbol{0}, 0)$. Indeed, taking $\underline{\tau} = \underline{\sigma}_N$, $\boldsymbol{v} = \boldsymbol{u}_N$, $q = p_N$ in (2.11) and adding the three equations yields

$$a(\underline{\sigma}_N, \underline{\sigma}_N) + c(\boldsymbol{u}_N, \boldsymbol{u}_N) + e(p_N, p_N) = 0,$$

which implies $\underline{\sigma}_N = \underline{0}$, $\underline{[\![\boldsymbol{u}_N]\!]} = \underline{0}$ on $\mathcal{E}_i$, $\boldsymbol{u}_N = \boldsymbol{0}$ on $\mathcal{E}_{\mathcal{D}}$, and $[\![p_N]\!] = 0$ on $\mathcal{E}_i$ since the coefficients C_{11} and D_{11} are positive. Consequently, the first equation in (2.11) reads as

$$\sum_{K \in \mathcal{T}} \int_K \nabla \boldsymbol{u}_N : \underline{\tau}\, d\boldsymbol{x} = 0 \qquad \forall \underline{\tau} \in \underline{\Sigma}_N.$$

Assumption (2.13) implies that $\nabla \boldsymbol{u}_N = \underline{0}$ on every $K \in \mathcal{T}$, and, since $\underline{[\![\boldsymbol{u}_N]\!]} = \underline{0}$ on $\mathcal{E}_i$ and $\boldsymbol{u}_N = \boldsymbol{0}$ on $\mathcal{E}_{\mathcal{D}}$, we must have $\boldsymbol{u}_N = \boldsymbol{0}$.

Taking $\underline{\sigma}_N = \underline{0}$ and $\boldsymbol{u}_N = \boldsymbol{0}$, the second equation in (2.11) becomes

$$\sum_{K \in \mathcal{T}} \int_K \boldsymbol{v} \cdot \nabla p_N \, d\boldsymbol{x} = 0 \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_N.$$

Analogously, we conclude from assumption (2.14) that $\nabla p_N = \boldsymbol{0}$ on every $K \in \mathcal{T}$ and, since $[\![p_N]\!] = 0$, that p_N is a constant. Since we also require that $\int_\Omega p_N \, d\boldsymbol{x} = 0$, we conclude that $p_N = 0$. ∎

2.4. A priori estimates. In this section we state and discuss our a priori error bounds for the LDG method. We assume that every element K of the triangulation $\mathcal{T}$ is *affinely equivalent* (see [10, section 2.3]) to one of several reference elements in an arbitrary but fixed set; this allows us to use elements of various shapes with possibly curved boundaries. For each $K \in \mathcal{T}$, we denote by h_K the diameter of K and by ρ_K the diameter of the biggest ball included in K; we set, as usual, $h := \max_{K \in \mathcal{T}} h_K$. The triangulations we consider can have hanging nodes but have to be *regular*; that is, there exists a positive constant σ_1 such that

$$\frac{h_K}{\rho_K} \leq \sigma_1 \qquad \forall\, K \in \mathcal{T} \tag{2.15}$$

(see [10, section 3.1]). Moreover, we let the maximum number of neighbors of a given element K be arbitrary but fixed. To formally state this property, we need to introduce the set $\langle K, K' \rangle$ defined as

$$\langle K, K' \rangle = \begin{cases} \emptyset & \text{if } \mathrm{meas}_{(d-1)}(\partial K \cap \partial K') = 0, \\ \text{interior of } \partial K \cap \partial K' & \text{otherwise.} \end{cases}$$

Thus, we assume that there exists a positive constant $\sigma_2 < 1$ such that, for each element $K \in \mathcal{T}$,

$$\sigma_2 \leq \frac{h_{K'}}{h_K} \leq \sigma_2^{-1} \qquad \forall K' : \langle K, K' \rangle \neq \emptyset\,. \tag{2.16}$$

These three hypotheses allow for quite general triangulations and are not restrictive in practice.

We assume that the local finite element spaces satisfy the following *inclusions* for $i = 1, \dots, d$:

$$\partial_i \mathcal{V}(K) \subseteq \mathcal{S}(K), \quad \partial_i \mathcal{S}(K) \subseteq \mathcal{V}(K), \quad \partial_i \mathcal{V}(K) \subseteq \mathcal{Q}(K), \quad \partial_i \mathcal{Q}(K) \subseteq \mathcal{V}(K). \tag{2.17}$$

Note that (2.17) also implies the assumptions (2.13) and (2.14) on the local spaces.

We denote by $P^\kappa(K)$ the set of all polynomials of degree at most κ on K and by $Q^\kappa(K)$ the polynomials of degree at most κ in each variable. Then, in order to guarantee certain approximation properties of the local spaces, we assume that they contain at least the following polynomial spaces:

$$P^k(K) \subseteq \mathcal{V}(K), \qquad P^l(K) \subseteq \mathcal{S}(K), \qquad P^m(K) \subseteq \mathcal{Q}(K), \tag{2.18}$$

with approximation orders $k \geq 1$ and $l, m \geq 0$. Since $\partial_i P^k(K) \subset P^{k-1}(K)$ and $\partial_i Q^k(K) \subset Q^k(K)$, conditions (2.17) and (2.18) are satisfied, for example, by

$$\mathcal{V}(K) = P^k(K), \qquad \mathcal{S}(K) = P^l(K), \qquad \mathcal{Q}(K) = P^m(K), \tag{2.19}$$

with $k \geq 1$, $l = k$ or $l = k - 1$, and $m = k$ or $m = k - 1$, or by

$$\mathcal{V}(K) = Q^k(K), \qquad \mathcal{S}(K) = Q^k(K), \qquad \mathcal{Q}(K) = Q^k(K), \qquad k \geq 1. \tag{2.20}$$

Next, we introduce a seminorm that appears in a natural way in the analysis of LDG methods. We denote by $H^s(D)$, D being a domain in $\mathbb{R}^d$, the Sobolev spaces of integer orders, and by $\|\cdot\|_{s,D}$ and $|\cdot|_{s,D}$ the usual norms and seminorms in $H^s(D)$, $H^s(D)^d$, and $H^s(D)^{d^2}$; we omit the dependence on the domain in the norms whenever $D = \Omega$. We define

$$|\!|\!| (\underline{\sigma}, \boldsymbol{u}, p) |\!|\!|_{\mathcal{A}}^2 := \|\underline{\sigma}\|_0^2 + \Theta^2(\boldsymbol{u}, p),$$

where

$$\Theta^2(\boldsymbol{u}, p) = \int_{\mathcal{E}_i} \left(C_{11} \underline{[\![\boldsymbol{u}]\!]}^2 + D_{11} [\![p]\!]^2 \right) ds + \int_{\mathcal{E}_\mathcal{D}} C_{11} (\boldsymbol{u} \otimes \boldsymbol{n})^2 \, ds.$$

We assume that the stabilization coefficients C_{11} and D_{11} defining the numerical fluxes in (2.7) and (2.9) are given by

$$C_{11}(\boldsymbol{x}) = \begin{cases} c_{11} \max\{h_{K^+}^{-1}, h_{K^-}^{-1}\} & \text{if } \boldsymbol{x} \in \langle K^+, K^- \rangle, \\ c_{11} h_{K^+}^{-1} & \text{if } \boldsymbol{x} \in \partial K^+ \cap \partial\Omega, \end{cases} \tag{2.21}$$

$$D_{11}(\boldsymbol{x}) = d_{11} \max\{h_{K^+}, h_{K^-}\}, \qquad \boldsymbol{x} \in \langle K^+, K^- \rangle, \tag{2.22}$$

with c_{11} and $d_{11} > 0$ independent of the meshsize and $|\boldsymbol{C}_{12}|$ as well as $|\boldsymbol{D}_{12}|$ of order one.

We are now ready to state our a priori error estimates for the LDG method. The first result is concerned with the error in the seminorm $|\!|\!| \cdot |\!|\!|_{\mathcal{A}}$ and the L^2-error in the pressure.

THEOREM 2.2. *Let* $(\underline{\sigma}, \boldsymbol{u}, p)$ *be the solution of* (1.2)–(1.5) *and let* $(\underline{\sigma}_N, \boldsymbol{u}_N, p_N)$ *be the approximate solution given by the* LDG *method* (2.4)–(2.6) *with numerical fluxes* (2.7)–(2.10). *Assume the hypotheses* (2.15), (2.16) *on the triangulations, the*

hypotheses (2.17), (2.18) *on the local spaces, with approximation orders* $k \geq 1$ *and* $l, m \geq 0$, *and the hypotheses* (2.21), (2.22) *on the form of the stabilization parameters. For* $\underline{\sigma} \in H^{l+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{k+1}(\Omega)^d$, *and* $p \in H^{m+1}(\Omega)$, *we have that the errors* $\underline{e}_\sigma = \underline{\sigma} - \underline{\sigma}_N$, $\boldsymbol{e}_u = \boldsymbol{u} - \boldsymbol{u}_N$, *and* $e_p = p - p_N$ *satisfy*

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} + \|e_p\|_0 \leq C\Big[h^{l+1}\|\underline{\sigma}\|_{l+1} + h^k\|\boldsymbol{u}\|_{k+1} + h^{m+1}\|p\|_{m+1}\Big],$$

where the constant C *solely depends on* Ω, σ_1, σ_2, c_{11}, d_{11}, d, *and the dimensions of the local spaces but is independent of the meshsize* h.

To prove a priori bounds for the L^2-error in $\boldsymbol{u}$, we assume elliptic regularity, that is, we assume that the solution $(\boldsymbol{z}, q)$ of the homogeneous Stokes problem

$$-\Delta \boldsymbol{z} + \nabla q = \boldsymbol{\lambda} \qquad \text{in } \Omega, \tag{2.23}$$

$$\nabla \cdot \boldsymbol{z} = 0 \qquad \text{in } \Omega, \tag{2.24}$$

$$\boldsymbol{z} = \boldsymbol{0} \qquad \text{on } \partial\Omega \tag{2.25}$$

with right-hand side $\boldsymbol{\lambda} \in L^2(\Omega)^d$ satisfies the estimate

$$\|\boldsymbol{z}\|_2 + \|q\|_1 \leq C\|\boldsymbol{\lambda}\|_0 \tag{2.26}$$

for a constant $C > 0$ just depending on Ω. For the inequality (2.26) to hold, certain restrictions on Ω are necessary; see, for example, Proposition 2.3 in Témam [32].

THEOREM 2.3. *Under the same assumptions as in Theorem* 2.2 *and the elliptic regularity assumption* (2.26), *we have that*

$$\|\boldsymbol{e}_u\|_0 \leq C\Big[h^{l+2}\|\underline{\sigma}\|_{l+1} + h^{k+1}\|\boldsymbol{u}\|_{k+1} + h^{m+2}\|p\|_{m+1}\Big]$$

with a constant C *that depends solely on* Ω, σ_1, σ_2, c_{11}, d_{11}, d, *the dimensions of the local spaces, and the constant in* (2.26) *but that is independent of the meshsize* h.

Let us briefly discuss the results of Theorems 2.2 and 2.3:

• When P^k- or Q^k-elements with $k \geq 1$ are used for all field variables, i.e, the local spaces are chosen as in (2.19) with $l = k$ and $m = k$ or as in (2.20), we obtain for smooth solutions $\underline{\sigma} \in H^{k+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{k+1}(\Omega)$, $p \in H^{k+1}(\Omega)$ the error bounds

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} + \|e_p\|_0 \leq Ch^k, \qquad \|\boldsymbol{e}_u\|_0 \leq Ch^{k+1}.$$

Although these rates are sharp in the sense that they are actually observed in the numerical experiments of section 4, they are not optimal in terms of the approximation properties of the finite element spaces.

• If the P-elements used for $\underline{\sigma}$ and p are of one order lower than the ones used for the velocities $\boldsymbol{u}$, i.e., if we consider P-elements as in (2.19) with $l = m = k - 1$ and $k \geq 1$, then an *optimal*-order error estimate is obtained: for $\underline{\sigma} \in H^k(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{k+1}(\Omega)^d$, and $p \in H^k(\Omega)$ we again have

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} + \|e_p\|_0 \leq Ch^k, \qquad \|\boldsymbol{e}_u\|_0 \leq Ch^{k+1},$$

which is optimal in terms of the approximation properties and of the regularity requirements of the exact solution.

• The stabilization parameters C_{11} and D_{11} are taken to be of order $\mathcal{O}(1/h)$ and $\mathcal{O}(h)$, respectively. As can be inferred from our analysis, this choice maximizes the

rates of convergence. For different selections of C_{11} and D_{11}, the obtained orders of convergence are summarized in Table 2.1 for smooth solutions and P^k- or Q^k-elements. On the other hand, our numerical results in Table 5.1 below indicate that for C_{11} and D_{11} of order one the parameters $\boldsymbol{C}_{12}$ and $\boldsymbol{D}_{12}$ can be chosen in such a way that the LDG method superconverges on Cartesian grids and for tensor product polynomials; we rigorously proved this result in [13] for the Laplacian. The extension of the study there to the Stokes system will be addressed in future work.

TABLE 2.1
Orders of convergence for P^k- or Q^k-elements in dependence of C_{11} and D_{11}.

C_{11}	D_{11}	$\lvert\!\lvert\!\lvert (\underline{e}_\sigma, \boldsymbol{e}_u, e_p) \rvert\!\rvert\!\rvert_{\mathcal{A}}$	$\lVert e_p \rVert_0$	$\lVert \boldsymbol{e}_u \rVert_0$
$\mathcal{O}(1), \mathcal{O}(1/h)$	$\mathcal{O}(1)$	k	k	$k+1/2$
$\mathcal{O}(1)$	$\mathcal{O}(h)$	k	k	$k+1/2$
$\mathcal{O}(1/h)$	$\mathcal{O}(h)$	k	k	$k+1$

2.5. The setting for the error analysis. The purpose of this section is to display as clearly as possible the main ingredients of the proof of our a priori results in section 2.4. To do so, we base our analysis on an abstract setting similar to the one introduced in [8] for the Laplacian.

We split the error $(\underline{e}_\sigma, \boldsymbol{e}_u, e_p) = (\underline{\sigma} - \underline{\sigma}_N, \boldsymbol{u} - \boldsymbol{u}_N, p - p_N)$ into the following sum:

$$(\underline{e}_\sigma, \boldsymbol{e}_u, e_p) = (\underline{\sigma} - \underline{\Pi}\underline{\sigma}, \boldsymbol{u} - \boldsymbol{\Pi}\boldsymbol{u}, p - \Pi p) + (\underline{\Pi}\underline{e}_\sigma, \boldsymbol{\Pi}\boldsymbol{e}_u, \Pi e_p),$$

where $\underline{\Pi} : \underline{\Sigma} \to \underline{\Sigma}_N$, $\boldsymbol{\Pi} : \boldsymbol{V} \to \boldsymbol{V}_N$, and $\Pi : Q \to Q_N$ are fixed projections onto the corresponding finite element spaces.

The basic ingredients. The basic ingredients of our error analysis are two. The first one is, as it is classical in finite element error analysis, the so-called Galerkin orthogonality property, namely,

$$\mathcal{A}(\underline{e}_\sigma, \boldsymbol{e}_u, e_p; \underline{\tau}, \boldsymbol{v}, q) = 0 \qquad \forall (\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N. \tag{2.27}$$

This property is a straightforward consequence of the consistency of the numerical fluxes and is valid since $(\boldsymbol{u}, p) \in H^2(\Omega)^d \times H^1(\Omega)$.

The second ingredient is a couple of inequalities that reflect the approximation properties of the projections $\underline{\Pi}$, $\boldsymbol{\Pi}$, and Π; namely, we assume that there exist error bounds $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ such that

$$\lvert\!\lvert\!\lvert \mathcal{A}(\underline{\sigma} - \underline{\Pi}\underline{\sigma}, \boldsymbol{u} - \boldsymbol{\Pi}\boldsymbol{u}, p - \Pi p; \underline{\tau} - \underline{\Pi}\underline{\tau}, \boldsymbol{v} - \boldsymbol{\Pi}\boldsymbol{v}, q - \Pi q) \rvert\!\rvert\!\rvert \le C K_{\mathcal{A}}(\underline{\sigma}, \boldsymbol{u}, p; \underline{\tau}, \boldsymbol{v}, q) \tag{2.28}$$

for any $(\underline{\sigma}, \boldsymbol{u}, p)$, $(\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma} \times \boldsymbol{V} \times Q$, and

$$\lvert\!\lvert\!\lvert \mathcal{A}(\underline{\sigma} - \underline{\Pi}\underline{\sigma}, \pm(\boldsymbol{u} - \boldsymbol{\Pi}\boldsymbol{u}), p - \Pi p; \underline{\tau}, \pm\boldsymbol{v}, q) \rvert\!\rvert\!\rvert \le C \lvert\!\lvert\!\lvert (\underline{\tau}, \boldsymbol{v}, q) \rvert\!\rvert\!\rvert_{\mathcal{A}} \, K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p) \tag{2.29}$$

for any $(\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$ and $(\underline{\sigma}, \boldsymbol{u}, p) \in \underline{\Sigma} \times \boldsymbol{V} \times Q$ and with constants C which are independent of the meshsize (specific forms for $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ shall be provided below). As we show next, all the error estimates we are interested in can be obtained in terms of $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$.

Error in the $\mathcal{A}$-seminorm. The error in $|\cdot|_{\mathcal{A}}$ can be estimated as follows.

LEMMA 2.4. *We have*

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} \le CK_{\mathcal{A}}^{1/2}(\underline{\sigma}, \boldsymbol{u}, p; \underline{\sigma}, \boldsymbol{u}, p) + CK_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p),$$

with C independent of the meshsize.

Proof. This is a straightforward extension of [8, Lemma 2.3]. We present the proof for the sake of completeness. Since $|\,(\cdot,\cdot,\cdot)\,|_{\mathcal{A}}$ is a seminorm, we have

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} \le |\,(\underline{\sigma} - \underline{\Pi\sigma}, \boldsymbol{u} - \boldsymbol{\Pi u}, p - \Pi p)\,|_{\mathcal{A}} + |\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}}.$$

By the definition of $\mathcal{A}$ in (2.12), by Galerkin orthogonality (2.27), and by assumption (2.29),

$$\begin{aligned}|\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}}^2 =& \mathcal{A}(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p; \underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\\ =& \mathcal{A}(\underline{\Pi\sigma} - \underline{\sigma}, \boldsymbol{\Pi u} - \boldsymbol{u}, \Pi p - p; \underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\\ \le& C|\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}} K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p),\end{aligned}$$

we have that

$$|\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}} \le CK_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p), \tag{2.30}$$

and so

$$|\,(\underline{e}_\sigma, \boldsymbol{e}_u, e_p)\,|_{\mathcal{A}} \le |\,(\underline{\sigma} - \underline{\Pi\sigma}, \boldsymbol{u} - \boldsymbol{\Pi u}, p - \Pi p)\,|_{\mathcal{A}} + CK_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p).$$

The estimate now follows from a simple application of assumption (2.28). This completes the proof. ∎

Error in the pressure. To obtain an error estimate in the pressure, we shall prove a stability result which allows us to measure the error of the pressure in the L^2-norm. It can be viewed as a discrete counterpart of the standard continuous inf–sup condition for the Stokes problem (see, e.g., [6, 24]), adapted to the discontinuous spaces considered here. Its proof is obtained by following the techniques used by Franca and Stenberg [23] in section 3.4 below.

PROPOSITION 2.5. *There exist positive constants κ_1 and κ_2 independent of the meshsize such that for all $(\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$ there is a $\boldsymbol{w} \in \boldsymbol{V}_N$ with*

$$\mathcal{A}(\underline{\tau}, \boldsymbol{v}, q; \underline{0}, \boldsymbol{w}, 0) \ge \kappa_1 \|q\|_0^2 - \kappa_2 |\,(\underline{\tau}, \boldsymbol{v}, q)\,|_{\mathcal{A}}^2, \quad |\,(\underline{0}, \boldsymbol{w}, 0)\,|_{\mathcal{A}} = \Theta(\boldsymbol{w}, 0) \le \|q\|_0. \tag{2.31}$$

Based on this inf-sup condition we obtain the following estimate for e_p.

LEMMA 2.6. *We have*

$$\|e_p\|_0 \le \|p - \Pi p\|_0 + CK_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p),$$

with C independent of the meshsize.

Proof. We only have to find an estimate for $\|\Pi e_p\|_0$, since we trivially have $\|p - p_N\|_0 \le \|p - \Pi p\|_0 + \|\Pi e_p\|_0$. To do that, we see that by Proposition 2.5 there exists a test function $\boldsymbol{w} \in \boldsymbol{V}_N$ such that (2.31) is satisfied for $(\underline{\tau}, \boldsymbol{v}, q) = (\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)$. By

(2.31), Galerkin orthogonality (2.27), assumption (2.29), estimate (2.30), the Cauchy–Schwarz inequality, and the properties of $\boldsymbol{w}$, we obtain

$$\begin{aligned}\kappa_1\|\Pi e_p\|_0^2 &\le \mathcal{A}(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p; \underline{0}, \boldsymbol{w}, 0) + \kappa_2 |\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}}^2 \\ &= \mathcal{A}(\underline{\Pi\sigma} - \underline{\sigma}, \boldsymbol{\Pi u} - \boldsymbol{u}, \Pi p - p; 0, \boldsymbol{w}, 0) + \kappa_2 |\,(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p)\,|_{\mathcal{A}}^2 \\ &\le C_1 |\,(\underline{0}, \boldsymbol{w}, 0)\,|_{\mathcal{A}} K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p) + C_2 K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p)^2 \\ &\le \frac{C_1}{2\varepsilon}\|\Pi e_p\|_0^2 + \Big(C_1\frac{\varepsilon}{2} + C_2\Big) K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p)^2\end{aligned}$$

for all $\varepsilon > 0$. We can now choose ε in such a way that

$$\|\Pi e_p\|_0 \le C K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p)$$

with a constant C depending on C_1 and C_2. The assertion follows. ∎

Error in the velocity. The estimate for the error $\|\boldsymbol{e}_u\|_0$ is based on a duality argument similar to the one used in [8].

LEMMA 2.7. *Assume that the elliptic regularity inequality* (2.26) *holds. Then we have*

$$\|\boldsymbol{e}_u\|_0 \le C \sup_{\boldsymbol{\lambda}\in L^2(\Omega)^d} \frac{K_{\mathcal{A}}(\underline{\sigma}, \boldsymbol{u}, p; \underline{\zeta}, \boldsymbol{z}, \widetilde{q})}{\|\boldsymbol{\lambda}\|_0} + C K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p) \sup_{\boldsymbol{\lambda}\in L^2(\Omega)^d} \frac{K_{\mathcal{B}}(\underline{\zeta}, \boldsymbol{z}, \widetilde{q})}{\|\boldsymbol{\lambda}\|_0}, \tag{2.32}$$

with $(\boldsymbol{z}, q)$ *denoting the solution of* (2.23)–(2.25) *with right-hand side* $\boldsymbol{\lambda}$ *and* $\underline{\zeta} = -\nabla\boldsymbol{z}$, $\widetilde{q} = -q$.

Proof. We introduce the linear functional $\Lambda(\boldsymbol{u}) = (\boldsymbol{\lambda}, \boldsymbol{u})$, where $(\cdot,\cdot)$ denotes the $L^2(\Omega)^d$-inner product. Then we have

$$\|\boldsymbol{e}_u\|_0 = \sup_{\boldsymbol{\lambda}\in L^2(\Omega)^d} \frac{\Lambda(\boldsymbol{e}_u)}{\|\boldsymbol{\lambda}\|_0}. \tag{2.33}$$

Now, let $(\boldsymbol{z}, q)$ be the solution of the adjoint equation (2.23)–(2.25) with right-hand side $\boldsymbol{\lambda}$. It is easy to verify that, if we set $\underline{\zeta} = -\nabla\boldsymbol{z}$, $\widetilde{q} = -q$, we have

$$\mathcal{A}(-\underline{\zeta}, \boldsymbol{z}, -\widetilde{q}; -\underline{\tau}, \boldsymbol{w}, -r) = \Lambda(\boldsymbol{w})$$

for all $(\underline{\tau}, \boldsymbol{w}, r) \in \underline{\Sigma} \times \boldsymbol{V} \times Q$. Taking $(\underline{\tau}, \boldsymbol{w}, r) = (\underline{e}_\sigma, \boldsymbol{e}_u, e_p)$, we get by the definition of $\mathcal{A}$ in (2.12) and by Galerkin orthogonality (2.27)

$$\begin{aligned}\Lambda(\boldsymbol{e}_u) =& \mathcal{A}(-\underline{\zeta}, \boldsymbol{z}, -\widetilde{q}; -\underline{e}_\sigma, \boldsymbol{e}_u, -e_p) \\ =& \mathcal{A}(\underline{e}_\sigma, \boldsymbol{e}_u, e_p; \underline{\zeta}, \boldsymbol{z}, \widetilde{q}) \\ =& \mathcal{A}(\underline{e}_\sigma, \boldsymbol{e}_u, e_p; \underline{\zeta} - \underline{\Pi\zeta}, \boldsymbol{z} - \boldsymbol{\Pi z}, \widetilde{q} - \Pi\widetilde{q}) \\ =& \mathcal{A}(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p; \underline{\zeta} - \underline{\Pi\zeta}, \boldsymbol{z} - \boldsymbol{\Pi z}, \widetilde{q} - \Pi\widetilde{q}) \\ &+ \mathcal{A}(\underline{\sigma} - \underline{\Pi\sigma}, \boldsymbol{u} - \boldsymbol{\Pi u}, p - \Pi p; \underline{\zeta} - \underline{\Pi\zeta}, \boldsymbol{z} - \boldsymbol{\Pi z}, \widetilde{q} - \Pi\widetilde{q}).\end{aligned}$$

We obtain with assumption (2.29) and estimate (2.30)

$$\begin{aligned}&|\,\mathcal{A}(\underline{\Pi e}_\sigma, \boldsymbol{\Pi e}_u, \Pi e_p; \underline{\zeta} - \underline{\Pi\zeta}, \boldsymbol{z} - \boldsymbol{\Pi z}, \widetilde{q} - \Pi\widetilde{q})| \\ &= |\,\mathcal{A}(\underline{\zeta} - \underline{\Pi\zeta}, -(\boldsymbol{z} - \boldsymbol{\Pi z}), \widetilde{q} - \Pi\widetilde{q}; \underline{\Pi e}_\sigma, -\boldsymbol{\Pi e}_u, \Pi e_p)\,| \le C K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p)\, K_{\mathcal{B}}(\underline{\zeta}, \boldsymbol{z}, \widetilde{q}),\end{aligned}$$

and hence

$$|\,\Lambda(\boldsymbol{e}_u)\,| \le CK_{\mathcal{B}}(\underline{\sigma},\boldsymbol{u},p)\,K_{\mathcal{B}}(\underline{\zeta},\boldsymbol{z},\widetilde{q}\,) + |\,\mathcal{A}(\underline{\sigma}-\underline{\Pi\sigma},\boldsymbol{u}-\boldsymbol{\Pi u},p-\Pi p;\underline{\zeta}-\underline{\Pi\zeta},\boldsymbol{z}-\boldsymbol{\Pi z},\widetilde{q}-\Pi\widetilde{q}\,)\,|.$$

The estimate now follows from a simple application of assumption (2.28) and from the characterization (2.33) of the L^2-norm. ∎

Conclusion. Thus, in order to prove our a priori estimates, all we need to do is to obtain the functionals $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ as well as the stability estimate in Proposition 2.5; this will be carried out in the next section. Then Theorems 2.2 and 2.3 will immediately follow after a simple application of Lemmas 2.4, 2.6, and 2.7.

3. Proofs. In this section, we prove our main results in the setting of section 2.5. We proceed as follows. After presenting some preliminary results, we obtain the functional $K_{\mathcal{A}}$ for general projection operators $\underline{\Pi}$, $\boldsymbol{\Pi}$, and Π. To obtain the functional $K_{\mathcal{B}}$, the projections $\underline{\Pi}$, $\boldsymbol{\Pi}$, and Π are chosen as L^2-projections.

3.1. Preliminaries. The following two lemmas contain all the information we actually use about our finite elements. The first one is a standard approximation result, valid for any linear continuous and polynomial preserving operator Π from $H^{s+1}(K)$ onto a finite-dimensional space $\mathcal{N}(K) \supset P^\kappa(K)$; it can be easily obtained by using the techniques of [10]. The second one is a standard inverse inequality.

LEMMA 3.1. *Let Π be a linear continuous operator from $H^{s+1}(K)$, $s \ge 0$, onto $\mathcal{N}(K) \supset P^\kappa(K)$ such that $\Pi w = w$ for all $w \in P^\kappa(K)$, $\kappa \ge 0$. Then we have*

$$|w-\Pi w|_{r,K} \le Ch_K^{\min(s,\kappa)+1-r}\|w\|_{s+1,K}, \qquad r=0,1,$$
$$\|w-\Pi w\|_{0,\partial K} \le Ch_K^{\min(s,\kappa)+\frac{1}{2}}\|w\|_{s+1,K}$$

for some constant C that solely depends on σ_1 in inequality (2.15)*, the dimension of $\mathcal{N}(K)$, d, and s.*

LEMMA 3.2. *There exists a positive constant C_{inv} that depends solely on σ_1 in inequality* (2.15)*, the dimension of $\mathcal{N}(K)$, and d such that for all $s \in \mathcal{N}(K)$ we have $\|s\|_{0,\partial K} \le C_{\mathrm{inv}}h_K^{-1/2}\|s\|_{0,K}$ for all $K \in \mathcal{T}$.*

Let $\underline{\Pi} : \underline{\Sigma} \to \underline{\Sigma}_N$, $\boldsymbol{\Pi} : \boldsymbol{V} \to \boldsymbol{V}_N$, and $\Pi : Q \to Q_N$ be projection operators onto the corresponding finite element spaces satisfying (componentwise) the assumptions in Lemma 3.1. We will make use of the following shorthand notation:

$$\underline{\xi}_\sigma = \underline{\sigma}-\underline{\Pi\sigma}, \qquad \boldsymbol{\xi}_u = \boldsymbol{u}-\boldsymbol{\Pi u}, \qquad \xi_p = p-\Pi p$$

for $(\underline{\sigma},\boldsymbol{u},p) \in \underline{\Sigma}\times\boldsymbol{V}\times Q$. We also define the quantities

$$\underline{C}_{11}^{\partial K} := \inf\{C_{11}(\boldsymbol{x}) : \boldsymbol{x} \in \partial K\}, \qquad \overline{C}_{11}^{\partial K} := \sup\{C_{11}(\boldsymbol{x}) : \boldsymbol{x} \in \partial K\},$$
$$\underline{D}_{11}^{\partial K} := \inf\{D_{11}(\boldsymbol{x}) : \boldsymbol{x} \in \partial K \setminus \partial\Omega\}, \qquad \overline{D}_{11}^{\partial K} := \sup\{D_{11}(\boldsymbol{x}) : \boldsymbol{x} \in \partial K \setminus \partial\Omega\}.$$

3.2. The functional $K_{\mathcal{A}}$. Using Cauchy–Schwarz's inequality, the approximation properties in Lemma 3.1 and the assumptions (2.15) and (2.16) on the meshes, we can prove, in exactly the same way as in [8, section 3.2], the following approximation results for the LDG forms.

LEMMA 3.3. *Assume* (2.15)*,* (2.16)*, and* (2.18)*. Let $\underline{\Pi}$, $\boldsymbol{\Pi}$, and Π be projection operators satisfying (componentwise with $\kappa = k$, $\kappa = l$, and $\kappa = m$, respectively) the*

assumptions in Lemma 3.1. *Let* $\underline{\sigma} \in H^{r+1}(\Omega)^{d^2}$, $\underline{\tau} \in H^{\overline{r}+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{s+1}(\Omega)^d$, $\boldsymbol{v} \in H^{\overline{s}+1}(\Omega)^d$, $p \in H^{t+1}(\Omega)$, *and* $q \in H^{\overline{t}+1}(\Omega)$ *for* r, $\overline{r}$, s, $\overline{s}$, t, $\overline{t} \geq 0$. *Then we have*

$$|a(\underline{\xi}_\sigma, \underline{\xi}_\tau)| \leq C \left(\sum_{K\in\mathcal{T}} h_K^{2\min(r,l)+2} \|\underline{\sigma}\|_{r+1,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} h_K^{2\min(\overline{r},l)+2} \|\underline{\tau}\|_{\overline{r}+1,K}^2 \right)^{\frac{1}{2}},$$

$$|b(\boldsymbol{\xi}_u, \underline{\xi}_\tau)| \leq C \left(\sum_{K\in\mathcal{T}} h_K^{2\min(s,k)} \|\boldsymbol{u}\|_{s+1,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} h_K^{2\min(\overline{r},l)+2} \|\underline{\tau}\|_{\overline{r}+1,K}^2 \right)^{\frac{1}{2}},$$

$$|c(\boldsymbol{\xi}_u, \boldsymbol{\xi}_v)| \leq C \left(\sum_{K\in\mathcal{T}} \overline{C}_{11}^{\partial K} h_K^{2\min(s,k)+1} \|\boldsymbol{u}\|_{s+1,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} \overline{C}_{11}^{\partial K} h_K^{2\min(\overline{s},k)+1} \|\boldsymbol{v}\|_{\overline{s}+1,K}^2 \right)^{\frac{1}{2}},$$

$$|d(\boldsymbol{\xi}_u, \xi_q)| \leq C \left(\sum_{K\in\mathcal{T}} h_K^{2\min(s,k)} \|\boldsymbol{u}\|_{s+1,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} h_K^{2\min(\overline{t},m)+2} \|q\|_{\overline{t}+1,K}^2 \right)^{\frac{1}{2}},$$

$$|e(\xi_p, \xi_q)| \leq C \left(\sum_{K\in\mathcal{T}} \overline{D}_{11}^{\partial K} h_K^{2\min(t,m)+1} \|p\|_{t+1,K}^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} \overline{D}_{11}^{\partial K} h_K^{2\min(\overline{t},m)+1} \|q\|_{\overline{t}+1,K}^2 \right)^{\frac{1}{2}},$$

with constants C *independent of the meshsize.*

For the special form of C_{11} and D_{11} proposed in (2.21) and (2.22), respectively, we have as a consequence of Lemma 3.3 and (2.16) the following result.

COROLLARY 3.4. *Under the same assumptions as in Lemma* 3.3 *and for coefficients* C_{11} *and* D_{11} *of the form* (2.21) *and* (2.22), *respectively, we have*

$$\begin{aligned}
|a(\underline{\xi}_\sigma, \underline{\xi}_\tau)| &\leq C h^{\min(r,l)+\min(\overline{r},l)+2} \|\underline{\sigma}\|_{r+1} \|\underline{\tau}\|_{\overline{r}+1},\\
|b(\boldsymbol{\xi}_u, \underline{\xi}_\tau)| &\leq C h^{\min(s,k)+\min(\overline{r},l)+1} \|\boldsymbol{u}\|_{s+1} \|\underline{\tau}\|_{\overline{r}+1},\\
|c(\boldsymbol{\xi}_u, \boldsymbol{\xi}_v)| &\leq c_{11} C h^{\min(s,k)+\min(\overline{s},k)} \|\boldsymbol{u}\|_{s+1} \|\boldsymbol{v}\|_{\overline{s}+1},\\
|d(\boldsymbol{\xi}_u, \xi_q)| &\leq C h^{\min(s,k)+\min(\overline{t},m)+1} \|\boldsymbol{u}\|_{s+1} \|q\|_{\overline{t}+1},\\
|e(\xi_p, \xi_q)| &\leq d_{11} C h^{\min(t,m)+\min(\overline{t},m)+2} \|p\|_{t+1} \|q\|_{\overline{t}+1},
\end{aligned}$$

with constants C *independent of the meshsize.*

From Corollary 3.4 we immediately obtain a general expression for the functional $K_{\mathcal{A}}$ since

$$\begin{aligned}
\mathcal{A}(\underline{\xi}_\sigma, \boldsymbol{\xi}_u, \xi_p; \underline{\xi}_\tau, \boldsymbol{\xi}_v, \xi_q) &= a(\underline{\xi}_\sigma, \underline{\xi}_\tau) + b(\boldsymbol{\xi}_u, \underline{\xi}_\tau) - b(\boldsymbol{\xi}_v, \underline{\xi}_\sigma)\\
&\quad + c(\boldsymbol{\xi}_u, \boldsymbol{\xi}_v) + d(\boldsymbol{\xi}_v, \xi_p) - d(\boldsymbol{\xi}_u, \xi_q) + e(\xi_p, \xi_q).
\end{aligned} \tag{3.1}$$

In the situations encountered in Lemmas 2.4 and 2.7 we obtain the following results.

COROLLARY 3.5. *Assume* (2.15), (2.16), *and* (2.18) *with approximation orders* $k \geq 1$, $l, m \geq 0$. *Assume the coefficients* C_{11} *and* D_{11} *to be of the form* (2.21) *and* (2.22), *respectively. Let* $\underline{\Pi}$, $\boldsymbol{\Pi}$, *and* Π *be projection operators as in Lemma* 3.3. *Let* $\underline{\sigma} \in H^{l+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{k+1}(\Omega)^d$, *and* $p \in H^{m+1}(\Omega)$. *Then we have in Lemma* 2.4

$$K_{\mathcal{A}}(\underline{\sigma}, \boldsymbol{u}, p; \underline{\sigma}, \boldsymbol{u}, p) \leq C \Big[h^{2l+2} \|\underline{\sigma}\|_{l+1}^2 + h^{2k} \|\boldsymbol{u}\|_{k+1}^2 + h^{2m+2} \|p\|_{m+1}^2 \Big].$$

Furthermore, assume the elliptic regularity inequality (2.26) *and let* $(\boldsymbol{z}, q)$ *denote the solution of* (2.23)–(2.25) *with right-hand side* $\boldsymbol{\lambda} \in L^2(\Omega)^d$, $\underline{\zeta} = -\nabla \boldsymbol{z}$, $\widetilde{q} = -q$. *Then*

we have in Lemma 2.7

$$K_{\mathcal{A}}(\underline{\sigma}, \boldsymbol{u}, p; \underline{\zeta}, \boldsymbol{z}, \widetilde{q}) \leq C\Big[h^{2+l}\|\underline{\sigma}\|_{l+1} + h^{1+k}\|\boldsymbol{u}\|_{k+1} + h^{2+m}\|p\|_{m+1}\Big]\|\boldsymbol{\lambda}\|_0.$$

Proof. The assertions follow immediately from Corollary 3.4, the identity (3.1), the choice of the coefficients C_{11} and D_{11}, and from the elliptic regularity estimate (2.26) which yields $\|\underline{\zeta}\|_1 + \|\boldsymbol{u}\|_2 + \|\widetilde{q}\|_1 \leq C\|\boldsymbol{\lambda}\|_0$. ∎

3.3. The functional $K_{\mathcal{B}}$. In this subsection we determine the functional $K_{\mathcal{B}}$ reflecting the approximation properties in (2.29). We start by investigating the forms a, c, and d. Lemma 3.1 and Cauchy–Schwarz's inequality immediately give the following estimates.

LEMMA 3.6. *Assume* (2.15), (2.16), *and* (2.18). *Let* $\underline{\Pi}$, $\boldsymbol{\Pi}$, *and* Π *be projection operators satisfying (componentwise with* $\kappa = k$, $\kappa = l$, *and* $\kappa = m$, *respectively) the assumptions in Lemma* 3.1. *Let* $\underline{\sigma} \in H^{r+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{s+1}(\Omega)^d$, *and* $p \in H^{t+1}(\Omega)$ *for* r, s, $t \geq 0$. *Then we have*

$$\begin{aligned}
|a(\underline{\xi}_\sigma, \underline{\tau})| &\leq C\left(\sum_{K\in\mathcal{T}} h_K^{2\min(r,l)+2}\|\underline{\sigma}\|_{r+1,K}^2\right)^{\frac{1}{2}} \|\underline{\tau}\|_0 && \forall\ \underline{\tau} \in \underline{\Sigma},\\
|c(\boldsymbol{\xi}_u, \boldsymbol{v})| &\leq C\left(\sum_{K\in\mathcal{T}} \overline{C}_{11}^{\partial K} h_K^{2\min(s,k)+1}\|\boldsymbol{u}\|_{s+1,K}^2\right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V},\\
|e(\xi_p, q)| &\leq C\left(\sum_{K\in\mathcal{T}} \overline{D}_{11}^{\partial K} h_K^{2\min(t,m)+1}\|p\|_{t+1,K}^2\right)^{\frac{1}{2}} \Theta(\boldsymbol{0}, q) && \forall q \in Q,
\end{aligned}$$

with constants C *independent of the meshsize.*

Next, we estimate the forms b and d in the case where $\underline{\Pi} : \underline{\Sigma} \to \underline{\Sigma}_N$, $\boldsymbol{\Pi} : \boldsymbol{V} \to \boldsymbol{V}_N$, and $\Pi : Q \to Q_N$ are chosen to be L^2-projections. Note that these projections clearly satisfy the assumptions of Lemma 3.1. It is also important to note that this is the *only* part of our analysis in which we actually use the inclusion properties (2.17).

LEMMA 3.7. *Assume* (2.15), (2.16) *and* (2.17), (2.18). *Let* $\underline{\Pi}$, $\boldsymbol{\Pi}$, *and* Π *be the (componentwise)* L^2*-projections onto the corresponding finite element spaces. Let* $\underline{\sigma} \in H^{r+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{s+1}(\Omega)^d$, *and* $p \in H^{t+1}(\Omega)$ *for* r, s, $t \geq 0$. *Then we have*

$$\begin{aligned}
|b(\boldsymbol{\xi}_u, \underline{\tau})| &\leq C\left(\sum_{K\in\mathcal{T}} h_K^{2\min(s,k)}\|\boldsymbol{u}\|_{s+1,K}^2\right)^{\frac{1}{2}} \|\underline{\tau}\|_0 && \forall\ \underline{\tau} \in \underline{\Sigma}_N,\\
|b(\boldsymbol{v}, \underline{\xi}_\sigma)| &\leq C\left(\sum_{K\in\mathcal{T}} \frac{1}{\underline{C}_{11}^{\partial K}} h_K^{2\min(r,l)+1}\|\underline{\sigma}\|_{r+1,K}^2\right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V}_N,\\
|d(\boldsymbol{\xi}_u, q)| &\leq C\left(\sum_{K\in\mathcal{T}} \frac{1}{\underline{D}_{11}^{\partial K}} h_K^{2\min(s,k)+1}\|\boldsymbol{u}\|_{s+1,K}^2\right)^{\frac{1}{2}} \Theta(\boldsymbol{0}, q) && \forall q \in Q_N,\\
|d(\boldsymbol{v}, \xi_p)| &\leq C\left(\sum_{K\in\mathcal{T}} \frac{1}{\underline{C}_{11}^{\partial K}} h_K^{2\min(t,m)+1}\|p\|_{t+1,K}^2\right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V}_N,
\end{aligned}$$

with constants C *independent of the meshsize.*

Proof. We start by proving the estimates for the form b. We note that $\int_K (\boldsymbol{u} - \boldsymbol{\Pi u}) \cdot \nabla \cdot \underline{\tau}\, d\boldsymbol{x} = 0$ due to the properties of the L^2-projection and the inclusion property $\partial_i \mathcal{S}(K) \subset \mathcal{V}(K)$ in (2.17). Therefore, using the fact that $\boldsymbol{C}_{12}$ is of order one, a repeated application of Cauchy–Schwarz's inequality gives

$$\begin{aligned}
|b(\boldsymbol{\xi}_u, \underline{\tau})| &= \left| \int_{\mathcal{E}_i} (\{\!\{\boldsymbol{\xi}_u\}\!\} + \underline{[\![\boldsymbol{\xi}_u]\!]} \cdot \boldsymbol{C}_{12}) \cdot [\![\underline{\tau}]\!]\, ds \right| \\
&\le C \left(\sum_{K \in \mathcal{T}} h_K^{-1} \|\boldsymbol{\xi}_u\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}} \widehat{h}_K \|\underline{\tau}\|_{0,\partial K}^2 \right)^{\frac{1}{2}},
\end{aligned}$$

where $\widehat{h}_K = \sup\{h_{K'} : \langle K, K' \rangle \neq \emptyset\}$. Assumption (2.16) implies that $\widehat{h}_K \le \sigma_2^{-1} h_K$, and therefore, the desired estimate follows from Lemma 3.1 and the inverse inequality in Lemma 3.2.

Furthermore, we also note that $\int_K \nabla \boldsymbol{v} : (\underline{\sigma} - \underline{\Pi\sigma})\, d\boldsymbol{x} = 0$, since $\underline{\Pi}|_K$ is the L^2-projection into $\mathcal{S}(K)$ and $\partial_i \mathcal{V}(K) \subset \mathcal{S}(K)$ in (2.17). Thus, we obtain

$$\begin{aligned}
|b(\boldsymbol{v}, \underline{\xi}_\sigma)| &= \left| \int_{\mathcal{E}_i} (\{\!\{\underline{\xi}_\sigma\}\!\} - [\![\underline{\xi}_\sigma]\!] \otimes \boldsymbol{C}_{12}) : \underline{[\![\boldsymbol{v}]\!]}\, ds + \int_{\mathcal{E}_\mathcal{D}} \underline{\xi}_\sigma : (\boldsymbol{v} \otimes \boldsymbol{n})\, ds \right| \\
&\le \left(\int_{\mathcal{E}_i} \frac{1}{C_{11}} (\{\!\{\underline{\xi}_\sigma\}\!\} - [\![\underline{\xi}_\sigma]\!] \otimes \boldsymbol{C}_{12})^2\, ds + \int_{\mathcal{E}_\mathcal{D}} \frac{1}{C_{11}} \underline{\xi}_\sigma^2\, ds \right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0). \\
&\le C \left(\sum_{K \in \mathcal{T}} \frac{1}{\underline{C}_{11}^{\partial K}} \|\underline{\xi}_\sigma\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0).
\end{aligned}$$

The second estimate for the form b follows from Lemma 3.1.

The estimates for d are obtained in a similar way from Lemma 3.1, observing again that $\boldsymbol{D}_{12}$ is of order one and that the volume terms vanish due to the properties of the L^2-projections and the inclusions in (2.17). Thus, using the inclusion $\partial_i \mathcal{Q}(K) \subset \mathcal{V}(K)$, we have

$$\begin{aligned}
|d(\boldsymbol{\xi}_u, q)| &= \left| \int_{\mathcal{E}_i} (\{\!\{\boldsymbol{\xi}_u\}\!\} + \boldsymbol{D}_{12} [\![\boldsymbol{\xi}_u]\!]) \cdot [\![q]\!]\, ds \right| \\
&\le \left(\int_{\mathcal{E}_i} \frac{1}{D_{11}} (\{\!\{\boldsymbol{\xi}_u\}\!\} + \boldsymbol{D}_{12} [\![\boldsymbol{\xi}_u]\!])^2\, ds \right)^{\frac{1}{2}} \Theta(\boldsymbol{0}, q) \\
&\le C \left(\sum_{K \in \mathcal{T}} \frac{1}{\underline{D}_{11}^{\partial K}} \|\boldsymbol{\xi}_u\|_{0,\partial K} \right)^{\frac{1}{2}} \Theta(\boldsymbol{0}, q),
\end{aligned}$$

and using the inclusion $\partial_i \mathcal{V}(K) \subset \mathcal{Q}(K)$,

$$\begin{aligned}
|d(\boldsymbol{v}, \xi_p)| &= \left| \int_{\mathcal{E}_i} (\{\!\{\xi_p\}\!\} - \boldsymbol{D}_{12} \cdot [\![\xi_p]\!]) [\![\boldsymbol{v}]\!]\, ds + \int_{\mathcal{E}_\mathcal{D}} \xi_p \boldsymbol{v} \cdot \boldsymbol{n}\, ds \right| \\
&\le C \left(\sum_{K \in \mathcal{T}} \frac{1}{\underline{C}_{11}^{\partial K}} \|\xi_p\|_{0,\partial K}^2 \right)^{\frac{1}{2}} \Theta(\boldsymbol{v}, 0).
\end{aligned}$$

The application of Lemma 3.1 completes the proof. ∎

For the special form of C_{11} and D_{11} proposed in (2.21) and (2.22), respectively, we have as a consequence of Lemma 3.6, Lemma 3.7, and (2.16) the following result.

COROLLARY 3.8. *Assume* (2.15), (2.16) *and* (2.17), (2.18). *Let the coefficients* C_{11} *and* D_{11} *be given by* (2.21), (2.22), *and let* $\underline{\Pi}$, $\mathbf{\Pi}$, *and* Π *be the (componentwise)* L^2*-projections onto the corresponding finite element spaces. Let* $\underline{\sigma} \in H^{r+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{s+1}(\Omega)^d$, *and* $p \in H^{t+1}(\Omega)$ *for* $r, s, t \geq 0$. *Then we have*

$$\begin{aligned}
|a(\underline{\xi}_\sigma, \underline{\tau})| &\leq Ch^{\min(r,l)+1}\|\underline{\sigma}\|_{r+1}\|\underline{\tau}\|_0 && \forall\ \underline{\tau} \in \underline{\Sigma},\\
|b(\boldsymbol{\xi}_u, \underline{\tau})| &\leq Ch^{\min(s,k)}\|\boldsymbol{u}\|_{s+1}\|\underline{\tau}\|_0 && \forall\ \underline{\tau} \in \underline{\Sigma}_N,\\
|b(\boldsymbol{v}, \underline{\xi}_\sigma)| &\leq c_{11}^{-\frac{1}{2}}Ch^{\min(r,l)+1}\|\underline{\sigma}\|_{r+1}\Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V}_N,\\
|c(\boldsymbol{\xi}_u, \boldsymbol{v})| &\leq c_{11}^{\frac{1}{2}}Ch^{\min(s,k)}\|\boldsymbol{u}\|_{s+1}\Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V},\\
|d(\boldsymbol{\xi}_u, q)| &\leq d_{11}^{-\frac{1}{2}}Ch^{\min(s,k)}\|\boldsymbol{u}\|_{s+1}\Theta(\mathbf{0}, q) && \forall q \in Q_N,\\
|d(\boldsymbol{v}, \xi_p)| &\leq c_{11}^{-\frac{1}{2}}Ch^{\min(t,m)+1}\|p\|_{t+1}\Theta(\boldsymbol{v}, 0) && \forall \boldsymbol{v} \in \boldsymbol{V}_N,\\
|e(\xi_p, q)| &\leq d_{11}^{\frac{1}{2}}Ch^{\min(t,m)+1}\|p\|_{t+1}\Theta(\mathbf{0}, q) && \forall q \in Q,
\end{aligned}$$

with constants C *independent of the meshsize.*

From Corollary 3.8 we are able to derive the following estimate for $K_{\mathcal{B}}$.

COROLLARY 3.9. *Assume* (2.15), (2.16) *and* (2.17), (2.18), *with approximation orders* $k \geq 1$, $l, m \geq 0$. *Let the coefficients* C_{11} *and* D_{11} *be given by* (2.21), (2.22), *and let* $\underline{\Pi}$, $\mathbf{\Pi}$, Π *denote* L^2*-projections. For* $\underline{\sigma} \in H^{l+1}(\Omega)^{d^2}$, $\boldsymbol{u} \in H^{k+1}(\Omega)^d$, *and* $p \in H^{m+1}(\Omega)$ *the error bound* (2.29) *is satisfied with*

$$K_{\mathcal{B}}(\underline{\sigma}, \boldsymbol{u}, p) \leq C\Big[h^{l+1}\|\underline{\sigma}\|_{l+1} + h^k\|\boldsymbol{u}\|_{k+1} + h^{m+1}\|p\|_{m+1}\Big].$$

Furthermore, assume the elliptic regularity (2.26) *and let* $(\boldsymbol{z}, q)$ *denote the solution of* (2.23)–(2.25) *with right-hand side* $\boldsymbol{\lambda} \in L^2(\Omega)^d$, $\underline{\zeta} = -\nabla\boldsymbol{z}$, $\widetilde{q} = -q$. *Then we have in Lemma* 2.7

$$K_{\mathcal{B}}(\underline{\zeta}, \boldsymbol{z}, \widetilde{q}) \leq Ch\|\underline{\lambda}\|_0.$$

Proof. The first assertion follows from the fact that

$$\begin{aligned}
\mathcal{A}(\underline{\xi}_\sigma, \pm\boldsymbol{\xi}_u, \xi_p; \underline{\tau}, \pm\boldsymbol{v}, q) = a(\underline{\xi}_\sigma, \underline{\tau}) \pm b(\boldsymbol{\xi}_u, \underline{\tau}) \mp b(\boldsymbol{v}, \underline{\xi}_\sigma) \pm c(\boldsymbol{\xi}_u, \boldsymbol{v})\\
\pm d(\boldsymbol{v}, \xi_p) \mp d(\boldsymbol{\xi}_u, q) + e(\xi_p, q),
\end{aligned}$$

from the definition of the $\mathcal{A}$-seminorm, and from Corollary 3.8.

The second assertion follows similarly from Corollary 3.8, substituting $(\underline{\sigma}, \boldsymbol{u}, p)$ by $(\underline{\zeta}, \boldsymbol{z}, \widetilde{q})$, observing the special form of C_{11} and D_{11}, and (2.26) which gives $\|\underline{\zeta}\|_1 + \|\boldsymbol{z}\|_2 + \|\widetilde{q}\|_1 \leq C\|\boldsymbol{\lambda}\|_0$. ∎

3.4. Proof of Proposition 2.5. We prove the stability result in Proposition 2.5. To do so, we fix $(\underline{\tau}, \boldsymbol{v}, q) \in \underline{\Sigma}_N \times \boldsymbol{V}_N \times Q_N$. Then, by the continuous inf-sup condition for the standard Stokes forms (see, e.g., [6, 24]) there is a velocity field $\boldsymbol{u} \in H_0^1(\Omega)^d = \{\boldsymbol{u} \in H^1(\Omega)^d : \boldsymbol{u}|_{\partial\Omega} = \mathbf{0}\}$ satisfying

$$-\int_\Omega q\,\nabla\cdot\boldsymbol{u}\,d\boldsymbol{x} \geq \kappa\|q\|_0^2, \qquad \|\boldsymbol{u}\|_1 \leq \|q\|_0, \tag{3.2}$$

with a constant $\kappa > 0$ just depending on Ω. Let $\mathbf{\Pi}\boldsymbol{u}$ be the L^2-projection of $\boldsymbol{u}$ onto the finite element space $\boldsymbol{V}_N$. By definition of $\mathcal{A}$, we have

$$\mathcal{A}(\underline{\tau}, \boldsymbol{v}, q; \underline{0}, \mathbf{\Pi}\boldsymbol{u}, 0) = -b(\mathbf{\Pi}\boldsymbol{u}, \underline{\tau}) + c(\boldsymbol{v}, \mathbf{\Pi}\boldsymbol{u}) + d(\Pi\boldsymbol{u}, q) =: T_1 + T_2 + T_3.$$

We set $\boldsymbol{\xi}_u := \boldsymbol{u} - \boldsymbol{\Pi}\boldsymbol{u}$ and estimate each of the terms T_1–T_3 separately.

For T_1 we have, by Corollary 3.8,

$$|T_1| \le |b(\boldsymbol{\xi}_u, \underline{\tau})| + |b(\boldsymbol{u}, \underline{\tau})| \le C\|\boldsymbol{u}\|_1\|\underline{\tau}\|_0 + \left|\int_\Omega \nabla\boldsymbol{u} : \underline{\tau}\, d\boldsymbol{x}\right| \le C\|\boldsymbol{u}\|_1\|\underline{\tau}\|_0,$$

and, by (3.2),

$$T_1 \ge -\frac{C_1}{\varepsilon_1}\|q\|_0^2 - C_1\varepsilon_1\|\underline{\tau}\|_0^2.$$

For the second term T_2 we have, analogously,

$$T_2 = c(\boldsymbol{v}, \boldsymbol{\Pi}\boldsymbol{u}) = c(\boldsymbol{v}, \boldsymbol{\xi}_u) \le c_{11}^{\frac{1}{2}} C\|\boldsymbol{u}\|_1 \Theta(\boldsymbol{v}, 0),$$

and hence

$$T_2 \ge -\frac{C_2 c_{11}}{\varepsilon_2}\|q\|_0^2 - C_2\varepsilon_2\Theta^2(\boldsymbol{v}, q).$$

For the third term, we write

$$T_3 = d(\boldsymbol{\Pi}\boldsymbol{u}, q) = d(\boldsymbol{u}, q) - d(\boldsymbol{\xi}_u, q).$$

Since, by Corollary 3.8 and (3.2)

$$|d(\boldsymbol{\xi}_u, q)| \le d_{11}^{-\frac{1}{2}} C\|\boldsymbol{u}\|_1 \Theta(\boldsymbol{0}, q) \le \frac{C d_{11}^{-1}}{\varepsilon_3}\|q\|_0^2 + C\varepsilon_3\Theta^2(\boldsymbol{v}, q),$$

and $d(\boldsymbol{u}, q) = -\int_\Omega q\nabla\cdot\boldsymbol{u}\, d\boldsymbol{x}$, we obtain

$$T_3 \ge \kappa\|q\|_0^2 - \frac{C_3 d_{11}^{-1}}{\varepsilon_3}\|q\|_0^2 - C_3\varepsilon_3\Theta^2(\boldsymbol{v}, q).$$

From the above estimates we conclude that

$$\begin{aligned}&\mathcal{A}(\underline{\tau}, \boldsymbol{v}, q; \underline{0}, \boldsymbol{\Pi}\boldsymbol{u}, 0)\\ &\quad\ge \Big(\kappa - \frac{C_1}{\varepsilon_1} - \frac{C_2 c_{11}}{\varepsilon_2} - \frac{C_3 d_{11}^{-1}}{\varepsilon_3}\Big)\|q\|_0^2 - C_1\varepsilon_1\|\underline{\tau}\|_0^2 - (C_2\varepsilon_2 + C_3\varepsilon_3)\Theta^2(\boldsymbol{v}, q).\end{aligned}$$

The parameters $\{\varepsilon_i\}_{i=1}^3$ can be chosen in such a way that

$$\mathcal{A}(\underline{\tau}, \boldsymbol{v}, q; \underline{0}, \boldsymbol{\Pi}\boldsymbol{u}, 0) \ge K_1\|q\|_0^2 - K_2|\,(\underline{\tau}, \boldsymbol{v}, q)\,|_{\mathcal{A}}^2,$$

with constants K_i independent of the meshsize.

Furthermore, we have by Corollary 3.4

$$|\,(\underline{0}, \boldsymbol{\Pi}\boldsymbol{u}, 0)\,|_{\mathcal{A}}^2 = c(\boldsymbol{\Pi}\boldsymbol{u}, \boldsymbol{\Pi}\boldsymbol{u}) = c(\boldsymbol{\xi}_u, \boldsymbol{\xi}_u) \le c_{11} C\|\boldsymbol{u}\|_1^2 \le K_3^2\|q\|_0^2.$$

The function $\boldsymbol{w} = \boldsymbol{\Pi}\boldsymbol{u}/K_3$ then satisfies the assertion in Proposition 2.5, with $\kappa_1 = K_1/K_3$ and $\kappa_2 = K_2/K_3$. This completes the proof.

3.5. Proof of the main results. Theorems 2.2 and 2.3 follow now immediately by choosing the projection operators $\underline{\Pi}$, $\boldsymbol{\Pi}$, and Π as L^2-projections, by combining Corollaries 3.5 and 3.9 with Lemmas 2.4, 2.6, and 2.7 and by taking into account the form of the coefficients C_{11} and D_{11}.

TABLE 4.1
Convergence rates for P^k-elements.

Degree	Grid	$\lvert e\rvert_{\mathcal{A}}$		$\lVert \underline{e}_\sigma \rVert_0$		$\lVert \boldsymbol{e}_u \rVert_0$		$\lVert e_p \rVert_0$	
k	level	Error	Order	Error	Order	Error	Order	Error	Order
	3	3.4e-1	0.94	2.1e-1	0.69	8.4e-3	2.07	2.0e-2	1.57
1	4	1.7e-1	0.96	1.2e-1	0.85	2.1e-3	2.03	8.2e-3	1.27
	5	8.8e-2	0.98	6.2e-2	0.93	5.1e-4	2.01	3.4e-3	1.25
	3	1.2e-2	1.87	9.1e-3	1.73	2.0e-4	3.07	5.1e-4	2.46
2	4	3.2e-3	1.84	2.5e-3	1.88	2.4e-5	3.06	1.2e-4	2.08
	5	9.2e-4	1.80	6.4e-4	1.94	2.9e-6	3.03	3.0e-5	2.00
	2	1.8e-3	2.86	1.4e-3	2.66	5.8e-5	3.98	2.4e-4	2.80
3	3	2.4e-4	2.91	1.9e-4	2.82	3.6e-6	4.02	3.9e-5	2.65
	4	3.0e-5	2.96	2.5e-5	2.91	2.2e-7	4.02	5.3e-6	2.87

4. Numerical results. The numerical experiments we present in this section are devised to verify our theoretical error estimates. We also explore the effect of the use of several combinations of polynomial spaces on the efficiency of the resulting LDG methods. The numerical tests are carried out by using the finite element library `deal.II` by Bangerth and Kanschat [3].

We consider the Stokes system (1.1) with $\Omega = (-1,1)^2$ and right-hand side $\boldsymbol{f}$ and Dirichlet boundary condition $\boldsymbol{g}_{\mathcal{D}}$ chosen such that the exact solution is

$$\begin{aligned} u_1(x_1,x_2) &= -e^{x_1}(x_2\cos x_2 + \sin x_2),\\ u_2(x_1,x_2) &= e^{x_1}x_2\sin x_2,\\ p(x_1,x_2) &= 2e^{x_1}\sin x_2. \end{aligned}$$

In all our experiments, we use uniform triangulations made of squares; the grid whose squares have size $h = 2^{-\nu+1}$ is called a grid of level ν.

4.1. Verifying the sharpness of the theoretical error estimates. We begin by considering LDG methods with the same polynomial spaces for $\underline{\sigma}$, $\boldsymbol{u}$, and p and taking $C_{11} = h^{-1}$, $D_{11} = h$, $\boldsymbol{C}_{12} = \boldsymbol{0}$, and $\boldsymbol{D}_{12} = \boldsymbol{0}$. The results are shown in Tables 4.1 and 4.2 for P^k- and Q^k-elements, respectively. The tables confirm that the orders of convergence predicted by the theory are sharp since they are actually achieved. However, one exception needs to be pointed out: The pressure converges better than expected for linear and bilinear shape functions since superlinear convergence is observed. This phenomenon is particularly well accentuated in the case of bilinear functions for which the order of convergence of 3/2 can be clearly seen. The same order of convergence has recently been observed by Berrone [5] for the stabilized P^1-P^1 SUPG method.

4.2. The effect of the use of different polynomial spaces. To get an idea of what is the effect of the use of P^k- versus Q^k-spaces on quadrilateral elements, the errors of quadratic and biquadratic elements are compared in relation to the numerical effort in Figure 4.1. We use the number of nonzero elements in the stiffness matrix as a measure of the solution cost of a discretization. The graphs show that it is possible to compute the velocities $\boldsymbol{u}$ with the same accuracy and effort with P^2- and Q^2-shape functions; however, the pressures are computed more efficiently with P^2-elements.

Finally, since the theoretical results predict the same orders of convergence for all quantities if we take lower order P-elements for $\underline{\sigma}$ and p, we compare the efficiency of LDG methods obtained with several combinations of local spaces $\mathcal{S}(K)/\mathcal{V}(K)/\mathcal{Q}(K)$ in Figures 4.2 and 4.3.

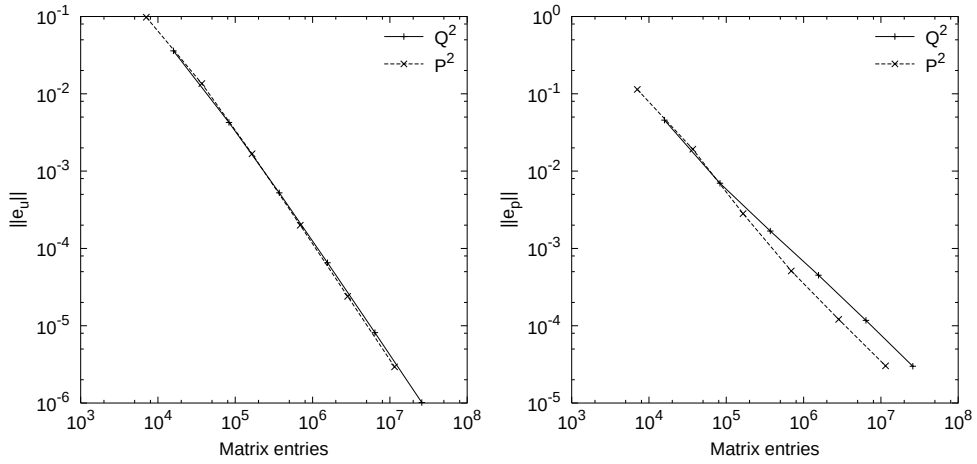


FIG. 4.1. *Comparison of quadratic P^2- and biquadratic Q^2-elements.*

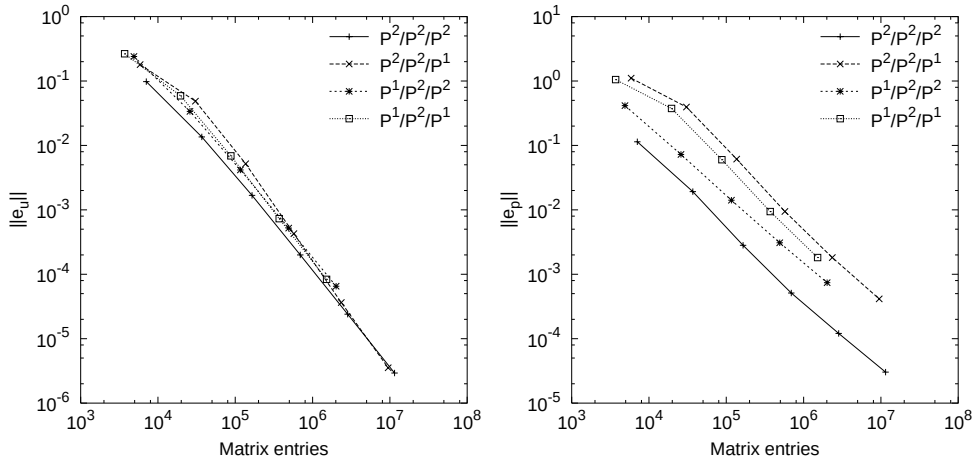


FIG. 4.2. *Comparison of mixed spaces for quadratic P^2-velocities.*

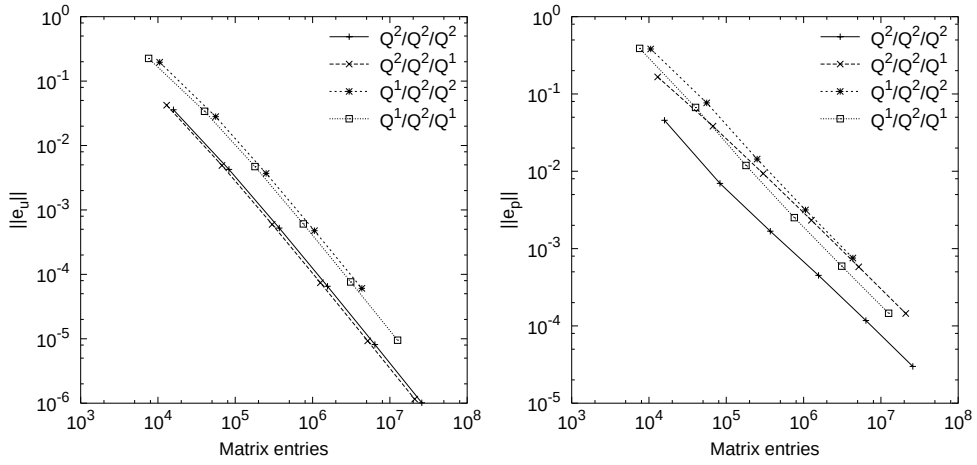


FIG. 4.3. *Comparison of mixed spaces for biquadratic Q^2-velocities.*

TABLE 4.2
Convergence rates for Q^k-elements.

Degree	Grid	$\lvert e\rvert_{\mathcal{A}}$		$\lVert \underline{e}_{\sigma}\rVert_0$		$\lVert \boldsymbol{e}_u\rVert_0$		$\lVert e_p\rVert_0$	
k	level	Error	Order	Error	Order	Error	Order	Error	Order
1	3	2.4e-1	0.94	2.2e-1	0.73	5.6e-3	2.06	2.9e-2	1.50
	4	1.3e-1	0.96	1.2e-1	0.86	1.4e-3	2.04	1.0e-2	1.52
	5	6.4e-2	0.97	6.2e-2	0.93	3.4e-4	2.01	3.8e-3	1.43
2	3	2.6e-3	2.00	6.3e-4	2.10	6.5e-5	3.01	4.5e-4	1.90
	4	6.4e-4	2.00	1.6e-4	2.02	8.1e-6	3.00	1.2e-4	1.94
	5	1.6e-4	2.00	3.9e-5	2.00	1.0e-6	3.00	3.0e-5	1.97
3	2	6.1e-4	2.76	3.8e-4	2.37	1.9e-5	3.82	2.4e-4	2.24
	3	8.1e-5	2.92	6.4e-5	2.55	1.1e-6	4.12	3.8e-5	2.63
	4	1.0e-5	2.98	9.3e-6	2.80	6.0e-8	4.19	5.2e-6	2.88

We can see that all these LDG discretizations converge with the same order, as expected and proved for P-elements, and that, in most cases, it is more efficient to use the same local approximating spaces for all quantities. In fact, only the velocities in the Q^2-case are computed slightly more efficiently using a lower degree for the pressure. On the other hand, using lower-order polynomials for $\underline{\sigma}$ and/or p *increases* the error in p such that at least one additional refinement is necessary to recover the accuracy corresponding to an LDG method using the same local spaces.

5. Extensions and concluding remarks. In this paper, we have introduced LDG methods for the Stokes system and have carried out an a priori error analysis. We have shown that if polynomial approximations of degree $k-1$ are used for the pressure p and the stress tensor $\underline{\sigma}$ and polynomial approximations of degree k for the velocity $\boldsymbol{u}$, then optimal error estimates are obtained when the stabilization parameters C_{11} and D_{11} are taken to be of order h^{-1} and h, respectively. Future work will be devoted to the extension of the LDG method to the incompressible Navier–Stokes equations.

Extensions of our analysis to curvilinear elements and to (nonconvex) polygonal domains as well as to error estimates in negative-order norms for both the velocity and the pressure can easily be carried out; see [8] for details of the corresponding extensions for the Laplacian. Here, we simply must note that, to take into account the presence of the pressure, we have to consider the following modified adjoint problem:

$$\begin{aligned} -\Delta \boldsymbol{z} + \nabla q &= \boldsymbol{\lambda} && \text{in } \Omega, \\ \nabla \cdot \boldsymbol{z} &= g && \text{in } \Omega, \\ \boldsymbol{z} &= \boldsymbol{0} && \text{on } \partial\Omega, \end{aligned}$$

where g is in $H^1(\Omega)$. The elliptic regularity result we have used in (2.26) is a particular case of the above more general case; see, for example, Proposition 3.14 in [2] and the references therein.

The technique of the analysis employed is an extension of that used in [8] for the simpler case of the Laplacian. This same technique was then used in [13] to get improved convergence estimates for a special LDG method on Cartesian grids by changing some auxiliary projections used in the analysis. In a forthcoming paper, we shall carry out a similar study for the Stokes system. The numerical results in Table 5.1 already suggest an improvement similar to that obtained for the Laplacian.

Indeed, the use of these special fluxes with quadratic shape functions increases the order of convergence of the pressure by 1/2; moreover, they improve the order of

TABLE 5.1
Orders of convergence for "superconvergent" fluxes.

Level	P^2-elements			Q^2-elements		
	$\|\underline{e}_\sigma\|_0$	$\|\boldsymbol{e}_u\|_0$	$\|e_p\|_0$	$\|\underline{e}_\sigma\|_0$	$\|\boldsymbol{e}_u\|_0$	$\|e_p\|_0$
1	1.77	2.85	2.38	2.28	2.80	2.05
2	1.86	2.94	2.49	2.43	2.90	2.33
3	1.92	2.98	2.60	2.48	2.95	2.45
4	1.96	2.99	2.66	2.49	2.98	2.48
5	1.98	3.00	2.68	2.50	2.99	2.49

convergence of the pressure *and* the stresses by 1/2 when biquadratic finite elements are used.

Acknowledgments. The authors are grateful to Leopoldo Franca for bringing to their attention the results of Berrone [5].

REFERENCES

[1] D.N. ARNOLD, F. BREZZI, B. COCKBURN, AND D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] G.A. BAKER, W.N. JUREIDINI, AND O.A. KARAKASHIAN, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.

[3] W. BANGERTH AND G. KANSCHAT, *Concepts for Object-Oriented Finite Element Software—The* `deal.`II *Library*, Preprint 99-43, Sonderforschungbereich 3-59, IWR, Universität Heidelberg, Heidelberg, Germany, 1999.

[4] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[5] S. BERRONE, *Adaptive discretization of stationary and incompressible Navier-Stokes equations by stabilized finite element methods*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4435–4455.

[6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[7] P. CASTILLO, *An optimal error estimate for the local discontinuous Galerkin method*, in First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G.E. Karniadakis, and C.W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 285–290.

[8] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[9] P. CASTILLO, B. COCKBURN, D. SCHÖTZAU, AND C. SCHWAB, *Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems*, Math. Comp., 71 (2002), pp. 455–478.

[10] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, The Netherlands, 1978.

[11] B. COCKBURN AND C. DAWSON, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions*, in The Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X, J.R. Whiteman, ed., Elsevier, Oxford, UK, 2000, pp. 225–238.

[12] B. COCKBURN, S. HOU, AND C.W. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV*: The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[13] B. COCKBURN, G. KANSCHAT, I. PERUGIA, AND D. SCHÖTZAU, *Superconvergence of the local discontinuous Galerkin method for elliptic problems on Cartesian grids*, SIAM J. Numer. Anal., 39 (2001), pp. 264–285.

[14] B. COCKBURN, G.E. KARNIADAKIS, AND C.W. SHU, *The development of discontinuous Galerkin methods*, in First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G.E. Karniadakis, and C.W. Shu, eds., Springer-

Verlag, Berlin, 2000, pp. 3–50.

[15] B. Cockburn, S.Y. Lin, and C.W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* III*: One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[16] B. Cockburn and C.W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* II*: General framework*, Math. Comp., 52 (1989), pp. 411–435.

[17] B. Cockburn and C.W. Shu, *The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[18] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin finite element method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[19] B. Cockburn and C.W. Shu, *The Runge-Kutta discontinuous Galerkin finite element method for conservation laws* V*: Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[20] C. Dawson and J. Proft, *A priori estimates for interior penalty versions of the local discontinuous Galerkin method applied to transport equations*, Numer. Methods Partial Differential Equations, to appear.

[21] M. Fortin, *Finite element solution of the Navier-Stokes equations*, Acta Numer., 5 (1993), pp. 239–284.

[22] L.P. Franca, T. Hughes, and R. Stenberg, *Stabilized finite element methods*, in Incompressible Computational Fluid Dynamics: Trends and Advances, M.D. Gunzburger and R.A. Nicolaides, eds., Cambridge University Press, Cambridge, UK, 1993, pp. 87–107.

[23] L.P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

[24] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, New York, 1986.

[25] T. Hughes, G. Engel, L. Mazzei, and M. Larson, *A comparison of discontinuous and continuous Galerkin methods*, in First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G.E. Karniadakis, and C.W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 135–146.

[26] T. Hughes and L. Franca, *A new finite element formulation for computational fluid dynamics,* VII*. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces*, Comput. Methods Appl. Mech. Engrg., 65 (1987), pp. 85–96.

[27] T. Hughes, L.P. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics,* V*. Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[28] O. Karakashian and T. Katsaounis, *A discontinuous Galerkin method for the incompressible Navier-Stokes equations*, in First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G.E. Karniadakis, and C.W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 157–166.

[29] N. Kechkar and D.J. Silvester, *Analysis of locally stabilized mixed finite element methods for the Stokes problem*, Math. Comp., 58 (1992), pp. 1–10.

[30] D. Pelletier, A. Fortin, and R. Camarero, *Are FEM solutions of incompressible flows really incompressible? (or how simple flows can cause headaches!)*, Internat. J. Numer. Methods Fluids, 9 (1989), pp. 99–112.

[31] D.J. Silvester and N. Kechkar, *Stabilized bilinear-constant velocity-pressure finite elements for the conjugate gradient solution of the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 79 (1990), pp. 71–86.

[32] R. Témam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, 1979.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 344–375

FAST COLLOCATION METHODS FOR SECOND KIND INTEGRAL EQUATIONS*

ZHONGYING CHEN†, CHARLES A. MICCHELLI‡, AND YUESHENG XU§

Dedicated to Professor Mike Powell
on the occasion of his sixty-fifth birthday with friendship and esteem

Abstract. In this paper we develop fast collocation methods for integral equations of the second kind with weakly singular kernels. For this purpose, we construct *multiscale interpolating functions* and *collocation functionals* having vanishing moments. Moreover, we propose a truncation strategy for the coefficient matrix of the corresponding discrete system which forms a basis for fast algorithms. An optimal order of convergence of the approximate solutions obtained from the fast algorithms is proved and the computational complexity of the algorithms is estimated. The stability of the numerical method and the condition number of the truncated coefficient matrix are analyzed.

Key words. fast collocation methods, Fredholm integral equations of the second kind, refinable sets, multiscale interpolation

AMS subject classifications. 65B05, 45L10

PII. S0036142901389372

1. Introduction.

The main purpose of this paper is to develop a fast collocation method for solving Fredholm integral equations of the second kind with weakly singular kernels. Among conventional numerical methods for solving integral equations, the collocation method receives more favorable attention from engineering applications due to lower computational cost in generating the coefficient matrix of the corresponding discrete equations. In comparison, the implementation of the Galerkin method requires much more computational effort for the evaluation of integrals (for example, see [At1, AC1, CX]). Nonetheless, it seems that most of the attention in wavelet methods for boundary integral equations has been paid to Galerkin methods or Petrov–Galerkin methods (see [BCR, CMX1, CMX3, DPS, MXZ, PS, PSS, R] and references cited therein). These methods are amenable to L_2 analysis and therefore the vanishing moments of the multiscale basis functions naturally lead to matrix truncation techniques.

For collocation methods, the appropriate context to work in is L_∞ and this provides challenging technical obstacles for the identification of good matrix truncation strategies. It is the goal of this paper to lay the foundations for fast collocation

*Received by the editors May 11, 2001; accepted for publication (in revised form) January 15, 2002; published electronically May 10, 2002.
http://www.siam.org/journals/sinum/40-1/38937.html

†Department of Scientific Computing and Computer Applications, Zhongshan University, Guangzhou 510275, P. R. China (lnsczy@zsu.edu.cn). The research of this author was supported in part by the Morningside Center of Mathematics at the Chinese Academy of Sciences, Guangdong Provincial Natural Science Foundation of China, and the Advanced Research Foundation of Zhongshan University.

‡Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY 12222 (cam@math.albany.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9973427.

§Department of Mathematics, West Virginia University, Morgantown, WV 26506 and Institute of Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, P. R. China (yxu@math.wvu.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9973427, by NASA under grant NCC5-399, and by the program of "One Hundred Distinguished Young Scientists" of the Chinese Academy of Sciences.

methods for solving integral equations of the second kind with weakly singular kernels. This paper is divided into two main parts. In part one we develop and analyze a fast collocation method for solving general multidimensional integral equations under a sequence of hypotheses on the basis functions and collocation functionals. In part two, we present a construction of such functions and functionals when the domain of interest is an invariant set relative to a finite set of contractive affine mappings. This construction draws upon ideas developed in [CMX2, MSX1, MSX2].

The equations we consider in this paper are Fredholm integral equations of the second kind with weakly singular kernels defined on a bounded domain in $\mathbb{R}^d$ for $d \geq 1$. These types of equations cover many important applications including boundary integral equations [At1, At2] and the radiosity equations [ACr, AC2]. In practice, these equations are solved numerically by using piecewise polynomial collocation methods and, when the order of the full coefficient matrix is large, the computational cost for generating the matrix and then solving the corresponding linear system is huge. We introduce a matrix truncation strategy by making a careful choice of basis functions and collocation functionals, which leads to a fast algorithm for solving the integral equations.

We organize this paper as follows. In section 2, we describe the setting of the collocation methods, including the multiscale basis for the approximate solution space and the collocation functionals. In section 3, based on an estimate of the entries of the coefficient matrix of the collocation method, a fast collocation method is proposed, and in section 4, we analyze the matrix truncation algorithm including the order of convergence, stability, and computational complexity. The fast collocation methods that we develop and analyze in sections 3 and 4 are based on a set of hypotheses on the basis functions of the approximation space and the collocation functionals. In section 5 we present a concrete construction of multiscale bases on an invariant set in $\mathbb{R}^d$ and collocation functionals needed for fast collocation algorithms. This construction fulfills all the hypotheses imposed in sections 2–4.

2. The collocation scheme. In this section we describe a general setup for multiscale bases and collocation functionals for solving Fredholm integral equations of the second kind. We begin establishing our notational conventions. For a compact subset E of the d-dimensional Euclidean space $\mathbb{R}^d$, we let $\mathbb{X} = L^\infty(E)$, $\mathbb{V} = C(E)$ and denote the dual space of $\mathbb{V}$ by $\mathbb{V}^*$. For $\ell \in \mathbb{V}^*$ and $v \in \mathbb{X}$, we use $\langle \ell, v \rangle$ to denote the value of the linear functional ℓ evaluated at the function v and use $\|\ell\|$, $\|v\|_\infty$ for their respective norms. For any $s \in E$ we use δ_s to denote the linear functional in $\mathbb{V}^*$ defined for $v \in \mathbb{V}$ by the equation $\langle \delta_s, v \rangle = v(s)$. We shall need to evaluate δ_s on functions in $\mathbb{X}$. Therefore, as in [AGS] we take any norm preserving extension of δ_s to $\mathbb{X}$ and use the same notation for the extension. In particular, this convention allows us to evaluate piecewise polynomials *anywhere* on E.

For a set $A \subset \mathbb{R}^d$, $d(A)$ represents the diameter of A, i.e.,

$$d(A) := \sup\{|x - y| : x, y \in A\}, \tag{2.1}$$

where $|\cdot|$ denotes the Euclidean norm on the space $\mathbb{R}^d$. We use the index sets $\mathbb{Z}_n := \{0, 1, \ldots, n-1\}$, $\mathbb{Z} := \{\ldots, -1, 0, 1, \ldots\}$, $\mathbb{N} := \{1, 2, \ldots\}$, and $\mathbb{N}_0 := \{0, 1, \ldots\}$ throughout the paper. A vector x in $\mathbb{R}^d$ is written in the form $x = [x_i : i \in \mathbb{Z}_d]$ and a lattice point α is an element in $\mathbb{N}_0^d$ written as $\alpha := [\alpha_i \in \mathbb{N}_0 : i \in \mathbb{Z}_d]$. As is usually the case we set $|\alpha| := \sum_{i \in \mathbb{Z}_d} \alpha_i$ and denote the space of polynomials of total degree less than k by π_k.

A set in $\mathbb{R}^d$ is called star-shaped if it contains a point, called the center of the set, for which the line segment connecting this point and any other point in the set is contained in the set. As usual, for k a positive integer $W^{k,\infty}(E)$ will denote the set of all functions v on E such that $D^\alpha v \in \mathbb{X}$, where we use the standard multi-index notation for derivatives

$$D^\alpha v(x) = \frac{\partial^{|\alpha|} v(x)}{\partial x_0^{\alpha_0} \cdots \partial x_{d-1}^{\alpha_{d-1}}}, \quad x \in \mathbb{R}^d,$$

and the norm

$$\|v\|_{k,\infty} := \max\{\|D^\alpha v\|_\infty : |\alpha| \le k\}$$

on $W^{k,\infty}(E)$. For a star-shaped set E it is easy to estimate the distance of a function $v \in W^{k,\infty}(E)$ to the space π_k. Specifically, there is a positive constant c such that

$$\operatorname{dist}(v, \pi_k) \le c(d(E))^k \|v\|_{k,\infty}. \tag{2.2}$$

Throughout the paper c will always stand for a generic constant whose value will change with the context. Its meaning will be clear from the order of the qualifiers used to describe its role in our estimates.

There are several ingredients required in the development of the fast collocation algorithms for solving integral equations. First, we require a *multiscale* of finite dimensional subspaces of $\mathbb{X}$ denoted by $\mathbb{F}_n$ where $n \in \mathbb{N}_0$ in which we do our approximation. These spaces are required to have the property that

$$\mathbb{F}_n \subseteq \mathbb{F}_{n+1}, \quad n \in \mathbb{N}_0, \tag{2.3}$$

and

$$\mathbb{V} \subseteq \overline{\bigcup_{n \in \mathbb{N}_0} \mathbb{F}_n}. \tag{2.4}$$

For efficient computation relative to a scale of spaces we express them as a direct sum of subspaces

$$\mathbb{F}_n = \mathbb{W}_0 \bigoplus \mathbb{W}_1 \bigoplus \cdots \bigoplus \mathbb{W}_n. \tag{2.5}$$

These spaces serve as *multiscale* subspaces of $\mathbb{X}$ and later will be constructed as piecewise polynomial functions on E. We use the notation $w(n) := \dim \mathbb{W}_n$ and so we have that

$$f(n) := \dim \mathbb{F}_n = \sum_{r \in \mathbb{Z}_{n+1}} w(r).$$

We need a multiscale partition of the set E. It consists of a family of partitions $\{E_n : n \in \mathbb{N}_0\}$ of E such that for each scale $n \in \mathbb{N}_0$ the partition E_n consists of a family of subsets $\{E_{ni} : i \in \mathbb{Z}_{e(n)}\}$ of E with the properties that

$$\operatorname{meas}(E_{ni} \cap E_{ni'}) = 0, \quad i, i' \in \mathbb{Z}_{e(n)}, i \ne i', \tag{2.6}$$

and

$$\bigcup_{i \in \mathbb{Z}_{e(n)}} E_{ni} = E. \tag{2.7}$$

At the appropriate time later, we will adjust the number $e(n)$ of elements and the maximum diameter of the cells in the nth partition to be commensurate with $\dim \mathbb{W}_n$. The family of partitions $\{E_n : n \in \mathbb{N}_0\}$ is used in two ways. First, we demand that there is a basis $W_n := \{w_{nm} : m \in \mathbb{Z}_{w(n)}\}$ for the spaces

$$\mathbb{W}_n := \operatorname{span} W_n, \quad n \in \mathbb{N}_0,$$

having the following property.

(I) There exist positive integers ρ and h such that for every $n > h$ and $m \in \mathbb{Z}_{w(n)}$ written in the form $m = j\rho + s$ where $s \in \mathbb{Z}_\rho$ and $j \in \mathbb{N}_0$

$$w_{nm}(x) = 0, \quad x \notin E_{n-h,j}. \tag{2.8}$$

Hypothesis (I) ensures that the basis functions w_{nm} are locally supported and their supports are shrinking as level n increases. For $n > h$ we use the notation $S_{nm} := E_{n-h,j}$ so that the support of the functions w_{nm} is contained in the set S_{nm}. Note that the supports of the basis functions at the nth level are *not* disjoint. However, for every $n > h$ and every function w_{nm} there are at most ρ other functions at level n whose support overlaps the support of w_{nm}.

To define the collocation method we need a set of linear functionals in $\mathbb{V}^*$ given by

$$L_n := \{\ell_{nm} : m \in \mathbb{Z}_{w(n)}\}, \quad n \in \mathbb{N}_0.$$

The multiscale of partitions $\{E_n : n \in \mathbb{N}_0\}$ is also used to specify the supports of the linear functionals by the requirement that the linear functional ℓ_{nm} is a finite sum of point evaluations

$$\ell_{nm} = \sum_{s \in \hat{E}_{n-h,j}} c_s \delta_s, \tag{2.9}$$

where c_s are constants and $\hat{E}_{ni}$ is a finite subset of distinct points in E_{ni} with the cardinality bounded *independent* of $n \in \mathbb{N}$ and $i \in \mathbb{Z}_{w(n)}$. As with the functions we set $\hat{S}_{nm} := \hat{E}_{n-h,j}$ and consider it as the "support" of the functionals ℓ_{nm}.

The linear functionals and multiscale basis functions are tied together by the next requirement.

(II) For any $n, n' \in \mathbb{N}_0$

$$\langle \ell_{n'm'}, w_{nm} \rangle = \delta_{nn'}\delta_{mm'}, \quad (n,m),(n',m') \in \mathbb{U}, \quad n \le n', \tag{2.10}$$

where $\mathbb{U} := \{(i,j) : i \in \mathbb{N}_0, j \in \mathbb{Z}_{w(i)}\}$ and $\delta_{ii'}$ is the Kronecker delta and there exists a positive constant γ for which

$$\sum_{m \in \mathbb{Z}_{w(n)}} |\langle \ell_{n'm'}, w_{nm} \rangle| \le \gamma, \quad (n,m),(n'm') \in \mathbb{U}, \quad n > n'. \tag{2.11}$$

To include the commonly used piecewise polynomial collocation methods with a change of bases as our important special cases, we do not require the linear functionals and the multiscale basis functions to be biorthogonal. Instead, we require them to have a "semibiorthogonality" property imposed by condition (2.10) with a controllable perturbation from the biorthogonality, which is ensured by condition (2.11). Specifically, (2.10) means that the basis functions vanish when they are applied by

collocation functionals of higher levels. We denote by $\mathbf{E}$ the semi-infinite matrix with entries

$$\mathbf{E}_{n'm',nm} := \langle \ell_{n'm'}, w_{nm} \rangle, \quad (n',m'),(n,m) \in \mathbb{U}.$$

We note by condition (2.10) that the matrix $\mathbf{E}$ can be viewed as a block upper triangular matrix with the diagonal blocks equal to identity matrices. Consequently, the infinite matrix $\mathbf{E}$ has an inverse $\mathbf{E}^{-1}$ of the same type, that is,

$$(\mathbf{E}^{-1})_{n'm',nm} = \delta_{nn'}\delta_{mm'}, \quad n \le n', \quad m \in \mathbb{Z}_{w(n)}, \quad m' \in \mathbb{Z}_{w(n')}.$$

These conditions are more than is needed to introduce the collocation method for solving integral equations of the second kind. For this purpose, we suppose that K is a weakly singular kernel, that is, for every $s \in E$, $K(s,\cdot) \in L^1(E)$. Therefore, the operator $\mathcal{K} : \mathbb{X} \to \mathbb{V}$ defined by

$$(\mathcal{K}u)(s) := \int_E K(s,t)u(t)dt, \quad s \in E,$$

is compact in $\mathbb{X}$. We consider Fredholm integral equations of the second kind in the form

$$u - \mathcal{K}u = f, \tag{2.12}$$

where $f \in \mathbb{X}$ is a given function and $u \in \mathbb{X}$ is the unknown to be determined. When one is not an eigenvalue of $\mathcal{K}$, (2.12) has a unique solution in $\mathbb{X}$. The collocation scheme for solving (2.12) seeks a vector $\mathbf{u}_n := [u_{ij} : (i,j) \in \mathbb{U}_n]^T$, where $\mathbb{U}_n$ is the set of lattice points in $\mathbb{R}^2$ defined as $\{(i,j) : j \in \mathbb{Z}_{w(i)}, i \in \mathbb{Z}_{n+1}\}$, such that the function

$$u_n := \sum_{(i,j)\in\mathbb{U}_n} u_{ij}w_{ij}$$

in $\mathbb{F}_n$ has the property that

$$\langle \ell_{i'j'}, u_n - \mathcal{K}u_n \rangle = \langle \ell_{i'j'}, f \rangle, \quad (i',j') \in \mathbb{U}_n. \tag{2.13}$$

Equivalently, we obtain the linear system of equations

$$(\mathbf{E}_n - \mathbf{K}_n)\mathbf{u}_n = \mathbf{f}_n,$$

where

$$\mathbf{K}_n := [\langle \ell_{i'j'}, \mathcal{K}w_{ij} \rangle]_{f(n)\times f(n)},$$

$$\mathbf{E}_n := [\langle \ell_{i'j'}, w_{ij} \rangle]_{f(n)\times f(n)},$$

and

$$\mathbf{f}_n := [\langle \ell_{ij}, f \rangle : (i,j) \in \mathbb{U}_n]^T.$$

By definition, we have that $(\mathbf{E}_n)_{i'j',ij} = \mathbf{E}_{i'j',ij}$, for $(i',j'),(i,j) \in \mathbb{U}_n$ and by (2.10) we see that

$$(\mathbf{E}_n^{-1})_{i'j',ij} = (\mathbf{E}^{-1})_{i'j',ij}, \quad (i',j'),(i,j) \in \mathbb{U}_n. \tag{2.14}$$

The collocation scheme defined in (2.13) which has a multiscale structure is different from the traditional collocation scheme. However, it can be viewed as a scheme obtained from the traditional one by an appropriate change of bases for both approximate spaces and collocation functionals.

Let us use condition (II) to estimate the *inverse* of the matrix $\mathbf{E}_n$. To this end, we introduce a weighted norm on the vector $\mathbf{x} := [x_{ij} : (i,j) \in \mathbb{U}_n]^T$. We will find it convenient for our study of the multiscale collocation methods to define the weighted vector norms in a way that the weights differ from level to level. Specifically, for any $i \in \mathbb{Z}_{n+1}$ we set

$$\mathbf{x}_i := [x_{ij} : j \in \mathbb{Z}_{w(i)}]^T,$$

$$\|\mathbf{x}_i\|_\infty := \max\{|x_{ij}| : j \in \mathbb{Z}_{w(i)}\},$$

and whenever $\nu \in (0,1)$ we define

$$\|\mathbf{x}\|_\nu := \max\{\|\mathbf{x}_i\|_\infty \nu^{-i} : i \in \mathbb{Z}_{n+1}\}.$$

We also use the notation

$$\|\mathbf{x}\|_\infty := \max\{\|\mathbf{x}_i\|_\infty : i \in \mathbb{Z}_{n+1}\}$$

for the max norm of the vector $\mathbf{x}$.

LEMMA 2.1. *If condition* (II) *holds,* $0 < \nu < 1$ *and* $(1+\gamma)\nu < 1$, *then for any integer* $n \in \mathbb{N}_0$ *and vector* $\mathbf{x} \in \mathbb{R}^{f(n)}$ *there holds that*

$$\|\mathbf{x}\|_\nu \le \frac{1-\nu}{1-(1+\gamma)\nu} \|\mathbf{E}_n \mathbf{x}\|_\nu.$$

Proof. Let $\mathbf{y} := \mathbf{E}_n \mathbf{x}$ so that

$$y_{ij} = \sum_{(i',j') \in \mathbb{U}_n} \langle \ell_{ij}, w_{i'j'} \rangle \, x_{i'j'}.$$

In particular, for $i = n$, we have that $y_{ij} = x_{ij}$, while for $0 \le l \le n-1$ we have from (2.10) that

$$x_{n-l-1,j} = y_{n-l-1,j} - \sum_{n-l \le i' \le n, j' \in \mathbb{Z}_{w(i')}} \langle \ell_{n-l-1,j}, w_{i'j'} \rangle \, x_{i'j'}, \quad j \in \mathbb{Z}_{w(n-l-1)}.$$

Using condition (2.11), we conclude that

$$\|\mathbf{x}_{n-l-1}\|_\infty \le \|\mathbf{y}_{n-l-1}\|_\infty + \gamma \sum_{i=0}^{l} \|\mathbf{x}_{n-i}\|_\infty.$$

By induction on j, it readily follows that

$$\|\mathbf{x}_{n-j}\|_\infty \le \sum_{l=0}^{j-1} \gamma (1+\gamma)^l \|\mathbf{y}_{n-j+l+1}\|_\infty + \|\mathbf{y}_{n-j}\|_\infty.$$

Thus, we have that

$$\begin{aligned}\|\mathbf{x}_{n-j}\|_\infty \nu^{-(n-j)} &\le \gamma\nu \sum_{l=0}^{j-1} [(1+\gamma)\nu]^l \|\mathbf{y}_{n-j+l+1}\|_\infty \nu^{-(n-j+l+1)} + \|\mathbf{y}_{n-j}\|_\infty \nu^{-(n-j)} \\ &\le \left[1 + \gamma\nu \sum_{l=0}^{j-1} [(1+\gamma)\nu]^l\right] \|\mathbf{E}_n \mathbf{x}\|_\nu \\ &\le \frac{1-\nu}{1-(1+\gamma)\nu} \|\mathbf{E}_n \mathbf{x}\|_\nu,\end{aligned}$$

from which the result is proved. □

3. Estimates for matrix $\mathbf{K}_n$. In this section, our goal is to obtain estimates for the entries of the matrix $\mathbf{K}_n$. This requires conditions on the regularity of kernel $K(\cdot,\cdot)$, the support of the basis functions for $\mathbb{W}_n$, and vanishing moments for both the basis functions and the linear functionals. We describe these conditions next.

(III) There is a positive integer k such that for all $p \in \pi_k$

$$\langle \ell_{nm}, p\rangle = 0, \quad (w_{nm}, p) = 0, \quad (n,m) \in \mathbb{U}, \quad n \ge 1,$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(E)$.

Condition (III) is crucial for establishing the matrix compression scheme. The vanishing moment condition for the collocation functionals restricts them to being a certain type of divided difference functional.

(IV) There exists a positive constant θ_0 such that for all $(n,m) \in \mathbb{U}$ there holds that

$$\|\ell_{nm}\| + \|w_{nm}\|_\infty \le \theta_0,$$

where the norm of linear functionals is defined as in the beginning of section 2.

Condition (IV) is equivalent to saying that both basis functions and collocation functionals are uniformly bounded.

(V) For $s,t \in E$, $s \ne t$, the kernel K has continuous partial derivatives $D_s^\alpha D_t^\beta K(s,t)$ for $|\alpha| \le k$, $|\beta| \le k$. Moreover, there exists positive constants σ and θ_1 with $\sigma < d$ such that for $|\alpha| = |\beta| = k$ there holds

$$\left| D_s^\alpha D_t^\beta K(s,t) \right| \le \frac{\theta_1}{|s-t|^{\sigma+|\alpha|+|\beta|}}. \tag{3.1}$$

In the next lemma, we present an estimate of the entries of the matrix $\mathbf{K}_n$. A similar estimate for the matrix obtained from the Galerkin method was proved in [MXZ]. For estimates in a different form for the entries of wavelet Galerkin matrices, see [Al, BCR, DPS, PS, PSS]. Such an estimate forms the basis for a truncation strategy. In the statement of the next lemma we use the quantities

$$d_i := \max\{d(S_{ij}) : j \in \mathbb{Z}_{w(i)}\}, \quad i \in \mathbb{N}_0.$$

LEMMA 3.1. *If conditions* (I), (III)–(V) *hold and there is a constant* $r > 1$ *such that*

$$\operatorname{dist}(S_{ij}, S_{i'j'}) \ge r(d_i + d_{i'}), \tag{3.2}$$

then there exists a positive constant c such that

$$|K_{i'j',ij}| \le c(d_i d_{i'})^k \sum_{s\in\hat{S}_{i'j'}} \int_{S_{ij}} \frac{1}{|s-t|^{2k+\sigma}} dt.$$

Proof. Let s_0, t_0 be centers of the sets $S_{i'j'}$ and S_{ij}, respectively. Using the Taylor theorem with remainder we write

$$K = K_1 + K_2 + K_3,$$

where $K_1(s,\cdot)$ and $K_2(\cdot,t)$ are polynomials of total degree $\le k-1$ in t and in s, respectively,

$$|K_3(s,t)| \le d_i^k d_{i'}^k v(s,t), \quad s \in S_{i'j'},\ t \in S_{ij},$$

where

$$v(s,t) := \sum_{|\alpha|=k} \sum_{|\beta|=k} \frac{|r_{\alpha\beta}(s,t)|}{\alpha!\beta!} \tag{3.3}$$

and

$$r_{\alpha\beta}(s,t) = \int_0^1 \int_0^1 D_s^\alpha D_t^\beta K(s_0 + t_1(s-s_0), t_0 + t_2(t-t_0))(1-t_1)^{k-1}(1-t_2)^{k-1} dt_1 dt_2. \tag{3.4}$$

Applying the vanishing moment conditions yields the bound

$$|K_{i'j',ij}| \le \|\ell_{i'j'}\| \|w_{ij}\|_\infty d_i^k d_{i'}^k \sum_{s\in\hat{S}_{i'j'}} \int_{S_{ij}} |v(s,t)| dt. \tag{3.5}$$

It follows from the mean-value theorem and condition (V) that

$$|r_{\alpha\beta}(s,t)| = k^{-2} |D_s^\alpha D_t^\beta K(s',t')| \le \frac{\theta_1}{k^2 |s'-t'|^{2k+\sigma}}$$

holds for some $s' \in S_{i'j'},\ t' \in S_{ij}$. For $s \in \hat{S}_{i'j'} \sqsubseteq S_{i'j'}$, $t \in S_{ij}$ the assumption (3.2) yields

$$|s'-t'| \ge |s-t| - d_i - d_{i'} \ge (1-r^{-1})|s-t|,$$

from which it follows that

$$|r_{\alpha\beta}(s,t)| \le \frac{c_1}{|s-t|^{2k+\sigma}},$$

where

$$c_1 := \frac{\theta_1}{k^2(1-r^{-1})^{2k+\sigma}}.$$

Substituting the above inequality into (3.5) completes the proof with

$$c := \frac{\theta_1 \theta_0^2 e^{\frac{2d}{1-r^{-1}}}}{k^2(1-r^{-1})^\sigma}. \qquad \square$$

To present the truncation strategy we partition matrix $\mathbf{K}_n$ into a block matrix

$$\mathbf{K}_n = [\mathbf{K}_{i'i}]_{i',i\in\mathbb{Z}_{n+1}},$$

with

$$\mathbf{K}_{i'i} = [K_{i'j',ij}]_{j'\in\mathbb{Z}_{w(i')},j\in\mathbb{Z}_{w(i)}}.$$

We truncate the block $\mathbf{K}_{i'i}$ by using a given positive number ϵ to form a matrix

$$\mathbf{K}(\epsilon)_{i'i} = [K(\epsilon)_{i'j',ij}]_{j'\in\mathbb{Z}_{w(i')},j\in\mathbb{Z}_{w(i)}},$$

with

$$K(\epsilon)_{i'j',ij} = \begin{cases} K_{i'j',ij}, & \operatorname{dist}(S_{i'j'}, S_{ij}) \le \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where ϵ is called the truncation parameters and it may depend on i', i, and n. In the next section, we will choose these truncation parameters so that the truncation scheme gives an optimal order of convergence and computational complexity up to a logarithmic factor. In the next lemma, we use the estimate for the entries of $\mathbf{K}_n$ presented in Lemma 3.1 to obtain estimates for the discrepancy between the blocks of $\mathbf{K}(\epsilon)$ and $\mathbf{K}_n$.

LEMMA 3.2. *If conditions* (I), (III)–(V) *hold, then given any constant* $r > 1$ *and* $0 \le \sigma' < \min\{2k, d-\sigma\}$ *there exists a positive constant* c *such that whenever* $\epsilon \ge r(d_i + d_{i'})$

$$\|\mathbf{K}_{i'i} - \mathbf{K}(\epsilon)_{i'i}\|_\infty \le c\epsilon^{-\eta}(d_i d_{i'})^k, \quad i', i \in \mathbb{Z}_{n+1},$$

where $\eta := 2k - \sigma'$.

Proof. We first note that

$$\|\mathbf{K}_{i'i} - \mathbf{K}(\epsilon)_{i'i}\|_\infty = \max_{j'\in\mathbb{Z}_{w(i')}} \sum_{j\in\mathbb{Z}_{i'j'}(\epsilon)} |K_{i'j',ij}|,$$

where

$$\mathbb{Z}_{i'j'}(\epsilon) := \{j : j \in \mathbb{Z}_{w(i)}, \operatorname{dist}(S_{ij}, S_{i'j'}) > \epsilon\}.$$

Therefore, by using Lemma 3.1 we have that

$$\|\mathbf{K}_{i'i} - \mathbf{K}(\epsilon)_{i'i}\|_\infty \le c(d_i d_{i'})^k \max_{j'\in\mathbb{Z}_{w(i')}} \sum_{s\in\hat{S}_{i'j'}} \sum_{j\in\mathbb{Z}_{i'j'}(\epsilon)} \int_{S_{ij}} \frac{1}{|s-t|^{2k+\sigma}} dt.$$

Although the sets S_{ij} are not disjoint we can use property (I) to conclude that

$$\sum_{j\in\mathbb{Z}_{i'j'}(\epsilon)} \int_{S_{ij}} \frac{1}{|s-t|^{2k+\sigma}} dt \le \rho\epsilon^{-\eta} \int_E \frac{1}{|s-t|^{\sigma+\sigma'}} dt.$$

Since $\sigma + \sigma' < d$ and E is a compact set, there holds that

$$\max_{s\in E} \int_E \frac{1}{|s-t|^{\sigma+\sigma'}} dt < \infty.$$

We conclude the above inequalities to obtain the desired estimate. ∎

4. Analysis of the truncation algorithm. In this section, we discuss the truncation strategy for the collocation method proposed in previous sections. We will analyze the order of convergence, stability, and computational complexity of the truncation algorithm. To this end, we let $\mathcal{P}_n$ be the projection operator from $\mathbb{X}$ onto $\mathbb{F}_n$ defined by the requirement that

$$\langle \ell_{ij}, \mathcal{P}_n x\rangle = \langle \ell_{ij}, x\rangle, \quad (i,j) \in \mathbb{U}_n.$$

It follows from (2.10) that $\mathcal{P}_n$ is well defined. We now introduce the operator from $\mathbb{F}_n$ into itself defined by the equation

$$\mathcal{K}_n := \mathcal{P}_n \mathcal{K}|_{\mathbb{F}_n}$$

and note that its matrix representation relative to the basis $\{w_{ij} : (i,j) \in \mathbb{U}_n\}$ is given by $\mathbf{E}_n^{-1}\mathbf{K}_n$. For each block $\mathbf{K}_{i'i}$, $i, i' \in \mathbb{Z}_{n+1}$ of $\mathbf{K}_n$, we shall specify later truncation parameters $\epsilon_{i'i}^n$ and reassemble the block to form a truncation matrix

$$\tilde{\mathbf{K}}_n = [\mathbf{K}(\epsilon_{i'i}^n)_{i'i}]_{i',i\in\mathbb{Z}_{n+1}}.$$

Using this truncation matrix, we let $\tilde{\mathcal{K}}_n : \mathbb{F}_n \to \mathbb{F}_n$ be the linear operator from $\mathbb{F}_n$ into itself relative to the basis $\{w_{ij} : (i,j) \in \mathbb{U}_n\}$ having the matrix representation $\mathbf{E}_n^{-1}\tilde{\mathbf{K}}_n$. Our goal is to provide an essential estimate for the difference of these two operators.

For $v \in L^\infty(E)$ we set

$$\mathcal{P}_n v = \sum_{(i,j)\in\mathbb{U}_n} v_{ij} w_{ij}$$

and note that the quantities v_{ij} are linear functionals of v. In the next lemma, we estimates these quantities under the following two additional requirements.

(VI) There exists a positive integer $\mu > 1$ and positive constants c_-, c_+ such that as $n \to \infty$

$$c_-\mu^n \le \dim\ \mathbb{F}_n \le c_+\mu^n,$$

$$c_-\mu^n \le \dim\ \mathbb{W}_n \le c_+\mu^n,$$

and

$$c_-\mu^{-n/d} \le d_n \le c_+\mu^{-n/d}.$$

Condition (VI) says that the dimension of spaces $\mathbb{F}_n$ and $\mathbb{W}_n$ grows exponentially in n and the diameters d_n decay exponentially in n. The next condition imposes an addition restriction on the constant γ appearing in (2.11).

(VII) The constant γ in hypothesis (II) satisfies the condition

$$(1+\gamma)\mu^{-k/d} < 1.$$

LEMMA 4.1. *Suppose that conditions* (I)–(IV), (VI), *and* (VII) *hold. If* $v \in W^{k,\infty}(E)$, *then there exists a positive constant* c *such that*

$$|v_{ij}| \le c\mu^{-ki/d}\|v\|_{k,\infty}, \quad (i,j) \in \mathbb{U}_n. \tag{4.1}$$

Proof. For $v \in W^{k,\infty}(E)$, we write

$$\mathcal{P}_n v = \sum_{(i,j)\in\mathbb{U}_n} v_{ij} w_{ij}$$

and denote $\mathbf{v} := [v_{ij} : (i,j) \in \mathbb{U}_n]$. By the definition of the projection $\mathcal{P}_n$, we have that

$$\mathbf{E}_n\mathbf{v} = \left[\left\langle \ell_{ij}, \sum_{(i',j')\in\mathbb{U}_n} v_{i'j'} w_{i'j'} \right\rangle : (i,j) \in \mathbb{U}_n\right] = [\langle \ell_{ij}, v\rangle : (i,j) \in \mathbb{U}_n].$$

Meanwhile, using Lemma 2.1 with $\nu := \mu^{-k/d}$ and condition (VII), we conclude that

$$\|\mathbf{v}\|_{\mu^{-k/d}} \le c\|\mathbf{E}_n\mathbf{v}\|_{\mu^{-k/d}},$$

where

$$c := \frac{1-\mu^{-k/d}}{1-(1+\gamma)\mu^{-k/d}} > 0$$

is a constant. Hence,

$$\|\mathbf{v}\|_{\mu^{-k/d}} \le c \max_{(i,j)\in\mathbb{U}_n} |\mu^{ik/d} \langle \ell_{ij}, v\rangle |. \tag{4.2}$$

On the other hand, recalling that the "support" of the functional ℓ_{ij} is the set $\hat{S}_{ij} \subseteq S_{ij}$, we use the Taylor theorem with remainder on the set S_{ij} for $v \in W^{k,\infty}(E)$ and conditions (III), (IV), and (VI) to conclude that there exists a positive constant c such that

$$| \langle \ell_{ij}, v\rangle | \le c d_i^k \|v\|_{k,\infty} \le c\mu^{-ki/d}\|v\|_{k,\infty}.$$

Combining this inequality with (4.2) we obtain the estimate

$$\|\mathbf{v}\|_{\mu^{-k/d}} \le c\|v\|_{k,\infty}.$$

Again, using the definition of the weighted norms, we have that

$$\|\mathbf{v}_i\|_\infty \le c\mu^{-ki/d}\|v\|_{k,\infty},$$

which proves the estimate of this lemma. □

Lemma 4.1 ensures that for a function $v \in W^{k,\infty}(E)$ the coefficients of its expansion in basis W_n and functionals L_n decay in order $\mathcal{O}(\mu^{-ik/d})$. This is an extension of a well-known result for orthogonal wavelets to the interpolating wavelets constructed in this paper.

For any real numbers α and β, we make use of the notation

$$\mu[\alpha,\beta;n] := \sum_{i\in\mathbb{Z}_{n+1}} \mu^{\alpha i/d} \sum_{i'\in\mathbb{Z}_{n+1}} \mu^{\beta i'/d}$$

to state the next lemma, which will play an important role in the analysis for the order of convergence and stability of the multiscale collocation method. To prove the

next lemma, we need to estimate the $L^\infty(E)$ norm of a typical element in $\mathbb{F}_n$ given by

$$v := \sum_{(i,j)\in\mathbb{U}_n} v_{ij} w_{ij} \tag{4.3}$$

by the norm $\|\mathbf{v}\|_\infty$ of its coefficients $\mathbf{v} := [v_{ij} : (i,j) \in \mathbb{U}_n]^T$. Specifically, we require the following condition.

(VIII) There exist positive constants θ_2 and θ_3 such that for all $n \in \mathbb{N}_0$ and v having form (4.3)

$$\theta_2 \|\mathbf{v}\|_\infty \le \|v\|_\infty \le \theta_3 (n+1) \|\mathbf{E}_n \mathbf{v}\|_\infty. \tag{4.4}$$

One way to satisfy this hypothesis is to consider the sequence of functions $\{\zeta_{ij} : (i,j) \in \mathbb{U}\}$ defined by the equation

$$\zeta_{ij} := \sum_{(i',j')\in\mathbb{U}} (\mathbf{E}^{-1})_{i'j',ij} w_{i'j'}, \quad (i,j) \in \mathbb{U}.$$

These functions are biorthogonal relative to the set of linear functionals $\{\ell_{ij} : j \in \mathbb{Z}_{w(i)}, i \in \mathbb{N}_0\}$, that is,

$$\langle \ell_{i'j'}, \zeta_{ij} \rangle = \delta_{ii'} \delta_{jj'}, \quad (i,j), (i',j') \in \mathbb{U}.$$

If in addition for all $i \in \mathbb{N}_0$

$$\sup_{t\in E} \sum_{j\in\mathbb{Z}_{w(i)}} |\zeta_{ij}(t)| \le \theta_3, \tag{4.5}$$

then the second inequality of (4.4) follows.

In the next lemma, we estimate the difference of operators $\mathcal{K}_n$ and $\tilde{\mathcal{K}}_n$ applying to $\mathcal{P}_n v$. It is an important step for both stability analysis and the convergence estimate.

LEMMA 4.2. *Suppose that conditions* (I)–(VIII) *hold,* $0 < \sigma' < \min\{2k, d - \sigma\}$, *and* $\eta := 2k - \sigma'$. *Let* b *and* b' *be real numbers, and let the truncation parameters* $\epsilon^n_{i'i}$, $i', i \in \mathbb{Z}_{n+1}$, *be chosen such that*

$$\epsilon^n_{i'i} \ge \max\{a\mu^{[-n+b(n-i)+b'(n-i')]/d}, r(d_i + d_{i'})\}, \quad i, i' \in \mathbb{Z}_{n+1},$$

for some constants $a > 0$ *and* $r > 1$. *Then there exists a positive constant* c *independent of* n *such that for any* $v \in W^{k,\infty}(E)$,

$$\|(\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n v\|_\infty \le c\mu[2k - b\eta, k - b'\eta; n](n+1)\mu^{-(k+\sigma')n/d}\|v\|_{k,\infty} \tag{4.6}$$

and for $v \in L^\infty(E)$,

$$\|(\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n v\|_\infty \le c\mu[k - b\eta, k - b'\eta; n](n+1)\mu^{-\sigma' n/d}\|v\|_\infty. \tag{4.7}$$

Proof. Since

$$\mathcal{P}_n v = \sum_{(i,j)\in\mathbb{U}_n} v_{ij} w_{ij},$$

we conclude that

$$(\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n v = \sum_{(i,j)\in\mathbb{U}_n} h_{ij} w_{ij},$$

where

$$\mathbf{h} := \mathbf{E}_n^{-1}(\mathbf{K}_n - \tilde{\mathbf{K}}_n)\mathbf{v}.$$

Thus, by hypothesis (VIII), we conclude that

$$\|(\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n v\|_\infty \le \theta_3(n+1)\|(\mathbf{K}_n - \tilde{\mathbf{K}}_n)\mathbf{v}\|_\infty. \tag{4.8}$$

We next estimate $\|(\mathbf{K}_n - \tilde{\mathbf{K}}_n)\mathbf{v}\|_\infty$. To this end, we introduce the matrix

$$\mathbf{\Delta}_n := [\Delta_{i'j',ij}]_{f(n)\times f(n)},$$

whose elements are given by

$$\Delta_{i'j',ij} := \nu\mu^{k(n-i)/d+\sigma' n/d}(K_{i'j',ij} - \tilde{K}_{i'j',ij}), \quad (i,j),(i',j') \in \mathbb{U}_n,$$

where $\nu = 1/\mu[2k - b\eta, k - b'\eta; n]$ and the vector

$$\mathbf{v}' := [v'_{ij} : (i,j) \in \mathbb{U}_n]$$

whose components are

$$v'_{ij} := \mu^{ki/d} v_{ij}, \quad (i,j) \in \mathbb{U}_n.$$

In this notation, we observe that

$$\|(\mathbf{K}_n - \tilde{\mathbf{K}}_n)\mathbf{v}\|_\infty \le \nu^{-1}\mu^{-(k+\sigma')n/d}\|\mathbf{\Delta}_n\|_\infty\|\mathbf{v}'\|_\infty. \tag{4.9}$$

By Lemma 4.1 we have that there exists a positive constant c such that for $v \in W^{k,\infty}(E)$,

$$\|\mathbf{v}'\|_\infty \le c\|v\|_{k,\infty}. \tag{4.10}$$

On the other hand, from Lemma 3.2, there exists a positive constant c such that

$$\begin{aligned}\sum_{(i,j)\in\mathbb{U}_n} \Delta_{i'j',ij} &\le \nu \sum_{i\in\mathbb{Z}_{n+1}} \mu^{k(n-i)/d+\sigma' n/d}\|\mathbf{K}_{i'i} - \tilde{\mathbf{K}}_{i'i}\|_\infty \\ &\le c\nu \sum_{i\in\mathbb{Z}_{n+1}} \mu^{k(n-i)/d+\sigma' n/d-k(i+i')/d}(\epsilon_{i'i}^n)^{-\eta}.\end{aligned}$$

Consequently, by the choice of $\epsilon_{i'i}$, we conclude that

$$\|\mathbf{\Delta}_n\|_\infty := \max_{(i',j')\in\mathbb{U}_n} \sum_{(i,j)\in\mathbb{U}_n} \Delta_{i'j',ij} \le c. \tag{4.11}$$

Combining (4.9)–(4.11) yields the first estimate.

To prove the second estimate, we proceed similarly and introduce the matrix

$$\mathbf{\Delta'}_n := [\Delta'_{i'j',ij}]_{f(n)\times f(n)},$$

whose entries are given by

$$\Delta'_{i'j',ij} := \nu' \mu^{\sigma' n/d}(K_{i'j',ij} - \tilde{K}_{i'j',ij}), \quad (i,j),(i',j') \in \mathbb{U}_n,$$

where $\nu' = 1/\mu[k - b\eta, k - b'\eta; n]$. With these quantities, we have the estimate

$$\|(\mathbf{K}_n - \tilde{\mathbf{K}}_n)\mathbf{v}\|_\infty \le (\nu')^{-1}\mu^{-\sigma' n/d}\|\mathbf{\Delta}'_n\|_\infty \|\mathbf{v}\|_\infty. \tag{4.12}$$

Condition (VIII) provides a positive constant c such that for $v \in L^\infty(E)$,

$$\|\mathbf{v}\|_\infty \le c\|v\|_\infty. \tag{4.13}$$

As before, Lemma 3.2 and the choice of $\epsilon_{i'i}$, $i, i' \in \mathbb{Z}_{n+1}$, ensure that there exists a positive constant c such that

$$\|\mathbf{\Delta}'_n\|_\infty \le c. \tag{4.14}$$

Combining this inequality with (4.12)–(4.13) yields the second estimate. □

We now turn our attention to the stability of the multiscale collocation method. For this purpose, we require the next hypothesis.

(IX) The operator $\mathcal{P}_n$ converges pointwise to the identity operator $\mathcal{I}$ in $L^\infty(E)$ as $n \to \infty$. In other words, for any $x \in L^\infty(E)$, there holds

$$\lim_{n\to\infty} \|\mathcal{P}_n x - x\|_\infty = 0.$$

Condition (IX) follows trivially if $\mathbb{F}_n$ is a space of piecewise polynomials. Because of this property and the fact that $\mathcal{K}$ is compact, we conclude for sufficiently large n that the operators $(\mathcal{I} - \mathcal{K}_n)^{-1}$ exist and are uniformly bounded in $L^\infty(E)$ (see, for example, [An, At2]). From this fact follows the stability estimate; that is, there exists a positive constant ρ and a positive integer m such that for $n \ge m$ and $x \in \mathbb{F}_n$ there holds that

$$\|(\mathcal{I} - \mathcal{K}_n)x\|_\infty \ge \rho\|x\|_\infty.$$

We shall establish a similar estimate for $\mathcal{I} - \tilde{\mathcal{K}}_n$.

THEOREM 4.3. *Suppose that* $0 < \sigma' < \min\{2k, d - \sigma\}$ *and* $\eta := 2k - \sigma'$. *If the conditions* (I)–(IX) *hold and* $\epsilon^n_{i'i}$, $i, i' \in \mathbb{Z}_{n+1}$, *are chosen as in Lemma* 4.2 *with*

$$b > \frac{k - \sigma'}{\eta}, \quad b' > \frac{k - \sigma'}{\eta}, \quad b + b' > 1,$$

then there exists a positive constant c *and a positive integer* m *such that when* $n \ge m$ *and* $x \in \mathbb{F}_n$,

$$\|(\mathcal{I} - \tilde{\mathcal{K}}_n)x\|_\infty \ge c\|x\|_\infty.$$

Proof. Note that for any real numbers α, β, and e,

$$\lim_{n\to\infty} \mu[\alpha, \beta; n](n+1)\mu^{-en/d} = 0$$

when $e > \max\{0, \alpha, \beta, \alpha+\beta\}$. Thus, our hypothesis ensures that there exists a positive integer m such that when $n \ge m$,

$$c\mu[k - b\eta, k - b'\eta; n](n+1)\mu^{-\sigma' n/d} < \rho/2, \tag{4.15}$$

where the constant c is the one appearing in (4.7). The stability of the collocation scheme and the second estimate in Lemma 4.2, together with (4.15), yield for $x \in \mathbb{F}_n$ that

$$\|(\mathcal{I} - \tilde{\mathcal{K}}_n)x\|_\infty \geq \|(\mathcal{I} - \mathcal{K}_n)x\|_\infty - \|(\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n x\|_\infty \geq \frac{\rho}{2}\|x\|_\infty.$$

This completes the proof. □

In particular, this theorem ensures for $n \geq m$ that the equation

$$(\mathcal{I} - \tilde{\mathcal{K}}_n)\tilde{u}_n = \mathcal{P}_n f \tag{4.16}$$

has a unique solution given by

$$\tilde{u}_n := \sum_{(i,j)\in\mathbb{U}_n} \tilde{u}_{ij} w_{ij}.$$

This equation is equivalent to the matrix equation

$$(\mathbf{E}_n - \tilde{\mathbf{K}}_n)\tilde{\mathbf{u}}_n = \mathbf{f}_n,$$

where $\tilde{\mathbf{u}}_n = [\tilde{u}_{ij} : (i,j) \in \mathbb{U}_n]^T$. The next theorem provides error bounds for $\|u - \tilde{u}_n\|_\infty$. For this purpose, we introduce the next condition.

(X) There exists a positive constant c such that for $u \in W^{k,\infty}(E)$

$$\operatorname{dist}(u, \mathbb{F}_n) \leq c\mu^{-kn/d}\|u\|_{k,\infty}.$$

When $\mathbb{F}_n$ contains the piecewise polynomials of order k, the estimate in condition (X) follows directly from (2.2) and condition (VI).

THEOREM 4.4. *Suppose that conditions* (I)–(X) *hold and that* $0 < \sigma' < \min\{2k, d - \sigma\}$ *and* $\eta := 2k - \sigma'$. *Let* $\epsilon^n_{i'i}$, $i, i' \in \mathbb{Z}_{n+1}$, *be chosen as in Lemma* 4.2 *with* b *and* b' *satisfying one of the following three conditions:*

(i) $b > 1$, $b' > \frac{k-\sigma'}{\eta}$, $b + b' > 1 + \frac{k}{\eta}$.

(ii) $b = 1$, $b' > \frac{k-\sigma'}{\eta}$, $b + b' \geq 1 + \frac{k}{\eta}$; $b > 1$, $b' = \frac{k-\sigma'}{\eta}$, $b + b' > 1 + \frac{k}{\eta}$; *or* $b > 1$, $b' = \frac{k-\sigma'}{\eta}$, $b + b' = 1 + \frac{k}{\eta}$.

(iii) $b = 1$, $b' = \frac{k}{\eta}$; *or* $b = \frac{2k}{\eta}$, $b' = \frac{k-\sigma'}{\eta}$.

Then there exists a positive constant c *and positive integer* m *such that for all* $n \geq m$,

$$\|u - \tilde{u}_n\|_\infty \leq c f(n)^{-k/d}(\log f(n))^\tau \|u\|_{k,\infty},$$

where $\tau = 0$ *in case* (i), $\tau = 1$ *in case* (ii), *and* $\tau = 2$ *in case* (iii).

Proof. It follows from Theorem 4.3 that there exists a positive constant c such that

$$\|u - \tilde{u}_n\|_\infty \leq \|u - \mathcal{P}_n u\|_\infty + c\|(\mathcal{I} - \tilde{\mathcal{K}}_n)(\mathcal{P}_n u - \tilde{u}_n)\|_\infty. \tag{4.17}$$

Using the equation

$$\mathcal{P}_n(\mathcal{I} - \mathcal{K})u = (\mathcal{I} - \tilde{\mathcal{K}}_n)\tilde{u}_n,$$

we find that

$$(\mathcal{I} - \tilde{\mathcal{K}}_n)(\mathcal{P}_n u - \tilde{u}_n) = \mathcal{P}_n(\mathcal{I} - \mathcal{K})(\mathcal{P}_n u - u) + (\mathcal{K}_n - \tilde{\mathcal{K}}_n)\mathcal{P}_n u. \tag{4.18}$$

From (4.17), (4.18), hypothesis (IX), and Lemma 4.2, there exist positive constants c, p such that

$$\|u - \tilde{u}_n\|_\infty \le (1 + p\|\mathcal{I} - \mathcal{K}\|)\|\mathcal{P}_n u - u\|_\infty + c\mu' \mu^{-kn/d}\|u\|_{k,\infty},$$

where

$$\mu' := \mu[2k - b\eta, k - b'\eta; n](n+1)\mu^{-\sigma' n/d}.$$

We estimate each term separately. For the first term, we note that conditions (IX) and (X) provide a positive constant c such that

$$\|\mathcal{P}_n u - u\|_\infty \le c\mu^{-kn/d}\|u\|_{k,\infty}.$$

Now we turn our attention to estimating the quantity μ'. To this end, we observe for any real numbers α, β, and e with $e > 0$, the asymptotic order

$$\mu[\alpha, \beta; n](n+1)\mu^{-en/d} = \begin{cases} o(1) & \text{if } \ e > \max\{\alpha, \beta, \alpha + \beta\}, \\ \mathcal{O}(n) & \text{if } \ \alpha = e, \beta < e, \alpha + \beta < e \\ & \text{or if } \ \alpha < e, \beta = e, \alpha + \beta < e \\ & \text{or if } \ \alpha < e, \beta < e, \alpha + \beta = e, \\ \mathcal{O}(n^2) & \text{if } \ \alpha = 0, \beta = e \ \text{ or if } \ \alpha = e, \beta = 0, \end{cases}$$

as $n \to \infty$. Using this fact with $\alpha := 2k - b\eta$, $\beta := k - b'\eta$, and $e := \sigma'$, we conclude that

$$\mu' = \begin{cases} o(1) & \text{in case (i)}, \\ \mathcal{O}(n) & \text{in case (ii)}, \\ \mathcal{O}(n^2) & \text{in case (iii)}, \end{cases}$$

which establishes the result of this theorem by noting that $n = \log f(n)$. ∎

We see from this theorem that the convergence order of the approximate solution $\tilde{u}$ obtained from the truncated collocation method is almost optimal.

We next estimate the condition number of the matrix $\tilde{\mathbf{A}}_n := \mathbf{E}_n - \tilde{\mathbf{K}}_n$.

THEOREM 4.5. *If the conditions of Theorem* 4.3 *hold, then there exists a positive constant c such that the condition number of the matrix $\tilde{\mathbf{A}}_n$ satisfies the estimate*

$$\mathrm{cond}_\infty(\tilde{\mathbf{A}}_n) \le c \log^2(f(n)),$$

where $\mathrm{cond}_\infty(\mathbf{A})$ *denotes the condition number of a matrix* $\mathbf{A}$ *in the* ℓ^∞ *matrix norm.*

Proof. For any $\mathbf{v} := [v_{ij} : (i,j) \in \mathbb{U}_n]^T \in \mathbb{R}^{f(n)}$, we define the vector $\mathbf{g} := [g_{ij} : (i,j) \in \mathbb{U}_n]^T \in \mathbb{R}^{f(n)}$ by the equation

$$\tilde{\mathbf{A}}_n \mathbf{v} = \mathbf{g} \tag{4.19}$$

and the function

$$g := \sum_{(i,j) \in \mathbb{U}_n} g_{ij} \zeta_{ij}.$$

Therefore, we have that

$$g_{ij} = \langle \ell_{ij}, g \rangle = \langle \ell_{ij}, \mathcal{P}_n g \rangle, \quad (i,j) \in \mathbb{U}_n.$$

It follows from (IV) that

$$\|\tilde{\mathbf{A}}_n \mathbf{v}\|_\infty \le \theta_0 \|\mathcal{P}_n g\|_\infty. \tag{4.20}$$

Let

$$v := \sum_{(i,j)\in \mathbb{U}_n} v_{ij} w_{ij},$$

and observe the equation

$$(\mathcal{I} - \tilde{\mathcal{K}}_n) v = \mathcal{P}_n g. \tag{4.21}$$

We conclude from (4.20) and (4.21) that there exists a positive constant c such that

$$\begin{aligned} \|\tilde{\mathbf{A}}_n \mathbf{v}\|_\infty &\le \theta_0 \|(\mathcal{I} - \tilde{\mathcal{K}}_n) v\|_\infty \\ &\le \theta_0 (\|(\mathcal{I} - \mathcal{K}_n) v\|_\infty + \|(\mathcal{K}_n - \tilde{\mathcal{K}}_n) v\|_\infty) \\ &\le c \|v\|_\infty, \end{aligned}$$

where the last inequality holds because of (4.7) and (4.15). Next, appealing to hypotheses (I) and (IV), we observe for any $t \in E$ and $i \in \mathbb{Z}_{n+1}$ that

$$\left| \sum_{j \in \mathbb{Z}_{w(i)}} v_{ij} w_{ij}(t) \right| \le \rho \theta_0 \|\mathbf{v}\|_\infty,$$

because there are at most ρ value of $j \in \mathbb{Z}_{w(i)}$ such that functions $w_{ij}(t) \neq 0$. Therefore, we conclude that

$$\|v\|_\infty \le \rho \theta_0 (n+1) \|\mathbf{v}\|_\infty. \tag{4.22}$$

Consequently, there exists a positive constant c such that

$$\|\tilde{\mathbf{A}}_n\|_\infty \le c(n+1). \tag{4.23}$$

Conversely, for any $\mathbf{g} \in \mathbb{R}^{f(n)}$, there exists a vector $\mathbf{v} \in \mathbb{R}^{f(n)}$ such that (4.19) holds. Similar to (4.22), we argue that there exists a positive constant c such that

$$\|g\|_\infty \le c(n+1) \|\mathbf{g}\|_\infty.$$

Hence, we obtain from condition (VIII) the inequality

$$\|\mathbf{v}\|_\infty \le c\|v\|_\infty \le c\|(\mathcal{I} - \tilde{\mathcal{K}}_n) v\|_\infty = c\|g\|_\infty \le c(n+1)\|\mathbf{g}\|_\infty$$

from which it follows that there exists a positive constant c such that

$$\|\tilde{\mathbf{A}}_n^{-1}\|_\infty \le c(n+1). \tag{4.24}$$

Recalling hypothesis (VI) we combine the estimates (4.23) and (4.24) to obtain the desired result, namely

$$\mathrm{cond}_\infty(\tilde{\mathbf{A}}_n) = \mathcal{O}\left((n+1)^2\right) = \mathcal{O}\left(\log^2(f(n))\right), \quad n \to \infty. \qquad \square$$

In the remainder of this section, we will estimate the number of nonzero entries of matrix $\tilde{\mathbf{A}}_n$, which shows that the truncation strategy embodied in Lemma 4.2 can

lead to a fast numerical algorithm for solving (2.12) while preserving nearly optimal order of convergence. For any matrix $\mathbf{A}$, we denote by $\mathcal{N}(\mathbf{A})$ the number of nonzero entries in $\mathbf{A}$. The proof of the following theorem closely follows a similar result in [MXZ].

THEOREM 4.6. *Suppose that hypotheses* (I) *and* (VI) *hold. Let b and b' be real numbers not larger than one and the truncation parameters $\epsilon^n_{i'i}$, $i', i \in \mathbb{Z}_{n+1}$, be chosen such that*

$$\epsilon^n_{i'i} \le \max\{a\mu^{[-n+b(n-i)+b'(n-i')]/d}, r(d_i + d_{i'})\}, \quad i, i' \in \mathbb{Z}_{n+1},$$

for some constants $a > 0$ and $r > 1$. Then

$$\mathcal{N}(\tilde{\mathbf{A}}_n) = \mathcal{O}(f(n)\log^\tau f(n)),$$

where $\tau = 1$ except for $b = b' = 1$, in which case $\tau = 2$.

Proof. We first estimate the number $\mathcal{N}(\tilde{\mathbf{A}}_{i'i})$. For fixed i, i', and j', if $\tilde{A}_{i'j',ij} \neq 0$, then $\operatorname{dist}(S_{i'j'}, S_{ij}) \le \epsilon^n_{i'i}$, so that

$$S_{ij} \subseteq S(i,i') := \{v : v \in \mathbb{R}^d, |v - v_0| \le d_i + d_{i'} + \epsilon^n_{i'i}\},$$

where v_0 is an arbitrary point in the set $S_{i'j'}$. Let $\mathcal{N}_{i,i'j'}$ be the number of such sets which are contained in $S(i,i')$. Using condition (VI) we conclude that there exists a positive constant c such that

$$\mathcal{N}_{i,i'j'} \le \frac{\operatorname{meas}(S(i,i'))}{\min\{\operatorname{meas}(S_{ij}) : S_{ij} \subseteq S(i,i')\}} \le c\mu^i(d_i + d_{i'} + \epsilon^n_{i'i})^d.$$

Next, we invoke condition (I) to conclude that the number of functions w_{ij} having support contained in S_{ij} is bounded by ρ, and appealing to condition (VI) we have that $w(i') = \mathcal{O}(\mu^{i'})$, $i' \to \infty$. Consequently, there exists a positive constant c such that

$$\mathcal{N}(\tilde{\mathbf{A}}_{i'i}) \le \rho \sum_{j' \in \mathbb{Z}_{w(i')}} N_{i,i'j'} \le c\mu^{i+i'}(d_i + d_{i'} + \epsilon^n_{i'i})^d, \quad i, i' \in \mathbb{Z}_{n+1},$$

from which it follows that

$$\mathcal{N}(\tilde{\mathbf{A}}_n) \le c \sum_{i,i' \in \mathbb{Z}_{n+1}} \mu^{i+i'} \left[(d_i)^d + (d_{i'})^d + (\epsilon^n_{i'i})^d\right].$$

This inequality and conditions (I) imply that if the truncation parameters have the bound

$$\epsilon^n_{i'i} \le a\mu^{[-n+b(n-i)+b'(n-i')]/d},$$

then

$$\begin{aligned}
\mathcal{N}(\tilde{\mathbf{A}}_n) &\le c \sum_{i' \in \mathbb{Z}_{n+1}} \sum_{i \in \mathbb{Z}_{n+1}} \mu^{i+i'} \left(\mu^{-i} + \mu^{-i'} + a^d \mu^{-n+b(n-i)+b'(n-i')}\right) \\
&\le c \left[2(n+1) \sum_{i \in \mathbb{Z}_{n+1}} \mu^i + a^d \mu^n \left(\sum_{i \in \mathbb{Z}_{n+1}} \mu^{(b-1)(n-i)}\right)\left(\sum_{i' \in \mathbb{Z}_{n+1}} \mu^{(b'-1)(n-i')}\right)\right] \\
&= \mathcal{O}\left(\mu^n (n+1)^\tau\right) = \mathcal{O}\left(f(n)\log^\tau f(n)\right),
\end{aligned}$$

as $n \to \infty$. If $\epsilon^n_{i'i} \le r(d_i + d_{i'})$, a similar argument leads to

$$\mathcal{N}(\tilde{\mathbf{A}}_n) = \mathcal{O}\left(f(n)\log f(n)\right), \quad n \to \infty.$$

This completes the proof. □

It follows from Theorems 4.3–4.6 that for the truncation scheme to have *all* the desired properties of stability, convergence, and complexity, we have to choose the truncation parameters to satisfy the equation

$$\epsilon^n_{i'i} = \max\{a\mu^{[-n+b(n-i)+b'(n-i')]/d}, r(d_i + d_{i'})\}, \quad i, i' \in \mathbb{Z}_{n+1},$$

with $b = 1$, $b' > \frac{k-\sigma}{\eta}$, $b + b' \ge 1\frac{k}{\eta}$ or with $b = 1$, $b' = \frac{k}{\eta}$, $\sigma' < k$.

5. A concrete construction of multiscale functions and functionals. Whenever the basis $W := \{w_{ij} : (i,j) \in \mathbb{U}\}$ and the collocation functionals $L := \{\ell_{ij} : (i,j) \in \mathbb{U}\}$ satisfy hypotheses (I)–(IV) and (VI)–(X), the results of the last three sections present convergence, complexity, and stability estimates for the truncated collocation method. In this section, we illustrate by an example of practical importance multiscale functions and functionals which satisfy all these hypotheses. Our point of view here is based on the notion of iterated function systems which we have recently investigated in the context of the numerical solutions of integral equations [MX1, MX2, CMX2, MSX2]. In our example below, the solution spaces will be piecewise polynomials on a multiscale partition. Let us first describe the method we use to generate a multiscale partition of an invariant set E.

We start with a positive integer μ and a family $\Phi := \{\phi_e : e \in \mathbb{Z}_\mu\}$ of contractive affine mappings on $\mathbb{R}^d$. There exists a unique compact subset E of $\mathbb{R}^d$ such that

$$\Phi(E) = E, \tag{5.1}$$

where

$$\Phi(E) := \bigcup_{e \in \mathbb{Z}_\mu} \phi_e(E)$$

(see [H]). This set E is called the invariant set associated with the family of mappings Φ. Generally, it has a complex *fractal* structure. For example, there are choices of Φ for which E is the Cantor subset of $[0,1]$, the Sierpinski gasket contained in an equilateral triangle, or the twin dragons from wavelet analysis. We are interested in the cases when E has a simple structure which include, for example, the cube and simplex in $\mathbb{R}^d$. With these cases in mind, we make the following additional restriction on the family of mappings Φ.

(a) For every $e \in \mathbb{Z}_\mu$, the mapping ϕ_e has a continuous inverse on E.

(b) The set E has nonempty interior and

$$\text{meas}\,(\phi_e(E) \cap \phi_{e'}(E)) = 0, \quad e, e' \in \mathbb{Z}_\mu, e \ne e'.$$

We use Φ to obtain a multiscale partition $\{E_n : n \in \mathbb{N}_0\}$ of the set E in the following way. Given any $\mathbf{e} := (e_0, e_1, \ldots, e_{n-1}) \in \mathbb{Z}^n_\mu := \mathbb{Z}_\mu \times \cdots \times \mathbb{Z}_\mu$, n times, we define the mappings

$$\phi_{\mathbf{e}} := \phi_{e_0} \circ \phi_{e_1} \circ \cdots \circ \phi_{e_{n-1}}$$

and the number

$$\mu(\mathbf{e}) := \mu^{n-1}e_0 + \cdots + \mu e_{n-2} + e_{n-1}.$$

Note that every $i \in \mathbb{Z}_{\mu^n}$ can be uniquely written as $i = \mu(\mathbf{e})$ for some $\mathbf{e} \in \mathbb{Z}_\mu^n$. From (5.1) and conditions (a) and (b) it follows that the collection of sets

$$E_n := \{E_{n,\mathbf{e}} : E_{n,\mathbf{e}} = \phi_{\mathbf{e}}(E),\ \mathbf{e} \in \mathbb{Z}_\mu^n\}$$

forms a partition of E. We require that this partition has the following property:

(c) There exist positive constants c_-, c_+ such that for all $n \in \mathbb{N}_0$

$$c_- \mu^{-n/d} \le \max\{d(E_{n,\mathbf{e}}) : \mathbf{e} \in \mathbb{Z}_\mu^n\} \le c_+ \mu^{-n/d}. \tag{5.2}$$

This requirement is fulfilled when the Jacobi of the contractive affine mappings ϕ_e, $e \in \mathbb{Z}_\mu$, have the property that

$$J_{\phi_e} \sim \mathcal{O}(\mu^{-1}).$$

In fact, for any $s, t \in \phi_e(E)$, there exist $\hat{s}, \hat{t} \in E$ such that $s = \phi_e(\hat{s})$ and $t = \phi_e(\hat{t})$, and thus we have that

$$|s - t| \sim (J_{\phi_e})^{1/d} |\hat{s} - \hat{t}|.$$

This with the hypothesis on the Jacobi of the mappings ensures that for any $e \in \mathbb{Z}_\mu$

$$d(E_{1,e}) \sim \mathcal{O}(\mu^{-1/d}),$$

and by induction we find that for any $\mathbf{e} \in \mathbb{Z}_\mu^n$,

$$d(E_{n,\mathbf{e}}) \sim \mathcal{O}(\mu^{-n/d}).$$

On the partition E_n, we consider piecewise polynomials. Choose a positive integer k and let $\mathbb{F}_n$ be the spaces of all functions such that their restriction to any cell $E_{n,\mathbf{e}}$, $\mathbf{e} \in \mathbb{Z}_\mu^n$, is a polynomial of total degree $\le k - 1$. Here we use the convention that for $n = 0$ the set E is the only cell in the partition and so

$$m := \dim \mathbb{F}_0 = \begin{pmatrix} k + d - 1 \\ d \end{pmatrix}.$$

We must generate a suitable multiscale decomposition of $\mathbb{F}_n$. To this end, let

$$G_0 = \{t_j : j \in \mathbb{Z}_m\}$$

be a finite set of distinct points in E, which is refinable relative to the mappings Φ; that is, G_0 satisfies

$$G_0 \subseteq \Phi(G_0).$$

Set

$$G_1 := \Phi(G_0), \quad V_1 := G_1 \setminus G_0 = \{t_{m+j} : j \in \mathbb{Z}_r\}$$

with $r := (\mu - 1)m$. Now, we require that there exists a basis of elements in $\mathbb{F}_0$, denoted by $\psi_0, \psi_1, \ldots, \psi_{m-1}$ such that

$$\mathbb{F}_0 := \operatorname{span}\{\psi_j : j \in \mathbb{Z}_m\},$$

and they satisfy Lagrange interpolation conditions

$$\psi_i(t_j) = \delta_{i,j}, \quad i,j \in \mathbb{Z}_m. \tag{5.3}$$

A construction of refinable points $\{t_j : j \in \mathbb{Z}_m\} \in E$ that admits a unique d-dimensional Lagrange interpolation is presented in [MSX1].

With this basis of $\mathbb{F}_0$ at hand, we will generate a multiscale basis for $\mathbb{F}_0$ in the following way. For this purpose, we introduce linear operators $\mathcal{T}_e : \mathbb{X} \to \mathbb{X}$, $e \in \mathbb{Z}_\mu$, defined by

$$(\mathcal{T}_e x)(t) := x(\phi_e^{-1}(t))\chi_{\phi_e(E)}(t),$$

where χ_S denotes the characteristic function of some set S, and observe that $\|\mathcal{T}_e\| = 1$. Therefore, it follows that

$$\mathbb{F}_n = \bigoplus_{e \in \mathbb{Z}_\mu} \mathcal{T}_e \mathbb{F}_{n-1}, \quad n \in \mathbb{N},$$

where $\mathbb{A} \oplus \mathbb{B}$ denotes the direct sum of the spaces $\mathbb{A}$ and $\mathbb{B}$. The functions $\psi_{m+j} \in \mathbb{F}_1$, $j \in \mathbb{Z}_r$, satisfying

$$\psi_{m+j}(t_i) = 0, \quad i \in \mathbb{Z}_m, j \in \mathbb{Z}_r, \quad \psi_{m+j}(t_{m+j'}) = \delta_{jj'}, \quad j,j' \in \mathbb{Z}_r \tag{5.4}$$

with ψ_j, $j \in \mathbb{Z}_m$, defined by (5.3), form a basis for $\mathbb{F}_1$.

We require another basis for $\mathbb{F}_1$ consisting of functions with vanishing moments. To this end, we set

$$w_{0j} := \psi_j, \quad j \in \mathbb{Z}_m.$$

We set $q := m + r$ and for $j \in \mathbb{Z}_r$ find a vector $[c_{js} : s \in \mathbb{Z}_q]^T \in \mathbb{R}^q$ such that

$$w_{1j} := \sum_{s \in \mathbb{Z}_q} c_{js}\psi_s, \quad j \in \mathbb{Z}_r, \tag{5.5}$$

satisfies the equation

$$(w_{1j}, \psi_{j'}) = 0, \quad j' \in \mathbb{Z}_m, \ j \in \mathbb{Z}_r. \tag{5.6}$$

Since for each $j \in \mathbb{Z}_r$, (5.6) is a linear system of rank m with m equations and q unknowns, there exist r linearly independent solutions of this system which we denote by $w_{1j}, j \in \mathbb{Z}_r$. These functions form a basis for the space $\mathbb{W}_1$.

Let us now turn our attention to a construction of collocation functionals. We begin by defining

$$\ell_{0j} := \delta_{t_j}, \quad j \in \mathbb{Z}_m,$$

and for $j' \in \mathbb{Z}_r$, we find the vector $[c'_{j's} : s \in \mathbb{Z}_q]$ such that

$$\ell_{1j'} := \sum_{s \in \mathbb{Z}_q} c'_{j's}\delta_{t_s}, \quad j' \in \mathbb{Z}_r, \tag{5.7}$$

satisfies the equations

$$\langle \ell_{1j'}, w_{0j} \rangle = 0, \quad j \in \mathbb{Z}_m, \quad j' \in \mathbb{Z}_r, \tag{5.8}$$

and

$$\langle \ell_{1j'}, w_{1j} \rangle = \delta_{jj'}, \quad j \in \mathbb{Z}_r, \quad j' \in \mathbb{Z}_r. \tag{5.9}$$

For $j \in \mathbb{Z}_r$, the matrix of order q for this linear system of equations is

$$\mathbf{A} := [\langle \delta_{t_{i'j'}}, w_{ij} \rangle]_{(i,j),(i',j') \in \mathbb{U}_1}.$$

Let us prove that the matrix $\mathbf{A}$ is nonsingular. To this end, we assume that there are constants $a_{ij}, (i,j) \in \mathbb{U}_1$, such that

$$\sum_{(i,j) \in \mathbb{U}_1} a_{ij} \langle \delta_{t_{i'j'}}, w_{ij} \rangle = 0, \quad (i',j') \in \mathbb{U}_1,$$

that is,

$$\left\langle \delta_{t_{i'j'}}, \sum_{(i,j) \in \mathbb{U}_1} a_{ij} w_{ij} \right\rangle = 0, \quad (i',j') \in \mathbb{U}_1.$$

Since the set G_1 is Lagrange admissible relative to $(\Phi, \mathbb{F}_1)$ (cf. [CMX2]), we conclude that

$$\sum_{(i,j) \in \mathbb{U}_1} a_{ij} w_{ij} = 0,$$

and therefore $a_{ij} = 0$, $(i,j) \in \mathbb{U}_1$. This proves that $\mathbf{A}$ is nonsingular.

We find it convenient to write (5.8)–(5.9) in a matrix form. For this purpose, we introduce matrices

$$\tilde{\mathbf{B}} := [\langle \delta_{t_i}, \psi_j \rangle]_{i,j \in \mathbb{Z}_q}, \quad \mathbf{B} := [\langle \delta_{t_{m+i}}, \psi_j \rangle]_{i \in \mathbb{Z}_r, j \in \mathbb{Z}_m},$$

$$\mathbf{C}_1 := [c_{js}]_{j \in \mathbb{Z}_r, s \in \mathbb{Z}_m}, \quad \mathbf{C}_2 := [c_{j,m+s}]_{j \in \mathbb{Z}_r, s \in \mathbb{Z}_r},$$

$$\mathbf{C}'_1 = [c'_{js}]_{j \in \mathbb{Z}_r, s \in \mathbb{Z}_m}, \quad \mathbf{C}'_2 := [c'_{j,m+s}]_{j \in \mathbb{Z}_r, s \in \mathbb{Z}_r},$$

and

$$\mathbf{C} := [\mathbf{C}_1, \mathbf{C}_2], \quad \mathbf{C}' := [\mathbf{C}'_1, \mathbf{C}'_2].$$

The next lemma gives a relationship between the matrices $\mathbf{C}$ and $\mathbf{C}'$.

LEMMA 5.1.

$$\mathbf{C}'_1 = -\mathbf{C}'_2 \mathbf{B}, \quad \mathbf{C}'_2 = (\mathbf{C}_2^T)^{-1}.$$

Proof. It follows from (5.8) and (5.9) that

$$\mathbf{C}' \tilde{\mathbf{B}} [\mathbf{I}_m \mathbf{O}_{m \times r}]^T = \mathbf{O}_{r \times m},$$

and

$$\mathbf{C}' \tilde{\mathbf{B}} \mathbf{C}^T = \mathbf{I}_r,$$

where $\mathbf{O}_{m\times r}$ denotes the $m\times r$ zero matrix and $\mathbf{I}_m$ denotes the $m\times m$ identity matrix. The properties of basis $\{\psi_j : j\in\mathbb{Z}_q\}$ and the functionals $\{\delta_{t_j} : j\in\mathbb{Z}_q\}$ described above in (5.3) and (5.4) imply that

$$\tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{I}_m & \mathbf{O}_{m\times r} \\ \mathbf{B} & \mathbf{I}_r \end{bmatrix},$$

$$\mathbf{C}_1' + \mathbf{C}_2'\mathbf{B} = \mathbf{O}_{r\times m},$$

and

$$(\mathbf{C}_1' + \mathbf{C}_2'\mathbf{B})\mathbf{C}_1^T + \mathbf{C}_2'\mathbf{C}_2^T = \mathbf{I}_r,$$

from which the result follows. □

We next describe the construction of a basis for $\mathbb{W}_i$, $i\in\mathbb{N}$. To this end, for $\mathbf{e} := (e_0,\dots,e_{n-1})\in\mathbb{Z}_\mu^n$ we introduce a composition operator $\mathcal{T}_\mathbf{e}$ by

$$\mathcal{T}_\mathbf{e} := \mathcal{T}_{e_0}\circ\cdots\circ\mathcal{T}_{e_{n-1}}.$$

For $i = 2,3,\dots,n$ we let

$$w_{ij} := \mathcal{T}_\mathbf{e} w_{1l},\quad j=\mu(\mathbf{e})r+l,\quad \mathbf{e}\in\mathbb{Z}_\mu^{i-1},\quad l\in\mathbb{Z}_r, \tag{5.10}$$

and

$$\mathbb{W}_i := \operatorname{span}\{w_{ij} : j\in\mathbb{Z}_{w(i)}\}.$$

Observe that the support of w_{ij} is contained in $S_{ij} := \phi_\mathbf{e}(E)$, $j\in\mathbb{Z}_{w(i)}$.

To generate multiscale collocation functionals, we introduce for any $e\in\mathbb{Z}_\mu$ a linear operator $\mathcal{L}_e : \mathbb{X}^*\to\mathbb{X}^*$ defined by the equation

$$\langle \mathcal{L}_e\ell, v\rangle = \langle \ell, v\circ\phi_e\rangle,\qquad v\in\mathbb{X},\ \ell\in\mathbb{X}^*,$$

and observe that $\|\mathcal{L}_e\| = 1$. Moreover, for $\mathbf{e} := (e_0,\dots,e_{n-1})\in\mathbb{Z}_\mu^n$, we define the composition operator

$$\mathcal{L}_\mathbf{e} := \mathcal{L}_{e_0}\circ\cdots\circ\mathcal{L}_{e_{n-1}}.$$

Consequently, for any $\mathbf{e},\mathbf{e}'\in\mathbb{Z}_\mu^i$, $w\in\mathbb{X}$, and $\ell\in\mathbb{X}^*$, we have that

$$\langle \mathcal{L}_\mathbf{e}\ell, \mathcal{T}_{\mathbf{e}'}w\rangle = \langle \ell, w\rangle\,\delta_{\mathbf{e}\mathbf{e}'}. \tag{5.11}$$

In addition, for $i>1$, $j=\mu(\mathbf{e})r+l$, $\mathbf{e}\in\mathbb{Z}_\mu^{i-1}$, $l\in\mathbb{Z}_r$, we define

$$\ell_{ij} := \mathcal{L}_\mathbf{e}\ell_{1l} \tag{5.12}$$

and observe that

$$\langle \ell_{ij}, v\rangle = \langle \ell_{1l}, v\circ\phi_\mathbf{e}\rangle = \sum_{s\in\mathbb{Z}_q} c_{ls}' v(\phi_\mathbf{e}(t_s)).$$

Note that the "support" of ℓ_{ij} is also contained in S_{ij}.

We partition the matrix $\mathbf{E}_n$ into a block matrix

$$\mathbf{E}_n := [\mathbf{E}_{i'i}]_{i',i\in\mathbb{Z}_{n+1}},$$

where

$$\mathbf{E}_{i'i} := [E_{i'j',ij}]_{j'\in\mathbb{Z}_{w(i')},j\in\mathbb{Z}_{w(i)}},$$

and in the next lemma relate the norm of the matrix $\mathbf{E}_{i'i}$ to that of $\mathbf{E}_{1,i-i'+1}$.

LEMMA 5.2. *If $i', i \in \mathbb{N}$ with $i > i'$, then*

$$\|\mathbf{E}_{i'i}\|_\infty = \|\mathbf{E}_{1,i-i'+1}\|_\infty. \tag{5.13}$$

Proof. From the definition of $\ell_{i'j'}$ and w_{ij}, for $(i',j'),(i,j) \in \mathbb{U}_n$ with $i > i'$, we obtain that there exist $\mathbf{e} \in \mathbb{Z}_\mu^{i-1}$, $\mathbf{e}' \in \mathbb{Z}_\mu^{i'-1}$, $l, l' \in \mathbb{Z}_r$, such that

$$\langle \ell_{i'j'}, w_{ij}\rangle = \langle \mathcal{L}_{\mathbf{e}'}\ell_{1l'}, \mathcal{T}_{\mathbf{e}}w_{1l}\rangle\,.$$

We introduce the vectors

$$\mathbf{e}_1 := (e_0,\dots,e_{i'-2}), \quad \mathbf{e}_2 := (e_{i'-1},\dots,e_{i-2})$$

and conclude from (5.11) that

$$\langle \ell_{i'j'}, w_{ij}\rangle = \langle \ell_{1l'}, \mathcal{T}_{\mathbf{e}_2}w_{1l}\rangle\,\delta_{\mathbf{e}'\mathbf{e}_1}.$$

Let $j_0 := \mu(\mathbf{e}_2)r + l$ and obtain that

$$\langle \ell_{i'j'}, w_{ij}\rangle = \langle \ell_{1l'}, w_{i-i'+1,j_0}\rangle\,\delta_{\mathbf{e}'\mathbf{e}_1}.$$

Consequently, we have that

$$\sum_{j\in\mathbb{Z}_{w(i)}} |\langle \ell_{i'j'}, w_{ij}\rangle| = \sum_{j\in\mathbb{Z}_{w(i-i'+1)}} |\langle \ell_{1l'}, w_{i-i'+1,j}\rangle|\,,$$

which proves the lemma. ∎

Let us use this lemma to estimate the constant γ appearing in condition (2.11).

LEMMA 5.3. *The condition* (2.11) *is satisfied with*

$$\gamma := \max\{\|\mathbf{C}_1\|_1, \|\mathbf{C}'\|_\infty\|\mathbf{C}_1\|_1\}.$$

Proof. By Lemma 5.2, it suffices to prove for $i \in \mathbb{N}$ that

$$\|\mathbf{E}_{0i}\|_\infty \le \|\mathbf{C}_1\|_1 \tag{5.14}$$

and

$$\|\mathbf{E}_{1,i+1}\|_\infty \le \|\mathbf{C}'\|_\infty\|\mathbf{C}_1\|_1. \tag{5.15}$$

Recall the definition

$$\mathbf{E}_{1,i+1} = [\langle \ell_{1l'}, w_{i+1,j}\rangle]_{l'\in\mathbb{Z}_r, j\in\mathbb{Z}_{w(i+1)}}\,.$$

We need to decompose this matrix. This is done by using (5.7) to write

$$\ell_{1l'} = \sum_{s'\in\mathbb{Z}_q} c'_{l's'}\delta_{t_{s'}}, \quad l' \in \mathbb{Z}_r.$$

Thus, it follows from (5.5) and (5.10) for any $j \in \mathbb{Z}_{w(i+1)}$ that there exists a unique pair $\mathbf{e}_j \in \mathbb{Z}_\mu^i$ and $l \in \mathbb{Z}_r$ such that

$$w_{i+1,j} = \sum_{s \in \mathbb{Z}_q} c_{ls} \mathcal{T}_{\mathbf{e}_j} \psi_s.$$

Since for any $s' \in \mathbb{Z}_q, \mathbf{e}_j \in \mathbb{Z}_\mu^i, i \in \mathbb{N}$, and $s = m, \ldots, q-1$,

$$\left\langle \delta_{t_{s'}}, \mathcal{T}_{\mathbf{e}_j} \psi_s \right\rangle = 0,$$

we conclude for $l' \in \mathbb{Z}_r$, $j = \mu(\mathbf{e}_j)r + l$, $\mathbf{e}_j \in \mathbb{Z}_\mu^i$, $l \in \mathbb{Z}_r$, that

$$\langle \ell_{1l'}, w_{i+1,j} \rangle = \sum_{s' \in \mathbb{Z}_q} \sum_{s \in \mathbb{Z}_m} c'_{l's'} c_{ls} \left\langle \delta_{t_{s'}}, \mathcal{T}_{\mathbf{e}_j} \psi_s \right\rangle.$$

We write this equation in matrix form by introducing for each $\mathbf{e} \in \mathbb{Z}_\mu^i$ matrix

$$\mathbf{D}_\mathbf{e} := \left[\left\langle \delta_{t_{s'}}, \mathcal{T}_\mathbf{e} \psi_s \right\rangle \right]_{s' \in \mathbb{Z}_q, s \in \mathbb{Z}_m}$$

and from these matrices build the matrix

$$\mathbf{D} := \left[\mathbf{D}_{\mathbf{e}_0}, \mathbf{D}_{\mathbf{e}_1}, \ldots, \mathbf{D}_{\mathbf{e}_{\mu^i - 1}} \right].$$

This notation allows us to write

$$\mathbf{E}_{1,i+1} = \mathbf{C}' \mathbf{D} \mathrm{diag}\{\mathbf{C}_1^T, \ldots, \mathbf{C}_1^T\},$$

where the rightmost matrix is a block diagonal matrix with μ^i identical blocks of $\mathbf{C}_1^T$. This formula will allow us to estimate the norm of the matrix $\mathbf{E}_{1,i+1}$.

Because the set G_0 is refinable relative to the contractive affine mappings Φ, we are assumed that for any $s' \in \mathbb{Z}_m$ there exist a unique $\mathbf{e}'' \in \mathbb{Z}_\mu^i$ and $s'' \in \mathbb{Z}_m$ such that $t_{s'} = \phi_{\mathbf{e}''}(t_{s''})$. Thus

$$\left\langle \delta_{t_{s'}}, \mathcal{T}_\mathbf{e} \psi_s \right\rangle = \left\langle \mathcal{L}_{\mathbf{e}''} \delta_{t_{s''}}, \mathcal{T}_\mathbf{e} \psi_s \right\rangle = \left\langle \delta_{t_{s''}}, \psi_s \right\rangle \delta_{\mathbf{e}''\mathbf{e}} = \delta_{ss''} \delta_{\mathbf{e}''\mathbf{e}},$$

which implies that $\|\mathbf{D}\|_\infty = 1$. Consequently, we conclude inequality (5.15). Similarly inequality (5.14) follows from

$$\mathbf{E}_{0i} = \left[\langle \ell_{0l'}, w_{ij} \rangle \right]_{l' \in \mathbb{Z}_m, j \in \mathbb{Z}_{w(i)}}.$$

This completes the proof of this lemma. □

We next show that the pair (W, L) of basis functions and collocation functionals constructed in this section satisfies hypotheses (I)–(IV) and (VI)–(X) described in the previous sections. This will be done in three propositions.

PROPOSITION 5.4. *The pair (W, L) satisfies hypotheses* (I)–(IV), (VI), (IX), *and* (X).

Proof. Hypothesis (I) is satisfied because for $(i, j) \in \mathbb{U}$, with $i > 1$, the support of w_{ij} is contained in $S_{ij} = \phi_\mathbf{e}(E) = E_{i-1,\mu(\mathbf{e})}$, where $j = \mu(\mathbf{e})r + l$, $l \in \mathbb{Z}_r$.

We now prove that the pair (W, L) satisfies hypothesis (II). For $(i, j) \in \mathbb{U}$, there exists a unique pair of $\mathbf{e} \in \mathbb{Z}_\mu^{i-1}$ and $l \in \mathbb{Z}_r$ such that $j = \mu(\mathbf{e})r + l$ and $w_{ij} = \mathcal{T}_\mathbf{e} w_{1l}$. Likewise, for $(i', j') \in \mathbb{U}$, there exists a unique pair of $\mathbf{e}' \in \mathbb{Z}_\mu^{i'-1}$ and $l' \in \mathbb{Z}_r$ such that $j' = \mu(\mathbf{e}')r + l'$ and $\ell_{i'j'} = \mathcal{L}_{\mathbf{e}'} \ell_{1l'}$. When $i = i'$, it follows from (5.11) and (5.9) that

$$\langle \ell_{i'j'}, w_{ij} \rangle = \langle \mathcal{L}_{\mathbf{e}'} \ell_{1l'}, \mathcal{T}_\mathbf{e} w_{1l} \rangle = \langle \ell_{1l'}, w_{1l} \rangle \delta_{\mathbf{e}'\mathbf{e}} = \delta_{l'l} \delta_{\mathbf{e}'\mathbf{e}} = \delta_{j'j}.$$

When $i < i'$, let $\mathbf{e}'_1 = (e'_0, \ldots, e'_{i-2}), \mathbf{e}'_2 = (e'_{i-1}, \ldots, e'_{i'-2})$; then

$$\langle \ell_{i'j'}, w_{ij} \rangle = \left\langle \mathcal{L}_{\mathbf{e}'_2} \ell_{1l'}, w_{1l} \right\rangle \delta_{\mathbf{e}'_1 \mathbf{e}} = \left\langle \ell_{1l'}, w_{1l} \circ \phi_{\mathbf{e}'_2} \right\rangle \delta_{\mathbf{e}'_1 \mathbf{e}}.$$

Since $\phi_{\mathbf{e}'_2} : E \to \phi_{\mathbf{e}'_2}(E)$ is an affine mapping, we conclude that $w_{1l} \circ \phi_{\mathbf{e}'_2}$ is a polynomial of total degree $\leq k-1$ in $\mathbb{F}_0$. By using (5.8), we have that

$$\langle \ell_{i'j'}, w_{ij} \rangle = 0, \quad (i,j), (i',j') \in \mathbb{U}, i < i'.$$

When $i > i'$, Lemma 5.3 ensures that condition (2.11) is satisfied. This proves hypothesis (II).

Next, we verify that hypothesis (III) is satisfied. Again, it follows from (5.8) that

$$\langle \ell_{i'j'}, \psi_j \rangle = \langle \ell_{1l'}, \psi_j \circ \phi_{\mathbf{e}'} \rangle = 0, \quad j \in \mathbb{Z}_m.$$

This proves the first equation of hypothesis (III). To prove the second equation, we consider $\mathcal{T}_{\mathbf{e}}$ as an operator from $L^2(E)$ to $L^2(E)$ and denote by $\mathcal{T}^*_{\mathbf{e}}$ the conjugate operator of $\mathcal{T}_{\mathbf{e}}$, which is defined by

$$(\mathcal{T}_{\mathbf{e}} x, y) = (x, \mathcal{T}^*_{\mathbf{e}} y), \quad x, y \in L^2(E).$$

It can be shown that for $y \in L^2(E)$

$$\mathcal{T}^*_{\mathbf{e}} y = J_{\phi_{\mathbf{e}}} y \circ \phi_{\mathbf{e}},$$

where $J_{\Phi_{\mathbf{e}}}$ is the Jacobi of mapping $\phi_{\mathbf{e}}$. Therefore, we have that

$$(w_{ij}, \psi_{j'}) = (\mathcal{T}_{\mathbf{e}} w_{1l}, \psi_{j'}) = (w_{1l}, \mathcal{T}^*_{\mathbf{e}} \psi_{j'}) = 0.$$

The last equality holds because $\mathcal{T}^*_{\mathbf{e}} \psi_{j'}$ is a polynomial of total degree $\leq k-1$ and w_{1l} satisfies condition (5.6).

From (5.12), (5.7), (5.10), and (5.5) we have that for $(i,j) \in \mathbb{U}, j = \mu(\mathbf{e})r + l$,

$$| \langle \ell_{ij}, v \rangle | = | \langle \ell_{1l}, v \circ \phi_{\mathbf{e}} \rangle | \leq \|\mathbf{C}'\|_\infty \|v\|_\infty$$

and

$$\|w_{ij}\|_\infty \leq \|w_{1l} \circ \phi^{-1}_{\mathbf{e}} \chi_{\phi_{\mathbf{e}}(E)}\|_\infty \leq \|\mathbf{C}\|_\infty \max_{j \in \mathbb{Z}_q} \|\psi_j\|_\infty,$$

which confirms hypothesis (IV).

By our construction, it is the case that

$$\dim \mathbb{F}_n = m\mu^n$$

and

$$\dim \mathbb{W}_n = m(\mu - 1)\mu^{n-1}.$$

These equations with (5.2) imply that hypothesis (VI) is satisfied.

The pointwise convergence of the interpolating projections $\mathcal{P}_n$ condition (IX) follows from a result of [AGS]. Finally, hypothesis (X) holds, since $\mathbb{F}_n$ are the spaces of piecewise polynomials of total degree $\leq k-1$. □

The next proposition regards hypothesis (VII).

PROPOSITION 5.5. *For any $k, d \in \mathbb{N}$, there exists an integer $\mu > 1$ such that hypothesis* (VII) *holds.*

Proof. We must show that there exists an integer $\mu > 1$ such that

$$1 + \gamma < \mu^{k/d},$$

where γ is defined in Lemma 5.3. This will be done by proving that γ is bounded from above independent of μ. For this purpose, we consider the matrices

$$\mathbf{H}_1 := [(\psi_i, \psi_j)]_{i\in\mathbb{Z}_m, j\in\mathbb{Z}_m}, \quad \mathbf{H}_2 := [(\psi_i, \psi_{m+j})]_{i\in\mathbb{Z}_m, j\in\mathbb{Z}_r}, \quad \mathbf{H} := [\mathbf{H}_1, \mathbf{H}_2].$$

Therefore, from (5.6) it follows that

$$\mathbf{C}\mathbf{H}^T = \mathbf{C}_1\mathbf{H}_1^T + \mathbf{C}_2\mathbf{H}_2^T = \mathbf{0},$$

where $\mathbf{C}_2$ is an arbitrary $r \times r$ nonsingular matrix. We choose $\mathbf{C}_2 := \mathbf{I}_r$, from which we have that

$$\mathbf{C}_1 = -\mathbf{H}_2^T(\mathbf{H}_1^T)^{-1}. \tag{5.16}$$

Moreover, from Lemma 5.1, we have that

$$\mathbf{C}' = [-\mathbf{B}, \mathbf{I}]$$

and thus

$$\|\mathbf{C}'\|_\infty = \|\mathbf{B}\|_\infty + 1. \tag{5.17}$$

For $j \in \mathbb{Z}_m$, the functions ψ_j are polynomials and therefore continuous, and thus, there exists a positive constant ρ such that

$$\max_{j\in\mathbb{Z}_m} \|\psi_j\|_\infty \le \rho.$$

Hence, recalling the definition of matrix $\mathbf{B}$ and (5.17) we have that

$$\|\mathbf{C}'\|_\infty = 1 + \max_{i\in\mathbb{Z}_r} \sum_{j\in\mathbb{Z}_m} |\psi_j(t_{m+i})| \le 1 + m \max_{j\in\mathbb{Z}_m} \|\psi_j\|_\infty \le 1 + m\rho.$$

On the other hand, we have by (5.16) that

$$\|\mathbf{C}_1\|_1 = \|\mathbf{H}_1^{-1}\mathbf{H}_2\|_\infty \le \|\mathbf{H}_1^{-1}\|_\infty \|\mathbf{H}_2\|_\infty.$$

Since $\|\mathbf{H}_1^{-1}\|_\infty$ is independent of μ, it remains to estimate $\|\mathbf{H}_2\|_\infty$ from above independent of μ. Therefore, we recall for $j \in \mathbb{Z}_r$ that

$$\psi_{m+j}(t) = \psi_l(\phi_e^{-1}(t))\chi_{\phi_e(E)}(t), \quad t \in E,$$

for some $l \in \mathbb{Z}_m$ and $e \in \mathbb{Z}_\mu$. Consequently, from (5.2) we conclude that there exists a positive constant c such that

$$|(\psi_i, \psi_{m+j})| \le \int_{\phi_e(E)} \left|\psi_i(t)\psi_l(\phi_e^{-1}(t))\right| dt \le \rho^2 \mathrm{meas}(\phi_e(E)) \le c\frac{\rho^2}{\mu}.$$

Noting that $r = (\mu - 1)m$, we obtain the desired estimate

$$\|\mathbf{H}_2\|_\infty = \max_{i \in \mathbb{Z}_m} \sum_{j \in \mathbb{Z}_r} |(\psi_i, \psi_{m+j})| \le c \frac{\rho^2}{\mu} (\mu - 1)m \le c\rho^2 m,$$

thereby proving the result. □

We have studied hypothesis (VII) in several cases of practical importance. We report below our finding for the cases when $d = 1$ and $E = [0, 1]$, as well as $d = 2$ and $E = \Delta$, where Δ is the triangle with vertices $(0, 0), (1, 0), (1, 0)$. When $d = 1$ and $E = [0, 1]$, hypothesis (VII) is satisfied for the following choices:

(1) $k = 2$, $\mu = 2$,

$$\phi_e(t) = \frac{t + e}{2}, \quad t \in E, \quad e = 0, 1,$$

and $t_i = (i + 1)/3$ for $i = 0, 1$;

(2) $k = 3$, $\mu = 2$,

$$\phi_e(t) = \frac{t + e}{2}, \quad t \in E, \quad e = 0, 1,$$

and $t_i = 2^i/7$ for $i = 0, 1, 2$;

(3) $k = 3$, $\mu = 3$,

$$\phi_e(t) = \frac{t + e}{3}, \quad t \in E, \quad e = 0, 1, 2,$$

and $t_i = (i + 1)/4$ for $i = 0, 1, 2$;

(4) $k = 4$, $\mu = 2$,

$$\phi_e(t) = \frac{t + e}{2}, \quad t \in E, \quad e = 0, 1,$$

and $t_i = (i + 1)/5$ for $i = 0, 1, 2, 3$.

In the other case hypothesis (VII) is also satisfied when $k = 2$, $\mu = 4$ for $(x, y) \in \Delta$

$$\phi_0(x, y) = \left(\frac{x}{2}, \frac{y}{2}\right), \quad \phi_1(x, y) = \left(\frac{x + 1}{2}, \frac{y}{2}\right),$$

$$\phi_2(x, y) = \left(\frac{x}{2}, \frac{y + 1}{2}\right), \quad \phi_3(x, y) = \left(\frac{1 - x}{2}, \frac{1 - y}{2}\right),$$

and $t_0 = (1/7, 4/7)$, $t_1 = (2/7, 1/7)$, $t_2 = (4/7, 2/7)$.

Finally, we turn our attention to hypothesis (VIII). We consider the sequence of functions $\{\zeta_{ij} : (i, j) \in \mathbb{U}\}$, biorthogonal to the linear functionals $\{\ell_{ij} : (i, j) \in \mathbb{U}\}$ and having property (4.5). Let

$$\zeta_{0j} := w_{0j}, \quad j \in \mathbb{Z}_m,$$

and observe that

$$\langle \ell_{0j}, \zeta_{0j'} \rangle = \delta_{jj'}, \quad j, j' \in \mathbb{Z}_m.$$

For each $j \in \mathbb{Z}_r$, we find vectors $\mathbf{c}_j'' := [c_{js}'' : s \in \mathbb{Z}_q]$ such that the function

$$\zeta_{1j} := \sum_{s \in \mathbb{Z}_q} c_{js}'' \psi_s$$

satisfies the system of linear equations

$$\langle \ell_{0j'}, \zeta_{1j} \rangle = 0, \quad j' \in \mathbb{Z}_m, \tag{5.18}$$

and

$$\langle \ell_{1j'}, \zeta_{1j} \rangle = \delta_{jj'}, \quad j' \in \mathbb{Z}_r. \tag{5.19}$$

Let us confirm $\mathbf{c}_j''$ exists and is unique. The coefficient matrix for (5.18) and (5.19) is

$$\tilde{\mathbf{A}} := [\langle \ell_{i'j'}, \psi_j \rangle]_{j \in \mathbb{Z}_q, (i',j') \in \mathbb{U}_1}. \tag{5.20}$$

Since $\{\psi_j : j \in \mathbb{Z}_q\}$ is a basis for the space $\mathbb{F}_1$, we conclude that matrix $\tilde{\mathbf{A}}$ is nonsingular since $\mathbf{A}$ is nonsingular. Thus, there exists a unique solution $\mathbf{c}''$ for equations (5.18) and (5.19). For $i > 1$, $j = \mu(\mathbf{e})r + l$, $\mathbf{e} \in \mathbb{Z}_\mu^{i-1}$, $l \in \mathbb{Z}_r$, we define functions

$$\zeta_{ij} := \mathcal{T}_{\mathbf{e}} \zeta_{1l}.$$

The functions constructed above will be used in the proof of the next result.

PROPOSITION 5.6. *The pair* (W, L) *satisfies hypothesis* (VIII).

Proof. We first verify that the sequences of functionals $\{\ell_{ij} : (i,j) \in \mathbb{U}\}$ and functions $\{\zeta_{ij} : (i,j) \in \mathbb{U}\}$ are biorthogonal, that is, they satisfy the condition

$$\langle \ell_{i'j'}, \zeta_{ij} \rangle = \delta_{ii'} \delta_{jj'}, \quad (i,j), (i',j') \in \mathbb{U}, \tag{5.21}$$

and, in addition, that there exists a positive constant θ_3 such that for any $i \in \mathbb{N}_0$ condition (4.5) is satisfied.

The proof of (5.21) for the case $i \le i'$ is similar to that for (II) in Proposition 5.4. Hence, we only present the proof for the case $i' < i$. In this case, we have

$$\langle \ell_{i'j'}, \zeta_{ij} \rangle = \langle \mathcal{L}_{\mathbf{e}'} \ell_{1l'}, \mathcal{T}_{\mathbf{e}} \zeta_{1l} \rangle = \langle \ell_{1l'}, \mathcal{T}_{\mathbf{e}_2} \zeta_{1l} \rangle \, \delta_{\mathbf{e}'\mathbf{e}_1},$$

where $j' = \mu(\mathbf{e}')r + l', j = \mu(\mathbf{e})r + l, \mathbf{e}_1 = (e_0, \ldots, e_{i'-2})$, and $\mathbf{e}_2 = (e_{i'-1}, \ldots, e_{i-2})$. From this, it follows that

$$\langle \ell_{i'j'}, \zeta_{ij} \rangle = 0$$

except for $\mathbf{e}' = \mathbf{e}_1$, in which case

$$\langle \ell_{i'j'}, \zeta_{ij} \rangle = \langle \ell_{1l'}, \zeta_{1l} \circ \phi_{\mathbf{e}_2}^{-1} \chi_{\phi_{\mathbf{e}_2}(E)} \rangle. \tag{5.22}$$

Since G_0 is a refinable set, we have that $\phi_e^{-1}(t) \in G_0$ when $t \in G_1 \cap \phi_e(E)$, $e \in \mathbb{Z}_\mu$, and thus

$$\phi_{\mathbf{e}_2}^{-1}(t_s) \in G_0 \ \text{ when } \ t_s \in \phi_{\mathbf{e}_2}(E), \ s \in \mathbb{Z}_q.$$

This observation with (5.18) yields the equation

$$\zeta_{1l}(\phi_{\mathbf{e}_2}^{-1}(t_s)) = \langle \delta_{\phi_{\mathbf{e}_2}^{-1}(t_s)}, \zeta_{1l} \rangle = 0 \tag{5.23}$$

whenever $t_s \in \phi_{\mathbf{e}_2}(E)$, $s \in \mathbb{Z}_q$. We appeal to (5.22) and (5.23) to conclude that $\langle \ell_{i'j'}, \zeta_{ij} \rangle = 0$.

Next, we show that condition (4.5) is satisfied. Without loss of generality, we consider only the case when $i \geq 1$. In this case the definition of ζ_{ij}, for $i \geq 1$, guarantees that

$$\begin{aligned}\sup_{t\in E} \sum_{j\in\mathbb{Z}_{w(i)}} |\zeta_{ij}(t)| &= \sup_{t\in E} \sum_{\mathbf{e}\in\mathbb{Z}_\mu^{i-1}} \sum_{l\in\mathbb{Z}_r} |\mathcal{T}_{\mathbf{e}}\zeta_{1l}(t)| \\ &= \sup_{t\in E} \sum_{\mathbf{e}\in\mathbb{Z}_\mu^{i-1}} \sum_{l\in\mathbb{Z}_r} |\zeta_{1l}(\phi_{\mathbf{e}}^{-1}(t))\chi_{\phi_{\mathbf{e}}(E)}(t)| \\ &\leq \sum_{l\in\mathbb{Z}_r} \|\zeta_{1l}\|_\infty,\end{aligned}$$

and therefore (4.5) holds with $\theta_3 := \sum_{l\in\mathbb{Z}_r} \|\zeta_{1l}\|_\infty$.

Finally, we verify the first inequality of (4.4) in condition (VIII). To this end, we note that for $v := \sum_{(i,j)\in\mathbb{U}_n} v_{i,j} w_{i,j}$ and $\mathbf{v} := [v_{ij} : (i,j) \in \mathbb{U}_n]^T$, there exists $(i_0, j_0) \in \mathbb{U}_n$ with $j_0 = \mu(\mathbf{e}_0)r + l_0, \mathbf{e}_0 \in \mathbb{Z}_\mu^{i_0-1}, l_0 \in \mathbb{Z}_r$, such that

$$\|\mathbf{v}\|_\infty = |v_{i_0 j_0}|. \tag{5.24}$$

For $l \in \mathbb{Z}_r$ we denote $\tilde{v}_l := v_{i_0 j}$ and $\tilde{w}_l := w_{i_0 j}$, where $j = \mu(\mathbf{e}_0)r + l$ and $\mathbf{e}_0 \in \mathbb{Z}_\mu^{i_0-1}$, and observe that

$$|v_{i_0 j_0}| \leq \left(\sum_{l\in\mathbb{Z}_r} |\tilde{v}_l|^2 \right)^{1/2}. \tag{5.25}$$

Recalling that $w_{i_0 j} = \mathcal{T}_{\mathbf{e}} w_{1l}$, $l \in \mathbb{Z}_r$, and that ϕ_e, $e \in \mathbb{Z}_\mu$, are affine, we conclude that

$$(\tilde{w}_{l'}, \tilde{w}_l) = J_{\phi_{\mathbf{e}_0}}(w_{1l'}, w_{1l}), \tag{5.26}$$

where J_{ϕ_e} denotes the Jacobi of the mapping ϕ_e. We introduce an $r \times r$ matrix

$$\mathbf{W} := [(w_{1l'}, w_{1l})]_{l',l\in\mathbb{Z}_r}$$

and note that it is the Gram matrix of the basis w_{1l}, $l \in \mathbb{Z}_r$, and thus it is positive definite. It follows that there exists a positive constant c_0 such that for $\tilde{v} := \sum_{l\in\mathbb{Z}_r} \tilde{v}_l \tilde{w}_l$ and $\tilde{\mathbf{v}} := [\tilde{v}_l : l \in \mathbb{Z}_r]^T$,

$$c_0 \sum_{l\in\mathbb{Z}_r} |\tilde{v}_l|^2 \leq \tilde{\mathbf{v}}^T \mathbf{W} \tilde{\mathbf{v}}. \tag{5.27}$$

By formula (5.26), we have that

$$\|\tilde{v}\|_2^2 = (\tilde{v}, \tilde{v}) = J_{\phi_{\mathbf{e}_0}} \tilde{\mathbf{v}}^T \mathbf{W} \tilde{\mathbf{v}}.$$

Combining this equation with (5.27) yields

$$\sum_{l\in\mathbb{Z}_r} |\tilde{v}_l|^2 \leq \frac{1}{c_0 J_{\phi_{\mathbf{e}_0}}} \|\tilde{v}\|_2^2. \tag{5.28}$$

Since the basis $\{w_{ij} : (i,j) \in \mathbb{U}_n\}$ that has been constructed in this section satisfies the conditions (III), we obtain that

$$\|\tilde{v}\|_2^2 = \int_{\phi_{\mathbf{e}}(E)} \tilde{v}(t)v(t)dt \le J_{\phi_{\mathbf{e}_0}} \|\tilde{v}\|_\infty \|v\|_\infty \le J_{\phi_{\mathbf{e}_0}} |v_{i_0 j_0}| \sum_{l \in \mathbb{Z}_r} \|\tilde{w}_l\|_\infty \|v\|_\infty.$$

Using condition (IV), we conclude that there exists a positive constant c_0 such that

$$\sum_{l \in \mathbb{Z}_r} \|\tilde{w}_l\|_\infty \le c,$$

which implies with the last inequality that

$$\|\tilde{v}\|_2^2 \le c J_{\phi_{\mathbf{e}_0}} \|\mathbf{v}\|_\infty \|v\|_\infty. \tag{5.29}$$

Combining (5.24), (5.25), (5.28), and (5.29) yields that there exists a positive constant c such that

$$\|\mathbf{v}\|_\infty \le c \|\mathbf{v}\|_\infty^{1/2} \|v\|_\infty^{1/2},$$

and thus

$$\|\mathbf{v}\|_\infty \le c \|v\|_\infty.$$

We have proved the first inequality of (4.4) with $\theta_2 := 1/c$. □

REFERENCES

[Al] B. K. Alpert, *A class of bases in L^2 for the sparse representation of integral operators*, SIAM J. Math. Anal., 24 (1993), pp. 246–262.

[An] P. M. Anselone, *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[At1] K. E. Atkinson, *A survey of boundary integral equation methods for the numerical solution of Laplace's equation in three dimensions*, in Numerical Solution of Integral Equations, M. Golberg, ed., Plenum Press, NY, 1990, pp. 1–34.

[At2] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[AC1] K. E. Atkinson and D. Chien, *Piecewise polynomial collocation for boundary integral equations*, SIAM J. Sci. Comput., 16 (1994), pp. 651–681.

[AC2] K. E. Atkinson and D. Chien, *A fast matrix-vector multiplication method for solving the radiosity equation*, Adv. Comput. Math., 12 (2000), pp. 151–174.

[ACr] K. E. Atkinson and G. Chandler, *The collocation method for solving the radiosity equation for unoccluded surfaces*, J. Integral Equations Appl., 10 (1998), pp. 253–290.

[AGS] K. Atkinson, I. Graham, and I. Sloan, *Piecewise continuous collocation for integral equations*, SIAM J. Numer. Anal., 20 (1983), pp. 172–186.

[BCR] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[CMX1] Z. Chen, C. A. Micchelli, and Y. Xu, *The Petrov-Galerkin methods for second kind integral equations* II: *Multiwavelet scheme*, Adv. Comput. Math., 7 (1997), pp. 199–233.

[CMX2] Z. Chen, C. A. Micchelli, and Y. Xu, *A construction of interpolating wavelets on invariant sets*, Math. Comp., 68 (1999), pp. 1569–1587.

[CMX3] Z. Chen, C. A. Micchelli, and Y. Xu, *Discrete wavelet Petrov-Galerkin methods*, Adv. Comput. Math., 16 (2002), pp. 1–28.

[CX] Z. Chen and Y. Xu, *The Petrov–Galerkin and iterated Petrov–Galerkin methods for second kind integral equations*, SIAM J. Numer. Anal., 35 (1998), pp. 406–434.

[DPS] W. Dahmen, S. Proessdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations* II: *Matrix compression and fast solutions*, Adv. Comput. Math., 1 (1993), pp. 259–335.

[H] J. E. Hutchinson, *Fractals and self similarity*, Indiana Univ. Math. J., 30 (1981), pp. 713–747.

[MSX1] C. A. Micchelli, T. Sauer, and Y. Xu, *A construction of refinable sets for interpolating wavelets*, Results Math., 34 (1998), pp. 359–372.

[MSX2] C. A. Micchelli, T. Sauer, and Y. Xu, *Subdivision schemes for iterated function systems*, Proc. Amer. Math. Soc., 129 (2001), pp. 1861–1872.

[MX1] C. A. Micchelli and Y. Xu, *Using the matrix refinement equation for the construction of wavelets on invariant sets*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 391–401.

[MX2] C. A. Micchelli and Y. Xu, *Reconstruction and decomposition algorithms for biorthogonal multiwavelets*, Multidimen. Systems Signal Process., 8 (1997), pp. 31–69.

[MXZ] C. A. Micchelli, Y. Xu, and Y. Zhao, *Wavelet Galerkin methods for second-kind integral equations*, J. Comput. Appl. Math., 86 (1997), pp. 251–270.

[PS] T. von Petersdorff and C. Schwab, *Wavelet approximation of first kind integral equations in a polygon*, Numer. Math., 74 (1996), pp. 479–516.

[PSS] T. von Petersdorff, C. Schwab, and R. Schneider, *Multiwavelets for second-kind integral equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2212–2227.

[R] A. Rathsfeld, *A wavelet algorithm for the solution of a singular integral equation over a smooth two-dimensional manifold*, J. Integral Equations Appl., 10 (1998), pp. 445–501.

SIAM J. NUMER. ANAL.
Vol. 40, No. 1, pp. 376–401

NUMERICAL TREATMENT OF DEFECTIVE BOUNDARY CONDITIONS FOR THE NAVIER–STOKES EQUATIONS*

L. FORMAGGIA†, J.-F. GERBEAU‡, F. NOBILE‡, AND A. QUARTERONI†§

Abstract. We present a formulation for accommodating defective boundary conditions for the incompressible Navier–Stokes equations where only averaged values are prescribed on measurable portions of the boundary. In particular we consider the case where the flow rate is imposed on several domain sections. This methodology has an interesting application in the numerical simulation of flow in blood vessels, when only a reduced set of boundary data are generally available for the upstream and downstream sections.

Key words. Navier–Stokes equations, boundary conditions, finite elements, Lagrange multipliers, fractional step methods, simulation of blood flow

AMS subject classifications. 35Q30, 65N30

PII. S003614290038296X

Introduction. A necessary condition for the existence of the solution of the incompressible Navier–Stokes equations on a bounded domain Ω is that an appropriate set of boundary conditions is imposed on $\partial\Omega$. In a classical setting, at each point on the boundary one needs a number of conditions equal to the spatial dimension of the problem. Typically, one can prescribe the components of the velocity (Dirichlet boundary condition) or those of the Cauchy normal stress (Neumann boundary condition) or an appropriate combination of velocity and normal stress.

In this work, we will consider the specific situation occurring when one has at their disposal only averaged quantities on portions of the domain boundary, a priori insufficient to "close" the differential problem at hand. We will refer to this incomplete set as *defective* boundary conditions. An important applicative field where this situation occurs is the numerical simulation of blood flow in the human vascular system. If we aim at computing the blood flow field in an isolated portion of an artery—for instance, reconstructed from medical images—we immediately face the problem of which boundary conditions to impose at the artificial upstream and downstream sections. A possibility is to exploit data coming from measurements. Yet the most common techniques are normally able to measure only flow rates or average pressures and not complete fields as would be required for the numerical computations.

A similar situation occurs when one wants to simulate the cardiovascular system by multiscale approaches such those proposed in [8] and [6]. In that case, the boundary data for the Navier–Stokes equations do not come from measurements, but rather come as the output of simplified models of the global cardiovascular system.

*Received by the editors December 22, 2000; accepted for publication (in revised form) October 26, 2001; published electronically May 10, 2002. The work has been partially supported by the Swiss National Science Foundation project 21-54139.98, by MURST Cofin. 1998 "Advanced Numerical Methods for Scientific Computing," and by the special project of the Politecnico di Milano "Multiscale Computing in Biofluidynamics."

http://www.siam.org/journals/sinum/40-1/38296.html

†Départment de Mathématiques, École Polytechnique Fédérale de Lausanne, CH-1015, Lausane, Switzerland (luca.formaggia@epfl.ch, fabio.nobile@epfl.ch, alfio.quarteroni@epfl.ch).

‡INRIA, Projet M3N, Rocquencourt B.P. 105, F-78153 Le Chesnay Cedex, France (Jean-Frederic.Gerbeau@inria.fr).

§Dipartimento di Matematica "Francesco Brioschi," Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy.

These simple models, which are typically based on the solution of either a system of ordinary differential equations or of one-dimensional differential problems, normally provide the evolution of mean pressure and velocity inside the various regions of the cardiovascular system. If we want to use them to feed boundary data into a more detailed local model based on the solution of the Navier–Stokes equations, we need to have a way to "translate" these mean quantities in mathematically sound boundary conditions for the Navier–Stokes problem. This issue has been addressed in the cited references. Another interesting application of the technique proposed in this work is in the simulation of a free interface problem, when one wants to ensure that the numerical approximation satisfies mass conservation within machine precision. An example of a problem of this type is presented in the section dedicated to numerical experiments. Another applicative field is the simulation of flow in pipes, when measuring sensors provide only flow rate information.

A viable approach to handle the case of defective boundary conditions is provided by the so-called *do-nothing* boundary conditions proposed in [10].

In this paper we analyze another, somewhat more flexible, alternative based on the use of Lagrange multipliers. In particular, we will consider the case when given flow rates Q_i are to be prescribed on several sections of the domain boundary. The corresponding variational formulation, augmented by the Lagrange multipliers, is presented in all generality and analyzed for the case of a Stokes problem, where a well-posedness result is given. We present several approaches for the numerical solution of this problem in the context of fractional step schemes and compare their properties with respect to the effective fulfillment of the imposed flow rate constraint and computational efficiency. Several numerical experiments prove the effectiveness of this technique, which may be implemented in existing software with little efforts. This is an advantage with respect to the do-nothing approach, whose implementation for the prescribed flow rate problem is not straightforward, as it would require the construction of suitable (nonstandard) test functions.

We also discuss how to impose an average pressure (or normal stresses) on measurable parts of the domain. We show how the Lagrange multiplier technique may be successfully implemented in the case of slip boundary conditions for the velocity.

The technique developed here is targeted to applications where it is important to match the solution at the inflow and/or the outflow with known average data. In the present form it is not directly applicable for far-field conditions in unbounded domains and in particular for devising "nonabsorbing" boundary treatment for vortex flow. The reader interested in this particular aspect may, for instance, refer to [3, 4, 5].

In the first section of this work we address the problem in general terms and we introduce the functional setting for the analysis. We also give an overview of the do-nothing approach applied to problems where either the flow rate or the average normal stress is imposed on a portion of the computational domain boundary. In section 2 we introduce the alternative formulation given by a Lagrange multiplier approach and carry out its analysis. In section 3 we propose several algorithms that are suitable for its implementation in the context of the solution of the Navier–Stokes equations by algebraic fractional step techniques. Finally, section 4 presents numerical results illustrating the effectiveness of the proposed methodology.

1. Problem formulation and defective boundary conditions. Let Ω be a bounded domain of $\mathbb{R}^d$, $d = 2$ or 3, whose boundary $\partial\Omega$ is decomposed into the union of Γ and several disjoint sections S_0, $S_1, \ldots, S_n$, $n \geq 1$ (see Figure 1).

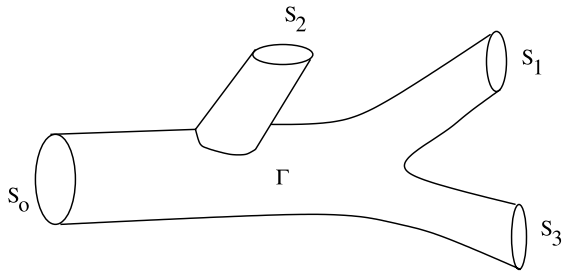


FIG. 1. *The partition of the boundary of the domain* Ω.

We are interested in solving the Navier–Stokes equations in Ω,

$$\begin{cases} \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p - \nu \triangle \mathbf{u} = \mathbf{f}, & t > 0, \\ \operatorname{div} \mathbf{u} = 0, & t > 0, \\ \mathbf{u} = \mathbf{u}_0, & t = 0, \end{cases} \tag{1.1}$$

supplemented by homogeneous boundary conditions on Γ,

$$\mathbf{u}|_\Gamma = 0, \tag{1.2}$$

while two different kinds of boundary conditions will be considered on the sections S_i, $i = 0, \ldots, n$. Both are well suited for blood flow simulations [18, 6], where Ω would represent the portion of an artery, Γ the vessel wall, and S_i the artificial upstream and downstream sections. Even if for this specific problem the vessel wall should be considered moving with time because of the flexibility of the vessel wall structure, here we will address only the case where Γ is fixed.

The first condition we consider is the so-called *prescribed mean pressure problem* which requires that

$$\frac{1}{\operatorname{meas}(S_i)} \int_{S_i} p \, ds = P_i, \quad i = 0, \ldots, n, \tag{1.3}$$

where each P_i is a prescribed function of the time t, constant on S_i.

The second condition we wish to address is the *prescribed flow rate problem*

$$\int_{S_i} \mathbf{u} \cdot \mathbf{n} \, ds = Q_i \quad \text{for } i = 0, \ldots, n, \tag{1.4}$$

where the flow rates Q_i's (also called *velocity fluxes*) are assigned functions of time. Due to the fluid incompressibility fluid, a compatibility relation must exist among the fluxes Q_i, namely Q_0 must be equal to $-\sum_{i=1}^n Q_i$.

The initial-boundary value problem (1.1)–(1.2) with either (1.3) or (1.4) is not well-posed from a physical point of view, since its solution is not unique. Indeed, on every section S_i, we are prescribing just one scalar condition rather than d conditions at every point $\mathbf{x} \in S_i$, as it should be.

In [10] the do-nothing approach was advocated as a way of solving the two situations just presented. By this technique, a particular weak formulation is devised which allows to fulfill conditions (1.3) (resp., (1.4)) at some extent, giving rise to a well-posed problem. In fact, this formulation contains also "implicit" (Neumann-like) boundary conditions which select one particular solution among all the physical solutions of the original differential problem.

We will here give a brief presentation of this approach. Let us introduce the functional spaces

$$V = \left\{ \mathbf{v} \in \left[H^1(\Omega)\right]^d, \mathbf{v}|_\Gamma = 0 \right\} \qquad \text{and} \qquad M = L^2(\Omega).$$

We suppose that $\mathbf{f} \in V'$ and we introduce the functional $\phi_i \in V'$, $i = 0, \ldots, n$, which measures the flux of a vector function through the surface S_i. Precisely,

$$\langle \phi_i, \mathbf{v} \rangle = \int_{S_i} \mathbf{v} \cdot \mathbf{n} \, ds \quad \forall \mathbf{v} \in V,$$

where $\mathbf{n}$ is the outward unit normal vector on $\partial\Omega$. For this reason ϕ_i is called the *flux functional* on S_i.

Then, the do-nothing formulation for the *mean pressure problem* reads as follows: find $\mathbf{u} \in V$ and $p \in M$ such that, for all $\mathbf{v} \in V$ and $q \in M$,

$$\text{(1.5)} \qquad \begin{cases} (\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}, \mathbf{v}) + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) - (p, \operatorname{div} \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle - \displaystyle\sum_{i=0}^{n} P_i \langle \phi_i, \mathbf{v} \rangle, \\ (q, \operatorname{div} \mathbf{u}) = 0, \end{cases}$$

for all $t > 0$, with $\mathbf{u} = \mathbf{u}_0$ for $t = 0$.

It follows easily, by using the Green formula, that the solution of (1.5) satisfies

$$\left(p - \nu \frac{\partial u_n}{\partial \mathbf{n}} \right) \Big|_{S_i} = P_i, \qquad \frac{\partial \mathbf{u}_\tau}{\partial \mathbf{n}} \Big|_{S_i} = 0 \quad \text{for } i = 0, \ldots, n,$$

where we have set $u_n = \mathbf{u} \cdot \mathbf{n}$ and $\mathbf{u}_\tau = \mathbf{u} - u_n \mathbf{n}$.

Thus

$$\text{(1.6)} \qquad \frac{1}{\operatorname{meas}(S_i)} \int_{S_i} p \, ds = P_i + \frac{\nu}{\operatorname{meas}(S_i)} \int_{S_i} \frac{\partial u_n}{\partial \mathbf{n}} \, ds.$$

We conclude that the desired condition (1.3) is recovered exactly only in those cases where the last integral in (1.6) vanishes. This occurs, for instance, when S_i is a plane section perpendicular to a cylindrical pipe. Otherwise, P_i will be the mean value of the normal component of the normal stresses on S_i.

For the *prescribed flow rate problem*, the do-nothing approach is formulated as follows. Let us introduce the space

$$V^\star = \{ \mathbf{v} \in V, \langle \phi_i, \mathbf{v} \rangle = 0, i = 0, \ldots, n \}$$

and the vector functions $\mathbf{b}_i \in V$, $i = 1, \ldots, n$, (called *flux-carriers*) that satisfy

$$\operatorname{div} \mathbf{b}_i = 0, \quad \int_{S_0} \mathbf{b}_i \cdot \mathbf{n} \, ds = -1, \quad \int_{S_j} \mathbf{b}_i \cdot \mathbf{n} \, ds = \delta_{ij} \quad \text{for } i, j = 1, \ldots, n.$$

The weak formulation of problem (1.1), (1.2), (1.4) proposed in [10] reads as follows: Find $\mathbf{u} = \mathbf{w} + \sum_{i=1}^{n} Q_i \mathbf{b}_i$, with $\mathbf{w} \in V^\star$ and $p \in M \setminus \mathbb{R}$ such that for all $\mathbf{v} \in V^\star$ and $q \in M$

$$\text{(1.7)} \qquad \begin{cases} (\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}, \mathbf{v}) + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) - (p, \operatorname{div} \mathbf{v}) = 0, \\ (q, \operatorname{div} \mathbf{u}) = 0 \end{cases}$$

for all $t > 0$, with $\mathbf{u} = \mathbf{u}_0$ for $t = 0$.

The corresponding solution satisfies

$$\left(p - \nu\frac{\partial u_n}{\partial \mathbf{n}}\right)|_{S_i} = P_i, \qquad \frac{\partial \mathbf{u}_\tau}{\partial \mathbf{n}}|_{S_i} = 0 \quad \text{for } i = 0, \dots, n, \tag{1.8}$$

where the P_i's are *a priori* unknown constants (in space).

The formulation of the *mean pressure problem* may be easily discretized as it can be regarded as a classical Navier–Stokes problem with Neumann boundary conditions. On the other hand, the definition of the functional space $V^\star$ makes the implementation of the *prescribed flow rate problem* less straightforward.

2. A Lagrange multiplier approach for flow rate boundary conditions. In this section, we propose a slightly different formulation of the prescribed flow rate problem presented above. We consider (1.1) and (1.2) and we prescribe the velocity flux on all but one section of $\partial\Omega$. More precisely, we aim at satisfying

$$\langle \phi_i, \mathbf{u}\rangle = \int_{S_i} \mathbf{u}\cdot\mathbf{n}\, ds = Q_i \quad \text{for } i = 1, \dots, n, \tag{2.1}$$

plus the following homogeneous Neumann boundary condition on S_0:

$$\left(-p\mathbf{n} + \frac{\partial \mathbf{u}}{\partial \mathbf{n}}\right)\Big|_{S_0} = 0. \tag{2.2}$$

The motivation of such an approach will be clarified in Remark 2.

Our goal is to formulate the initial-boundary value problem (1.1), (1.2), (2.1), (2.2) in a way that its numerical approximation be as simple as possible to implement. We look for $\mathbf{u} \in V$, $p \in M$, and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ such that, for all $\mathbf{v} \in V$ and $q \in M$,

$$\begin{cases} (\partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u}, \mathbf{v}) + \nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + \displaystyle\sum_{i=1}^{n} \lambda_i \langle \phi_i, \mathbf{v}\rangle - (p, \operatorname{div}\mathbf{v}) = \langle \mathbf{f}, \mathbf{v}\rangle, \\ (q, \operatorname{div}\mathbf{u}) = 0, \\ \langle \phi_i, \mathbf{u}\rangle = Q_i, \qquad i = 1, \dots, n, \end{cases} \tag{2.3}$$

for all $t > 0$, with $\mathbf{u} = \mathbf{u}_0$ for $t = 0$.

Note that now the test functions $\mathbf{v}$ are taken in V, a space which is more straightforward to discretize than $V^\star$.

PROPOSITION 2.1. *Any smooth solution of* (2.3) *satisfies the additional boundary conditions*

$$\left(p - \nu\frac{\partial u_n}{\partial \mathbf{n}}\right)|_{S_i} = \lambda_i \quad \textit{and} \quad \frac{\partial \mathbf{u}_\tau}{\partial \mathbf{n}}|_{S_i} = 0, \quad i = 1, \dots, n. \tag{2.4}$$

In particular, this yields that both $\frac{\partial \mathbf{u}_\tau}{\partial \mathbf{n}}$ *and* $p - \nu\frac{\partial u_n}{\partial \mathbf{n}}$ *are indeed constant over* S_i *for* $i = 1, \dots, n$. *Furthermore,* $(\mathbf{u}, p)$ *satisfies* (1.1), (1.2), (2.1), (2.2).

Proof. Conditions (1.2) and (2.1) are obviously satisfied. Integrating by parts the first equation of (2.3) yields for any $\mathbf{v} \in V$

$$\begin{aligned} \langle \partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} - \nu\triangle\mathbf{u} + \nabla p, \mathbf{v}\rangle + \int_{S_0} \left(\nu\frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n}\right)\cdot\mathbf{v}\, ds \\ + \sum_{i=1}^{n} \int_{S_i} \left(\nu\frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} + \lambda_i\mathbf{n}\right)\cdot\mathbf{v}\, ds = \langle \mathbf{f}, \mathbf{v}\rangle. \end{aligned} \tag{2.5}$$

Now taking $\mathbf{v} \in \mathcal{D}(\Omega)$, we recover the momentum equation (1.1) in the sense of $\mathcal{D}'(\Omega)$. Consequently, from (2.5), it follows that

$$\int_{S_0} \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} \right) \cdot \mathbf{v}\, ds + \sum_{i=1}^{n} \int_{S_i} \left(\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} + \lambda_i \mathbf{n} \right) \cdot \mathbf{v}\, ds = 0$$

for all $\mathbf{v} \in V$. Now using the splitting of the trace of u on $\partial\Omega$ in its normal and tangential component $\mathbf{u}|_{\partial\Omega} = u_n \mathbf{n} + \mathbf{u}_\tau$, we deduce relations (2.2) and (2.4). ∎

Remark 1. Among all possible solutions of (1.1), (1.2), (2.1), (2.2), problem (2.3) selects the one that satisfies the additional boundary condition (2.4).

Remark 2. In (2.3), we could have imposed a flux Q_0 on S_0, instead of the homogeneous Neumann condition (2.2). The corresponding problem would have been as follows: find $\mathbf{u} \in V$, $\tilde{p} \in M$, and $\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_n \in \mathbb{R}$ such that, for all $\mathbf{v} \in V$ and $q \in M$,

$$\text{(2.6)} \qquad \begin{cases} (\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}, \mathbf{v}) + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + \displaystyle\sum_{i=0}^{n} \tilde{\lambda}_i \langle \phi_i, \mathbf{v} \rangle - (p, \operatorname{div} \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle, \\ (q, \operatorname{div} \mathbf{u}) = 0, \\ \langle \phi_i, \mathbf{u} \rangle = Q_i, \qquad i = 0, \ldots, n, \end{cases}$$

for all $t > 0$, with $\mathbf{u} = \mathbf{u}_0$ for $t = 0$.

Due to the incompressibility of the fluid, the value of Q_0 must be equal to $-\sum_{i=1}^{n} Q_i$ (otherwise the problem has no solution). If $(\mathbf{u}, \tilde{p}, \tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ is a solution of (2.6), then, for any constant $C \in \mathbb{R}$, $(\mathbf{u}, \tilde{p} + C, \tilde{\lambda}_0 + C, \tilde{\lambda}_1 + C, \ldots, \tilde{\lambda}_n + C)$ is also a solution. The solution of (2.6) is thus defined up to an additive constant. Now, we set this constant C equal to $-\tilde{\lambda}_0$, and we denote $\tilde{p} + C$ and $\tilde{\lambda}_i + C$ by p and λ_i. Then, $(\mathbf{u}, p, \lambda_1, \ldots, \lambda_n)$ is the solution of (2.3) and, according to the "implicit" boundary condition (2.4), $\lambda_0 = 0$ yields simply the Neumann boundary condition (2.2). In other words, problem (2.3) (with the Neumann condition on S_0) and (2.6) (with the flux condition on S_0) are equivalent as soon as the "free" constant of problem (2.6) is well chosen.

Remark 3. From a theoretical viewpoint, our approach is very close to the do-nothing formulation recalled in the previous section. Comparing (1.8) and (2.4), we may note that the Lagrange multipliers corresponding to the constraints on the flux are in fact equal to the "a priori unknown" constants of the do-nothing formulation (1.8). Yet, our approach uses a standard functional space V which can be more straightforwardly discretized than the space V^*.

Now, for the sake of simplicity, we restrict ourselves to the analysis of the stationary Stokes problem (which embodies however all relevant difficulties of our Lagrange multiplier approach): find $(\mathbf{u}, p, \lambda_1, \ldots, \lambda_n) \in V \times M \times \mathbb{R}^n$ such that for all $(\mathbf{v}, q) \in V \times M$

$$\text{(2.7)} \qquad \begin{cases} \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + \displaystyle\sum_{i=1}^{n} \lambda_i \langle \phi_i, \mathbf{v} \rangle - (p, \operatorname{div} \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle, \\ (q, \operatorname{div} \mathbf{u}) = 0, \\ \langle \phi_i, \mathbf{u} \rangle = Q_i, \ i = 1, \ldots, n. \end{cases}$$

The extension of the analysis to the complete time-dependent, nonlinear problem (2.3) can then be carried out by usual techniques (see, e.g., [20, 9, 19]).

PROPOSITION 2.2. *Problem* (2.7) *is well-posed.*

Proof. In order to prove existence, let us denote by $(\tilde{\mathbf{u}}, \tilde{p}) \in V \times M$ the solution of

$$\begin{cases} \nu(\nabla\tilde{\mathbf{u}}, \nabla\mathbf{v}) - (\tilde{p}, \operatorname{div}\mathbf{v}) = \langle \mathbf{f}, \mathbf{v}\rangle, \\ (q, \operatorname{div}\tilde{\mathbf{u}}) = 0 \end{cases} \tag{2.8}$$

for all $(\mathbf{v}, q) \in V \times M$. This is the weak formulation of the Stokes problem with homogeneous Dirichlet conditions on Γ and homogeneous Neumann conditions on $\partial\Omega \setminus \Gamma$.

Moreover, for $i = 1, \dots, n$, let $(\mathbf{w}_i, \pi_i) \in V \times M$ be the solution of the problem

$$\begin{cases} \nu(\nabla\mathbf{w}_i, \nabla\mathbf{v}) - (\pi_i, \operatorname{div}\mathbf{v}) = -\langle \phi_i, \mathbf{v}\rangle, \\ (q, \operatorname{div}\mathbf{w}_i) = 0, \end{cases} \tag{2.9}$$

for all $(\mathbf{v}, q) \in V \times M$.

Both systems (2.8) and (2.9) admit a unique solution. Note that the equations satisfied by $(\tilde{\mathbf{u}}, \tilde{p})$ are the unconstrained counterpart of (2.7) and that the solution $(\mathbf{w}_i, \pi_i)$ of (2.9) depends only on the geometry and not on the data of the Stokes problem. In some sense, the functions $\mathbf{w}_i$ are related to the flux-carriers $\mathbf{b}_i$ introduced in the do-nothing formulation (1.7).

We set, then, $\mathbf{u} = \tilde{\mathbf{u}} + \sum_{i=1}^n \lambda_i \mathbf{w}_i$ and $p = \tilde{p} + \sum_{i=1}^n \lambda_i \pi_i$. No matter how the λ_i, $i = 1, \dots, n$, are chosen, $(\mathbf{u}, p)$ satisfies the first two equations of (2.7). If we further require $\mathbf{u}$ to satisfy the third equation of (2.7), we obtain the following equations:

$$\langle \phi_i, \tilde{\mathbf{u}}\rangle + \sum_{j=1}^n \lambda_j \langle \phi_i, \mathbf{w}_j\rangle = Q_i, \quad i = 1, \dots, n,$$

whose compact form reads as

$$B\Lambda = Q - \tilde{Q}, \tag{2.10}$$

where $Q = (Q_1, \dots, Q_n) \in \mathbb{R}^n$, $\tilde{Q} = (\langle \phi_1, \tilde{\mathbf{u}}\rangle, \dots, \langle \phi_n, \tilde{\mathbf{u}}\rangle) \in \mathbb{R}^n$, $\Lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n\times n}$, $B_{ij} = \langle \phi_i, \mathbf{w}_j\rangle$. The matrix B is nonsingular (as will be proven in Lemma 2.3); thus, once the $\{\lambda_i\}$ are computed through relation (2.10), $(\mathbf{u}, p, \lambda_1, \dots, \lambda_n)$ will provide a solution of (2.7).

To prove uniqueness, let $(\mathbf{u}_1, p_1, \lambda_1^{(1)}, \dots, \lambda_n^{(1)})$ and $(\mathbf{u}_2, p_2, \lambda_1^{(2)}, \dots, \lambda_n^{(2)})$ be two solutions of (2.7). Then

$$\begin{cases} \nu(\nabla(\mathbf{u}_1 - \mathbf{u}_2), \nabla\mathbf{v}) + \displaystyle\sum_{i=1}^n (\lambda_i^{(1)} - \lambda_i^{(2)})\langle \phi_i, \mathbf{v}\rangle - (p_1 - p_2, \operatorname{div}\mathbf{v}) = 0, \\ (q, \operatorname{div}(\mathbf{u}_1 - \mathbf{u}_2)) = 0, \\ \langle \phi_i, \mathbf{u}_1 - \mathbf{u}_2\rangle = 0, \qquad i = 1, \dots, n, \end{cases} \tag{2.11}$$

for all $\mathbf{v} \in V$ and $q \in M$.

Taking $\mathbf{v} = \mathbf{u}_1 - \mathbf{u}_2$ in (2.11) we obtain $\nu\|\nabla(\mathbf{u}_1 - \mathbf{u}_2)\|_{L^2(\Omega)} = 0$, from which $\mathbf{u}_1 = \mathbf{u}_2$ a.e. in Ω. Consequently,

$$\sum_{i=1}^n (\lambda_i^{(1)} - \lambda_i^{(2)})\langle \phi_i, \mathbf{v}\rangle - (p_1 - p_2, \operatorname{div}\mathbf{v}) = 0 \qquad \forall \mathbf{v} \in V. \tag{2.12}$$

For all $i = 1, \dots, n$ we can construct $\mathbf{w}_i \in V$ which satisfy

$$\operatorname{div} \mathbf{w}_i = 0, \qquad \mathbf{w}_i|_{S_j} = 0,\ j = 1, \dots, n,\ j \neq i, \quad \text{and} \quad \langle \phi_i, \mathbf{w}_i \rangle = 1.$$

Note that $\langle \phi_0, \mathbf{w}_i \rangle = -1$ for all $i = 1, \dots, n$.

Taking in (2.12) $\mathbf{v} = \mathbf{w}_i$ we obtain

$$\lambda_i^{(1)} = \lambda_i^{(2)} \qquad \forall i = 1, \dots, n.$$

Finally, choosing $\mathbf{z} \in V$ such that

$$\operatorname{div} \mathbf{z} = p_1 - p_2, \qquad \mathbf{z}|_{S_i} = 0 \quad \forall i = 1, \dots, n,$$

(such a function exists; see, e.g., [9, 13]) and taking $\mathbf{v} = \mathbf{z}$ in (2.12), we obtain $\|p_1 - p_2\|_{L^2(\Omega)} = 0$. Henceforth $p_1 = p_2$ a.e. in Ω. ∎

LEMMA 2.3. *The matrix B introduced in* (2.10) *is nonsingular.*

Proof. Given an arbitrary vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, for any $i = 1, \dots, n$ we can multiply each problem (2.9) by α_i and sum from $i = 1$ to n. Owing to the linearity of (2.9) we have

$$\begin{cases} \nu(\sum_{i=1}^n \alpha_i \nabla \mathbf{w}_i, \nabla \mathbf{v}) - (\sum_{i=1}^n \alpha_i \pi_i, \operatorname{div} \mathbf{v}) + \sum_{i=1}^n \alpha_i \langle \phi_i, \mathbf{v} \rangle = 0, \\ (q, \operatorname{div} \sum_{i=1}^n \alpha_i \mathbf{w}_i) = 0. \end{cases} \tag{2.13}$$

Now taking in $\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{w}_i$, we obtain

$$\nu \left\| \nabla \left(\sum_{i=1}^n \alpha_i \mathbf{w}_i \right) \right\|^2_{L^2(\Omega)} + \sum_{i=1}^n \alpha_i \left\langle \phi_i, \sum_{j=1}^n \alpha_j \mathbf{w}_j \right\rangle = 0,$$

which implies, for all $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\boldsymbol{\alpha} \neq \mathbf{0}$,

$$\boldsymbol{\alpha}^T B \boldsymbol{\alpha} = -\nu \left\| \nabla \left(\sum_{i=1}^n \alpha_i \mathbf{w}_i \right) \right\|^2_{L^2(\Omega)} \leq -\frac{\nu}{1 + C_p} \left\| \sum_{i=1}^n \alpha_i \mathbf{w}_i \right\|^2_{H^1(\Omega)} < 0. \tag{2.14}$$

In the last relation we have used the Poincaré inequality

$$\int_\Omega \mathbf{v}^2 \leq C_p \int_\Omega |\nabla \mathbf{v}|^2 \quad \forall \mathbf{v} \in V.$$

From (2.14) we infer that B is negative definite and then nonsingular. ∎

Remark 4. Let us point out a difficulty that may be encountered if one wants to impose on the sections S_i a mean value $P \in \mathbb{R}$ for the pressure (or, for the normal stresses) by following a route similar to that presented for the flow rate. The case where one wants to impose a pointwise value for the pressure on the boundary for the Stokes problem has been analyzed in [1]. For the sake of simplicity, we restrict ourselves to the Stokes problem. A possible formulation is as follows: find $(\mathbf{u}, p) \in V \times M$ and $\lambda_0, \dots, \lambda_n \in \mathbb{R}$ such that, for all $(\mathbf{v}, q) \in V \times M$,

$$\begin{cases} \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + (\nabla p, \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle, \\ (\nabla q, \mathbf{u}) - \displaystyle\sum_{i=0}^n \lambda_i \int_{S_i} q \, ds = 0, \\ \displaystyle\int_{S_i} p \, ds = P_i \operatorname{meas}(S_i), \quad i = 0, \dots, n. \end{cases} \tag{2.15}$$

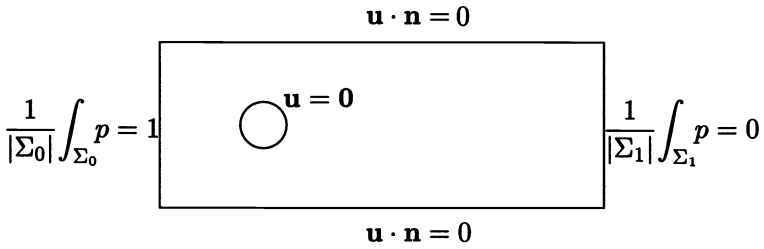


FIG. 2. *Boundary conditions for the Stokes flow around a cylinder.*

This formulation is, in some sense, the dual of the flux formulation as it imposes constraints on the dual problem (the pressure equation), whereas the flux boundary conditions yields a constraint on the primal problem (the velocity equations). Therefore, it can be regarded as the natural counterpart of our formulation for the flux problem. Unfortunately, it may be recognized that from (2.15) it follows that $\mathbf{u}\cdot\mathbf{n}|_{S_i} = \lambda_i$ on each S_i, whereas $\mathbf{u}\cdot\mathbf{n}|_{S_i}$ cannot be a constant different from 0 (since we assume no-slip boundary conditions on Γ). This formulation is therefore not unsuitable for the problem we are interested in. Nevertheless, it may be adopted in those cases where a slip boundary condition is imposed on the wall. In this case system (2.15) will effectively impose a mean value for p, thus differing from the do-nothing approach (1.5) which is instead equivalent to imposing the much stronger condition (1.6).

A numerical test is given hereafter to illustrate how the mean pressure formulation may be used when it is consistent with the velocity boundary conditions. We consider a Stokes flow around a cylinder between two flat plates. We have imposed a homogeneous Dirichlet boundary condition on the cylinder surface, a pure slip condition (i.e., $\mathbf{u}\cdot\mathbf{n} = 0$) on the plates. Moreover, we will prescribe the mean pressure at the inlet and the outlet (see Figure 2). In Figure 3 we show the pressure profile obtained at the inlet. It may be noted how it varies around the imposed mean value of 1.

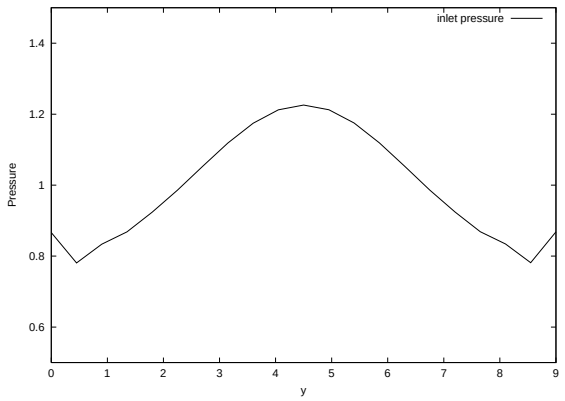


FIG. 3. *Inlet pressure distribution for the Stokes flow around a cylinder. It may be noted that the pressure is not uniform but is distributed around the imposed mean value of* 1.

3. The numerical solution of the Lagrange multipliers problem. In order to discretize formulation (2.7), we introduce a Galerkin approximation based on the finite-dimensional spaces $V_h \subset V$ and $M_h \subset M$, which we assume to satisfy the well-known LBB condition

$$\forall q_h \in M_h \quad \exists \mathbf{v}_h \in V_h,\ \mathbf{v}_h \neq 0 : \quad (q_h, \operatorname{div} \mathbf{v}_h) \geq \beta_h \|q_h\|_{L^2} \|\mathbf{v}_h\|_{H^1}. \tag{3.1}$$

Let $(\mathbf{u}_h, p_h, \lambda_{1h}, \dots, \lambda_{nh})$ be the solution of the discrete problem. We denote by $(u_i)_{i=1\dots dN}$ (resp., $(p_i)_{i=1\dots M}$) the components of $\mathbf{u}_h$ (resp., p_h) with respect to a basis $\{\mathbf{v}_i\}$ of V_h (resp., $\{q_i\}$ of M_h). Finally, we introduce the vectors $U = (u_1, \dots, u_{dN}) \in \mathbb{R}^{dN}$, $P = (p_1, \dots, p_M) \in \mathbb{R}^M$, and $\Lambda = (\lambda_{1h}, \dots, \lambda_{nh}) \in \mathbb{R}^n$.

Then the discrete counterpart of (2.7) gives rise to the following algebraic system of equations:

$$\begin{cases} AU + D^T P + \Phi^T \Lambda = F, \\ D\,U = 0, \\ \Phi\,U = Q, \end{cases} \tag{3.2}$$

where $A \in \mathbb{R}^{dN \times dN}$ is the stiffness matrix, $D \in \mathbb{R}^{M \times dN}$ is the matrix associated with the divergence operator, and Φ is the $n \times dN$ matrix whose lines are given by the vectors $\phi_i = (\int_{S_i} \mathbf{v}_1 \cdot \mathbf{n}\, ds, \dots, \int_{S_i} \mathbf{v}_{dN} \cdot \mathbf{n}\, ds)$, $i = 1, \dots, n$.

PROPOSITION 3.1. *System* (3.2) *admits a unique solution.*

Proof. The proof of the existence of the discrete solution $(\mathbf{u}_h, p_h, \lambda_{ih},\ i = 1, \dots, n)$ as well as that of the uniqueness of $\mathbf{u}_h$ and of the λ_{ih} follows the same lines of Proposition 2.2 by substituting V and M by their discrete counterpart. The uniqueness of p_h is assured by condition (3.1). □

If we discretize in time the Navier–Stokes system (2.3) by, for instance, a semi-implicit Euler scheme and then in space by the finite element method, we will produce an algebraic system analogous to (3.2), where the matrix A is now given by

$$A = \frac{1}{\delta t} M + B + K,$$

δt being the time step, M and B the mass and advection matrices, and K the stiffness matrix. Here and in the following it is assumed that the matrix A is positive definite, which is always the case if δt is chosen appropriately.

3.1. Solution algorithms. Here, we present and analyze four possible algorithms which may be adopted for the solution of system (3.2) and which are computationally more efficient than solving simultaneously for U, P, and Λ.

1. *Solution of additional Stokes problems.* If one wishes to introduce the proposed approach on a Navier–Stokes solver with as few modifications as possible to an existing code for Navier–Stokes equations with "classical" boundary conditions, a possibility is to follow the constructive proof of Proposition 2.2. As we have seen, the solution of the constrained problem can be obtained by combining the solutions of $n + 1$ unconstrained Stokes problems, given by (2.8) and (2.9). In particular, for a fixed geometry, the n solutions $\mathbf{w}_i$, $i = 1, \dots, n$, of problem (2.9) can be computed only once, so that the additional computational cost at each time step is just that of the solution of (2.8). The drawback is that the memory requirement to store the $\mathbf{w}_i$ may become prohibitive particularly in a three-dimensional computation and with a large number of Lagrange multipliers.

2. *Schur complement + Iterative solver.* An alternative algorithm is based on using an iterative solution of a Schur complement system. We rewrite (3.2) in the form

$$\begin{bmatrix} S & \tilde{\Phi}^T \\ \tilde{\Phi} & 0 \end{bmatrix} \cdot \begin{bmatrix} X \\ \Lambda \end{bmatrix} = \begin{bmatrix} G \\ Q \end{bmatrix}, \tag{3.3}$$

where $\tilde{\Phi} = [\Phi, 0] \in \mathbb{R}^{n \times (dN+M)}$, $X = [U, P]^T$, $G = [F, 0]^T$. The matrix

$$S = \begin{bmatrix} A & D^T \\ D & 0 \end{bmatrix}$$

has a standard Stokes form and, since the two discrete spaces V_h and M_h satisfy the LBB condition (3.1), S is nonsingular (see, e.g., [19, 2]). We can then eliminate the unknown X from (3.3), obtaining

$$\tilde{\Phi} S^{-1} \tilde{\Phi}^T \Lambda = \tilde{\Phi} S^{-1} G - Q, \tag{3.4}$$

which can be solved by an appropriate iterative method. Any matrix-vector multiplication will imply the solution of a Stokes problem with homogeneous Neumann conditions on the sections S_i.

If A is symmetric, then $R = \tilde{\Phi} S^{-1} \tilde{\Phi}^T$ is symmetric and positive definite (see Proposition 3.2 below). Consequently, the conjugate gradient (CG) algorithm may be used, which will converge to the exact solution in n iterations, n being the number of Lagrange multipliers (which coincides with the number of boundary sections S_i minus one). For instance, in the case of just one Lagrange multiplier, one iteration of the CG algorithm suffices to obtain the solution. (Note that in this case the linear system (3.4) reduces to just one scalar equation.)

Remark 5. The computational cost of the procedure depends on the number of matrix-vector multiplications required for every iteration of the chosen iterative solver and on the number of iterations necessary for convergence. For each matrix-vector multiplication we need to solve a Stokes problem. In the case of the CG algorithm, it is known that it converges to the exact solution in n steps. In addition, two extra Stokes problems have to be solved to obtain the initial residual (required to start up the procedure) and the final solution X. Therefore, if CG is adopted the computational cost would be equal to the solution of $n + 2$ Stokes problems (at each time step), which is higher than that of procedure 1. On the other hand, there is no need to store intermediate solutions.

PROPOSITION 3.2. *The matrix $R = \tilde{\Phi} S^{-1} \tilde{\Phi}^T$ is positive semidefinite; moreover, if A is symmetric, then R is symmetric and positive definite.*

Proof. System (3.3) (which is equivalent to system (3.2)) admits a unique solution, as shown in Proposition 3.1. Then, we necessarily have that $\ker(\tilde{\Phi}^T) = \{0\}$.

The matrix

$$S^* = \begin{bmatrix} A & D^T \\ -D & 0 \end{bmatrix}$$

is positive semidefinite (see, e.g., [19]), thus S^{*-1} and $R^* = \tilde{\Phi} S^{*-1} \tilde{\Phi}^T$ are also positive semidefinite.

On the other hand, $S = PS^*$, where

$$P = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix},$$

is such that $P^{-1} = P$. Then

$$R = \tilde{\Phi} S^{-1} \tilde{\Phi}^T = \tilde{\Phi} (P S^*)^{-1} \tilde{\Phi}^T = \tilde{\Phi} S^{*-1} P \tilde{\Phi}^T = \tilde{\Phi} S^{*-1} \tilde{\Phi}^T = R^*.$$

We conclude that also R is positive semidefinite. Moreover, if A is symmetric, R turns out to be symmetric, positive semidefinite, and nonsingular, so it is also positive definite. □

In the case where there is just one Lagrange multiplier (which occurs whenever we have just one input section and one output section), the CG algorithm reads as follows:

given $\lambda_0 \in \mathbb{R}$,

$$\begin{aligned}
&\text{(i)} && SX_1 = G - \tilde{\Phi}^T \lambda_0, \\
&\text{(ii)} && r_0 = \tilde{\Phi} X_1 - Q, \\
&\text{(iii)} && SX_2 = \tilde{\Phi}^T r_0, \\
&\text{(iv)} && \lambda = \lambda_0 + \frac{r_0^2}{r_0 \tilde{\Phi} X_2} r_0 = \lambda_0 + \frac{r_0^2}{\tilde{\Phi} X_2}, \\
&\text{(v)} && SX = G - \tilde{\Phi}^T \lambda.
\end{aligned}$$

Since in this case the CG method converges in one iteration, λ and X are the solutions of (3.3). This algorithm requires the solution of 3 Stokes problems at steps (i), (iii), and (v).

Remark 6. By a closer inspection, it can be noted that by taking $\lambda_0 = 0$, the CG algorithm just presented effectively reduces to procedure 1.

3. *Reordering + fractional step* I. We recall that any Stokes system of the form

$$\begin{bmatrix} A & D^T \\ D & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}$$

can be solved exactly by the following three step algorithm:

(i) $AU_0 = F$,
(ii) $DA^{-1}D^T P = DU_0$,
(iii) $U = U_0 - A^{-1} D^T P$.

Let $H^{(1)}$ and $H^{(2)}$ denote two suitable approximations of A^{-1}. We can write an approximate factorization scheme as

(i) $AU_0 = F$,
(ii) $DH^{(1)}D^T P = DU_0$,
(iii) $U = U_0 - H^{(2)} D^T P$.

In this way we can recover many projection or quasi-compressibility methods for the Navier–Stokes equations (see [14, 17]). In particular, we will focus on the Yosida projection scheme [16], which consists of adopting $H^{(1)} = \delta t M^{-1}$ (while no approximation is made in step (iii), i.e., $H^{(2)} = A^{-1}$), and on the algebraic version of the Chorin–Temam scheme in which we have $H^{(1)} = H^{(2)} = \delta t M^{-1}$ [17].

System (3.2) can be reordered in a Stokes-like form as

$$\begin{bmatrix} \tilde{A} & \tilde{D}^T \\ \tilde{D} & 0 \end{bmatrix} \begin{bmatrix} \tilde{U} \\ P \end{bmatrix} = \begin{bmatrix} \tilde{F} \\ 0 \end{bmatrix}, \tag{3.5}$$

where

$$\tilde{A} = \begin{bmatrix} A & \Phi^T \\ \Phi & 0 \end{bmatrix} \in \mathbb{R}^{(dN+n)\times(dN+n)}, \qquad \tilde{D} = \begin{bmatrix} D & 0 \end{bmatrix} \in \mathbb{R}^{M\times(dN+n)},$$
$$\tilde{U} = \begin{bmatrix} U \\ \Lambda \end{bmatrix} \in \mathbb{R}^{dN+n}, \qquad \tilde{F} = \begin{bmatrix} F \\ Q \end{bmatrix} \in \mathbb{R}^{dN+n}.$$

We can then write an approximate factorization scheme as follows:

(i) $\tilde{A}\tilde{U}_0 = \tilde{F}$. Since A is positive definite and $\ker(\Phi^T) = \emptyset$, $\tilde{A}$ is nonsingular, too, and this system admits a unique solution. In particular, we have $\Phi U_0 = Q$.
(ii) $\tilde{D}\tilde{H}^{(1)}\tilde{D}^T P = \tilde{D}\tilde{U}_0$,
(iii) $\tilde{U} = \tilde{U}_0 - \tilde{H}^{(2)}\tilde{D}^T P$, where now $\tilde{H}^{(1)}$ and $\tilde{H}^{(2)}$ are possible approximations of $\tilde{A}^{-1}$.

We are now in the same setting of factorization schemes for the Stokes problem. We will then use the term *Yosida* scheme when $\tilde{A}^{-1}$ is approximated only in step (ii), while a scheme where $\tilde{H}^{(1)} = \tilde{H}^{(2)} \neq \tilde{A}^{-1}$ will be called a *Chorin–Temam* scheme.

We now detail a possible way to approximate $\tilde{A}^{-1}$. First, we note that if we write $\tilde{H}^{(1)}$ in the block form

$$\tilde{H}^{(1)} = \begin{bmatrix} H_{11}^{(1)} & H_{12}^{(1)} \\ H_{21}^{(1)} & H_{22}^{(1)} \end{bmatrix},$$

step (ii) of the algorithm is equivalent to

$$DH_{11}^{(1)}D^T P = DU_0,$$

where just the first diagonal block of $\tilde{H}^{(1)}$ is actually involved in the computation. Therefore, we need only look for an approximation $H_{11}^{(1)}$ of the corresponding term C_{11} of the following block decomposition of $\tilde{A}^{-1}$:

$$\tilde{A}^{-1} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

The term C_{11} is equal to

$$C_{11} = A^{-1}\left(I - \Phi^T V^{-1}\Phi A^{-1}\right), \quad \text{where } V = \Phi A^{-1}\Phi^T.$$

A natural approximation of C_{11} is then

$$H_{11}^{(1)} = \delta t M^{-1}\left(I - \delta t \Phi^T V_{appr}^{-1}\Phi M^{-1}\right), \quad V_{appr} = \delta t \Phi M^{-1}\Phi^T, \tag{3.6}$$

and in particular we have $H_{11}^{(1)} = C_{11} + O(\delta t^2)$. The matrix V_{appr} is an $n \times n$ matrix. In general, the number n of Lagrange multipliers is very small so that V_{appr} can be easily inverted. Furthermore, if the lumped form of the mass matrix is used, V_{appr} is diagonal. In the case of just one Lagrange multiplier, V_{appr} reduces to a scalar.

The Yosida scheme then becomes

(i) $\tilde{A}\tilde{U}_0 = \tilde{F}$,
(ii) $DH_{11}^{(1)}D^T P = DU_0$, and
(iii) $\tilde{A}\tilde{U} = \tilde{A}\tilde{U}_0 - \tilde{D}^T P$.

We observe that, in this case, we recover exactly the constraints on the fluxes. Indeed, step (iii) implies $\Phi U = \Phi U_0 = Q$ in particular.

In the Chorin–Temam case, we note that step (iii) is equivalent to

$$U = U_0 - H_{11}^{(1)} D^T P, \tag{3.7}$$

$$\Lambda = \Lambda_0 - H_{21}^{(1)} D^T P. \tag{3.8}$$

Since we are interested only in the velocity field, we can neglect (3.8) and we need only compute the block $H_{11}^{(1)}$ as in (3.6). The algebraic Chorin–Temam scheme then becomes

(i) $\tilde{A}\tilde{U}_0 = \tilde{F}$,
(ii) $DH_{11}^{(1)} D^T P = DU_0$, and
(iii) $U = U_0 - H_{11}^{(1)} D^T P$.

We observe that, also in this case, the constraints on the fluxes are recovered exactly. Indeed, by multiplying step (iii) by Φ we have

$$\begin{aligned}\Phi U &= \Phi U_0 - \Phi H_{11}^{(1)} D^T P \\ &= \Phi U_0 - \left(\delta t \Phi M^{-1} - \delta t V_{appr} V_{appr}^{-1} \Phi M^{-1}\right) D^T P = \Phi U_0 = Q.\end{aligned}$$

4. *Reordering + fractional step* II. System (3.2) can also be reordered in a different manner as

$$\begin{bmatrix} A & \tilde{D}^T \\ \tilde{D} & 0 \end{bmatrix} \begin{bmatrix} U \\ \tilde{P} \end{bmatrix} = \begin{bmatrix} F \\ \tilde{Q} \end{bmatrix}, \tag{3.9}$$

where

$$\tilde{D} = \begin{bmatrix} D \\ \Phi \end{bmatrix} \in \mathbb{R}^{(M+n)\times dN}, \qquad \tilde{P} = \begin{bmatrix} P \\ \Lambda \end{bmatrix} \in \mathbb{R}^{M+n}, \qquad \tilde{Q} = \begin{bmatrix} 0 \\ Q \end{bmatrix} \in \mathbb{R}^{M+n}.$$

The three step algorithm then reads as

(i) $AU_0 = F$, which is unperturbed with respect to the Stokes system without constraints,
(ii) $\tilde{D} H^{(1)} \tilde{D}^T \tilde{P} = \tilde{D} U_0 - \tilde{Q}$,
(iii) $U = U_0 - H^{(2)} \tilde{D}^T \tilde{P}$.

Again, we consider the approximation $H^{(1)} = \delta t M^{-1}$ and either $H^{(2)} = A^{-1}$ (Yosida) or $H^{(2)} = \delta t M^{-1}$ (algebraic Chorin–Temam).

Remark 7. This algorithm can be easily implemented starting from an existing Navier–Stokes solver which uses factorization methods. It suffices to add to the matrix D the few lines of matrix Φ and apply the chosen factorization method.

Remark 8. Step (ii) is equivalent to

$$DH^{(1)}(D^T P + \Phi^T \Lambda) = DU_0, \tag{3.10}$$

$$\Phi H^{(1)}(D^T P + \Phi^T \Lambda) = \Phi U_0 - Q. \tag{3.11}$$

On the other hand, the third step gives

$$U = U_0 - H^{(2)}(D^T P + \Phi^T \Lambda),$$

from which we can infer that

$$\begin{aligned}\Phi U &= \Phi U_0 - \Phi H^{(2)}(D^T P + \Phi^T \Lambda) \\ &= \Phi U_0 - \Phi H^{(1)}(D^T P + \Phi^T \Lambda) + \Phi(H^{(1)} - H^{(2)})(D^T P + \Phi^T \Lambda).\end{aligned} \tag{3.12}$$

By exploiting (3.11), we finally have

$$\Phi U = Q + \Phi(H^{(1)} - H^{(2)})(D^T P + \Phi^T \Lambda). \tag{3.13}$$

Whenever $H^{(1)} = H^{(2)}$, like in the algebraic Chorin–Temam scheme, we recover the constraint on the fluxes exactly.

On the contrary, in the Yosida scheme, (3.13) becomes

$$\Phi U = Q + \Phi(\delta t M^{-1} - A^{-1})(D^T P + \Phi^T \Lambda) = Q + O(\delta t^2).$$

4. Numerical tests and algorithm assessment.

4.1. Womersley flow. In order to assess the proposed methodology, we consider a case where the analytical solution of the Navier–Stokes equations is known. More precisely, we consider the *Womersley* solution, which describes the transient flow in a cylindrical pipe associated to a time-periodic pressure gradient (see, e.g., [12]). As such, it may be considered as a transient counterpart of the Poiseuille solution.

If the pressure gradient is given by

$$\nabla p = \frac{dp}{dz}(t)\mathbf{e}_z = -\rho a \sin(\omega t)\mathbf{e}_z,$$

z being the pipe axial coordinate and ρ the fluid density, the velocity $\mathbf{u}$ reduces only to its axial component, i.e., $\mathbf{u} = u_z \mathbf{e}_z$, and the analytical expression for u_z is as follows:

- In the *two-dimensional* $(2D)$ *case* (flow between two infinite planes),

$$u_z(r,t) = \sum_0^\infty \gamma_{2k+1} \sin\left(\frac{(2k+1)\pi}{2r_0} r\right),$$

  where

$$\gamma_l = \frac{4a}{\pi l(l^4\sigma^2 + \omega^2)} \left(l^2\sigma \sin(\omega t) + \omega e^{-l^2\sigma t} - \omega \cos(\omega t)\right).$$

  Here $\sigma = \frac{\mu\pi^2}{4\rho r_0^2}$, r is the transverse coordinate, $2r_0$ the distance between the two planes, and μ the dynamic fluid viscosity.

- In the *three-dimensional* $(3D)$ *case* (flow in a cylindrical pipe),

$$u_z(r,t) = \mathrm{Re}\left\{ -\frac{a}{\omega}\left(1 - \frac{J_0\left(i^{3/2}\sqrt{\frac{\rho\omega}{\mu}} r\right)}{J_0\left(i^{3/2}\sqrt{\frac{\rho\omega}{\mu}} r_0\right)}\right) e^{i\omega t} \right\},$$

  where r is the radial coordinate, r_0 the cylinder radius, and J_0 the Bessel function of first kind and of order zero.

In both the 2D and 3D test cases, we have imposed homogeneous Neumann boundary conditions at the inflow, while at the outflow we have prescribed the flow rate associated to the Womersley solution. In Figure 4 we show the axial velocity field for the 2D case at two different times, together with the velocity profile at the inflow. The solution obtained agrees very well with the analytical Womersley solution. Therefore, a single condition on the flow rate at the outflow, imposed through a Lagrange multiplier, is sufficient to recover the Womersley flow.

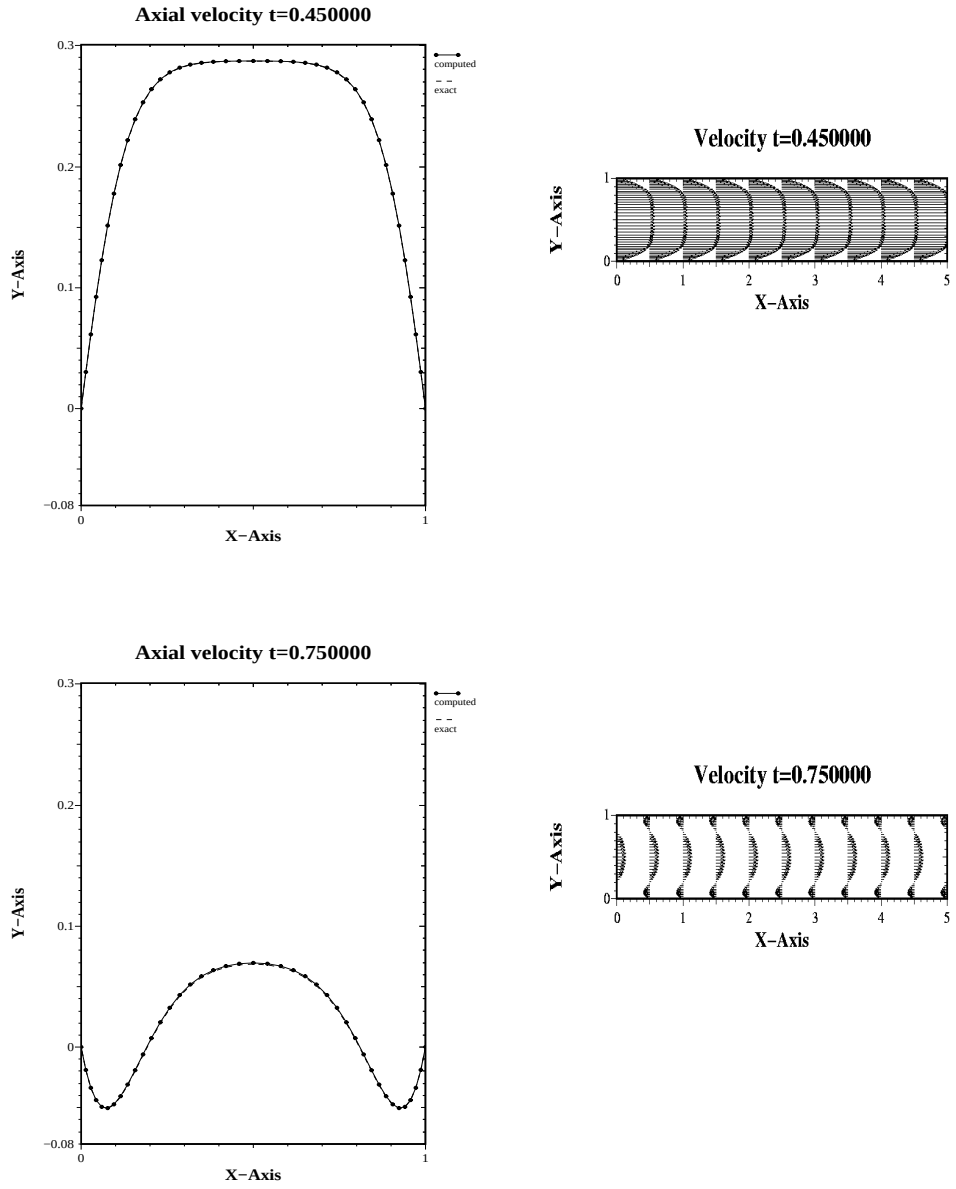


FIG. 4. *2D numerical solution obtained imposing the flux of the Womersley solution at the outflow section.*

The same experiment has been carried out in 3D and the result is shown in Figure 5. Here, the computed velocity field at three different times is illustrated, together with the corresponding axial velocity profile on the inflow section. Again, we outline the excellent agreement with the analytical solution.

Finally, we have carried out the same experiment in 2D using the numerical schemes "reordering + fractional step I" and "reordering + fractional step II" proposed in the previous section, both for the Yosida and the algebraic Chorin–Temam approximations, and we have evaluated the errors introduced on the fluxes. As expected, for the first scheme the difference between the flux we wish to impose and the one actually computed is of the order of the machine round-off error, both for the Yosida and the Chorin–Temam approximation.

For the latter scheme, this instead is true only when adopting the Chorin–Temam approximation. The behavior of the error on the fluxes for the Yosida approximation is shown in Figure 6. The error is decreasing with the time step size, with a convergence rate that appears to be even higher than quadratic.

4.2. Mass conservation in free interface simulations. We present here an application where it may be useful to impose the mass flow rate through a surface. We consider two immiscible and incompressible fluids, with the same viscosity, confined in a closed tank Ω and separated by a free interface (see Figure 7). We denote by $\Omega_i(t)$ and $\Sigma(t)$, for $i = 1, 2$, the domain occupied by the fluid i and the interface at time t, respectively. We adopt an arbitrary Lagrangian Eulerian (ALE) formulation [7] and we denote by $\mathbf{w}$ the domain velocity which satisfies $\mathbf{w} \cdot \mathbf{n} = \mathbf{u} \cdot \mathbf{n}$ on $\Sigma(t)$ and $\mathbf{w} \cdot \boldsymbol{\nu} = 0$ on $\partial\Omega$, where $\mathbf{n}$ denotes the normal to $\Sigma(t)$ directed from $\Omega_1(t)$ to $\Omega_2(t)$ and $\boldsymbol{\nu}$ is the outward normal on $\partial\Omega$. Because of the incompressibility and the immiscibility of the two fluids, the volume of $\Omega_1(t)$ (or equivalently $\Omega_2(t)$) must be preserved. At the continuous level, this property is satisfied. Indeed,

$$\begin{aligned} \operatorname{meas}(\Omega_1(t_2)) - \operatorname{meas}(\Omega_1(t_1)) &= \int_{t_1}^{t_2}\!\!\int_{\Sigma(t)} \mathbf{w} \cdot \mathbf{n}\, d\sigma = \int_{t_1}^{t_2}\!\!\int_{\Sigma(t)} \mathbf{u} \cdot \mathbf{n}\, d\sigma \\ &= \int_{t_1}^{t_2}\!\!\int_{\Omega_1(t)} \operatorname{div} \mathbf{u}\, dx = 0, \end{aligned} \tag{4.1}$$

since $\operatorname{div} \mathbf{u}$ vanishes almost everywhere in $\Omega_1(t)$.

At the discrete level, the relation $\int_{\Omega_1(t)} \operatorname{div} \mathbf{u}_h\, dx = 0$ is still verified if the pressure is discretized using *discontinuous* functions (as in the Q2/P1 or Q1/P0 finite elements) [2].

If instead the pressure is discretized using continuous functions, as in Taylor–Hood (P2/P1 or Q2/Q1), P1-isoP2, or Q1/Q1 stabilized finite elements [2, 9], there is no guarantee that (4.1) still holds at the discrete level. Numerical tests indeed confirm that those discretizations fail to conserve the measure of $\Omega_1(t)$. A possible strategy for the solution of this problem is to impose the condition

$$\int_{\Sigma(t)} \mathbf{u}_h \cdot \mathbf{n}_h\, d\sigma = 0$$

by a Lagrange multiplier, using the techniques presented in the previous section.

Let us show the results obtained on a 2D test case. In the following, all quantities are given in International System (IS) units. The two fluids are initially at rest and they are subjected to an oscillating body force $\mathbf{f} = (a\, g \sin(2\pi\nu\, t), -g)$ with $a = 0.05$,

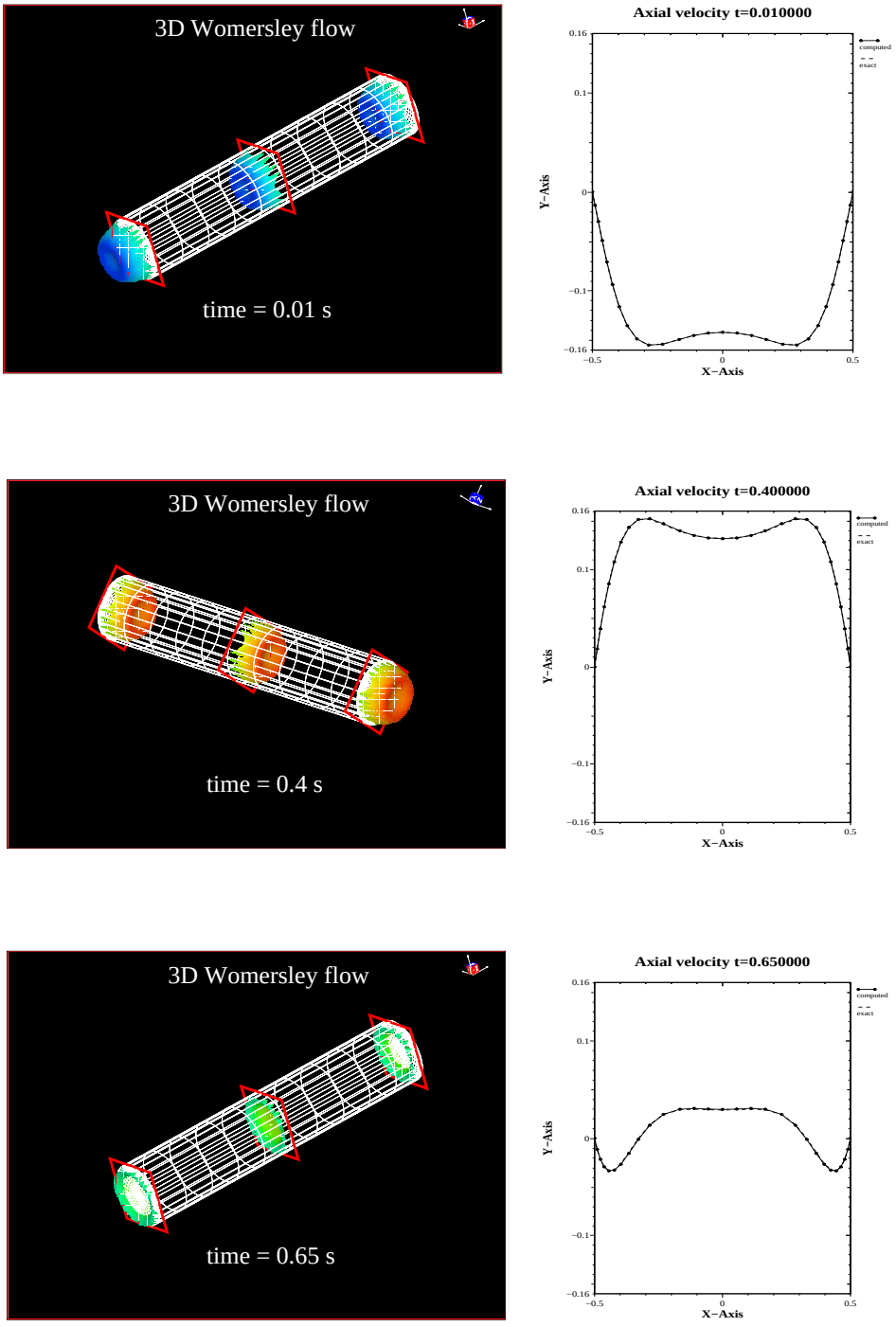


FIG. 5. *3D numerical solution obtained imposing the flux of the Womersley solution at the outflow section.*

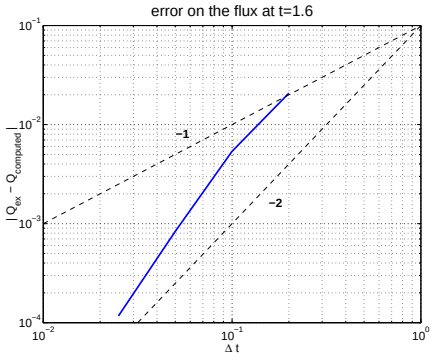


FIG. 6. *Behavior of the error on the flux for the scheme "reordering + fractional step* II*" with the Yosida approximation; the dotted lines are the reference lines for the error decrease rate.*

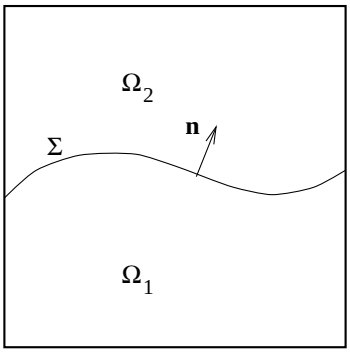


FIG. 7. *Two fluids separated by a free interface.*

$g = 10$, $\nu = 0.0625$. The kinematic viscosity of both fluids is taken to be $\nu = 0.005$, the density of the upper and the lower fluid is 0.91 and 1, respectively, $\text{meas}(\Omega_1) = \text{meas}(\Omega_2) = 4$ at $t = 0$. The mesh is allowed to move only along the vertical direction and the typical size of the mesh elements is $h = 0.1$. At time $t = 220$, if we use Q1/P0 or Q2/P1 elements, $\text{meas}(\Omega_1)$ is still equal to 4 (within machine precision), whereas it drops to approximately 3.9 when we adopt Q2/Q1 elements (see Figure 8). We have obtained analogous results with stabilized Q1/Q1 finite elements. Clearly, this lack of mass conservation decreases as h goes to zero, yet for many practical applications a mass loss is not acceptable and the use of an extremely fine mesh is not economical (or even not feasible).

Figure 8 shows that a perfect mass conservation is also obtained with Q2/Q1 elements if we impose a zero flow rate through Σ by the Lagrange multiplier technique. Finally, Figure 9 shows the elevation of a point on the interface obtained on the same mesh with the Q2/P1 elements and the Q2/Q1 elements with flux constraint. The difference is barely visible. The use of a Lagrange multiplier technique thus allows to adopt continuous pressure elements for this type of problem.

4.3. Multiscale domain decomposition. An application in which it is necessary to impose defective boundary conditions to a Navier–Stokes problem arises in the hemodynamics context when the cardiovascular system is simulated by a multiscale model. A multiscale technique couples detailed models, based on the solution

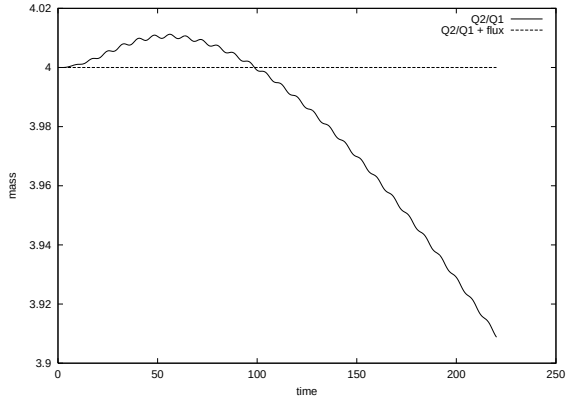


FIG. 8. *Mass conservation fails using $Q2/Q1$ elements, while it holds if we add a constraint on the velocity flux.*

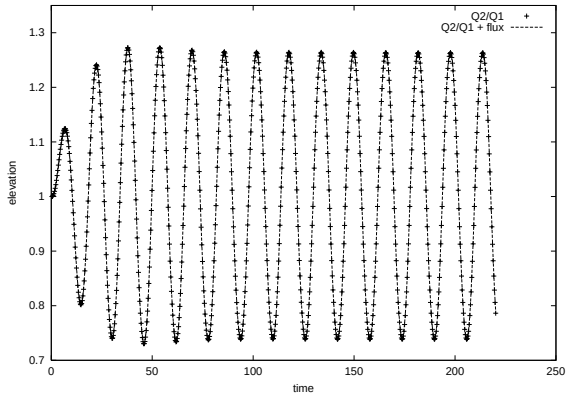


FIG. 9. *The time history of the elevation of a point of the interface. The results obtained with $Q2/P1$ elements are almost the same of that given by $Q2/Q1$ elements with flux constraint.*

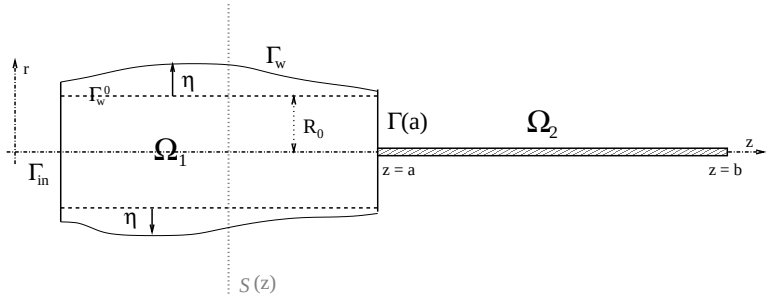


FIG. 10. *Coupled $2D$-$1D$ problem. On the left, the $2D$ model (Ω_1), where Γ_w represents the arterial wall and η the wall displacement with respect to the reference configuration Γ_w^0. On the right, the $1D$ model (Ω_2) defined on the interval $a \leq z \leq b$.*

of 2D or 3D fluid-structure interaction problems, with reduced models based on one-dimensional (1D) approximations or on systems of ordinary differential equations [8]. The simpler models normally provide the evolution of mean pressure and mean velocity in various regions of the cardiovascular system. The boundary data for the detailed model, which is based on the solution of the Navier–Stokes equations coupled with the vessel wall dynamics, must be obtained from these averaged quantities. This is a typical case of defective boundary conditions.

4.3.1. A 2D-1D coupling. Here we will give an illustrative example which consists in the coupling of a 2D and a 1D model. Let us consider the domain illustrated in Figure 10. In Ω_1 we define for $t > 0$ the fluid-structure model as

$$
\begin{cases}
\rho\partial_t \mathbf{u} + \rho\mathbf{u}\cdot\nabla\mathbf{u} + \nabla p - \mu\triangle\,\mathbf{u} = 0 & \text{in } \Omega_1,\\
\operatorname{div}\mathbf{u} = 0 & \text{in } \Omega_1,\\
\rho_w h\dfrac{\partial^2\eta}{\partial t^2} - kGh\dfrac{\partial^2\eta}{\partial z^2} + \dfrac{Eh}{1-\nu^2}\dfrac{\eta}{R_0^2} - \gamma\dfrac{\partial^3\eta}{\partial z^2\partial t} = f(t,z) & \text{on } \Gamma_w^0,\\
\partial_t\eta\,\mathbf{e}_r = \mathbf{u} & \text{on } \Gamma_w,\\
f(t,z) = \left(p\mathbf{n} - \mu\dfrac{\partial\mathbf{u}}{\partial\mathbf{n}}\right)\cdot\mathbf{e}_r\sqrt{1+\left(\dfrac{\partial\eta}{\partial z}\right)^2} & \text{on } \Gamma_w^0,
\end{cases}
\tag{4.2}
$$

where the unknowns are the fluid velocity $\mathbf{u}$, the fluid pressure p, and the wall displacement η. Here, ρ is the fluid density, h the vessel wall thickness, E the Young modulus, G the Timoshenko factor, and ρ_w the wall density. At $t = 0$ initial conditions $\mathbf{u}_0$, η_0, $\dot\eta_0$ are provided for the velocity, displacement, and displacement rate, respectively. We refer to [6] for a more detailed description and analysis of this problem. In Ω_2 we consider the following 1D problem for the velocity flux Q and the vessel section area A:

$$
\begin{cases}
\dfrac{\partial A}{\partial t} + \dfrac{\partial Q}{\partial z} = 0, & a < z < b,\ \ t > 0,\\
\dfrac{\partial Q}{\partial t} + \dfrac{\partial}{\partial z}\left(\alpha\dfrac{Q^2}{A}\right) + \dfrac{A}{\rho}\dfrac{\partial\bar p}{\partial z} + K_R\dfrac{Q}{A} = 0, & a < z < b,\ \ t > 0,
\end{cases}
\tag{4.3}
$$

with the algebraic relation $\bar p = \beta(A - A_0)$, A_0 being the reference area $A_0 = 2R_0$. The system is supplemented by initial conditions for A and Q at $t = 0$. This 1D reduced model is basically derived from (4.2), integrating the Navier–Stokes equations over each axial section $\mathcal{S}(z)$ and adopting a simplified version of the equation for the wall dynamics. A is the measure of $\mathcal{S}(z)$, the velocity flux is given by $Q(z) = \int_{\mathcal{S}(z)} u_z dr$, and $\bar p(z) = (\int_{\mathcal{S}(z)} p dr)/A(z)$; see [8] for more details.

System (4.2) has been discretized in space using P1-isoP2 finite elements for the fluid and P1 elements for the structure. For time discretization, we have adopted an ALE formulation to account for the domain movement with an implicit Euler discretization for the fluid equations and a Newmark scheme for the structure. On the other hand, system (4.3), which is hyperbolic, has been discretized using a second-order Taylor–Galerkin scheme with a characteristic treatment of the boundary.

At each time step t^n, we look for a solution of (4.2) and (4.3) which satisfies at $\Gamma(a)$ the coupling conditions

$$
\operatorname{meas}(\Gamma^n(a)) = A^n(a), \quad \int_{\Gamma^n(a)} u_z^n = Q^n(a), \quad \frac{1}{\operatorname{meas}(\Gamma^n(a))}\int_{\Gamma(a)} p^n = \bar p^n(a).
\tag{4.4}
$$

We have solved iteratively at each time step the two subproblems in Ω_1 and Ω_2. Given the approximate solution $\mathbf{u}^n$, p^n, η^n, Q^n, and A^n of the coupled problem at time $t = t^n$, we look for the solution $\mathbf{u}^{n+1}$, p^{n+1}, η^{n+1}, Q^{n+1}, and A^{n+1} using the following iterative algorithm:

We set $\mathbf{u}_{(0)} = \mathbf{u}^n$, $p_{(0)} = p^n$, and $\eta_{(0)} = \eta^n$, and for $k = 0, 1, \ldots$ we do the following:

1. We solve the 1D model (4.3) imposing at $z = a$

$$A_{(k+1)}(a) = A_0 + \frac{1}{\beta \,\mathrm{meas}(\Gamma_{(k)}(a))} \int_{\Gamma(a)} p_{(k)}$$

   and at $z = b$ absorbing boundary conditions based on characteristic analysis. We obtain $Q_{(k+1)}$ and $A_{(k+1)}$ in Ω_2.
2. We then solve the 2D problem imposing on $\Gamma(a)$ for the Navier–Stokes equations the defective condition

$$\int_{\Gamma(a)} \mathbf{u}_{(k+1)} \cdot \mathbf{e}_z = Q_{(k+1)}(a)$$

   and for the structure at $z = a$

$$\eta_{(k+1)}(a) = \frac{1}{2} A_{(k+1)}(a) - R_0.$$

   We obtain $\mathbf{u}_{(k+1)}$, $p_{(k+1)}$, $\eta_{(k+1)}$ in Ω_1.

We iterate until the coupling conditions are satisfied within a fixed tolerance and we finally set the solution at time t^{n+1} equal to the converged value. We may eventually add a relaxation step on the variable $A_{(k)}(a)$.

We observe that in step 2 of this algorithm we have to solve Navier–Stokes equations with flux boundary conditions on $\Gamma(a)$. A different algorithm for the same coupled 2D/1D problem, which allows us to impose a mean pressure condition on $\Gamma(a)$, has been proposed and analyzed in [6].

We present here the numerical results relative to the following test case: we have considered a fluid initially at rest and we have imposed a pressure of $15 mmHg$ ($2{\cdot}10^4\, dynes/cm^2$) at the inlet ($\Gamma_{in}$) for 0.005 seconds. For the fluid we have taken $\mu = 0.035\, poise$ and $\rho = 1\, g/cm^3$, while for the structure we have $E = 0.75{\cdot}10^6\, dynes/cm^2$, $\nu = 0.5$, $\rho_w = 1.1\, g/cm^3$, and $h = 0.1\, cm$. Figure 11 shows the fluid pressure and the domain deformation at different times. We may note how the "pressure wave" crosses the interface between the two models with little spurious reflections.

4.3.2. A two-dimensional–zero-dimensional coupling. This time, a by-pass anastomosis in a coronary, modeled by the incompressible Navier–Stokes equations in a fixed domain, is coupled with a lumped parameter model for the rest of the cardiovascular system. Lumped parameters models are well established tools [21] and able to provide an approximation of the time evolution of average pressure and flow rate in different compartments of the cardiovascular system. They are based on the solution of a system of algebraic-ordinary differential equations derived by using an analogue with an electrical circuit. In this analogy, electrical currents and voltage are interpreted as flow rate and mean pressure, respectively. The model here adopted is the one proposed in [11]. Its coupling with a 2D description of a coronary by-pass is shown in Figure 12.

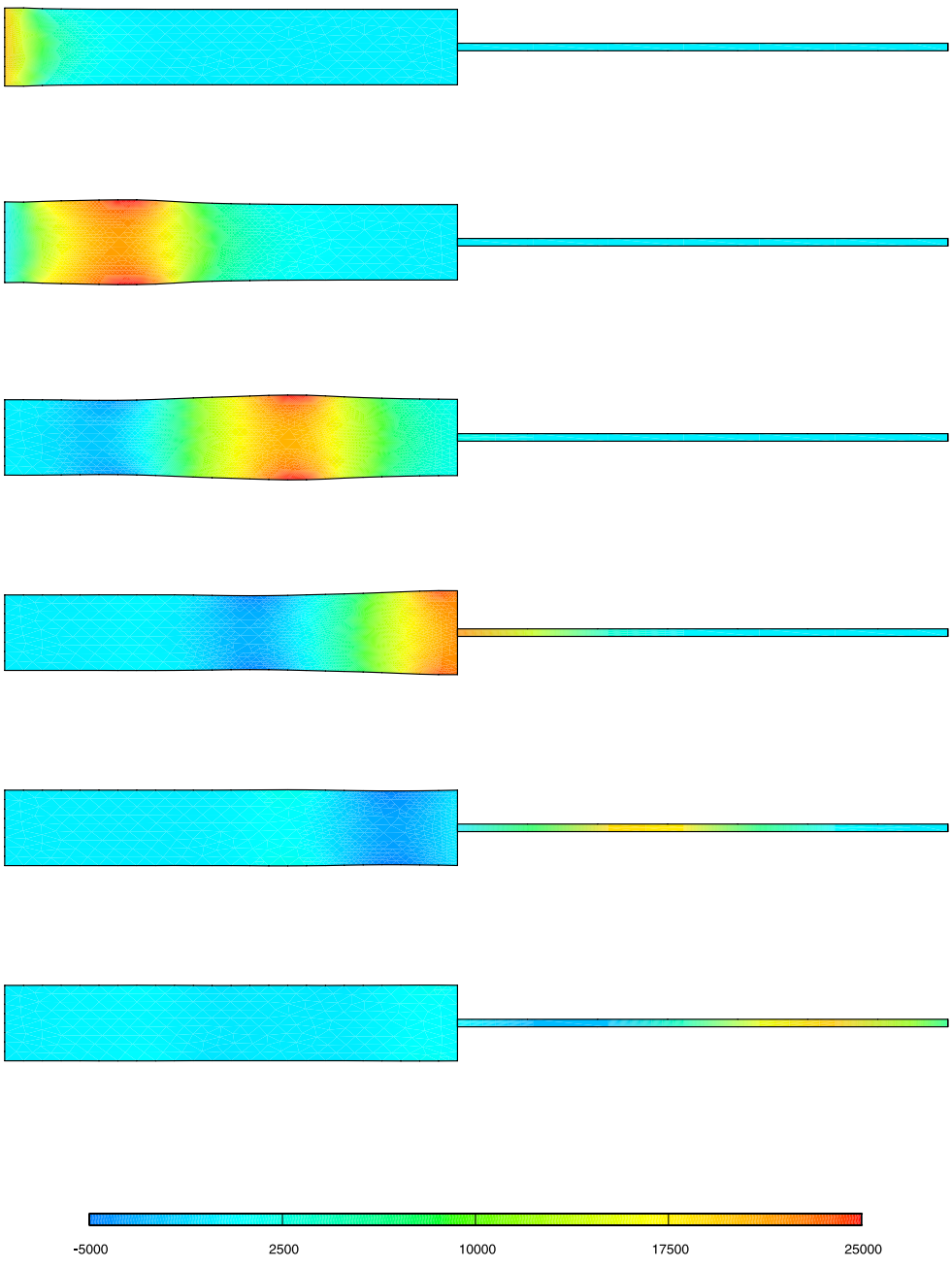


FIG. 11. *Coupling* $2D$ *simulation with the* $1D$ *reduced model; pressure distribution every* $5\,ms$*, starting from* $t = 1\,ms$.

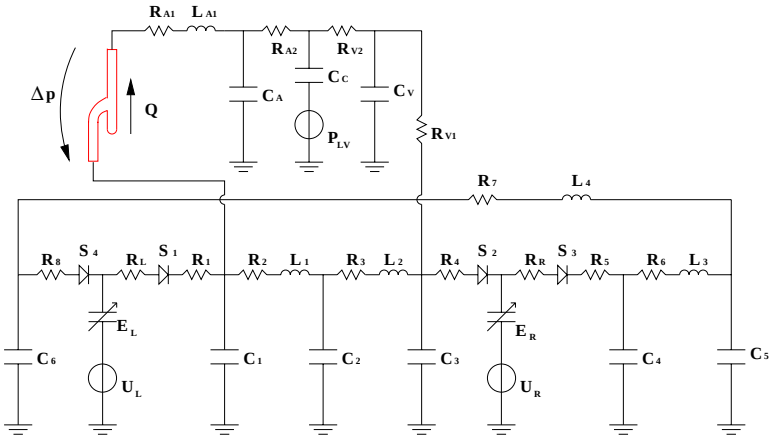


FIG. 12. *Lumped parameter model of the circulatory system coupled with a* 2*D description of a coronary by-pass.*

The interface condition we wish to impose between the two models are the continuity of flow rate and mean pressure. The numerical scheme employs a staggered algorithm where the pressure drop between inflow and outflow calculated by the Navier–Stokes model at a generic time step $t = t^k$ is imposed to the lumped parameter model which, in turn, is used to obtain the flow rate to advance the Navier–Stokes solution to $t = t^{k+1}$. We are therefore facing the case where we need to employ the technique presented in section 2. Since we are using a rigid wall model for the by-pass, the incoming and outgoing flow rates are equal, due to the incompressibility constraint. Actually, we have prescribed to the Navier–Stokes equations only the flow rate on the inflow section while homogeneous Neumann boundary conditions have been imposed on the outflow section. On the other hand, the pressure drop between inflow and outflow, needed to advance the lumped parameter model, is simply provided by the Lagrange multiplier.

An alternative coupling algorithm, based on imposing a mean pressure to the inflow and outflow sections of the Navier–Stokes problem while prescribing the flow rate to the lumped model, has been described in [15].

At the top of Figure 13 we show the flow rate and the pressure drop in the by-pass computed by the coupled system. The marks indicate the values at the times corresponding to the four snapshots of the fluid speed found in the lower part of the same figure.

5. Conclusions. In this work we have considered defective boundary conditions for Navier–Stokes equations. In particular, we have addressed the case where one wants to impose the flow rate on a measurable subset of the domain boundary. We have proposed a formulation based on a Lagrange multiplier technique and we have shown that it is well-posed for the Stokes and the linearized Navier–Stokes equations. Moreover, we have considered some numerical algorithms to effectively solve the mixed problem thus obtained. Finally, we have presented several applications in which the technique may be advantageously used and we have shown some numerical results illustrating its effectiveness.

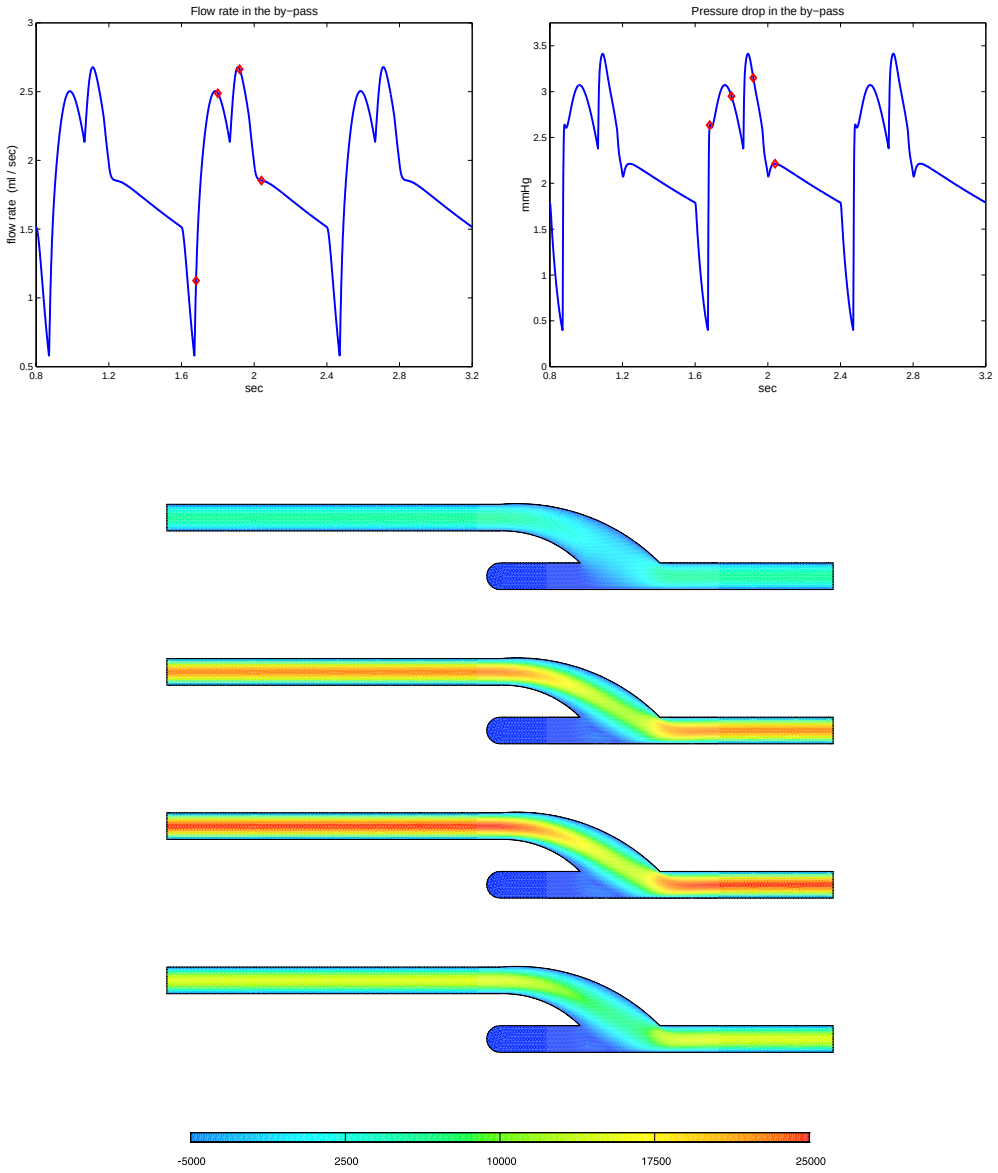


FIG. 13. *On the top: flow rate and pressure drop in the by-pass. On the bottom: fluid speed at* $t = 1.68\,s, 1.8\,s, 1.92\,s, 2.04\,s$.

REFERENCES

[1] C. BÈGUE, C. CONCA, F. MURAT, AND O. PIRONNEAU, *Les équations de Stokes et de Navier-Stokes avec des conditions aux limites sur la pression*, in Nonlinear partial differential equations and their applications, Collège de France Seminar, Vol. IX (Paris, 1985–1986), Longman Scientific and Technical, Harlow, UK, 1988, pp. 179–264.

[2] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Elements*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[3] C.H. BRUNEAU, *Boundary conditions on artificial frontiers for incompressible and compressible Navier-Stokes equations,* M2AN Math. Model. Numer. Anal., 34 (2000), pp. 303–314.

[4] C.H. Bruneau and P. Fabre, *Effective downstream boundary conditions for incompresisble Navier-Stokes equations,* Internat. J. Numer. Methods Fluids, 19 (1994), pp. 693–705.
[5] C. Conca, C. Pares, O. Pironneau, and M. Thieret, *A computational model of Navier-Stokes equations with imposed pressure and velocity fluxes,* Internat. J. Numer. Methods Fluids, 20 (1995), pp. 267–287.
[6] L. Formaggia, J.-F. Gerbeau, F. Nobile, and A. Quarteroni, *On the coupling of* 3*D and* 1*D Navier-Stokes equations for flow problems in compliant vessels,* Comput. Methods Appl. Mech. Engrg., 191 (2001), pp. 561–582.
[7] L. Formaggia and F. Nobile, *A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements,* East-West J. Numer. Math., 7 (1999), pp. 105–132.
[8] L. Formaggia, F. Nobile, A. Quarteroni, and A. Veneziani, *Multiscale Modelling of the Circulatory System: A preliminary analysis,* Comput. Vis. Science, 2 (1999), pp. 75–83.
[9] V. Girault and P.-A. Raviart, *Navier–Stokes Equations: Theory and Numerical Analysis*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.
[10] J.G. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations,* Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.
[11] F. Inzoli, F. Migliavacca, and S. Mantero, *Pulsatile flow in an aorto-coronary bypass* 3*-D model*, in Biofluid Mechanics, Proceedings of the Third International Symposium, Munich, Germany, D. Liepsch, ed., VDI Verlag, Dusseldorf, 1994.
[12] D.A. McDonald, *Blood Flow in Arteries*, 3rd ed., W.W. Nichols and M.F. O'Rourke, eds., Edward Arnold, London, 1990.
[13] F.C. Otto and G. Lube, *A non-overlapping domain decomposition method for the Oseen equations,* Math. Models Methods Appl. Sci., 8 (1998), pp. 1091–1117.
[14] B. Perot, *An analysis of the fractional step method,* J. Comput. Phys., 108 (1993), pp. 51–58.
[15] A. Quarteroni, S. Ragni, and A. Veneziani, *Coupling between lumped and distributed models for blood flow problems,* Comput. Vis. Science, 4 (2001), pp. 111–124.
[16] A. Quarteroni, F. Saleri, and A. Veneziani, *The Yosida method for the numerical approximation of Navier-Stokes equations,* J. Math. Pures Appl. (9), 78 (1999), pp. 473–503.
[17] A. Quarteroni, F. Saleri, and A. Veneziani, *Factorization methods for the numerical approximation of Navier-Stokes equations,* Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 505–526.
[18] A. Quarteroni, M. Tuveri, and A. Veneziani, *Computational vascular fluid dynamics: Problems, models and methods,* Comput. Vis. Science, 2 (2000), pp. 163–197.
[19] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
[20] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, 3rd ed., North-Holland, Amsterdam, 1984.
[21] N. Westerhof, F. Bosman, C.D. Vries, and A. Noordergraaf, *Analog studies of the human systemic arterial tree,* J. Biomech., 2 (1969), pp. 121–143.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 403–415

FLUX RECOVERY FROM PRIMAL HYBRID FINITE ELEMENT METHODS*

SO-HSIANG CHOU†, DO Y. KWAK‡, AND KWANG Y. KIM‡

Abstract. A flux recovery technique is introduced and analyzed for the computed solution of the primal hybrid finite element method for second-order elliptic problems. The recovery is carried out over a single element at a time while ensuring the continuity of the flux across the interelement edges and the validity of the discrete conservation law at the element level. Our construction is general enough to cover all degrees of polynomials and grids of triangular or quadrilateral type. We illustrate the principle using the Raviart–Thomas spaces, but other well-known related function spaces such as the Brezzi–Douglas–Marini (BDM) or Brezzi–Douglas–Fortin–Marini (BDFM) space can be used as well. An extension of the technique to the nonlinear case is given. Numerical results are presented to confirm the theoretical results.

Key words. recovery technique, primal hybrid method, nonconforming method, conservative method

AMS subject classifications. 65N15, 65N30

PII. S0036142900381266

1. Introduction.

We consider the second-order elliptic boundary value problem

$$\begin{cases} -\operatorname{div}(\mathcal{K}\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded polygonal domain in $\mathbb{R}^2$ with the boundary $\partial\Omega$, and $\mathcal{K} = \mathcal{K}(\mathbf{x})$ is assumed to be symmetric and uniformly positive definite, i.e., there exist two positive constants c_1 and c_2 such that

$$c_1\xi^T\xi \le \xi^T\mathcal{K}(\mathbf{x})\xi \le c_2\xi^T\xi \qquad \forall \xi \in \mathbb{R}^2,\ \mathbf{x} \in \overline{\Omega}.$$

In many applications, it is more important to gain accurate approximation for the vector variable $\boldsymbol{\sigma} = -\mathcal{K}\nabla u$ (e.g., Darcy velocity) rather than the scalar variable u (e.g., pressure). A common way of achieving that goal is to use the mixed finite element methods, which have been a very active area of research since the late 1970s; see, for example, [4, 5, 6, 9, 20, 22]. All mixed methods have the further advantage of maintaining the discrete conservation law at the element level.

However, mixed methods lead to an indefinite symmetric algebraic system which may be hard to solve iteratively. An efficient way to solve for the mixed finite element method is to further introduce the Lagrange multipliers on the edges of the mesh to ensure the continuity of normal components of the velocity variable. This is sometimes called the mixed-hybrid method. In this fashion the velocity and the pressure finite element spaces have no continuity constraints at all, and thus both variables can be

*Received by the editors November 15, 2000; accepted for publication (in revised form) January 10, 2002; published electronically May 29, 2002.
http://www.siam.org/journals/sinum/40-2/38126.html

†Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH (chou@bgnet.bgsu.edu). The research of this author was supported by NSF grant DMS–0074259.

‡Department of Mathematics, Korea Advanced Institute of Science and Technology, Taejon, Korea 305-701 (dykwak@math.kaist.ac.kr, kky@mathx.kaist.ac.kr). This work was supported by grant 2000-2-10300-001-5 from the Basic Research Program of the Korea Science & Engineering Foundation.

eliminated to obtain a symmetric and positive definite matrix system which involves only the Lagrange multipliers. It can be shown that this matrix system is equivalent to some nonconforming finite element method for the original problem (1.1); see, for example, [1, 2, 7]. The nonconforming method for the pressure requires fewer degrees of freedom than the mixed finite element method, and moreover, its solution can be computed by a fast solver such as multigrid algorithms (cf. [3, 8, 10, 11]). By using this nonconforming solution, the vector approximation can be obtained through a simple formula, for example, as done in [17]. The rectangular case with piecewise constant diagonal tensor was considered in [8]. For the triangular case, see [1, 2, 7, 8, 17].

Our objective in this paper is to show that similar equivalence results can be derived through the primal hybrid finite element methods for the problem (1.1) which were analyzed in [21] as a general approach of constructing nonconforming finite element approximations. We first present a technique of recovering from the primal hybrid solution an optimal flux approximation $\boldsymbol{\sigma}_h$ based on the local Raviart–Thomas spaces. Although the construction is carried out in a local manner (over a single element at a time), it is ensured that $\boldsymbol{\sigma}_h$ is continuous across the interelement boundaries and that the discrete conservation law holds locally. Also, instead of the Raviart–Thomas spaces, other mixed finite element spaces such as the Brezzi–Douglas–Marini (BDM) spaces or the Brezzi–Douglas–Fortin–Marini (BDFM) spaces can be used as well.

The main advantage of our technique is that it is general enough to cover all degrees of polynomials and all types of grids, triangular or quadrilateral. In particular, our technique can be applied to any nonconforming finite element method which can be viewed as a primal hybrid finite element method.

As good examples of how the technique can be applied, we will derive simple formulas for $\boldsymbol{\sigma}_h$ in the lowest-order cases on triangular and quadrilateral grids, which lead to the $P1$ and the rotated $Q1$ nonconforming finite element methods, respectively (see [21] or section 4).

The rest of the paper is organized as follows. In the next section the primal finite element methods are introduced for the problem (1.1). In section 3, we present a technique of flux recovery from the primal hybrid finite element methods and establish optimal error estimates for the vector approximation thus obtained, and in section 4 a detailed description of how the technique can be applied for the lowest-order elements is given. In section 5, our results are extended to nonlinear problems. Finally, in section 6, some numerical results are presented to confirm the theoretical results.

2. Primal hybrid finite element methods. In this section we give a brief description of the primal hybrid finite element method for the problem (1.1). The reader can find much more detail on this subject in [21].

Let $\mathcal{T}_h$ be a partition of Ω into triangles or convex quadrilaterals which satisfies the usual regularity assumption

$$C_1 h_T^2 \le |T| \le C_2 h_T^2 \qquad \forall T \in \mathcal{T}_h,$$

where h_T denotes the diameter of T, $|T|$ is the area of T, and $h = \max_{T\in\mathcal{T}_h} h_T$.

Denote by $\hat{T}$ a standard reference element, i.e., the unit square or the unit triangle with the vertices $\hat{\mathbf{x}}_i$'s. Then there exists a unique bijective bilinear or linear transformation $F_T : \hat{T} \to T$ such that $\mathbf{x}_i = F_T(\hat{\mathbf{x}}_i)$ for all i. We set

$$\mathcal{J}_T = \text{ Jacobian matrix of } F_T, \qquad J_T = \det \mathcal{J}_T.$$

Based on the triangulation $\mathcal{T}_h$ we define the spaces

$$X = \{v \in L^2(\Omega) : v|_T \in H^1(T) \quad \forall T \in \mathcal{T}_h\} = \prod_{T\in\mathcal{T}_h} H^1(T),$$

$$M = \Big\{\mu \in \prod_{T\in\mathcal{T}_h} H^{-1/2}(\partial T) : \text{ there exists } \boldsymbol{\tau} \in \mathbf{H}(\mathrm{div},\Omega) \text{ such that } \boldsymbol{\tau}\cdot\mathbf{n}_T = \mu \text{ on } \partial T, \qquad \forall T \in \mathcal{T}_h\Big\},$$

where $\mathbf{n}_T$ is the unit outward normal along ∂T. Let $|\cdot|_{m,\Omega}$ and $\|\cdot\|_{m,\Omega}$ denote the usual seminorm and norm, respectively, on the Sobolev space $H^m(\Omega)$. We define the mesh-dependent norms

$$|||v|||_X = \left(\sum_{T\in\mathcal{T}_h} |||v|||_{1,T}^2\right)^{1/2}, \qquad v \in X,$$

$$\|\mu\|_h = \left(\sum_{T\in\mathcal{T}_h} h_T\|\mu\|_{0,\partial T}^2\right)^{1/2}, \qquad \mu \in \prod_{T\in\mathcal{T}_h} L^2(\partial T),$$

where

$$|||v|||_{1,T}^2 = |v|_{1,T}^2 + h_T^{-2}\|v\|_{0,T}^2.$$

Now the primal hybrid formulation for the problem (1.1) is given as follows: find a pair $(u,\lambda) \in X \times M$ such that

$$a(u,v) + b(v,\lambda) = (f,v) \qquad \forall v \in X, \tag{2.1a}$$

$$b(u,\mu) = 0 \qquad \forall \mu \in M, \tag{2.1b}$$

where

$$a(u,v) = \sum_{T\in\mathcal{T}_h} \int_T \mathcal{K}\nabla u\cdot\nabla v\,dx, \qquad b(v,\mu) = \sum_{T\in\mathcal{T}_h} \int_{\partial T} v\mu\,ds, \tag{2.2}$$

$$(f,v) = \int_\Omega fv\,dx. \tag{2.3}$$

It was shown in [21] that u belongs to $H_0^1(\Omega)$ and is the unique solution of the standard weak formulation

$$\int_\Omega \mathcal{K}\nabla u\cdot\nabla v\,dx = \int_\Omega fv\,dx, \qquad v \in H_0^1(\Omega),$$

and that

$$\lambda = -\mathcal{K}\nabla u\cdot\mathbf{n}_T \ \text{ on } \partial T \qquad \forall T \in \mathcal{T}_h. \tag{2.4}$$

We use the standard notation for the spaces of polynomials, i.e., $P_r(T)$ denotes the space of polynomials on T of total degrees at most r, and $Q_{r,s}(T)$ denotes the space of polynomials on T of degrees at most r and s in x and y, respectively. We also set $Q_r(T) = Q_{r,r}(T)$. On any element T we define

$$R_r(T) = \begin{cases} P_r(T) & \text{if } T \text{ is a triangle,} \\ Q_r(\hat{T})\circ F_T^{-1} & \text{if } T \text{ is a quadrilateral} \end{cases}$$

and

$$S_r(\partial T) = \{\mu \in L^2(\partial T) : \mu|_e \in P_r(e) \qquad \forall e \text{ edges of } T\}.$$

In order to discretize the primal hybrid formulation (2.1), we introduce a finite-dimensional subspace $\hat{X}$ of $H^1(\hat{T})$ such that $R_k(\hat{T}) \subset \hat{X}$ for some $k \geq 1$. Then we define a pair $X_h \times M_h$ of finite element spaces on $\mathcal{T}_h$ by

$$X_h = \{v \in X : v|_T \in X_h(T) \qquad \forall T \in \mathcal{T}_h\}, \tag{2.5}$$

$$M_h = \left\{ \mu \in \prod_{T \in \mathcal{T}_h} S_{k-1}(\partial T) : \mu|_{\partial T_1} + \mu|_{\partial T_2} = 0 \text{ on } \partial T_1 \cap \partial T_2 \right. \\ \left. \text{if } T_1 \text{ and } T_2 \text{ are adjacent elements} \right\}, \tag{2.6}$$

where we set $X_h(T) = \{\hat{v} \circ F_T^{-1} : \hat{v} \in \hat{X}\}$.

Now the primal hybrid finite element method is defined as follows: find a pair $(u_h, \lambda_h) \in X_h \times M_h$ such that

$$a(u_h, v) + b(v, \lambda_h) = (f, v) \qquad \forall v \in X_h, \tag{2.7a}$$

$$b(u_h, \mu) = 0 \qquad \forall \mu \in M_h. \tag{2.7b}$$

Examples for the space $\hat{X}$ are given in [21] for all $k \geq 1$ which ensures the existence and uniqueness of a solution (u_h, λ_h) for the system (2.7) and satisfy the following optimal error estimates (cf. [15, 21]).

THEOREM 2.1. *For $u \in H^{k+1}(\Omega)$ we have*

$$|\!|\!| u - u_h |\!|\!|_X + \|\lambda - \lambda_h\|_h \leq C h^k |u|_{k+1}.$$

The following observation is crucial to decouple the mixed system (2.7): u_h is the solution of the *nonconforming* finite element method

$$a(u_h, v) = (f, v), \qquad v \in V_h, \tag{2.8}$$

where

$$V_h = \{v \in X_h : b(v, \mu) = 0 \qquad \forall \mu \in M_h\}. \tag{2.9}$$

This implies that we may compute u_h directly from (2.8) and then compute λ_h from u_h locally by (2.7a), which reduces to

$$\int_{\partial T} v \lambda_h \, ds = \int_T f v \, dx - \int_T \mathcal{K} \nabla u_h \cdot \nabla v \, dx, \qquad v \in X_h(T). \tag{2.10}$$

Thus, the Lagrange multiplier λ_h may be interpreted as the (weak) *local residuals* of the nonconforming approximation u_h.

3. Flux recovery technique. To begin with, we define the Raviart–Thomas space of index $r \geq 0$ on $\mathcal{T}_h$ as follows:

$$RT_r = \{\boldsymbol{\tau} \in \mathbf{H}(\mathrm{div}, \Omega) : \boldsymbol{\tau}|_T \in RT_r(T)\},$$

where the local space $RT_r(T)$ is defined as

$$RT_r(T) = \{\mathcal{P}_T\hat{\boldsymbol{\tau}} : \hat{\boldsymbol{\tau}} \in (R_r(\hat{T}))^2 + (x,y)R_r(\hat{T})\},$$

and $\mathcal{P}_T\hat{\boldsymbol{\tau}} = J_T^{-1}\mathcal{J}_T\hat{\boldsymbol{\tau}} \circ F_T^{-1}$. We also set for $r \geq 1$

$$\boldsymbol{\Psi}_r(T) = \begin{cases} (P_{r-1}(T))^2 & \text{if } T \text{ is a triangle,} \\ \{\mathcal{J}_T^{-t}\hat{\boldsymbol{\tau}} \circ F_T^{-1} : \hat{\boldsymbol{\tau}} \in Q_{r-1,r}(\hat{T}) \times Q_{r,r-1}(\hat{T})\} & \text{if } T \text{ is a quadrilateral.} \end{cases}$$

Let us point out that if T is a rectangle, then

$$\boldsymbol{\Psi}_r(T) = Q_{r-1,r}(T) \times Q_{r,r-1}(T).$$

Now we present a technique of recovering an optimal vector approximation. Once the solution (u_h, λ_h) of the system (2.7) is computed, one can construct a unique $\boldsymbol{\sigma}_h \in RT_{k-1}(T)$ on each $T \in \mathcal{T}_h$:

$$\boldsymbol{\sigma}_h \cdot \mathbf{n}_T = \lambda_h \quad \text{on } \partial T, \tag{3.1a}$$

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K}\nabla u_h) \cdot \boldsymbol{\tau}\, dx = 0, \quad \boldsymbol{\tau} \in \boldsymbol{\Psi}_{k-1}(T) \quad (k \geq 2) \tag{3.1b}$$

(cf. [6, 20, 22]). By construction we immediately obtain the following two propositions.

PROPOSITION 3.1. *The normal components of $\boldsymbol{\sigma}_h$ are continuous across the interelement boundaries, i.e., we have $\boldsymbol{\sigma}_h \in RT_{k-1}$.*

Proof. This is a direct consequence of (3.1a). ∎

PROPOSITION 3.2. *We have for all $v \in R_{k-1}(T)$*

$$\int_T \operatorname{div} \boldsymbol{\sigma}_h\, v\, dx = \int_T f v\, dx.$$

This implies that the discrete conservation law holds locally.

Proof. By using (3.1a) and Green's theorem, (2.7a) becomes

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K}\nabla u_h) \cdot \nabla v\, dx + \int_T \operatorname{div} \boldsymbol{\sigma}_h\, v\, dx = \int_T f v\, dx \qquad \forall v \in X_h(T). \tag{3.2}$$

There is nothing to be done for $k = 1$, since $\nabla v = 0$ for $v \in R_0(T)$. For $k \geq 2$ we have $\nabla v \in \boldsymbol{\Psi}_{k-1}(T)$ for $v \in R_{k-1}(T)$, which proves the desired result by (3.1b). ∎

Remark 3.1. We could use other mixed finite elements instead of RT_{k-1}. For example, when one wants to use the $BDM_{k-1}(T)$ space on a triangle T ($k \geq 2$), (3.1) is replaced by

$$\boldsymbol{\sigma}_h \cdot \mathbf{n}_T = \lambda_h \quad \text{on } \partial T, \tag{3.3a}$$

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K}\nabla u_h) \cdot \nabla v\, dx = 0, \quad v \in P_{k-2}(T), \tag{3.3b}$$

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K}\nabla u_h) \cdot \operatorname{curl}(b_T v)\, dx = 0, \quad v \in P_{k-3}(T) \quad (k \geq 3), \tag{3.3c}$$

where b_T is the cubic bubble function on T. If the space $\hat{X}$ contains only $P_k(\hat{T})$ on a quadrilateral T, one may use the $BDFM_k(T)$ space, in which case (3.1) is replaced by

$$\boldsymbol{\sigma}_h \cdot \mathbf{n}_T = \lambda_h \quad \text{on } \partial T, \tag{3.4a}$$

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K}\nabla u_h) \cdot \boldsymbol{\tau}\, dx = 0, \quad \boldsymbol{\tau} \in \boldsymbol{\Psi}_{k-1}(T) \quad (k \geq 2), \tag{3.4b}$$

where we set

$$\boldsymbol{\Psi}_r(T) = \{\mathcal{J}_T^{-t}\hat{\boldsymbol{\tau}} \circ F_T^{-1} : \hat{\boldsymbol{\tau}} \in (P_{r-1}(\hat{T}))^2\}.$$

For a review of the degrees of freedom (3.3) and (3.4), we refer to [4, 5, 6].

Before going to an error estimate, we prove the following key lemma.

LEMMA 3.3. *Given $\beta \in L^2(\partial T)$ and $\mathbf{q} \in (L^2(T))^2$, let $\boldsymbol{\xi}_h \in RT_r(T)$ satisfy*

$$\int_{\partial T} \boldsymbol{\xi}_h \cdot \mathbf{n}_T\, \mu\, ds = \int_{\partial T} \beta\mu\, ds \qquad \forall \mu \in S_r(\partial T),$$
$$\int_T \boldsymbol{\xi}_h \cdot \boldsymbol{\tau}\, dx = \int_T \mathbf{q} \cdot \boldsymbol{\tau}\, dx \qquad \forall \boldsymbol{\tau} \in \boldsymbol{\Psi}_r(T) \quad (r \geq 1).$$

Then we obtain

$$\|\boldsymbol{\xi}_h\|_{0,T} \leq C(\|\mathbf{q}\|_{0,T} + h_T^{1/2}\|\beta\|_{0,\partial T}).$$

Proof. By considering the L^2 projections, we may assume that $\beta \in S_r(\partial T)$ and $\mathbf{q} \in RT_r(T)$. Then the proof is done by using a simple scaling argument [2, 6]. ∎

Now we derive an error estimate for the vector approximation $\boldsymbol{\sigma}_h$ constructed by (3.1). It is well known (see, e.g., [6, 20, 22, 23]) that the Raviart–Thomas projection $\Pi_h : (H^1(\Omega))^2 \to RT_r$ can be defined by

$$\int_{\partial T} \Pi_h\boldsymbol{\sigma} \cdot \mathbf{n}\, \mu\, ds = \int_{\partial T} \boldsymbol{\sigma} \cdot \mathbf{n}\, \mu\, ds, \qquad \mu \in S_r(\partial T), \tag{3.5}$$

$$\int_T \Pi_h\boldsymbol{\sigma} \cdot \boldsymbol{\tau}\, dx = \int_T \boldsymbol{\sigma} \cdot \boldsymbol{\tau}\, dx, \qquad \boldsymbol{\tau} \in \boldsymbol{\Psi}_r(T), \tag{3.6}$$

possessing the following approximation properties:

$$\|\boldsymbol{\sigma} - \Pi_h\boldsymbol{\sigma}\|_0 \leq Ch^l\, \|\boldsymbol{\sigma}\|_l, \qquad 1 \leq l \leq r+1,$$

for all $\boldsymbol{\sigma} \in (H^l(\Omega))^2$, and

$$\|\operatorname{div}(\boldsymbol{\sigma} - \Pi_h\boldsymbol{\sigma})\|_0 \leq Ch^l\, \|\operatorname{div}\boldsymbol{\sigma}\|_l, \qquad 0 \leq l \leq r+1,$$

for all $\boldsymbol{\sigma} \in (H^l(\Omega))^2$ with $\operatorname{div}\boldsymbol{\sigma} \in H^l(\Omega)$.

THEOREM 3.4. *Let $\boldsymbol{\sigma} = -\mathcal{K}\nabla u$. Then we have for $u \in H^{k+1}(\Omega)$*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_0 \leq Ch^k(|\boldsymbol{\sigma}|_k + |u|_{k+1}).$$

Proof. It suffices to prove that

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,T} \leq Ch^k(|\boldsymbol{\sigma}|_k + |u|_{k+1}).$$

From (3.5) and (3.6) it follows that

$$\int_{\partial T}(\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\mathbf{n}_T\,\mu\,ds=\int_{\partial T}(\lambda-\lambda_h)\mu\,ds,\quad \mu\in S_{k-1}(\partial T),$$
$$\int_T(\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\cdot\boldsymbol{\tau}\,dx=-\int_T\mathcal{K}\nabla(u-u_h)\cdot\boldsymbol{\tau}\,dx,\quad \boldsymbol{\tau}\in\boldsymbol{\Psi}_{k-1}(T)\quad (k\geq 2).$$

By applying Lemma 3.3 and then Theorem 2.1, we obtain

$$\begin{aligned}\|\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,T}&\leq C(|u-u_h|_{1,T}+h_T^{1/2}\|\lambda-\lambda_h\|_{0,\partial T})\\&\leq Ch^k(|\boldsymbol{\sigma}|_k+|u|_{k+1}).\end{aligned}$$

This completes the proof. ∎

4. Examples. In this section $\bar{f}$ indicates the piecewise constant average of f on $\mathcal{T}_h$, i.e.,

$$\bar{f}|_T=\frac{1}{|T|}\int_T f\,dx.$$

4.1. *P*1 nonconforming method. Let $\mathcal{T}_h$ be composed of triangles. We consider the lowest-order element, i.e., $k=1$:

$$X_h(T)=P_1(T),\qquad M_h=\prod_{T\in\mathcal{T}_h}S_0(\partial T)\bigcap M.$$

Then it is easy to see that V_h (defined by (2.9)) is the $P1$ nonconforming finite element space.

Let T be an arbitrary element of $\mathcal{T}_h$ with the edges e_1,e_2,e_3 and the barycenter $\mathbf{x}_T$, and let $\phi_i\in X_h(T)$ be the basis function associated with the edge e_i, namely, $\frac{1}{|e_i|}\int_{e_i}\phi_j\,ds=\delta_{ij}$. Then $\lambda_h|_{e_i}$ is given by (see (2.10))

$$\lambda_h|_{e_i}=\frac{1}{|e_i|}\left(\int_T f\phi_i\,dx-\int_T\mathcal{K}\nabla u_h\cdot\nabla\phi_i\,dx\right).$$

Using the formula $\nabla\phi_i=\frac{\mathbf{n}_i|e_i|}{|T|}$ results in

$$\lambda_h|_{e_i}=-\bar{\mathcal{K}}\nabla u_h\cdot\mathbf{n}_i+\frac{1}{|e_i|}\int_T f\phi_i\,dx.$$

By comparing the normal components $\boldsymbol{\sigma}_h\cdot\mathbf{n}_i$ for each i, one can show that the vector $\boldsymbol{\sigma}_h$ constructed by (3.1) is identical to the one given in [13]:

$$\boldsymbol{\sigma}_h=-\bar{\mathcal{K}}\nabla u_h+\frac{\bar{f}}{2}(\mathbf{x}-\mathbf{x}_T)+\mathbf{C}_T,\tag{4.1}$$

where $\mathbf{C}_T$ is determined by any two of the three equations

$$|e_i|\mathbf{n}_i\cdot\mathbf{C}_T=\int_T f\phi_i\,dx-\frac{|T|}{3}\bar{f},\qquad i=1,2,3.$$

In particular, when f is a constant on T, we obtain $\mathbf{C}_T=0$ and

$$\boldsymbol{\sigma}_h|_T=-\bar{\mathcal{K}}\nabla u_h|_T+\frac{f}{2}(\mathbf{x}-\mathbf{x}_T),\tag{4.2}$$

which is the formula obtained by Marini [17].

Remark 4.1. It is straightforward to extend the above results to higher-order elements of odd degrees $k \geq 3$:

$$X_h(T) = P_k(T), \qquad M_h = \prod_{T \in \mathcal{T}_h} S_{k-1}(\partial T) \bigcap M.$$

Once u_h is computed, λ_h can be computed at k Gauss–Legendre points on each edge by using the basis functions associated with these points.

Now let us consider the primal hybrid finite element method with the right-hand side f replaced by $\bar{f}$. By Proposition 3.2 we then have

$$\int_T \operatorname{div} \boldsymbol{\sigma}_h \, dx = \int_T \bar{f} \, dx, \tag{4.3a}$$

or $\operatorname{div} \boldsymbol{\sigma}_h = \bar{f}$. This, together with (3.2), implies that

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K} \nabla u_h) \, dx = 0. \tag{4.3b}$$

The equations (4.3a)–(4.3b) form the finite volume box method introduced by Courbet and Croisille [14]. Thus the primal hybrid finite element method along with our technique of flux recovery provide an alternative approach to the finite volume box method. A different approach is given in [13]; see also [12].

4.2. Rotated $Q1$ nonconforming method. Let $\mathcal{T}_h$ be composed of quadrilaterals, and

$$\hat{X} = \operatorname{span}\{1, \hat{x}, \hat{y}, \hat{x}^2 - \hat{y}^2\}, \qquad M_h = \prod_{T \in \mathcal{T}_h} S_0(\partial T) \bigcap M.$$

Then it is easy to see that V_h is the *parametric* rotated $Q1$ nonconforming finite element space introduced by Rannacher and Turek [19]. One could use the nonparametric version as well which, on rectangular grids, is given by

$$X_h(T) = \operatorname{span}\{1, x, y, x^2 - y^2\}, \qquad M_h = \prod_{T \in \mathcal{T}_h} S_0(\partial T) \bigcap M.$$

As in the $P1$ nonconforming method, $\lambda_h|_{e_i}$ is given by

$$\lambda_h|_{e_i} = \frac{1}{|e_i|} \left(\int_T f \phi_i \, dx - \int_T \nabla u_h \cdot \nabla \phi_i \, dx \right),$$

where $\phi_i \in X_h(T)$ is the basis function associated with the edge e_i.

Now we derive a simple formula for $\boldsymbol{\sigma}_h$ for the nonparametric version on rectangular grids. Suppose that f is piecewise constant. Then we obtain $\operatorname{div} \boldsymbol{\sigma}_h = f$ and

$$\int_T (\boldsymbol{\sigma}_h + \mathcal{K} \nabla u_h) \cdot \nabla v \, dx = 0 \qquad \forall v \in X_h(T). \tag{4.4}$$

Let us decompose $\boldsymbol{\sigma}_h|_T$ into

$$\boldsymbol{\sigma}_h|_T = \boldsymbol{\sigma}_{h,0}|_T + a_T(h_{T_y}^2 (x - x_T), h_{T_x}^2 (y - y_T)),$$

where $\boldsymbol{\sigma}_{h,0}|_T$ belongs to

$$\nabla X_h(T) = \{(a + bx, c - by) : a, b, c \in \mathbb{R}\},$$

and h_{T_x} and h_{T_y} are the width and the height of T, respectively. Since we have $\operatorname{div} \boldsymbol{\sigma}_{h,0}|_T = 0$, it follows that

$$a_T = \frac{f|_T}{h_{T_x}^2 + h_{T_y}^2}.$$

Now, by means of the orthogonality relation

$$\int_T (h_{T_y}^2(x - x_T), h_{T_x}^2(y - y_T)) \cdot \nabla v \, dx = 0, \qquad v \in X_h(T),$$

it is easy to see from (4.4) that

$$\boldsymbol{\sigma}_{h,0} = -P_0(\mathcal{K}\nabla u_h), \tag{4.5}$$

where P_0 is the L^2 projection which locally maps onto the space $\nabla X_h(T)$. Combining the results obtained thus far, we obtain

$$\boldsymbol{\sigma}_h|_T = -P_0(\mathcal{K}\nabla u_h)|_T + \frac{f|_T}{h_{T_x}^2 + h_{T_y}^2}(h_{T_y}^2(x - x_T), h_{T_x}^2(y - y_T)). \tag{4.6}$$

Remark 4.2. Similar results using the lowest-order rectangular Raviart–Thomas mixed finite element method can be found in [1, 8]. Our results show the primal hybrid approach provides a clear way of constructing the vector approximation from the Q_1 nonconforming solution.

5. Extension to nonlinear problems. The previous results can be extended to the nonlinear second-order elliptic boundary value problem

$$\begin{cases} -\operatorname{div} \mathbf{a}(u, \nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.1}$$

The primal hybrid formulation is to find a pair $(u, \lambda) \in X \times M$ such that

$$a(u, v) + b(v, \lambda) = (f, v) \qquad \forall v \in X, \tag{5.2a}$$

$$b(u, v) = 0 \qquad \forall \mu \in M, \tag{5.2b}$$

where

$$a(u, v) = \sum_{T \in \mathcal{T}_h} \int_T \mathbf{a}(u, \nabla u) \cdot \nabla v \, dx. \tag{5.3}$$

Analysis of the primal hybrid finite element methods for this nonlinear problem is given in [18] for $k \geq 2$.

The previous technique of recovering a vector approximation can be applied as well. After computing the primal hybrid solution (u_h, λ_h), one constructs a unique $\boldsymbol{\sigma}_h \in RT_{k-1}(T)$ on each $T \in \mathcal{T}_h$ by

$$\boldsymbol{\sigma}_h \cdot \mathbf{n}_T = \lambda_h \quad \text{on } \partial T, \tag{5.4a}$$

$$\int_T [\boldsymbol{\sigma}_h + \mathbf{a}(u_h, \nabla u_h)] \cdot \boldsymbol{\tau} \, dx = 0, \quad \boldsymbol{\tau} \in \boldsymbol{\Psi}_{k-1}(T) \quad (k \geq 2). \tag{5.4b}$$

Setting $\boldsymbol{\sigma} = -\mathbf{a}(u, \nabla u)$, we can derive the bound for $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_0$ in the same way as before. By applying Lemma 3.3 to the error equations

$$\int_{\partial T} (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \cdot \mathbf{n}_T \, \mu \, ds = \int_{\partial T} (\lambda - \lambda_h) \mu \, ds, \quad \mu \in S_{k-1}(\partial T),$$
$$\int_T (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \cdot \boldsymbol{\tau} \, dx = -\int_T [\mathbf{a}(u, \nabla u) - \mathbf{a}(u_h, \nabla u_h)] \cdot \boldsymbol{\tau} \, dx,$$
$$\boldsymbol{\tau} \in \boldsymbol{\Psi}_{k-1}(T) \quad (k \geq 2),$$

we obtain

$$\|\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,T} \leq C(\|\mathbf{a}(u, \nabla u) - \mathbf{a}(u_h, \nabla u_h)\|_{0,T} + h_T^{1/2} \|\lambda - \lambda_h\|_{0,\partial T}).$$

Note that if $\mathbf{a}$ has bounded derivatives, then there exists a constant $C > 0$ independent of h such that

$$|\mathbf{a}(u, \nabla u) - \mathbf{a}(u_h, \nabla u_h)| \leq C(|u - u_h| + |\nabla(u - u_h)|),$$

which implies that

$$\|\mathbf{a}(u, \nabla u) - \mathbf{a}(u_h, \nabla u_h)\|_{0,T} \leq C \|u - u_h\|_{1,T}.$$

Thus it follows by Theorem 2.1 that

$$\|\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,T} \leq C h^k (|\boldsymbol{\sigma}|_k + |u|_{k+1}).$$

THEOREM 5.1. *We have for* $u \in H^{k+1}(\Omega)$

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_0 \leq C h^k (|\boldsymbol{\sigma}|_k + |u|_{k+1}).$$

6. Numerical results. To confirm the theoretical results established in the previous sections, numerical experiments are carried out on the unit square $\Omega = (0, 1)^2$ for three test problems. The first problem has a discontinuous tensor coefficient, and the second one has a smooth coefficient, but its solution has a very weak "layer" near the right boundary. Finally the third problem is taken from [16]. For numerical results on triangular grids, we refer to [13].

Errors for the velocity and the pressure approximations are computed in the discrete L^2 norms

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,h}^2 = \sum_{T \in \mathcal{T}_h} \sum_{e \in \partial T} \left[\int_e (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \cdot \mathbf{n} \, ds \right]^2, \tag{6.1}$$

$$\|u - u_h\|_{0,h}^2 = \sum_{T \in \mathcal{T}_h} \int_T (u - u_h)^2 \, dxdy, \tag{6.2}$$

where the integrals are evaluated by the midpoint rule, i.e., if S denotes an edge e or an area T, then we evaluate $\int_S g$ by $|S| \times g(\mathbf{x}_S)$, where $\mathbf{x}_S$ is the mass center of S. All the results below show second-order convergence in the velocity. They are tabulated as Tables 6.1–6.3.

PROBLEM 1.

$$\mathcal{K} = \begin{pmatrix} 10^4 & 0 \\ 0 & 1 \end{pmatrix} \text{ for } 0 < x < .5, \qquad \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \text{ for } .5 < x < 1,$$

TABLE 6.1
Problem 1. *Discontinuous tensor coefficients.*

h	$\|\mathfrak{E} - \mathfrak{E}_h\|_{0,h}$	$\|u - u_h\|_{0,h}$
1/8	2.58135e-1	5.60052e-4
1/16	2.64380e-1	1.39916e-4
1/32	7.91115e-2	3.49786e-5
1/64	1.97629e-2	8.74365e-6
1/128	4.93600e-3	2.18584e-6

TABLE 6.2
Problem 2. *Weak layer at the right boundary.*

h	$\|\mathfrak{E} - \mathfrak{E}_h\|_{0,h}$	$\|u - u_h\|_{0,h}$
1/8	9.30579e-2	8.34233e-2
1/16	2.67894e-2	2.21723e-2
1/32	7.01984e-3	5.63162e-3
1/64	1.77762e-3	1.41354e-3
1/128	4.45865e-4	3.53740e-4

TABLE 6.3
Problem 3. *Distorted grids,* $\beta = 60°, \theta = 45°$.

Grid size	$\|\mathfrak{E} - \mathfrak{E}_h\|_{0,h}$	$\|u - u_h\|_{0,h}$
8×8	2.0878e-1	4.1977e-2
16×16	5.2684e-2	1.0989e-2
32×32	1.3526e-2	2.7816e-3
64×64	3.4843e-3	6.9757e-4
128×128	8.9701e-4	1.7453e-4

and

$$u(x, y) = x(1 - x)y(1 - y).$$

The domain Ω *is partition into the squares of size* h. *By simple calculations it is easy to see that the velocity* $\boldsymbol{\sigma} = -\mathcal{K}\nabla u$ *has continuous normal components across the line of discontinuity* $x = 1/2$. *We use the parametric rotated* $Q1$ *nonconforming method for this problem.*

PROBLEM 2. *In this problem we let* $\mathcal{K} = I$, *the identity matrix. The exact solution is*

$$u(x, y) = x(1 - x)y(1 - y)\exp(5x),$$

which has a boundary layer. We use rectangular grids.

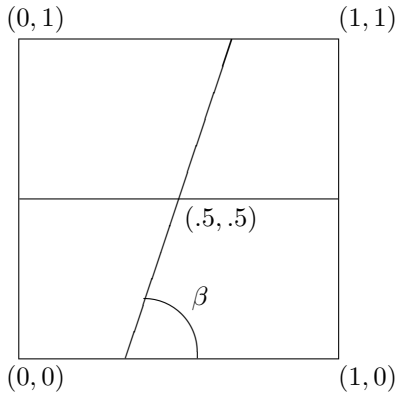


FIG. 6.1. *Distorted grids for Problem* 3.

PROBLEM 3.

$$\mathcal{K} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0.01 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

and $u(x, y) = \cos(\pi x)\cos(2\pi y)$. *The grids are obtained through successive refinements of the initial grid shown in Figure* 6.1. *The refinement is done by connecting the midpoints of opposite edges of every quadrilateral. We use the parametric rotated* $Q1$ *nonconforming method for this problem.*

REFERENCES

[1] T. ARBOGAST AND Z. CHEN, *On the implementation of mixed methods as nonconforming methods for second order elliptic problems*, Math. Comp., 66 (1997), pp. 85–104.

[2] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[3] S. C. BRENNER, *An optimal order multigrid for* $P1$ *nonconforming finite elements*, Math. Comp., 52 (1989), pp. 1–15.

[4] F. BREZZI, J. DOUGLAS, M. FORTIN, AND L. MARINI, *Efficient rectangular mixed finite elements in two and three variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[5] F. BREZZI, J. DOUGLAS, AND L. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New-York, 1991.

[7] Z. CHEN, *Analysis of mixed methods using conforming and nonconforming finite element methods*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 9–34.

[8] Z. CHEN, *Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems*, East-West J. Numer. Math., 4 (1996), pp. 1–33.

[9] Z. CHEN AND J. DOUGLAS, JR., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[10] Z. CHEN AND D. Y. KWAK, *Convergence of multigrid methods for nonconforming finite elements without regularity assumptions*, Comput. Appl. Math., 17 (1998), pp. 283–302.

[11] Z. CHEN AND P. OSWALD, *Multigrid and multilevel methods for nonconforming* $Q1$ *elements*, Math. Comp., 67 (1998), pp. 667–693.

[12] S. H. CHOU, D. Y. KWAK, AND K. Y. KIM, *Mixed finite volume methods on non-staggered quadrilateral grids for elliptic problems*, Math. Comp., to appear.

[13] S.-H. CHOU AND S. TANG, *Conservative* $P1$ *conforming and nonconforming Galerkin FEMs: Effective flux evaluation via a nonmixed method approach*, SIAM J. Numer. Anal., 38 (2000), pp. 660–680.

[14] B. Courbet and J. P. Croisille, *Finite volume box schemes on triangular meshes*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 631–649.

[15] J. Douglas, C. P. Gupta, and G. Y. Li, *Global estimates for a primal hybrid finite element method for second order elliptic problems in the plane*, Mat. Apl. Comput., 2 (1983), pp. 273–283.

[16] J. E. Jones, *A Mixed Finite Volume Element Method for Accurate Computation of Fluid Velocities in Porous Media*, Ph.D. thesis, University of Colorado at Denver, Denver, CO, 1995.

[17] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.

[18] E. J. Park, *A primal hybrid finite element method for a strongly nonlinear second-order elliptic problem*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 61–75.

[19] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.

[20] P. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in *Proceedings of the Conference on Mathematical Aspects of Finite Element Methods*, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977, pp. 292–315.

[21] P. A. Raviart and J. M. Thomas, *Primal hybrid finite element methods for 2nd order elliptic equations*, Math. Comp., 31 (1977), pp. 391–413.

[22] J. E. Roberts and J. M. Thomas, *Mixed and hybrid methods*, in Handb. Numer. Anal., Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[23] J. Wang and T. Mathew, *Mixed finite element methods over quadrilaterals*, in the Proceedings of the Third International Conference on Advances in Numerical Methods and Applications, I. T. Dimov, Bl. Sendov, and P. Vassilevski, eds., World Scientific, Singapore, 1994, pp. 203–214.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 416–430

ON THE NUMERICAL ANALYSIS OF THE IMPERFECT BIFURCATION OF CODIM ≤ 3*

KLAUS BÖHMER†, DÁŠA JANOVSKÁ‡, AND VLADIMÍR JANOVSKÝ§

Abstract. A postprocessing analysis of a previously computed bifurcation point of codim ≤ 3, corank $= 1$ is proposed. Some cases with codim > 3 also are included. Our aim is to predict quantitatively imperfect bifurcation phenomena. Our idea is to compute the differential of the diffeomorphism that links the state and parameter variables of the actual problem with its normal form.

Key words. bifurcation points, imperfect bifurcation diagrams, quantitative analysis

AMS subject classifications. 34A34, 35B32, 37M20, 65L99, P30, 58F14

PII. S0036142900369283

1. Introduction. This paper deals with unusual numerical techniques for the bifurcation analysis of steady states. The classical numerical methods for continuation of parameter dependent solutions are presented, e.g., in [25], [20], [26]; for their up-to-date versions, see, e.g., [12], [14], [13], [15], [27], [19], [31]. The latter in particular yield the following types of results. The bifurcation points (u^*, λ^*) in our notation below are detected along the solution curve, where solutions bifurcate or become singular. The numbers of bifurcating branches, their stability properties, and often their tangents, are determined (see, e.g., [31], [19], and here, in particular, section 7.8). The evaluation of the defining equations (see below) allows us to classify this type of singularity. Hence, the structure of the bifurcation scenario is known; see [17, Chap. III, Figs. 7.1, 7.2, 8.1 ff. and Chap. IV, Figs. 4.1 ff]. So, let (u^*, λ^*) for the original problem be classified, e.g., as a pitchfork or a winged cusp; see Figure 4.3 in [17, Chap. IV]. For the classified normal form, all details of the scenario are well known. However, it is not known where exactly in the neighborhood of the original (u^*, λ^*), e.g., the hysteresis effects really do occur. In the case of chemical processes, this is important information for the production process; see [28], [29].

To present this problem in more detail, we give a short outline. A unifying approach via a *generalized Liapunov–Schmidt reduction* was proposed and justified in [24]. Discretization methods for operator equations as elliptic PDEs, including Navier–Stokes operators, are applied here to the generalized Liapunov–Schmidt reductions. Their convergence was proved in [3], [10], [32], [2], [4], [5], [11]. Finally, the link between reduction techniques and *bordered matrices* was considered, e.g., in [18]. For some recent developments in this direction see [33], [1], [31], [19].

*Received by the editors March 6, 2000; accepted for publication (in revised form) November 15, 2001; published electronically May 29, 2002. This research was supported by the Deutsche Forschungsgemeinschaft Bo-622/13-3. The research of the second and third authors was partially supported by the Grant Agency of the Czech Republic (grants 201/98/0220 and 201/98/0528) and by grants CEZ J13/98 : 113200007 and J19/98 : 223400007.

http://www.siam.org/journals/sinum/40-2/36928.html

†Fachbereich Mathematik und Informatik, Philipps Universität, 35032 Marburg, Germany (boehmer@mathematik.uni-marburg.de).

‡Institute of Chemical Technology, Technická 5, 166 28 Prague, Czech Republic (janovskd@vscht.cz).

§Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 180 00 Prague 8, Czech Republic (janovsky@karlin.mff.cuni.cz).

For bifurcation points we adopt the classification with one distinguished parameter; see [17], [24], [23], [31], [19].

The current state of the art of numerical bifurcation analysis (see, e.g., [24], [31], [19]) could be briefly described as follows: Going down the hierarchy of bifurcation points, find an *organizing center*. This is a singular point with the highest *codimension* locally available. In our approach it is codim = 3; it is a higher codim in other specific cases.

Going up the hierarchy, we start with the *postprocessing* of a previously discovered organizing center: in [19, section 7.8] this consists of finding tangents to specific curves of solutions passing through the organizing center by reducing the codim by one. In [30], [34], and [31, Chap. 6], scaling techniques are employed to determine, even for corank ≥ 1, the tangents of bifurcating solutions.

In our paper, we propose a much more complex postprocessing analysis based on the bifurcation equation. The technique presented will supply a first-order approximation to the complete bifurcation scenario in a neighborhood of the organizing center of the original problem. Hence, we (approximately) transform the scenario in a neighborhood of, e.g., a pitchfork from the normal form to the original situation. Golubitsky and Schaeffer in [17, pp. 146, 147; Figs. 7.1, 7.2] demonstrate the value of this normal form analysis. Our postprocessing allows us to transform this back to the original situation. This is important, e.g., for the chemical and pharmaceutical industries in connection with the continuously stirred tank reactor, in particular, with many different reactants, and also is important in other applications; see, e.g., [17], [28], [29].

Our immediate motivation is the theory for imperfect bifurcation; see [16] and [17]. It yields *qualitative information* concerning the behavior of the bifurcation when the problem is subjected to an arbitrary sufficiently small perturbation. In this context we use the well-known *unfolded normal forms* of the bifurcation scenarios. They are models of the actual bifurcation problems under perturbations. However, there exists no *quantitative* link between the models, i.e., the unfolded normal form and the actual computed bifurcation problem.

So, the main goal of our approach is to compute (or at least to approximate) a diffeomorphism that links the unfolded normal form to the real bifurcation problem. For the general case of codim ≤ 3, this is achieved for the first time in our Theorem 3.6. Its use, in combination with Tables 2–11 listed below, is demonstrated in Example 3.1. In the case studies [21] and [8] for simple and pitchfork bifurcations, resp., we have verified that the idea really works numerically.

The outline of the paper is as follows. In our preliminaries (section 2), we construct a kind of contact equivalence between the unfolded normal form and the actual (reduced) bifurcation problem (Lemma 2.1). Hence, the roots of the particular unfolded normal form are diffeomorphic to the solutions of the actual bifurcation problem. The differential of the diffeomorphism yields first-order approximations to every kind of bifurcation point in a neighborhood of the organizing center.

In section 3, we show how to compute this differential for all organizing centers with codim ≤ 3, corank $= 1$; see Theorem 3.6, Example 3.1, and Tables 2–11. For the extremely technical proofs of the necessary lemmas and the determination of the tables, we refer to [9]. We assume that a dimensional reduction has already been performed.

We conclude with a brief review of the generalized Liapunov–Schmidt reduction and computation of data for our a posteriori analysis (section 4). This is needed for

the final transformation of the reduced-to-the-original bifurcation scenario.

2. Preliminaries. We consider a smooth parameter dependent mapping $F : \mathbb{R}^N \times \mathbb{R}^n \to \mathbb{R}^N$, e.g., obtained by discretizing a PDE. Let $F = F(u, \beta)$, $\beta = (\lambda, \alpha) \in \mathbb{R}^1 \times \mathbb{R}^k$, $n = 1 + k$. In the bifurcation context [17], u is the state variable, and λ and α are the control and unfolding parameters.

Let $(u^*, \lambda^*, \alpha^*) \in \mathbb{R}^N \times \mathbb{R}^{1+k}$ be a bifurcation point of F with corank $= 1$, i.e.,

$$F(u^*, \lambda^*, \alpha^*) = 0, \qquad \dim \operatorname{Ker} F_u(u^*, \lambda^*, \alpha^*) = 1. \tag{2.1}$$

The point $(u^*, \lambda^*, \alpha^*)$ plays the role of an organizing center.

Note that here and in the following, subscripts of a mapping denote partial differentials of the mapping w.r.t. the variable indicated by the subscript.

We consider a *Liapunov–Schmidt reduction*

$$g : \mathbb{R}^1 \times \mathbb{R}^{1+k} \to \mathbb{R}^1, \quad g = g(x, y), \quad y = (t, z),$$

of F at the point $(u^*, \lambda^*, \alpha^*)$; see, e.g., [17, p. 25] and also [24], [31], [19] for a computational version of the reduction, which we shall outline later in section 4. As a consequence of the reduction, the solution sets $F(u, \lambda, \alpha) = 0$ and $g(x, t, z) = 0$ are locally one-to-one (isomorphic) in neighborhoods of $(u^*, \lambda^*, \alpha^*)$ and $0 \in \mathbb{R}^{2+k}$. This isomorphism is described in section 4. It also links the singular roots of F and g; see, e.g., [19, Prop. 6.2.7] and [31, Thm. 6.1.1].

Let $h : \mathbb{R}^2 \to \mathbb{R}^1$ be defined as the restriction $h(x, t) \equiv g(x, t, 0)$ so that $h(x, t) = 0$ defines the *perfect bifurcation scenario*. Note that for all singularities $h = h_x = 0$ at the origin.

The classification of g is realized by the classification of the perfect bifurcation scenario $h(x, t) = 0$. This can be achieved by algebraic/geometric means, namely, by linking the map h to a suitable *normal form* $h^* : \mathbb{R}^2 \to \mathbb{R}^1$. It guarantees the existence, but does not allow the computation of the diffeomorphism. The bifurcation scenario of the normal form is usually well understood.

The link is formally defined as a *contact equivalence*: there exist a smooth $M : \mathbb{R}^2 \to \mathbb{R}^1$ and a local diffeomorphism $\Psi : \mathbb{R}^2 \to \mathbb{R}^2$, $\Psi(x, t) = (\chi(x, t), \tau(t))$, such that

$$\chi = 0,\ M > 0,\ \chi_x > 0 \text{ at } 0 \in \mathbb{R}^2 \text{ and } \tau = 0,\ \tau_t > 0 \text{ at } 0 \in \mathbb{R}^1, \tag{2.2}$$

and

$$h = M h^* \circ \Psi \tag{2.3}$$

in a neighborhood of $0 \in \mathbb{R}^2$.

We shall abbreviate (2.2), (2.3) by saying $h \sim h^*$; the relation $\sim$ is a well-defined equivalence on germs of smooth functions $\mathbb{R}^2 \to \mathbb{R}^1$; see [17, p. 104].

In Table 1, we list the normal forms h^* considered in this paper: the relevant g^* is a *universal unfolding* and k is the codimension; see [17, p. 196].

The role of g^* becomes clear from the following lemma.

LEMMA 2.1. *Let $M : \mathbb{R}^2 \to \mathbb{R}^1$ and a local diffeomorphism $\Psi : \mathbb{R}^2 \to \mathbb{R}^2$ satisfy* (2.2) *and* (2.3) *in a neighborhood of $0 \in \mathbb{R}^2$. Let $g^* : \mathbb{R}^2 \times \mathbb{R}^k \to \mathbb{R}^1$ be a universal unfolding of h^*. Then there exist smooth extensions S, Φ of M, Ψ,*

$$S : \mathbb{R}^{2+k} \to \mathbb{R}^1, \quad \Phi : \mathbb{R}^{2+k} \to \mathbb{R}^{2+k}, \quad \Phi(x, t, z) = (X(x, t, z), T(t, z), Z(z)) \tag{2.4}$$

TABLE 1

Normal forms h^* and universal unfoldings g^* for codim $= k \leq 3$, corank $= 1$. Here, $|p| = |q| = 1$. For details see Tables 2–11 below, indicating the following cases $= 1, \ldots, 10$.

Case	Normal form h^*	k	Universal unfolding g^* of h^*
1	$px^2 + qt$	0	$px^2 + qt$
2	$px^2 + qt^2$	1	$px^2 + qt^2 + z_1$
3	$px^3 + qt$	1	$px^3 + qt + z_1$
4	$px^2 + qt^3$	2	$px^2 + qt^3 + z_1 + z_2 t$
5	$px^3 + qtx$	2	$px^3 + qtx + z_1 + z_2 x^2$
6	$px^4 + qt$	2	$px^4 + qt + z_1 x + z_2 x^2$
7	$px^2 + qt^4$	3	$px^2 + qt^4 + z_1 + z_2 t + z_3 t^2$
8	$px^3 + qt^2$	3	$px^3 + qt^2 + z_1 + z_2 x + z_3 tx$
9	$px^4 + qtx$	3	$px^4 + qtx + z_1 + z_2 t + z_3 x^2$
10	$px^5 + qt$	3	$px^5 + qt + z_1 x + z_2 x^2 + z_3 x^3$

such that

$$S(\cdot,\cdot,0) = M(\cdot,\cdot), \qquad \Phi(\cdot,\cdot,0) = \Psi(\cdot,\cdot) \tag{2.5}$$

satisfy (2.2) *and*

$$g = S\ g^* \circ \Phi \tag{2.6}$$

in a neighborhood of $0 \in \mathbb{R}^{2+k}$.

Proof. Let us extend $\Psi : \mathbb{R}^2 \to \mathbb{R}^2$ as $\bar{\Psi} : \mathbb{R}^{2+k} \to \mathbb{R}^{2+k}$, by the local bijection $\bar{\Psi}(x,t,z) = (\chi(x,t), \tau(t), z)$; see (2.2). We define $f : \mathbb{R}^{2+k} \to \mathbb{R}^1$ as

$$f(x,t,z) = \big(M \circ \Psi^{-1}(x,t)\big)^{-1}\, g \circ \bar{\Psi}^{-1}(x,t,z).$$

Then, $f(x,t,0) = h \circ \Psi^{-1}(x,t) / \big(M \circ \Psi^{-1}(x,t)\big) = h^*(x,t)$; i.e., $f = f(x,t,z)$ is a k-parameter unfolding of $h^* = h^*(x,t)$; see [17, p. 120]. Since we assume $g^* = g^*(x,t,z)$ to be a universal unfolding of $h^* = h^*(x,t)$, the f factors through g^* [17, p. 120, Def. 1.2], and hence there exist smooth $\widetilde{S} : \mathbb{R}^{2+k} \to \mathbb{R}^1$ and $\widetilde{\Phi} : \mathbb{R}^{2+k} \to \mathbb{R}^{2+k}$, $\widetilde{\Phi}(x,t,z) = (\widetilde{X}(x,t,z), \widetilde{T}(t,z), \widetilde{Z}(z))$ satisfying the following conditions:

$$\widetilde{\Phi}(x,t,0) = (x,t,0)\,, \quad \widetilde{S}(x,t,0) = 1\,, \quad f = \widetilde{S} g^* \circ \widetilde{\Phi}$$

in a neighborhood of $0 \in \mathbb{R}^{2+k}$. Then it is easy to verify that

$$S = M(x,t)\widetilde{S} \circ \bar{\Psi}(x,t,z)\,, \quad \Phi = \widetilde{\Phi} \circ \bar{\Psi}$$

satisfy (2.5) and (2.6). □

In the particular applications of Lemma 2.1 we have to assume that Φ is a diffeomorphism in a neighborhood of the origin; i.e., $D\Phi(0) \in \mathcal{L}(\mathbb{R}^{(2+k)}, \mathbb{R}^{(2+k)})$ is regular. In section 3 we will explicitly determine the Ψ and the Φ for the four families of singularities of codim ≤ 3.

REMARK 2.1. *The assumption concerning the regularity of $D\Phi(0)$ is equivalent to the assumption that the gradient of the defining conditions is regular at the origin; see* [9], [19]. *The mentioned gradients are listed in* [17, Table 3.2, p. 204]; *for the defining conditions, see* [17, Table 2.3, p. 198].

Following the discussion in [17, Chap. III, section 4], *the gradient of the defining conditions is regular if and only if g is a universal unfolding of $h \equiv g(x,t,0)$.*

In less formal language, we will assume that g depicts all quantitatively significant perturbations of h.

Hence, assume $\det D\Phi(0) \neq 0$. Due to Lemma 2.1, $g(x,t,z) = 0$ if and only if $g^*(X,T,Z) = 0$, where $\Phi(x,t,z) = (X(x,t,z), T(t,z), Z(z))$; the statement holds in the obvious local sense. Recall the canonical structure of the diffeomorphism Φ, namely, that $z \mapsto Z(z)$ and $(t,z) \mapsto (T(t,z), Z(z))$ are diffeomorphisms in a neighborhood of the origin.

These facts suggest the following solution of $g = 0$: Choose an imperfection $z \in \mathbb{R}^k$ and a control parameter $t \in \mathbb{R}^1$; map them as $z \mapsto Z(z)$, $(t,z) \mapsto T(t,z)$; solve $g^*(X, T(t,z), Z(z)) = 0$ for X, which amounts to finding roots of an algebraic equation; and define $\Phi^{-1}(X, T(t,z), Z(z))$ as (x,t,z). Then (x,t,z) solves $g = 0$.

The same applies to singular roots of g and g^* since Φ^{-1} provides a one-to-one link between stratified manifolds of singular points of g and those of g^*.

EXAMPLE 2.1. *Consider the classical pitchfork (case 5 in Table 1) as an example. Note that* ($g =$) $g_x = g_t = g_{xx} = 0$ *and* $g_{xxx} g_{xt} \neq 0$ *are the defining equations and nondegeneracy conditions, resp., and $z \in \mathbb{R}^2$ the unfolding parameters. Since, e.g., g^*_{tt} is none of the conditions, the $g^*_{tt} = 0$ in Table 1 does not imply $g_{tt} = 0$ in Table 6.*

The set of all singular points of g is stratified as $\mathcal{L} = \{(x,t,z) : g = g_x = 0\}$, $\mathcal{B} = \{(x,t,z) : g = g_x = g_t = 0\}$ and $\mathcal{H} = \{(x,t,z) : g = g_x = g_{xx} = 0\}$; these are called limit points, simple bifurcation points, and hysteresis points. Obviously, $\mathcal{L}$ is the image $\Phi^{-1}(\mathcal{L}^*)$, $\mathcal{L}^* = \{(x,t,z) : g^* = {g^*}_x = 0\}$. Similarly, $\mathcal{B} = \Phi^{-1}(\mathcal{B}^*)$ and $\mathcal{H} = \Phi^{-1}(\mathcal{H}^*)$. The sets $\mathcal{L}^*$, $\mathcal{B}^*$, and $\mathcal{H}^*$ can be constructed explicitly.

The diffeomorphism Φ is not available in general. The objective of this paper is to compute $D\Phi(0)$. With $\Phi(0) = 0$ this yields the natural first-order approximation for Φ. We will come back to this problem in section 3.

Regarding the solution of $g = 0$, the following (partially) linearized approximation procedure replaces the above nonlinear approach: Choose $z \in \mathbb{R}^k$ and $t \in \mathbb{R}^1$; map them as $z \mapsto Z_z\, z$ and $(t,z) \mapsto T_t\, t + T_z\, z$; solve the algebraic equation $g^*(X, T_t\, t + T_z\, z, Z_z\, z) = 0$ for X; and invert $D\Phi(0)$ and define the action of $(D\Phi(0))^{-1}$ at $(X, T_t\, t + T_z\, z, Z_z\, z) \in \mathbb{R}^{2+k}$ as (x,t,z). Then (x,t,z) are first-order approximations to the roots of $g = 0$. Similarly, all *singular* solutions of $g = 0$ in a neighborhood of the origin (e.g., limit points $\mathcal{L}$) can be approximated by the singular solutions of $g^* = 0$ (e.g., limit points $\mathcal{L}^*$).

The analysis of singular roots of g has to be lifted from a small dimensional $\mathbb{R}^{2+k}$ to the actual state space $\mathbb{R}^{N+1+k}$, where the roots of F live. This will be done in section 4.

In summary, by processing $D\Phi(0)$, we will obtain a sort of first-order predictor for all singular points in the neighborhood of $(u^*, \lambda^*, \alpha^*)$. The implementation of this idea was tested numerically in two case studies, where organizing centers 2 and 5 of Table 1 were considered; see [21] and [8].

3. Computing the differential $D\Phi(0)$. The crucial step towards computing $D\Phi(0)$ is an explicit solution of the recognition problem [17]. Note that the solution of the recognition problem in [17] is not constructive; see Chapter II, section 11. We need a solution based on the implicit function theorem. For motivation and discussion see both case studies [21] and [8].

We can give only a rough idea for the procedure: To obtain an equation for $D\,\Phi\,(0)$, we compute the partials in (2.6) w.r.t. x, t in a neighborhood of $0 \in \mathbb{R}^{k+2}$ to obtain the vector of defining conditions, e.g., $(g, g_x, g_t, g_{xx}) = 0$, for the pitchfork. With $g_{xxx} g_{tx} \neq 0$ this characterizes the origin $0 \in \mathbb{R}^{k+2}$ as an appropriate bifurcation

point of g and g^*. Next, we compute the gradients w.r.t. x, t, z of these vectors, e.g., $(g, g_x, g_t, g_{xx})^T$ for g [24], [23] and g^* [17]. They are collected to define $\mathbf{B}$ and $\mathbf{B}^*$. By [19], g and g^* represent a universal unfolding of h and h^* if and only if the corresponding matrix $\mathbf{B}$ and $\mathbf{B}^*$ is regular, respectively. Then we obtain for $D\,\Phi\,(0)$ the following equation:

$$\mathbf{B} \;=\; \mathbf{A}\,\mathbf{B}^*\,D\,\Phi\,(0); \tag{3.1}$$

in $\mathbf{A}$ we collect all other terms, e.g., products and sums of $S,\ S_x,\ S_t,\ X,\ X_x,\ X_t$, a.s.o., obtained in the above differentiation [9]. Note that the block structure of $D\Phi(0)$ reflects the canonical structure of Φ:

$$D\Phi(0) = \begin{pmatrix} X_x & X_t & X_z \\ 0 & T_t & T_z \\ 0 & 0 & Z_z \end{pmatrix} \in \mathcal{L}(\mathbb{R}^{(2+k)}, \mathbb{R}^{(2+k)})\,, \qquad Z_z \in \mathcal{L}(\mathbb{R}^k, \mathbb{R}^k)\,; \tag{3.2}$$

here and in (3.1) the required partial derivatives of $X(x,t,z)$, $T(t,z)$, and $Z(z)$ are evaluated at the origin. Now, the unknown terms in $\mathbf{A}, D\,\Phi\,(0)$ have to be determined by (3.1). For a naive approach this system usually has more unknowns than equations. So we need additional information.

Following [19], normal forms for singularities of codim$\leq$ 3 are organized in four families: pitchfork family (cases 5, 9 of Table 1), hysteresis family (cases 1, 3, 6, 10), asymmetric cusp family (cases 2, 4, 7), and a singleton winged cusp (case 8). For each family the structure of the $\chi(x,t), \tau(t)$ in (2.2) is described in detail in the following lemmas, finally allowing us to solve (3.1). We need the following conditions for the lemmas.

(3.3) Let $M = M(x,t),\ \chi = \chi(x,t),\ H = H(x),\ \omega = \omega(x), \psi = \psi(t),\ \tau = \tau(t),$ $a = a(t),\ b = b(t),\ c = c(t)$ be smooth real-valued functions in a neighborhood of the origin and $M(0,0) > 0$.

LEMMA 3.1. Pitchfork family. *Let* $h^* = px^n + qxt$, $n \geq 3$, $|p| = |q| = 1$. *Then* $h \sim h^*$ *if and only if* (*see* (3.3))

$$\begin{aligned} h(x,t) &= M \cdot \left(p\,\chi^k \,+\, q\,\chi\,\tau\right), \ \textit{where} \\ \tau(t) &= c\,t\,, \quad c > 0, \\ \chi(x,t) &= (x - \psi(t))\,H(x - \psi(t)) \ \textit{with}\ \psi(0) = 0,\ H(0) = 1. \end{aligned} \tag{3.4}$$

LEMMA 3.2. Hysteresis family. *Let* $h^* = px^n + qt$, $n \geq 2$, $|p| = |q| = 1$. *Then* $h \sim h^*$ *if and only if* (*see* (3.3))

$$\begin{aligned} h(x,t) &= M \cdot (p\,\chi^n \,+\, q\,\tau)\,, \ \textit{where} \\ \tau(t) &= c^n\,t\,, \quad c = \frac{1}{\omega(0)} > 0, \\ \chi(x,t) &= \chi(x) \,=\, cx\,\omega(x)\,. \end{aligned} \tag{3.5}$$

LEMMA 3.3. Asymmetric cusp family. *Let* $h^* = px^2 + qt^n$, $n \geq 2$, $|p| = |q| = 1$. *Then* $h \sim h^*$ *if and only if* (*see* (3.3))

$$h(x,t) = M \cdot \left(p\,\chi^2 \,+\, q\,\tau^n\right), \ \textit{where} \tag{3.6}$$

$$\tau(t) = (pqb(t))^{1/n}\,, \tag{3.7}$$

$$\chi(x,t) = x + a(t)\,, \tag{3.8}$$

with $a(0) = 0$ *and* $b(0) = b'(0) = 0$, $sgn(b''(0)) = pq$.

LEMMA 3.4. Winged cusp. *Let* $h^* = px^3 + qt^2$, $|p| = |q| = 1$. *Then* $h \sim h^*$ *if and only if* (*see* (3.3))

$$h(x,t) = M \cdot \left(p\,\chi^3 + q\,\tau^2\right), \text{ where} \tag{3.9}$$

$$\tau(t) = (pqc(t))^{1/2}, \tag{3.10}$$

$$\chi(x,t) = \left((x+a(t))((x+a(t))^2 + b(t))\right)^{1/3}, \tag{3.11}$$

with $a(0) = 0$, $b(0) = b'(0) = 0$ *and* $c(0) = c'(0) = 0$, $sgn(c''(0)) = pq$.

Proof. Lemma 3.1 follows from Lemma 2.7 in [16]. Lemmas 3.2–3.4 can be proved similarly; see [9]. □

For $h \sim h^*$, these lemmas yield sufficient information to construct a pair M, $\Psi = (\chi, \tau)$ such that (2.2), (2.3) hold. Note that the construction is certainly not unique: M/c, $c\Psi$ would be another pair, provided that $c \neq 0$. However, imposing the scaling condition $\chi_x(0) = 1$ makes the choice unique. Then Lemma 2.1 can be used to determine the $\Phi(x,t,z) = (X(x,t,z), T(t,z), Z(z))$ for the four families of singularities of codim ≤ 3.

DEFINITION 3.5. *Let* $h \sim h^*$; *i.e.,* (2.2), (2.3), *and* $\chi_x(0) = 1$ *hold. We say that* $M : \mathbb{R}^2 \to \mathbb{R}^1$, $\Psi(x,t) \equiv (\chi(x,t), \tau(t)) : \mathbb{R}^2 \to \mathbb{R}^2$ *are the scaled solutions of the recognition problem, provided that they are defined as suggested in Lemmas* 3.1–3.4, *respectively.*

Recall Lemma 2.1 and its assumptions. The scaled solution of the recognition problem is a legitimate candidate for a choice of M and Ψ.

THEOREM 3.6. *Consider a Liapunov–Schmidt reduction* g *of* F *at a singular point* $(u^*, \lambda^*, \alpha^*)$. *Let* $h \equiv g(x,t,0)$ *be equivalent to a normal form,* $h \sim h^*$, *from Table* 1. *Let* g *be a universal unfolding of* h. *Construct the diffeomorphism* Φ *in Lemma* 2.1, *employing the scaled solution of the recognition problem.*

Then for every family of singularities the $D\Phi(0)$ *is uniquely defined. In order to compute* $D\Phi(0)$, *the evaluation of a finite number of partial derivatives of* g *at the origin is needed.*

Proof. This is constructive. Tables 2–11 summarize the explicit relationship between data (selected partial derivatives of g at the origin) and entries of $D\Phi(0)$. Details are given in [9]. □

We demonstrate how to use Tables 2–11 in order to compute $D\Phi(0)$.

EXAMPLE 3.1. *We consider the pitchfork case* 5 *in Table* 1 *with* $g = g_x = g_t = g_{xx} = 0$, $g_{xxx}g_{xt} \neq 0$. *Table* 6 *is the table which is relevant to this case. The data represent* 16 *partial derivatives* g_{xxx}, g_{xt}, g_{tt}, g_{xtt}, g_{xxt}, g_{xxxx}, g_{xxxt}, $\partial^5 g/\partial x^5 \in \mathbb{R}^1$, $(g_z)^\top$, $(g_{xz})^\top$, $(g_{tz})^\top$, *and* $(g_{xxz})^\top \in \mathbb{R}^2$. *The claim is that the data define the matrix* $D\Phi(0) \in \mathcal{L}(\mathbb{R}^4, \mathbb{R}^4)$; *see* (3.2). *The entry* $X_x \equiv \chi_x(0)$ *is scaled to* 1. *Table* 6 *should be interpreted as a set of* 16 *nonlinear equations for*

- X_t, X_z, T_t, T_z, *and* Z_z *which are the entries of the scaled* $D\Phi(0)$. $X_t \in \mathbb{R}^1$, $(X_z)^\top \in \mathbb{R}^2$, $T_t \in \mathbb{R}^1$, $(T_z)^\top \in \mathbb{R}^2$, *and* $Z_z = ((Z_1)_z, (Z_2)_z) \in \mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$;
- S, S_x, S_t, S_{xx}, X_{xx}, $X_{xxx} \in \mathbb{R}^1$ *resulting from the chain rule and necessary to determine the above first partials;*

hence, 16 *equations for* 16 *unknowns altogether. This (and the other) system is solvable step by step in only linear subsystems of one or two equations with nonvanishing determinants. This is indicated in the last column of Table* 6 *starting with* S, T_t. *The pairs* S_x, X_{xx} *and* S_{xx}, X_{xxx} *are coupled linearly. To indicate that, we have introduced the horizontal lines in Tables* 6 *and* 10.

TABLE 2
Limit point, $X_t = 0$.

Data			⇒
g_{xx}	=	$2pS$	S
g_t	=	qST_t	T_t

TABLE 3
Simple bifurcation point and isola center, $z \in \mathbb{R}^1$.

Data			⇒
g_{xx}	=	$2pS$	S
g_z	=	SZ_z	Z_z
g_{xt}	=	$2pSX_t$	X_t
g_{tt}	=	$2pSX_t^2 + 2qST_t^2$	T_t
g_{xxx}	=	$6pS_x$	S_x
g_{xz}	=	$S_xZ_z + 2pSX_z$	X_z
g_{xxt}	=	$4pS_xX_t + 2pS_t$	S_t
g_{tz}	=	$S_tZ_z + 2pSX_tX_z + 2qST_tT_z$	T_z

4. Remarks on implementation. Let us briefly review the Liapunov–Schmidt reduction via *bordered systems* following [19, section 7.4].

Consider $(u^*, \lambda^*, \alpha^*)$, satisfying (2.1). Let L and M be a pair of fixed vectors $M \in \mathbb{R}^N$, $L^\top \in \mathbb{R}^N$. Let the matrix

$$\mathcal{J}(u^*, \lambda^*, \alpha^*) \equiv \begin{pmatrix} F_u(u^*, \lambda^*, \alpha^*) & M \\ L & 0 \end{pmatrix} \in \mathcal{L}(\mathbb{R}^{(N+1)}, \mathbb{R}^{(N+1)})$$

be regular (which is the case for a generic choice of the *bordering* vectors M and L and necessary for the approach in [21], [7]. For appropriate choices of M, L in $\mathcal{J}$ with good condition numbers see [35], [34].) The nonlinear system

$$F\,(u^* + v, \lambda^* + t, \alpha^* + z) - M\,g\, = 0\,, \qquad L\,v\, =\, x \tag{4.1}$$

implicitly defines $g = g(x,t,z)$ and $v = v(x,t,z)$ as germs of smooth mappings $g : \mathbb{R}^1 \times \mathbb{R}^{1+k} \to \mathbb{R}^1$ and $v : \mathbb{R}^1 \times \mathbb{R}^{1+k} \to \mathbb{R}^N$ centered at the origin, i.e., satisfying $g(0) = 0$ and $v(0) = 0$. This particular definition of g is called Liapunov–Schmidt reduction.

Note that $F(u, \lambda, \alpha) = 0$ if and only if $g(x,t,z) = 0$, where $u = u^* + v(x,t,z)$, $\lambda = \lambda^* + t$, and $\alpha = \alpha^* + z$. In other words, there exists a local isomorphism between the roots of F and g. Hence, as a consequence of Lemma 2.1, the roots of F and g^* are in one-to-one correspondence. The same argument can be used regarding *singular* roots of F, g, and g^*.

Let $g(x,t,z) = 0$. Consider $(u, \lambda, \alpha) \in \mathbb{R}^{N+1+k}$,

$$u = u^* + v_x\, x + v_t\, t + v_z\, z\,, \quad \lambda = \lambda^* + t\,, \quad \alpha = \alpha^* + z, \tag{4.2}$$

where the partial differentials of v are evaluated at $(0,0,0) \in \mathbb{R}^{2+k}$. The point (u, λ, α) approximates a root of F. This approximation is to the first order.

TABLE 4
Hysteresis, $X_t = 0$, $z \in \mathbb{R}^1$.

Data			⇒
g_{xxx}	=	$3!pS$	S
g_t	=	qST_t	T_t
g_z	=	$qST_z, z \in \mathbb{R}^1$	T_z
g_{xt}	=	qS_xT_t	S_x
g_{xz}	=	$qS_xT_z + SZ_z$	Z_z
g_{xxxx}	=	$4!pS_x + 6 \cdot 3!pSX_{xx}$	X_{xx}
g_{xxt}	=	$qS_{xx}T_t$	S_{xx}
g_{xxz}	=	$qS_{xx}T_z + 2S_xZ_z + 3!pSX_z + SZ_zX_{xx}$	X_z

TABLE 5
Asymmetric cusp, $z \in \mathbb{R}^2$.

Data			⇒
g_{xx}	=	$2pS$	S
g_{xt}	=	$2pSX_t$	X_t
g_z	=	$S(Z_1)_z$	$(Z_1)_z$
g_{xxx}	=	$6pS_x$	S_x
g_{xz}	=	$2pSX_z + S_x(Z_1)_z$	X_z
g_{xxt}	=	$4pS_xX_t + 2pS_t$	S_t
$(D_2(g))_x$	=	$4S^2X_{tt}$	X_{tt}
$(D_2(g))_t$	=	$4S^2X_tX_{tt} + 12pqS^2T_t^3$	T_t
g_{tz}	=	$2pSX_tX_z + S_t(Z_1)_z + ST_t(Z_2)_z$	$(Z_2)_z$
g_{xxxx}	=	$12pS_{xx}$	S_{xx}
g_{xxxt}	=	$6pS_{xx}X_t + 6pS_{xt}$	S_{xt}
g_{xxtt}	=	$2pS_{xx}X_t^2 + 8pS_{xt}X_t + 4pS_xX_{tt} + 2pS_{tt}$	S_{tt}
g_{xttt}	=	$6pS_{xt}X_t^2 + 6pS_{tt}X_t + 6pS_xX_tX_{tt} + 6qS_xT_t^3$	
		$+6pS_tX_{tt} + 2pSX_{ttt}$	X_{ttt}
g_{tttt}	=	$12pS_{tt}X_t^3 + 24pS_tX_tX_{tt} + 24qS_tT_t^3 + 6pSX_{tt}^2$	
		$+8pSX_tX_{ttt} + 30qST_t^2T_{tt}$	T_{tt}
$(D_2(g))_z$	=	$12pqS^2T_t^2T_z + 2pS(Z_1)_z(X_t^2S_{xx} + S_{tt} - 2X_tS_{xt})$	
		$+2pS(Z_2)_z(2T_tS_t - 2S_xX_tT_t) + 4S^2X_zX_{tt}$	T_z

TABLE 6
Pitchfork, $z \in \mathbb{R}^2$.

Data			$\Rightarrow$
g_{xxx}	$=$	$6pS$	$S \neq 0$
g_{xt}	$=$	qST_t	$T_t \neq 0$
g_z	$=$	$S(Z_1)_z$	$(Z_1)_z$
g_{tt}	$=$	$2qST_tX_t$	X_t
g_{xxt}	$=$	$qSX_{xx}T_t + 2qT_tS_x + 6pSX_t$	$S_x,\ X_{xx}$
g_{xxxx}	$=$	$36pSX_{xx} + 24pS_x$	
g_{xz}	$=$	$qST_z + S_x(Z_1)_z$	T_z
g_{xtt}	$=$	$2qT_t(SX_tX_{xx} + X_tS_x + S_t) + 6pSX_t^2$	S_t
g_{tz}	$=$	$qSX_zT_t + qST_zX_t + (Z_1)_zS_t$	X_z
g_{xxxt}	$=$	$qT_t(SX_{xxx} + 3S_xX_{xx} + 3S_{xx})$	
		$+18pX_t(2SX_{xx} + S_x) + 6pS_t$	$S_{xx},\ X_{xxx}$
$\frac{\partial^5}{\partial x^5}g$	$=$	$30p(2SX_{xxx} + 3SX_{xx}^2 + 6S_xX_{xx} + 2S_{xx})$	
g_{xxz}	$=$	$qSX_{xx}T_z + 2qS_xT_z + 6pSX_z + (Z_1)_zS_{xx}$ $+2S(Z_2)_z$	$(Z_2)_z$

TABLE 7
Quartic fold, $X_t = 0$, $z \in \mathbb{R}^2$.

Data			$\Rightarrow$
g_{xxxx}	$=$	$4!pS$	S
g_t	$=$	qST_t	T_t
g_z	$=$	qST_z	T_z
g_{xt}	$=$	qS_xT_t	S_x
g_{xz}	$=$	$qS_xT_z + S(Z_1)_z$	$(Z_1)_z$
g_{xxt}	$=$	$qS_{xx}T_t$	S_{xx}
$\frac{\partial^5}{\partial x^5}g$	$=$	$5!pS_x + 2 \cdot 5!pSX_{xx}$	X_{xx}
g_{xxz}	$=$	$qS_{xx}T_z + 2S_x(Z_1)_z + S(Z_1)_zX_{xx} + 2S(Z_2)_z$	$(Z_2)_z$
$\frac{\partial^6}{\partial x^6}g$	$=$	$9 \cdot 5!pSX_{xx}^2 + 4 \cdot 5!pSX_{xxx} + 3 \cdot 5!pS_{xx} + 2 \cdot 6!pS_xX_{xx}$	X_{xxx}
g_{xxxt}	$=$	$qS_{xxx}T_t$	S_{xxx}
g_{xxxz}	$=$	$qS_{xxx}T_z + 3S_{xx}(Z_1)_z + 3S_xX_{xx}(Z_1)_z + 3!S_x(Z_2)_z$	
		$+4!pSX_z + 3(Z_1)_zX_{xxx} + 3!SX_{xx}(Z_2)_z$	X_z

TABLE 8
Asymmetric cusp, $z \in \mathbb{R}^3$.

Data			$\Rightarrow$
g_{xx}	$=$	$2pS$	S
g_{xt}	$=$	$2pSX_t$	X_t
g_z	$=$	$S(Z_1)_z$	$(Z_1)_z$
$(D_2(g))_x$	$=$	$4S^2X_{tt}$	X_{tt}
g_{xxx}	$=$	$6pS_x$	S_x
g_{xz}	$=$	$2pSX_z + S_x(Z_1)_z$	X_z
g_{xxt}	$=$	$4pS_xX_t + 2pS_t$	S_t
$(D_3(g))_x$	$=$	$8pS^2(SX_{ttt} + S_tX_{tt} - S_xX_tX_{tt})$	X_{ttt}
$(D_3(g))_t$	$=$	$96qS^3T_t^4 + 8pS^2X_tX_{tt}(S_t - S_xX_t) + 8pS^3(X_{tt}^2 + X_tX_{ttt})$	T_t
g_{tz}	$=$	$2pSX_tX_z + S_t(Z_1)_z + ST_t(Z_2)_z$	$(Z_2)_z$
g_{xxxx}	$=$	$12pS_{xx}$	S_{xx}
g_{xxxt}	$=$	$6pS_{xx}X_t + 6pS_{xt}$	S_{xt}
g_{xxtt}	$=$	$2pS_{xx}X_t^2 + 8pS_{xt}X_t + 4pS_xX_{tt} + 2pS_{tt}$	S_{tt}
$\frac{\partial^5}{\partial x^5}g$	$=$	$20pS_{xxx}$	S_{xxx}
$\frac{\partial^5}{\partial x^4\partial t}g$	$=$	$8pS_{xxx}X_t + 12pS_{xxt}$	S_{xxt}
$\frac{\partial^5}{\partial x^3\partial t^2}g$	$=$	$2pS_{xxx}X_t^2 + 12pS_{xxt}X_t + 6pS_{xtt} + 6pS_{xx}X_{tt}$	S_{xtt}
$\frac{\partial^5}{\partial x^2\partial t^3}g$	$=$	$6pS_{xxt}X_t^2 + 12pS_{xtt}X_t + 6pS_{xx}X_tX_{tt} + 12pS_{xt}X_{tt}$ $+2pS_{ttt} + 4pS_xX_{ttt}$	S_{ttt}
$\frac{\partial^5}{\partial x\partial t^4}g$	$=$	$8pS_{ttt}X_t + 12pS_{xtt}X_t^2 + 12pS_{tt}X_{tt} + 24pS_{xt}X_tX_{tt}$	
		$+8pS_tX_{ttt} + 6pS_xX_{tt}^2 + 8pS_xX_tX_{ttt} + 24qS_xT_t^4 + 2pS\frac{\partial^4}{\partial t^4}X$	$\frac{\partial^4}{\partial t^4}X$
$\frac{\partial^5}{\partial t^5}g$	$=$	$20pS_{ttt}X_t^2 + 60pS_{tt}X_tX_{tt} + 30pS_tX_{tt}^2 + 40pS_tX_tX_{ttt}$	
		$+20pSX_{tt}X_{ttt} + 10pSX_t\frac{\partial^4}{\partial t^4}X + 120qS_tT_t^4 + 240qST_t^3T_{tt}$	T_{tt}
$(D_2(g))_z$	$=$	$2pS(Z_1)_z(X_t^2S_{xx} + S_{tt} - 2X_tS_{xt}) + 4pS^2T_t^2(Z_3)_z$	
		$+2pS(Z_2)_z(2T_tS_t - 2S_xX_tT_t + ST_{tt}) + 4S^2X_zX_{tt}$	$(Z_3)_z$
$(D_3(g))_z$	$=$	$96qS^3T_t^3T_z + 16pSS_xX_t^2(S_xX_tX_z - S_xX_tT_z - 2S_tX_z)$	
		$+ 8pS^2(SX_zX_{ttt} + S_tX_zX_{tt} - S_xX_tX_zX_{tt})$	
		$+ 4S(Z_1)_z[S(S_{ttt} - 3S_{xtt}X_t + 3S_{xxt} + X_t^2 - S_{xxx}X_t^3)$	
		$+ S_{xx}X_t(3SX_{tt} + S_tX_t - S_xX_t^2) + S_{tt}(S_t - S_xX_t)$	T_z
		$+ S_{xt}(2S_xX_t^2 - 2S_tX_t - 3SX_{tt})]$	
		$+ 4S(Z_2)_z[3ST_t(S_{xx}X_t^2 + S_{tt} - 2S_{xt}X_t)$	
		$+ 2T_t(X_t^2S_x^2 + S_t^2 - 2S_xS_tX_t)$	
		$+ S(4S_tT_{tt} - 4S_xX_tT_{tt} - 3S_xX_{tt}T_t) + S^2T_{ttt}]$	
		$+ 8(Z_3)_zS^2T_t(3ST_{tt} + 4S_tT_t - 4S_xX_tT_t)$	

TABLE 9
Winged cusp, $z \in \mathbb{R}^3$.

Data			⇒
g_{xxx}	=	$6pS$	S
g_{xxt}	=	$6pSX_t$	X_t
g_{tt}	=	$2qST_t^2$	T_t
g_{xtt}	=	$2qS_xT_t^2 + 6pSX_t^2$	S_x
g_z	=	$S(Z_1)_z$	$(Z_1)_z$
g_{xz}	=	$S_x(Z_1)_z + S(Z_2)_z$	$(Z_2)_z$
g_{xxxt}	=	$18pS_xX_t + 6pS_t$	S_t
g_{ttt}	=	$6pSX_t^3 + 6qS_tT_t^2 + 6qST_tT_{tt}$	T_{tt}
g_{tz}	=	$2qST_tT_z + S_t(Z_1)_z$	T_z
$\frac{\partial^5}{\partial x^5}g$	=	$60pS_{xx}$	S_{xx}
g_{xxz}	=	$6pSX_z + (Z_1)_zS_{xx} + 2S_x(Z_2)_z$	X_z
g_{xxtt}	=	$2qS_{xx}T_t^2 + 12pS_xX_t^2 + 12pS_tX_t + 6pSX_{tt}$	X_{tt}
g_{xttt}	=	$6qS_{xt}T_t^2 + 6pS_xX_t^3 + 18pS_tX_t^2 + 18pSX_tX_{tt} + 6qS_xT_tT_{tt}$	S_{xt}
g_{xtz}	=	$2qS_xT_tT_z + 6pSX_tX_z + (Z_1)_zS_{xt}$	
		$+(Z_2)_z(S_xX_t + S_t) + ST_t(Z_3)_z$	$(Z_3)_z$

At the end of section 2, we described a first-order approximation procedure for (possibly singular) roots of g. Hence, whatever root of g is *approximated*, the transformation (4.2) is adequate to lift its coordinates from $\mathbb{R}^{1+1+k}$ to the original state space $\mathbb{R}^{N+1+k}$.

As far as the proposed analysis is concerned, the main cost represents a computation of the partial derivatives of g which are required in Tables 2–11. All these derivatives are to be evaluated at the origin.

Note that the majority of these derivatives is contained in the gradient of the defining equations; see Remark 2.1. This gradient is obtained as a byproduct of the computation of $(u^*, \lambda^*, \alpha^*)$ by the Newton (or a Newton-like) method; see [19, section 7.4] and also the case study [8].

There are algorithms for a systematic computation of these partials; see [19, section 7.4]. They are computed by solving canonical linear systems. These systems have the same matrix, namely, $\mathcal{J}(u^*, \lambda^*, \alpha^*)$. Moreover, the partials of g are naturally computed with the relevant partial derivatives of v. So, by computing g_x we compute simultaneously v_x (from (4.2)), etc.

Having this fact in mind, one may ask why we have not used the higher derivatives of v (which we had computed anyway) for a higher order approximation in (4.2). Consider, say, the quadratic approximation. Observe that in Tables 2–11 the g_{zz} never shows up in the data, which means that we do not need g_{zz} for the first-order analysis of g. Therefore, computing v_{zz} would mean making an extra and "unsystematic" effort.

TABLE 10
Pitchfork, $z \in \mathbb{R}^3$.

Data			$\Rightarrow$
$\frac{\partial^4}{\partial x^4} g$	$=$	$4!pS$	S
g_{xt}	$=$	qST_t	T_t
g_z	$=$	$S(Z_1)_z$	$(Z_1)_z$
g_{tt}	$=$	$2qST_tX_t$	X_t
g_{xxt}	$=$	$qST_tX_{xx} + 2qT_tS_x$	$S_x,\ X_{xx}$
$\frac{\partial^5}{\partial x^5} g$	$=$	$5!pS_x + 2 \cdot 5!pSX_{xx}$	
g_{xz}	$=$	$qST_z + S_x(Z_1)_z$	T_z
g_{xtt}	$=$	$2qT_t(SX_tX_{xx} + S_xX_t + S_t) + 6pSX_t^2$	S_t
g_{tz}	$=$	$qST_tX_z + qSX_tT_z + S_t(Z_1)_z$	X_z
g_{xxxt}	$=$	$qT_t(SX_{xxx} + 3S_xX_{xx} + 3S_{xx}) + 24pSX_t$	$S_{xx}\ X_{xxx}$
$\frac{\partial^6}{\partial x^6} g$	$=$	$5!p(3S_{xx} + 12S_xX_{xx} + 9SX_{xx}^2 + 4SX_{xxx})$	
g_{xxz}	$=$	$qST_zX_{xx} + 2qS_xT_z + (Z_1)_zS_{xx} + 2S(Z_2)_z$	$(Z_2)_z$
$\frac{\partial^5}{\partial x^4 \partial t} g$	$=$	$qT_t(4S_{xxx} + 6S_{xx}X_{xx} + 4S_xX_{xxx} + SX_{xxxx})$ $+4!pS_t + 2 \cdot 4!pX_t(2S_x + 5SX_{xx})$	$S_{xxx},\ \frac{\partial^4}{\partial x^4} X$
$\frac{\partial^7}{\partial x^7} g$	$=$	$7 \cdot 5!p(S_{xxx} + 6S_{xx}X_{xx} + 9S_xX_{xx}^2 + 4S_xX_{xxx}$ $+3SX_{xx}^3 + 6SX_{xx}X_{xxx} + SX_{xxxx})$	
g_{xxxz}	$=$	$qT_z(3S_{xx} + 3S_xX_{xx} + SX_{xxx}) + 4!pSX_z$ $+S_{xxx}(Z_1)_z + 6(Z_2)_z(S_x + SX_{xx}) + 6S(Z_3)_z$	$(Z_3)_z$

The required derivatives of g at the origin could be approximated by classical finite differences. Nevertheless, except for the origin where g is pinned to zero, the function $(x,t,z) \mapsto g(x,t,z) \in \mathbb{R}^1$ is not known, and its values have to be computed, as shown below.

Given (x,t,z), set $v^{(0)} = v_x\, x + v_t\, t + v_z\, z$ to be an initial approximation of $v(x,t,z)$; for v_x, v_t, and v_z, see (4.2). Consider the following iterations: Find $\delta v \in \mathbb{R}^N$ and $g^{(j+1)} \in \mathbb{R}^1$ as the solution of the linear system

$$\mathcal{J}\big(u^* + v^{(j)}, \lambda^* + t, \alpha^* + z\big) \begin{pmatrix} \delta v \\ -g^{(j+1)} \end{pmatrix} = \begin{pmatrix} -F(u^* + v^{(j)}, \lambda^* + t, \alpha^* + z) \\ 0 \end{pmatrix}$$

and update $v^{(j+1)} := v^{(j)} + \delta v$.

If (x,t,z) is sufficiently close to $0 \in \mathbb{R}^{2+k}$, then the iteration process is locally quadratically convergent. In particular, $v^{(j)} \to v(x,t,z)$ and $g^{(j)} \to g(x,t,z)$ as $j \to \infty$. For details see [22], [6], [7].

TABLE 11
Hysteresis $X_t = 0$, $z \in \mathbb{R}^3$.

Data			$\Rightarrow$
$\frac{\partial^5}{\partial x^5} g$	=	$5!pS$	S
g_t	=	qST_t	T_t
g_z	=	qST_z	T_z
g_{xt}	=	qS_xT_t	S_x
g_{xz}	=	$qS_xT_z + S(Z_1)_z$	$(Z_1)_z$
g_{xxt}	=	$qS_{xx}T_t$	S_{xx}
g_{xxxt}	=	$qS_{xxx}T_t$	S_{xxx}
$\frac{\partial^5}{\partial x^4 \partial t} g$	=	$qS_{xxxx}T_t$	S_{xxxx}
$\frac{\partial^6}{\partial x^6} g$	=	$6!pS_x + 15 \cdot 5!pSX_{xx}$	X_{xx}
g_{xxz}	=	$qS_{xx}T_z + 2S_x(Z_1)_z + S(Z_1)_zX_{xx} + 2S(Z_2)_z$	$(Z_2)_z$
$\frac{\partial^7}{\partial x^7} g$	=	$21 \cdot 5!pS_{xx} + 105 \cdot 5!pS_xX_{xx}$	
		$+105 \cdot 5!pSX_{xx}^2 + 35 \cdot 5!pSX_{xxx}$	X_{xxx}
g_{xxxz}	=	$qS_{xxx}T_z + 3S_{xx}(Z_1)_z + 3S_x(Z_1)_zX_{xx} + 6S_x(Z_2)_z$	
		$+S(Z_1)_zX_{xxx} + 6S(Z_2)_zX_{xx} + 6S(Z_3)_z$	$(Z_3)_z$
$\frac{\partial^8}{\partial x^8} g$	=	$56 \cdot 5!pS_{xxx} + 10 \cdot 7!pS_{xx}X_{xx} + 20 \cdot 7!pS_xX_{xx}^2$	
		$+280 \cdot 5!pS_xX_{xxx} + 10 \cdot 7!pSX_{xx}^3$	X_{xxxx}
		$+560 \cdot 5!pSX_{xx}X_{xxx} + 70 \cdot 5!pSX_{xxxx}$	
$\frac{\partial^5}{\partial x^4 \partial z} g$	=	$qS_{xxxx}T_z + 4S_{xxx}(Z_1)_z + 6S_{xx}(Z_1)_zX_{xx}$	
		$+12S_{xx}(Z_2)_z + 4S_xX_{xxx}(Z_1)_z + 24S_xX_{xx}(Z_2)_z$	X_z
		$+4!pS_x(Z_3)_z + 5!pSX_z + S(Z_1)_zX_{xxxx}$	
		$+6S(Z_2)_zX_{xx}^2 + 8S(Z_2)_zX_{xxx} + 36(Z_3)_zX_{xx}$	

REFERENCES

[1] E. L. ALLGOWER AND H. SCHWETLICK, *A general view of minimally extended systems for simple bifurcation points*, Z. Angew. Math. Mech., 77 (1997), pp. 83–97.

[2] K. BÖHMER, *On hybrid methods for bifurcation studies for general operator equations*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, Berlin, Heidelberg, New York, 2001, pp. 73–107.

[3] K. BÖHMER, *On a numerical Liapunov-Schmidt method for operator equations*, Computing, 51 (1993), pp. 237–269.

[4] K. BÖHMER, *On numerical bifurcation studies for general operator equations*, in Proceedings of International Conference on Differential Equations, 1999, Vol. 2, J. Sprekels, B. Fiedler, and K. Gröger, eds., World Scientific, Singapore, 2000, pp. 877–883.

[5] K. BÖHMER, *Finite Element and Collocation Methods with Variational Crimes*, Technical report, Fachbereich Mathematik, Philipps-Universität Marburg, Marburg, Germany, 2001.

[6] K. BÖHMER, W. GOVAERTS, AND V. JANOVSKY, *Numerical detection of symmetry breaking bifurcation points with nonlinear degeneracies*, Math. Comp., 68 (1999), pp. 1097–1108.

[7] K. BÖHMER, D. JANOVSKÁ, AND V. JANOVSKY, *Numerical analysis of the imperfect bifurcation diagrams*, Z. Angew. Math. Mech., 77 (1997), pp. 445–448.

[8] K. BÖHMER, D. JANOVSKÁ, AND V. JANOVSKY, *Computer aided analysis of imperfect bifurcation diagrams*, East-West J. Numer. Math., 6 (1998), pp. 207–222.

[9] K. Böhmer, D. Janovská, and V. Janovsky, *A Postprocessing Analysis for Bifurcation Singularities with codim* ≤ 3 *and corank* $= 1$, Technical report, University of Marburg, Marburg, Germany, 1999.

[10] K. Böhmer and N. Sassmannshausen, *Numerical Liapunov-Schmidt spectral method for k-determined problems*, Comput. Methods Appl. Mech. Engrg., 170 (1999), pp. 277–312.

[11] K. Böhmer and N. Sassmannshausen, *Stability for generalized Petrov-Galerkin methods applied to bifurcation*, ZAMM Z. Angew. Math. Mech., submitted.

[12] E. Doedel and J. P. Kernevez, *AUTO. Software for Continuation and Bifurcation Problems in Ordinary Differential Equations*, California Institute of Technology, Pasadena, CA, 1986.

[13] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. J. Wang, *AUTO* 97: *Continuation and Bifurcation Software for Ordinary Differential Equations (with Homcont)*, Technical report, California Institute of Technology, Pasadena, CA, 1997.

[14] E. J. Doedel, X. J. Wang, and T. F. Fairgrieve, *AUTO* 94: *Software for Continuation and Bifurcation in Ordinary Differential Equations*, California Institute of Technology, Pasadena, CA, 1994.

[15] E. Doedel, *Numerical Analysis of Bifurcation Problems*, Technical report, University of Hamburg, Hamburg, Germany, 1997.

[16] M. Golubitsky and D. G. Schaeffer, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[17] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. 1, Appl. Math. Sci. 51, Springer-Verlag, Berlin, Heidelberg, New York, 1985.

[18] W. Govaerts, *Computation of singularities in large nonlinear systems*, SIAM J. Numer. Anal., 34 (1997), pp. 867–880.

[19] W. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equlibria*, SIAM, Philadelphia, 2000.

[20] A. Griewank and G. W. Reddien, *Characterization and computation of generalized turning points*, SIAM J. Numer. Anal., 21 (1984), pp. 176–185.

[21] V. Janovský and P. Plecháč, *Computer-aided analysis of imperfect bifurcation diagrams,* I. *Simple bifurcation point and isola formation centre*, SIAM J. Numer. Anal., 29 (1992), pp. 498–512.

[22] V. Janovský and V. Seige, *A global analysis of Newton iterations for determining turning points*, Appl. Math., 38 (1993), pp. 323–360.

[23] A. D. Jepson and A. Spence, *The numerical solution of nonlinear equations having several parameters.* I: *Scalar equations*, SIAM J. Math. Anal., 22 (1985), pp. 736–759.

[24] A. D. Jepson and A. Spence, *On a reduction process for nonlinear equations*, SIAM J. Math. Anal., 20 (1989), pp. 39–56.

[25] H. B. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, London, 1977, pp. 359–384.

[26] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 1987.

[27] Y. A. Kuznetsov and V. V. Levitin, *CONTENT: A Multiplatform Environment for Analyzing Dynamical Systems*, Technical report, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands, 1996.

[28] H. G. Kwatny and B. C. Chang, *Constructing linear families from parameter-dependent nonlinear dynamics*, IEEE Trans. Automat. Control, 43 (1998), pp. 1143–1147.

[29] W. Marquardt, M. Friedrich, and E. D. Gilles, *Analyse und Synthese mehrphasiger Reaktionssysteme mit Methoden der Nichtlinearen Dynamik,* Technical report, R.W.-Techn. Hochschule, Aachen, Germany, 1999.

[30] Z. Mei and A. Schwarzer, *Scaling solution branches of bifurcation problems*, J. Math. Anal. Appl., 204 (1996), pp. 404–433.

[31] Z. Mei, *Numerical Bifurcation Analysis for Reaction-Diffusion Equations*, Springer Ser. Comput. Math. 28, Springer-Verlag, Berlin, Germany, 2000.

[32] N. Sassmannshausen and K. Böhmer, *Petrov-Galerkin Methods for Projected Linear Operator Equation, Stability and Convergence*, preprint, DFG-Schwerpunktprogramm Danse, 1998.

[33] U. Schnabel, G. Pönisch, and H. Schwetlick, *Computing Multiple Pitchfork Bifurcation Points*, Technical report, Technical University of Dresden, Dresden, Germany, 1995.

[34] A. Schwarzer, *Skalierungstechniken für k-bestimmte Verzweigungsprobleme und Iterationsmethoden für grosse, dünn besetzte Eigenwertprobleme*, Ph.D. thesis, Philipps-Universität Marburg, Marburg, Germany, 1997.

[35] R. Sebastian, *Anwendung von Krylov-Verfahren auf Verzweigungs- und Fortsetzungsprobleme*, Dissertation am Fachbereich Mathematik, Universität Marburg, Marburg, Germany, 1995.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 431–450

VARIANCE REDUCTION METHODS FOR SIMULATION OF DENSITIES ON WIENER SPACE*

ARTURO KOHATSU-HIGA† AND ROGER PETTERSSON‡

Abstract. We develop a general error analysis framework for the Monte Carlo simulation of densities for functionals in Wiener space. We also study variance reduction methods with the help of Malliavin derivatives. For this, we give some general heuristic principles which are applied to diffusion processes. A comparison with kernel density estimates is made.

Key words. stochastic differential equations, weak approximation, variance reduction, kernel density estimation

AMS subject classifications. 60H10, 65C05, 65C30, 68U20, 60H07, 60H35

PII. S0036142901385507

1. Introduction.

The Monte Carlo simulation method is used to estimate quantities of the type $E[f(X)]$, where f is a somewhat regular function and X is a random variable that can be simulated.

In this article, we are interested in the case when f is a generalized function such as the Dirac delta function δ_x or a discontinuous function such as an indicator function. In the first case the expectation will become the density of the random variable X and in the second the distribution function. If f is not regular, then the Monte Carlo method has to be slightly modified using $\frac{1}{n}\sum_{i=1}^{n} f_n(X^i)$, where f_n is a smooth function that approximates f and X^i are independent copies of X. This approximation converges to the desired quantity but a big error is produced due to the nonsmoothness of the general function f. In this framework it becomes important to devise methods for reducing the variance of the Monte Carlo estimation. This problem has been extensively studied by statisticians (although in a slightly different situation) in the theory of kernel density estimation; see, e.g., [12].

Here we propose to analyze the above problem using Malliavin calculus for Wiener space. More explicitly, using the integration by parts formula of Malliavin calculus one has that $E[f(X)] = E[F(X)H(X,1)]$, where $H(X,1)$ is an appropriate random variable and F is an antiderivative of f. In this way we gain smoothness in the function to be evaluated but the simulation of $H(X,1)$ is now required. The above formula can be explained as the integration by parts of $\int_{\mathbb{R}} f(x)p(x)dx = -\int_{\mathbb{R}} F(x)p'(x)dx$, where p is the density of X, i.e., $H(X,1) = -p'(X)/p(X)$. This looks simple as long as one knows the density of X. Here we deal with cases where p is not known explicitly. Still, we show that there are ways to simulate $H(X,1)$ and that some variance reduction is in fact achieved.

The typical example that we treat here is when X is the final value of a diffusion.

*Received by the editors February 22, 2001; accepted for publication (in revised form) December 27, 2001; published electronically May 29, 2002.
http://www.siam.org/journals/sinum/40-2/38550.html

†Departament d'Economia, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005 Barcelona, Spain (arturo.kohatsu@econ.upf.es). The research of this author was partially supported by grants PB98-1059, BFM 2000-807, and BFM 2000-0598 of the Ministerio de Ciencia y Tecnologia.

‡Matematiska och Systemtekniska Institutionen, Växjö Universitet, Vejdes Plats 7, S-351 95 Växjö, Sweden (rpe@msi.vxu.se). The research of this author was partially supported by EU grant ERBF MRX CT96 0075A.

That is, $X = X_1$ where

$$X_t = x_0 + \int_0^t b(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}W_s, \qquad t \in [0,1]. \tag{1.1}$$

Here $x_0 \in \mathbb{R}$ and b and σ are smooth functions. If the Hörmander hypothesis is satisfied, then the density of X_1 exists and is smooth. In [1] and [2] the approximation error for the density is studied when the random variable X_1 is replaced by the Euler–Maruyama approximation $\bar{X}$.

Obviously, the density of X is explicitly known only in particular cases and therefore the simulation of $H(X,1)$ is not a trivial matter. This is exactly the merit of Malliavin calculus. One can use this technique to develop an expression for $H(X,1)$ that can be simulated. In order to simulate $E[f(X)]$, our Monte Carlo method with variance reduction is to calculate $\frac{1}{n}\sum_{i=1}^n F(\bar{X}^i)H(\bar{X}^i,1)$, where $\bar{X}^i$ are independent Euler approximations of X. We concentrate on the particular case when f is the delta function which therefore generates the density of the diffusion process, but this methodology can be applied also when approximating the price of an option or its Greeks in mathematical finance. In fact, this idea appeared first in [5] applied to the calculation of Greeks called delta, vega, and gamma. Also in [6] a more careful study of the simulation of the density is carried out. An optimal variance reduction method is devised but it requires the knowledge of the density itself and is therefore not amenable to direct application.

In this article, we introduce a control variate method and a tuning method, similar to the ones used in kernel density estimation, that helps to reduce the variance substantially. The main difference with respect to kernel density estimation methods is that our tuning does not require that the window size goes to 0 as the sample size increases. Furthermore, the same simulated paths give good estimates for densities at any point. That is, one can compute the density over the whole real line with the same number of simulated paths.

We focus in the one-dimensional case just to avoid cumbersome notation. The results are also valid in multidimensions with appropriate modifications. The importance of these methods is obvious when the dimension is relatively large. See also [9] on variance reduction of smooth functions of diffusions, where methods of importance sampling and control variates are developed without the use of the integration by parts formula.

In section 2 after some preliminaries on Malliavin calculus we explain the general method and give a control variate method for variance reduction. In section 3 we estimate the error of the approximating expectations. The error is estimated when there is an Itô–Taylor expansion for the functional in the spirit of [8]. In section 4 we consider, as an application of section 3, the case of diffusion processes with a Hörmander condition. We also define the different approximations and give bounds on the approximation error. In section 5 we study the mean square error of the kernel density method. In section 6 a similar study for the integration by parts method is made and a comparison is made. In section 7 numerical implementations are described.

Throughout, let c denote a generic constant which may differ from line to line.

2. Malliavin derivative and density by duality. Let $W = \{W_t\}_{t\in[0,1]}$ be a standard one-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$. Assume $\mathcal{F} = \{\mathcal{F}_t\}_{t\in[0,1]}$ is generated by W. Let $\mathcal{S}$ be the space of random variables of the form $F = f(W_{t_1}, \dots, W_{t_n})$, where f is smooth. For $F \in \mathcal{S}$, $D_tF =$

$\sum_{i=1}^n \frac{\partial}{\partial x_i} f(W_{t_1}, \dots, W_{t_n}) 1_{[0,t_i]}(t)$. For $k \in \mathbb{Z}_+$, $p \geq 1$, let $\mathbb{D}^{k,p}$ be the completion of $\mathcal{S}$ with respect to the norm

$$\|F\|_{k,p} = \left(E[|F|^p] + E\left[\left(\sum_{j=1}^k \int_0^1 \dots \int_0^1 |D^j_{s_1,\dots,s_j} F|^2 ds_1 \dots ds_j \right)^{p/2} \right] \right)^{1/p},$$

where $D^j_{t_1,\dots,t_j} F = D_{t_1} \dots D_{t_j} F$. We let $\|F\|_{0,p} = (E[F^p])^{1/p} = \|F\|_p$ and $\mathbb{D}^\infty = \cap_{k,p} \mathbb{D}^{k,p}$. For processes $u = \{u_t\}_{t\in[0,1]}$ on $(\Omega, \mathcal{F}, P)$, $\mathbb{D}^{k,p}_{L^2([0,1])}$ is defined as $\mathbb{D}^{k,p}$ but with norm $\|u\|_{k,p,L^2([0,1])} = (E[\|u\|^p_{L^2([0,1])}] + E[(\sum_{j=1}^k \int_0^1 \dots \int_0^1 \|D^j_{s_1,\dots,s_j} u\|^2_{L^2([0,1])} ds_1 \dots ds_j)^{p/2}])^{1/p}$. For two-parameter processes $u = \{u_{s,t}\}_{s,t\in[0,1]}$, $\mathbb{D}^{k,p}_{L^2([0,1]^2)}$ is defined analogously. $\mathbb{D}^\infty_{L^2([0,1])}$ and $\mathbb{D}^\infty_{L^2([0,1]^2)}$ are defined similarly to $\mathbb{D}^\infty$.

We denote by $\delta(u)$ the Skorokhod integral, the dual operator of D. If u_t is $\mathcal{F}_t$ adapted, then $\delta(u) = \int_0^1 u_t dW_t$, the Itô integral of u; see, e.g., [11]. Here we write $\delta(u) = \int_0^1 u_t dW_t$, even if u_t is not $\mathcal{F}_t$ adapted. This integral satisfies that

$$\int_0^1 F u_t dW_t = F \int_0^1 u_t dW_t - \int_0^1 (D_t F) u_t dt \tag{2.1}$$

for $F \in \mathbb{D}^{1,2}$ and $E(F^2 \int_0^1 u_t^2 dt) < \infty$ (see, e.g., Nualart [11, (1.49), p. 40]) and

$$E\left[\int_0^1 (D_t F) u_t dt \right] = E[F\delta(u)]. \tag{2.2}$$

For F, G in $\mathbb{D}^{1,2}$ and h a stochastic process such that $E\int_0^1 h_t^2 dt < \infty$, we use the notation

$$H^h(F, G) = \int_0^1 \tilde{h}_t G dW_t, \tag{2.3}$$

where $\tilde{h}_t = h_t / \int_0^1 h_s D_s F ds$, whenever the integrand in (2.3) is Skorokhod integrable. The usefulness of $H^h(F, G)$ can be seen in the integration by parts formula of Malliavin calculus which can be expressed as

$$E[f'(F)G] = E[f(F)H^h(F, G)]. \tag{2.4}$$

We let $H(F, G) \equiv H^{DF}(F, G)$ (i.e., $h_t = D_t F$ in (2.3)). Arguments similar to those in [11, pp. 78, 97] give the following theorem.

THEOREM 2.1. *Assume $F \in \mathbb{D}^{1,2}$ and $E\int_0^1 h_t^2 dt < \infty$. Let φ be a function on $\mathbb{R}$ such that φ, $\frac{d}{dx}\varphi \in L^2(\mathbb{R})$, $\varphi(0) = 1$, and $c \in L^2(\mathbb{R})$. Let r be a positive number and assume that $\tilde{h}\varphi((F - x)/r)$ is Skorokhod integrable. Then the density of F, f exists, is continuous, and f has the representation $f(x) = E[\xi_{c,r}(x)]$, where $\xi_{c,r}(x) = (1_{\{F>x\}} - c(x))H^h(F, \varphi(\frac{F-x}{r}))$. Furthermore,*

$$H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right) = \varphi\left(\frac{F-x}{r}\right) H^h(F, 1) - \frac{1}{r}\varphi'\left(\frac{F-x}{r}\right). \tag{2.5}$$

Proof of Theorem 2.1. We first observe that taking $F = 1$ and $u_t = \varphi((F-x)/r)\tilde{h}_t$ in (2.2) yields

$$E\left[H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right)\right] = 0. \tag{2.6}$$

We now show the existence of the density of F. For this we assume that F and u are sufficiently smooth as well as φ. The general argument follows by a density argument. Taking $a < b$ and using (2.1), (2.2), and (2.6) we have that

$$\begin{aligned}
&\int_a^b E\left[(1_{\{F>x\}} - c(x))H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right)\right] dx \\
&= \int_a^b E\left[1_{\{F>x\}} \int_0^1 \tilde{h}_t \varphi\left(\frac{F-x}{r}\right) dW_t\right] dx \\
&= \int_a^b E\left[1_{\{F>x\}} \left(\varphi\left(\frac{F-x}{r}\right)\int_0^1 \tilde{h}_t dW_t - \int_0^1 \tilde{h}_t D_t \varphi\left(\frac{F-x}{r}\right) dt\right)\right] dx \\
&= E\left[\int_{-\infty}^F 1_{[a,b]}(x)\varphi\left(\frac{F-x}{r}\right) dx \int_0^1 \tilde{h}_t dW_t\right] \\
&\quad - \int_a^b E\left[\int_0^1 1_{\{F>x\}} \tilde{h}_t D_t \varphi\left(\frac{F-x}{r}\right) dt\right] dx \\
&= E\left[\int_0^1 D_t\left(\int_{-\infty}^F 1_{[a,b]}(x)\varphi\left(\frac{F-x}{r}\right)\right) dx \tilde{h}_t dt\right] \\
&\quad - \int_a^b E\left[\int_0^1 1_{\{F>x\}} \tilde{h}_t D_t \varphi\left(\frac{F-x}{r}\right) dt\right] dx \\
&= E\left[\int_0^1 1_{[a,b]}(F)\varphi(0)\tilde{h}_t D_t F dt\right] = E[1_{[a,b]}(F)],
\end{aligned}$$

which shows the absolute continuity of the law of F with respect to the Lebesgue measure. The right continuity of the density of F, f follows from the fact that $\frac{d}{dx}\varphi \in L^2([0,1])$ and the right continuity of the indicator function $1_{\{\cdot>x\}}$. Now f can also be written with $1_{\{F>x\}}$ replaced by $1_{\{F\geq x\}}$ from which the left continuity follows. Finally, taking $F = \varphi((F-x)/r)$ and $u_t = \tilde{h}_t$ in (2.1) the claim (2.5) follows. That is,

$$\begin{aligned}
&H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right) \\
&= \int_0^1 \varphi\left(\frac{F-x}{r}\right)\tilde{h}_t dW_t \\
&= \varphi\left(\frac{F-x}{r}\right)\int_0^t \tilde{h}_t dW_t - \int_0^1 D_t\left(\varphi\left(\frac{F-x}{r}\right)\right)\tilde{h}_t dt. \qquad \square
\end{aligned}$$

Taking $F = \varphi((F-x)/r)/\int_0^1 h_s D_s F ds$ and $u_t = h_t$ in (2.1) and assuming enough smoothness one obtains that

$$H^h(F,1) = \frac{\int_0^1 h_t dW_t}{\int_0^1 h_s D_s F ds} + \frac{\int_0^1\int_0^1 D_t(h_s D_s F)h_t ds dt}{(\int_0^1 h_s D_s F ds)^2}. \tag{2.7}$$

Note that the variance of $\xi_{c,r}(x)$ is finite under reasonable assumptions, while $\mathrm{var}(\delta_x(f)) = \infty$. The issue of the variance of $\xi_{c,r}(x)$ will be further discussed in section 5. The representation of the density f introduced in Theorem 2.1 has the additional benefit that it allows us to develop a control variate method for the reduction of variance; see also [5], [6], and [7] for related results.

REMARK 2.2 (control variate method). *Assume the same hypotheses as in Theorem* 2.1. *If* $f(x) > 0$, *then* $E[H^h(F, \varphi(\frac{F-x}{r}))^2] > 0$ *and, for fixed* φ *and* r, *the variance of* $\xi_{c,r}(x) = (1_{\{F \geq x\}} - c(x))H^h(F, \varphi(\frac{F-x}{r}))$ *is minimized by*

(2.8)
$$c(x) = c^h_{loc}(x) = E\left[1_{\{F\geq x\}} H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right)^2\right] \Big/ E\left[H^h\left(F, \varphi\left(\frac{F-x}{r}\right)\right)^2\right]. \qquad \square$$

The case of $f(x) = 0$ can also be dealt with using some extra changes. To simplify our discussion we will focus on the case when $f(x) > 0$.

3. Convergence of approximative functionals. In this section we present a general theory of approximation for random variables F on Wiener space that gives as a result rates of convergence to the density of F. This theory is based on Itô–Taylor expansions in the spirit of [8]. Later we consider as an application the case when F is the terminal value of the solution of a stochastic differential equation. Other examples that satisfy the following conditions will be treated in forthcoming publications. To simplify we use the notation $dW^1_s = dW_s$ and $dW^0_s = ds$.

CONDITION 3.1. (i) $\{F_n\}_{n\geq 0}$ *and* F *are in* $\mathbb{D}^\infty$ *and satisfy*

$$F - F_n = \sum_{i,j=0}^{1} \iint_{A_n^{i,j}} u_n^{i,j}(s_1, s_2) dW^i_{s_1} dW^j_{s_2},$$

where $A_n^{i,j}$ *are subsets of* $[0,1]^2$ *with mean area* $\sum_{i,j=0}^1 |A_n^{i,j}|/4 \leq a_n$ *for a sequence* $a_n \to 0$ *as* $n \to \infty$, *and* $\sup_n \sup_{s_1,s_2} \sum_{i,j=0}^1 \|u_n^{i,j}(s_1,s_2)\|_{k,p} < \infty$ *for all* $k \in \mathbb{Z}_+$, $p > 1$. *The processes* $u_n^{i,j}$ *are measurable, not necessarily adapted but with enough properties so that the above integrals are well defined.*

(ii) *For* $\alpha \in [0,1]$ *there exist processes* $h_n \equiv h_{n,\alpha}$ *in* $\mathbb{D}^\infty(L^2[0,1])$, *uniformly bounded in* n, *and a process* $h \in \mathbb{D}^\infty(L^2([0,1])$, *such that* $E[|\int_0^1 h(s)D_sFds|^{-p}] < \infty$ *for all* $p > 1$ *and* $E\|h_n - h\|^p_{L^2[0,1]} \leq e_n(p)$ *for a sequence* $e_n(p) \equiv e_{n,\alpha}(p) \to 0$, $n \to \infty$.

(iii) *For* $\alpha \in [0,1]$ *there exist positive random variables* d_n *and positive bounded constants* b_n *and* c *such that* $|b_n + \int_0^1 h(s)D_sFds| > c|\int_0^1 h(s)D_sFds|$ *and*

$$\left|b_n + \int_0^1 h_n(s)(\alpha D_sF_n + (1-\alpha)D_sF)ds\right| \geq d_n,$$

where for any $p > 1$ *there exists* $k(p) \equiv k(p,\alpha) \in \mathbb{Z}_+$ *such that* $\sup_n \sup_{\alpha\in[0,1]} E[d_n^{-p}] \times (a_n^{k(p)/2} + e_n(4k(p))^{1/2} + b_n^{2k(p)}) < \infty$.

Without loss of generality we assume that all the sequences a_n, b_n, e_n, and d_n are smaller than 1. Next we give the main approximation result in this section.

THEOREM 3.2. *Assume Condition* 3.1. *Then for any distribution* T,

$$|E[T(F) - T(F_n + Y_n)]| \leq c(a_n + b_n), \tag{3.1}$$

where Y_n *is an independent normal random variable with mean zero and variance* b_n. *Furthermore, if*

$$\sup_{n\geq 1}\sup_{0<\alpha<1} E\left[\left|\int_0^1 h_n(s)(\alpha D_s F_n + (1-\alpha)D_s F)ds\right|^{-p}\right] < \infty \tag{3.2}$$

for all $p > 1$, *then*

$$|E[T(F) - T(F_n)]| \leq ca_n. \tag{3.3}$$

We say that the approximation problem is *uniformly elliptic* when (3.2) is satisfied. If we instead only assume Condition 3.1, we will say that the approximation problem is of *Hörmander* type. See section 4 for more explanation about this terminology.

The above theorem will be usually applied to $T(y) = 1_{\{y\geq x\}}$ or $T(y) = \delta_x^{(k)}(y)$, the kth derivative of the Dirac delta measure. We will do the proof in the second case for $x = 0$, $k = 0$. The general case is proved similarly. The application of Theorem 3.2 to diffusion processes and its Euler approximation will be given in section 4. At the end of this section we also give a generalization of Theorem 3.2, where $F - F_n$ may be expressed as a sum of higher order stochastic multiple integrals.

We start with some technical results.

LEMMA 3.3. *Assume Condition* 3.1(i). *Then* $E\|D(F_n - F)\|^p_{L^2[0,1]} \leq ca_n^{p/4}$ *for any* $p > 4$.

The above rate is not optimal in most cases. But for our purposes it will suffice as a rate of convergence.

LEMMA 3.4. *Assume Condition* 3.1. *Then*

$$\sup_{n\geq 1}\sup_{0<\alpha<1} E\left[\left|b_n + \int_0^1 h_n(s)(\alpha D_s F_n + (1-\alpha)D_s F)ds\right|^{-p}\right] < \infty \quad \text{for all } p > 1.$$

Proof of Lemma 3.3. We consider one of the terms in Condition 3.1(i) ($i = 1$, $j = 1$). By Proposition 1.4.5 of [11, p. 69], which is a consequence of Meyer's inequality,

$$\begin{aligned}
&E\left\|D\iint_{A_n^{1,1}} u_n^{1,1}(s_1,s_2)dW_{s_1}^1 dW_{s_2}^1\right\|^p_{L^2([0,1])}\\
&\leq c_1\left\|\iint_{A_n^{1,1}} u_n^{1,1}(s_1,s_2)dW_{s_1}^1 dW_{s_2}^1\right\|^p_{1,p}\\
&\leq c_2\|1_{A_n^{1,1}}u_n^{1,1}\|^p_{3,p}\\
&= c_3\Bigg(E[\|1_{A_n^{1,1}}u_n^{1,1}\|^p_{L^2([0,1]^2)}]\\
&\qquad + E\left[\left(\sum_{j=1}^3\int_0^1\cdots\int_0^1 \|1_{A_n^{1,1}}D^j_{s_1,\ldots,s_j}u_n^{1,1}\|^2_{L^2([0,1]^2)}ds_1\ldots ds_j\right)^{p/2}\right]\Bigg)
\end{aligned}$$

$$\le E[\|1_{A_n^{1,1}} u_n^{1,1}\|^p_{L^2([0,1]^2)}]$$
$$+ E\left[\left(\|1_{A_n^{1,1}}\|^{1/2}_{L^2([0,1]^2)}\right.\right.$$
$$\left.\left.\times\left(\iint_{[0,1]^2}\left(\sum_{j=1}^{3}\int_0^1\cdots\int_0^1 (D^j_{s_1,\dots,s_j}u_n^{1,1}(t_1,t_2))^2 ds_1\dots ds_j\right)^2 dt_1 dt_2\right)^{1/2}\right)^{p/2}\right]$$
$$\le a_n^{p/4}\sup_{(t_1,t_2)\in[0,1]^2}\|u_n^{1,1}(t_1,t_2)\|^p_{3,p} \le c a_n^{p/4}. \qquad \square$$

Proof of Lemma 3.4. Define the set

$$A \equiv \left\{ \left|\int_0^1 (h_n(s)D_sF_n - h(s)D_sF)ds\right| \vee \left|\int_0^1 (h_n(s)-h(s))D_sF ds\right| \right.$$
$$\left. \vee\, b_n < \frac{1}{4}\left|\int_0^1 h(s)D_sF ds\right| \right\}.$$

On A, $|\int_0^1 h_n(s)(\alpha D_sF_n + (1-\alpha)D_sF)ds| \ge \frac{1}{2}|\int_0^1 h(s)D_sF ds|$; hence

$$E\left[\left|b_n + \int_0^1 h_n(s)(\alpha D_sF_n + (1-\alpha)D_sF)ds\right|^{-p}; A\right] \le 4^p E\left[\left|\int_0^1 h(s)D_sF ds\right|^{-p}\right].$$

By Chebyshev's inequality and Condition 3.1(iii), $E[\|b_n + \int_0^1 h_n(s)(\alpha D_sF_n + (1-\alpha)D_sF)ds\|^{-p}; A^c] \le c_p(E[d_{n,\alpha}^{-2p}]P(A^c))^{1/2}$. For any $k \in \mathbb{Z}_+$ so that $kp > 1$, we have by Condition 3.1(ii) and Lemma 3.3 that $P(A^c)$ is less than or equal to

$$4^{2kp}c_{k,p}E\left[\left|\int_0^1 h(s)D_sF ds\right|^{-2kp}\left(\|h_nDF_n - hDF\|^{2kp}_{L^1[0,1]}\right.\right.$$
$$\left.\left. + \|(h_n-h)DF\|^{2kp}_{L^1[0,1]} + b_n^{2kp}\right)\right]$$
$$\le c_{k,p}\left(E\left|\int_0^1 h(s)D_sF ds\right|^{-8kp}\right)^{1/4}\left\{\left(E\|D(F_n-F)\|^{4kp}_{L^2[0,1]}\right)^{1/2}\left(E\|h_n\|^{8kp}_{L^2[0,1]}\right)^{1/4}\right.$$
$$\left. + \left(E\|h_n-h\|^{4kp}_{L^2[0,1]}\right)^{1/2}\left(E\|DF\|^{8kp}_{L^2[0,1]}\right)^{1/4}\right\}$$
$$+ c_{k,p}E\left(\left|\int_0^1 h(s)D_sF ds\right|^{-2kp}\right)b_n^{2kp}$$
$$\le c_{k,p}\left(a_n^{kp/2} + e_n(4kp)^{1/2} + b_n^{2kp}\right).$$

The result follows by Condition 3.1(ii) and (iii). $\square$

Recall (2.3) and inductively define $H^{(n)}$ by $H^{(n)}(F,G) = H^h(F, H^{(n-1)}(F,G))$ and $H^{(0)}(F,G) = G$. We then have that for any $m \in \mathbb{Z}_+$ and $p > 1$,

$$\|H^{(m)}(F,G)\|_p$$
$$\le c\|G\|_{m+1,p_0}(\|F\|^{\alpha_1}_{m+1,p_1} + \|h\|^{\alpha_2}_{m,p_2})\left\|\left(\int_0^1 h(s)D_sF ds\right)^{-1}\right\|^{\alpha_3}_{p_3} \tag{3.4}$$

for a constant c and indices $\alpha_1,\ \alpha_2\ ,\ \alpha_3,\ p_0,\dots,p_3$ depending on m, p; see, e.g., [10, Proposition 3.3.2]. We consider the extended Wiener space $\mathcal{W}$ generated by $(W,\tilde{W})$, where $\tilde{W}$ is a Brownian motion independent of W. Let $Y_n=\sqrt{b_n}\tilde{W}_1$. For $G\in\mathbb{D}^{1,2}(\mathcal{W})$ (which has as a norm the natural extension for the product space $\mathbb{D}^{1,2}(W)\times\mathbb{D}^{1,2}(\tilde{W})$-norm) we deduce using (2.4) that

$$E[g'(Y_n+F)G]=E[g(Y_n+F)\bar{H}(F,G)], \tag{3.5}$$

where

$$\bar{H}(F,G)\equiv\bar{H}^h(F,G)=\int_0^1\frac{Gh(t)}{\int_0^1 h(s)D_sFds+b_n}dW_t+\int_0^1\frac{G\sqrt{b_n}}{\int_0^1 h(s)D_sFds+b_n}d\tilde{W}_t.$$

Similarly, define by induction $\bar{H}^{(k)}$ by $\bar{H}^{(k)}(F,G)=\bar{H}(F,\bar{H}^{(k-1)}(F,G))$, where $\bar{H}^{(0)}(F,G)=G$. Also if instead of h we use h_n in the definition of $\bar{H}$, then we use the notation $\bar{H}_n$. Using similar arguments as to those of the proof of (3.4) the following result is deduced.

LEMMA 3.5. *Assume $F\in\mathbb{D}^\infty(W)$ and $G\in\mathbb{D}^\infty(\mathcal{W})$. Then for $m\in\mathbb{Z}_+$ and $p>1$,*

$$\begin{aligned}\|\bar{H}^{(m)}(F,G)\|_p\le c\|G\|_{m+1,p_0}(\|F\|_{m+1,p_1}^{\alpha_1}+\|h\|_{m,p_2}^{\alpha_2}+b_n^{\alpha_3})\\ \times\left\|\left(\int_0^1 h(s)D_sFds+b_n\right)^{-1}\right\|_{p_4}^{\alpha_4}\end{aligned}$$

for a constant c and indices $p_0,p_1,p_2,p_4,\ \alpha_1,\dots,\alpha_4$, depending on m and p.

Proof of Lemma 3.5. We use induction. For $\bar{H}_k\equiv\bar{H}^{(k)}(F,G)$,

$$\bar{H}_k=\int_0^1\frac{\bar{H}_{k-1}h(t)}{\int_0^1 h(s)D_sFds+b_n}dW_t+\int_0^1\frac{\bar{H}_{k-1}\sqrt{b_n}}{\int_0^1 h(s)D_sFds+b_n}d\tilde{W}_t,$$

where $\bar{H}_0=G$. Applying Meyer's inequality (see, e.g., [11, p. 69]) and Hölder's inequality, for $k=0,\dots,m$, we have

$$\begin{aligned}&\|\bar{H}_k\|_{m-k,p}\\ &\le c\|\bar{H}_{k-1}\|_{m-k+1,\alpha p}(\|h\|_{m-k+1,\beta p}+\sqrt{b_n})\left\|\left(\int_0^1 h(s)D_sFds+b_n\right)^{-1}\right\|_{m-k+1,\gamma p},\end{aligned}$$

where $\alpha^{-1}+\beta^{-1}+\gamma^{-1}=1$. Furthermore, using the Cauchy–Schwarz inequality in the calculation of terms of the form $\int_{[0,1]^i}(D_{t_1,\dots,t_i}(\int_0^1 h(s)D_sFds+b_n)^{-1})^2dt_1\dots dt_i$ gives

$$\begin{aligned}&\left\|\left(\int_0^1 h(s)D_sFds+b_n\right)^{-1}\right\|_{m+1,\gamma p}\\ &\le c(\|F\|_{m+1,q_1}^{\beta_1}+\|h\|_{m,q_2}^{\beta_2}+b_n^{\beta_3})\left\|\left(\int_0^1 h(s)D_sFds+b_n\right)^{-1}\right\|_{q_4}^{\beta_4}\end{aligned}$$

for some indices $q_1,q_2,q_4,\beta_1,\dots,\beta_4$. The result follows by induction. □

Proof of Theorem 3.2. Let $f_n(x) = \phi_{\sqrt{b_n}}(x) = \exp(-x^2/2b_n)/\sqrt{2\pi b_n}$ and $G_n = \int_0^1 f_n'(Y_n+\alpha F_n+(1-\alpha)F)d\alpha$. Using the mean value theorem and the duality between the Skorokhod integral and the derivative operator (see, e.g., [11, equation (1.41), p. 35]) yields

(3.6)

$$
\begin{aligned}
&E[f_n(F+Y_n)-f_n(F_n+Y_n)]\\
&= E[G_n(F-F_n)] = \sum_{i,j=0}^{1} E\left[G_n \iint_{A_n^{i,j}} u_n^{i,j}(s_1,s_2)dW_{s_1}^i dW_{s_2}^j\right]\\
&= \iint_{A_n^{0,0}} E[G_n u_n^{0,0}(s_1,s_2)]ds_1ds_2 + \iint_{A_n^{0,1}} E[D_{s_2}(G_n)u_n^{0,1}(s_1,s_2)]ds_1ds_2\\
&\quad + \iint_{A_n^{1,0}} E[D_{s_1}(G_n)u_n^{1,0}(s_1,s_2)]ds_1ds_2 + \iint_{A_n^{1,1}} E[D^2_{s_1,s_2}(G_n)u_n^{1,1}(s_1,s_2)]ds_1ds_2.
\end{aligned}
$$

We will compute one of these terms as they are all similar. Using (3.5) three times, we get that $|E[D_{s_2}G_n u_n^{0,1}(s_1,s_2)]|$ equals

$$
\begin{aligned}
&\left|\int_0^1 E[f_n''(Y_n+\alpha F_n+(1-\alpha)F)(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F)u_n^{0,1}(s_1,s_2)]d\alpha\right|\\
&= \left|\int_0^1 E[\Phi_n(Y_n+\alpha F_n+(1-\alpha)F)\bar{H}^{(3)}(\alpha F_n+(1-\alpha)F,\right.\\
&\qquad\qquad \left.(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F)u_n^{0,1}(s_1,s_2))]d\alpha\right|\\
&\le \int_0^1 E|\bar{H}_n^{(3)}(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F,(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F)u_n^{0,1}(s_1,s_2))|d\alpha,
\end{aligned}
$$

where Φ_n is the distribution function associated with f_n. By Lemma 3.5,

$$
\begin{aligned}
&\sup_n \sup_{s_1,s_2} \sup_{0<\alpha<1} E|\bar{H}_n^{(3)}(\alpha F_n+(1-\alpha)F,(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F)u_n^{0,1}(s_1,s_2))|\\
&\le c\sup_n \sup_{s_1,s_2} \sup_{0<\alpha<1} \|(\alpha D_{s_2}F_n+(1-\alpha)D_{s_2}F)u_n^{0,1}(s_1,s_2)\|_{4,p_0}\\
&\quad\times(\|\alpha F_n+(1-\alpha)F\|_{4,p_1}^{\alpha_1}+\|h_n\|_{3,p_2}^{\alpha_2}+b_n^{\alpha_3})\\
&\quad\times\left\|\left(b_n+\int_0^1 h_n(s)(\alpha D_sF_n+(1-\alpha)D_sF)ds\right)^{-1}\right\|_{p_4}^{\alpha_4},
\end{aligned}
$$

which is finite by Condition 3.1(i), (ii) and Lemma 3.4. Similar considerations lead to the conclusion that the other terms in (3.6) have a similar bound. In conclusion one has that $|E[f_n(F+Y_n)-f_n(F_n+Y_n)]| \le ca_n$.

Now consider $E[\delta_0(F)-f_n(F+Y_n)]$. Observe that

$$
\begin{aligned}
E[f_n(F+Y_n)] &= \int_{\mathbb{R}} E[f_n(F+y)]f_n(y)dy = E[\phi_{\sqrt{2b_n}}(F)]\\
&= E\int_{\mathbb{R}} \delta_0(F+z)\phi_{\sqrt{2b_n}}(z)dz = E\delta_0(F+\sqrt{2}Y_n).
\end{aligned}
$$

By Condition 3.1(ii) and Theorem 2.1, the densities of F and $F+\sqrt{2}Y_n$ are continuous and hence $E[\delta_0(F)-\delta_0(F+\sqrt{2}Y_n)] = \lim_{m\to\infty} E[f_m(F)-f_m(F+\sqrt{2}Y_n)]$. A Taylor

expansion of f_m around F yields

$$E[f_m(F) - f_m(F + \sqrt{2}Y_n)] = \sqrt{2}E[f_m'(F)Y_n] + 2E\left[\int_0^1 f_m''(F + \alpha\sqrt{2}Y_n)(1-\alpha)Y_n^2 d\alpha\right].$$

Clearly, $E[f_m'(F)Y_n] = 0$ by the independence of F and Y_n. By (2.4), (3.4), and Condition 3.1(ii),

$$\begin{aligned}|E[f_m''(F + \alpha\sqrt{2}Y_n)Y_n^2]| &= \left|\int E[f_m''(F + \alpha\sqrt{2}y)]y^2 f_n(y)dy\right| \\ &= \left|\int_{\mathbb{R}} E[\Phi_m(F + \alpha\sqrt{2}y)H^3(F,1)]y^2 f_n(y)dy\right| \\ &\le c\big(\|F\|_{4,p_1}^{\alpha_1} + \|h\|_{3,p_2}^{\alpha_2}\big)\left\|\left(\int_0^1 h(s)D_sF ds\right)^{-1}\right\|_{p_3}^{\alpha_3} E[Y_n^2] \le cb_n.\end{aligned}$$

Hence, $|E[\delta_0(F) - f_n(F + Y_n)]| \le cb_n$. Similarly, for an independent copy $\bar{Y}_n$ of Y_n, we obtain by Lemma 3.5

$$\begin{aligned}&|E[\delta_0(F_n + Y_n) - f_n(F_n + Y_n)]| \\ &= |E[\delta_0(F_n + Y_n) - \delta_0(F_n + Y_n + \sqrt{2}\bar{Y}_n)]| \\ &= 2\lim_{m\to\infty}\left|\int_0^1\int_{\mathbb{R}} E[f_m''(\alpha\sqrt{2}y + F_n + Y_n)]y^2 f_n(y)dy(1-\alpha)d\alpha\right| \\ &= 2\lim_{m\to\infty}\left|\int_0^1\int_{\mathbb{R}} E[\Phi_m(\alpha\sqrt{2}y + F_n + Y_n)\bar{H}_n^{(3)}(F_n,1)]y^2 f_n(y)dy(1-\alpha)d\alpha\right| \\ &\le c(\|F_n\|_{4,p_1}^{\alpha_1} + \|h_n\|_{3,p_2}^{\alpha_2} + b_n^{\alpha_3})\left\|\left(\int_0^1 h_n(s)D_sF_n ds + b_n\right)^{-1}\right\|_{p_4}^{\alpha_4} EY_n^2 \le cb_n.\end{aligned}$$

Furthermore, if (3.2) is satisfied, (3.3) follows as above but with $Y_n \equiv 0$. □

With the above technique and a further generalization of Condition 3.1(i) one can obtain a power expansion of the error.

THEOREM 3.6. *Assume Condition* 3.1 *but with* (i) *replaced by*

$$\begin{aligned}\text{(i)}' \qquad F - F_n = &\sum_{i=2}^{l}\sum_{j_1,\dots,j_i=0,1}\int_{A_n^{j_1,\dots,j_i}} u_{j_1,\dots,j_i}(s_1,\dots,s_i)dW_{s_1}^{j_1}\dots dW_{s_i}^{j_i} \\ &+ \sum_{j_1,\dots,j_{l+1}=0,1}\int_{R_n^{j_1,\dots,j_{l+1}}} u_n(s_1,\dots,s_{l+1})dW_{s_1}^{j_1}\dots dW_{s_{l+1}}^{j_{l+1}}\end{aligned}$$

for $l \ge 2$, where $A_n^{j_1,\dots,j_i}$ is a subset of $[0,1]^i$ with $\sum_{i=2}^{l}\sum_{j_1,\dots,j_i=0,1}|A_n^{j_1,\dots,j_i}|/[2(2^l - 2)] \le a_n \to 0$ as $n \to \infty$, $R_n^{j_1,\dots,j_{l+1}}$ is a subset of $[0,1]^{l+1}$, and $u_{j_1,\dots,j_i}$ as well as u_n are two measurable stochastic processes not necessarily adapted. Assume

$$\max_i \sup_{s_1,\dots,s_i}\|u_{j_1,\dots,j_i}(s_1,\dots,s_i)\|_{k,p} + \sup_n \sup_{s_1,\dots,s_{l+1}}\|u_n(s_1,\dots,s_{l+1})\|_{k,p} < \infty$$

for $k \in \mathbb{Z}_+$, $p > 1$. Let Y_n be an independent normal random variable with mean 0 and variance $b_n \le \sum_{i=2}^{l}\sum_{j_1,\dots,j_i=0,1}|A_n^{j_1,\dots,j_i}|/[2(2^l-2)]$. Then for any distribution

T, there exist deterministic functions $c_{j_1,\dots,j_i}$ and a constant c such that

$$\sup_n \Bigg| E[T(F) - T(F_n + Y_n)] - \sum_{i=2}^{l} \sum_{j_1,\dots,j_i=0,1} \int_{A_n^{j_1,\dots,j_i}} c_{j_1,\dots,j_i}(s_1,\dots,s_i)ds_1\dots ds_i \Bigg| \le c \sum_{j_1,\dots,j_{l+1}=0,1} |R_n^{j_1,\dots,j_{l+1}}| + o\left(\sum_{i=2}^{l} \sum_{j_1,\dots,j_i=0,1} |A_n^{j_1,\dots,j_i}| \right).$$

Furthermore, if (3.2) *is satisfied, then Y_n can be replaced by* 0.

Note that the second integral in Theorem 3.6 is interpreted as the anticipating multiple Skorokhod integral. In this theorem we have not used the coefficients a_n because this was just a bound for the sum of the areas of the sets $A_n^{i,j}$ (see also section 5).

COROLLARY 3.7. *If condition* (i)$'$ *in the above theorem is replaced by*

$$\text{(i)}'' \qquad F - F_n = \sum_{i=2}^{l} \sum_{j_1,\dots,j_i=0,1} \int_{A_n^{j_1,\dots,j_i}} u_{j_1,\dots,j_i}^n(s_1,\dots,s_i)dW_{s_1}^{j_1}\dots dW_{s_i}^{j_i},$$

where $\sup_n \max_i \sup_{s_1,\dots,s_i} \|u_{j_1,\dots,j_i}^n(s_1,\dots,s_i)\|_{k,p} \le c(k,p)$, *then for any distribution T,* $|E[T(F) - T(F_n + Y_n)]| \le c\sum_{i=2}^{l} \sum_{j_1,\dots,j_i=0,1} |A_n^{j_1,\dots,j_i}|$. *Furthermore, if* (3.2) *is valid, then Y_n can be replaced by* 0.

4. Application to diffusion processes. We assume for convenience throughout in this section that $b \in C_b^\infty(\mathbb{R})$ and $\sigma \in C_b^\infty(\mathbb{R})$. Consider the particular case when $F = X_1$ is given by (1.1) and $F_n = \bar{X}_1^n$ is given by its Euler approximation $\bar{X}_{t_i}^n = \bar{X}_{t_{i-1}}^n + b(\bar{X}_{t_{i-1}}^n)\Delta t_i + \sigma(\bar{X}_{t_{i-1}}^n)\Delta W_i$, where $\pi_n = \{0 = t_0 < t_1 < \cdots < t_n = 1\}$ is a partition of $[0,1]$ with mesh $m(\pi_n) = \max\{t_{i+1} - t_i : 0 \le i \le n-1\}$ and $\Delta W_i = W_{t_i} - W_{t_{i-1}}$. We interpolate $\bar{X}^n$ between the grid points by $\bar{X}_t^n = x_0 + \int_0^t b(\bar{X}_{\eta_s}^n)ds + \int_0^t \sigma(\bar{X}_{\eta_s}^n)dW_s$, where $\eta_s = \max\{t_i : t_i < s\}$. We first prove that Condition 3.1(i) is satisfied.

LEMMA 4.1. *Let $b \in C_b^\infty(\mathbb{R})$ and $\sigma \in C_b^\infty(\mathbb{R})$. Then Condition* 3.1(i) *is satisfied for $a_n = m(\pi_n)$.*

Proof.

$$X_t - \bar{X}_t^n = \int_0^t b'(\xi_s^0)(X_s - \bar{X}_s^n)ds + \int_0^t \sigma'(\xi_s^1)(X_s - \bar{X}_s^n)dW_s + \int_0^t b(\bar{X}_s^n) - b(\bar{X}_{\eta_s}^n)ds + \int_0^t \sigma(\bar{X}_s^n) - \sigma(\bar{X}_{\eta_s}^n)dW_s. \tag{4.1}$$

Here ξ_s^0 and ξ_s^1 are random points in the interval determined by X_s and $\bar{X}_s^n$. In particular we understand the expression $b'(\xi_s^0)$ in its integral form $b'(\xi_s^0) = \int_0^1 b'(\bar{X}_s^n + \lambda(X_s - \bar{X}_s^n))d\lambda$ and similarly for $\sigma'(\xi_s^1)$. Note that (4.1) is linear in $X - \bar{X}^n$. Therefore, if we define $\mathcal{E}$ as the unique solution to $\mathcal{E}_t = 1 + \int_0^t b'(\xi_s^0)\mathcal{E}_s ds + \int_0^t \sigma'(\xi_s^1)\mathcal{E}_s dW_s$, we

have

$$X_t - \bar{X}_t^n = \mathcal{E}_t \int_0^t \mathcal{E}_s^{-1}\sigma'(\epsilon_s^0)\{b(\bar{X}_{\eta_s}^n)(s-\eta_s) + \sigma(\bar{X}_{\eta_s}^n)(W_s - W_{\eta_s})\}ds$$
$$+\mathcal{E}_t \int_0^t \mathcal{E}_s^{-1}b'(\epsilon_s^1)\{b(\bar{X}_{\eta_s}^n)(s-\eta_s) + \sigma(\bar{X}_{\eta_s}^n)(W_s - W_{\eta_s})\}dW_s$$
$$- \mathcal{E}_t \int_0^t \mathcal{E}_s^{-1}\sigma'(\xi_s^1)\sigma'(\epsilon_s^0)\{b(\bar{X}_{\eta_s}^n)(s-\eta_s) + \sigma(\bar{X}_{\eta_s}^n)(W_s - W_{\eta_s})\}ds.$$

Here $b'(\epsilon_s^1) = \int_0^1 b'(\bar{X}_{\eta_s}^n + \lambda(\bar{X}_s^n - \bar{X}_{\eta_s}^n))d\lambda$, and similarly for $\sigma'(\epsilon_s^0)$. By using the integration by parts formula (see, e.g., [11, equation (1.49), p. 40])

$$X_t - \bar{X}_t^n = \sum_{i,j\in\{0,1\}} \int_0^t \int_{\eta_{s_2}}^{s_2} u_n^{i,j}(s_1,s_2)dW_{s_1}^i dW_{s_2}^j.$$

It is straightforward to show that $\|u_n^{i,j}(s_1,s_2)\|_{k,p}$ is uniformly bounded in (s_1,s_2) and n. Clearly $|A_n^{i,j}| = \int_0^1 \int_{\eta_s}^s duds \le m(\pi_n)$. Condition 3.1(i) is satisfied. □

Now we introduce sufficient conditions that ensure the smoothness of the density of X_t. This also explains the terminology introduced for Condition 3.1 and (3.2).

CONDITION 4.2 (Hörmander condition). *$|\sigma(x_0)| \ge \epsilon > 0$ or $|b(x_0)\sigma^{(k)}(x_0)| \ge \epsilon > 0$ for some $k \in \mathbb{N}$ and for some $\epsilon > 0$.*

CONDITION 4.3 (uniform ellipticity condition). *$|\sigma(x)| \ge \epsilon > 0$ for all $x \in \mathbb{R}$ and for some $\epsilon > 0$.*

LEMMA 4.4. (i) *If Condition 4.2 is satisfied, then Condition 3.1 is satisfied for $F = X_1$, $F_n = \bar{X}_1^n$ with $h_n(s) = h_{n,\alpha}(s) = \alpha D_s F_n + (1-\alpha)D_s F$, $h(s) = D_s F$, $e_n(p) = c_p a_n^{p/4}$ for some constant c_p, $b_n = d_n = a_n = m(\pi_n)$.*

(ii) *If Condition 4.3 is satisfied and $1 - t_{n-1} \ge cm(\pi_n)$ for some $c > 0$, then (3.2) is satisfied with the same choices for h_n and h as above.*

Results similar to Lemma 4.4 for h and F are well known; see, e.g., [11, p. 111].

Proof of Lemma 4.4. First we prove Lemma 4.4(i). Condition 3.1(i) is satisfied by Lemma 4.1. Condition 3.1(ii) is satisfied by Lemma 3.3. In fact,

$$E\,\|h_{n,\alpha} - h\|_{L^2[0,1]} = (1-\alpha)E\left\|D(\bar{X}_1^n - X_1)\right\|_{L^2[0,1]} \le c_p a_n^{p/4}.$$

Furthermore, by Condition 4.2 and the proof of [11, Theorem 2.3.2], $E[(\int_0^1 (D_s X_1)^2 ds)^{-p}] < \infty$ for all $p > 1$, and Condition 3.1(ii) follows.

To prove that Condition 3.1(iii) is satisfied we note that obviously $|a_n + \int_0^1 (D_s X_1)^2 ds| > |\int_0^1 (D_s X_1)^2 ds|$ and $|a_n + \int_0^1 (\alpha D_s \bar{X}_1^n + (1-\alpha)D_s X_1)^2 ds| \ge d_n \equiv a_n$. Clearly $\sup_n d_n^{-p}(a_n^{k(p)/2} + e_n(4k(p))^{1/2} + b_n^{2k(p)}) = 2a_n^{k(p)/2-p} + b_n^{2k(p)-p} < \infty$ if $2k(p) \ge p$.

Next we prove Lemma 4.4(ii). Similar to the proof of Lemma 3.4 we define the set $A \equiv \{\int_0^1 \left(D_s(\bar{X}_1^n - X_1)\right)^2 ds < \frac{1}{4}\int_0^1 (D_s X_1)^2 ds\}$, and we have that for any $p > 1$,

$$\sup_n \sup_\alpha E\left[\left|\int_0^1 (\alpha D_s \bar{X}_1^n + (1-\alpha)D_s X_1)^2 ds\right|^{-p}; A\right] \le 4^{-p} E\left[\left|\int_0^1 (D_s X_1)^2 ds\right|^{-p}\right] < \infty.$$

Next we find a similar bound for the expectation taken over the set A^c. Note that without loss of generality we can suppose that $\sigma(x) \ge \epsilon > 0$ for all $x \in \mathbb{R}$. Then

$D_sX_1 > \epsilon \exp(\int_s^1 \bar{b}'(X_u)du + \int_s^1 \sigma'(X_u)dW_u)$ (see (7.1)). Also notice that $D_s\bar{X}_1^n = \sigma(\bar{X}(t_{n-1}^n)) \geq \epsilon > 0$ for $t_{n-1} < s \leq 1$ (see (7.3) and (7.5)). Hence for $\alpha > 1/2$,

$$\left| \int_0^1 (\alpha D_s\bar{X}_1^n + (1-\alpha)D_sX_1)^2 ds \right| \geq \frac{1}{4}\epsilon^2(1-t_{n-1}).$$

For $\alpha \leq 1/2$ we use that

$$\begin{aligned} &\left| \int_0^1 (\alpha D_s\bar{X}_1^n + (1-\alpha)D_sX_1)^2 ds \right| \\ &\geq \frac{1}{4}\epsilon^2 \int_{t_{n-1}}^1 \exp\left(2\int_s^1 \bar{b}'(X_u)du + 2\int_s^1 \sigma'(X_u)dW_u\right) ds. \end{aligned}$$

In A^c the above estimates together with Chebyshev's inequality and Lemma 3.3 complete the proof as in the proof of Lemma 3.4. That is,

$$\begin{aligned} &\sup_n \sup_\alpha E\left[\left| \int_0^1 (\alpha D_s\bar{X}_1^n + (1-\alpha)D_sX_1)^2 ds \right|^{-p}; A^c \right] \\ &\leq 4^p\epsilon^{-2p}\Bigg((1-t_{n-1})^{-p}P(A^c) \\ &\quad + E\left[\left(\int_{t_{n-1}}^1 \exp\left(2\int_s^1 \bar{b}'(X_u)du + 2\int_s^1 \sigma'(X_u)dW_u\right) ds \right)^{-2p} \right]^{1/2} P(A^c)^{1/2} \Bigg) \\ &\leq cm(\pi_n)^{-p}(P(A^c) + P(A^c)^{1/2}). \end{aligned}$$

From here the result follows as $P(A^c) \leq c_k(m(\pi_n))^{k/2}$ for any $k > 1$. Taking k big enough finishes the proof of the lemma. □

Section 3 gives the rate of convergence of the Euler approximation. The same proof gives the following stronger result.

PROPOSITION 4.5. *Assume Condition* 4.2. *Then*

$$\sup_x |E\delta_x(X_1) - E\delta_x(\bar{X}_1^n + Y_n)| \leq cm(\pi_n),$$

where Y_n is an independent normal random variable with zero mean and variance $m(\pi_n)$. Furthermore, if Condition 4.3 *is valid and $1 - t_{n-1} \geq cm(\pi_n)$ for some $c > 0$, then*

$$\sup_x |E\delta_x(X_1) - E\delta_x(\bar{X}_1^n)| \leq cm(\pi_n).$$

This stronger version follows because the antiderivative of the delta function is the indicator function which is bounded in x. Applying Remark 2.2 to our current setting gives the following.

REMARK 4.6 (control variate method). (i) *Assume Condition* 4.2 *with the choices for h_n and h in Lemma* 4.4(i). *Let $\xi_{n,r}^{h_n}(x) = (1_{\{\bar{X}_1^n + Y_n > x\}} - c(x))H^{h_n}(\bar{X}_1^n + Y_n, \varphi(\frac{\bar{X}_1^n + Y_n - x}{r}))$. Then $E(\xi_{n,r}^{h_n}(x)) = E\delta_x(\bar{X}_1^n + Y_n)$ and $\sup_{x,n} E(\xi_{n,r}^{h_n}(x)^2) < \infty$. If $E\delta_x(\bar{X}_1^n + Y_n) > 0$, then $E[H^{h_n}(\bar{X}_1^n + Y_n, \varphi(\frac{\bar{X}_1^n - x}{r}))^2] > 0$ and, for fixed φ and r, the variance of $\xi_{n,r}^h(x)$ is minimized by*

$$c(x) = c_{n,r}^h(x) = \frac{E[1_{\{\bar{X}_1^n + Y_n > x\}} H^{h_n}(\bar{X}_1^n + Y_n, \varphi(\frac{\bar{X}_1^n + Y_n - x}{r}))^2]}{E[H^{h_n}(\bar{X}_1^n + Y_n, \varphi(\frac{\bar{X}_1^n + Y_n - x}{r}))^2]}.$$

(ii) *Assume Condition* 4.3 *with the choices for* h_n *and* h *in Lemma* 4.4(ii). *Then* Y_n *above can be replaced by* 0. *Furthermore* $E\delta_x(\bar{X}_1^n) > 0$. □

Results similar to Proposition 4.5 have already been obtained in [1] and [2]. The main difference with the results here is that the method of proof is somewhat different and that our Proposition 4.5 is the result of a general theory based on Itô–Taylor expansions which can also be applied to other situations. In fact, under further restrictions on the structure of the sets A_n, R_n, and the continuity of the processes u and u_n, one can improve Theorem 3.6 to obtain Taylor expansions of the errors. For example in the uniformly elliptic case we have

$$E[T(F_n) - T(F)] = c_1 a_n + c_2^{(n)} a_n^2$$

for any distribution T. In the general Hörmander case,

$$E[T(F_n + Y_n) - T(F)] = c_1 a_n + c_2 b_n + c_3^{(n)} a_n b_n + c_4^{(n)} a_n^2 + c_5^{(n)} b_n^2,$$

where Y_n is a mean zero normal random variable with variance b_n, independent of W, and $\sup_n |c_i^{(n)}| < \infty$, $i = 2, \dots, 5$. This result will be proven elsewhere.

5. Kernel density estimation method. So far we have given convergence results for the density approximation by integration by parts. In this section we discuss heuristically the "most natural" approach by kernel density estimates and compare the asymptotic variances.

The kernel density estimation technique is a very well known method used in statistics. The main difference between this and our situation is that in statistics the amount of data is limited while here the amount of simulations can be fixed by the user. Nevertheless, the same theory gives us some insight into the optimal use of this method for simulation of densities.

That is, let ϕ be a smooth positive even function with $\int_{\mathbb{R}} \phi(x)dx = 1$. Then the approximation of the density is obtained by computing $\sum_{i=1}^N \phi(\frac{F_n^i - x}{h})/(Nh)$. The error is measured through the $L^2(\mathbb{R})$-norm of the variance. Estimating this error requires the study of various errors.

The first error is the difference between the expectations of the simulated approximation and the limit random variable,

$$\frac{1}{h} E\left[\phi\left(\frac{F_n - x}{h}\right) - \phi\left(\frac{F - x}{h}\right)\right] = c_1(x) a_n + c_2^{(h,n)}(x) a_n^2, \tag{5.1}$$

where $\sup_{n,h} |c_2^{(h,n)}(x)| < \infty$. Here the constants obviously depend also on ϕ. To obtain this result it is enough to notice that

$$\frac{1}{h} E\left[\phi\left(\frac{F_n - x}{h}\right) - \phi\left(\frac{F - x}{h}\right)\right] = E[\delta_x(F_n + hY) - \delta_x(F + hY)],$$

where Y is a smooth random variable with density given by ϕ. This converts the estimation of the error into the uniformly elliptic case. Therefore the same method of proof as in Theorem 3.6 can be used.

The second error is the difference between the density to be approximated and the approximation with the kernel function (see, e.g., [12]):

$$\frac{1}{h} E\left[\phi\left(\frac{F - x}{h}\right) - \delta_x(F)\right] = \frac{1}{2} h^2 p''(x) \int u^2 \phi(u) du + O(h^4) p^{(4)}(x),$$

where p is the density of F. Similarly for the mean square error,

$$\begin{aligned}
&E\left[\left(\frac{1}{Nh}\sum_{i=1}^{N}\phi\left(\frac{F_n^i-x}{h}\right)-p(x)\right)^2\right]\\
&=\operatorname{Var}\left[\frac{1}{Nh}\sum_{i=1}^{N}\phi\left(\frac{F_n^i-x}{h}\right)\right]+\left(\frac{1}{h}E\left[\phi\left(\frac{F_n-x}{h}\right)-\phi\left(\frac{F-x}{h}\right)\right]\right)^2\\
&\quad+\left(E\left[\frac{1}{h}\phi\left(\frac{F-x}{h}\right)-p(x)\right]\right)^2\\
&\quad+2\frac{1}{h}E\left[\phi\left(\frac{F_n-x}{h}\right)-\phi\left(\frac{F-x}{h}\right)\right]E\left[\frac{1}{h}\phi\left(\frac{F-x}{h}\right)-p(x)\right]\\
&=p_n(x)\frac{1}{Nh}\int\phi^2(u)du+c_1(x)^2a_n^2+\frac{h^4}{4}\left(p''(x)\int u^2\phi(u)du\right)^2\\
&\quad+c_1(x)h^2a_np''(x)\int u^2\phi(u)du+\text{higher order terms},
\end{aligned}$$

where p_n is the density of F_n. If one considers as a minimization criterion the L^1-norm of the mean squared error, this gives the classical criterion of kernel density estimation. That is,

$$\begin{aligned}
&\int E\left[\left(\frac{1}{Nh}\sum_{i=1}^{N}\phi\left(\frac{F_n^i-x}{h}\right)-p(x)\right)^2\right]dx\\
&\quad\approx\frac{1}{Nh}\int\phi^2(u)du+c_1^2a_n^2+\frac{h^4}{4}\int p''(x)^2dx\left(\int u^2\phi(u)du\right)^2.
\end{aligned}$$

The optimum is therefore obtained when ϕ is the Epanechnikov kernel and $h\sim N^{-1/5}$ and $N\sim a_n^{-5/2}$.

6. Optimal choice for the integration by parts method. As in the previous section we will find heuristically an optimal choice of localization function φ and localization parameter r for the integration by parts method introduced in section 2. In order to do this we will find an asymptotical expression for the variance of the simulations.

Let $\varphi\in C_b^1$ with $\varphi(0)=1$. One criteria for optimality may be to choose φ and r so that they minimize

$$\int_{\mathbb{R}}E\left[1(F_n+Y_n\geq x)H^h\left(F_n+Y_n,\varphi\left(\frac{F_n+Y_n-x}{r}\right)\right)^2\right]dx \tag{6.1}$$

under the general Hörmander condition. This criteria can be studied but is cumbersome as the optimal choices will depend on n. Instead one may study the limit assuming that the error terms are small. Therefore for simplicity we consider for small r, under convenient smoothness and boundedness conditions, the asymptotic limit of (6.1) which, using (2.5), equals $\int_{\mathbb{R}}I(x)dx$, where

$$I(x)=E\left[1(F\geq x)\left(\varphi\left(\frac{F-x}{r}\right)H^h(F,1)-\frac{1}{r}\varphi'\left(\frac{F-x}{r}\right)\right)^2\right].$$

Let $H_i(x) = E[H^h(F,1)^i|F=x]$ for $i=1,2$, and let p be the density of F. Then

$$\begin{aligned} I(x) &= \int_x^\infty E\left[\left(\varphi\left(\frac{y-x}{r}\right)H^h(F,1) - \frac{1}{r}\varphi'\left(\frac{y-x}{r}\right)\right)^2 |F=y\right] p(y)dy \\ &= \int_x^\infty \left(\frac{1}{r^2}\varphi'\left(\frac{y-x}{r}\right)^2 - \frac{2}{r}\varphi\varphi'\left(\frac{y-x}{r}\right)H_1(y) + \varphi\left(\frac{y-x}{r}\right)^2 H_2(y)\right)p(y)dy \\ &= r\int_0^\infty \left(\frac{1}{r^2}\varphi'(z)^2 - \frac{2}{r}\varphi\varphi'(z)H_1(x+rz) + \varphi(x+rz)^2 H_2(x+rz)\right)p(x+rz)dz. \end{aligned}$$

Under smoothness and boundedness conditions of H_i, φ, and p, $I(x) = I_2(x) + O(r^2)$ for small r, where

$$\begin{aligned} I_2(x) = {} & \frac{1}{r}p(x)\int_0^\infty \varphi'(z)^2 dz + H_1(x)p(x)\int_0^\infty (\varphi'(z)^2 z p'(x) - 2\varphi\varphi'(z))dz \\ & + rp''(x)\int_0^\infty \frac{1}{2}\varphi'(z)^2 z^2 dz + H_2(x)p(x)\int_0^\infty \varphi^2(z)dz \\ & - 2(H_1(x)p'(x) + H_1'(x)p(x))\int_0^\infty \varphi\varphi'(z)z dz \end{aligned} \tag{6.2}$$

and

$$\int_{\mathbb{R}} I_2(x)dx = \frac{1}{r}\int_0^\infty \varphi'(z)^2 dz + r\int_0^\infty \varphi^2(z)dz \int_{\mathbb{R}} H_2(x)p(x)dx, \tag{6.3}$$

where $\int_{\mathbb{R}} H_2(x)p(x)dx = E[H^h(F,1)^2]$ and $\int_{\mathbb{R}} H_1(x)p(x)dx = E[H^h(F,1)] = 0$. An optimal value for r which minimizes (6.3) is given by

$$r = \left(\frac{\int_0^\infty \varphi'(z)^2 dz}{E[H^h(F,1)^2]\int_0^\infty \varphi(z)^2 dz}\right)^{1/2}.$$

Replacing this r in (6.3) yields

$$\int_{\mathbb{R}} I_2(x)dx = 2\left(E\left[H^h(F,1)^2\right]\int_0^\infty \varphi(z)^2 dz \int_0^\infty \varphi'(z)^2 dz\right)^{1/2},$$

which, by variational analysis, is minimized for φ solving

$$\varphi(z)\int_0^\infty \varphi'(z)^2 dz - \varphi''(z)\int_0^\infty \varphi(z)^2 dz = 0. \tag{6.4}$$

For any $\lambda > 0$, the function $\varphi(z) = e^{-\lambda|z|}$ is symmetric, solves (6.4), and satisfies $\varphi(0) = 1$. Hence we propose as a natural choice of r and φ

$$r = \left(\frac{\int_0^\infty \varphi'(z)^2 dz}{E[H^h(F,1)^2]\int_0^\infty \varphi(z)^2 dz}\right)^{1/2}, \qquad \varphi(x) = e^{-\lambda|x|}, \tag{6.5}$$

where $\lambda > 0$ may be arbitrarily chosen. Note that the main error term (6.3) with optimal φ is independent of the value of λ.

After the minimization in r and φ is done one can apply the control variate method introduced in Remark 4.6. Therefore the variance error for the integration by parts with control variates and localization is in the uniformly elliptic case

$$E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(1(F_n^i \geq x) - c_n(x,r))\bar{H}_n\left(F_n^i, \varphi\left(\frac{F_n^i - x}{r}\right)\right) - p(x)\right)^2\right]$$
$$\approx C_1(x)^2 a_n^2 + \frac{\mathrm{Var}[(1(F_n \geq x) - c_n(x,r))\bar{H}_n(F_n, \varphi(\frac{F_n - x}{r}))]}{N}.$$

The optimal choice is therefore $N \sim a_n^{-2}$. For the general Hörmander case, $N \sim (a_n + b_n)^{-2}$.

6.1. Comparison of the kernel density estimate and the integration by parts method: Some conclusions and remarks. A first look at both methods shows that kernel density estimation has a square bias asymptotically equal to $h^4 p''(x)^2 \int u^2 \phi(u) du$ due to the fact of using ϕ, besides the square bias $c_1(x)^2 a_n^2 + c_2(x)^2 b_n^2$ from the approximation of F. If the first type of error is much smaller than the second one, then only the second one is important when comparing the two methods.

In order to compare both methods, suppose that $a_n = n^{-1}$. Then the optimal sample size for the integration by parts method is $N = n^2$, which is significantly less than the optimal sample size $N = n^{5/2}$ for the kernel density method. Furthermore, the kernel density method creates bias while the integration by parts does not, at least theoretically. Nevertheless, the amount of calculations in the integration by parts method is higher.

The optimal parameter $r_{n,N}$ does not go to 0 as n, N increase. In fact, r could remain constant throughout the calculations with little increase of the variance. It seems that $r_{n,N} \to r > 0$ in most of the cases. Numerical experiments indicate that the choice of r does not look to be sensitive. Kernel density estimation often requires a fine tuning of the bandwidth h.

There is no clear way to apply a control variate method to kernel density estimation methods.

In higher dimensions the kernel density estimate rate of convergence deteriorates typically to $N^{-\frac{4}{d+4}}$ while the integration by parts keeps the same rate.

Constants in the integration by parts methods increase in value as the degree of hypoellipticity increases.

Similar variance reductions could be studied on other environments where an integration by parts formula is available. For example, in the Poisson case one could use the same ideas as shown here; see, e.g., [3].

7. Numerical implementation. We consider the particular case when $F = X_1$ is given by (1.1) and F_n is its Euler approximation. We first note that

$$D_s X_t = \begin{cases} \sigma(X_s) e^{\int_s^t \bar{b}'(X_v)dv + \int_s^t \sigma'(X_v)dW_v}, & s \leq t, \\ 0, & s > t, \end{cases} \tag{7.1}$$

where $\bar{b}'(X_v) = b'(X_v) - \frac{1}{2}\sigma'(X_v)^2$; see, e.g., [11, p. 107]. Using (7.1), it follows that

(7.2)
$$D_s D_t X_1 = D_s[X_t]\sigma'(X_t) e^{\int_t^1 \bar{b}'(X_v)dv + \int_t^1 \sigma'(X_v)dW_v}$$

$$+\left[\sigma'(X_s)1_{\{t\le s\}}+\int_t^1 \bar{b}''(X_v)D_sX_v dv+\int_t^1 \sigma''(X_v)D_sX_v dW_v\right]D_tX_1.$$

Since

$$D_{t_j}\bar{X}^n_{t_k}=D_{t_j}\bar{X}^n_{t_{k-1}}+[b'(\bar{X}^n_{t_{k-1}})\Delta t+\sigma'(\bar{X}^n_{t_{k-1}})\Delta W_k]D_{t_j}\bar{X}^n_{t_{k-1}}\\ +\sigma(\bar{X}^n_{t_{k-1}})D_{t_j}\Delta W_k,$$

and $D_s\Delta W_{t_k}=1_{\{t_{k-1}<s\le t_k\}}$, it follows that $D_{t_k}\bar{X}^n_{t_k}=\sigma(\bar{X}^n_{t_{k-1}})$. By induction,

(7.3)
$$D_{t_j}\bar{X}^n_{t_k}=\begin{cases}0, & j=0,\\ \sigma(\bar{X}^n_{t_{k-1}}), & 1\le j=k,\\ \sigma(\bar{X}^n_{t_{j-1}})\Pi_{l=j}^{k-1}(1+b'(\bar{X}^n_{t_l})\Delta t_{l+1}+\sigma'(\bar{X}^n_{t_l})\Delta W_{l+1}), & 1\le j\le k-1,\\ 0, & j\ge k+1.\end{cases}$$

Note that (7.3) is a discrete version of (7.1) ($\Pi_j(1+\varepsilon_j)\approx e^{\sum_j(\epsilon_j-\epsilon_j^2/2)}$). Using similar arguments one obtains

(7.4)
$$D_{t_i}D_{t_j}\bar{X}^n_{t_k}=\begin{cases}0, & j=0,\\ 0, & j\ge k+1,\\ b'(\bar{X}^n_{t_{k-1}})D_{t_i}\bar{X}^n_{t_{k-1}}, & j=k,\\ D_{t_i}[\bar{X}^n_{t_{j-1}}]b'(\bar{X}^n_{t_{j-1}})\Pi_{l=j}^{k-1}(1+a'(\bar{X}^n_{t_l})\Delta t+b'(\bar{X}^n_{t_l})\Delta W_{l+1})\\ \quad+\Big(\sum_{l=j}^{k-1}\frac{[a''(\bar{X}^n_{t_l})\Delta t+b''(\bar{X}^n_{t_l})\Delta W_{l+1}]D_{t_i}\bar{X}^n_{t_l}}{1+a'(\bar{X}^n_{t_l})\Delta t+b'(\bar{X}^n_{t_l})\Delta W_{l+1}}\\ \quad+\frac{b'(\bar{X}^n_{t_{i-1}})1_{\{j\le i-1\le k-1\}}}{1+a'(\bar{X}^n_{t_{i-1}})\Delta t+b'(\bar{X}^n_{t_{i-1}})\Delta W_i}\Big)D_{t_j}\bar{X}^n_{t_k},\\ 2\le j\le k-1\end{cases}$$

which is a discrete version of (7.2).

To apply the above formulas to the integration by parts method we need to use that

$$D_s\bar{X}^n_{t_k}=D_{\eta_s^+}\bar{X}^n_{t_k},\tag{7.5}$$

where $\eta_s^+=\min\{t_i:t_i\ge s\}$. This follows because

$$D_s\bar{X}^n_{t_k}=D_s[\bar{X}^n_{t_{k-1}}+b(\bar{X}^n_{t_{k-1}})\Delta t_k+\sigma(\bar{X}^n_{t_{k-1}})\Delta W_k]\\ =[1+b'(\bar{X}^n_{t_{k-1}})\Delta t_k+\sigma'(\bar{X}^n_{t_{k-1}})\Delta W_k]D_s\bar{X}^n_{t_{k-1}}+\sigma(\bar{X}^n_{t_{k-1}})1_{\{t_{k-1}<s\le t_k\}}.$$

We then have that for $t_{k-1}<s<t_k$, $D_s\bar{X}^n_{t_k}=\sigma(\bar{X}^n_{t_{k-1}})=D_{t_k}\bar{X}^n_{t_k}$. By induction we have in general that for $t_{j-1}<s<t_j$, $D_s\bar{X}^n_{t_k}=D_{t_j}\bar{X}^n_{t_k}$ for $j=1,\dots,k$. We also have that

$$D_sD_t\bar{X}^n_{t_k}=D_{\eta_s^+}D_{\eta_t^+}\bar{X}^n_{t_k}.\tag{7.6}$$

Using (7.5), (7.6), and (2.7) gives

$$H^1(\bar{X}^n_1,1)=\frac{W_1}{\sum_1^n D_{t_i}\bar{X}^n_1\Delta t_i}+\frac{\sum_{i,j=1}^n D_{t_i}D_{t_j}\bar{X}^n_1\Delta t_i\Delta t_j}{(\sum_1^n D_{t_i}\bar{X}^n_1\Delta t_i)^2},$$

from which $H^1(\bar{X}_t^n, \varphi((\bar{X}_t^n - x)/r))$ can be computed by (2.5). An approximation to the density using the integration by parts formula can now be explicitly written. For example,

$$f^{n,N}(x) = \frac{1}{N}\sum_{i=1}^{N}(1_{\{\bar{X}_1^{n,i}\geq x\}} - \hat{c}_{loc}^1(x))H^1\left(\bar{X}_1^{n,i}, \varphi\left(\frac{\bar{X}_1^i - x}{r}\right)\right), \tag{7.7}$$

where

$$\hat{c}_{loc}^1(x) = \frac{\frac{1}{N}\sum_{i=1}^{N} 1_{\{\bar{X}_1^{n,i}\geq x\}} H^1\left(\bar{X}_1^{n,i}, \varphi\left(\frac{\bar{X}_1^{n,i}-x}{r}\right)\right)^2}{\frac{1}{N}\sum_{i=1}^{N} H^1\left(\bar{X}_1^{n,i}, \varphi\left(\frac{\bar{X}_1^{n,i}-x}{r}\right)\right)^2}$$

is a natural estimate of (2.9).

We perform the simulation (7.7) with optimal φ and r from (6.5) with $\lambda = 1$ and equidistant partition $m(\pi_n) = n^{-1}$ and compare with a locally optimal r (numerically obtained optimal r for fixed x) and the kernel density estimate; see Figure 1. We also compare the convergences in Figure 2. The computations are made in MATLAB.

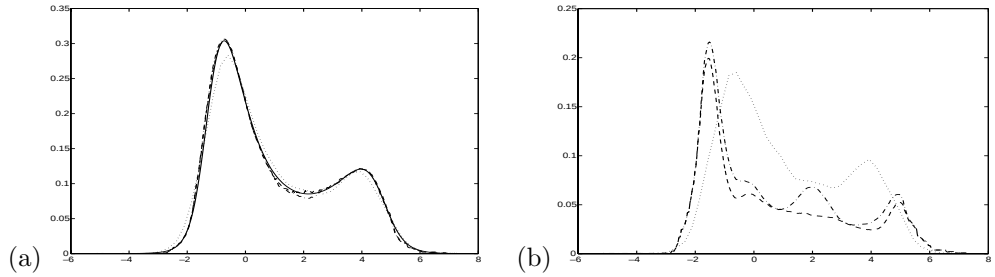


FIG. 1. *Monte Carlo simulation of* $dX = dt + (\sin X + 2)dW$, $X_0 = 0$, $n = m(\pi_n)^{-1} = 3000$, $N = 1000$. (a) *Approximation using the integration by parts formula with control variate; local search of optimal r - - - (optimal r for given x), and global search of optimal r - · - (minimizing* (6.1)*), respectively. Gaussian kernel density estimate with optimal bandwidth* [4, p. 47] *. . .. The numerical solution of the Fokker Planck equation —.* (b) *Corresponding sample variances of the estimates.*

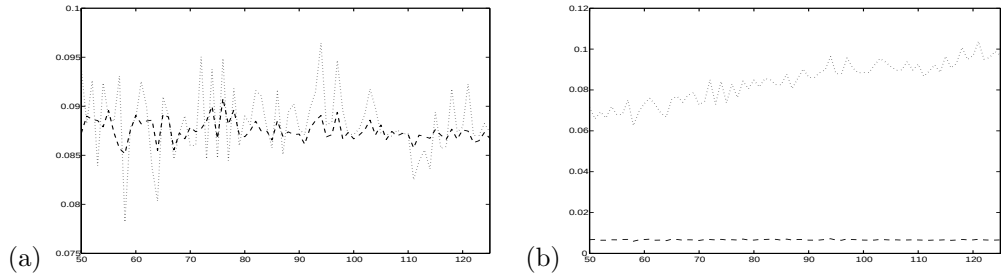


FIG. 2. *Monte Carlo simulation of* $dX = dt + (\sin X + 2)dW$, $X_0 = 0$. *In* (a)*, convergence of approximations to the density at* $x = 2$ *for* $n = m(\pi_n)^{-1} = 50, 51, \ldots, 125$, $N = n^2$. *Integration by parts method with control variate, local search of optimal r - - - (optimal r for given x), Gaussian kernel density estimate with optimal bandwidth* [4, p. 47] *. . .. In* (b)*, corresponding sample variances.*

REFERENCES

[1] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations* (I): *Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[2] V. Bally and D. Talay, *The law of the Euler scheme for stochastic differential equations* (II): *Convergence rate of the distribution function*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

[3] K. Bichteler, J. B. Gravereaux, and J. Jacod, *Malliavin Calculus for Processes with Jumps*, Gordon and Breach Science, New York, 1987.

[4] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1997.

[5] E. Fournié, J. M. Lasry, J. Lebuchoux, P. L. Lions, and N. Touzi, *Applications of Malliavin calculus to Monte-Carlo methods in finance*, Finance Stoch., 3 (1999), pp. 391–412.

[6] E. Fournié, J. M. Lasry, J. Lebuchoux, and P. L. Lions, *Applications of Malliavin calculus to Monte-Carlo methods in finance* II, Finance Stoch., 5 (2001), pp. 201–236.

[7] P. Glasserman, P. Heidelberger, and P. Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Math. Finance, 9 (1999), pp. 17–152.

[8] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer-Verlag, Berlin, 1992.

[9] N. J. Newton, *Variance reduction for simulated diffusions*, SIAM J. Appl. Math., 54 (1994), pp. 1780–1805.

[10] D. Nualart, *Analysis on Wiener space and anticipating stochastic calculus*, in Lectures on Probability Theory and Statistics (Saint-Flour, 1995), Lecture Notes in Math. 1690, Springer-Verlag, Berlin, 1998, pp. 123–227.

[11] D. Nualart, *The Malliavin Calculus and Related Topics*, Springer-Verlag, Berlin, 1995.

[12] W. Silverman, *Density Estimation*, Chapman and Hall, London, 1986.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 451–468

CONVERGENCE ANALYSIS OF A TRANSMISSION ALGORITHM FOR CONVECTION-DIFFUSION PROBLEMS*

M. TIDRIRI†

Abstract. We prove the convergence of the transmission time marching algorithm for the β–monoscale-multimodel methods for hydrodynamics-type problems. Our analysis is based on the author's transmission multiplier method we introduced in [*C. R. Acad. Sci. Paris Sér.* I *Math.*, 328 (1999), pp. 637–642] and the local, global, and trace estimates we developed in [*J. Math. Anal. Appl.*, 229 (1999), pp. 137–157; *Abstr. Appl. Anal.*, 3 (2001), pp. 131–150].

Key words. boundary layers, compressible Navier–Stokes equations, convection-diffusion problems, Dirichlet–Dirichlet problems, Dirichlet–Neumann problems, local, global, and trace estimates, transmission multiplier method, transmission time marching algorithm, α–monoscale-multimodel method, β–monoscale-multimodel method

AMS subject classifications. 35J25, 35Q30, 65N12, 76R05

PII. S0036142900382843

1. Introduction.

In this paper we prove the convergence of the transmission time marching algorithm (TTMA) for the β–monoscale-multimodel methods for hydrodynamics-type problems. The α– and β–monoscale-multimodel methods were introduced by the author [7, 8, 9] in order to separately handle local and global phenomena in the modeling of physical systems. These methods offer an easy way of supplementing and testing a large variety of boundary conditions. Moreover, in the local domain the model can be chosen to handle phenomena such as boundary layers and turbulence and can incorporate models of chemistry. We have already proved the validity of these methods and their superiority to classical methods in many (important) real life applications. More details about this type of methods can be found in [8, 9]. The TTMA [8, 9] is an algorithm that allows us to solve the multimodels obtained by applying the α– and β–monoscale-multimodel methods. In this paper we are interested in studying the convergence properties of the TTMA for the β–monoscale-multimodel methods for hydrodynamics-type problems.

Because of the practical importance of this algorithm, the establishment of its mathematical foundations is of crucial importance. The mathematical theory of such an algorithm started in [2, 3]. The analysis of the general TTMA for hydrodynamics-type applications remained open. We shall provide in this paper a complete analysis of such an algorithm. Our analysis is based on a general method we have introduced in [10]. Because of its practical importance we shall term it "the transmission multiplier method." We also use the local, global, and trace estimates we developed in [11, 12].

In section 2, we describe the α– and β–monoscale-multimodel methods for convection-diffusion problems and the TTMA that we propose for their solutions together with some applications in fluid mechanics. In section 3, we state the main local, global, and trace estimates of [11, 12] we shall need for the analysis of the TTMA. In

*Received by the editors December 27, 2000; accepted for publication (in revised form) November 1, 2001; published electronically May 29, 2002. The research of this author was supported by the Air Force Office of Scientific Research under contract F49620-99-1-0197.

http://www.siam.org/journals/sinum/40-2/38284.html

†Department of Mathematics, Iowa State University, 400 Carver Hall, Ames, IA 50011-2064 (tidriri@iastate.edu).

section 4, we study the convergence properties of the TTMA. We conclude this paper by some comments in section 5.

2. The α- and β-methods for convection-diffusion problems and the TTMA.

2.1. The convection-diffusion model. Let Ω be a connected bounded domain of $\mathbb{R}^n$, such that its boundary $\partial\Omega$ is Lipschitzian and Ω_l is a connected domain of $\mathbb{R}^n$ with $\Omega_l \subset \Omega$ (Figure 2.1). Let

$$\Gamma_b = \partial\Omega \cap \partial\Omega_l \quad \text{(internal boundary)},$$

$$\Gamma_i = \partial\Omega_l \cap \Omega \quad \text{(interface)},$$

$$\Gamma_\infty = \partial\Omega \backslash \Gamma_b \quad \text{(farfield boundary)}.$$

We denote by n the external unit normal vector to $\partial\Omega$ or $\partial\Omega_l$.

Let $v \in (L^\infty(\Omega))^n$ be a given velocity field of an inviscid incompressible flow such that

$$\begin{cases} \operatorname{div} v = 0 & \text{in} \quad \Omega, \\ v.n = 0 & \text{on} \quad \Gamma_b. \end{cases} \tag{2.1}$$

The model problem we want to solve is the following one:

$$\text{Find } \phi : \Omega \times (0,T) \to \mathbb{R} \text{ such that}$$

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \operatorname{div}(v\phi) - \nu\Delta\phi = 0 \ \text{ in } \ \Omega \times (0,T), \\ \phi = \phi^\infty \ \text{ on } \ \Gamma_\infty \times (0,T), \\ \phi = 0 \ \text{ on } \ \Gamma_b \times (0,T), \\ \phi(0) = \phi_0 \ \text{ in } \ \Omega, \end{cases} \tag{2.2}$$

where v is the velocity field given by (2.1) and ν is the diffusion coefficient. Assuming that $\phi_0 \in L^2(\Omega)$ and $\phi^\infty \in L^2(0,T; H^{1/2}(\Gamma_\infty))$, problem (2.2) has a unique solution $\phi \in L^2(0,T; H^1(\Omega))$. The corresponding stationary problem follows:

Find φ, a real valued function, defined on Ω and satisfying

$$\begin{cases} \operatorname{div}(v\varphi) - \nu\Delta\varphi = 0 \text{ in } \Omega, \\ \varphi = \varphi^\infty \text{ on } \Gamma_\infty, \\ \varphi = 0 \text{ on } \Gamma_b. \end{cases} \tag{2.3}$$

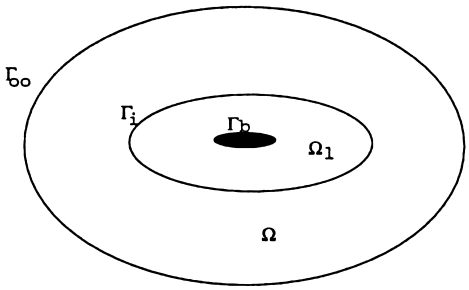


FIG. 2.1. *Description of the domain* Ω *and its splitting.*

2.2. The α- and β-methods for the convection-diffusion model. Let Ω_l be a domain of $\mathbb{R}^n$ such that $\Omega_l \subset \Omega$ (see Figure 2.1) and has as an external boundary Γ_i. The β–monoscale-multimodel method applied to the model problem consists in replacing the evolution problem (2.2) by the following evolution system:

Find ϕ (resp., ϕ_{loc}) : $\Omega \to \mathbb{R}$ (resp., $\Omega_l \to \mathbb{R}$) satisfying

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \operatorname{div}(v\phi) - \nu \Delta \phi = 0 \ \text{ in } \ \Omega \times (0,T), \\ \phi = \phi^{\infty} \ \text{ on } \ \Gamma_{\infty} \times (0,T), \\ \nu \dfrac{\partial \phi}{\partial n} = \nu \dfrac{\partial \phi_{loc}}{\partial n} \ \text{ on } \ \Gamma_b \times (0,T), \end{cases} \tag{2.4}$$

$$\begin{cases} \dfrac{\partial \phi_{loc}}{\partial t} + \operatorname{div}(v\phi_{loc}) - \nu \Delta \phi_{loc} = 0 \ \text{ in } \ \Omega_l \times (0,T), \\ \phi_{loc} = 0 \ \text{ on } \ \Gamma_b \times (0,T), \\ \phi_{loc} = \phi \ \text{ on } \ \Gamma_i \times (0,T), \end{cases} \tag{2.5}$$

$$\phi(0) = \phi_0 \ \text{ in } \ \Omega \ \ \phi_{loc}(0) = \phi_{0loc} \ \text{ in } \ \Omega_l. \tag{2.6}$$

The resulting coupled problem is referred to as the β–monoscale-multimodel problem for the convection-diffusion equations. We should notice here that this methodology yields coupled problems through transmission boundary conditions, which are obtained at the modeling level of the physical phenomena. Therefore this method *cannot and should not* be classified as a domain decomposition method.

REMARK 2.1. *The global problem has no no-slip boundary condition. This suppresses the boundary layer which appears at low viscosity and facilitates the numerical solution of this problem. The boundary layers are modeled by the local problem* (2.5), (2.6) *which are to be solved only on a small domain, with a very fine discretization if needed. This clearly indicates that our method here is not a domain decomposition method since the objective is to be able to handle phenomena such as boundary layers at the local level.*

REMARK 2.2. *The "Neumann" version of this method consists of replacing in* (2.5) *the boundary condition on* Γ_i *with the corresponding Neumann condition*

$$\frac{\partial \phi_{loc}}{\partial n} = \frac{\partial \phi}{\partial n} \ \text{ on } \ \Gamma_i.$$

REMARK 2.3. *If in* (2.4) *we replace* Ω *with* Ω_E *defined*

$$\Omega_E = \Omega \backslash \Omega_l$$

(in this case $\partial\Omega_E = \Gamma_\infty \cup \Gamma_i$*), we obtain the* α*–monoscale-multimodel problem for convection-diffusion equations.*

The asymptotic behavior of problems (2.4)–(2.6) for large time is given in the following theorem.

THEOREM 2.1. *The solution of the coupled problems* (2.4)–(2.6) *converges towards the solution of the stationary problem* (2.3) *as* t *goes to infinity.*

For the proof of this theorem we refer to [2].

REMARK 2.4. *Using the same method as in* [2] *we can prove similar theorems for the variant of our method mentioned in Remarks* 2.2 *and* 2.3.

2.3. TTMA and the main result. The TTMA is an algorithm that iterates back and forth between the local and global problem. For the β–monoscale-multimodel problem (2.4)–(2.6) this corresponds to a time integration scheme which yields the following semi-implicit algorithm:

- set $\phi^0_{loc} = \phi_0$ and $\phi^0 = \phi_0$;
- then, for $n \geq 0, \phi^n_{loc}$ and ϕ^n being known, solve successively

$$\text{(2.7)} \qquad \begin{cases} \dfrac{\phi^{n+1}_{loc} - \phi^n_{loc}}{\Delta t} + \operatorname{div}(v\phi^{n+1}_{loc}) - \nu\Delta\phi^{n+1}_{loc} = 0 \ \text{ in } \ \Omega_l, \\ \phi^{n+1}_{loc} = \phi^n \ \text{ on } \ \Gamma_i, \\ \phi^{n+1}_{loc} = 0 \ \text{ on } \ \Gamma_b, \end{cases}$$

$$\text{(2.8)} \qquad \begin{cases} \dfrac{\phi^{n+1} - \phi^n}{\Delta t} + \operatorname{div}(v\phi^{n+1}) - \nu\Delta\phi^{n+1} = 0 \ \text{ in } \ \Omega, \\ \phi^{n+1} = \phi^\infty \ \text{ on } \ \Gamma_\infty, \\ \nu\dfrac{\partial\phi^{n+1}}{\partial n} = \nu\dfrac{\partial\phi^{n+1}_{loc}}{\partial n} \ \text{ on } \ \Gamma_b. \end{cases}$$

REMARK 2.5. *We have a full decoupling between* (2.7) *and* (2.8)*. They can be solved by two independent solution techniques.*

REMARK 2.6. *The "Neumann" version of this method consists of replacing the Dirichlet boundary condition on* Γ_i *with the corresponding Neumann condition*

$$\frac{\partial\phi^{n+1}_{loc}}{\partial n} = \frac{\partial\phi^n}{\partial n} \ \text{ on } \ \Gamma_i.$$

The convergence properties of the resulting algorithm can be obtained using the same method developed in this paper.

REMARK 2.7. *The fully implicit version of this method consists of replacing the condition*

$$\phi^{n+1}_{loc} = \phi^n \ \text{ on } \ \Gamma_i$$

(resp., $\frac{\partial\phi^{n+1}_{loc}}{\partial n} = \frac{\partial\phi^n}{\partial n}$ *on* Γ_i*) with the condition*

$$\phi^{n+1}_{loc} = \phi^{n+1} \ \text{ on } \ \Gamma_i$$

(resp., $\frac{\partial \phi_{loc}^{n+1}}{\partial n} = \frac{\partial \phi^{n+1}}{\partial n}$ *on* Γ_i*). Since both coupling conditions are implicit we proposed applying a fixed-point iteration to each time step. The resulting algorithm has been studied in* [2, 3] *using the local, global, and trace estimates developed by the author in* [11]. *The semi-implicit algorithm studied here leads directly to uncoupled problems in each time step; see Remark* 2.5. *The analysis of the resulting algorithm is more technical and relies on the methods we introduced in* [10, 12].

REMARK 2.8. *If in* (2.8) *we replace* Ω *with* Ω_E *defined as*

$$\Omega_E = \Omega \backslash \Omega_l$$

(in this case $\partial\Omega_E = \Gamma_\infty \cup \Gamma_i$*), we obtain the TTMA for the* α*–monoscale-multimodel problem. The convergence properties of the resulting algorithm can be obtained using similar methods as those developed in this paper.*

We state now the main result of this paper.

THEOREM 2.2. *Assume that* $\phi^{(0)} \in L^n(\Omega) \cup L^\infty(\Omega)$ *and* $\phi_l^{(0)} \in L^n(\Omega_l) \cup L^\infty(\Omega_l)$. *For* τ *sufficiently small and for all choices of* Ω_l*, the solution of* (2.7)–(2.8) *converges linearly in* $H^1(\Omega)$ *to the solution of the stationary problem* (2.3).

Before we prove this theorem, we shall give in the next subsection some important applications of the α- and β-methods in fluid mechanics.

2.4. Applications to fluid mechanics. We describe in this subsection some important applications of the α- and β-methods in fluid mechanics. More details about these applications can be found in [8, 9].

Let us consider the compressible Navier–Stokes equations which we formally write either as

$$\frac{\partial W}{\partial t} + \operatorname{div}[F(W)] = 0 \quad \text{on} \quad \Omega \text{ (conservative form)}$$

or as

$$\frac{\partial U}{\partial t} + T(U) + D(U) = 0 \quad \text{on} \quad \Omega \text{ (nonconservative form)},$$

with $W = (\rho, \rho v, \rho E)$ and $U = (\rho, v, \theta)$ the conservative and nonconservative variables, $F = F_C + F_D$ the total flux (convective and viscous part), T and D the convective and viscous terms in the nonconservative form of the Navier–Stokes equations. The problem consists of computing a steady solution of these equations, with boundary conditions

$$\begin{gathered} \rho v, \rho E \text{ given on } \Gamma_\infty \text{ (exterior limit of the domain)}, \\ \rho \text{ given on } \Gamma_\infty \cap \{x, v(x) \cdot n \le 0\} \text{ (inflow)}, \\ v = 0 \text{ on the body } \Gamma_b \text{ (no-slip)}, \\ \theta = \theta_0 \text{ on the body } \Gamma_b. \end{gathered}$$

The global numerical treatment of these equations faces the following difficulties:

(i) In a conservative calculation, the numerical viscosity of the discretization scheme interferes with the physical viscosity and for a mesh of reasonable size leads to an overprediction of the boundary layer. Moreover, no-slip boundary conditions on the body are difficult to handle for many TVD schemes.

(ii) In a nonconservative calculation, the correct calculation of a shock requires locally a very fine grid if we want to satisfy the Rankine–Hugoniot conditions.

In this framework, our strategy is to couple a *global conservative scheme*, defined on the whole domain and based, for example, on a finite volume space discretization [1], and *a local approximation*, defined in the neighborhood of the body and based, for example, on a mixed finite element approximation of the nonconservative Navier–Stokes equations [6].

The coupling problem corresponds then to solving the following systems. In Ω, we solve the conservative Navier–Stokes equations

$$\frac{\partial W}{\partial t} + \operatorname{div}[F(W)] = 0 \text{ in } \Omega,$$
$$F(W)\cdot n = \begin{bmatrix} 0 \\ n\cdot\sigma(W)\cdot n \\ \tau\cdot\sigma(U_{loc})\cdot n \\ -q(U_{loc})\cdot n \end{bmatrix} \quad \text{on the wall,}$$
$$W = \text{given imposed value on } \Gamma_\infty.$$

In Ω_l, we solve the nonconservative Navier–Stokes equations

$$\frac{\partial U}{\partial t} + T(U) + D(U) = 0 \text{ in } \Omega,$$
$$U_{loc} = 0 \text{ on } \Gamma_b,$$
$$U_{loc} = W \text{ on } \Gamma_i.$$

Above, $n.\sigma.n$ and $\tau.\sigma.n$, respectively, denote the normal and the tangential forces exerted by the body on the flow, with n the unit normal vector to the wall oriented towards its interior. Notice that in the global conservative problem the matching conditions are of Neumann type as in (2.4), while for the local nonconservative problem these matching boundary conditions are of Dirichlet type as in (2.8) (but with an explicit boundary condition on Γ_i). From this, we see how it is possible to generalize algorithms (2.7)–(2.8) to a more complex system, such as the Navier–Stokes equations.

This coupling provides an efficient strategy to circumvent the difficulties mentioned in (i) and (ii) at the beginning of this subsection. These methods offer an easy way of supplementing and testing a large variety of boundary conditions. Moreover, in the local domain the model can be chosen to handle phenomena such as boundary layers and turbulence and can incorporate models of chemistry. We have already proved the validity of these methods and their superiority to classical methods in many (important) real life applications. More details about this type of coupling can be found in [8, 9].

3. Preliminary results. In this section we state local, global, and trace estimates for the solutions of elliptic equations. They are obtained by the author in [11, 12]. They play an important role in the proof of the main result of this paper. The first estimates are obtained for the solution of the following Dirichlet–Neumann problem:

$$\mathcal{L}u = -\nu\Delta u + v\cdot\nabla u + \frac{1}{\tau}u = \frac{1}{\tau}f \text{ in } \Omega, \tag{3.1}$$
$$u = 0 \text{ on } \Gamma_\infty, \tag{3.2}$$
$$\frac{\partial u}{\partial n} = g \text{ on } \Gamma_b, \tag{3.3}$$

where the function g is given in $H^{-1/2}(\Gamma_b)$, the coefficient τ is strictly positive, and ν is the diffusion coefficient. We assume that $f \in L^n(\Omega) \cap L^\infty(\Omega)$ and the coefficients ν and τ satisfy the relation

$$\nu\tau \le \frac{1}{2} \text{ and } \tau \le 1. \tag{3.4}$$

This hypothesis is neither necessary nor restrictive (see [11, 12]). Let $d > 0$ denote the distance between Γ_b and Γ_i. Let β be a real number such that $0 < \beta < 3\sqrt{\nu}/d$, and set $k = \beta/(\nu\sqrt{\tau})$. In the proof of Theorem 2.2 no additional restrictions on β were required. Thus, the inequality $d < 3\sqrt{\nu}/\beta$ relating the local domain Ω_l to the viscosity ν shows that all choices of Ω_l are possible. For small viscosity the domain Ω_l can be chosen small.

We have the following global H^1 estimate of the solution u of the Dirichlet–Neumann problem (3.1)–(3.3) in terms of the boundary data g and the data f.

LEMMA 3.1. *There exists a constant c_0 such that*

$$\|u\|_{1,\Omega} \le c_0\|g\|_{-1/2,\Gamma_b} + \frac{1}{\nu\tau}\|f\|_{0,\Omega}. \tag{3.5}$$

We now state the trace estimate.

THEOREM 3.1. *The solution u of the Dirichlet–Neumann problem* (3.1)–(3.3) *satisfies*

$$\|u\|_{1/2,\Gamma_i} \le C_1\sqrt{d}\left(d + \frac{\|v\|_\infty}{\nu}\right)^{1/2} \exp\left(-kd^2/36\right) \times \left[\|g\|_{-1/2,\Gamma_b} + \frac{1}{\nu\tau}\|f\|_{0,\Omega} + \frac{d}{\nu\tau}\|f\|_{L^n(\Omega)} + 2\|f\|_{\infty,\Omega}\right] + \frac{C_2}{\tau\sqrt{\nu}}\|f\|_{0,\Omega},$$

where C_1 and C_2 are constants, with C_1 depending only on n and $(\|V\|_\infty d/\nu)^2$ but not on τ.

We now consider the following Dirichlet–Dirichlet problem:

$$\mathcal{L}u_l = -\nu\Delta u_l + v \cdot \nabla u_l + \frac{1}{\tau}u_l = \frac{1}{\tau}f_l \text{ in } \Omega_l, \tag{3.6}$$

$$u_l = h \text{ on } \Gamma_i, \tag{3.7}$$

$$u_l = 0 \text{ on } \Gamma_b, \tag{3.8}$$

where the function h is given in $H^{1/2}(\Gamma_i)$, the coefficient τ is strictly positive, and ν is the diffusion coefficient. We assume that $f_l \in L^n(\Omega_l) \cap L^\infty(\Omega_l)$. The velocity field v is given by (2.1). Let Γ_V be the center surface of Ω_l defined as the surface whose distance from Γ_b and Γ_i is at least $d/2$. Let Ω_{il} be the subdomain of Ω_l of width $d/6$ centered at Γ_V (see Figure 3.1).

We have the following global and local estimate of the solution u_l of the Dirichlet–Dirichlet problem (3.6)–(3.8).

LEMMA 3.2. *The solution u_l of the Dirichlet–Dirichlet problem* (3.6)–(3.8) *satisfies*

$$\|u_l\|_{1,\Omega_l} \le \left[1 + \frac{c_1}{\nu}\|v\|_{\infty,\Omega_l} + \frac{1}{\nu\tau}\right]\|h\|_{1/2,\Gamma_i} + \frac{c_1}{\nu\tau}\|f_l\|_{0,\Omega_l}. \tag{3.9}$$

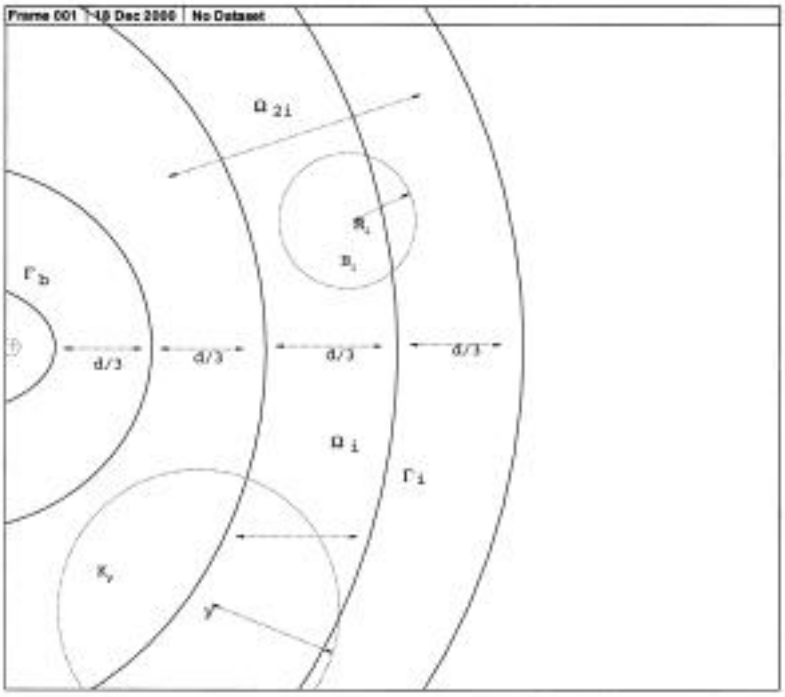


FIG. 3.1. *Description of the domain Ω_l and the splitting used in the majorization of the local solution.*

LEMMA 3.3. *There exists a constant c_2 such that*

$$\|u_l\|_{\infty,\Omega_{il}} \le c_2\|u_l\|_{0,\Omega_l} + c_2\frac{d}{\nu}\|f_l\|_{L^n(\Omega_l)}, \tag{3.10}$$

where c_2 depends only on n and $(\|V\|_\infty d/\nu)^2$.

We now state the trace estimate.

THEOREM 3.2. *The solution u_l of the Dirichlet–Dirichlet problem* (3.6)–(3.8) *satisfies*

$$\begin{aligned}\|\partial u_l/\partial n\|_{-1/2,\Gamma_b} \le{}& C_1\alpha_1^2\alpha_2\exp(-kd^2/36)\|h\|_{1/2,\Gamma_i}\\ &+ C_1\alpha_1\alpha_2\exp(-kd^2/36)\frac{d}{\nu\tau}\|f_l\|_{L^n(\Omega_l)}\\ &+ \alpha_1\left(C_1\alpha_2\exp(-kd^2/36)\frac{1}{\nu\tau} + C_2\frac{1}{\tau\sqrt{\nu}}\right)\|f_l\|_{0,\Omega_l}\\ &+ C_1\alpha_1\alpha_2\exp(-kd^2/36)\|f_l\|_{\infty,\Omega_{il}},\end{aligned}$$

where C_1 and C_2 are constants with C_1 depending only on n and $(n\|v\|_\infty d/\nu)^2$, $\alpha_1 = [1 + \frac{1}{\nu}\|v\|_{\infty,\Omega_l} + \frac{1}{\nu\tau}]$, and $\alpha_2 = \sqrt{d}(d + \frac{\|v\|_\infty}{\nu})^{1/2}$.

For the proof of these local, global, and trace estimates we refer to [12].

4. Proof of the main result. The proof of this theorem is based in a crucial manner on the method we have introduced in [10] and the local, global, and trace

estimates we developed in [12]. This proof is divided into four steps. In the first step we use the method we introduced in [10]. We then obtain a newly transformed problem. In the second step we give estimates of the boundary terms using the local, global, and trace estimates of the previous section. In the third step we study the transformed problem. In the last step we conclude the proof of our theorem.

Step 1. Use of the method of [10]. Without loss of generality, we may assume that $\phi_\infty = 0$ on Γ_∞. Setting $u = \phi^{n+1}$, $u_l = \phi_{loc}^{n+1}$, $f = \phi^n$, $f_l = \phi_{loc}^n$, $f_{lm} = \phi_{loc}^{n-1}$, $\tau = \Delta t$, and $a = \frac{1}{\tau}$ the algorithm in system (2.7)–(2.8) becomes the following:

For $n \geq 0$, f_l and f being known, solve

$$\begin{cases} a(u-f) + \operatorname{div}(vu) - \nu\Delta u = 0 \quad \text{in} \quad \Omega, \\ u = 0 \quad \text{on} \quad \Gamma_\infty, \\ \nu\dfrac{\partial u}{\partial n} = \nu\dfrac{\partial u_l}{\partial n} \quad \text{on} \quad \Gamma_b, \end{cases} \tag{4.1}$$

$$\begin{cases} a(u_l - f_l) + \operatorname{div}(vu_l) - \nu\Delta u_l = 0 \quad \text{in} \quad \Omega_l, \\ u_l = f \quad \text{on} \quad \Gamma_i, \\ u_l = 0 \quad \text{on} \quad \Gamma_b. \end{cases} \tag{4.2}$$

Let φ_1 and φ_2 be two positive functions defined, respectively, on Ω, Ω_l, to be precisely determined later. Multiplying the first equation in (4.1) by $\varphi_1 u$ and using Green's formula, we obtain

$$\int a(u-f)\varphi_1 u + \int \operatorname{div}(vu)\varphi_1 u - \nu\int \Delta u\varphi_1 u = 0. \tag{4.3}$$

Since the vector field v is given by (2.1), we have, using Green's formula,

$$\int \operatorname{div}(vu)\varphi_1 u = -\frac{1}{2}\int u^2 \nabla\varphi_1 \cdot v + \frac{1}{2}\int_\Gamma v\cdot n\varphi_1 u^2, \tag{4.4}$$

where $\Gamma = \partial\Omega$. We also have, using Green's formula,

$$\int \Delta u\varphi_1 u = -\int \nabla u\nabla\varphi_1 u - \int \varphi_1|\nabla u|^2 + \int_\Gamma \frac{\partial u}{\partial n}\varphi_1 u. \tag{4.5}$$

Combining (4.3), (4.4), and (4.5), we obtain

$$\begin{aligned} &\int \left(a\varphi_1 - \frac{1}{2}\nabla\varphi_1\cdot v\right)u^2 + \nu\int u\nabla u\cdot\nabla\varphi_1 + \nu\int \varphi_1|\nabla u|^2 \\ &+ \frac{1}{2}\int_\Gamma v\cdot n\varphi_1 u^2 - \nu\int_\Gamma \frac{\partial u}{\partial n}\varphi_1 u = a\int f\varphi_1 u. \end{aligned} \tag{4.6}$$

Similarly, we obtain

$$\begin{aligned} &\int \left(a\varphi_2 - \frac{1}{2}\nabla\varphi_2\cdot v\right)u_l^2 + \nu\int u_l\nabla u_l\cdot\nabla\varphi_2 + \int \varphi_2|\nabla u_l|^2 \\ &+ \frac{1}{2}\int_{\Gamma_l} v\cdot n\varphi_2 u_l^2 - \nu\int_{\Gamma_l}\frac{\partial u_l}{\partial n}\varphi_2 u_l = a\int f_l\varphi_2 u_l, \end{aligned} \tag{4.7}$$

where $\Gamma_l = \partial\Omega_l$.

Combining (4.6) and (4.7), we obtain

$$\begin{aligned}
&\int \left(a\varphi_1 - \frac{1}{2}\nabla\varphi_1 \cdot v\right) u^2 + \int \left(a\varphi_2 - \frac{1}{2}\nabla\varphi_2 \cdot v\right) u_l^2 \\
&+ \nu \int u\nabla u \cdot \nabla\varphi_1 + \nu \int \varphi_1 |\nabla u|^2 + \nu \int u_l \nabla u_l \cdot \nabla\varphi_2 + \nu \int \varphi_2 |\nabla u_l|^2 \\
&+ \frac{1}{2}\int_\Gamma v \cdot n\varphi_1 u^2 - \nu \int_\Gamma \frac{\partial u}{\partial n}\varphi_1 u + \frac{1}{2}\int_{\Gamma_l} v \cdot n\varphi_2 u_l^2 - \nu \int_{\Gamma_l} \frac{\partial u_l}{\partial n}\varphi_2 u_l \\
&= a\int f\varphi_1 u + a\int f_l \varphi_2 u_l. \qquad (4.8)
\end{aligned}$$

Using the coupling boundary conditions, we obtain

$$\begin{aligned}
&\frac{1}{2}\int_\Gamma v \cdot n\varphi_1 u^2 - \nu \int_\Gamma \frac{\partial u}{\partial n}\varphi_1 u + \frac{1}{2}\int_{\Gamma_l} v \cdot n\varphi_2 u_l^2 - \nu \int_{\Gamma_l} \frac{\partial u_l}{\partial n}\varphi_2 u_l \\
&\quad = \frac{1}{2}\int_{\Gamma_b} v \cdot n\varphi_1 u^2 - \nu \int_{\Gamma_b} \frac{\partial u}{\partial n}\varphi_1 u + \frac{1}{2}\int_{\Gamma_i} v \cdot n\varphi_2 u_l^2 - \nu \int_{\Gamma_i} \frac{\partial u_l}{\partial n}\varphi_2 u_l \\
&\quad = -\nu \int_{\Gamma_b} \frac{\partial u_l}{\partial n}\varphi_1 u + \frac{1}{2}\int_{\Gamma_i} v \cdot n\varphi_2 f^2 - \nu \int_{\Gamma_i} \frac{\partial u_l}{\partial n}\varphi_2 f \\
&\quad = BC_1 + BC_2 + BC_3. \qquad (4.9)
\end{aligned}$$

Combining (4.8) and (4.9) and using Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&\int \left(\frac{a}{2}\varphi_1 - \frac{1}{2}\nabla\varphi_1 \cdot v\right) u^2 + \int \left(\frac{a}{2}\varphi_2 - \frac{1}{2}\nabla\varphi_2 \cdot v\right) u_l^2 + \nu \int u\nabla u \cdot \nabla\varphi_1 \\
&+ \nu \int \varphi_1 |\nabla u|^2 + \nu \int u_l \nabla u_l \cdot \nabla\varphi_2 + \nu \int \varphi_2 |\nabla u_l|^2 + \Sigma BC_i \\
&\leq \frac{a}{2}\int \varphi_1 f^2 + \frac{a}{2}\int \varphi_2 f_l^2. \qquad (4.10)
\end{aligned}$$

Using Green's formula we obtain

$$\int \nabla u \cdot \nabla\varphi_1 u = -\frac{1}{2}\int u^2 \Delta\varphi_1 + \frac{1}{2}\int_\Gamma u^2 \frac{\partial\varphi_1}{\partial n}. \qquad (4.11)$$

Similarly, we obtain

$$\int \nabla u_l \cdot \nabla\varphi_1 u_l = -\frac{1}{2}\int u_l^2 \Delta\varphi_1 + \frac{1}{2}\int_\Gamma u_l^2 \frac{\partial\varphi_1}{\partial n}. \qquad (4.12)$$

Equation (4.10) then becomes

$$\begin{aligned}
&\int \left(\frac{a}{2}\varphi_1 - \frac{1}{2}\nabla\varphi_1 \cdot v\right) u^2 + \int \left(\frac{a}{2}\varphi_2 - \frac{1}{2}\nabla\varphi_2 \cdot v\right) u_l^2 + \nu \int |\nabla u|^2 \varphi_1 \\
&- \frac{\nu}{2}\int u^2 \Delta\varphi_1 + \frac{\nu}{2}\int_\Gamma u^2 \frac{\partial\varphi_1}{\partial n} + \nu \int |\nabla u_l|^2 \varphi_2 \\
&- \frac{\nu}{2}\int u_l^2 \Delta\varphi_2 + \frac{\nu}{2}\int_{\Gamma_l} u_l^2 \frac{\partial\varphi_2}{\partial n} + \Sigma BC_i \\
&\leq \frac{a}{2}\int \varphi_1 f^2 + \frac{a}{2}\int \varphi_2 f_l^2. \qquad (4.13)
\end{aligned}$$

Using the boundary conditions, (4.13) becomes

$$\begin{aligned}&\frac{1}{2}\int(-\nu\Delta\varphi_1 - v\cdot\nabla\varphi_1 + a\varphi_1)u^2 + \frac{1}{2}\int(-\nu\Delta\varphi_2 - v\cdot\nabla\varphi_2 + a\varphi_2)u_l^2\\ &+\frac{\nu}{2}\int_{\Gamma_b} u^2\frac{\partial\varphi_1}{\partial n} + \frac{\nu}{2}\int_{\Gamma_i} u_l^2\frac{\partial\varphi_2}{\partial n} + \nu\int|\nabla u|^2\varphi_1 + \nu\int|\nabla u_l|^2\varphi_2 + \Sigma BC_i\\ &\qquad\le \frac{a}{2}\int\varphi_1 f^2 + \frac{a}{2}\int\varphi_2 f_l^2.\end{aligned}\tag{4.14}$$

Step 2. Use of the local, global, and trace estimates of [11, 12]. We shall give here estimates of the boundary terms in (4.8). These estimates are based on the local, global, and trace estimates we obtained in [11, 12].

Estimate of BC_1. Using the coupling boundary conditions and Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}|BC_1| &= \nu\left|\int_{\Gamma_b}\frac{\partial u_l}{\partial n}\varphi_1 u\right|\\ &\le \nu\|\varphi_1|_{\Gamma_b}\|_\infty\|u\|_{\frac{1}{2},\Gamma_b}\left\|\frac{\partial u_l}{\partial n}\right\|_{-\frac{1}{2},\Gamma_b}\\ &\le \nu\|\varphi_1|_{\Gamma_b}\|_\infty\left(\|u\|^2_{\frac{1}{2},\Gamma_b} + \left\|\frac{\partial u_l}{\partial n}\right\|^2_{-\frac{1}{2},\Gamma_b}\right).\end{aligned}$$

Using the trace theorem and Lemma 3.1, we obtain

$$\begin{aligned}\|u\|_{1/2,\Gamma_b} &\le C(\Omega)\|u\|_{1,\Omega}\\ &\le c_0 C(\Omega)\left(\|g\|_{-1/2,\Gamma_b} + \frac{1}{\nu\tau}\|f\|_{0,\Omega}\right),\end{aligned}$$

where $g = \frac{\partial u}{\partial n}$. Using now the coupling boundary conditions in (4.1) and Theorem 3.2, we obtain

$$\begin{aligned}\|g\|_{-1/2,\Gamma_b} &= \|\partial u_l/\partial n\|_{-1/2,\Gamma_b}\\ &\le \beta_1\|h\|_{1/2,\Gamma_i} + \beta_2\|f_l\|_{0,\Omega_l} + \beta_3\|f_l\|_{L^n(\Omega_l)} + \beta_4\|f_l\|_{\infty,\Omega_{il}},\end{aligned}$$

where

$$\begin{aligned}\beta_1 &= C_1\alpha_1^2\alpha_2\exp(-kd^2/36),\\ \beta_2 &= C_1\alpha_1\alpha_2\exp(-kd^2/36)\frac{1}{\nu\tau} + C_2\alpha_1\frac{1}{\tau\sqrt{\nu}},\\ \beta_3 &= C_1\alpha_1\alpha_2\exp(-kd^2/36)\frac{d}{\nu\tau},\\ \beta_4 &= C_1\alpha_1\alpha_2\exp(-kd^2/36),\\ \alpha_1 &= 1 + \frac{1}{\nu}\|v\|_{\infty,\Omega_l} + \frac{1}{\nu\tau},\\ \alpha_2 &= \sqrt{d}\left(d + \frac{\|v\|_\infty}{\nu}\right)^{1/2}.\end{aligned}$$

Using now the fact that $h = u_l = f$ on Γ_i and the trace theorem, we obtain

$$\begin{aligned}\|\partial u_l/\partial n\|_{-1/2,\Gamma_b} &\le \beta_1\|f\|_{1/2,\Gamma_i} + \beta_2\|f_l\|_0 + \beta_3\|f_l\|_{L^n(\Omega_l)} + \beta_4\|f_l\|_{\infty,\Omega_i}\\ &\le C(\Omega)\beta_1\|f\|_{1,\Omega} + \beta_2\|f_l\|_{0,\Omega_l} + \beta_3\|f_l\|_{L^n(\Omega_l)} + \beta_4\|f_l\|_{\infty,\Omega_{il}}.\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
|BC_1| &\le \nu\|\varphi_1|_{\Gamma_b}\|_\infty \left(\|u\|^2_{1/2,\Gamma_b} + \left\| \frac{\partial u_l}{\partial n} \right\|^2_{-1/2,\Gamma_b} \right) \\
&\le C(\Omega)\nu\|\varphi_1|_{\Gamma_b}\|_\infty \left(\|u\|^2_{1,\Omega} + \left\| \frac{\partial u_l}{\partial n} \right\|^2_{-1/2,\Gamma_b} \right) \\
&\le c_0 C(\Omega)\nu\|\varphi_1|_{\Gamma_b}\|_\infty \left(\|g\|^2_{-1/2,\Gamma_b} + \frac{1}{(\nu\tau)^2}\|f\|^2_{0,\Omega} + \left\| \frac{\partial u_l}{\partial n} \right\|^2_{-1/2,\Gamma_b} \right) \\
&\le C_{\varphi_1}\beta_1^2\|f\|^2_{1,\Omega} + C_{\varphi_1}\beta_2^2\|f_l\|^2_{0,\Omega_l} + C_{\varphi_1}\beta_3^2\|f_l\|^2_{L^n(\Omega_l)} + C_{\varphi_1}\beta_4^2\|f_l\|^2_{\infty,\Omega_{il}} \\
&\quad + C_{\varphi_1}\frac{1}{(\nu\tau)^2}\|f\|^2_{0,\Omega}, \qquad (4.15)
\end{aligned}
$$

where

$$
C_{\varphi_1} = c_0 C(\Omega)\nu\|\varphi_1|_{\Gamma_b}\|_\infty. \qquad (4.16)
$$

Estimate of BC_2. For the term BC_2 we have, using the fact that $u_l = f$ on Γ_i and the trace theorem,

$$
\begin{aligned}
|BC_2| &= \frac{1}{2}\left| \int_{\Gamma_i} v\cdot n\varphi_2 f^2 \right| \\
&\le \frac{1}{2}\|v\|_\infty\|\varphi_2|_{\Gamma_i}\|_\infty\|f\|^2_{\frac{1}{2},\Gamma_i} \\
&\le C(\Omega)\|v\|_\infty\|\varphi_2|_{\Gamma_i}\|_\infty\|f\|^2_{1,\Omega} \\
&\le C_v\|f\|^2_{1,\Omega}, \qquad (4.17)
\end{aligned}
$$

where

$$
C_v = C(\Omega)\|v\|_\infty\|\varphi_2|_{\Gamma_i}\|_\infty.
$$

Estimate of BC_3. Finally to get an estimate of BC_3 we proceed as follows:

$$
\begin{aligned}
|BC_3| &= \nu\left| \int_{\Gamma_i} \frac{\partial u_l}{\partial n}\varphi_2 f \right| \\
&\le \nu\|\varphi_2|_{\Gamma_i}\|_\infty\|f\|_{\frac{1}{2},\Gamma_i}\left\| \frac{\partial u_l}{\partial n} \right\|_{-\frac{1}{2},\Gamma_i}.
\end{aligned}
$$

The term $\|\partial u_l/\partial n\|_{-1/2,\Gamma_i}$ is estimated as follows. Using the weak formulation of problem (4.2), (2.1), and the boundary conditions in (4.2), we obtain

$$
\int_{\Gamma_i} \frac{\partial u_l}{\partial n} w = \int_{\Omega_l} \nabla u_l \nabla w + \frac{1}{\nu}\int_{\Omega_l} wv\cdot\nabla u_l - \frac{1}{\nu\tau}\int_{\Omega_l} f_l w + \frac{1}{\nu\tau}\int_{\Omega_l} u_l w,
$$

where $w \in H^1(\Omega_l)$ with $w = 0$ on Γ_b. Using the trace theorem and (2.1), we obtain

$$
\left| \int_{\Gamma_i} \frac{\partial u_l}{\partial n} w \right| \le \left(\left(1 + \frac{1}{\nu}\|v\|_{\infty,\Omega_l} \right)\|\nabla u_l\|_{0,\Omega_l} + \frac{1}{\nu\tau}\|f_l\|_{0,\Omega_l} + \frac{1}{\nu\tau}\|u_l\|_{0,\Omega_l} \right)\|w\|_{1,\Omega_l}.
$$

Therefore, we have

$$\left\|\frac{\partial u_l}{\partial n}\right\|_{-1/2,\Gamma_i} \le \left(1+\frac{1}{\nu}\|v\|_{\infty,\Omega_l}\right)\|\nabla u_l\|_{0,\Omega_l} + \frac{1}{\nu\tau}\|u_l\|_{0,\Omega_l} + \frac{1}{\nu\tau}\|f_l\|_{0,\Omega_l}$$

$$\le \alpha_1\|u_l\|_{1,\Omega_l} + \frac{1}{\nu\tau}\|f_l\|_{0,\Omega_l}. \tag{4.18}$$

Moreover, using Lemma 3.2 and the trace theorem, we obtain

$$\begin{aligned}\|u_l\|_{1,\Omega_l} &\le c_1\alpha_1\|h\|_{1/2,\Gamma_i} + \frac{c_1}{\nu\tau}\|f_l\|_{0,\Omega_l}\\ &\le c_1\alpha_1\|f\|_{1/2,\Gamma_i} + \frac{c_1}{\nu\tau}\|f_l\|_{0,\Omega_l}\\ &\le C(\Omega)c_1\alpha_1\|f\|_{1,\Omega} + \frac{c_1}{\nu\tau}\|f_l\|_{0,\Omega_l}.\end{aligned}$$

Hence we obtain

$$\begin{aligned}|BC_3| &\le C(\Omega)\nu\|\varphi_2|_{\Gamma_i}\|_\infty\alpha_1(\|u_l\|_{1,\Omega_l} + \|f_l\|_{0,\Omega_l})\|f\|_{1,\Omega}\\ &\le C(\Omega)\nu\|\varphi_2|_{\Gamma_i}\|_\infty\alpha_1^2c_1(\|f\|_{1,\Omega}^2 + \|f_l\|_{0,\Omega_l}^2)\\ &\le \alpha_3(\|f\|_{1,\Omega}^2 + \|f_l\|_{0,\Omega_l}^2),\end{aligned} \tag{4.19}$$

where

$$\alpha_3 = C(\Omega)\nu\|\varphi_2|_{\Gamma_i}\|_\infty\alpha_1^2c_1.$$

Estimate of $\Sigma|BC_i|$. Combining (4.15), (4.17), and (4.19), we obtain

$$\Sigma|BC_i| \le (C_{\varphi_1}\beta_1^2 + C_v + \alpha_3)\|f\|_{1,\Omega}^2 + (C_{\varphi_1}\beta_2^2 + \alpha_3)\|f_l\|_{0,\Omega_l}^2 + C_{\varphi_1}\beta_3^2\|f_l\|_{L^n(\Omega_l)}^2$$

$$+ C_{\varphi_1}\beta_4^2\|f_l\|_{\infty,\Omega_{il}}^2 + C_{\varphi_1}\frac{1}{(\nu\tau)^2}\|f\|_{0,\Omega}^2. \tag{4.20}$$

To proceed further we need to give an estimate of $\|f_l\|_{L^n(\Omega_l)}$ and $\|f_l\|_{\infty,\Omega_{il}}$. We use the Gagliardo–Nirenberg interpolation inequality (see, for example, [4] and references therein); we obtain for $n \ge 2$

$$\|f_l\|_{L^n(\Omega_l)} \le \epsilon_1\|f_l\|_{1,\Omega_l} + C_{\epsilon_1}\|f_l\|_{0,\Omega_l}. \tag{4.21}$$

Using now Lemma 3.3 and (4.21), we obtain

$$\begin{aligned}\|f_l\|_{\infty,\Omega_{il}} &\le c_2\|f_l\|_{0,\Omega_l} + c_2\frac{d}{\nu\tau}\|f_{lm}\|_{L^n(\Omega_l)}\\ &\le c_2\|f_l\|_{0,\Omega_l} + c_2\frac{d}{\nu\tau}(\epsilon_1\|f_{lm}\|_{1,\Omega_l} + C_{\epsilon_1}\|f_{lm}\|_{0,\Omega_l}).\end{aligned} \tag{4.22}$$

Combining (4.20), (4.21), and (4.22) we obtain

$$
\begin{aligned}
\Sigma|BC_i| &\le (C_{\varphi_1}\beta_1^2 + C_v + \alpha_3)\|f\|_{1,\Omega}^2 + (C_{\varphi_1}\beta_2^2 + \alpha_3)\|f_l\|_{0,\Omega_l}^2 \\
&\quad + C_{\varphi_1}\beta_3^2({\epsilon_1}^2\|f_l\|_{1,\Omega_l}^2 + C_{\epsilon_1}\|f_l\|_{0,\Omega_l}^2) \\
&\quad + c_2 C_{\varphi_1}\beta_4^2\left(\|f_l\|_{0,\Omega_l}^2 + \left(\frac{d}{\nu\tau}\right)^2({\epsilon_1}^2\|f_{lm}\|_{1,\Omega_l}^2 + C_{\epsilon_1}\|f_{lm}\|_{0,\Omega_l}^2)\right) \\
&\quad + C_{\varphi_1}\frac{1}{(\nu\tau)^2}\|f\|_{0,\Omega}^2 \\
&\le (C_{\varphi_1}\beta_1^2 + C_v + \alpha_3)\|\nabla f\|_{0,\Omega}^2 + \left[(C_{\varphi_1}\beta_1^2 + C_v + \alpha_3) + C_{\varphi_1}\frac{1}{(\nu\tau)^2}\right]\|f\|_{0,\Omega}^2 \\
&\quad + C_{\varphi_1}\beta_3^2{\epsilon_1}^2\|\nabla f_l\|_{0,\Omega_l}^2 + [C_{\varphi_1}\beta_3^2({\epsilon_1}^2 + C_{\epsilon_1}) + (C_{\varphi_1}\beta_2^2 + \alpha_3) + \alpha_2 C_{\varphi_1}\beta_4^2]\|f_l\|_{0,\Omega_l}^2 \\
&\quad + c_2 C_{\varphi_1}\beta_4^2\left(\frac{d}{\nu\tau}\right)^2{\epsilon_1}^2\|\nabla f_{lm}\|_{0,\Omega_l}^2 + c_2 C_{\varphi_1}\beta_4^2\left(\frac{d}{\nu\tau}\right)^2(\epsilon_1^2 + C_{\epsilon_1})\|f_{lm}\|_{0,\Omega_l}^2 \\
&\le \delta_1\|\nabla f\|_{0,\Omega}^2 + \delta_1'\|f\|_{0,\Omega}^2 + \delta_2\|\nabla f_l\|_{0,\Omega_l}^2 + \delta_3\|f_l\|_{0,\Omega_l}^2 + \delta_4\|f_{lm}\|_{0,\Omega_l}^2 \\
&\quad + \delta_5\|\nabla f_{lm}\|_{0,\Omega_l}^2, \qquad (4.23)
\end{aligned}
$$

where

$$
\begin{aligned}
\delta_1 &= C_{\varphi_1}\beta_1^2 + C_v + \alpha_3, \\
\delta_1' &= (C_{\varphi_1}\beta_1^2 + C_v + \alpha_3) + C_{\varphi_1}\frac{1}{(\nu\tau)^2}, \\
\delta_2 &= C_{\varphi_1}\beta_3^2{\epsilon_1}^2, \\
\delta_3 &= C_{\varphi_1}\beta_3^2({\epsilon_1}^2 + C_{\epsilon_1}) + (C_{\varphi_1}\beta_2^2 + \alpha_3) + \alpha_2 C_{\varphi_1}\beta_4^2, \\
\delta_4 &= c_2 C_{\varphi_1}\beta_4^2\left(\frac{d}{\nu\tau}\right)^2(\epsilon_1^2 + C_{\epsilon_1}), \\
\delta_5 &= c_2 C_{\varphi_1}\beta_4^2\left(\frac{d}{\nu\tau}\right)^2{\epsilon_1}^2. \qquad (4.24)
\end{aligned}
$$

Step 3. Study of the transformed problem and use of the method of [10]. Using (4.14) we obtain

$$
\begin{aligned}
&\frac{1}{2}\int(-\nu\Delta\varphi_1 - v\cdot\nabla\varphi_1 + a\varphi_1)u^2 + \frac{1}{2}\int(-\nu\Delta\varphi_2 - v\cdot\nabla\varphi_2 + a\varphi_2)u_l^2 \\
&+ \frac{\nu}{2}\int_{\Gamma_b} u^2\frac{\partial\varphi_1}{\partial n} + \frac{\nu}{2}\int_{\Gamma_i} u_l^2\frac{\partial\varphi_2}{\partial n} + \nu\int|\nabla u|^2\varphi_1 + \nu\int|\nabla u_l|^2\varphi_2 \\
&\qquad \le \frac{a}{2}\int\varphi_1 f^2 + \frac{a}{2}\int\varphi_2 f_l^2 + \Sigma|BC_i|. \qquad (4.25)
\end{aligned}
$$

Combining (4.23) and (4.25), we obtain

$$
\begin{aligned}
&\frac{1}{2}\int(-\nu\Delta\varphi_1 - v\cdot\nabla\varphi_1 + a\varphi_1)u^2 + \frac{1}{2}\int(-\nu\Delta\varphi_2 - v\cdot\nabla\varphi_2 + a\varphi_2)u_l^2 \\
&+\frac{\nu}{2}\int_{\Gamma_b} u^2\frac{\partial\varphi_1}{\partial n} + \frac{\nu}{2}\int_{\Gamma_i} u_l^2\frac{\partial\varphi_2}{\partial n} + \nu\int|\nabla u|^2\varphi_1 + \nu\int|\nabla u_l|^2\varphi_2 \\
&\qquad \le \int_\Omega\left(\frac{a}{2}\varphi_1 + \delta_1'\right)f^2 + \int_{\Omega_l}\left(\frac{a}{2}\varphi_2 + \delta_3\right)f_l^2 + \int_\Omega\delta_1|\nabla f|^2 \\
&\qquad\quad + \int_{\Omega_l}\delta_2|\nabla f_l|^2 + \int_{\Omega_l}\delta_4 f_{lm}^2 + \int_{\Omega_l}\delta_5|\nabla f_{lm}|^2. \qquad (4.26)
\end{aligned}
$$

We shall now construct φ_1 and φ_2 positive bounded below and above by positive constants such that we have

$$\begin{aligned}\frac{1}{2}(-\nu\Delta\varphi_1 - v\cdot\nabla\varphi_1 + a\varphi_1) &= \left(\frac{a}{2}\varphi_1 + \delta_1'\right)(1+\epsilon_2) + \delta_1'',\\ \frac{1}{2}(-\nu\Delta\varphi_2 - v\cdot\nabla\varphi_2 + a\varphi_2) &= \left(\frac{a}{2}\varphi_2 + \delta_3\right)(1+\epsilon_2) + \delta_3'',\end{aligned} \tag{4.27}$$

where ϵ_2, δ_1'', and δ_3'' are positive constants. This corresponds to

$$-\nu\Delta\varphi_1 - v\cdot\nabla\varphi_1 - a\epsilon_2\varphi_1 = 2\delta_1'(1+\epsilon_2) + 2\delta_1'', \tag{4.28}$$

$$-\nu\Delta\varphi_2 - v\cdot\nabla\varphi_2 - a\epsilon_2\varphi_2 = 2\delta_3(1+\epsilon_2) + 2\delta_3''. \tag{4.29}$$

For ϵ_2 small, we can use the generalized maximum principle [5]. We then can choose (see the appendix) $\varphi_{1|\Gamma_b} > 0$, $\varphi_{2|\Gamma_i} > 0$, $\delta_1'' > 0$, and $\delta_3'' > 0$, such that

$$\varphi_1 > 0, \quad \frac{\partial\varphi_1}{\partial n}_{|\Gamma_b} > 0, \quad \nu\varphi_1 > (1+\epsilon_2)\delta_1, \tag{4.30}$$

$$\varphi_2 > 0, \quad \frac{\partial\varphi_2}{\partial n}_{|\Gamma_i} > 0, \quad \nu\varphi_2 > (1+\epsilon_2)\delta_2. \tag{4.31}$$

We then obtain

$$\begin{aligned}&\int\left(\left(\frac{a}{2}\varphi_1 + \delta_1'\right)(1+\epsilon_2) + \delta_1''\right)u^2 + \int\left(\left(\frac{a}{2}\varphi_2 + \delta_3\right)(1+\epsilon_2) + \delta_3''\right)u_l^2\\ &\quad + \frac{\nu}{2}\int_{\Gamma_b} u^2\frac{\partial\varphi_1}{\partial n} + \frac{\nu}{2}\int_{\Gamma_i} u_l^2\frac{\partial\varphi_2}{\partial n} + \int(1+\epsilon_2)\delta_1|\nabla u|^2 + \int(1+\epsilon_2)\delta_2|\nabla u_l|^2\\ &\le \int_\Omega\left(\frac{a}{2}\varphi_1 + \delta_1'\right)f^2 + \int_{\Omega_l}\left(\frac{a}{2}\varphi_2 + \delta_3\right)f_l^2 + \int_\Omega\delta_1|\nabla f|^2 + \int_{\Omega_l}\delta_2|\nabla f_l|^2\\ &\quad + \int_{\Omega_l}\delta_4 f_{lm}^2 + \int_{\Omega_l}\delta_5|\nabla f_{lm}|^2.\end{aligned} \tag{4.32}$$

Step 4. Conclusions. From (4.32) we deduce

$$\begin{aligned}&\int\left(\left(\frac{a}{2}\varphi_1 + \delta_1'\right) + \frac{1}{1+\epsilon_2}\delta_1''\right)u^2 + \int\left(\left(\frac{a}{2}\varphi_2 + \delta_3\right) + \frac{1}{1+\epsilon_2}\delta_3''\right)u_l^2\\ &+ \frac{1}{1+\epsilon_2}\frac{\nu}{2}\int_{\Gamma_b} u^2\frac{\partial\varphi_1}{\partial n} + \frac{1}{1+\epsilon_2}\frac{\nu}{2}\int_{\Gamma_i} u_l^2\frac{\partial\varphi_2}{\partial n} + \int\delta_1|\nabla u|^2 + \int\delta_2|\nabla u_l|^2\\ &\quad\le \int_\Omega\frac{1}{1+\epsilon_2}\left(\frac{a}{2}\varphi_1 + \delta_1'\right)f^2 + \int_{\Omega_l}\frac{1}{1+\epsilon_2}\left(\frac{a}{2}\varphi_2 + \delta_3\right)f_l^2 + \int_\Omega\frac{1}{1+\epsilon_2}\delta_1|\nabla f|^2\\ &\qquad + \int_{\Omega_l}\frac{1}{1+\epsilon_2}\delta_2|\nabla f_l|^2 + \int_{\Omega_l}\frac{1}{1+\epsilon_2}\delta_4 f_{lm}^2 + \int_{\Omega_l}\frac{1}{1+\epsilon_2}\delta_5|\nabla f_{lm}|^2.\end{aligned} \tag{4.33}$$

Because of our special construction of the functions φ_1 and φ_2, we obtain

$$\int \left(\left(\frac{a}{2}\varphi_1 + \delta_1' \right) + \frac{1}{1+\epsilon_2}\delta_1'' \right) u^2 + \int \left(\left(\frac{a}{2}\varphi_2 + \delta_3 \right) + \frac{1}{1+\epsilon_2}\delta_3'' \right) u_l^2$$
$$+ \int \delta_1 |\nabla u|^2 + \int \delta_2 |\nabla u_l|^2$$
$$\leq \int_\Omega \frac{1}{1+\epsilon_2} \left(\frac{a}{2}\varphi_1 + \delta_1' \right) f^2 + \int_{\Omega_l} \frac{1}{1+\epsilon_2} \left(\frac{a}{2}\varphi_2 + \delta_3 \right) f_l^2$$
$$+ \int_\Omega \frac{1}{1+\epsilon_2}\delta_1 |\nabla f|^2 + \int_{\Omega_l} \frac{1}{1+\epsilon_2}\delta_2 |\nabla f_l|^2$$
$$+ \int_{\Omega_l} \frac{1}{1+\epsilon_2}\delta_4 f_{lm}^2 + \int_{\Omega_l} \frac{1}{1+\epsilon_2}\delta_5 |\nabla f_{lm}|^2. \tag{4.34}$$

Setting

$$a_{n+1} = \int \left(\left(\frac{a}{2}\varphi_1 + \delta_1' \right) + \frac{1}{1+\epsilon_2}\delta_1'' \right) u^2 + \int \delta_1 |\nabla u|^2$$
$$+ \int \left(\left(\frac{a}{2}\varphi_2 + \delta_3 \right) + \frac{1}{1+\epsilon_2}\delta_3'' \right) u_l^2 + \int \delta_2 |\nabla u_l|^2,$$

equation (4.34) then becomes

$$\begin{aligned} a_{n+1} &\leq s_1 a_n + s_2 a_{n-1}, \\ s_1 &= \frac{1}{1+\epsilon_2}, \\ s_2 &= \frac{1}{1+\epsilon_2}\max(\delta_4, \delta_5). \end{aligned} \tag{4.35}$$

Setting

$$b_n = \begin{pmatrix} a_n \\ a_{n-1} \end{pmatrix}$$

and

$$A = \begin{pmatrix} s_1 & s_2 \\ 1 & 0 \end{pmatrix},$$

we then obtain, using the Euclidean norm $\|.\|_2$,

$$\begin{aligned} \|b_n\|_2 \leq \|Ab_{n-1}\|_2 &= \sqrt{(s_1 a_{n-1} + s_2 a_{n-2})^2 + a_{n-1}^2} \\ &\leq \|A\|_2 \|b_{n-1}\|_2 \\ &\leq \|A\|_2^{n-1} \|b_1\|_2. \end{aligned}$$

Then the sequence a_n converges if the spectral radius of the matrix A is less than 1. This is true if

$$s_2 \leq 1 - \frac{1}{1+\epsilon_2} = \frac{\epsilon_2}{1+\epsilon_2}.$$

This last inequality is true if τ is taken to be sufficiently small. (See the definition of δ_4 and δ_5.) This concludes the proof of the theorem.

5. Conclusions. In this paper we proved the convergence of the TTMA for the case of transmission problems of hydrodynamics type. The proof is based in an essential way on the method we introduced in [10], and termed here "the transmission multiplier method," and the local, global, and trace estimates we developed in [11, 12]. We have described some applications of the TTMA to solving problems in fluid mechanics resulting from the applications of the α– and β–monoscale-multimodel methods. Our proof is obtained under a smallness condition on the time step only.

Appendix. Equation (4.28) can be written in the form

$$(L+h)\varphi_1 = f, \tag{A.1}$$

where

$$L\varphi_1 = \nu\Delta\varphi_1 + v\cdot\nabla\varphi_1,$$
$$h = a\epsilon_2, \quad f = -2\delta_1'(1+\epsilon_2) - 2\delta_1''.$$

Since Ω_l is bounded we can find a_1 and a_2, two real numbers such that Ω_l is contained in the slab $a_1 < x_1 < a_2$, where x_1 is the first coordinate of $x = (x_1, \ldots, x_n)$. Let

$$w(x) = 1 - \eta_2 e^{\eta_1(x_1 - a_1)}. \tag{A.2}$$

The numbers η_1 and η_2 are to be selected so that

$$w(x) > 0 \quad \text{on} \quad \Omega_l \cup \partial\Omega_l, \tag{A.3}$$
$$(L+h)w \le 0 \quad \text{in} \quad \Omega_l. \tag{A.4}$$

The operator $L+h$ applied to w yields

$$(L+h)w = -\eta_2(\eta_1^2\nu + \eta_1 v_1 + a\epsilon_2)e^{\eta_1(x_1-a_1)} + a\epsilon_2. \tag{A.5}$$

Let $m > 0$ be such that $v_1(x) \ge -m \ \ \forall x \in \Omega_l$. We may choose $m = \|v\|_{\infty,\Omega}$. We then choose η_1 and η_2 such that

$$\eta_1^2\nu - \eta_1 m + a\epsilon_2 > 0,$$
$$\eta_2 = \frac{a\epsilon_2}{\eta_1^2\nu - \eta_1 m + a\epsilon_2}.$$

This choice of η_1 and η_2 yields

$$(L+h)w \le 0 \quad \text{in} \quad \Omega_l. \tag{A.6}$$

Since we also want w to be positive on $\Omega_l \cup \partial\Omega_l$, we must have $\eta_2 e^{\eta_1(a_2-a_1)} < 1$. That is the inequality

$$\frac{a\epsilon_2}{\eta_1^2\nu - \eta_1 m + a\epsilon_2} < e^{-\eta_1(a_2-a_1)}. \tag{A.7}$$

This is satisfied if

$$\begin{aligned}\epsilon_2 &< \frac{1}{a}\eta_1 e^{-\eta_1(a_2-a_1)}\frac{(\eta_1\nu - m)}{1 - e^{-\eta_1(a_2-a_1)}} \\ &< \tau\eta_1 e^{-\eta_1(a_2-a_1)}\frac{(\eta_1\nu - m)}{1 - e^{-\eta_1(a_2-a_1)}},\end{aligned} \tag{A.8}$$

where we have used the definition of a. For ϵ_2 satisfying (A.8), the function w satisfies (A.3) and (A.4). Therefore [5] the solution of

$$(L+h)\varphi = -1 \text{ in } \Omega_l, \tag{A.9}$$

$$\frac{\partial \varphi}{\partial n} + \eta\varphi = 1 \text{ on } \Gamma_i, \tag{A.10}$$

$$\varphi = 1 \text{ on } \Gamma_b \tag{A.11}$$

with $\eta > 0$ satisfies $\varphi \geq w > 0$. By appropriate choice of $f = -2\delta_1'(1+\epsilon_2) - 2\delta_1'' \leq 0$, $g_1 > 0$, and $g_2 > 0$, the solution of

$$(L+h)\varphi = f \text{ in } \Omega_l, \tag{A.12}$$

$$\frac{\partial \varphi}{\partial n} + \eta\varphi = g_1 \text{ on } \Gamma_i, \tag{A.13}$$

$$\varphi = g_2 \text{ on } \Gamma_b \tag{A.14}$$

satisfies $\varphi > 0$ in $\Omega_l \cup \Gamma_i \cup \Gamma_b$ and $\frac{\partial \varphi}{\partial n} > 0$ on Γ_b. Thus $a_3\varphi_1$ with appropriate choice of $a_3 > 0$ satisfies the requirements in (4.30). Similar construction can be used to find φ_2 satisfying (4.31)

REFERENCES

[1] M. O. Bristeau, R. Glowinski, L. Dutto, J. Périaux, and G. Rogé, *Compressible viscous flow calculations using compatible finite element approximations*, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 719–749.

[2] P. Le Tallec and M. D. Tidriri, *Convergence analysis of domain decomposition algorithms with full overlapping for the advection-diffusion problems*, Math. Comp., 68 (1999), pp. 585–606.

[3] P. Le Tallec and M. D. Tidriri, *Application of maximum principles to the analysis of a coupling time marching algorithm,* J. Math. Anal. Appl., 229 (1999), pp. 158–169.

[4] L. Nirenberg, *On elliptic partial differential equations,* Ann. Swóla. Norm. Sup. Pisa (3), 13 (1959), pp. 115–162.

[5] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations,* Prentice Hall, Englewood Cliffs, NJ, 1967.

[6] Ph. Rostand and B. Stoufflet, *Finite Volume Galerkin Methods for Viscous Gas Dynamics*, Technical report 863, INRIA, Rocquencourt, France, 1988.

[7] M. D. Tidriri, *Couplage d'approximations et de modèles de types différents dans le calcul d'écoulements externes*, thèse, Université de Paris IX, Paris, France, 1992.

[8] M. D. Tidriri, *Domain Decomposition for Incompatible Nonlinear Models*, Technical report 2435, INRIA, Rocquencourt, France, 1994.

[9] M. D. Tidriri, *Domain decompositions for compressible Navier–Stokes equations,* J. Comput. Phys., 119 (1995), pp. 271–282.

[10] M. D. Tidriri, *Asymptotic analysis of a coupled system of kinetic equations*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 637–642, 1999.

[11] M. D. Tidriri, *Local and global estimates for the solutions of convection-diffusion problems,* J. Math. Anal. Appl., 229 (1999), pp. 137–157.

[12] M. D. Tidriri, *Development of local, global, and trace estimates for the solutions of elliptic equations*, Abstr. Appl. Anal., 3 (2001), pp. 131–150.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 469–491

A NEW CLASS OF OPTIMAL HIGH-ORDER STRONG-STABILITY-PRESERVING TIME DISCRETIZATION METHODS*

RAYMOND J. SPITERI† AND STEVEN J. RUUTH‡

Abstract. Strong-stability-preserving (SSP) time discretization methods have a nonlinear stability property that makes them particularly suitable for the integration of hyperbolic conservation laws where discontinuous behavior is present. Optimal SSP schemes have been previously found for methods of order 1, 2, and 3, where the number of stages s equals the order p. An optimal low-storage SSP scheme with $s = p = 3$ is also known. In this paper, we present a new class of optimal high-order SSP and low-storage SSP Runge–Kutta schemes with $s > p$. We find that these schemes are ultimately more efficient than the known schemes with $s = p$ because the increase in the allowable time step more than offsets the added computational expense per step. We demonstrate these efficiencies on a set of scalar conservation laws.

Key words. strong stability preserving, total variation diminishing, Runge–Kutta methods, high-order accuracy, time discretization

AMS subject classifications. 65L06, 65M20

PII. S0036142901389025

1. Introduction.

The method of lines is a popular semidiscretization method for the solution of time-dependent partial differential equations (PDEs). The idea behind it is to first suitably discretize the spatial variables (e.g., by finite differences, finite volumes, finite elements, or spectral methods) to yield a set of ordinary differential equations (ODEs) in time. Then, this set of ODEs can be integrated using standard time-stepping techniques such as linear multistep or Runge–Kutta methods.

Standard stability analysis for the solvers of such systems generally focuses on linear stability. Indeed, such analysis is often adequate when the desired solutions are smooth. However, solutions to hyperbolic PDEs may not be smooth: shock waves or other discontinuous behavior can develop even from smooth initial data. In such cases, standard discretizations based on linear stability analysis suffer from poor performance due to the presence of spurious oscillations, overshoots, and progressive smearing. The numerical solutions obtained from these discretizations often exhibit a weak form of instability (called *nonlinear instability*) resulting in unphysical behavior. Accordingly, numerical methods based on a nonlinear stability requirement are very desirable. Such methods were originally referred to as *total variation diminishing* (TVD) [17]; see also the subsequent articles [18, 6]. However, following the more recent article [7], we refer to them in this paper as *strong-stability-preserving* (SSP) methods.

We are interested in the development, implementation, and analysis of a new class of optimal SSP Runge–Kutta (SSPRK) time-stepping schemes for the system of

*Received by the editors May 6, 2001; accepted for publication (in revised form) December 18, 2001; published electronically May 29, 2002.

http://www.siam.org/journals/sinum/40-2/38902.html

†Department of Computer Science, Dalhousie University, Halifax, Nova Scotia, B3H 1W5 Canada (spiteri@cs.dal.ca). The work of this author was partially supported by grants from NSERC Canada and Imperial Oil.

‡Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada (sruuth@sfu.ca). The work of this author was partially supported by a grant from NSERC Canada.

ODEs

$$\dot{U} = L(U)$$

subject to suitable initial conditions, obtained from applying the method of lines to the hyperbolic conservation law

$$u_t + f(u)_x = 0. \tag{1.1}$$

Here, we assume that (1.1) has been suitably discretized in its spatial variables (e.g., using essentially nonoscillatory (ENO) schemes [10], TVD schemes [9], or monotonic upstream-centered schemes for conservation laws (MUSCL) methods [19]) and $U = U(t)$ is a vector of discretized variables; i.e., $[U(t)]_j = U_j(t) = u(x_j, t)$. In particular, if u_j^n is the numerical approximation to $u(x_j, t_n)$, then TVD discretizations have the property that the total variation

$$TV(U^n) = \sum_j |u_j^n - u_{j-1}^n| \tag{1.2}$$

of the numerical solution does not increase with time; i.e.,

$$TV(U^{n+1}) \leq TV(U^n).$$

When combined with a suitable SSP time-stepping scheme, the numerical solution obtained typically does not exhibit nonlinear instabilities. However, nonlinear instabilities can occur in a numerical solution obtained with, e.g., a TVD or MUSCL spatial discretization scheme, but with a standard (i.e., linearly stable) time-stepping scheme [6]. Hence, SSP time-stepping schemes are a critical part of the overall solution strategy to (1.1).

It has been known for some time from a result of Goodman and LeVeque [5] that any method that is TVD in two dimensions is at most first-order accurate. However, if we relax the strict requirement of TVD schemes, higher-order methods can be constructed that preserve stability in another suitable norm, such as the maximum norm. These schemes are what we call SSP, and their favorable properties are derived only from convexity arguments. In particular, if the forward Euler method is strongly stable with a certain CFL number, higher-order SSPRK methods with a modified CFL number can be constructed as convex combinations of forward Euler steps with various step sizes [18].

Optimal SSP schemes based on Runge–Kutta methods have been found for accuracy orders 1, 2, and 3, where the number of stages s is assumed to be equal to the order p. Gottlieb and Shu [6] recently proved that, unfortunately, no such four-stage, fourth-order SSPRK method exists involving just evaluations of $L(\cdot)$. Fourth-order accuracy has only been obtained at the additional expense of introducing two additional evaluations of a related operator $\tilde{L}(\cdot)$, leading to suboptimal efficiency both in terms of time-step restriction and memory usage (see section 2). This appears to be where the search for higher-order SSPRK methods has stopped, thus leaving researchers to focus on third-order accurate SSPRK methods.

In this paper, we derive a new class of optimal high-order SSPRK schemes where the restriction $s = p$ is lifted. For s-stage methods of orders 1 and 2, we provide proofs of optimality. The SSPRK scheme (4,3) is also proven to be optimal. The remaining schemes of order 3 and higher and the low-storage schemes are the results

of numerical optimization. We investigate the performance of our new schemes on a few test problems designed to capture solution features that pose particular difficulties to numerical methods. These features include contact discontinuities, expansion fans, compressive shocks, and sonic points. The results from these investigations indicate that both the standard and low-storage versions of our schemes offer significant advantages over methods currently available. In particular, our new schemes have significantly better stability restrictions than the best SSPRK schemes currently known. Thus, step-size selection can be based more on accuracy requirements rather than stability requirements, ultimately leading to more efficient integrators. Indeed, the results based on three important test cases indicate that our new fourth-order SSPRK scheme offers between 40% and 80% improvement in the effective time-step restriction over the most popular fourth-order schemes currently in use.

The remainder of this paper unfolds as follows. In section 2, we describe SSP schemes and motivate their use. In section 3, we determine optimal families of SSPRK schemes up to 5 stages and order 4. We also give optimal low-storage versions of these schemes. In section 4, we investigate the performance of our new SSPRK schemes on a set of scalar conservation laws having solutions that commonly cause numerical problems. The success of the new methods is measured relative to the most popular schemes currently in use. Finally, in section 5, we summarize our findings and offer plans for future work.

2. SSP schemes. The concept of strong stability is central to our discussion, so we begin with its definition.

DEFINITION 2.1. *A sequence* $\{U^n\}$ *is said to be* strongly stable *in a given norm* $||\cdot||$ *provided that* $||U^{n+1}|| \le ||U^n||$ *for all* $n \ge 0$.

We tacitly assume that U^n represents a vector of solution values on a mesh obtained from a method-of-lines approach to solving a PDE. The choice of norm is arbitrary,[1] with the TV-norm (1.2) and the infinity norm being two natural possibilities. Clearly, strong stability may not be relevant to the solution of an arbitrary PDE. However, the class of PDEs (1.1) forms a notable exception. Exact solutions for this class of problems have a range-diminishing property that forbids existing maxima from increasing, existing minima from decreasing, and new maxima or minima from forming. Although not precisely a discrete analogue to the range-diminishing property, the strong-stability property is a useful property to require of a numerical solution to (1.1): by imposing such a condition on the numerical solution, we can suppress the formation of spurious oscillations under a suitable restriction on the time step. Such oscillations are termed *nonlinear instabilities* and are often a precursor for the numerical solution itself to become completely unstable.

The authors in [7] prove the somewhat surprising result that, under rather general assumptions, high-order SSP methods must in fact be explicit. Fortunately, many researchers in fact prefer explicit time discretization methods in order to avoid the expense[2] of solving systems of nonlinear equations at each step. Accordingly, in this paper we will focus on the development of explicit Runge–Kutta methods. Consider an s-stage, explicit Runge–Kutta method written in the form

[1] Indeed, the results of this paper still apply if we replace the norm $||\cdot||$ by *any* convex function that maps into the nonnegative real line.

[2] Both in terms of computation time and software development.

$$U^{(0)} = U^n, \tag{2.1a}$$

$$U^{(i)} = \sum_{k=0}^{i-1} (\alpha_{ik} U^{(k)} + \Delta t \beta_{ik} L(U^{(k)})), \qquad i = 1, 2, \ldots, s, \tag{2.1b}$$

$$U^{n+1} = U^{(s)}, \tag{2.1c}$$

where all $\alpha_{ik} \geq 0$ and $\alpha_{ik} = 0$ only if $\beta_{ik} = 0$ [17]. This representation of a Runge–Kutta method can be converted to the standard Butcher array form (see, e.g., [8]) in a straightforward manner; see also [6]. However, the conversion from the Butcher array form to (2.1) is not unique. For example, the modified Euler scheme

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

has a one-parameter family of representations of the form (2.1):

$$\alpha_{10} = 1,\ \alpha_{20} = 1 - \lambda,\ \alpha_{21} = \lambda, \qquad \beta_{10} = 1,\ \beta_{20} = \frac{1}{2} - \lambda,\ \beta_{21} = \frac{1}{2},$$

where $\lambda \in [0, 1]$. All of these representations are algebraically equivalent [18]; i.e., the only differences noticeable between stable implementations of any scheme would be due to round-off errors. However, different choices of λ may lend themselves more easily to implementation, memory management, or determination of stability restrictions. Throughout this article, we give representations that naturally allow stability restrictions to be read from the coefficients of the scheme. Standard Butcher array forms of the schemes presented are given in Appendix B.

For consistency, we must have that $\sum_{k=0}^{i-1} \alpha_{ik} = 1$, $i = 1, 2, \ldots, s$. Hence, if both sets of coefficients α_{ik}, β_{ik} are positive, then (2.1) is a convex combination of forward Euler steps with various step sizes $\frac{\beta ik}{\alpha_{ik}} \Delta t$. The Runge–Kutta scheme written in this form is particularly convenient to make use of the following result [18, 7].

THEOREM 2.2. *If the forward Euler method is strongly stable under the CFL restriction $\Delta t \leq \Delta t_{FE}$, then the Runge–Kutta method* (2.1) *with $\beta_{ik} \geq 0$ is SSP, provided*

$$\Delta t \leq c \Delta t_{FE},$$

where c is the CFL coefficient

$$c \equiv \min_{i,k} \frac{\alpha_{ik}}{\beta_{ik}}.$$

Thus, we can use the result of this theorem to provide a theoretical criterion according to which we can optimize a given SSPRK method.

SSPRK schemes with negative coefficients β_{ik} are also possible with the appropriate interpretation. Following the procedure first suggested in [17], whenever $\beta_{ik} < 0$, the operator $L(\cdot)$ is replaced with the related operator $\tilde{L}(\cdot)$, where $\tilde{L}(\cdot)$ is assumed to be strongly stable for Euler's method solved *backward* in time for a suitable time-step restriction. This allows the following generalization of Theorem 2.2.

THEOREM 2.3. *Let Euler's method solved forward in time combined with the spatial discretization $L(\cdot)$ be strongly stable under the CFL restriction $\Delta t \leq \Delta t_{FE}$.*

Let Euler's method solved backward in time combined with the spatial discretization $\tilde{L}(\cdot)$ *also be strongly stable under the same CFL restriction* $\Delta t \le \Delta t_{FE}$. *Then the Runge–Kutta method* (2.1) *is SSP, provided*

$$\Delta t \le c\Delta t_{FE},$$

where c *is the CFL coefficient*

$$c \equiv \min_{i,k} \frac{\alpha_{ik}}{|\beta_{ik}|},$$

where $\beta_{ik}L(\cdot)$ *is replaced by* $\beta_{ik}\tilde{L}(\cdot)$ *whenever* β_{ik} *is negative.*

Note. If both $L(U^{(i)})$ and $\tilde{L}(U^{(i)})$ are required, then the computational cost and storage requirements for that stage are typically doubled. Moreover, there is the added inconvenience of having to code the spatial discretization represented by $\tilde{L}(\cdot)$. These reasons provide the incentive for us to want to avoid negative β_{ik} as much as possible when searching for the most efficient SSPRK methods.

We also note that the quantity

$$\min_{i,k} \frac{\alpha_{ik}}{|\beta_{ik}|} \tag{2.2}$$

obviously depends on the particular representation (2.1) of a given Runge–Kutta scheme. Accordingly, the CFL restriction is determined by the choice of coefficients $\alpha_{ij},\ \beta_{ij}$ that maximizes (2.2); other choices render bounds that are not as sharp.

We will be comparing a new class of SSPRK methods with s stages and order p with $s > p$ to methods known in the literature where $s = p$ or some $\beta_{ik} < 0$. We note that if a method requires n_- extra evaluations of $\tilde{L}(\cdot)$, then the effective number of stages of that method is $m = s + n_-$. We find that the new SSPRK methods can have a significantly greater CFL coefficient (as given in Theorem 2.2) than the methods currently used in practice. However, we must make a fair comparison as to the computational cost of a step. This motivates the following definition.

Definition 2.4. *The* effective CFL coefficient *of an SSPRK method of order* p *is* cs^*/s, *where* c *is the CFL coefficient of the method,* s^* *is the minimum number of stages to theoretically achieve order* p, *and* s *is the number of stages required for one step of the method.*

It is well known (see, e.g., [8]) that a Runge–Kutta method having s stages can achieve order p for $s = p \le 4$. For $p > 4$, it is required that $s > p$. Because the cases we consider in this paper involve only $p \le 4$, we always take $s^* = p$ here.

As conjectured in Shu and Osher [18] and subsequently proven in Gottlieb and Shu [6], the optimal two-stage, order-2 SSPRK scheme is the modified Euler scheme

$$\begin{aligned} U^{(1)} &= U^n + \Delta t L(U^n), \\ U^{n+1} &= \frac{1}{2}U^n + \frac{1}{2}U^{(1)} + \frac{1}{2}\Delta t L(U^{(1)}). \end{aligned}$$

It has a CFL restriction $\Delta t \le \Delta t_{FE}$, which implies a CFL coefficient of 1. Henceforth, we will refer to this scheme as SSP(2,2). In general, we adopt the convention of referring to an s-stage, order-p SSPRK scheme as SSP(s,p).

Shu and Osher [18] also conjectured that the optimal three-stage, order-3 SSPRK scheme is

$$
\begin{aligned}
U^{(1)} &= U^n + \Delta t L(U^n),\\
U^{(2)} &= \frac{3}{4}U^n + \frac{1}{4}U^{(1)} + \frac{1}{4}\Delta t L(U^{(1)}),\\
U^{n+1} &= \frac{1}{3}U^n + \frac{2}{3}U^{(2)} + \frac{2}{3}\Delta t L(U^{(2)}),
\end{aligned}
$$

which has a CFL coefficient of 1 as well. The optimality of this scheme was later proved by Gottlieb and Shu [6]. This scheme is commonly called the *third-order TVD Runge–Kutta scheme*, but we will simply refer to it as SSP(3,3).

To achieve fourth order, Shu and Osher provide a four-stage method that contains two negative coefficients β_{ik} [18]. A slightly improved scheme (but also containing two negative coefficients β_{ik}) was proposed by Gottlieb and Shu [6]:

$$
\begin{aligned}
U^{(1)} &= U^n + \frac{1}{2}\Delta t L(U^n),\\
U^{(2)} &= \frac{649}{1600}U^n - \frac{10890423}{25193600}\Delta t \tilde{L}(U^n) + \frac{951}{1600}U^{(1)} + \frac{5000}{7873}\Delta t L(U^{(1)}),\\
U^{(3)} &= \frac{53989}{2500000}U^n - \frac{102261}{5000000}\Delta t \tilde{L}(U^n) + \frac{4806213}{20000000}U^{(1)}\\
&\quad - \frac{5121}{20000}\Delta t \tilde{L}(U^{(1)}) + \frac{23619}{32000}U^{(2)} + \frac{7873}{10000}\Delta t L(U^{(2)}),\\
U^{n+1} &= \frac{1}{5}U^n + \frac{1}{10}\Delta t L(U^n) + \frac{6127}{30000}U^{(1)} + \frac{1}{6}\Delta t L(U^{(1)})\\
&\quad + \frac{7873}{30000}U^{(2)} + \frac{1}{3}U^{(3)} + \frac{1}{6}\Delta t L(U^{(3)}).
\end{aligned}
$$

This scheme has a CFL coefficient of 0.936 and an effective CFL coefficient of $0.936 \times 4/6 = 0.624$ because 6 function evaluations are required per step. Because this seems to be the best four-stage, order-4 SSPRK scheme known, we will refer to it as SSP(4^{**},4), with the two asterisks meant to convey two negative coefficients β_{ik}. Gottlieb and Shu [6] subsequently proved that no four-stage, order-4 SSPRK scheme exists with positive coefficients.

Gottlieb and Shu [6] have also carried out an investigation of SSP time discretization methods for generalized Runge–Kutta methods (also known as pseudo-Runge–Kutta methods or hybrid methods [8]).[3] They report that they were unable to find effective SSP methods in this wider class of methods. It is from this point that we start our derivations of improved SSPRK schemes where generally $s > p$. The details of these derivations are provided in the next section.

3. Optimal SSP schemes. We now turn to the task of finding optimal SSPRK schemes. To begin, we seek to optimize an s-stage, order-p SSPRK scheme by maximizing its CFL coefficient according to Theorem 2.2. That is, we seek the global maximum of the nonlinear programming problem,

$$
\max_{(\alpha_{ik},\beta_{ik})} \min \frac{\alpha_{ik}}{\beta_{ik}}, \tag{3.1}
$$

where $\alpha_{ik},\ \beta_{ik}, k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$, are real and nonnegative. The case $\alpha_{ik} = \beta_{ik} = 0$ is defined as `NaN` in the sense that it is not included in the minimization

[3] All of these methods also are special cases of methods known as *general linear methods.*

process if it occurs. Besides the nonnegativity constraints on the variables $\alpha_{ik},\ \beta_{ik}$, the objective function (3.1) is subject to the constraints

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \qquad i = 1, 2, \ldots, s, \tag{3.2}$$

$$\sum_{j=1}^{s} b_j \Phi_j(t) = \frac{1}{\gamma(t)}, \quad t \in T_q, \quad q = 1, 2, \ldots, p. \tag{3.3}$$

Here, the functions $\Phi_j(t)$ are nonlinear constraints that are polynomial in $\alpha_{ik},\ \beta_{ik}$ and that correspond to the order conditions for a Runge–Kutta method to be of order p (see, e.g., [8]); i.e., T_q stands for the set of all rooted trees of order equal to q. The number of constraints represented by the Runge–Kutta order conditions is equal to

$$\sum_{q=1}^{p} \mathrm{card}(T_q),$$

where $\mathrm{card}(T_q)$ is the cardinality of T_q. Also, we use the notation b_j in the usual sense of the Butcher array representation of a Runge–Kutta method; again this would be a polynomial function of the coefficients α_{ik} and β_{ik}. It can be expected that the particular choice of coefficients $\alpha_{ik},\ \beta_{ik}$ that maximizes the quantity (2.2) for a given Runge–Kutta method will be naturally produced by the solution to this nonlinear programming problem; hence, the result will be a sharp estimate of the CFL coefficient.

In this form, the optimization problem does not lend itself easily to numerical solution. The difficulty due to the high degree of nonlinearity in the constraints is compounded by the following two considerations. First, the objective function (3.1) is nonsmooth and so an optimization strategy that uses gradient information will have difficulty obtaining reliable numerical estimates of the derivatives. Second, the $\min(\cdot)$ function can be quite insensitive to its arguments. This also contributes to the poor performance of optimization software on this problem. We found that even optimizers that do not rely on gradient information were unable to consistently converge to the same optimum with this formulation.

The performance of optimization software on this problem is greatly enhanced through the following standard reformulation. By introducing a dummy variable z, the nonlinear programming problem can be reformulated as

$$\max_{(\alpha_{ik},\beta_{ik})} z \tag{3.4a}$$

subject to

$$\alpha_{ik} \geq 0, \tag{3.4b}$$

$$\beta_{ik} \geq 0, \tag{3.4c}$$

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \qquad i = 1, 2, \ldots, s, \tag{3.4d}$$

$$\sum_{j=1}^{s} b_j \Phi_j(t) = \frac{1}{\gamma(t)}, \quad t \in T_q, \quad q = 1, 2, \ldots, p, \tag{3.4e}$$

$$\alpha_{ik} - z\beta_{ik} \geq 0, \quad k = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, s. \tag{3.4f}$$

It is easy to see that the dummy variable z corresponds to the CFL coefficient. This reformulation is a standard technique that is widely used in the context of linear programming problems with objective functions of the form $\max(\cdot)$ or $\min(\cdot)$ (see, e.g., [2]). It is also a common reformulation of the so-called *feasibility problem*, where any feasible solution to a set of equality or inequality constraints is desired (e.g., as in the first phase of a two-phase simplex algorithm for linear programming [3]).

The reformulated problem (3.4) was solved directly using the `fmincon` function from Matlab's Optimization Toolbox for $s = 1, 2, 3, 4, 5$ and $p = 1, 2, 3, 4$, and the results are shown below. Table 3.1 shows the optimal values for the CFL coefficients for given pairs (s, p). The * in the $(4, 4)$ position denotes the fact that no such SSPRK method exists with all coefficients $\alpha_{ik},\ \beta_{ik}$ positive.

TABLE 3.1
Optimal CFL coefficients for s-stage, order-p SSPRK methods.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$p = 1$	1	2	3	4	5
$p = 2$		1	2	3	4
$p = 3$			1	2	2.65
$p = 4$				*	1.51

Table 3.2 gives the theoretical efficiencies of these new schemes relative to the ones where $s = p$. We note that there is no efficiency gain for the first-order methods. For example, although the CFL coefficient of the (2,1) method is twice that of that (1,1) method (forward Euler), it also requires twice as much work. The percentages quoted refer to the theoretical increases in allowable step size of the new methods relative to the methods with $s = p$. For example, the (3,2) method has twice the allowable step size compared to the (2,2) method (the modified Euler method), but it requires 3/2 times more work. We thus report that the net effect is a relative increase in step size of $((2/1)/(3/2) - 1) \times 100\% = 33\%$. Equivalently, assuming the CFL coefficient is the exact bound on the time step, the new $(3, 2)$ scheme can produce a comparable second-order accurate answer with only 75% of the computational effort as the $(2, 2)$ scheme.

TABLE 3.2
Theoretical efficiency improvement over standard pth order SSPRK schemes.

	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$p = 2$		33%	50%	60%
$p = 3$			50%	59%
$p = 4$				94%

We draw particular attention to the efficiency of the $(5, 4)$ scheme in Table 3.2. As mentioned earlier, a $(4, 4)$ SSPRK scheme does not exist for any positive CFL coefficient. The figure of 94% is measured relative to the $(4^{**}, 4)$ scheme reported in [6] as the best scheme of order 4 that could be found. Recall that this scheme had a CFL coefficient of 0.936 and effectively used 6 stages because it involved 2 coefficients β_{ik} that are negative (hence leading to a 50% increase of the storage requirement per step and the overhead of coding $\tilde{L}(\cdot)$). The new $(5, 4)$ scheme thus compares very favorably.

The first few optimal SSPRK schemes of orders 1 and 2 are given in Tables 3.3

TABLE 3.3
The first few optimal SSPRK schemes of order 1.

Stages	α_{ik}	β_{ik}	CFL coefficient
1	1	1	1
2	$\begin{matrix}1 & \\ 0 & 1\end{matrix}$	$\begin{matrix}\frac{1}{2} & \\ 0 & \frac{1}{2}\end{matrix}$	2
3	$\begin{matrix}1 & & \\ 0 & 1 & \\ 0 & 0 & 1\end{matrix}$	$\begin{matrix}\frac{1}{3} & & \\ 0 & \frac{1}{3} & \\ 0 & 0 & \frac{1}{3}\end{matrix}$	3

TABLE 3.4
The first few optimal SSPRK schemes of order 2.

Stages	α_{ik}	β_{ik}	CFL coefficient
2	$\begin{matrix}1 & \\ \frac{1}{2} & \frac{1}{2}\end{matrix}$	$\begin{matrix}1 & \\ 0 & \frac{1}{2}\end{matrix}$	1
3	$\begin{matrix}1 & & \\ 0 & 1 & \\ \frac{1}{3} & 0 & \frac{2}{3}\end{matrix}$	$\begin{matrix}\frac{1}{2} & & \\ 0 & \frac{1}{2} & \\ 0 & 0 & \frac{1}{3}\end{matrix}$	2
4	$\begin{matrix}1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ \frac{1}{4} & 0 & 0 & \frac{3}{4}\end{matrix}$	$\begin{matrix}\frac{1}{3} & & & \\ 0 & \frac{1}{3} & & \\ 0 & 0 & \frac{1}{3} & \\ 0 & 0 & 0 & \frac{1}{4}\end{matrix}$	3

and 3.4. Here we give the schemes in terms of the coefficients α_{ik}, β_{ik}; the Butcher form of these schemes is given in Appendix B. It is interesting to note that Gerisch and Weiner have independently proposed the SSP(3,2) scheme; see [4] for details.

From Tables 3.1, 3.3, and 3.4, we can conjecture the form of the optimal SSPRK methods with s stages and orders 1 and 2; namely, the optimal SSPRK method with s stages and order 1 has CFL coefficient s; and the optimal SSPRK method with s stages and order 2 has CFL coefficient $s-1$. Shu [17] has given a proof of the first-order result, and Gottlieb and Shu [6] have given a proof of the second-order result for $s=2$. We provide a new proof of the first-order result below as well as a proof of the second-order result for arbitrary s. These low-order methods with large CFL coefficients are useful when seeking a time-independent (steady-state) solution of (1.1), given that in such problems the accuracy considerations in time are typically less critical than those in space [17].

THEOREM 3.1. *For $s=1,2,3,\ldots$, the optimal s-stage SSPRK method of order 1 with $\beta_{ik}\geq 0$ has CFL coefficient s and can be represented in the form*

$$\alpha_{ik}=\begin{cases}1 & k=i-1,\\ 0 & \textit{otherwise.}\end{cases}\qquad \beta_{ik}=\begin{cases}\frac{1}{s} & k=i-1,\\ 0 & \textit{otherwise.}\end{cases}\qquad i=1,2,\ldots,s.$$

Before giving the proof of Theorem 3.1, we introduce the following notation and give a useful lemma. We find it convenient to write the general s-stage explicit Runge–Kutta method in the following form (cf. [6]):

$$U^{(0)}=U^n, \tag{3.5a}$$

$$U^{(i)}=U^{(0)}+\Delta t\sum_{k=0}^{i-1}c_{ik}L(U^{(k)}),\qquad i=1,2,\ldots,s, \tag{3.5b}$$

$$U^{n+1}=U^{(s)}. \tag{3.5c}$$

The coefficients c_{ik} are related to the coefficients $\alpha_{ik},\ \beta_{ik}$ recursively by

$$c_{ik} = \sum_{j=k+1}^{i-1} \alpha_{ij} c_{jk} + \beta_{ik}. \tag{3.6}$$

It is also easy to see that the coefficients c_{ik} are related to the Butcher array quantities $a_{ik},\ b_k$ by

$$\begin{aligned} a_{ik} &= c_{i-1,k-1}, \qquad k = 1, 2, \ldots, i-1, \quad i = 1, 2, \ldots, s-1, \\ b_k &= c_{s,k-1}, \qquad k = 1, 2, \ldots, s. \end{aligned}$$

LEMMA 3.2. *If a method of the form* (2.1) *with* $\alpha_{ik},\ \beta_{ik} \geq 0$ *has a CFL coefficient* $c > m > 0$, *then* $0 \leq c_{ik} < \frac{1}{m}$ *for all* $k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$.

Proof. From Theorem 2.2, if $c > m > 0$, then $\alpha_{ik} > m\beta_{ik}$, or equivalently $\beta_{ik} < \frac{1}{m}\alpha_{ik}$, for all $i,\ k$ such that $\alpha_{ik} \neq 0$.

Now,

$$\alpha_{ik} \geq 0, \quad \sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad i = 1, 2, \ldots, s, \quad \Rightarrow \quad \alpha_{ik} \leq 1$$

for all $i,\ k$. Hence, $\beta_{ik} < \frac{1}{m}$ for all $i,\ k$. In particular, $c_{10} = \beta_{10} < \frac{1}{m}$ for any valid SSPRK method.

We now proceed by induction on stage ℓ of an s-stage method. Assume $c_{ij} < \frac{1}{m}$ for $j = 0, 1, \ldots, \ell-1;\ i = 1, 2, \ldots, \ell$. (We have just shown that this result holds for $\ell = 1$.) Now consider stage $(\ell+1)$ of a valid SSPRK method; i.e., consider coefficients $c_{\ell+1,k}$ for $k = 0, 1, \ldots, \ell$ with

$$\sum_{k=0}^{\ell} \alpha_{\ell+1,k} = 1.$$

Then using (3.6),

$$\begin{aligned} c_{\ell+1,0} &= \sum_{k=1}^{\ell} \alpha_{\ell+1,k} c_{k0} + \beta_{\ell+1,0} \\ &< \frac{1}{m} \sum_{k=1}^{\ell} \alpha_{\ell+1,k} + \frac{1}{m}\alpha_{\ell+1,0} \\ &= \frac{1}{m}. \end{aligned}$$

Similar arguments can be used to show $c_{\ell+1,j} < \frac{1}{m}$ for $j = 1, 2, \ldots, \ell$. The lemma now follows by induction. ∎

Proof of Theorem 3.1. By contradiction, suppose there exists an s-stage, order-1 SSPRK method with CFL coefficient $c > s$. Because the method is order 1, we have

$$\sum_{k=0}^{s-1} c_{sk} = 1. \tag{3.7}$$

But from Lemma 3.2, we have

$$c_{ik} < \frac{1}{s}, \qquad k = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, s.$$

Thus,

$$\sum_{k=0}^{s-1} c_{sk} < \sum_{k=0}^{s-1} \frac{1}{s} = 1,$$

contradicting (3.7). Thus, no s-stage, order-1 SSPRK method can exist with CFL coefficient $c > s$. Because the SSPRK methods proposed in Theorem 3.1 have $c = s$, they must be optimal representations. ∎

THEOREM 3.3. *For $s = 2, 3, 4, \ldots$, the optimal s-stage SSPRK method of order 2 with $\beta_{ik} \geq 0$ has CFL coefficient $s - 1$ and can be represented in the form*

$$\alpha_{ik} = \begin{cases} 1 & k = i-1, \\ 0 & \textit{otherwise.} \end{cases} \qquad \beta_{ik} = \begin{cases} \frac{1}{s-1} & k = i-1, \\ 0 & \textit{otherwise.} \end{cases} \qquad i = 1, 2, \ldots, s-1.$$

$$\alpha_{ik} = \begin{cases} \frac{1}{s} & k = 0, \\ \frac{s-1}{s} & k = s-1, \\ 0 & \textit{otherwise.} \end{cases} \qquad \beta_{ik} = \begin{cases} \frac{1}{s} & k = s-1, \\ 0 & \textit{otherwise.} \end{cases} \qquad i = s.$$

Proof. By contradiction, suppose there exists an s-stage, order-2 SSPRK method with CFL coefficient $c > s - 1$. Because it is order 2, the coefficients of the method must satisfy (3.7) and

$$\sum_{i=1}^{s-1} c_{si} \sum_{k=0}^{i-1} c_{ik} = \frac{1}{2}. \tag{3.8}$$

Also, using Lemma 3.2 with $c > s - 1$ implies that

$$c_{ik} < \frac{1}{s-1}, \qquad k = 0, 1, \ldots, i-1, \quad i = 1, 2, \ldots, s. \tag{3.9}$$

Using (3.9) in (3.8) for $k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s-1$, leads to

$$\sum_{i=1}^{s-1} \frac{i}{s-1} c_{si} > \frac{1}{2},$$

and using this result in (3.7) yields

$$\sum_{k=0}^{s-2} \frac{s-k-1}{s-1} c_{sk} < \frac{1}{2}.$$

Thus,

$$\begin{aligned} \frac{1}{2} &> \sum_{k=0}^{s-2} \frac{s-k-1}{s-1} c_{sk} \\ &= \sum_{k=0}^{s-2} \frac{s-k-1}{s-1} \left(\sum_{j=k+1}^{s-1} \alpha_{sj} c_{jk} + \beta_{sk} \right). \end{aligned}$$

Now we substitute recursively for c_{jk} using (3.6) in the right-hand side of the above equation and (3.8), and recalling that $\alpha_{ik} > (s-1)\beta_{ik}$ and $\alpha_{ik} \geq 0$ for $k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$, we can use (3.8) to write

$$\frac{1}{2} > \frac{1}{2} + \sum_{j=1}^{s-2} j \sum_{l=s-j-1}^{s-1} \beta_{sl}\beta_{l,s-2-j} + \sum_{j=0}^{s-2} \frac{s-j-1}{s-1}\beta_{sj}.$$

This now contradicts the fact that $\beta_{ik} \geq 0$ for all $k = 0, 1, \ldots, i-1,\ i = 1, 2, \ldots, s$. Thus, no s-stage, order-2 SSPRK method can have CFL coefficient $c > s-1$. The proof is now completed by noting that because the schemes proposed have $c = s-1$, they must be optimal representations. □

In Tables A.1–A.2 in Appendix A, we give results for the coefficients of the optimal schemes of order $p = 3, 4$ in terms of their numerical values up to double precision.

A proof of optimality for the SSP(4,3) scheme follows easily from a result in [16], where it is proved that the optimal CFL coefficient of an s-stage SSPRK method of order p applied to a linear, constant-coefficient problem $\dot{U} = LU$ is $s - p + 1$. Thus if a nonlinear scheme can attain this optimal bound, then it must also be optimal. It is easy to see that SSP(4,3) is such a scheme.

We do not offer formal proofs of optimality in the remaining cases; however, these are the results of extensive numerical searches.

Finally, we describe our results for optimal low-storage SSPRK schemes. There are computational problems for which memory management considerations are at least as important as stability considerations when choosing a numerical time discretization method, e.g., direct numerical simulation of Navier–Stokes equations requiring high spatial resolution in three dimensions. In such cases, s-stage explicit Runge–Kutta methods that use less than the usual s units of storage are very desirable (see, e.g., [20]). We focus our discussion on SSPRK schemes that require only two units of storage per step,[4] although more general methods requiring more storage per step are possible. These schemes take the form

$$dU^{(i)} = A_i dU^{(i-1)} + \Delta t L(U^{(i-1)}), \tag{3.10a}$$

$$U^{(i)} = U^{(i-1)} + B_i dU^{(i-1)}, \qquad i = 1, 2, \ldots, s, \tag{3.10b}$$

where $U^{(0)} = U^n$, $U^{n+1} = U^{(s)}$, and $A_1 \equiv 0$. Again, we note that there is a relation between the coefficients $A_i,\ B_i$ and the coefficients $\alpha_{ik},\ \beta_{ik}$ or, equivalently, the usual quantities in the Butcher array. We denote the general s-stage, order-p low-storage SSPRK scheme simply by LS(s,p).

We have solved the corresponding nonlinear programming problems to optimize the CFL coefficient for the low-storage schemes defined by (3.10). The results for the coefficients $A_i,\ B_i$ are given in Tables 3.5–3.7 for up to 5 stages and order 3. Again, only numerical values of the coefficients are given to double precision. The Butcher array form of these schemes is given in Appendix C. Of course, a traditional implementation of any 2-stage scheme must be low-storage in the sense we are considering, so the optimal low-storage method with $s = p = 2$ corresponds to the optimal SSPRK scheme in Table 3.4. We note that the optimal 3-stage, order-3 low-storage method reported in Table 3.7 agrees with that reported in [6]. We also note that we were not successful in finding a 5-stage, order-4 scheme in this family, and we strongly suspect that such a method does not exist.

[4] We note that if some form of error control is envisaged, perhaps using an embedded [8] SSPRK scheme, then additional storage for the current solution vector is also required.

TABLE 3.5
The coefficients of the first few optimal low-storage schemes of order 1.

Stages	A_i	B_i	CFL coefficient
1	0	1	1
2	0	0.25471543653218	1
	0.66323286721269	0.44809394647120	
3	0	0.26237801705341	1
	0.42645094785793	0.20169056000013	
	0.45339958582027	0.27321697994061	
4	0	0.14142439246204	
	0.42623204099143	0.35397016495696	1
	0.38851833123083	0	
	0.01694135866933	0.34465757966021	
5	0	0.03368800719745	1
	0.61573074220688	0.13960527476637	
	0.24191712486786	0.22864919232774	
	0.16549924932085	0.26079330982391	
	-0.04239297405834	0.10750824432183	

TABLE 3.6
The coefficients of the first few optimal low-storage schemes of order 2.

Stages	A_i	B_i	CFL coefficient
2	0	1	1
	-1	$\frac{1}{2}$	
3	0	0.79609964254616	1
	-0.86514937424574	0.47921739051941	
	-0.01459406292961	0.13955204452449	
4	0	0.08820909208788	
	0.34143758512319	0.62773790223092	1
	-0.80189834090053	0.43908735985479	
	-0.26868602239001	0.10090483677631	
5	0	0.24064789292000	1
	-0.35363900948812	0.28813102587031	
	0.23144682054640	0.15490366543216	
	0.30287923513739	0.33623843526263	
	-0.90122396243589	0.27101878032131	

TABLE 3.7
The coefficients of the first few optimal low-storage schemes of order 3.

Stages	A_i	B_i	CFL coefficient
3	0	0.92457411523577	0.32234930738853
	-2.91549398859489	0.28771294148749	
	0.00000000151682	0.62653829645172	
4	0	1.03216665875130	0.52841816101829
	-4.94661981618529	0.18793881263711	
	0.00000000050902	0.15215751854315	
	-0.15127914578976	0.65675174856653	
5	0	0.67892607116139	1
	-2.60810978953486	0.20654657933371	
	-0.08977353434746	0.27959340290485	
	-0.60081019321053	0.31738259840613	
	-0.72939715170280	0.30319904778284	

4. Numerical studies. In this section, we study the numerical behavior of our schemes and Shu–Osher SSP schemes for a few test problems designed to capture solution features that pose particular difficulties to numerical methods. Experiments for the classical fourth-order explicit Runge–Kutta method are also included because this method is commonly used in method-of-lines discretizations of hyperbolic conservation laws but is not SSP.

4.1. Test problems. There are a variety of solution features in computational fluid dynamics that commonly cause numerical problems. For example, many numerical methods produce significant errors near sonic points (points where the wavespeed equals zero). Upwind methods in particular are forced to give sonic points special consideration since the upwind direction changes at sonic points. Shock waves, contact discontinuities, and expansion fans may also lead to a variety of serious problems including oscillations, overshoots, and smearing that can spread discontinuities over several cells. In particular, contact discontinuities do not have any physical compression and thus smearing increases progressively with the number of time steps. Even when approximating smooth solutions, most numerical methods exhibit obvious flaws. For example, many stable numerical methods continuously erode the solution, leading to amplitude and dissipation errors [13].

To investigate the behavior of our time-stepping schemes, we consider three of Laney's five test problems [13]. These three problems involve all of the important flow features identified above: shocks, contacts, expansion fans, sonic points, and smooth solutions. Similar to Laney, we focus on the behavior of the numerical scheme for interior regions rather than boundaries and impose periodic boundary conditions on the domain $[-1, 1]$. It is known that sometimes a conventional (and intuitive!) treatment of the boundary data (especially in the case of inflow boundary conditions) within the stages of a Runge–Kutta method can lead to deterioration in the overall accuracy of the integration. We refer to [1] and references therein for a discussion of this problem and a method for its resolution. The spatial discretization and the results of the three test cases follow.

4.2. Spatial discretization. SSPRK schemes are natural candidates for any method-of-lines discretization involving nonsmooth solutions. Similar to the original paper on SSPRK methods [18], we choose finite-difference Shu–Osher methods (ENO) to spatially discretize the equations. These methods are derived using flux reconstruction and have a variety of desirable properties. For example, they naturally extend to an arbitrary order of accuracy in space, and they are independent of the time discretization, thus allowing experimentation with different time discretization methods. Moreover, educational codes are also freely available [13, 12], an attribute which is desirable for standardizing numerical studies. Our simulations are carried out with a discretization that has the same order of accuracy in Δx as the time discretization accuracy p. We further note that flux splitting is carried out according to

$$f^+(U) = \frac{1}{2}(f(U) + \alpha^n_{i+1/2}U),$$
$$f^-(U) = \frac{1}{2}(f(U) - \alpha^n_{i+1/2}U),$$

where $\alpha^n_{i+1/2} = \max\{|f'(U^n_{i+1})|, |f'(U^n_i)|\}$. For full details on the discretization as well as code, see [13, 12].

It is noteworthy that high-order, fully TVD spatial discretization schemes are also available; see Osher and Chakravarthy [15]. In these numerical studies, we

choose Shu–Osher spatial discretization schemes rather than TVD schemes since TVD schemes obtain only between first- and second-order accuracy at extrema and they have "been largely superseded by Shu and Osher's class of high-order ENO methods" [13].

It is also noteworthy that recent variations on Shu–Osher methods such as methods based on weighted essentially nonoscillatory (WENO) reconstructions (e.g., [14, 11]) also naturally combine with SSPRK schemes. See [13] for detailed discussions on these and other spatial discretizations appropriate for hyperbolic conservation laws.

4.3. Test Case 1: Linear advection of a sinusoid. In this test case, the smooth initial conditions

$$u(x,0) = -\sin(\pi x)$$

are evolved to time $t = 30$ according to the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

using a constant grid spacing of $\Delta x = 1/320$. Since this evolution causes the initial conditions to travel around the periodic domain $[-1, 1]$ exactly 15 times, it is clear that the exact solution is just $u(x, 30) = -\sin(\pi x)$. Test Case 1 effectively illustrates the evolution of a smooth solution with no sonic points and is useful for verifying convergence rates for high-order schemes. Moreover, even on completely smooth solutions most numerical methods designed for hyperbolic conservation laws exhibit obvious flaws [13]. This test case is quite helpful for understanding phase and amplitude errors but should not be used to study dispersion because only one frequency is present in the exact solution. It is also informative to contrast these results with those derived for problems involving shocks and other discontinuities.

To quantify the accuracy of the computed solution, we use the logarithm of the l_1 errors, i.e.,

$$\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}|U_i - u(x_i, 30)|\right),$$

where N is the number of grid points and x_i is the ith grid node. A plot of the error is given in Figure 4.1. To ensure a fair comparison for methods with a different number of stages, the error is plotted as a function of the effective CFL number[5] rather than the CFL number itself. This implies that for a particular plot, the total number of function evaluations at a particular abscissa value will be the same for each scheme. We start calculating errors for an effective CFL number of 0.6 and continue until the numerical method is so unstable that a value of `NaN` is returned; i.e., the scheme has become completely unstable.

In this smooth test example, the new second-order schemes give improved stability and accuracy over the original SSP(2,2). Also, SSP(5,3) gives improved stability over SSP(3,3) and SSP(4,3). Calculations for low-storage schemes show that LS(5,3) outperforms both LS(4,3) and LS(3,3). (For clarity, we use arrows to indicate the exact points at which SSP(3,3) and LS(3,3) go completely unstable.)

[5]Similar to the definition of an effective CFL coefficient, the *effective CFL number* of an SSPRK method of order p is $\frac{\Delta t}{\Delta x}\frac{s^*}{s}$, where s^* is the minimum number of stages to theoretically achieve order p, and s is the number of stages required for one step of the method.

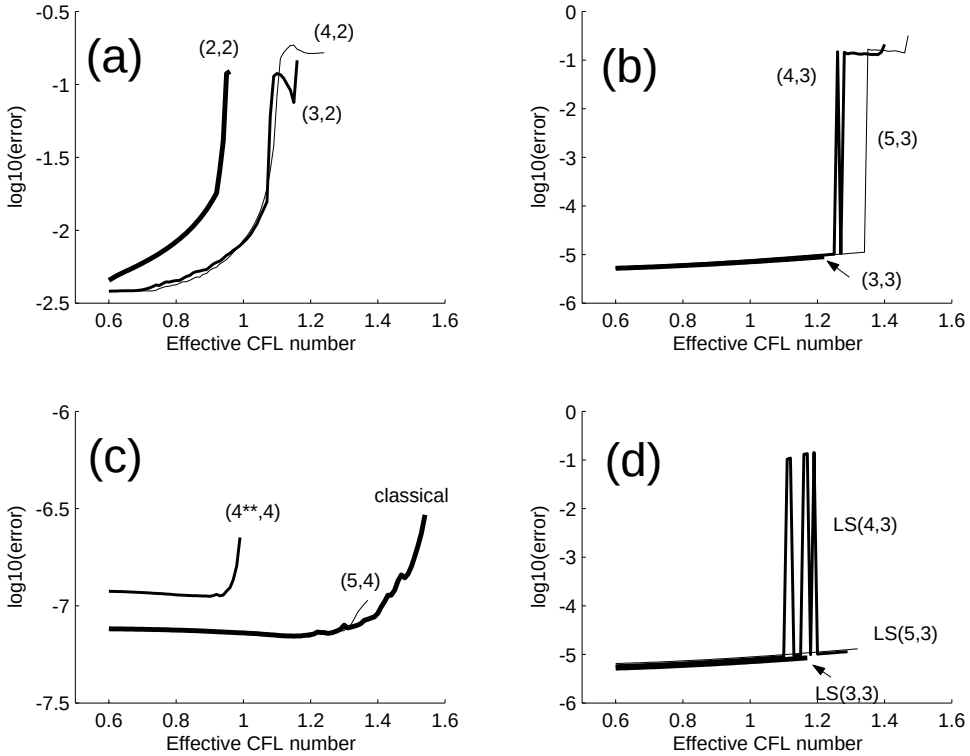


FIG. 4.1. l_1 *errors as a function of the effective CFL number.* (a) *Second-order schemes;* (b) *third-order schemes;* (c) *fourth-order schemes;* (d) *low-storage schemes.*

Based on these plots, we see that the second-order, third-order, and low-storage schemes all give stability restrictions that are within 20% of one another. This contrasts sharply with the results for fourth-order schemes (plot (c)). Here the new SSP(5,4) scheme gives more than a 40% improvement in the stability time-step restriction over the original SSP(4**,4). Moreover, it produces a marked reduction in the error, signifying a smaller error constant for this problem. It is noteworthy that in this case the classical fourth-order Runge-Kutta scheme outperforms even SSP(5,4): on *smooth* problems, schemes based purely on a linear stability analysis are expected to perform well. SSP schemes are designed to outperform on problems involving discontinuities in the solution or its derivatives, so in this case there is no reason to expect that schemes derived using nonlinear stability analysis will necessarily outperform classical schemes based on linear stability analysis.

4.4. Test Case 2: Linear advection of a square wave. In this test case, the discontinuous initial conditions

$$u(x,0) = \begin{cases} 1 & \text{for } |x| < 1/3, \\ 0 & \text{for } 1/3 < |x| \leq 1 \end{cases}$$

are evolved to time $t = 4$ according to the linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

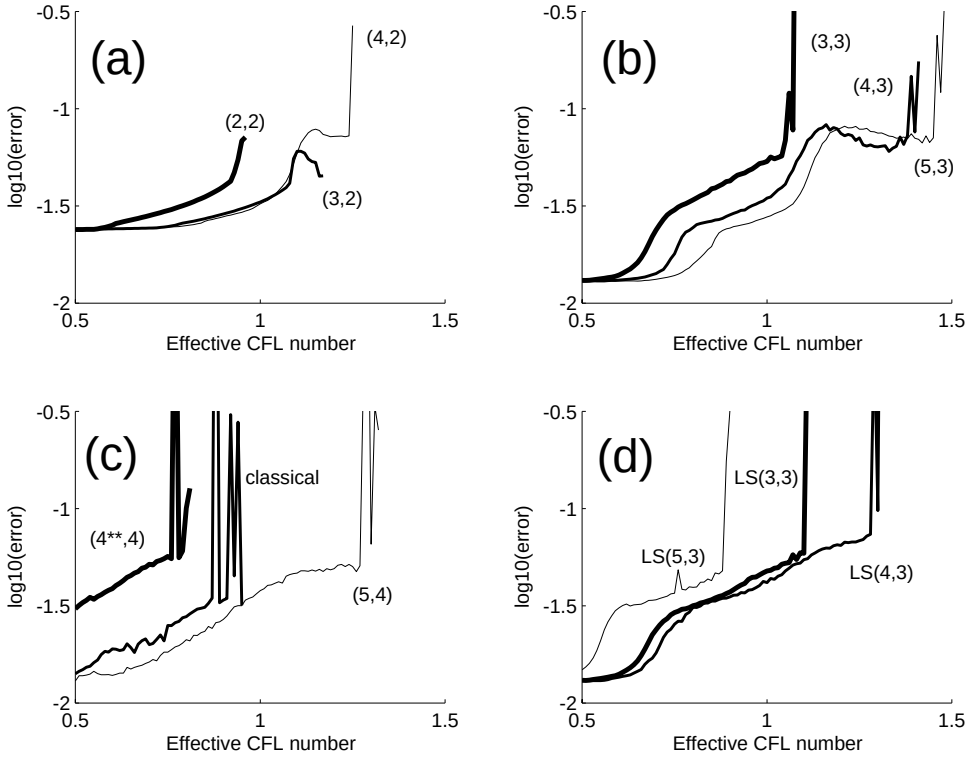


FIG. 4.2. *l_1 errors as a function of the effective CFL number.* (a) *Second-order schemes;* (b) *third-order schemes;* (c) *fourth-order schemes;* (d) *low-storage schemes.*

using a constant grid spacing of $\Delta x = 1/320$. Since this evolution causes the initial conditions to travel around the periodic domain $[-1, 1]$ exactly 2 times, it is clear that the exact solution at the final time is just $u(x, 4) = u(x, 0)$. Test Case 2 exhibits two jump discontinuities in the solution that correspond to contact discontinuities. This test case nicely illustrates progressive contact smearing and dispersion.

The log of the l_1 errors as a function of the effective CFL number is plotted in Figure 4.2. Based on these plots, it is immediately clear that a material improvement in both stability and accuracy is obtained using our new schemes.

For example, plot (a) shows that SSP(3,2) and SSP(4,2) allow about a 20–30% improvement in the time-step restriction over the original SSP(2,2). It is also clear that the new schemes also give a substantial improvement in stability and accuracy in the third-order case (b). Here we find that the optimal SSP(5,3) scheme gives about a 40% improvement in the stability time-step restriction over the usual SSP(3,3).

In the fourth-order case (c), even greater improvements are observed. SSP(5,4) gives more than a 60% improvement in the stability time-step restriction and requires only half the number of function evaluations to achieve an error of $10^{-1.5}$. Moreover, SSP(5,4) is clearly superior to the classical fourth-order Runge–Kutta scheme, with more than a 40% improvement in the observed time-step restriction. As conjectured, the best SSP schemes outperform classical (but generally non-SSP) schemes when discontinuities in the solution arise.

Out of the low-storage schemes (plot (d)), the new LS(4,3) gives the best

performance. It is interesting that the best-performing scheme LS(4,3) requires one-third less storage and is more CPU-efficient than the standard SSP(3,3) in this test example.

4.5. Test Case 3: Evolution of a square wave by Burgers's equation. In this test case, the discontinuous initial conditions

$$u(x,0) = \begin{cases} 1 & \text{for } |x| < 1/3, \\ -1 & \text{for } 1/3 < |x| \le 1 \end{cases}$$

are evolved to time $t = 0.3$ according to Burgers's equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right) = 0$$

using a constant grid spacing of $\Delta x = 1/320$. In this example, the jump at $x = -1/3$ creates a simple centered expansion fan and the jump at $x = 1/3$ creates a steady shock. Until the shock and expansion fan intersect (at time $t = 2/3$), the exact solution is

$$u(x,t) = \begin{cases} -1 & \text{for } -\infty < x < b_1, \\ -1 + 2\frac{x-b_1}{b_2-b_1} & \text{for } b_1 < x < b_2, \\ 1 & \text{for } b_2 < x < b_{shock}, \\ -1 & \text{for } b_{shock} < x < \infty, \end{cases}$$

where $b_1 = -1/3 - t$, $b_2 = -1/3 + t$, and $b_{shock} = 1/3$ [13]. Test Case 3 is particularly interesting because it illustrates the behaviors near sonic points ($u = 0$) that correspond to an expansion fan and a compressive shock.

The log of the l_1 errors as a function of the effective CFL number is plotted in Figure 4.3. Based on these plots, it is clear that a marked improvement in both stability and accuracy is obtained in the second-, third-, and fourth-order cases using our new schemes.

Once again, plot (a) shows that SSP(3,2) and SSP(4,2) show about a 20–30% improvement in the time-step restriction over the original SSP(2,2). It is also clear that the new schemes also give a substantial improvement in stability and accuracy in the third-order case (b). Here we find that the optimal SSP(5,3) scheme gives about a 20% improvement in the stability time-step restriction over the usual SSP(3,3).

In the fourth-order case (c), even greater improvements are observed than in Test Case 2. SSP(5,4) gives an 80% improvement in the stability time-step restriction and requires only one-third the number of function evaluations to achieve an error of $10^{-2.6}$. Moreover, SSP(5,4) is clearly superior to the classical fourth-order Runge–Kutta scheme, with more than a 60% improvement in the observed time-step restriction. Similar to the previous example, SSP(5,4) outperforms classical (but generally non-SSP) schemes when discontinuities in the solution arise.

Out of the low-storage schemes (plot (d)), LS(3,3) and the new LS(4,3) give the best performance. In this test case, the best low-storage schemes are nearly as CPU-efficient as SSP(3,3) but require one-third less storage.

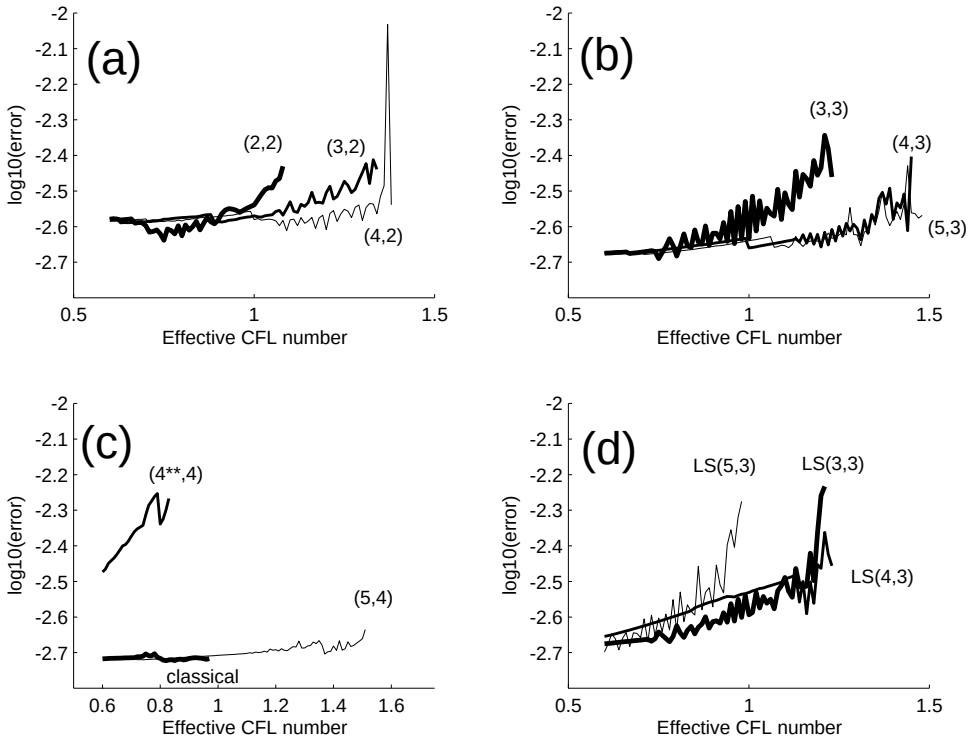


FIG. 4.3. l_1 *errors as a function of the effective CFL number.* (a) *Second-order schemes;* (b) *third-order schemes;* (c) *fourth-order schemes;* (d) *low-storage schemes.*

5. Summary and future work. We have presented new optimal SSPRK time discretization methods of orders 1 through 4 and stages 1 through 5. We find that, by allowing the number of stages to differ from the order of the method, it is possible to derive schemes with better, more effective CFL coefficients than those that are most commonly used. We have also performed a comparison of the new methods with Runge–Kutta methods (both SSP and non-SSP) most commonly used in practice on three problems involving scalar conservation laws. Our new methods compare favorably in terms of computational efficiency per time step, especially when the solution exhibits discontinuous behavior. The improvements are the greatest for the new fourth-order scheme with 5 stages (SSP(5,4)), where the allowable time step is significantly greater than the Shu–Osher fourth-order scheme and the classical fourth-order explicit Runge–Kutta scheme.

We also give results of a similar treatment of low-storage SSPRK schemes, where again we find significant improvements over the schemes most commonly used. The results are for orders 1 through 3 and stages 1 through 5. We were unable to find a low-storage scheme of order 4 having only 5 stages.

We have already examined the possibility of finding even more efficient SSPRK schemes by lifting the positivity constraint on the coefficients β_{ij}. Not surprisingly, improvements in the raw CFL coefficient are possible; however, the reduction in the effective CFL coefficient necessitated by the introduction of $\tilde{L}(\cdot)$ whenever $\beta_{ij} < 0$ causes these methods to be uncompetitive.

We are currently extending our investigation of optimal SSPRK methods to

methods having more than 5 stages and to orders 4 and 5. This work includes the study of low-storage SSPRK methods of order 4. We have also derived families of embedded SSPRK schemes for local error estimation and step-size control. We report on these findings elsewhere.

Appendix A. Optimal $(\alpha_{ik}, \beta_{ik})$ for $p = 3, 4$. Tables A.1 and A.2, respectively, give the optimal SSPRK methods of orders 3 and 4 and up to 5 stages in the representation (2.1).

TABLE A.1
The first few optimal SSPRK schemes of order 3.

Stages	α_{ik}				β_{ik}				CFL coefficient
3	1				1				1
	$\frac{3}{4}$	$\frac{1}{4}$			0	$\frac{1}{4}$			
	$\frac{1}{3}$	0	$\frac{2}{3}$		0	0	$\frac{2}{3}$		
4	1				$\frac{1}{2}$				2
	0	1			0	$\frac{1}{2}$			
	$\frac{2}{3}$	0	$\frac{1}{3}$		0	0	$\frac{1}{6}$		
	0	0	0	1	0	0	0	$\frac{1}{2}$	

Stages			5		
	1				
	0	1			
α_{ik}	0.56656131914033	0	0.43343868085967		
	0.09299483444413	0.00002090369620	0	0.90698426185967	
	0.00736132260920	0.20127980325145	0.00182955389682	0	0.78952932024253
	0.37726891511710				
	0	0.37726891511710			
β_{ik}	0	0	0.16352294089771		
	0.00071997378654	0	0	0.34217696850008	
	0.00277719819460	0.00001567934613	0	0	0.29786487010104
CFL coefficient			2.65062919294483		

TABLE A.2
The coefficients of the optimal SSPRK (5,4) *scheme.*

Stages			5		
	1				
	0.44437049406734	0.55562950593266			
α_{ik}	0.62010185138540	0	0.37989814861460		
	0.17807995410773	0	0	0.82192004589227	
	0.00683325884039	0	0.51723167208978	0.12759831133288	0.34833675773694
	0.39175222700392				
	0	0.36841059262959			
β_{ik}	0	0	0.25189177424738		
	0	0	0	0.54497475021237	
	0	0	0	0.08460416338212	0.22600748319395
CFL coefficient			1.50818004975927		

Appendix B. Butcher array forms of SSPRK schemes. The following are the Butcher array representations of the optimal SSPRK schemes given in this paper.

Order 1:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \hline & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}$$

Order 2:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \hline & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array}$$

Order 3:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2} \end{array}$$

0	0	0	0	0	0
0.37726891511710	0.37726891511710	0	0	0	0
0.75453783023419	0.37726891511710	0.37726891511710	0	0	0
0.49056882269314	0.16352294089771	0.16352294089771	0.16352294089771	0	0
0.78784303014311	0.14904059394856	0.14831273384724	0.14831273384724	0.34217696850008	0
	0.19707596384481	0.11780316509765	0.11709725193772	0.27015874934251	0.29786487010104

Order 4:

0	0	0	0	0	0
0.39175222700392	0.39175222700392	0	0	0	0
0.58607968896779	0.21766909633821	0.36841059262959	0	0	0
0.47454236302687	0.08269208670950	0.13995850206999	0.25189177424738	0	0
0.93501063100924	0.06796628370320	0.11503469844438	0.20703489864929	0.54497475021237	0
	0.14681187618661	0.24848290924556	0.10425883036650	0.27443890091960	0.22600748319395

Appendix C. Butcher array forms of low-storage schemes. The optimal low-storage SSPRK schemes of order 1 and order 2 occur when $s = p$ and have already been given both in terms of representation (2.1) and Butcher arrays. Here we provide the Butcher array representation of the third-order schemes presented in Table 3.7.

Order 3:

0	0	0	0
0.92457411523577	0.92457411523577	0	0
0.37346170537554	0.08574876388805	0.28771294148749	0
	0.08574876111733	0.28771294243783	0.62653829645172

0	0	0	0	0
1.03216665875130	1.03216665875130	0	0	0
0.29044361656735	0.10250480393024	0.18793881263711	0	0
0.44260113480482	0.10250480354712	0.18793881271456	0.15215751854315	0
	0.10250480379728	0.18793881266399	0.05280467502407	0.65675174856653

0	0	0	0	0	0
0.67892607116139	0.67892607116139	0	0	0	0
0.34677649493991	0.14022991560621	0.20654657933371	0	0	0
0.66673359500982	0.20569370073026	0.18144649137471	0.27959340290485	0	0
0.76590087429032	0.16104646283838	0.19856511041100	0.08890670263481	0.31738259840613	0
	0.19215670424132	0.18663683901393	0.22177739201759	0.09623007655432	0.30319904778284

Acknowledgments. The authors would like to express their thanks to J. Borwein and W. Sutherland for helpful discussions.

REFERENCES

[1] S. Abarbanel, D. Gottlieb, and M. H. Carpenter, *On the removal of boundary errors caused by Runge–Kutta integration of nonlinear partial differential equations*, SIAM J. Sci. Comput., 17 (1996), pp. 777–782.
[2] V. Chvátal, *Linear Programming*, W. H. Freeman and Company, New York, 1983.
[3] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons Ltd., Chichester, 1987.
[4] A. Gerisch and R. Weiner, *On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods*, Comput. Math. Appl., to appear.
[5] J. B. Goodman, R. J. LeVeque, and J. Randall, *On the accuracy of stable schemes for 2D scalar conservation laws*, Math. Comp., 45 (1985), pp. 15–21.
[6] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.
[7] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
[8] E. Hairer, S. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.
[9] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.
[10] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high-order accurate essentially nonoscillatory schemes.* III, J. Comput. Phys., 71 (1987), pp. 231–303.
[11] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[12] C. Laney, *CFD Recipes: Software for Computational Gasdynamics*, Cambridge University Press, Cambridge, UK, 1998. Available online at http://capella.colorado.edu/∼laney/booksoft.htm.

[13] C. Laney, *Computational Gasdynamics*, Cambridge University Press, Cambridge, UK, 1998.
[14] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially nonoscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.
[15] S. Osher and S. Chakravarthy, *Very high order accurate TVD schemes*, in Oscillation Theory, Computation, and Methods of Compensated Compactness, IMA Vol. Math. Appl. 2, C. Dafermos, J. Erikson, D. Kinderlehrer, and M. Slemrod, eds., Springer-Verlag, New York, 1986, pp. 229–271.
[16] S. Ruuth and R. Spiteri, *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211-220.
[17] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.
[18] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[19] B. van Leer, *Towards the ultimate conservative difference scheme.* V. *A second-order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.
[20] J. H. Williamson, *Low-storage Runge-Kutta schemes*, J. Comput. Phys., 35 (1980), pp. 48–56.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 492–515

GALERKIN PROPER ORTHOGONAL DECOMPOSITION METHODS FOR A GENERAL EQUATION IN FLUID DYNAMICS*

K. KUNISCH† AND S. VOLKWEIN†

Abstract. Error estimates for Galerkin proper orthogonal decomposition (POD) methods for nonlinear parabolic systems arising in fluid dynamics are proved. For the time integration the backward Euler scheme is considered. The asymptotic estimates involve the singular values of the POD snapshot set and the grid-structure of the time discretization as well as the snapshot locations.

Key words. proper orthogonal decomposition, evolution problems, Navier–Stokes equations, error estimates

AMS subject classifications. 35K20, 65N

PII. S0036142900382612

1. Introduction.

Proper orthogonal decomposition (POD) provides a method for deriving low order models of dynamical systems. It can be thought of as a Galerkin approximation in the spatial variable, built from functions corresponding to the solution of the physical system at prespecified time instances. These are called the snapshots. Due to possible linear dependence or almost linear dependence, the snapshots themselves are not appropriate as a basis. Instead, a singular value decomposition is carried out and the leading generalized eigenfunctions are chosen as a basis, referred to as the POD basis.

POD was successfully used in a variety of fields including signal analysis and pattern recognition (see, e.g., [10]), fluid dynamics and coherent structures (see, e.g., [7, 22, 23, 24]), and more recently in control theory (see, e.g., [1, 3, 6, 15, 17, 18, 21]) and inverse problems (see [5]). Good approximation properties are reported for POD based schemes in several articles; see [8, 9, 14, 19], for example. Symmetry preserving properties of POD approximations are analyzed in [4].

As soon as one uses POD, questions concerning the quality of the approximation properties, convergence, and rate of convergence become relevant. It appears that, except for the work in [16], these issues have not been addressed. This may be due, in part, to the fact that for POD based approximation of partial differential equations one cannot rely on results clarifying the approximation properties of the POD-subspaces to elements in function spaces as, e.g., L^p or C. Such results are an essential building block for, e.g., finite element approximations to partial differential equations. In our work we propose a strategy to describe and analyze convergence and rate of convergence approximations based on POD approximation in space and backwards Euler discretization in time for a class of nonlinear evolutionary partial differential equations including the Navier–Stokes equation in two dimensions. Due to the lack of typical function space approximation results described above, these results are not a priori in the format that one is familiar with from finite difference, finite element, or spectral approximations, for example. While we believe that the estimates that we propose reflect well the properties of POD based approximations,

*Received by the editors December 18, 2000; accepted for publication (in revised form) December 7, 2001; published electronically May 29, 2002.
http://www.siam.org/journals/sinum/40-2/38261.html
†Karl-Franzens-Universität Graz, Institut für Mathematik, Heinrichstraße 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at, stefan.volkwein@uni-graz.at).

they are certainly up for discussion and future improvement. Roughly, it will be shown that the approximation error can be decomposed in a contribution that arises due to the POD approximation in space, which is measured in terms of spectral properties specifying the POD basis, and in the usual approximation error due to the backwards Euler scheme with respect to time integration.

Concerning the availability of snapshots, two situations can be considered: in the first one the snapshots are obtained by an independent numerical method and then used within a POD approach for the sake of system reduction. In the second situation the snapshots could be obtained and digitalized from actual physical phenomena.

In this article we analyze the case in which snapshots are assumed to be available for the same system as that for which the approximation properties are analyzed. The value of this analysis is to understand and justify analytically the observed high approximation quality of POD schemes. Certainly the problem of quantifying the approximation properties of POD based schemes where the snapshots are taken from processes which are possibly "nearby," but different from the system under consideration, is of considerable interest. This situation can occur, for example, in the case of control and optimal control of systems: snapshots are taken from a system at a nominal control value, and a POD model reduction and subsequent optimal control step are performed. The dynamics of the optimally controlled system then differ from the original system. The analysis of these problems can be the focus of future research.

As already mentioned, the present work is a continuation of our efforts to approximation properties of POD based schemes. We extend our earlier results of [16] in three directions. First, the results of the present paper are asymptotic results in the sense that the constants appearing in the estimates do not depend on the snapshot set. Second, we now utilize two time discretizations, one for the set of snapshots and a second one for the numerical integration. The effect of the two grids on the convergence rate is kept separate in the estimates. Third, we focus in this paper on a different class of nonlinearities, including the Navier–Stokes equations in dimension two, which were not included in [16].

The paper is organized as follows. In section 2 the nonlinear evolution problem is introduced and necessary prerequisites are given. The POD method is reviewed in section 3. Convergence of the backward Euler scheme is studied in section 4. Technical proofs are deferred to appendices.

2. General equations in fluid dynamics. In this section we specify the abstract nonlinear evolution problem that will be considered in this paper and present an existence and uniqueness result.

Let V and H be real separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle \cdot\,, \cdot \rangle_H$ we denote the inner product in H. The inner product in V is given by a symmetric bounded, coercive, bilinear form $a : V \times V \to \mathbb{R}$:

$$\langle \varphi, \psi \rangle_V = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V \tag{2.1}$$

with an associated norm given by $\| \cdot \|_V = \sqrt{a(\cdot\,, \cdot)}$. Since V is continuously injected into H, there exists a constant $c_V > 0$ such that

$$\|\varphi\|_H \le c_V \, \|\varphi\|_V \quad \text{for all } \varphi \in V. \tag{2.2}$$

We associate with a the linear operator A:

$$\langle A\varphi, \psi \rangle_{V',V} = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V,$$

where $\langle \cdot, \cdot \rangle_{V',V}$ denotes the duality pairing between V and its dual. Then A is an isomorphism from V onto V'. Alternatively, A can be considered as a linear unbounded self-adjoint operator in H with domain

$$D(A) = \{\varphi \in V : A\varphi \in H\}.$$

By identifying H and its dual H' it follows that

$$D(A) \hookrightarrow V \hookrightarrow H = H' \hookrightarrow V',$$

each embedding being continuous and dense, when $D(A)$ is endowed with the graph norm of A.

We introduce the continuous operator $R : V \to V'$, which maps $D(A)$ into H and satisfies

$$\begin{aligned} \|R\varphi\|_H &\le c_R \, \|\varphi\|_V^{1-\delta_1} \|A\varphi\|_H^{\delta_1} \quad \text{for all } \varphi \in D(A), \\ |\langle R\varphi, \varphi \rangle_{V',V}| &\le c_R \, \|\varphi\|_V^{1+\delta_2} \|\varphi\|_H^{1-\delta_2} \quad \text{for all } \varphi \in V \end{aligned} \tag{2.3}$$

for a constant $c_R > 0$ and for $\delta_1, \delta_2 \in [0, 1)$. We also assume that $A + R$ is coercive on V; i.e., there exists a constant $\eta > 0$ such that

$$a(\varphi, \varphi) + \langle R\varphi, \varphi \rangle_{V',V} \ge \eta \, \|\varphi\|_V^2 \quad \text{for all } \varphi \in V. \tag{2.4}$$

Moreover, let $B : V \times V \to V'$ be a bilinear continuous operator mapping $D(A) \times D(A)$ into H such that there exist constants $c_B > 0$ and $\delta_3, \delta_4, \delta_5 \in [0, 1)$ satisfying

$$\begin{aligned} \langle B(\varphi, \psi), \psi \rangle_{V',V} &= 0, \\ \big|\langle B(\varphi, \psi), \phi \rangle_{V',V}\big| &\le c_B \, \|\varphi\|_H^{\delta_3} \|\varphi\|_V^{1-\delta_3} \|\psi\|_V \|\phi\|_V^{\delta_3} \|\phi\|_H^{1-\delta_3}, \\ \|B(\varphi, \chi)\|_H + \|B(\chi, \varphi)\|_H &\le c_B \, \|\varphi\|_V \|\chi\|_V^{1-\delta_4} \|A\chi\|_H^{\delta_4}, \\ \|B(\varphi, \chi)\|_H &\le c_B \, \|\varphi\|_H^{\delta_5} \|\varphi\|_V^{1-\delta_5} \|\chi\|_V^{1-\delta_5} \|A\chi\|_H^{\delta_5} \end{aligned} \tag{2.5}$$

for all $\varphi, \psi, \phi \in V$, for all $\chi \in D(A)$. To simplify the notation we set $B(\varphi) = B(\varphi, \varphi)$ for $\varphi \in V$.

For given $f \in L^2(0, T; H)$ and $y_0 \in V$ we consider the nonlinear evolution problem

$$\frac{d}{dt} \, \langle y(t), \varphi \rangle_H + a(y(t), \varphi) + \big\langle B(y(t)) + Ry(t), \varphi \big\rangle_{V',V} = \langle f(t), \varphi \rangle_H \tag{2.6a}$$

for all $\varphi \in V$ and $t \in (0, T]$ a.e. and

$$y(0) = y_0 \quad \text{in } H. \tag{2.6b}$$

The following theorem guarantees the existence of a unique solution to (2.6).

THEOREM 2.1. *Assume that* (2.3) *and* (2.5) *hold. Then for every* $f \in L^2(0, T; H)$ *and* $y_0 \in V$ *there exists a unique solution of* (2.6) *satisfying*

$$y \in C([0, T]; V) \cap L^2(0, T; D(A)) \cap H^1(0, T; H). \tag{2.7}$$

Proof. The proof is analogous to that of Theorem 2.1 in [25, p. 111], where the case with time-independent f was treated. □

Condition (2.4) will not be needed before section 4. Let us present an example for the nonlinear evolution system (2.6).

EXAMPLE 2.2. *Let Ω denote a bounded domain in $\mathbb{R}^2$ with boundary Γ and let $T > 0$. The two-dimensional Navier–Stokes equations are given by*

$$\varrho\left(u_t + (u\cdot\nabla)u\right) - \nu\Delta u + \nabla p = f \quad \text{in } Q = (0,T)\times\Omega, \tag{2.8a}$$

$$\operatorname{div} u = 0 \quad \text{in } Q, \tag{2.8b}$$

where $\varrho > 0$ is the density of the fluid, $\nu > 0$ is the kinematic viscosity, f represents volume forces, and

$$(u\cdot\nabla)u = \Big(u_1\frac{\partial u_1}{\partial x_1} + u_2\frac{\partial u_1}{\partial x_2}, u_1\frac{\partial u_2}{\partial x_1} + u_2\frac{\partial u_2}{\partial x_2}\Big)^{\mathsf{T}}.$$

The unknowns are the velocity field $u = (u_1, u_2)$ and the pressure p. Together with (2.8) *we consider nonslip boundary conditions*

$$u = u_d \quad \text{on } \Sigma = (0,T)\times\Gamma \tag{2.8c}$$

and the initial condition

$$u(0,\cdot) = u_0 \quad \text{in } \Omega. \tag{2.8d}$$

In [25, pp. 104–107, 116–117] *it was proved that* (2.8) *can be written in the form* (2.6).

Next we recall Young's inequality, which will frequently be used in our work. For a proof we refer to [2, p. 28], for instance.

LEMMA 2.3 (Young's inequality). *For all $a, b, \varepsilon > 0$ and for all $p \in (1,\infty)$ we have*

$$ab \le \frac{\varepsilon a^p}{p} + \frac{b^q}{q\varepsilon^{q/p}},$$

where $q = p/(p-1)$.

3. The POD method. This section is devoted to a discussion of the POD method for the nonlinear evolution problem (2.6). Throughout we denote by y the unique solution to (2.6) satisfying (2.7). Moreover, we suppose that $f \in C([0,T];H)$.

3.1. Computation of the POD basis. For given $n \in \mathbb{N}$ let

$$0 = t_0 < t_2 < \cdots < t_n \le T$$

denote a grid in the interval $[0,T]$ and set $\delta t_j = t_j - t_{j-1}$, $j = 1,\dots,n$. Define

$$\Delta t = \max\ (\delta t_1,\dots,\delta t_n) \quad \text{and} \quad \delta t = \min\ (\delta t_1,\dots,\delta t_n).$$

Suppose that the snapshots $y(t_j)$ of (2.6) at the given time instances t_j, $j = 0,\dots,n$, are known. We set

$$\mathcal{V} = \operatorname{span}\{y(t_0),\dots,y(t_n)\}$$

and refer to $\mathcal{V}$ as the ensemble consisting of the snapshots $\{y(t_j)\}_{j=0}^n$, at least one of which is assumed to be nonzero. Notice that $\mathcal{V} \subset V$ by construction. Throughout the remainder of this section we let X denote either the space V or H.

Let $\{\psi_i\}_{i=1}^d$ denote an orthonormal basis for $\mathcal{V}$ with $d = \dim \mathcal{V}$. Then each member of the ensemble can be expressed as

$$y(t_j) = \sum_{i=1}^{d} \langle y(t_j), \psi_i \rangle_X \psi_i \quad \text{for } j = 0, \dots, n. \tag{3.1}$$

The method of POD consists of choosing an orthonormal basis such that for every $\ell \in \{1, \dots, d\}$ the mean square error between the elements $y(t_j)$, $0 \le j \le n$, and the corresponding ℓth partial sum of (3.1) is minimized on average:

$$\begin{aligned} &\min_{\{\psi_i\}_{i=1}^{\ell}} \sum_{j=0}^{n} \alpha_j \left\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i \rangle_X \psi_i \right\|_X^2 \\ &\text{subject to } \langle \psi_i, \psi_j \rangle_X = \delta_{ij} \quad \text{for } 1 \le i \le \ell, 1 \le j \le i. \end{aligned} \tag{3.2}$$

Here $\{\alpha_j\}_{j=0}^n$ are positive weights, which for our purposes are chosen to be

$$\alpha_0 = \frac{\delta t_1}{2}, \quad \alpha_j = \frac{\delta t_j + \delta t_{j+1}}{2} \text{ for } j = 1, \dots, n-1, \quad \text{and} \quad \alpha_n = \frac{\delta t_n}{2}.$$

A solution $\{\psi_i\}_{i=1}^{\ell}$ to (3.2) is called a POD basis of rank ℓ. The subspace spanned by the first ℓ POD basis functions is denoted by V^ℓ.

Remark 3.1. Note that

$$\mathcal{I}_n(y) = \sum_{j=0}^{n} \alpha_j \left\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i \rangle_X \psi_i \right\|_X^2$$

is the trapezoidal approximation for the integral

$$\mathcal{I}(y) = \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_X \psi_i \right\|_X^2 dt.$$

For all $y \in C([0,T]; X)$ it follows that $\lim_{n\to\infty} \mathcal{I}_n(y) = \mathcal{I}(y)$.

The solution of (3.2) is characterized by the necessary optimality condition. For that purpose we endow $\mathbb{R}^{n+1}$ with the weighted inner product

$$\langle v, w \rangle_{\mathbb{R}^{n+1}} = \sum_{j=0}^{n} \alpha_j v_j w_j \quad \text{for } v = (v_0, \dots, v_n)^{\mathsf{T}},\ w = (w_0, \dots, w_n)^{\mathsf{T}} \in \mathbb{R}^{n+1}.$$

Let us introduce the bounded linear operator $\mathcal{Y}_n : \mathbb{R}^{n+1} \to X$ by

$$\mathcal{Y}_n v = \sum_{j=0}^{n} \alpha_j v_j y(t_j) \quad \text{for } v \in \mathbb{R}^{n+1}.$$

Then the adjoint $\mathcal{Y}_n^* : X \to \mathbb{R}^{n+1}$ is given by

$$\mathcal{Y}_n^* z = (\langle z, y(t_0) \rangle_X, \dots, \langle z, y(t_n) \rangle_X)^{\mathsf{T}} \quad \text{for } z \in X.$$

It follows that $\mathcal{R}_n = \mathcal{Y}_n \mathcal{Y}_n^* \in \mathcal{L}(X)$ and $\mathcal{K}_n = \mathcal{Y}_n^* \mathcal{Y}_n \in \mathbb{R}^{(n+1)\times(n+1)}$ are given by

$$\mathcal{R}_n z = \sum_{j=0}^{n} \alpha_j \langle z, y(t_j) \rangle_X y(t_j) \text{ for } z \in X \quad \text{and} \quad (\mathcal{K}_n)_{ij} = \langle y(t_j), y(t_i) \rangle_X,$$

respectively. Here $\mathcal{L}(X)$ denotes the Banach space of all bounded linear operators on X.

Using a Lagrangian framework we derive the following optimality conditions for the optimization problem (3.2):

$$\mathcal{R}_n\psi = \lambda\psi; \tag{3.3}$$

compare, e.g., [7, 26]. Note that $\mathcal{R}_n$ is a bounded, self-adjoint and nonnegative operator. Moreover, since the image of $\mathcal{R}_n$ has finite dimensions, $\mathcal{R}_n$ is also compact. By Hilbert–Schmidt theory (see, e.g., [20, p. 203]) there exist an orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for X and a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

$$\mathcal{R}_n\psi_i = \lambda_i\psi_i, \quad \lambda_1 \geq \cdots \geq \lambda_d > 0, \quad \text{and } \lambda_i = 0 \text{ for } i > d. \tag{3.4}$$

Moreover, $\mathcal{V} = \operatorname{span}\{\psi_i\}_{i=1}^d$.

Note that $\{\lambda_i\}_{i\in\mathbb{N}}$ as well as $\{\psi_i\}_{i\in\mathbb{N}}$ depend on n. Contents permitting the notation of this dependence are dropped.

Remark 3.2. Setting

$$v_i = \frac{1}{\sqrt{\lambda_i}}\,\mathcal{Y}_n^*\psi_i \quad \text{for } i = 1, \dots, d$$

we find $\mathcal{K}_n v_i = \lambda_i v_i$ and $\langle v_i, v_j\rangle_{\mathbb{R}^{n+1}} = \delta_{ij}$ for $1 \leq i, j \leq d$. Thus, $\{v_i\}_{i=1}^d$ is an orthonormal basis of eigenvectors of $\mathcal{K}_n$ for the image of $\mathcal{K}_n$. Conversely, if $\{v_i\}_{i=1}^d$ is a given orthonormal basis for the image of $\mathcal{K}_n$, then it follows that the first d eigenfunctions of $\mathcal{R}_n$ can be determined by

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\,\mathcal{Y}_n v_i \quad \text{for } i = 1, \dots, d.$$

The sequence $\{\psi_i\}_{i=1}^\ell$ solves the optimization problem (3.2). This fact as well as the error formula below were proved in [7, 26], for example.

PROPOSITION 3.3. *Let $\lambda_1 \geq \cdots \geq \lambda_d > 0$ denote the positive eigenvalues of $\mathcal{R}^n$ with the associated eigenvectors $\psi_1, \dots, \psi_d \in X$. Then, $\{\psi_i^n\}_{i=1}^\ell$ is a POD basis of rank $\ell \leq d$, and we have the error formula*

$$\sum_{j=0}^{n} \alpha_j \left\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i\rangle_X \psi_i \right\|_X^2 = \sum_{i=\ell+1}^{d} \lambda_i. \tag{3.5}$$

3.2. Perturbation analysis for $\sum_{i=\ell+1}^{d} \lambda_i$. The eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ depend on the time instances $\{t_j\}_{j=0}^n$. Next we investigate $\sum_{i=\ell+1}^d \lambda_i$ as Δt tends to zero, i.e., $n \to \infty$. Let us define the bounded linear operator $\mathcal{Y} : L^2(0, T; \mathbb{R}) \to X$ by

$$\mathcal{Y}\varphi = \int_0^T \varphi(t) y(t)\, dt \quad \text{for } \varphi \in L^2(0, T; \mathbb{R}).$$

The adjoint $\mathcal{Y}^* : X \to L^2(0, T; \mathbb{R})$ is given by

$$\big(\mathcal{Y}^* z\big)(t) = \langle z, y(t)\rangle_X \quad \text{for } z \in X.$$

For $\mathcal{R} = \mathcal{Y}\mathcal{Y}^* \in \mathcal{L}(X)$ we find

$$\mathcal{R}z = \int_0^T \langle z\,, y(t)\rangle_X y(t)\, dt \quad \text{for } z \in X. \tag{3.6}$$

Notice that $\mathcal{R}_n\varphi$ is the trapezoidal approximation for the integral $\mathcal{R}\varphi$. If $y_t \in L^2(0,T;X)$, then we obtain

$$\lim_{\Delta t\to\infty} \|\mathcal{R}_n - \mathcal{R}\|_{\mathcal{L}(X)} = 0. \tag{3.7}$$

Let us mention that as far as the following analysis is concerned any other choice of positive weights α_j is possible provided that (3.7) holds.

We proceed to investigate the relationship between $\mathcal{R}_n$ and $\mathcal{R}$. Notice that $\mathcal{R}$ is self-adjoint and nonnegative. Since $y \in C([0,T];V)$, the Kolmogorov compactness criterion in $L^2(0,T;\mathbb{R})$ implies that $\mathcal{Y}^* : X \to L^2(0,T;X)$ is compact. Boundedness of $\mathcal{Y}$ implies that $\mathcal{R}$ is a compact operator as well. From the Hilbert–Schmidt theorem it follows that there exists a complete orthonormal basis $\{\psi_i^\infty\}_{i\in\mathbb{N}}$ for X and a sequence $\{\lambda_i^\infty\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

$$\mathcal{R}\psi_i^\infty = \lambda_i^\infty\psi_i^\infty, \quad \lambda_1^\infty \ge \lambda_2^\infty \ge \cdots, \quad \text{and } \lambda_i^\infty \to 0 \text{ as } i\to\infty. \tag{3.8}$$

Remark 3.4. Analogous to Remark 3.2 we set

$$v_i^\infty = \frac{1}{\sqrt{\lambda_i^\infty}}\,\mathcal{Y}^*\psi_i^\infty = \frac{1}{\sqrt{\lambda_i^\infty}}\,\langle\psi_i^\infty, y(t)\rangle_X\,dt \quad \text{for } i \in \{j\in\mathbb{N} : \lambda_j^\infty > 0\}.$$

Let $\mathcal{K} = \mathcal{Y}^*\mathcal{Y} \in \mathcal{L}(L^2(0,T;\mathbb{R}))$ be given by

$$\mathcal{K}\varphi = \int_0^T \langle y(s), y(t)\rangle_X \varphi(s)\,ds \quad \text{for } \varphi \in L^2(0,T;\mathbb{R}).$$

It follows that

$$\begin{aligned}
\big(\mathcal{K}v_i^\infty\big)(t) &= \int_0^T \langle y(s), y(t)\rangle_X v_i^\infty(s)\,ds \\
&= \frac{1}{\sqrt{\lambda_i^\infty}}\left\langle \int_0^T \langle\psi_i^\infty, y(s)\rangle_X y(s)\,ds, y(t)\right\rangle_X = \frac{1}{\sqrt{\lambda_i^\infty}}\,\langle\mathcal{R}\psi_i^\infty, y(t)\rangle_X \\
&= \frac{1}{\sqrt{\lambda_i^\infty}}\,\langle\mathcal{R}\psi_i^\infty, y(t)\rangle_X = \lambda_i^\infty v_i^\infty(t)
\end{aligned}$$

and, consequently, the v_i^∞'s are the eigenfunctions of $\mathcal{K}$ for $i\in\mathbb{N}$ with $\lambda_i^\infty > 0$.

The spectra of $\mathcal{R}$ and $\mathcal{R}_n$ are pure point spectra except for possibly 0. Each nonzero eigenvalue of $\mathcal{R}$ has finite multiplicity and 0 is the only possible accumulation point of the spectrum of $\mathcal{R}$; see [13, p. 185]. These facts together with (3.7) will allow us to draw important conclusions on the term $\sum_{i=\ell+1}^d \lambda_i^n$ in our estimates below. Henceforth we denote by $\{\lambda_i^n\}_{i=1}^{d(n)}$ the positive eigenvalues of $\mathcal{R}_n$ with associated eigenfunctions $\{\psi_i^n\}_{i=1}^{d(n)}$. Similarly $\{\lambda_i^\infty\}_{i\in\mathbb{N}}$ denotes the positive eigenvalues of $\mathcal{R}$ with associated eigenfunctions $\{\psi_i^\infty\}_{i\in\mathbb{N}}$. In each case the eigenvalues are considered according to their multiplicity. Let us note that

$$\int_0^T \|y(t)\|_X^2\,dt = \sum_{i=1}^\infty \lambda_i^\infty. \tag{3.9}$$

In fact,

$$\mathcal{R}\psi_i^\infty = \int_0^T \langle\psi_i^\infty, y(t)\rangle_X y(t)\,dt \quad \text{for every } i\in\mathbb{N}.$$

Taking the inner product with ψ_i^∞ and summing over i we arrive at

$$\sum_{i=1}^{\infty} \int_0^T \left| \langle \psi_i^\infty, y(t) \rangle_X \right|^2 dt = \sum_{i=1}^{\infty} \langle \mathcal{R}\psi_i^\infty, \psi_i^\infty \rangle_X = \sum_{i=1}^{\infty} \lambda_i^\infty.$$

Expanding $y(t) \in X$ in terms of $\{\psi_i^\infty\}_{i\in\mathbb{N}}$ we have

$$y(t) = \sum_{i=1}^{\infty} \langle \psi_i^\infty, y(t) \rangle_X \psi_i^\infty$$

and hence

$$\int_0^T \|y(t)\|_X^2 \, dt = \sum_{i=1}^{\infty} \int_0^T \left| \langle \psi_i^\infty, y(t) \rangle_X \right|^2 dt = \sum_{i=1}^{\infty} \lambda_i^\infty,$$

which is (3.9). From Proposition 3.3 and (3.4) we obtain

$$\sum_{j=0}^{n} \alpha_j \, \|y(t_j)\|_X^2 = \sum_{i=1}^{\infty} \lambda_i^n \quad \text{for every } n \in \mathbb{N}. \tag{3.10}$$

For convenience we do not indicate the dependence of α_j on n. Note that for $y \in C([0,T], X)$

$$\sum_{j=0}^{n} \alpha_j \, \|y(t_j)\|_X^2 \to \int_0^T \|y(t)\|_X^2 \, dt \quad \text{as } \Delta t \to 0.$$

Combining this fact with (3.9) and (3.10) we find

$$\sum_{i=1}^{\infty} \lambda_i^n \to \sum_{i=1}^{\infty} \lambda_i^\infty \quad \text{as } \Delta t \to 0. \tag{3.11}$$

Now choose and fix

$$\ell \quad \text{such that} \quad \lambda_\ell^\infty \neq \lambda_{\ell+1}^\infty. \tag{3.12}$$

Then by spectral analysis of compact operators [13, pp. 212–214] and (3.7) it follows that

$$\lambda_i^n \to \lambda_i^\infty \quad \text{for } 1 \le i \le \ell \text{ as } \Delta t \to 0. \tag{3.13}$$

Combining (3.11) and (3.13) there exists $\overline{\Delta t} > 0$ such that

$$\sum_{i=\ell+1}^{\infty} \lambda_i^n \le 2 \sum_{i=\ell+1}^{\infty} \lambda_i^\infty \quad \text{for all } \Delta t \le \overline{\Delta t} \tag{3.14}$$

if $\sum_{i=\ell+1}^{\infty} \lambda_i^\infty \neq 0$. Moreover, for ℓ as above, $\overline{\Delta t}$ can also be chosen such that

$$\sum_{i=\ell+1}^{d(n)} \left| \langle \psi_i^n, y_0 \rangle_X \right|^2 \le 2 \sum_{i=\ell+1}^{\infty} \left| \langle \psi_i^\infty, y_0 \rangle_X \right|^2 \quad \text{for all } \Delta \le \overline{\Delta t}, \tag{3.15}$$

provided that $\sum_{i=\ell+1}^{\infty} |\langle y_0, \psi_i^\infty \rangle_X|^2 \neq 0$. To verify (3.15) let us first note that $y_0 = y(0) \in \overline{\text{range } \mathcal{R}} = (\ker \mathcal{R})^\perp$. In fact, if $v \in \ker \mathcal{R}$, then $t \mapsto \langle v, y(t) \rangle_X$ is the zero function in $L^2(0,T;X)$. Since by assumption $y \in C([0,T];X)$ it follows that $\langle v, y(0) \rangle_X = 0$. But $v \in \ker \mathcal{R}$ was chosen arbitrarily and hence $y_0 \in (\ker \mathcal{R})^\perp$. As a consequence we have

$$\|y_0\|_X^2 = \sum_{i=1}^{\infty} \big|\langle y_0, \psi_i^\infty \rangle_X\big|^2. \tag{3.16}$$

Since $t_0 = 0$ holds, we have $y_0 \in \mathcal{V}^{(n)}$ for every n and

$$\|y_0\|_X^2 = \sum_{i=1}^{d(n)} \big|\langle y_0, \psi_i^n \rangle_X\big|^2. \tag{3.17}$$

Therefore, for $\ell < d(n)$ by (3.16) and (3.17)

$$\begin{aligned}\sum_{i=\ell+1}^{d(n)} \big|\langle y_0, \psi_i^n \rangle_X\big|^2 &= \sum_{i=1}^{d(n)} \big|\langle y_0, \psi_i^n \rangle_X\big|^2 - \sum_{i=1}^{\ell} \big|\langle y_0, \psi_i^n \rangle_X\big|^2 + \sum_{i=1}^{\ell} \big|\langle y_0, \psi_i^\infty \rangle_X\big|^2 \\ &\quad + \sum_{i=\ell+1}^{\infty} \big|\langle y_0, \psi_i^\infty \rangle_X\big|^2 - \sum_{i=1}^{\infty} \big|\langle y_0, \psi_i^\infty \rangle_X\big|^2 \\ &= \sum_{i=1}^{\ell} \Big(\big|\langle y_0, \psi_i^\infty \rangle_X\big|^2 - \big|\langle y_0, \psi_i^n \rangle_X\big|^2\Big) + \sum_{i=\ell+1}^{\infty} \big|\langle y_0, \psi_i^\infty \rangle_X\big|^2.\end{aligned}$$

As a consequence of (3.7) and (3.12) we have $\lim_{\Delta t \to 0} \psi_i^n = \psi_i^\infty$ for $i = 1, \dots, \ell$ and hence (3.15) follows.

4. Backward Euler Galerkin method. This section is devoted to error estimates for the Galerkin POD method applied to (2.6) combined with the backward Euler method for the time integration. Throughout, (2.3)–(2.5) are assumed to hold.

4.1. Case $X = V$. Let us choose $X = V$ in the context of section 3. To study the backward Euler Galerkin POD method for (2.6), we introduce the Ritz projection $P^\ell : V \to V^\ell$, $1 \leq \ell \leq d$, by

$$a(P^\ell \varphi, \psi) = a(\varphi, \psi) \quad \text{for all } \psi \in V^\ell, \tag{4.1}$$

where $\varphi \in V$. Since the Hilbert space V is endowed with the inner product (2.1), P^ℓ is the orthogonal projection of V on V^ℓ. In particular, this implies that P^ℓ has norm one.

LEMMA 4.1. *For every $\ell \in \{1, \dots, d\}$ the projection operators P^ℓ satisfy*

$$\sum_{j=0}^{n} \alpha_j \, \|y(t_j) - P^\ell y(t_j)\|_V^2 \leq \sum_{i=\ell+1}^{d} \lambda_i, \tag{4.2}$$

where λ_i denote the eigenvalues introduced in (3.4).

Proof. For arbitrary $\varphi \in V$ we deduce from (2.1) and (4.1) that

$$\|\varphi - P^\ell \varphi\|_V^2 = a(\varphi - P^\ell \varphi, \varphi - P^\ell \varphi) = a(\varphi - P^\ell \varphi, \varphi - \psi) \leq \|\varphi - P^\ell \varphi\|_V \|\varphi - \psi\|_V$$

for all $\psi \in V^\ell$ so that

$$\|\varphi - P^\ell \varphi\|_V \le \|\varphi - \psi\|_V \quad \text{for all } \psi \in V^\ell. \tag{4.3}$$

Using (4.3) and (3.5) we obtain

$$\sum_{j=0}^{n} \alpha_j \ \|y(t_j) - P^\ell y(t_j)\|_V^2 \le \sum_{j=0}^{n} \alpha_j \ \left\| y(t_j) - \sum_{i=1}^{\ell} a(y(t_j), \psi_i)\psi_i \right\|_V^2 = \sum_{i=\ell+1}^{d} \lambda_i,$$

which is estimate (4.2). □

The Galerkin POD method for (2.6) is described next. For $m \in \mathbb{N}$ we introduce the time grid

$$0 = \tau_0 < \tau_1 < \cdots < \tau_m = T, \qquad \delta\tau_j = \tau_j - \tau_{j-1} \text{ for } j = 1, \dots, m$$

and set

$$\delta\tau = \min\{\delta\tau_j : 1 \le j \le m\} \quad \text{and} \quad \Delta\tau = \max\{\delta\tau_j : 1 \le j \le m\}.$$

Throughout we assume that $\Delta\tau/\delta\tau$ is bounded uniformly with respect to m. To relate the two time discretizations $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$ we set for every τ_k, $0 \le k \le m$, an associated index $\bar{k} = \operatorname{argmin}\{|\tau_k - t_j| : 0 \le j \le n\}$ and define $\sigma_n \in \{1, \dots, n\}$ as the maximum of the occurrence of the same value $t_{\bar{k}}$ as k ranges over $0 \le k \le m$.

The problem consists of finding a sequence $\{Y_k\}_{k=0}^m$ in V^ℓ satisfying

$$\langle Y_0, \psi \rangle_H = \langle y_0, \psi \rangle_H \quad \text{for all } \psi \in V^\ell \tag{4.4a}$$

and

$$\langle \overline{\partial}_\tau Y_k, \psi \rangle_H + a(Y_k, \psi) + \langle B(Y_k) + RY_k, \psi \rangle_{V',V} = \langle f(\tau_k), \psi \rangle_H \tag{4.4b}$$

for all $\psi \in V^\ell$ and $k = 1, \dots, m$, where we have set

$$\overline{\partial}_\tau Y_k = \frac{Y_k - Y_{k-1}}{\delta\tau_k}.$$

In the following theorem, existence and a priori estimates for the solution $\{Y_k\}_{k=0}^m$ are established. For the proof we refer to Appendix A.

THEOREM 4.2. *For every $k = 1, \dots, m$ there exists at least one solution Y_k of* (4.4b). *If $\Delta\tau$ is sufficiently small, the sequence $\{Y_k\}_{k=1}^m$ is uniquely determined. Moreover, the following estimates are satisfied:*

$$\|Y_k\|_H^2 \le (1 + \gamma\delta\tau)e^{-\gamma k\delta\tau} \ \|y_0\|_H^2 + \frac{1 - e^{-\gamma k\Delta\tau}}{\gamma} \ \|f\|_{C([0,T];H)}^2 \tag{4.5a}$$

for $k = 0, \dots, m$, where $\gamma = \eta/c_V^2$, c_V, η were introduced in (2.2) *and* (2.4), *respectively, and*

$$\sum_{k=1}^{m} \|Y_k - Y_{k-1}\|_H^2 + \eta \sum_{k=1}^{m} \delta\tau_k \ \|Y_k\|_V^2 \le \|y_0\|_H^2 + \frac{T}{\gamma} \ \|f\|_{C([0,T];H)}^2. \tag{4.5b}$$

Our next goal is to derive an error estimate for the expression

$$\sum_{k=0}^{m} \beta_k \ \|Y_k - y(\tau_k)\|_H^2,$$

where $y(\tau_k)$ is the solution of (2.6) at the time instances $t = \tau_k$, $k = 1, \dots, m$, and the positive weights β_j are given by

$$\beta_0 = \frac{\delta\tau_1}{2}, \quad \beta_j = \frac{\delta\tau_j + \delta\tau_{j+1}}{2} \text{ for } j = 1, \dots, m-1, \quad \text{and} \quad \beta_m = \frac{\delta\tau_m}{2}. \tag{4.6}$$

We make use of the following assumptions:

(A1) $y_t \in L^2(0,T;V)$ and $y_{tt} \in L^2(0,T;H)$.

(A2) There exists a normed linear space W continuously embedded in V and a constant $c_a > 0$ such that $y \in C([0,T];W)$ and

$$a(\varphi,\psi) \le c_a \ \|\varphi\|_H \|\psi\|_W \quad \text{for all } \varphi \in V \text{ and } \psi \in W. \tag{4.7}$$

(A3) $y \in W^{2,2}(0,T;V)$.

EXAMPLE 4.3. *For $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, with Ω a bounded domain in $\mathbb{R}^l$ and*

$$a(\varphi,\psi) = \int_\Omega \nabla\varphi \cdot \nabla\psi \, dx \quad \textit{for all } \varphi,\psi \in H_0^1(\Omega),$$

choosing $W = H^2(\Omega) \cap H_0^1(\Omega)$ implies $a(\varphi,\psi) \le \|\varphi\|_W \|\psi\|_H$ for all $\varphi \in W$, $\psi \in V$, and (4.7) holds with $c_a = 1$.

Remark 4.4. Note that (A2) implies the existence of a constant $c_P > 0$ depending on ℓ and λ_ℓ such that

$$\|P^\ell\|_{\mathcal{L}(H)} \le c_P \quad \text{for all } 1 \le \ell \le d. \tag{4.8}$$

In fact, using (2.2) and (4.7) we find

$$\|P^\ell\varphi\|_H \le \sum_{i=1}^{\ell} |a(\psi_i,\varphi)| \ \|\psi_i\|_H \le c_a c_V \ \|\varphi\|_H \sum_{i=1}^{\ell} \|\psi_i\|_W.$$

Now we estimate the term $\|\psi_i\|_W$ for $i = 1, \dots, \ell$. Using $\sum_{j=0}^{n} \alpha_j = T$ and (3.4) we have

$$\begin{aligned} \|\psi_i\|_W = \frac{1}{\lambda_i} \ \|\mathcal{R}_n\psi_i\|_W &\le \frac{1}{\lambda_\ell} \sum_{j=0}^{n} \alpha_j \big|a(\psi_i, y(t_j))\big| \|y(t_j)\|_W \\ &\le \frac{1}{\lambda_\ell} \ \|y\|_{C([0,T];W)} \sum_{j=0}^{n} \alpha_j \|y(t_j)\|_V \le \frac{T}{\lambda_\ell} \ \|y\|_{C([0,T];W)} \|y\|_{C([0,T];V)}. \end{aligned}$$

This bound implies

$$\|P^\ell\varphi\|_H \le \frac{c\ell}{\lambda_\ell} \ \|\varphi\|_H \tag{4.9}$$

with $c = c_a c_V T \ \|y\|_{C([0,T];W)} \|y\|_{C([0,T];V)}$.

Throughout we shall use the decomposition

$$Y_k - y(\tau_k) = Y_k - P^\ell y(\tau_k) + P^\ell y(\tau_k) - y(\tau_k) = \vartheta_k + \varrho_k, \tag{4.10}$$

where $\vartheta_k = Y_k - P^\ell y(\tau_k)$ and $\varrho_k = P^\ell y(\tau_k) - y(\tau_k)$. The following lemma establishes an error estimate for ϑ_k. For the proof we refer to Appendix B.

LEMMA 4.5. *Assume that $\Delta\tau$ is sufficiently small and that* (A1), (A2) *hold. Then there exist constants $C_1, C_2 > 0$ independent of the grids $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$ such that*

$$\begin{aligned} \|\vartheta_k\|_H^2 \;\le\; C_1 e^{C_2 k\delta\tau}\Bigg(& \|y_0 - P^\ell y_0\|_H^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\sum_{i=\ell+1}^{d}\lambda_i \\ & + \sigma_n\Delta\tau\big(1 + c_P^2\big)(\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,t_{\bar{k}+1};H)} \\ & + \sigma_n\Delta\tau\Delta t\ \|y_t\|^2_{L^2(0,t_{\bar{k}+1};V)}\Bigg) \end{aligned} \tag{4.11}$$

for each $1 \le k \le m$.

Remark 4.6. Since $y_0 \in \mathcal{V}$ we infer from (2.2) that

$$\|y_0 - P^\ell y_0\|_H^2 \le c_V^2 \sum_{i=\ell+1}^{d} \big|\langle \psi_i, y_0\rangle_V\big|^2.$$

We turn to the term $\|\varrho_k\|_H^2$. Observe that

$$\begin{aligned} \|\varrho_k\|_H^2 &= \|P^\ell y(\tau_k) - y(\tau_k)\|_H^2 \\ &\le 3\left(\|P^\ell y(\tau_k) - P^\ell y(t_{\bar{k}})\|_H^2 + \|P^\ell y(t_{\bar{k}}) - y(t_{\bar{k}})\|_H^2 + \|y(t_{\bar{k}}) - y(\tau_k)\|_H^2\right) \\ &\le 3\big(1 + c_P^2\big)\ \|y(t_{\bar{k}}) - y(\tau_k)\|_H^2 + 3\ \|P^\ell y(t_{\bar{k}}) - y(t_{\bar{k}})\|_H^2 \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} \|y(t_{\bar{k}}) - y(\tau_k)\|_H^2 &\le \left(\int_{t_{\bar{k}-1}}^{t_{\bar{k}+1}} \|y_t(s)\|_H\, ds\right)^2 \\ &\le (\delta t_{\bar{k}} + \delta t_{\bar{k}+1})\ \|y_t\|^2_{L^2(t_{\bar{k}-1},t_{\bar{k}+1};H)}, \end{aligned} \tag{4.13}$$

where we set $t_{m+1} = T$ whenever $\bar{k} = m$. Using (4.13) and $\beta_k \le \Delta\tau$ we obtain

$$\sum_{k=0}^{m} \beta_k\ \|y(t_{\bar{k}}) - y(\tau_k)\|_H^2 \le 2\sigma_n\Delta\tau\Delta t\ \|y_t\|^2_{L^2(0,T;H)}.$$

From (2.2), $\beta_k \le \Delta\tau$, $\alpha_j \ge \delta t/2$, and Lemma 4.1 we infer that

$$\begin{aligned} \sum_{k=0}^{m} \beta_k\ \|P^\ell y(t_{\bar{k}}) - y(t_{\bar{k}})\|_H^2 &\le \frac{2c_V^2\sigma_n\Delta\tau}{\delta t}\sum_{j=0}^{n}\alpha_j\ \|P^\ell y(t_j) - y(t_j)\|_V^2 \\ &\le \frac{2c_V^2\sigma_n\Delta\tau}{\delta t}\sum_{i=\ell+1}^{d}\lambda_i. \end{aligned} \tag{4.14}$$

Combining the last two bounds and (4.12) it follows that

$$\sum_{k=0}^{m} \beta_k \, \|\varrho_k\|_H^2 \le 6\sigma_n(1+c_P^2)\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,T;H)} + \frac{6c_V^2\sigma_n\Delta\tau}{\delta t} \sum_{i=\ell+1}^{d} \lambda_i. \tag{4.15}$$

Note that $\sum_{k=0}^{m} \beta_k = T$ holds. By Lemma 4.5 we have

$$\begin{aligned} \sum_{k=0}^{m} \beta_k \, \|\vartheta_k\|_H^2 \quad &\le \quad C_3 \left(\|\vartheta_0\|_H^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big) \sum_{i=\ell+1}^{d} \lambda_i \right) \\ &\quad + C_3\Delta\tau(1+c_P^2)(\Delta\tau + \sigma_n\Delta t)\|y_{tt}\|^2_{L^2(0,T;H)} \\ &\quad + C_3\sigma_n\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,T;V)}, \end{aligned} \tag{4.16}$$

where $C_3 = C_1 T e^{C_2 T}$. From (4.10), (4.15), (4.16), and Remark 4.4 we obtain the first part of the following theorem.

THEOREM 4.7.

(a) *Assume that* (A1), (A2) *hold and that* $\Delta\tau$ *is sufficiently small. Then there exists a constant* C *depending on* T*, but independent of the grids* $\{t_j\}_{j=0}^n$ *and* $\{\tau_j\}_{j=0}^m$*, such that*

(4.17)
$$\begin{aligned} &\sum_{k=0}^{m} \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \\ &\le C \sum_{i=\ell+1}^{d} \left(\big|\langle \psi_i, y_0\rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i \right) + C\sigma_n\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,T;V)} \\ &\quad + C\sigma_n(1+c_P^2)\Delta\tau \left(\Delta t\|y_t\|^2_{L^2(0,T;H)} + (\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,T;H)} \right). \end{aligned}$$

(b) *If* (A3) *is satisfied and* $\Delta\tau$ *sufficiently small, then there exists a constant* C *depending on* T*, but independent of the grids* $\{t_j\}_{j=0}^n$ *and* $\{\tau_j\}_{j=0}^m$*, such that*

(4.18)
$$\begin{aligned} &\sum_{k=0}^{m} \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \le C\sigma_n\Delta\tau(\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,T;V)} \\ &\quad + C \left(\sum_{i=\ell+1}^{d} \left(\big|\langle \psi_i, y_0\rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i \right) + \sigma_n\Delta\tau\Delta t\|y_t\|^2_{L^2(0,T;V)} \right). \end{aligned}$$

Proof. The proof of part (b) is obtained from that for (a) by utilizing (2.2) and $\|P^\ell\|_{\mathcal{L}(V)} = 1$ and by simple modifications of the estimates for the two terms $\sum_{k=0}^{m} \beta_k \|\vartheta_k\|_H^2$ and $\sum_{j=1}^{k} \delta\tau_j \|z_j\|_H^2$ in (B.16) of Appendix B. ∎

Compared to standard finite difference, finite element, or spectral element approximation results in the basic Galerkin POD backward Euler convergence, the result of Theorem 4.7 has an unusual format. This is due, in part, to the fact that one cannot rely on function space rate of convergence results, which are typically the basis for approximation theory of partial differential equations. The terms in the second line of (4.17) depend (through ψ_i, λ_i, d) on the way in which the snapshots are taken, on the number ℓ of basis elements, and on the relative locations of the snapshots and the

time discretization (through σ_n). In the remainder of this section we shall analyze these terms and show how they can be simplified if further assumptions are admitted.

Remark 4.8. In (4.17) and (4.18) the eigenvalues and eigenfunctions depend on n, i.e., $\lambda_i = \lambda_i^n$ and $\psi_i = \psi_i^n$. As proved in section 3, if ℓ satisfies (3.12) and $\sum_{i=\ell+1}^{\infty} \lambda_i^\infty \neq 0$ or $\sum_{i=\ell+1}^{\infty} |\langle \psi_i, y_0 \rangle_V|^2 \neq 0$, then by (3.14), (3.15) we have

$$\sum_{i=\ell+1}^{d} \Big(\big|\langle \psi_i, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i \Big) \leq 2 \sum_{i=\ell+1}^{\infty} \Big(\big|\langle \psi_i^\infty, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i^\infty \Big) \quad \text{for all } \Delta t \leq \overline{\Delta t},$$

and the dependence of the estimates of eigenvalues and eigenfunctions on n in (4.17) and (4.18) is thus eliminated.

Let us next derive some corollaries to the proof of Theorem 4.7. At first we consider the case in which the two grids coincide so that $n = m$ and $\tau_j = t_j$ for $j = 0, \ldots, m$.

COROLLARY 4.9. *Suppose that the assumptions of Theorem* 4.7(a) *hold. If the two time discretizations coincide, then there exists a constant $C > 0$ depending on T, but independent of the grid $\{\tau_j\}_{j=0}^m$, such that*

$$\begin{aligned} \sum_{k=0}^{m} \beta_k \, \|Y_k - y(\tau_k)\|_H^2 &\leq C(1 + c_P^2)\Delta\tau^2 \, \|y_{tt}\|_{L^2(0,T;H)}^2 \\ &\quad + C\left(\sum_{i=\ell+1}^{d} \Big(\big|\langle \psi_i, y_0 \rangle_V\big|^2 + \Big(\frac{1}{\delta\tau^2} + 1\Big)\lambda_i \Big) + \Delta\tau^2 \, \|y_t\|_{L^2(0,T;V)} \right). \end{aligned} \tag{4.19}$$

Proof. We proceed as in the proof of Theorem 4.2. Since the two time discretizations coincide, we obtain $n = m$, $\sigma_n = 1$, $\delta t = \delta\tau$, and $\alpha_j = \beta_j$ for $j = 0, \ldots, n$. In place of the estimate (4.15) we now have

$$\sum_{k=0}^{m} \beta_k \, \|\varrho_k\|_H^2 \leq c_V^2 \sum_{i=\ell+1}^{d} \lambda_i,$$

which gives the claim. ∎

Remark 4.10. Again, as in Theorem 4.7(b) compared to (a), the factor $1 + c_P^2$ can be avoided in (4.19) if in place of (A1), (A2) we assume (A3) and replace the term $\|y_{tt}\|_{L^2(0,T;H)}$ with $\|y_{tt}\|_{L^2(0,T;V)}$.

Let us briefly reflect on the behavior of the right-hand side of (4.17) and (4.18). First we note that if the number of POD elements for the Galerkin scheme coincides with the dimension of $\mathcal{V}$, then the first additive term on the right-hand side disappears. Second, if the number of snapshots is refined so that $\Delta t \to 0$, then the factor multiplying $\sum_{i=\ell+1}^{d} \lambda_i$ blows up. As noted above, the term $\sum_{i=\ell+1}^{d} \lambda_i$ itself changes as the snapshots are refined. While computations for many concrete situations show that $\sum_{i=\ell+1}^{d} \lambda_i$ is small compared to $\Delta\tau$, the question nevertheless arises of whether the term $1/(\delta\tau\delta t)$ can be avoided in the estimates. For this purpose we choose

$$\mathcal{V} = \text{span}\,\{y(t_0), \ldots, y(t_n), \overline{\partial}_t y(t_1), \ldots, \overline{\partial}_t y(t_n)\}, \tag{4.20}$$

where

$$\overline{\partial}_t y(t_j) = \frac{y(t_j) - y(t_{j-1})}{\delta t_j} \quad \text{for } j = 1, \dots, n.$$

Equation (3.5) must be replaced by

$$\sum_{j=0}^{n} \alpha_j \left\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \hat{\psi}_i \rangle_V \hat{\psi}_i \right\|_V^2 + \sum_{j=1}^{n} \alpha_j \left\| \overline{\partial}_t y(t_j) - \sum_{i=1}^{\ell} \langle \overline{\partial}_t y(t_j), \hat{\psi}_i \rangle_V \hat{\psi}_i \right\|_V^2 = \sum_{i=\ell+1}^{d} \hat{\lambda}_i,$$

where $\{\hat{\lambda}_i\}_{i\in\mathbb{N}}$, $\{\hat{\psi}_i\}_{i\in\mathbb{N}}$ are the eigenvalues and eigenfunctions of $\hat{\mathcal{R}}_n \in \mathcal{L}(V)$ given by

$$\hat{\mathcal{R}}_n z = \sum_{j=0}^{n} \alpha_j \left(\langle z, y(t_j) \rangle_V y(t_j) + \langle z, \overline{\partial}_t y(t_j) \rangle_V \overline{\partial}_t y(t_j) \right)$$

and satisfying

$$\hat{\mathcal{R}}_n \hat{\psi}_i = \hat{\lambda}_i \hat{\psi}_i, \quad \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{d(n)} > 0, \quad \text{and } \lambda_i = 0 \text{ for } i > d(n).$$

As a consequence, estimate (B.16) in Appendix B can be replaced by

$$\sum_{j=1}^{k} \delta\tau_j \ \|z_j\|_H^2 \leq 14\sigma_n(1 + c_P^2)(\Delta\tau^2 + \Delta\tau\Delta t)\|y_{tt}\|^2_{L^2(0,t_{\bar{k}+1};H)} + \frac{14\sigma_n c_V^2 \Delta\tau}{\delta t} \sum_{i=\ell+1}^{d} \hat{\lambda}_i$$

in the case of (A1), (A2) holding, and by

$$\sum_{j=1}^{k} \delta\tau_j \ \|z_j\|_H^2 \leq 28\sigma_n(\Delta\tau^2 + \Delta\tau\Delta t)\|y_{tt}\|^2_{L^2(0,t_{\bar{k}+1};V)} + \frac{28\sigma_n c_V^2 \Delta\tau}{\delta t} \sum_{i=\ell+1}^{d} \hat{\lambda}_i$$

in the case of (A3). We obtain the following corollary.

COROLLARY 4.11. *If in addition to the assumptions of Theorem* 4.7(a) *the snapshots set is taken as in* (4.20)*, then*

$$\begin{aligned} &\sum_{k=0}^{m} \beta_k \ \|Y_k - y(\tau_k)\|_H^2 \\ &\leq C \sum_{i=\ell+1}^{d} \left(\left| \langle \hat{\psi}_i, y_0 \rangle_V \right|^2 + \frac{\sigma_n \Delta\tau}{\delta t} \hat{\lambda}_i \right) + C\sigma_n \Delta\tau\Delta t \ \|y_t\|^2_{L^2(0,T;V)} \\ &\quad + C(1 + c_P^2)\Delta\tau \big((\Delta\tau + \sigma_n \Delta t)\|y_{tt}\|^2_{L^2(0,T;H)} + \sigma_n \Delta t \|y_t\|^2_{L^2(0,T;H)} \big), \end{aligned} \tag{4.21}$$

where C has the same properties as in Theorem 4.7.

Remark 4.12. In [12] a laser surface hardening problem was considered. The numerical experiments show that the inclusion of the difference quotients into the snapshot set leads to better results.

In estimate (4.21) the term $1 + c_P^2$ can be avoided if (A3) in place of (A1), (A2) holds and $\|y_{tt}\|_{L^2(0,T;H)}$ is replaced by $\|y_{tt}\|_{L^2(0,T;V)}$. Note that the terms $\{\hat{\lambda}_i\}_{i\in\mathbb{N}}$,

$\{\hat{\psi}_i\}_{i\in\mathbb{N}}$, and σ_n depend on the time discretization of $[0,T]$ for the snapshots as well as the numerical integration. We address this dependence next.

If we suppose that

$$\Delta t = O(\delta\tau) \quad \text{and} \quad \Delta\tau = O(\delta t), \tag{4.22}$$

then there exists a constant $c_1 > 0$ independent of $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$ such that

$$\max\left(\sigma_n, \frac{\sigma_n \Delta\tau}{\delta t}\right) \le c_1. \tag{4.23}$$

To obtain an estimate that is independent of the spectral values of a specific snapshot set $\{y(t_j)\}_{j=0}^n$ we follow the analysis of section 3.2. We assume that $y \in W^{2,2}(0,T;V)$, so that in particular (A3) holds, and introduce the operator $\hat{\mathcal{R}} \in \mathcal{L}(V)$ corresponding to $\mathcal{R}$ by

$$\hat{\mathcal{R}} z = \int_0^T \langle z, y(t)\rangle_V y(t) + \langle z, y_t(t)\rangle_V y_t(t)\, dt \quad \text{for } z \in V.$$

Note that $\hat{\mathcal{R}} = \hat{\mathcal{Y}}\hat{\mathcal{Y}}^*$, where $\hat{\mathcal{Y}}^* : V \to W^{1,2}(0,T;\mathbb{R})$ is given by

$$(\hat{\mathcal{Y}}^* z)(t) = \langle z, y(t)\rangle_V.$$

Since $y \in W^{2,2}(0,T;V)$ it is simple to argue that $\hat{\mathcal{Y}}^*$ is compact and hence $\hat{\mathcal{R}}$ is compact. Let us denote the positive eigenvalues and corresponding eigenfunctions of $\hat{\mathcal{R}}$ by $\{\hat{\lambda}_i^\infty\}_{i\in\mathbb{N}}$ and $\{\hat{\psi}_i^\infty\}_{i\in\mathbb{N}}$. Since $t_0 = 0$, we proceed as in section 3.2, that $y_0 \in \text{range } \hat{\mathcal{R}}_n$ for all n and $y_0 \in \overline{\text{range } \hat{\mathcal{R}}}$. The assumption $y \in W^{2,2}(0,T;V)$ allows us to argue that the analogue of (3.7), i.e.,

$$\lim_{\Delta t \to 0} \|\hat{\mathcal{R}}_n - \hat{\mathcal{R}}\|_{\mathcal{L}(V)} = 0,$$

holds. Let us choose and fix ℓ such that

$$\hat{\lambda}_\ell^\infty \ne \hat{\lambda}_{\ell+1}^\infty. \tag{4.24}$$

We can now proceed precisely as in section 3.2 to assert that there exists $\overline{\Delta t} > 0$ such that

$$\sum_{i=\ell+1}^{d(n)} \hat{\lambda}_i^n \le 2 \sum_{i=\ell+1}^{\infty} \hat{\lambda}_i^\infty \quad \text{and} \quad \sum_{i=\ell+1}^{d(n)} \left|\langle y_0, \hat{\psi}_i^n\rangle_V\right|^2 \le 2 \sum_{i=\ell+1}^{\infty} \left|\langle y_0, \hat{\psi}_i^\infty\rangle_V\right|^2 \tag{4.25}$$

for all $\Delta t \le \overline{\Delta t}$, provided, of course, that the terms on the right-hand side of (4.25) are different from zero. We summarize the above discussion in the following corollary.

COROLLARY 4.13. *Assume that $y \in W^{2,2}(0,T;V)$ and let the snapshots be chosen as in* (4.20). *If* (4.22) *holds and ℓ satisfies* (4.24), *then there exists a constant $C > 0$, independent of ℓ and the grids $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$, and a $\overline{\Delta t} > 0$, depending on ℓ, such that*

$$\begin{aligned}\sum_{k=0}^{m} \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \le C \sum_{i=\ell+1}^{\infty} \left(\left|\langle y_0, \hat{\psi}_i^\infty\rangle_V\right|^2 + \hat{\lambda}_i^\infty\right) \\ + C\left(\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,T;V)} + \Delta\tau(\Delta\tau + \Delta t) \, \|y_{tt}\|^2_{L^2(0,T;V)}\right)\end{aligned} \tag{4.26}$$

for all $\Delta t \le \overline{\Delta t}$.

Remark 4.14. In (4.26) the first term on the right-hand side of the inequality reflects the spatial approximation error of the Galerkin POD scheme and the second reflects the approximation error due to the temporal backward Euler scheme. If the latter is replaced by the Crank–Nicolson method, then, assuming $\Delta\tau = \Delta t$ and appropriate regularity on y, it can be shown with the techniques of this section that an estimate analogous to (4.26) holds with the first additive term on the right-hand side unchanged and the second one of fourth order in $\Delta\tau$.

4.2. Case $X = H$. Here we consider the case in which the POD basis is constructed with respect to the H-norm. Differently from the situation where the POD basis was constructed in V, the right-hand side of the estimate will involve the stiffness matrix

$$S = ((S_{ij})) \in \mathbb{R}^{d\times d} \quad \text{with} \quad S_{ij} = a(\psi_j, \psi_i).$$

We shall require the following lemma.

LEMMA 4.15. *For every* $\ell \in \{1, \dots, d\}$ *the projection operator* $P^\ell : V \to V^\ell$ *satisfies*

$$\sum_{j=0}^{n} \alpha_j \ \|y(t_j) - P^\ell y(t_j)\|_V^2 \le \|S\|_2 \sum_{i=\ell+1}^{d} \lambda_i, \tag{4.27}$$

where λ_i *denote the eigenvalues introduced in* (3.4) *and* $\|\cdot\|_2$ *stands for the spectral norm for symmetric matrices.*

Proof. Using the fact that $\|\varphi\|_V^2 \le \|S\|_2 \ \|\varphi\|_H^2$ for all $\varphi \in \mathcal{V}$ (see [16, Lemma 2]), we can proceed as in the proof of Lemma 4.1 and, utilizing (3.5) with $X = H$, we obtain the desired result. ∎

THEOREM 4.16. *Suppose that* (A3) *holds and that* $\Delta\tau$ *is sufficiently small. Then there exists a constant* $C > 0$ *depending on* T*, but independent of the grids* $\{t_j\}_{j=0}^n$ *and* $\{\tau_j\}_{j=0}^m$*, such that*

$$\begin{aligned}\sum_{k=0}^{m} \beta_k \ \|Y_k - y(\tau_k)\|_H^2 \le C \sum_{i=\ell+1}^{d} \|S\|_2 \Big(\big|\langle \psi_i, y_0\rangle_H\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i\Big) \\ + C\sigma_n\Delta\tau\left((\Delta\tau + \Delta t)\|y_{tt}\|_{L^2(0,T;V)}^2 + \Delta t \ \|y_t\|_{L^2(0,T;V)}^2\right).\end{aligned} \tag{4.28}$$

Proof. We proceed as in the proofs of Lemma 4.5 and Theorem 4.7 and indicate only the necessary changes. Estimate (B.15) requires no change. For (B.16) we utilize $\|P^\ell\varphi\| \le c_V\|\varphi\|_V$ for $\varphi \in V$ and obtain, by applying Lemma 4.15,

$$\begin{aligned}\sum_{j=1}^{k} \delta\tau_j \ \|z_j\|_H^2 \quad &\le \quad 14\sigma_n(1 + c_V^2)(\Delta\tau^2 + \Delta\tau\Delta t)\|y_{tt}\|_{L^2(0,t_{k+1};V)}^2 \\ &\quad + \frac{56\sigma_n c_V^2\|S\|_2}{\delta t\delta\tau} \sum_{i=\ell+1}^{d} \lambda_i.\end{aligned} \tag{4.29}$$

The analogue of (B.17) is again obtained by Lemma 4.15. Summarizing the ϑ_k-terms we have

$$\begin{aligned}\sum_{j=1}^{k} \delta\tau_j \ \|\vartheta_k\|_H^2 \le C\left(\|\vartheta_0\|_H^2 + \frac{\sigma_n\|S\|_2}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\sum_{i=\ell+1}^{d}\lambda_i\right) \\ + C\sigma_n(1 + c_V^2)\Delta\tau\big((\Delta\tau + \Delta t)\|y_{tt}\|_{L^2(0,T;V)}^2 + \Delta t \ \|y_t\|_{L^2(0,T;V)}^2\big).\end{aligned} \tag{4.30}$$

Turning to the ϱ_k-terms we find, following the estimates after (4.12),

$$\begin{aligned}
&\sum_{k=0}^{m} \beta_k \ \|\varrho_k\|_H^2 \\
&\quad \le 6\sigma_n(1+c_V^2)\Delta\tau\Delta t \ \|y_t\|^2_{L^2(0,T;V)} + 3\sum_{k=0}^{m} \beta_k \|P^\ell y(t_{\bar{k}}) - y(t_{\bar{k}})\|_H^2 \\
&\quad \le 6\sigma_n(1+c_V^2)\Delta\tau\Delta t \ \|y_t\|^2_{L^2(0,T;V)} + \frac{6c_V^2\sigma_n\Delta\tau}{\delta t} \sum_{j=0}^{n} \alpha_j \ \|P^\ell y(t_j) - y(t_j)\|_V^2.
\end{aligned}$$

Thus by Lemma 4.15

$$\sum_{k=0}^{m} \beta_k \ \|\varrho_k\|_H^2 \le 6\sigma_n(1+c_V^2)\Delta\tau\Delta t \ \|y_t\|^2_{L^2(0,T;V)} + \frac{6c_V^2\sigma_n\Delta\tau\|S\|_2}{\delta t} \sum_{i=\ell+1}^{d} \lambda_i. \tag{4.31}$$

Finally $\vartheta_0 = y_0 - P^\ell y_0$ can be estimated as follows:

$$\begin{aligned}
\|y_0 - P^\ell y_0\|_H \ &\le \ c_V \ \|y_0 - P^\ell y_0\|_V \le c_V \ \left\| y_0 - \sum_{i=1}^{\ell} \langle y_0, \psi_i\rangle_H \psi_i \right\|_V \\
&= \ c_V \left(\|S\|_2 \sum_{i=\ell+1}^{d} \left|\langle y_0, \psi\rangle_H\right|^2 \right)^{1/2}.
\end{aligned}$$

Combining the last estimate with (4.30)–(4.31) we obtain (4.28). ∎

Remark 4.17. Let us briefly discuss the asymptotic properties of the expression on the right-hand side of (4.28), which are restricted due to the appearance of $\delta t\delta\tau$ in the denominator and the terms σ_n and $\|S\|_2$. As in section 4.1 the factor $1/\delta\tau$ can be eliminated by adding the set $\{\overline{\partial} y(t_j)\}_{j=1}^n$ to the set of snapshots. Assuming that $\Delta t = O(\delta\tau)$ and $\Delta\tau = O(\delta t)$ implies (4.23), and consequently, σ_n and $\sigma_n\Delta\tau/\delta t$ are uniformly bounded with respect to refinement of the t- and τ-grids. The factor $\|S\|_2$, which tends to infinity as $m \to \infty$, appears to be unavoidable in case the POD basis is computed in H.

Appendix A. Proof of Theorem 4.2.

A.1. Existence. Existence of a solution $\{Y_k\}_{k=1}^m$ can be proved by using the Schauder fixed point theorem; see [11, p. 222], for instance. For that purpose we define $z = \mathcal{T}_k w$ via the mappings $\mathcal{T}_k : V^\ell \to V^\ell$, $k = 1, \dots, m$, as follows: $z \in V^\ell$ is the solution to

$$\langle z, \psi\rangle_H + \delta\tau_k\big(a(z,\psi) + \langle B(w,z) + Rz, \psi\rangle_{V',V}\big) = \langle \delta\tau_k f(\tau_k) + Y_{k-1}, \psi\rangle_H \tag{A.1}$$

for all $\psi \in V^\ell$. The bilinear form

$$\langle \cdot\,, \cdot\rangle_H + \delta\tau_k\big(a(\cdot\,,\cdot) + \langle B(w,\cdot) + R(\cdot), \cdot\rangle_{V',V}\big)$$

is continuous and coercive in $V^\ell \times V^\ell$ by (2.3)–(2.5). The existence and uniqueness of a solution to (A.1) can thus be shown by the Lax–Milgram theorem. The fixed points of $\mathcal{T}_k$ are the solutions of (4.4b). Taking $\psi = z$ in (A.1) above and using (2.2) and (2.4) we derive

$$\|z\|_V \le \frac{c_V}{\eta} \ \Big(\|f(\tau_k)\|_H + \frac{1}{\delta\tau_k} \ \|Y_{k-1}\|_H \Big). \tag{A.2}$$

Let us introduce the set

$$M_k = \Big\{ w \in V^\ell : \|w\|_V \le \frac{c_V}{\eta} \Big(\|f(\tau_k)\|_H + \frac{1}{\delta\tau_k} \|Y_{k-1}\|_H \Big) \Big\} \subset V^\ell.$$

From (A.2) we infer that $\mathcal{T}_k$ maps M_k into itself. Since M_k is a closed ball in V^ℓ, the set M_k is bounded, closed, and convex. Since the image of $\mathcal{T}_k$ is finite dimensional, $\mathcal{T}_k$ is compact. Thus, the existence of a fixed point Y_k follows from the Schauder fixed point theorem.

A.2. Uniqueness. To prove the uniqueness we assume that the two sequences $\{Y_k^1\}_{k=0}^m$, $\{Y_k^2\}_{k=0}^m$ in V^ℓ are solutions of (4.4b). Then $\delta Y_k = Y_k^1 - Y_k^2 \in V^\ell$ solves

$$\langle \delta Y_k, \psi \rangle_H + \delta\tau_k \big(a(\delta Y_k, \psi) + \langle R\delta Y_k, \psi \rangle_{V',V} \big) = \delta\tau_k \, \langle B(Y_k^2) - B(Y_k^1), \psi \rangle_{V',V}$$

for all $\psi \in V^\ell$. Setting $\psi = \delta Y_k$ and using (2.4), (2.5), and Young's inequality we obtain

$$\begin{aligned}
\|\delta Y_k\|_H^2 + \eta\delta\tau_k \, \|\delta Y_k\|_V^2 &\le \delta\tau_k \, \langle B(Y_k^2) - B(Y_k^1), \delta Y_k \rangle_{V',V} \\
&= -\delta\tau_k \, \langle B(\delta Y_k, Y_k^2) + B(Y_k^1, \delta Y_k), \delta Y_k \rangle_{V',V} \\
&= -\delta\tau_k \, \langle B(\delta Y_k, Y_k^2), \delta Y_k \rangle_{V',V} \\
&\le c_B \delta\tau_k \, \|Y_k^2\|_V \|\delta Y_k\|_H \|\delta Y_k\|_V \\
&\le \|\delta Y_k\|_H^2 + \frac{c_B^2 \delta\tau_k^2}{4} \, \|Y_k^2\|_V^2 \|\delta Y_k\|_V^2.
\end{aligned}$$

It follows that

$$\Big(1 - \frac{c_B^2 \delta\tau_k}{4\eta} \, \|Y_k^2\|_V^2 \Big) \, \|\delta Y_k\|_V^2 \le 0.$$

Let $c = \max\{\|Y_k^2\|_V : k = 1, \dots, m\}$. Then $\delta Y_k = 0$ and hence $Y_k^1 = Y_k^2$, provided that $\Delta\tau \le 4\eta/(c^2 c_B^2)$.

A.3. A priori estimates. To prove the estimates (4.5) we take $\psi = Y_k$ in (4.4b). Due to (2.3)–(2.5) and the identity

$$2 \, \langle \varphi - \psi, \varphi \rangle_H = \|\varphi\|_H^2 - \|\psi\|_H^2 + \|\varphi - \psi\|_H^2 \quad \text{for all } \varphi, \psi \in H \tag{A.3}$$

we obtain

$$\|Y_k\|_H^2 - \|Y_{k-1}\|_H^2 + \|Y_k - Y_{k-1}\|_H^2 + 2\eta\delta\tau_k \, \|Y_k\|_V^2 \le 2\delta\tau_k \, \|f(\tau_k)\|_H \|Y_k\|_H.$$

Using (2.2) and Young's inequality it follows that

$$\|Y_k\|_H^2 + \|Y_k - Y_{k-1}\|_H^2 + \eta\delta\tau_k \, \|Y_k\|_V^2 \le \|Y_{k-1}\|_H^2 + \frac{c_V^2 \delta\tau_k}{\eta} \, \|f(\tau_k)\|_H^2. \tag{A.4}$$

From (A.4) and (2.2) we infer that

$$(1 + \gamma\delta\tau_k) \, \|Y_k\|_H^2 \le \|Y_{k-1}\|_H^2 + \frac{\delta\tau_k}{\gamma} \, \|f(\tau_k)\|_H^2,$$

where $\gamma = \eta/c_V^2$, which yields

$$\|Y_k\|_H^2 \le \frac{1}{1 + \gamma\delta\tau} \, \|Y_{k-1}\|_H^2 + \frac{\delta\tau_k}{\gamma(1 + \gamma\delta\tau_k)} \, \|f(\tau_k)\|_H^2. \tag{A.5}$$

From

$$\frac{\delta\tau_k}{1+\gamma\delta\tau_k} = \frac{1}{\gamma}\Big(1-\frac{1}{1+\gamma\delta\tau_k}\Big) \le \frac{1}{\gamma}\Big(1-\frac{1}{1+\gamma\Delta\tau}\Big) = \frac{\Delta\tau}{1+\gamma\Delta\tau}$$

and (A.5) we infer upon summation that

$$\|Y_k\|_H^2 \le \Big(\frac{1}{1+\gamma\delta\tau}\Big)^k \|Y_0\|_H^2 + \frac{\Delta\tau}{\gamma}\,\|f\|^2_{C([0,T];H)} \sum_{j=1}^{k}\Big(\frac{1}{1+\gamma\Delta\tau}\Big)^j. \tag{A.6}$$

Recall that

$$\Big(\frac{1}{1+\gamma\delta\tau}\Big)^k \le (1+\gamma\delta\tau)e^{-\gamma k\delta\tau} \quad\text{and}\quad \Big(\frac{1}{1+\gamma\Delta\tau}\Big)^k \ge e^{-\gamma k\Delta\tau}. \tag{A.7}$$

Moreover, setting $\zeta = 1/(1+\gamma\Delta\tau)$ we find

$$\Delta\tau\sum_{j=1}^{k}\Big(\frac{1}{1+\gamma\Delta\tau}\Big)^j = \Delta\tau\,\frac{1-\zeta^k}{\zeta^{-1}-1} = \frac{1-\zeta^k}{\gamma} \le \frac{1-e^{-\gamma k\Delta\tau}}{\gamma}.$$

Inserting this estimate and (A.7) in (A.6) and utilizing the fact that $\|Y_0\|_H \le \|y_0\|_H$ yield (4.5a). Summing (A.4) over k we find

$$\|Y_m\|_H^2 + \sum_{k=1}^{m}\|Y_k - Y_{k-1}\|_H^2 + \eta\sum_{k=1}^{m}\delta\tau_k\|Y_k\|_V^2 \le \|Y_0\|_H^2 + \frac{c_V T}{\gamma}\|f\|^2_{C([0,T];H)},$$

which is estimate (4.5b).

Appendix B. Proof of Lemma 4.5. Using the notation $\overline{\partial}_\tau\vartheta_k = (\vartheta_k - \vartheta_{k-1})/\delta\tau_k$, $k = 1,\dots,m$, we obtain

$$\begin{aligned}&\langle\overline{\partial}_\tau\vartheta_k,\psi\rangle_H + a(\vartheta_k,\psi) + \langle R\vartheta_k,\psi\rangle_{V',V}\\ &\qquad = \langle v_k,\psi\rangle_H + \langle B(y(\tau_k)) - B(Y_k) + R(y(\tau_k) - P^\ell y(\tau_k)),\psi\rangle_{V',V},\end{aligned} \tag{B.1}$$

where

$$v_k = y_t(\tau_k) - \overline{\partial}_\tau P^\ell y(\tau_k) = y_t(\tau_k) - \overline{\partial}_\tau y(\tau_k) + \overline{\partial}_\tau y(\tau_k) - \overline{\partial}_\tau P^\ell y(\tau_k).$$

We put $w_k = y_t(\tau_k) - \overline{\partial}_\tau y(\tau_k)$ and $z_k = \overline{\partial}_\tau y(\tau_k) - \overline{\partial}_\tau P^\ell y(\tau_k)$. Choosing $\psi = \vartheta_k \in V^\ell$ in (B.1), using (2.4) and (A.3) we infer that

$$\begin{aligned}&\|\vartheta_k\|_H^2 - \|\vartheta_{k-1}\|_H^2 + \|\vartheta_k - \vartheta_{k-1}\|_H^2 + 2\eta\delta\tau_k\,\|\vartheta_k\|_V^2\\ &\le 2\delta\tau_k\big(\|v_k\|_H\|\vartheta_k\|_H + \big|\langle B(y(\tau_k)) - B(Y_k),\vartheta_k\rangle_{V',V}\big| + \|R\varrho_k\|_{V'}\|\vartheta_k\|_V\big).\end{aligned} \tag{B.2}$$

Applying Young's inequality it follows that

$$\|R\varrho_k\|_{V'}\|\vartheta_k\|_V \le \|R\|_{\mathcal{L}(V,V')}\|\varrho_k\|_V\|\vartheta_k\|_V \le \frac{\eta}{4}\,\|\vartheta_k\|_V^2 + c_0\,\|\varrho_k\|_V^2 \tag{B.3}$$

for a constant $c_0 > 0$ depending on $\|R\|_{\mathcal{L}(V,V')}$ and η. We proceed by estimating the nonlinear terms on the right-hand side of (B.2). Note that

$$\begin{aligned}&B(y(\tau_k)) - B(Y_k)\\ &\qquad = -B(y(\tau_k), Y_k - y(\tau_k)) - B(Y_k - y(\tau_k)) - B(Y_k - y(\tau_k), y(\tau_k)).\end{aligned} \tag{B.4}$$

Applying (2.5), (2.2), and Young's inequality we obtain the existence of two constants $c_1, c_2 > 0$ satisfying

$$\begin{aligned} &\big|\langle B(y(\tau_k), Y_k - y(\tau_k)), \vartheta_k\rangle_{V',V}\big| \\ &= \big|\langle B(y(\tau_k), \varrho_k), \vartheta_k\rangle_{V',V}\big| \le c_B c_V^{\delta_3}\, \|y\|_{C([0,T];V)} \|\varrho_k\|_V \|\vartheta_k\|_H^{1-\delta_3} \|\vartheta_k\|_V^{\delta_3} \\ &\le \frac{\eta}{4}\, \|\vartheta_k\|_V^2 + c_1\, \|\vartheta_k\|_H^2 + c_2\, \|\varrho_k\|_V^2. \end{aligned} \tag{B.5}$$

Again utilizing (2.5), Young's inequality, and (2.2) we find that there exist constants $c_3, c_4 > 0$ such that

$$\begin{aligned} &\big|\langle B(Y_k - y(\tau_k), y(\tau_k)), \vartheta_k\rangle_{V',V}\big| \\ &\quad = \big|\langle B(\vartheta_k, y(\tau_k)) + B(\varrho_k, y(\tau_k)), \vartheta_k\rangle_{V',V}\big| \\ &\quad \le c_B\, \|y\|_{C([0,T];V)} \Big(\|\vartheta_k\|_H \|\vartheta_k\|_V + c_V^{\delta_3} \|\varrho_k\|_V \|\vartheta_k\|_H^{1-\delta_3} \|\vartheta_k\|_V^{\delta_3}\Big) \\ &\quad \le \frac{\eta}{4}\, \|\vartheta_k\|_V^2 + c_3\, \|\vartheta_k\|_H^2 + c_4\, \|\varrho_k\|_V^2. \end{aligned} \tag{B.6}$$

From $y \in C([0,T];V)$ it follows that there exists a constant $c_5 > 0$ such that

$$\max_{1\le k\le m} \Big(\|\varrho_k\|_H^{\delta_3} \|\varrho_k\|_V^{1-\delta_3}, \|\varrho_k\|_V\Big) \le c_5. \tag{B.7}$$

Using (2.5) and (4.10) we conclude that

$$\langle B(Y_k - y(\tau_k)), \vartheta_k\rangle_{V',V} = \langle B(\vartheta_k, \varrho_k) + B(\varrho_k, \varrho_k), \vartheta_k\rangle_{V',V}. \tag{B.8}$$

Applying (2.5), (B.7), (B.8), and Young's inequality we find that

$$\begin{aligned} &\big|\langle B(Y_k - y(\tau_k)), \vartheta_k\rangle_{V',V}\big| \\ &\qquad \le c_B c_5 \Big(\|\vartheta_k\|_H \|\vartheta_k\|_V + \|\varrho_k\|_V \|\vartheta_k\|_H^{1-\delta_3} \|\vartheta_k\|_V^{\delta_3}\Big) \\ &\qquad \le \frac{\eta}{4}\, \|\vartheta_k\|_V^2 + c_6\, \|\vartheta_k\|_H^2 + c_7\, \|\varrho_k\|_V^2 \end{aligned} \tag{B.9}$$

for two constants $c_6, c_7 > 0$. From (B.2)–(B.9), Young's inequality, and $v_k = w_k + z_k$ we obtain

$$\|\vartheta_k\|_H^2 \le \|\vartheta_{k-1}\|_H^2 + \delta\tau_k\big(\|w_k\|_H^2 + \|z_k\|_H^2 + c_8\, \|\vartheta_k\|_H^2 + c_9\, \|\varrho_k\|_V^2\big), \tag{B.10}$$

where $c_8 = 2 + c_1 + c_3 + c_6$ and $c_9 = c_0 + c_2 + c_4 + c_7$. Suppose that

$$\Delta\tau \le \frac{1}{2c_8}. \tag{B.11}$$

With (B.11) holding we have $0 < 1 - c_8\delta\tau_k \le 1/2$ and

$$\frac{1}{1 - c_8\delta\tau_k} \le \frac{1}{1 - c_8\Delta\tau} \le 1 + 2c_8\Delta\tau. \tag{B.12}$$

From (B.10) and (B.12) we find that

$$\|\vartheta_k\|_H^2 \le (1 + 2c_8\Delta\tau)\big(\|\vartheta_{k-1}\|_H^2 + \delta\tau_k\big(\|w_k\|_H^2 + \|z_k\|_H^2 + c_9\, \|\varrho_k\|_V^2\big)\big) \tag{B.13}$$

holds. By summation on k we obtain

$$\begin{aligned}\|\vartheta_k\|_H^2 &\le \Big(1+\frac{2c_8\Delta\tau}{\delta\tau}\frac{k\delta\tau}{k}\Big)^k \\ &\quad\cdot\left(\|\vartheta_0\|_H^2+\sum_{j=1}^k\delta\tau_j\left(\|w_j\|_H^2+\|z_j\|_H^2+c_9\ \|\varrho_j\|_V^2\right)\right) \\ &\le e^{c_{10}k\delta\tau}\left(\|\vartheta_0\|_H^2+\sum_{j=1}^k\delta\tau_j\left(\|w_j\|_H^2+\|z_j\|_H^2+c_9\ \|\varrho_k\|_V^2\right)\right),\end{aligned}\tag{B.14}$$

where $c_{10}=2c_8\Delta\tau/\delta\tau$. Recall that by assumption, $\Delta\tau/\delta\tau$ is bounded uniformly with respect to m. We next estimate the terms involving w_j and z_j:

$$\begin{aligned}\sum_{j=1}^k\delta\tau_j\|w_j\|_H^2 &= \sum_{j=1}^k\delta\tau_j\ \|y_t(\tau_j)-\overline{\partial}_\tau y(\tau_j)\|_H^2 \\ &= \sum_{j=1}^k\frac{1}{\delta\tau_j}\ \|\delta\tau_j y_t(\tau_j)-(y(\tau_j)-y(\tau_{j-1}))\|_H^2 \\ &= \sum_{j=1}^k\frac{1}{\delta\tau_j}\left\|\int_{\tau_{j-1}}^{\tau_j}(s-\tau_{j-1})y_{tt}(s)\,ds\right\|_H^2 \\ &\le \sum_{j=1}^k\frac{1}{\delta\tau_j}\int_{\tau_{j-1}}^{\tau_j}(s-\tau_{j-1})^2\,ds\int_{\tau_{j-1}}^{\tau_j}\|y_{tt}(s)\|_H^2\,ds \\ &= \sum_{j=1}^k\frac{\delta\tau_j^2}{3}\ \|y_{tt}\|^2_{L^2(\tau_{j-1},\tau_j;H)}\le\frac{\Delta\tau^2}{3}\ \|y_{tt}\|^2_{L^2(0,\tau_k;H)}.\end{aligned}\tag{B.15}$$

The term $\|z_j\|_H^2$ can be estimated as follows:

$$\begin{aligned}\|z_j\|_H^2 &= \|\overline{\partial}_\tau y(\tau_j)-\overline{\partial}_\tau P^\ell y(\tau_j)\|_H^2 \\ &= \big\|\overline{\partial}_\tau y(\tau_j)-y_t(\tau_j)+y_t(\tau_j)-y_t(t_{\bar j})+y_t(t_{\bar j})-\overline{\partial}_\tau y(t_{\bar j}) \\ &\qquad+\overline{\partial}_\tau y(t_{\bar j})-\overline{\partial}_\tau P^\ell y(t_{\bar j})+\overline{\partial}_\tau P^\ell y(t_{\bar j})-P^\ell y_t(t_{\bar j}) \\ &\qquad+P^\ell y_t(t_{\bar j})-P^\ell y_t(\tau_j)+P^\ell y_t(\tau_j)-\overline{\partial}_\tau P^\ell y(\tau_j)\big\|_H^2 \\ &\le 7\big(1+\|P^\ell\|^2_{\mathcal{L}(H)}\big)\ \|\overline{\partial}_\tau y(\tau_j)-y_t(\tau_j)\|_H^2 \\ &\quad+7\ \big(1+\|P^\ell\|^2_{\mathcal{L}(H)}\big)\ \|y_t(t_{\bar j})-\overline{\partial}_\tau y(t_{\bar j})\|_H^2 \\ &\quad+7\big(1+\|P^\ell\|^2_{\mathcal{L}(H)}\big)\ \|y_t(\tau_j)-y_t(t_{\bar j})\|_H^2 \\ &\quad+7\ \|\overline{\partial}_\tau y(t_{\bar j})-\overline{\partial}_\tau P^\ell y(t_{\bar j})\|_H^2.\end{aligned}$$

Note that

$$\|\overline{\partial}_\tau y(\tau_j)-y_t(\tau_j)\|_H^2=\frac{1}{\delta\tau_j^2}\ \left\|\int_{\tau_{j-1}}^{\tau_j}(t-\tau_{j-1})y_{tt}(t)\,dt\right\|_H^2\le\frac{\delta\tau_j}{3}\ \|y_{tt}\|^2_{L^2(\tau_{j-1},\tau_j;H)}.$$

Analogously, we find

$$\|y_t(t_{\bar j})-\overline{\partial}_\tau y(t_{\bar j})\|_H^2\le\frac{\delta\tau_j}{3}\ \|y_{tt}\|^2_{L^2(t_{\bar j-1},t_{\bar j};H)}.$$

From

$$\|y_t(\tau_j) - y_t(t_{\bar{j}})\|_H^2 \le \left(\int_{t_{\bar{j}-1}}^{t_{\bar{j}+1}} \|y_{tt}(s)\|_H \, ds\right)^2 \le \big(\delta t_{\bar{j}} + \delta t_{\bar{j}+1}\big)\|y_{tt}\|^2_{L^2(t_{\bar{j}-1},t_{\bar{j}+1};H)}$$
$$\le 2\Delta t \, \|y_{tt}\|^2_{L^2(t_{\bar{j}-1},t_{\bar{j}+1};H)},$$

where we set $t_{m+1} = T$ whenever $j = m$, we find

$$\|z_j\|_H^2 \le \frac{7}{3}\big(1 + c_P^2\big)\delta\tau_j\big(\|y_{tt}\|^2_{L^2(\tau_{j-1},\tau_j;H)} + \|y_{tt}\|^2_{L^2(t_{\bar{j}-1},t_{\bar{j}};H)}\big)$$
$$+ 14\big(1 + c_P^2\big)\Delta t \, \|y_{tt}\|^2_{L^2(t_{\bar{j}-1},t_{\bar{j}+1};H)}$$
$$+ \frac{14}{\delta\tau_j^2}\, \big(\|y(t_{\bar{j}}) - P^\ell y(t_{\bar{j}})\|_H^2 + \|y(t_{\bar{j}-1}) - P^\ell y(t_{\bar{j}-1})\|_H^2\big),$$

where we set $t_{n+1} = T$. Note that $\alpha_j \ge \delta t/2$. Using (2.2) and Lemma 4.1 we estimate

$$\sum_{j=1}^k \frac{1}{\delta\tau_j} \, \|y(t_{\bar{j}}) - P^\ell y(t_{\bar{j}})\|_H^2 \le \frac{2\sigma_n}{\delta\tau\delta t} \sum_{j=0}^n \alpha_j \|y(t_j) - P^\ell y(t_j)\|_H^2 \le \frac{2\sigma_n c_V^2}{\delta\tau\delta t} \sum_{i=\ell+1}^d \lambda_i$$

and, analogously,

$$\sum_{j=1}^k \frac{1}{\delta\tau_j} \, \|y(t_{\bar{j}-1}) - P^\ell y(t_{\bar{j}-1})\|_H^2 \le \frac{2\sigma_n c_V^2}{\delta\tau\delta t} \sum_{i=\ell+1}^d \lambda_i.$$

Hence,

$$\sum_{j=1}^k \delta\tau_j \, \|z_j\|_H^2 \le 14\sigma_n(1 + c_P^2)(\Delta\tau^2 + \Delta\tau\Delta t)\|y_{tt}\|^2_{L^2(0,t_{\bar{k}+1};H)}$$
$$+ \frac{56\sigma_n c_V^2}{\delta t\delta\tau} \sum_{i=\ell+1}^d \lambda_i. \tag{B.16}$$

Using $\alpha_j \ge 2/\delta t$ and $\|P^\ell\|_{\mathcal{L}(V)} = 1$ we obtain for the terms $\|\varrho_k\|_V^2$

$$\|\varrho_j\|_V^2 = \|P^\ell y(\tau_j) - y(\tau_j)\|_V^2$$
$$= \|P^\ell y(\tau_j) - P^\ell y(t_{\bar{j}}) + P^\ell y(t_{\bar{j}}) - y(t_{\bar{j}}) + y(t_{\bar{j}}) - y(\tau_j)\|_V^2$$
$$\le 4 \, \|y(t_{\bar{j}}) - y(\tau_j)\|_V^2 + 2 \, \|P^\ell y(t_{\bar{j}}) - y(t_{\bar{j}})\|_V^2$$
$$\le 8\Delta t \, \|y_t\|^2_{L^2(t_{\bar{j}-1},t_{\bar{j}+1};V)} + \frac{4\alpha_{\bar{j}}}{\delta t} \, \|P^\ell y(t_{\bar{j}}) - y(t_{\bar{j}})\|_V^2.$$

Thus, we get

$$\sum_{j=1}^k \delta\tau_j \, \|\varrho_j\|_V^2 \le 4\sigma_n\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,t_{\bar{k}+1};V)} + \frac{4\sigma_n\Delta\tau}{\delta t} \sum_{i=\ell+1}^d \lambda_i. \tag{B.17}$$

Combining (B.14)–(B.17) the claim follows.

REFERENCES

[1] K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, in Shape Optimization and Optimal Design, Lecture Notes in Pure and Appl. Math. 216, Marcel Dekker, New York, 2001, pp. 317–332.

[2] H. W. Alt, *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*, Springer-Verlag, Berlin, 1992.

[3] J. A. Atwell and B. B. King, *Reduced order controllers for spatially distributed systems via proper orthogonal decomposition*, SIAM J. Sci. Comput., submitted.

[4] N. Aubry, W.-Y. Lian, and E. S. Titi, *Preserving symmetries in the proper orthogonal decomposition*, SIAM J. Sci. Comput., 14 (1993), pp. 483–505.

[5] H. T. Banks, M. L. Joyner, B. Winchesky, and W. P. Winfree, *Nondestructive evaluation using a reduced-order computational methodology,* Inverse Problems, 16 (2000), pp. 1–17.

[6] H. T. Banks, R. C. H. del Rosario, and R. C. Smith, *Reduced Order Model Feedback Control Design: Computational Studies for Thin Cylindrical Shells*, Technical report CRSC-TR98-25, North Carolina State University, Raleigh, NC, 1998.

[7] G. Berkooz, P. Holmes, and J. L. Lumley, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monogr. Mech., Cambridge University Press, Cambridge, UK, 1996.

[8] F. Diwoky and S. Volkwein, *Nonlinear boundary control for the heat equation utilizing proper orthogonal decomposition*, in Fast Solutions of Discretized Optimization Problems, Internat. Ser. Numer. Math. 138, K.-H. Hoffmann, R. H. W. Hoppe, and V. Schulz, eds., Birkhäuser, Basel, 2001, pp. 73–87.

[9] M. Fahl, *Computation of POD basis functions for fluid flows with Lanczos methods*, Math. Comput. Modelling, 34 (2001), pp. 91–107.

[10] K. Fukunaga, *Introduction to Statistical Recognition*, Academic Press, New York, 1990.

[11] D. Gilbarg and N. S. Trudinger, *Elliptic Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[12] D. Hömberg and S. Volkwein, *Suboptimal Control of Laser Surface Hardening Using Proper Orthogonal Decomposition*, Technical report 217, Special Research Center F 003 Optimization and Control, Project area Continuous Optimization and Control, University of Graz and Technical University of Graz, Graz, Austria, 2001.

[13] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[14] G. M. Kepler, H. T. Tran, and H. T. Banks, *Compensator control for chemical vapor deposition film growth used reduced order design models*, IEEE Trans. on Semiconducter Manufacturing, to appear.

[15] K. Kunisch and S. Volkwein, *Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1999), pp. 345–371.

[16] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math., 90 (2001), pp. 117–148.

[17] H. V. Ly and H. T. Tran, *Proper orthogonal decomposition for flow calculations and optimal control in a horizontal CVD reactor*, Quart. Appl. Math., to appear.

[18] H. V. Ly and H. T. Tran, *Modelling and control of physical processes using proper orthogonal decomposition*, Math. Comput. Modelling, 33 (2001), pp. 223–236.

[19] M. Manhart, *Umströmung einer Halbkugel in turbulenter Grenzschicht,* VDI-Verlag, Düsseldorf, 1996.

[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* I: *Functional Analysis*, Academic Press, New York, 1980.

[21] S.Y. Shvartsman and Y. Kevrikidis, *Nonlinear model reduction for control of distributed parameter systems: A computer-assisted study*, AIChE J., 44 (1998), pp. 1579–1595.

[22] L. Sirovich, *Turbulence and the dynamics of coherent structures.* I. *Coherent structures,* Quart. Appl. Math., 45 (1987), pp. 561–571.

[23] L. Sirovich, *Turbulence and the dynamics of coherent structures.* II. *Symmetries and transformations,* Quart. Appl. Math., 45 (1987), pp. 573–582.

[24] L. Sirovich, *Turbulence and the dynamics of coherent structures.* III. *Dynamics and scaling,* Quart. Appl. Math., 45 (1987), pp. 583–590.

[25] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci., 68, Springer-Verlag, New York, 1988.

[26] S. Volkwein, *Optimal control of a phase-field model using the proper orthogonal decomposition,* Z. Angew. Math. Mech., 81 (2001), pp. 83–97.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 516–542

LYAPUNOV SPECTRAL INTERVALS: THEORY AND COMPUTATION*

LUCA DIECI† AND ERIK S. VAN VLECK‡

Abstract. Different definitions of spectra have been proposed over the years to characterize the asymptotic behavior of nonautonomous linear systems. Here, we consider the spectrum based on exponential dichotomy of Sacker and Sell [*J. Differential Equations,* 7 (1978), pp. 320–358] and the spectrum defined in terms of upper and lower Lyapunov exponents. A main goal of ours is to understand to what extent these spectra are computable. By using an orthogonal change of variables transforming the system to upper triangular form, and the assumption of integral separation for the diagonal of the new triangular system, we justify how popular numerical methods, the so-called continuous QR and SVD approaches, can be used to approximate these spectra. We further discuss how to verify the property of integral separation, and hence how to a posteriori infer stability of the attained spectral information. Finally, we discuss the algorithms we have used to approximate the Lyapunov and Sacker–Sell spectra and present some numerical results.

Key words. Lyapunov exponents, Sacker–Sell spectrum, integral separation, numerical computation

AMS subject classification. 65L

PII. S0036142901392304

1. Introduction.

Lyapunov exponents, or Lyapunov characteristic numbers, characterize growth rates of time dependent linear differential equations and, by linearizing about trajectories, measure rates of convergence or divergence of nearby trajectories for nonlinear differential equations. For an n-dimensional problem, there are n Lyapunov exponents: these are the natural generalization to time dependent linear differential equations of the eigenvalues for autonomous linear systems. Although Lyapunov exponents are a set of n points, it is perhaps more natural to think of the spectrum of a linear nonautonomous system as possibly being a continuum. For example, consider the linear scalar differential equation $\dot{x} = (\sin(\ln(t)) + \cos(\ln(t)))x$ for $t \geq t_0 > 0$: the solution is $x(t) = \exp(t\sin(\ln(t)))\kappa_0$, $\kappa_0 = x(t_0)\exp(-t_0\sin(\ln(t_0)))$, so that all growth rates in the interval $[-1, +1]$ are attained.

This work is an attempt to blend the numerical techniques developed to approximate Lyapunov exponents with stability theory for Lyapunov exponents developed over 30 years ago. Characteristic exponents were developed by Lyapunov in his thesis [22] that was first published in 1892. Many of the ideas from Lyapunov's thesis and further developments on Lyapunov exponents are contained in the monograph of Adrianova [1] which serves as an excellent accessible introduction to the use of Lyapunov exponents in stability theory. Important results on stability of Lyapunov exponents that we use are due to Bylov [6], Bylov et al. [5], Bylov and Izobov [7], and Millionshchikov [24, 25]. An alternative to the spectrum of Lyapunov is based upon defining a spectrum in terms of exponential dichotomy. Important works are the book

*Received by the editors July 16, 2001; accepted for publication (in revised form) January 4, 2002; published electronically May 29, 2002. This work was supported in part by NSF grants DMS-9973226 and DMS-9973393.

http://www.siam.org/journals/sinum/40-2/39230.html

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 80302 (dieci@math.gatech.edu).

‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (evanvlec@mines.edu).

of Coppel [9] on exponential dichotomy in stability theory, the work of Sacker and Sell [30] which defines a spectrum in terms of exponential dichotomy, and the work of Palmer [28] who showed that the structurally stable linear systems on the half-line are those with exponential dichotomy.

A contribution of this paper is to show, under certain natural conditions, the relationship between three definitions of spectra. The first spectrum is commonly referred to as the Sacker–Sell spectrum and its origin may be traced back to [30]. The second spectrum generalizes the original definition of Lyapunov [22] so that it may be viewed as a continuous spectrum. The third spectrum is motivated by computational considerations, since its definition is based upon the information one may be able to retrieve when using the so-called QR method to approximate Lyapunov exponents.

The assumption under which we are able to show the relationship between these three spectra is *integral separation*. It has been well known in the theoretical community (see the results summarized in [1]) that, for systems with distinct Lyapunov exponents, integral separation is a necessary and sufficient condition for stability of the exponents, i.e., for continuity of the exponents with respect to changes in the coefficient matrix. Thus, it is natural to assume such a condition if we are interested in numerical approximation of the Lyapunov exponents.

We will emphasize how integral separation can be characterized for the numerical techniques that have been proposed to approximate Lyapunov exponents.

1. The *continuous QR method* is based upon finding an orthogonal change of variables transforming the system to upper triangular form. Then, the Lyapunov exponents are determined from the diagonal elements of the new system. The approach can be made legitimate under the assumption of *regularity* of the system. However, in spite of being a strong assumption, regularity does not ensure stability of the exponents. This motivated us to consider integral separation of the diagonal of the upper triangular coefficient matrix: we prove that this is sufficient for stability of the Lyapunov exponents.
2. We also consider a method for finding Lyapunov exponents based upon decomposing a fundamental matrix solution via a smooth singular value decomposition, the *SVD approach*. If such decomposition is feasible,[1] then the system is transformed to diagonal form, and the Lyapunov exponents are extracted from time averages of the diagonal system. Again, this can be justified under the assumption of regularity. But, rather, we show that if the new diagonal system has an integrally separated diagonal, then the Lyapunov exponents can be found from the diagonal system and are stable.

In spite of their importance in the physical sciences, Lyapunov exponents have received little attention from the numerical community. This is certainly due to the inherent difficulties (and uncertainties) present in the task, but we believe that it is also due to the fact that stability theory for Lyapunov exponents is not as well known as it should be. For this reason, and also to make the present work self-contained, the first two sections of this paper present background information. Sections 2 and 3 summarize results from [1] on Lyapunov exponents and on equivalence between stability of distinct Lyapunov exponents and integral separation. Section 4 summarizes the three spectra we consider. Sections 5, 6, and 7 contain our main results: under assumptions of integral separation, we show some relationships between the three spectra. Further, we validate the QR and SVD techniques to find the Lyapunov spectra. In section 8, we detail numerical techniques based on the continuous QR method to approximate

[1] E.g., it is feasible if the singular values stay distinct for all times t.

the spectra, and we also discuss how we can attempt to verify integral separation of a system. Finally, we give some new results on the relation between integral separation and the Sacker–Sell spectrum and outline a computational procedure to approximate such a spectrum. In Section 9 we present numerical experiments. Section 10 contains conclusions.

2. Lyapunov exponents theory. The *characteristic exponent* of a (nonvanishing) function $f(t)$ is defined as

$$\chi(f) = \limsup_{t\to\infty} \frac{1}{t} \ln |f(t)|. \tag{2.1}$$

The following equalities relate the upper and lower characteristic exponents of f and $1/f$ and will be useful when relating the exponents of a linear system and of its adjoint:

$$\begin{array}{rcl} \limsup_{t\to\infty} \frac{1}{t} \ln |f(t)| & = & -\liminf_{t\to\infty} \frac{1}{t} \ln |1/f(t)|, \\ \liminf_{t\to\infty} \frac{1}{t} \ln |f(t)| & = & -\limsup_{t\to\infty} \frac{1}{t} \ln |1/f(t)|. \end{array} \tag{2.2}$$

We now summarize some results on properties of characteristic exponents.

THEOREM 2.1 ([1, Thms. 2.1.2 and 2.1.4]). *The characteristic exponent of a product does not exceed the sum of the characteristic exponents, i.e.,* $\chi(fg) \le \chi(f) + \chi(g)$. *Moreover, if* $\chi(f) + \chi(1/f) = 0$, *then* $\chi(fg) = \chi(f) + \chi(g)$.

DEFINITION 2.2. *The Lyapunov exponent of a vector valued function* $x : t \in \mathbb{R} \to \mathbb{R}^n$ *is defined as the Lyapunov exponent of the norm:* $\chi(x) = \chi(||x||)$.

In this work, we restrict our consideration to the 2-norm, $\|x(t)\|_2$, and similarly for matrix valued functions. The advantage is that these are invariant under orthogonal transformations, but similar results would hold for different norms.

Consider now an n-dimensional linear system

$$\dot{x} = A(t)x\,, \tag{2.3}$$

where A is continuous and bounded: $\sup_t \|A(t)\| < \infty$. Given a fundamental matrix solution X of (2.3), consider the quantities

$$\lambda_i = \limsup_{t\to\infty} \frac{1}{t} \ln ||X(t)e_i||, \qquad i = 1, \ldots, n, \tag{2.4}$$

where e_i denotes the ith standard unit vector. When $\sum_{i=1}^n \lambda_i$ is minimized with respect to all possible fundamental matrix solutions, then the λ_i are called the Lyapunov exponents, or Lyapunov characteristic numbers, and the corresponding fundamental matrix solution is called a *normal basis*. In general, the Lyapunov exponents satisfy

$$\sum_{i=1}^{n} \lambda_i \ge \limsup_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{Tr}(A(s))ds, \tag{2.5}$$

where $\mathrm{Tr}(A(s))$ is the trace of the matrix $A(s)$.

Remark 2.1. The Lyapunov exponents are unaffected by what happens to X on a finite interval. For this reason, in (2.5) and elsewhere in this paper, one may replace 0 with any other (finite) value of t. With this in mind, we will continue using 0 as the lower limit of integration.

Along with (2.3), we will also need to consider the associated adjoint equation

$$\dot{y}(t) = -A^T(t)y(t). \tag{2.6}$$

Similarly to (2.4), one can define the Lyapunov exponents for (2.6); call them $\{-\mu_i\}_{i=1}^n$. We will henceforth restrict our consideration to the system (2.3) and the λ_i exponents only, but of course everything can be formulated also in terms of the adjoint system (2.6) and the μ_i's.

Given any fundamental matrix solution, Lyapunov showed how to construct a normal fundamental matrix solution.

THEOREM 2.3 (see Lyapunov's construction of a normal basis [22]). *Consider a matrix solution $Z(\cdot) = [Z_1, \ldots, Z_n]$ such that the Lyapunov exponents of the columns of Z are ordered as $\chi(Z_1) \geq \cdots \geq \chi(Z_n)$. Then, there exists a unit upper triangular matrix C such that $X(\cdot) = Z(\cdot)C$ is normal. Similarly, if the Lyapunov exponents of the columns of Z are ordered as $\chi(Z_1) \leq \cdots \leq \chi(Z_n)$, then there exists a unit lower triangular matrix C such that $X(\cdot) = Z(\cdot)C$ is normal.*

Remark 2.2. The assumption of ordered characteristic exponents for the columns of Z is not stringent, since it can be trivially achieved via column permutation of any matrix solution. In the original work of Lyapunov (see also [1]), the matrix C was taken as a unit lower triangular with the corresponding assumption that the growth rates of the columns of Z are ordered as $\chi(Z_1) \leq \cdots \leq \chi(Z_n)$. However, the ordering in which C is taken to be unit upper triangular is more natural for us, since often we end up working with upper triangular systems, and we should expect that the growth rates will be ordered from largest down to smallest. On the other hand, when working with the adjoint, it is the reverse ordering which is more natural; hence the use of a unit lower triangular C is more appropriate in this case. Indeed (see [1, Cor. 3.6.2]), if the basis X is normal for (2.3), then the basis X^{-T} is normal for the adjoint system; here and elsewhere in this work, X^{-T} is shorthand notation for $(X^{-1})^T$. Conceptually, then, we can always work with a normal basis and assume to have ordered Lyapunov exponents for a system and its adjoint:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \quad \text{and} \quad -\mu_n \geq \cdots \geq -\mu_2 \geq -\mu_1 .$$

Indeed, we will henceforth assume that we are working with a normal matrix solution X.

A fundamental property of Lyapunov exponents is that they (and their stability properties) are preserved under *Lyapunov transformations*.

DEFINITION 2.4. *A smooth invertible change of variables $y \leftarrow T^{-1}x$ is called a Lyapunov transformation if T, T^{-1}, and $\dot{T}$ are bounded.*

Clearly, under a Lyapunov transformation, (2.3) is transformed into

$$\dot{y} = B(t)y, \qquad B = T^{-1}AT - \dot{T}T^{-1} . \tag{2.7}$$

For example, it has been known since Perron [29] and Diliberto [17] that there exists a Lyapunov, and orthogonal, change of variables for which B is upper triangular. To see this, write a fundamental matrix solution $X(t)$ as $Q(t)R(t)$, where Q is an orthogonal matrix valued function and R is an upper triangular matrix valued function with positive diagonal entries. Upon differentiating we have

$$AQR = Q\dot{R} + \dot{Q}R \quad \text{or} \quad \dot{Q} = AQ - QB. \tag{2.8}$$

Since $\dot{R} = BR$, then B is upper triangular. Since Q is orthogonal, if we let $S(Q) := Q^T\dot{Q} = Q^TAQ - B$, then the strict lower triangular piece of the skew symmetric

function S can be defined as the corresponding piece of Q^TAQ, and the rest of S is given by skew-symmetry.

Remark 2.3. In what follows, when considering upper triangular systems $\dot{R} = BR$, we will always assume that the diagonal entries of R are positive.

Linear systems for which the Lyapunov exponents exist as limits were called *regular* by Lyapunov.

DEFINITION 2.5. *A system is regular (Lyapunov) if the time average of the trace has a finite limit and equality holds in* (2.5).

Example 2.1. A simple example of a linear system where a strict inequality holds in (2.5) is

$$\begin{aligned}\dot{x} &= (\sin(\ln t) + \cos(\ln t))y,\\ \dot{y} &= (\sin(\ln t) + \cos(\ln t))x\end{aligned}$$

which has Lyapunov exponents $\lambda_1 = \lambda_2 = 1$, but $\limsup_{t\to\infty} \frac{1}{t}\int_0^t \operatorname{trace}(A(s))ds = 0$.

It was shown by Lyapunov that regularity is maintained under Lyapunov transformations and, in particular, for a regular triangular system $\dot{R} = B(t)R$ the Lyapunov exponents may be obtained as time averages of the diagonal elements of B:

$$\lambda_j = \lim_{t\to\infty} \frac{1}{t}\int_0^t B_{jj}(s)ds, \qquad j = 1,\dots,n. \tag{2.9}$$

Further, in this regular case, the μ_i exponents of the adjoint system equal the λ_i exponents.

3. Stability of Lyapunov exponents and integral separation. In this section we summarize results on the relation between stability of the exponents and the property of integral separation.

DEFINITION 3.1. *The characteristic exponents $\lambda_1 \geq \cdots \geq \lambda_n$ of system* (2.3) *are said to be stable if for any $\epsilon > 0$ there exists $\delta > 0$ such that $\sup_{t\in\mathbb{R}^+} ||E(t)|| < \delta$ implies*

$$|\lambda_i - \gamma_i| < \epsilon, \qquad i = 1,\dots,n, \tag{3.1}$$

where the γ_i's are the (ordered) Lyapunov exponents of the perturbed system $\dot{x} = [A(t) + E(t)]x$.

Naturally, since Lyapunov transformations preserve the exponents and the smallness of perturbations, stability of the characteristic exponents is invariant under Lyapunov transformations.

THEOREM 3.2 (see [1, Thm. 5.2.1]). *If the λ_i exponents of* (2.3) *are stable, and $E \to 0$ as $t \to \infty$, then the exponents of the perturbed system are also given by the λ_i's.*

DEFINITION 3.3 (see [1, Def. 5.3.2] and [6]). *Write a fundamental matrix solution columnwise $X(t) = [X_1(t),\dots,X_n(t)]$. Then, X is integrally separated if for $i = 1,\dots,n-1$ there exist $a > 0$ and $d > 0$ such that*

$$\frac{||X_i(t)||}{||X_i(s)||} \cdot \frac{||X_{i+1}(s)||}{||X_{i+1}(t)||} \geq de^{a(t-s)} \tag{3.2}$$

for all $t, s \ : \ t \geq s$.

Again, if a matrix solution X is integrally separated, and T is a Lyapunov transformation, then the matrix solution $Y \leftarrow T^{-1}X$ associated with (2.7) is also integrally separated; i.e., integral separation is kept under Lyapunov transformations.

THEOREM 3.4 (see [1, Props. 5.3.1 and 5.3.3]). *Integrally separated systems have distinct Lyapunov exponents.*

DEFINITION 3.5. *The functions g_i, $i = 1, \ldots, n$, are said to be integrally separated if for $i = 1, \ldots, n-1$,*

$$\int_s^t (g_i(\tau) - g_{i+1}(\tau))d\tau \geq a(t-s) - d, \qquad t \geq s,\ a > 0,\ d \in \mathbb{R}. \tag{3.3}$$

THEOREM 3.6 (see [1, Thm. 5.4.7] and [7]). *If the system* (2.3) *has distinct characteristic exponents $\lambda_1 > \cdots > \lambda_n$, then they are stable if and only if there exists a Lyapunov transformation $z \leftarrow T^{-1}x$ transforming* (2.3) *to the diagonal form*

$$\dot{z} = \operatorname{diag}[p_1(t), \ldots, p_n(t)]z, \tag{3.4}$$

where the diagonal elements, the p_i, are integrally separated functions.

THEOREM 3.7 (see [1, Thm. 5.4.8] and [7]). *If the system* (2.3) *has distinct characteristic exponents $\lambda_1 > \cdots > \lambda_n$, then they are stable if and only if there exists a fundamental matrix solution with integrally separated columns, as in Definition* 3.3.

Given the implications of integral separation, it is a comforting fact that it is a natural condition to have. This is because of a result of Palmer [28, p. 21]. Palmer considered the Banach space $\mathcal{B}$, of continuous bounded matrix valued functions A, with norm $||A|| = \sup_{t\geq 0} ||A(t)||$, and—using results from [24] and [5]—he showed that the systems with integral separation form an open and dense subset of $\mathcal{B}$. Therefore, integral separation is a generic property in $\mathcal{B}$.

Regularity (see Definition 2.5), however, is not enough to ensure stability and hence integral separation, as the following example from [1, p. 171] shows. Consider the regular system

$$\begin{aligned} \dot{x}_1 &= \left(1 + \frac{\pi}{2}\sin(\pi\sqrt{t})\right) x_1, \\ \dot{x}_2 &= 0, \end{aligned} \tag{3.5}$$

which has distinct Lyapunov exponents $\lambda_1 = 1$ and $\lambda_2 = 0$. Since for any $n \in \mathbb{N}$,

$$\int_{(2n-1)^2}^{(2n)^2} \left(1 + \frac{\pi}{2}\sin(\pi\sqrt{t})\right) d\tau = 0, \tag{3.6}$$

then the system (3.5) is not integrally separated and hence the Lyapunov exponents are not stable.

Remark 3.1. In all numerical works on approximation of Lyapunov exponents of which we are aware, it is assumed that system (2.3) is regular; e.g., see [2, 3, 12, 13, 18, 19, 20, 21]. This is justified on the grounds that regularity is a prevalent condition in a measure theoretic sense; see [27]. However, from the numerical point of view, we need to insist that the Lyapunov exponents be stable, and for this to be true we need integral separation, not regularity.

We now show that the adjoint system (2.6) has an integrally separated fundamental matrix solution if the original system (2.3) does.

LEMMA 3.8. *If* (2.3) *has a fundamental matrix solution with integrally separated columns, then the adjoint* (2.6) *has a fundamental matrix solution with integrally separated columns.*

Proof. Because of (2.7) and (2.8), we may consider, without loss of generality, an upper triangular system $\dot{R} = BR$ with an integrally separated fundamental matrix solution R. Then $S = R^{-T}$ satisfies $\dot{S} = -B^T S$. Since R has integrally separated columns, by Theorems 3.6 and 3.7, there exists a Lyapunov transformation L such that $D = \text{diag}(p_i) = L^{-1}BL - L^{-1}\dot{L}$ and the p_i are integrally separated, i.e., they satisfy (3.3). Let $Y = L^{-1}R$; then Y is integrally separated and $Y = \text{diag}(\exp(\int_0^t p_i(s)ds))$. Let $Z = (L^{-T})^{-1}S$ so that $Z = Y^{-T}$ and Z satisfies $\dot{Z} = -D^T Z = -DZ$. Then $Z_{ii}(t) = \exp(-\int_0^t p_i(s)ds)$ for $i = 1, \ldots, n$, and so

$$\frac{Z_{i,i}(t)}{Z_{i,i}(s)} \cdot \frac{Z_{i+1,i+1}(s)}{Z_{i+1,i+1}(t)} = \frac{Y_{i+1,i+1}(t)}{Y_{i+1,i+1}(s)} \cdot \frac{Y_{i,i}(s)}{Y_{i,i}(t)} \geq d\exp(a(t-s)),$$

$$a > 0,\ t \geq s,\ \ i = 1, \ldots, n-1.$$

Thus, by Theorems 3.6 and 3.7 the adjoint equation has an integrally separated fundamental matrix solution. ∎

4. Three definitions of spectra. Consider (2.3). It is well known that if $A(\cdot)$ is constant, then the asymptotic stability properties of the zero solution of (2.3) are determined by the real parts of the eigenvalues of A and the corresponding eigenvectors. In the case in which A is periodic in t, the Floquet theory effectively reduces the question of stability to the constant coefficient case. For the general case, we recall the next two classical concepts of stability, and we introduce a third related one.

4.1. Sacker–Sell spectrum. In [30], Sacker and Sell introduced a spectrum for (2.3) based upon exponential dichotomy: the Sacker–Sell spectrum is given by those values $\lambda \in \mathbb{R}$ such that the shifted system $\dot{x} = [A(t) - \lambda I]x$ does not have exponential dichotomy. We will indicate the Sacker–Sell spectrum with Σ_{ED}. Recall that the system (2.3) has *exponential dichotomy* if for a fundamental matrix solution X there exists a projection P and constants $\alpha, \beta > 0$ and $K, L \geq 1$, such that

$$\begin{array}{rcll} \|X(t)PX^{-1}(s)\| & \leq & Ke^{-\alpha(t-s)}, & t \geq s, \\ \|X(t)(I-P)X^{-1}(s)\| & \leq & Le^{\beta(t-s)}, & t \leq s. \end{array} \tag{4.1}$$

It is shown in [30] that Σ_{ED} is given by the union of at most n closed intervals. Thus, it can be written, for some k: $1 \leq k \leq n$, as

$$\Sigma_{\text{ED}} := [a_1, b_1] \cup \cdots \cup [a_k, b_k]. \tag{4.2}$$

4.2. Lyapunov spectrum. Another characterization of spectrum is based on the characteristic exponents of (2.3) and (2.6), the λ_i's and $-\mu_i$'s which we can consider as being ordered: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. We define the Lyapunov spectrum, written Σ_{L}, as

$$\Sigma_{\text{L}} := \bigcup_{j=1}^{n} [\lambda_j^i, \lambda_j^s], \tag{4.3}$$

where $\lambda_j^i = \mu_j$ and $\lambda_j^s = \lambda_j$ and, in fact, $\lambda_j \geq \mu_j$ for $j = 1, \ldots, n$. The last statement is a consequence of the fact that the normal bases for (2.3) and (2.6) are X and X^{-T}, so, if $\lambda_j = \chi(Xe_j)$, then $-\mu_j = \chi(X^{-T}e_j)$. But obviously, $(X^{-T}e_j)^T(Xe_j) = 1$ for all t so that Theorem 2.1 gives $\lambda_j \geq \mu_j$.

Remark 4.1. Our definition of Lyapunov spectrum is strictly related to the *coefficient of irregularity* of Perron, who proved that a system is regular if and only if $\lambda_j = \mu_j$.

Remark 4.2. It must be appreciated that Σ_{L} and Σ_{ED} provide information on related, but different, questions. In particular, $\lambda \notin \Sigma_{\mathrm{L}}$ implies the existence of a bounded solution to the homogeneous problem $\dot{x} = (A(t) - \lambda I)x$, for some initial condition $x(0)$. Instead, $\lambda \notin \Sigma_{\mathrm{ED}}$ implies both the existence of a bounded solution to the homogeneous problem $\dot{x} = (A(t) - \lambda I)x$ for some $x(0)$ and the existence of a bounded solution to the nonhomogeneous problem $\dot{x} = (A(t) - \lambda I)x + f(t)$ for any (continuous and bounded) function $f(t)$, a condition which is not guaranteed by $\lambda \notin \Sigma_{\mathrm{L}}$. Obviously, both properties are quite important, and it depends on the particular application in which we are interested whether we need to know Σ_{ED} or whether knowledge of Σ_{L} is sufficient.

4.3. Computed Lyapunov spectrum. The third spectrum we consider is what we will call the *computed Lyapunov spectrum*, since it is close to what traditionally has been approximated. Its definition rests on the transformation of (2.3) to upper triangular form via an orthogonal change of variables; see (2.7) and (2.8). Consider the upper triangular system $\dot{R} = BR$. We define the computed Lyapunov spectrum, written Σ_{CL}, as

$$\Sigma_{\mathrm{CL}} := \bigcup_{j=1}^{n} [\lambda_{jj}^i, \lambda_{jj}^s], \quad \lambda_{jj}^i = \liminf_{t\to\infty} \frac{1}{t}\int_0^t B_{jj}(s)ds, \quad \lambda_{jj}^s = \limsup_{t\to\infty} \frac{1}{t}\int_0^t B_{jj}(s)ds. \tag{4.4}$$

5. The Lyapunov and computed Lyapunov spectra. In this section we prove that for upper triangular systems, integral separation of the diagonal elements implies that the Lyapunov spectrum, Σ_{L}, and the computed Lyapunov spectrum, Σ_{CL}, coincide. We prove this by constructing a bounded Lyapunov transformation that transforms the upper triangular system to a diagonal system given by the diagonal of the upper triangular system.

THEOREM 5.1. *For an upper triangular system $\dot{R} = BR$ with B smooth and bounded, integral separation of the diagonal of B implies $\Sigma_{\mathrm{L}} = \Sigma_{\mathrm{CL}}$.*

Proof. The proof is by induction. Write B in block form and define a transformation T_1 using the same blocking:

$$B = \begin{pmatrix} b_{11} & b_{12} & B_{13} \\ 0 & b_{22} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix} \quad \text{and} \quad T_1 = \begin{pmatrix} 1 & x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{5.1}$$

We want to take x such that

$$T_1^{-1}BT_1 - T_1^{-1}\dot{T}_1 = \begin{pmatrix} b_{11} & 0 & B_{13} - xB_{23} \\ 0 & b_{22} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix}. \tag{5.2}$$

To obtain this, we take x satisfying

$$\begin{cases} \dot{x} = b_{11}x - xb_{22} + b_{12}, \\ \lim\limits_{T\to\infty} x(T) = 0; \end{cases} \tag{5.3}$$

that is,

$$x(t) = -\lim_{T\to\infty} \int_t^T \exp\left(-\int_t^s (b_{11}(\tau) - b_{22}(\tau))d\tau\right) b_{12}(s)ds. \tag{5.4}$$

Since the diagonal elements of B are integrally separated (see (3.3)), we have

$$-\int_t^s (b_{11}(\tau) - b_{22}(\tau))d\tau \le -a(s-t) + d, \qquad a > 0,\ s \ge t, \tag{5.5}$$

which implies that x is bounded and the transformed coefficient matrix is bounded.

Now we assume that the matrix function B has been progressively diagonalized in its first p columns so that the transformed coefficient matrix has the form

$$B = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ 0 & b_{p+1,p+1} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix}, \tag{5.6}$$

where $B_{11} : t \to \mathbb{R}^{p\times p}$ is diagonal, and $B_{12} : t \to \mathbb{R}^{p\times 1}$, $B_{13} : t \to \mathbb{R}^{p\times(n-p-1)}$, $B_{23} : t \to \mathbb{R}^{1\times(n-p-1)}$ are all continuous and bounded. Consider the transformation T_p and transformed coefficient matrix of the form

$$T_p = \begin{pmatrix} I_p & x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I_{n-p-1} \end{pmatrix}, \qquad T_p^{-1}BT_p - T_p^{-1}\dot{T}_p = \begin{pmatrix} B_{11} & 0 & B_{13} - xB_{23} \\ 0 & b_{p+1,p+1} & B_{23} \\ 0 & 0 & B_{33} \end{pmatrix}, \tag{5.7}$$

where we require that x satisfies

$$\dot{x} = B_{11}x - xb_{p+1,p+1} + B_{12} = (B_{11} - b_{p+1,p+1}I)x + B_{12} \tag{5.8}$$

and $\lim_{T\to\infty} x(T) = 0$. Then, since B_{11} is diagonal,

$$x(t) = -\lim_{T\to\infty} \int_t^T \exp\left(-\int_t^s (B_{11}(\tau) - I \cdot b_{p+1,p+1}(\tau))d\tau\right) B_{12}(s)ds\,.$$

Since B_{12} is bounded and the diagonal of B is integrally separated, we have that x is bounded and T_p is Lyapunov. Since Lyapunov transformations preserve the Lyapunov spectrum, the result follows. ∎

The following corollary is an immediate consequence of the above proof and Theorems 3.6 and 3.7.

COROLLARY 5.1. *Given an upper triangular system $\dot{R} = BR$ with B smooth, bounded, and with integrally separated diagonal, then there exists an integrally separated fundamental matrix solution.*

As a partial converse to Theorem 5.1 we have the following.

THEOREM 5.2. *Suppose the system $\dot{R} = BR$, with B bounded, continuous and upper triangular, has an integrally separated fundamental matrix solution R. Then for all $\epsilon > 0$ there exists a permutation π such that $|\lambda^s_{\pi(i)} - \limsup_{t\to\infty} \frac{1}{t}\int_0^t B_{ii}(s)ds| < \epsilon$.*

Proof. Consider the system $\dot{D} = \mathrm{diag}(B)D$ and let $\lambda_i(D) = \limsup_{t\to\infty} \frac{1}{t}\int_0^t B_{ii}(s)ds$. Let L be the Lyapunov transformation defined by $L = \mathrm{diag}(\eta^{i-1}, i = 1,\dots,n)$ for $\eta \ge \eta_0 > 0$. Then $L^{-1}B(t)L - L^{-1}\dot{L} = \mathrm{diag}(B(t)) + E(t)$, where $E(t)$ is the strictly upper triangular function of entries $E_{ij}(t) = \eta^{j-i}B_{ij}(t)$, for $i = 1,\dots,n-1$ and $j = i+1,\dots,n$. Since L is Lyapunov, and stability of the exponents is preserved

under Lyapunov transformations, then, for $\epsilon > 0$ as given in the statement of the theorem, there exists $\delta = \delta(\epsilon)$ such that $\sup_t |E(t)| < \delta$ implies $|\lambda_i^s - \lambda_i'| < \epsilon$, where $\{\lambda_i'\}_{i=1}^n$ denote the Lyapunov exponents of $\dot{D} = \operatorname{diag}(B)D$.

We claim that there exists a permutation π such that $\lambda_i' = \lambda_{\pi(i)}(D)$. Let Π denote a permutation matrix such that $\hat{D} = \Pi D \Pi^T$ defines an ordering such that $\chi(\hat{D}_{11}) \geq \chi(\hat{D}_{22}) \geq \cdots \geq \chi(\hat{D}_{nn})$. Notice that $\hat{D}$ satisfies $\dot{\hat{D}} = \hat{B}_D \hat{D}$, where $\hat{B}_D = \Pi \operatorname{diag}(B) \Pi^T$ and $\chi(\hat{D}_{ii}) = \chi(\hat{D}e_i) = \chi(\Pi \operatorname{diag}(B)\Pi^T e_i) = \chi(\operatorname{diag}(B)e_{\pi(i)})$. To complete the claim, we need to show that the diagonal fundamental matrix solution $\hat{D}$ is normal. By the Lyapunov construction of a normal basis, there exists a unit upper triangular matrix C such that $\hat{D} \cdot C$ is normal, but since setting $C = I$ minimizes the sum of the characteristic exponents of the columns, we have that $\hat{D}$ is normal. ∎

Remark 5.1. The Lyapunov exponents of $\dot{D} = \operatorname{diag}(B)D$ are not necessarily stable. The ordering of the Lyapunov exponents is not necessarily preserved; hence the need for the permutation π.

6. Sacker–Sell spectrum and Lyapunov spectral intervals. In this section we state and prove results relating the Sacker–Sell spectrum, Σ_{ED}, the Lyapunov spectrum, Σ_{L}, and the computed Lyapunov spectrum, Σ_{CL}. The following lemma shows that if a system has exponential dichotomy, then the principal matrix solution and an orthogonal projection may be assumed.

LEMMA 6.1. *Suppose the linear system* (2.3) *admits an exponential dichotomy for some fundamental matrix solution. Then it also admits an exponential dichotomy for the principal matrix solution. Moreover, the projection P may be taken to be an orthogonal matrix.*

Proof. Assume that (2.3) admits an exponential dichotomy for a fundamental matrix solution $X(t) \equiv X(t; X_0)$ with $X(0) = X_0$. Then $X(t; X_0) = X(t, I)X_0$ and

$$\begin{aligned} X(t;X_0)PX^{-1}(s;X_0) &= X(t;I)(X_0PX_0^{-1})X^{-1}(s;I), \\ X(t;X_0)(I-P)X^{-1}(s;X_0) &= X(t;I)(X_0(I-P)X_0^{-1})X^{-1}(s;I). \end{aligned} \tag{6.1}$$

Let $\tilde{P} = X_0 P X_0^{-1}$ and observe that $\tilde{P}^2 = \tilde{P}$, so $\tilde{P}$ is a projection and hence we have that the principal matrix solution admits an exponential dichotomy.

Let $S = \operatorname{range}(\tilde{P})$ and let V denote an orthonormal basis for S so that $P_1 = VV^T$ is the unique orthogonal projection onto S. From [9, pp. 16–17], it follows that the principal matrix solution admits an exponential dichotomy with orthogonal projection P_1. ∎

The following is essentially in [30], but we give a different proof.

THEOREM 6.2. *The computed Lyapunov spectrum is contained within the Sacker–Sell spectrum.*

Proof. Consider $\dot{X} = A(t)X$ with principal matrix solution X and the shifted system $\dot{X}_\lambda = [A(t) - \lambda I]X_\lambda$ with fundamental matrix solution X_λ. Fix λ such that X_λ has exponential dichotomy. Then there exists a projection P, constants $\alpha, \beta > 0$ and $K, L \geq 1$ such that

$$\begin{aligned} \|X_\lambda(t)PX_\lambda^{-1}(s)\| &\leq Ke^{-\alpha(t-s)}, \qquad t \geq s, \\ \|X_\lambda(t)(I-P)X_\lambda^{-1}(s)\| &\leq Le^{\beta(t-s)}, \qquad t \leq s. \end{aligned} \tag{6.2}$$

By Lemma 6.1, the projection P can be chosen orthogonal and there exists an orthogonal matrix U such that $U^T P U = P_1$, where P_1 is a diagonal matrix with entries

either 0 or 1. Thus,

$$\begin{array}{rcll} \|X_\lambda(t)UP_1U^TX_\lambda^{-1}(s)\| & \leq & Ke^{-\alpha(t-s)}, & t\geq s,\\ \|X_\lambda(t)U(I-P_1)U^TX_\lambda^{-1}(s)\| & \leq & Le^{\beta(t-s)}, & t\leq s, \end{array} \tag{6.3}$$

or equivalently,

$$\begin{array}{rcll} e^{-\lambda(t-s)}\|W(t)P_1W^{-1}(s)\| & \leq & Ke^{-\alpha(t-s)}, & t\geq s,\\ e^{-\lambda(t-s)}\|W(t)(I-P_1)W^{-1}(s)\| & \leq & Le^{\beta(t-s)}, & t\leq s, \end{array} \tag{6.4}$$

where $W(t)=X(t)U$ satisfies $\dot W = A(t)W$. Let Π denote a column permutation such that $Z=W\Pi$ implies $\chi(Z_1)\geq\cdots\geq\chi(Z_n)$, where Z_i denotes the ith column of Z. Decompose Z as $Z(t)=Q(t)R(t)$, where $Q(0)=Z(0)=U\Pi$ and $R(0)=I$, and notice that $\chi(R_1)\geq\cdots\geq\chi(R_n)$. For this ordering of growth rates of the columns of R the Lyapunov construction of a normal basis (see Theorem 2.3) takes the form $R(t)\,C$, where C is a unit upper triangular matrix so the Lyapunov construction does not change the diagonal elements of R.

In terms of R, the exponential dichotomy relationship for the shifted system is

$$\begin{array}{rcll} \|R_\lambda(t)P_2R_\lambda^{-1}(s)\| & \leq & Ke^{-\alpha(t-s)}, & t\geq s,\\ \|R_\lambda(t)(I-P_2)R_\lambda^{-1}(s)\| & \leq & Le^{\beta(t-s)}, & t\leq s, \end{array} \tag{6.5}$$

where $P_2=\Pi^TP_1\Pi$. Recall that the computed Lyapunov spectrum is defined as $\bigcup_j[\lambda^i_{jj},\lambda^s_{jj}]$, where

$$\lambda^s_{jj}=\limsup_{t\to\infty}\frac{1}{t}\ln(R_{jj}(t)) \quad\text{and}\quad \lambda^i_{jj}=\liminf_{t\to\infty}\frac{1}{t}\ln(R_{jj}(t)). \tag{6.6}$$

Assume that rank of P, hence of P_1 and P_2, is m. In (6.5) set $s=0$ so we have $|R_\lambda(t)P_2|\leq Ke^{-\alpha t}$. Since P_2 is a permutation matrix (plus rows and columns of 0's), $R_\lambda(t)P_2$ is a matrix containing m columns of $R_\lambda(t)$ and $n-m$ zero columns. Thus, there must be m columns of R for which $\lambda^s_{jj}-\lambda=\chi(R_{jj})-\lambda\leq\chi(R_{\bullet,j})-\lambda\leq-\alpha$, while for $n-m$ rows of R^{-1} we have $-\lambda^i_{kk}+\lambda=\chi(R^{-1}_{kk})+\lambda\leq\chi(R^{-1}_{k,\bullet})+\lambda\leq-\beta$. Thus, for m indices j we have $\lambda^i_{jj}\leq\lambda^s_{jj}\leq\lambda-\alpha<\lambda$ and for $n-m$ indices k we have $\lambda<\lambda+\beta\leq\lambda^i_{kk}\leq\lambda^s_{kk}$. Hence, $\lambda\notin\bigcup_j[\lambda^i_{jj},\lambda^s_{jj}]$. ∎

THEOREM 6.3. *Assume that for a linear homogeneous n-dimensional system the Sacker–Sell spectrum is given by n disjoint intervals. Then there exists a fundamental matrix solution with integrally separated columns.*

Proof. Write the Sacker–Sell spectrum as $\bigcup_{i=1}^n[a_i,b_i]$, and for $i=1,\ldots,n-1$ choose $\lambda_i=(a_{i+1}+b_i)/2$. Obviously $\lambda_i\notin\Sigma_{\rm ED}$, and there exists a fundamental matrix solution X_{λ_i} that has exponential dichotomy. Using the argument from Theorem 6.2, there exists a projection P_i of the form $P_i=\left(\begin{smallmatrix}0&0\\0&I\end{smallmatrix}\right)$ and $K_i,L_i,\alpha_i,\beta_i>0$ such that

$$\begin{array}{rcl} K_ie^{-\alpha_i(t-s)} & \geq & ||X_{\lambda_i}(t)PX^{-1}_{\lambda_i}(s)||\geq\dfrac{|X_{\lambda_i}(t)PX^{-1}_{\lambda_i}(s)X_{\lambda_i}(s)Pc|}{|X_{\lambda_i}(s)Pc|}\\ & = & \dfrac{\|X_{\lambda_i}(t)Pc\|}{\|X_{\lambda_i}(s)Pc\|}=\dfrac{\|X_j(t)\|}{\|X_j(s)\|}\cdot e^{-\lambda_i(t-s)} \end{array} \tag{6.7}$$

for $t\geq s$, and $c=e_j$, $j=i+1,\ldots,n$, and

$$\begin{array}{rcl} L_ie^{\beta_i(t-s)} & \geq & ||X_{\lambda_i}(t)(I-P)X^{-1}_{\lambda_i}(s)||\geq\dfrac{\|X_{\lambda_i}(t)(I-P)X^{-1}_{\lambda_i}(s)X_{\lambda_i}(s)(I-P)c\|}{\|X_{\lambda_i}(s)(I-P)c\|}\\ & = & \dfrac{\|X_{\lambda_i}(t)(I-P)c\|}{\|X_{\lambda_i}(s)(I-P)c\|}=\dfrac{\|X_j(t)\|}{\|X_j(s)\|}\cdot e^{-\lambda_i(t-s)} \end{array} \tag{6.8}$$

for $t \le s$, and $c = e_j$, $j = 1, \ldots, i$. Then

$$\frac{\|X_i(t)\|}{\|X_i(s)\|} \cdot \frac{\|X_{i+1}(s)\|}{\|X_{i+1}(t)\|} \ge \frac{1}{L_i} e^{\beta_i (t-s)} \cdot \frac{1}{K_i} e^{\alpha_i (t-s)} = \frac{1}{L_i K_i} e^{(\alpha_i + \beta_i)(t-s)}. \tag{6.9}$$

Repeating for all $i = 1, \ldots, n-1$, and taking $a = \min_i\{\alpha_i + \beta_i\}$ and $d = \min_i\{\frac{1}{L_i K_i}\}$, completes the proof. □

Example 6.1. As a counterexample to a converse of Theorem 6.3, consider the diagonal system with $\dot{x}_1 = (\cos(\ln t) + \sin(\ln t))x_1$ and $\dot{x}_2 = (-1 + \cos(\ln t) + \sin(\ln t))x_2$ so that $\Sigma_{\mathrm{CL}} = \Sigma_{\mathrm{L}} = [-1, 1] \cup [0, 2]$. Then, because of Theorem 6.2, the Sacker–Sell intervals overlap, but

$$\frac{|x_1(t)|}{|x_1(s)|} \cdot \frac{|x_2(s)|}{|x_2(t)|} = e^{t-s}, \qquad t \ge s, \tag{6.10}$$

so that x_1 and x_2 are integrally separated. □

Even in the case of stable Lyapunov exponents, in general, the Lyapunov and computed Lyapunov spectra are contained in the Sacker–Sell spectrum (see [30]). The following example modeled after one of Perron (see [1, Ex. 4.4.1]) clarifies this fact and it will be important in order to understand how we may approximate Σ_{ED}.

Example 6.2. Consider the linear differential equation $\dot{x} = c(t)x$, $c(t) = \sin(\ln(t)) + \cos(\ln(t))$, for $t \ge t_0 > 0$. The exact solution is $x(t) = \exp(t \sin(\ln(t)))\kappa_0$, $\kappa_0 = x(t_0)\exp(-t_0 \sin(\ln(t_0)))$, and it is easily seen that the Lyapunov and computed Lyapunov spectra coincide and are given by the interval $[-1, +1]$. Since the problem is scalar, this Lyapunov spectrum is necessarily stable.

We will show that $[-1, 1] \subset \Sigma_{\mathrm{ED}}$, that is, that there are values of $\lambda > 1$, and $\lambda < -1$, for which the shifted system does not have exponential dichotomy. Consider $\lambda > 1$; the case $\lambda < -1$ is similar. Then, to have exponential dichotomy in the shifted system means that there exist constants $\alpha > 0$ and $K \ge 1$ such that

$$e^{-\lambda(t-s)} e^{\int_s^t c(r)dr} = x_\lambda(t) x_\lambda^{-1}(s) \le K e^{-\alpha(t-s)}, \qquad t \ge s \ge t_0,\ \lambda > 1\,. \tag{6.11}$$

We rewrite this in the equivalent form

$$\frac{e^{\lambda t}}{e^{\lambda s}} \, \frac{e^{\int_{t_0}^s c\,d\tau}}{e^{\int_{t_0}^t c\,d\tau}} \ge \frac{1}{K} e^{\alpha(t-s)} \tag{6.12}$$

and consider the diagonal system

$$\dot{X} = \begin{pmatrix} \lambda & 0 \\ 0 & c(t) \end{pmatrix} X\,. \tag{6.13}$$

Thus, to have exponential dichotomy is the same as asking that the principal matrix solution of this system be integrally separated with constants $\frac{1}{K} < 1$ and $\alpha > 0$. This is equivalent to the requirement that

$$\int_s^t (\lambda - c(\tau))d\tau \ge a(t-s) - d, \qquad t \ge s,\ a > 0\,,\ \ d \ge 0\,, \tag{6.14}$$

which, in general, is not true. Let $\lambda_M = \frac{e^{\pi/2} + e^{-\pi/2}}{e^{\pi/2} - e^{-\pi/2}} = \coth(\pi/2)$. If the functions λ and c were integrally separated, then we should have

$$\int_s^t (\lambda - \sin(\ln(\tau)) - \cos(\ln(\tau)))d\tau \ge a(t-s) - d\,,$$

or

$$\lambda(t-s) - (t\sin(\ln(t)) - s\sin(\ln(s))) \geq a(t-s) - d\,.$$

Now, consider the following sequences for t and s:

$$t_k = \exp(2k\pi + \pi/2), \qquad s_k = \exp(2k\pi - \pi/2). \tag{6.15}$$

Then, along these sequences we would need to have

$$\lambda e^{2k\pi}(e^{\pi/2} - e^{-\pi/2}) - e^{2k\pi}(e^{\pi/2} + e^{-\pi/2}) \geq a e^{2k\pi}(e^{\pi/2} - e^{-\pi/2}) - d$$

or

$$a(e^{\pi/2} - e^{-\pi/2}) \leq \lambda(e^{\pi/2} - e^{-\pi/2}) - (e^{\pi/2} + e^{-\pi/2}) + de^{-2k\pi}.$$

Thus, for $1 < \lambda < \lambda_M$ and k sufficiently large, $\lambda(e^{\pi/2} - e^{-\pi/2}) - (e^{\pi/2} + e^{-\pi/2}) + de^{-2k\pi} < 0$, and so no positive a exists and the system cannot have exponential dichotomy for $1 < \lambda < \lambda_M$, where $\lambda_M \geq 1.09$. A similar argument for $\lambda < -1$ leads us to consider the diagonal system

$$\dot{X} = \begin{pmatrix} c(t) & 0 \\ 0 & \lambda \end{pmatrix} X\,, \tag{6.16}$$

so that having exponential dichotomy is equivalent to integral separation of the principal matrix solution of (6.16) or (which is the same) to

$$\int_s^t (c(\tau) - \lambda)d\tau \geq a(t-s) - d, \qquad t \geq s\ ,\ a > 0\,,\ \ d \geq 0\,. \tag{6.17}$$

Similarly to the above, we now obtain that we cannot have exponential dichotomy for $-1.09 \leq \lambda < -1$. Therefore, $[-1.09, 1.09] \subseteq \Sigma_{\mathrm{ED}}$. This argument can be easily improved by replacing $\pi/2$ in the definition of t_k and s_k in (6.15) with $\omega \approx 1.25$ (find ω to maximize $\coth(\omega)\cdot\sin(\omega)$ so that ω is the positive root of $\cos(\omega)\cdot\sinh(\omega) - 2\sin(\omega) = 0$). This shows that $[-1.1187, 1.1187] \subseteq \Sigma_{\mathrm{ED}}$.

For the sake of completeness, we point out that in [16] we actually prove that—for this example—$\Sigma_{\mathrm{ED}} = [-\sqrt{2}, \sqrt{2}]$. We will use this fact in Example 9.1. ∎

7. The SVD. To approximate Lyapunov exponents, an alternative to QR-based techniques is based on the SVD of a fundamental matrix solution. This approach has been used in [19, 20, 23]. Here we explore the feasibility of this approach, in particular, the role of integral separation in this case. So, we will assume that we have an integrally separated fundamental matrix solution X with ordered growth rates: $\chi(X_1) > \cdots > \chi(X_n)$.

Techniques based on the SVD need to assume that X admits a smooth SVD for all $t \geq t_0$: $X(t) = U(t)\Sigma(t)V^T(t)$, where $U^TU = I, V^TV = I, \Sigma = \mathrm{diag}(\sigma_i,\ i = 1, \ldots, n)$ and U, V, Σ are all C^p functions, $p \geq 1$. Unlike the QR factorization of X, the existence of such a smooth SVD is not obvious except in the case where the singular values stay distinct. Still, some results are known: (i) If X is analytic, then the factors U, V, Σ exist and are analytic (see [4]); (ii) if $X \in C^p, p \geq 1$, then there exist smooth $U,\ V,\ \Sigma$ as long as the singular values do not coalesce with too high a degree of contact (in general, U and V lose some degree of differentiability, while Σ stays C^p; see [11] for a precise statement); (iii) generically (i.e., for a generic one-parameter family of

nonsingular C^p functions X), then U, V, Σ are C^p, and in fact the σ_i singular values are distinct for all t (see [11]).

To make some progress, let us henceforth assume that a smooth (at least C^1) SVD of X exists. Let $G = U^T A U$ for all t. We notice that since $X = U\Sigma V^T$ for all t, and all factors are smooth, then we must also have

$$\dot{X} = AU\Sigma V^T = \dot{U}\Sigma V^T + U\dot{\Sigma}V^T + U\Sigma\dot{V}^T\,,$$

so that by letting $H = U^T\dot{U}$ and $K = V^T\dot{V}$, and noticing that H and K must be skew-symmetric, one must have

$$\dot{\Sigma} = G\Sigma - H\Sigma + \Sigma K\,,$$

and so we must have

$$\dot{\sigma}_i \;=\; G_{ii}\sigma_i \to \sigma_i(t) = \sigma_i(s)e^{\int_s^t G_{ii}(\tau)d\tau}\,, \qquad i = 1, \ldots, n\,. \tag{7.1}$$

In [19, 20] under the assumption of distinct singular values, the authors derived differential equations for U and Σ, integrated these numerically, and then set[2]

$$\lambda_i = \limsup_{t\to\infty} \frac{1}{t}\ln(|\sigma_i(t)|)\,, \qquad i = 1, \ldots, n\,. \tag{7.2}$$

Under the assumption of distinct singular values, the differential equations describing the evolution of U, V, Σ have been derived many times before (e.g., see [32]) and are

$$\dot{U} = UH, \quad \dot{V}^T = -KV^T, \quad \dot{\Sigma} = D\Sigma, \tag{7.3}$$

where $D = \mathrm{diag}(G)$, $H^T = -H$, $K^T = -K$ and, for $i \neq j$,

$$H_{ij} = \frac{G_{ij}\sigma_j^2 + G_{ji}\sigma_i^2}{\sigma_j^2 - \sigma_i^2}, \qquad K_{ij} = \frac{(G_{ij} + G_{ji})\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}. \tag{7.4}$$

On the other hand, from the SVD of X the Lyapunov exponents may be obtained as

$$\chi(X_i) = \limsup_{t\to\infty} \frac{1}{t}\ln||\Sigma(t)V^T(t)e_i||\,. \tag{7.5}$$

Here, we explore the "equivalence" between (7.2) and (7.5) and at the same time validate the methods based upon differential equations for the U, V, and Σ factors. We will do this under the assumption that D, the diagonal of G, is integrally separated:

$$\int_s^t (G_{kk}(\tau) - G_{k+1,k+1}(\tau))d\tau \geq a(t-s) - d, \tag{7.6}$$

$$a > 0,\ d \in \mathbb{R},\ t \geq s,\ k = 1, 2, \ldots, n-1.$$

For some of the results here, we can assume also the following condition (simply (7.6) with $s = 0$) that is weaker and easier to verify than (7.6):

[2] In fact, in [19, 20], it was assumed that the λ_i's existed as limits.

$$\int_0^t (G_{k,k}(\tau) - G_{k+1,k+1}(\tau))d\tau \geq at - d, \tag{7.7}$$

$$a > 0,\ d \in \mathbb{R},\ t \geq 0,\ k = 1, \ldots, n-1.$$

LEMMA 7.1. *For all t, let $X = U\Sigma V^T$ be a C^p SVD of X, $p \geq 1$. Let $G = U^T AU$ satisfy* (7.7). *Then, for t sufficiently large, we eventually have*

$$\sigma_k(t) \ > \ \sigma_{k+1}(t), \qquad k = 1, \ldots, n-1. \tag{7.8}$$

Proof. Take $k = 1, \ldots, n-1$. From (7.1) and (7.7), we have

$$\begin{aligned}\sigma_k(t) &= \frac{\sigma_k(0)}{\sigma_{k+1}(0)}\ \sigma_{k+1}(0)e^{\int_0^t G_{kk}(\tau)d\tau} \\ &\geq \frac{\sigma_k(0)}{\sigma_{k+1}(0)}\ \sigma_{k+1}(0)e^{\int_0^t G_{k+1,k+1}(\tau)d\tau}\ e^{at}e^{-d}\,.\end{aligned}$$

That is,

$$\sigma_k(t) \geq \left[\frac{\sigma_k(0)}{\sigma_{k+1}(0)}e^{at}e^{-d}\right]\ \sigma_{k+1}(t), \qquad t \geq 0\,.$$

Now, let t_k be sufficiently large so that the term in brackets is greater than 1. Repeating the argument for all $k = 1, \ldots, n-1$ gives the result. ∎

Based upon Lemma 7.1, as long as (7.7) holds, we may as well assume that all singular values are distinct, and ordered, for all times $t \geq 0$. In particular, the differential equations (7.3) with H and K defined by (7.4) hold. Having done this, we now show the equivalence between (7.2) and (7.5).

THEOREM 7.2. *Under the assumption* (7.6), *we have* $\chi(X_i) = \limsup_{t\to\infty} \frac{1}{t}\ln(|\sigma_i(t)|)$.

Proof. Let $X = U\Sigma V^T$ be rewritten as $X = UP$, $P = \Sigma V^T$, so that

$$\dot{P} = (U^T AU - U^T\dot{U})P, \tag{7.9}$$

and since U is a Lyapunov transformation $\chi(P_i) = \chi(X_i)$. Consider also the system for Σ given in (7.3). We want to show that $\chi(\Sigma_{ii}) = \chi(P_i)$. If we rewrite the differential equations for P using the differential equations for Σ and V, then $\dot{P} = (D - \Sigma K\Sigma^{-1})P$.

Let $E = \Sigma K\Sigma^{-1}$, so for $i > j$, $E_{ij} = K_{ij}\frac{\sigma_i}{\sigma_j}$ and thus

$$E_{ij} = (G_{ij} + G_{ji})\frac{1}{\frac{\sigma_j^2}{\sigma_i^2} - 1}\,. \tag{7.10}$$

We have

$$\frac{\sigma_j^2}{\sigma_i^2} = \frac{\sigma_j^2(0)}{\sigma_i^2(0)}\exp\left(2\int_0^t (G_{jj}(\tau) - G_{ii}(\tau))\right)d\tau\,, \tag{7.11}$$

and since (7.6) holds, then $\frac{\sigma_j^2}{\sigma_i^2} \to \infty$ and thus $E_{ij} \to 0$ as $t \to \infty$ for $i > j$. Obviously, we also have $E_{ii} = 0$ for all i. Finally, for $j < i$, $E_{ji} = K_{ji}\frac{\sigma_j}{\sigma_i} = -K_{ij}\frac{\sigma_j}{\sigma_i}$. But

$$K_{ij}\frac{\sigma_j}{\sigma_i} = \frac{(G_{ij} + G_{ji})}{1 - \frac{\sigma_i^2}{\sigma_j^2}},$$

and now $1 - \frac{\sigma_i^2}{\sigma_j^2}$ does not go to ∞, and hence in general E_{ji} does not approach 0 as t approaches infinity. So, we write

$$E = \text{low}(E) + \text{upp}(E),$$

where $\text{upp}(E)$ is the strictly upper triangular part of E and $\text{low}(E)$ is the strictly lower triangular part of E, and consider the system

$$\dot{\overline{P}} = (D - \text{upp}(E))\overline{P}\,. \tag{7.12}$$

Since $\text{low}(E) \to 0$ as $t \to \infty$, and the exponents of the P system (7.9) are stable, then the Lyapunov exponents of the P system and of the $\overline{P}$ system (7.12) are the same by Theorem 3.2. In other words, $\chi(P_i) = \chi(\overline{P}_i)$ for $i = 1, \ldots, n$. Finally, with assumption (7.6) we can apply Theorem 5.1 to obtain

$$\chi(\overline{P}_i) = \chi(\overline{P}_{ii}) = \chi(\Sigma_{ii}), \qquad i = 1, \ldots, n. \tag{7.13}$$ □

Remark 7.1. From the proof of Theorem 7.2, it is apparent that for the Lyapunov exponents of the systems (7.9) and (7.12) to coincide it suffices to assume (7.7). However, the stronger condition (7.6) was needed to prove $\chi(\overline{P}_i) = \chi(\overline{P}_{ii})$.

When (7.7) holds, and a fortiori when (7.6) holds, we have the following result.

LEMMA 7.3. *Let* (7.7) *hold. Then, the orthogonal matrix function* $V(t) \to \overline{V}$*, as* $t \to \infty$*, where* $\overline{V}$ *is a constant orthogonal matrix.*

Proof. Recall that V satisfies $\dot{V}^T = -KV^T$, where K is defined in (7.4), $K_{ij} = (G_{ij} + G_{ji})\frac{\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}$ for $i \neq j$, and $K_{ii} = 0$ for all i. We claim that, under assumption (7.7), $K_{ij} \to 0$ exponentially fast as $t \to \infty$. For $i > j$, we have

$$\begin{aligned} K_{ij} &= (G_{ij} + G_{ji})\frac{\sigma_i(0)}{\sigma_j(0)}\frac{\exp(\int_0^t (G_{jj}(\tau) - G_{ii}(\tau))d\tau)}{\exp(2\int_0^t (G_{jj}(\tau) - G_{ii}(\tau))d\tau) - \frac{\sigma_i^2(0)}{\sigma_j^2(0)}} \\ &= (G_{ij} + G_{ji})\frac{\sigma_i(0)}{\sigma_j(0)}\left[\frac{1}{\exp(\int_0^t (G_{jj}(\tau) - G_{ii}(\tau))d\tau) + \frac{\sigma_i(0)}{\sigma_j(0)}}\right. \\ &\qquad \left. + \frac{\sigma_i(0)}{\sigma_j(0)}\frac{1}{\exp(2\int_0^t (G_{jj}(\tau) - G_{ii}(\tau))d\tau) - \frac{\sigma_i^2(0)}{\sigma_j^2(0)}}\right], \end{aligned} \tag{7.14}$$

and so by (7.7) we have that for $i > j$, $K_{ij} \to 0$ exponentially fast, as $t \to \infty$. By skew-symmetry, the same holds true also for $i < j$. The result now follows from [10, Thm. 2, p. 90].[3] □

Remark 7.2. Lemma 7.3 may be used indirectly to determine if (7.7) holds, a fact which was apparently used in [19].

The condition (7.6) is very similar to the condition (3.3) on the diagonal of the upper triangular coefficient matrix B obtained when finding the QR factorization of a

[3]Theorem 2 of [10] is concerned with the system $\dot{X} = (A + B(t))X$ when A is constant with simple eigenvalues λ_i and associated eigenvectors ξ_i, $i = 1, \ldots, n$, and B is continuous such that $\int_{t_0}^{\infty} \|B(t)\|dt < \infty$. In such a case, the cited theorem states that X converges to $\text{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t})[\xi_1, \ldots, \xi_n]C$, where C is a constant invertible matrix. However, the result and the proof in [10] hold true by just requiring that A is diagonalizable (not necessarily with distinct eigenvalues). We have used this fact with $A = 0$.

fundamental matrix solution. In fact, these two conditions lead to similar outcomes. The following is the QR analogue of Lemma 7.3.

LEMMA 7.4. *Consider the upper triangular system $\dot{R} = BR$, where B is bounded and continuous, and assume that the diagonal of B is integrally separated, as in* (3.5). *Then, $R \to \mathrm{diag}(R)\overline{Z}$ as $t \to \infty$, where $\overline{Z}$ is a constant upper triangular matrix with* 1*'s on the diagonal.*

Proof. Write $R = DZ$, where $Z = D^{-1}R$, and $D = \mathrm{diag}(R)$. Then D satisfies $\dot{D} = \mathrm{diag}(B)D$ and Z satisfies $\dot{Z} = EZ$, where $E = D^{-1}(B - \mathrm{diag}(B))D$. Then, $E_{ij} = B_{ij} \cdot \frac{R_{jj}}{R_{ii}}$ for $i < j$ and $E_{ij} = 0$ for $i \geq j$. Now,

$$\frac{R_{jj}}{R_{ii}} = \frac{R_{jj}}{R_{ii}}(0)\ e^{\int_0^t (B_{jj} - B_{ii})d\tau}\,.$$

Let $j = i + k$ for some $k = 1, \ldots$. The diagonal of B is integrally separated (see (3.5)), and so

$$\int_0^t (B_{jj} - B_{ii})d\tau \ \leq\ -k(at - d)\,,$$

from which $E_{ij} \to 0$ exponentially fast as $t \to \infty$, and the result follows. ∎

Remark 7.3. We notice that the assumption of integral separation (7.7) (or (7.6)) does not preclude singular values from coinciding at some (early) time t, in which case the computation of the factors U, V, Σ remains by and large unexplored territory. Indeed, if one chooses to use the SVD technique for approximating the Lyapunov exponents, even if (7.7) (or (7.6)) is satisfied, it is probably advisable to integrate for X for awhile prior to writing down and integrating the differential equations for the factors U, Σ, and V.

8. Numerical techniques. We only outline the continuous QR technique (see [8, 12, 13, 14, 21]), which is the one we used for the experiments in the next section.

Consider the linear homogeneous problem

$$\dot{x}(t) = A(t)x(t). \tag{8.1}$$

The key task is to find Q which transforms the upper left $p \times p$ ($p \leq n$) corner of A, $B = Q^T AQ - Q^T\dot{Q}$, to upper triangular form. From B, one can then approximate p Lyapunov exponents using (2.9) if the system is regular or the spectral intervals in the case where the system is not regular.

To find Q, one writes p columns of a fundamental matrix solution of (8.1) as $X = QR$, where Q is an $n \times p$ orthonormal function (i.e., for all $t \geq 0$: $Q^T(t)Q(t) = I_p$), and R is a $p \times p$ upper triangular function with positive entries on the diagonal. Upon differentiating $X = QR$, we have

$$AQR = \dot{X} = \dot{Q}R + Q\dot{R} \quad \text{or} \quad AQ = \dot{Q} + Q\dot{R}R^{-1}. \tag{8.2}$$

Let B denote the upper triangular function $\dot{R}R^{-1}$ and set $S(Q) = Q^T\dot{Q}$, which is skew symmetric. Then

$$Q^T AQ = S(Q) + B\,, \tag{8.3}$$

and since $S(Q)$ is skew symmetric and B is upper triangular, we have

$$S(Q)_{ij} = \begin{cases} (Q^T AQ)_{ij}, & i > j, \\ 0, & i = j, \\ -(Q^T AQ)_{ji}, & i < j. \end{cases} \tag{8.4}$$

Then, from (8.2), the equation for Q is

$$\dot{Q} = AQ - QB = AQ - Q(Q^T AQ - S(Q)) = (I - QQ^T)AQ + QS(Q). \tag{8.5}$$

Initial conditions for Q are obtained from a QR factorization of the initial conditions for X (and the most typical choice is to take $X(0) = \binom{I_p}{0}$).

To repeat this reasoning relative to a trajectory of a nonlinear problem, one must integrate

$$\dot{x} \;=\; f(x), \qquad x(0) = x_0\,, \tag{8.6}$$

and then use $A(t) = Df(x(t))$ in (8.2).

Once we have the triangular function B, we can compute the p Lyapunov exponents from (2.9) if the system is regular. Alternatively, one may (in principle) compute λ^s_{jj} (and λ^i_{jj}) in Σ_{CL}, $j = 1, \ldots, p$, from

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t B_{jj}(s)ds \quad \text{and} \quad \liminf_{t\to\infty} \frac{1}{t}\int_0^t B_{jj}(s)ds\ ,\ \ j = 1, \ldots, p\ . \tag{8.7}$$

Regardless of whether the B-system is regular, we found it convenient to work with the variables $\nu_j(t) = \int_0^t B_{jj}(s)ds$ so that we end up with the differential equations

$$\dot{\nu}_j = B_{jj}, \quad \nu_j(0) = 0, \quad j = 1, \ldots, p, \tag{8.8}$$

from which the exponents may be approximated as limits (or lim sups and lim infs) of

$$\frac{1}{t}\nu_j(t), \qquad j = 1, \ldots, p\,.$$

At this point, the skeleton of the method is clear: for nonlinear problems, integrate (8.6), (8.5), and (8.8); for linear problems, integrate just (8.5) and (8.8).

8.1. Numerical implementation. When approximating (8.5) numerically it is important to maintain Q orthonormal. Several choices are possible to achieve this; e.g., in [13, 3] techniques are discussed to directly integrate (8.5), whereas in [8, 21] a continuous Gram–Schmidt procedure is proposed. We have used the technique described in [14, 15]. The idea of this technique is to locally decompose Q in a way analogous to the numerical linear algebra context using elementary Givens rotations or Householder reflectors. Integration for these elementary factors can be done adaptively, and we refer to [15] for details.

So, in the end, the differential equations (8.6), (8.5), and (8.8) are all integrated with adaptive time stepping, controlled by the tolerance values `TOLX, TOLQ, TOLL`, respectively. The basic integrator in all cases is our implementation of the Dormand–Prince 4/5 embedded Runge–Kutta pair modeled after the pattern adopted in [15], to which we again refer for details.

8.2. Testing integral separation. It is clearly desirable to infer whether the system has integral separation. This is needed to gain some confidence in the answers one obtains (since it implies stable exponents), and also (see below) to obtain a computational procedure to approximate Σ_{ED}. We have used the following construction which is motivated by Steklov function considerations. Recall that, given a continuous

bounded function f, the Steklov function or Steklov average of f with step $H > 0$ is defined as (see [1, Def. 5.4.1] and [5])

$$f^H(t) = \frac{1}{H}\int_t^{t+H} f(\tau)d\tau. \tag{8.9}$$

Now, consider two bounded functions f_1 and f_2 (presently, think of them as diagonal elements of the upper triangular coefficient matrix B), and suppose we want to know if they are integrally separated:

$$\int_s^t (f_1(\tau) - f_2(\tau))d\tau \geq a(t-s) - d, \qquad a > 0,\, d \in \mathbb{R},\, t \geq s. \tag{8.10}$$

The importance of Steklov functions resides in the fact that (8.10) can be inferred from the Steklov average of the difference $f_1 - f_2$. This is the content of Lemma 5.4.1 in [1].

LEMMA 8.1. *Let f_1 and f_2 be two bounded functions. Then, f_1 and f_2 are integrally separated, i.e.,* (8.10) *holds if and only if for sufficiently large H their Steklov functions are separated, i.e.,*

$$f_1^H(t) - f_2^H(t) = \frac{1}{H}\int_t^{t+H} (f_1(\tau) - f_2(\tau))d\tau \geq a > 0, \qquad t \geq 0\,. \tag{8.11}$$

In practice, to check (8.11) will require a careful choice of H. We refer to Examples 9.1 and 9.2 for practical considerations.

8.3. Numerical computation of spectral intervals. In the case in which the system is not regular, it would be clearly desirable to approximate Σ_{CL} and/or Σ_{L}. Furthermore, it is clearly of interest to be able to approximate Σ_{ED} in the case in which the system does not have *point spectrum* (i.e., the Sacker–Sell intervals reduce to single points, the Lyapunov exponents of the system). As far as we know, the computational task of approximating spectral intervals has not been previously undertaken. This is most likely because in many problems the Lyapunov exponents appear to exist as limits (see Remark 3.1) and also because the numerical approximation of spectral intervals is an even more delicate task than approximation of Lyapunov exponents of regular systems. Naturally, this is due to the asymptotic nature of the quantities being computed. Further, for Σ_{CL}, there is the added difficulty that lim sups and lim infs must be approximated, which is more complicated than approximating limits. And, for Σ_{ED}, it is the uniformity (i.e., for all $t \geq s$) in the definition of exponential dichotomy which causes additional difficulties. These difficulties notwithstanding, below we present the strategies we have adopted to approximate spectral intervals. In what follows, we will restrict our attention to triangular systems: $\dot{R} = B(t)R$, where B is an upper triangular continuous and bounded function. As previously remarked, this restriction is no loss of generality. In order to further validate the results of the procedures to approximate the spectral intervals, we need to restrict our attention to triangular functions B whose diagonal is integrally separated.

8.3.1. Approximating Σ_{CL}. To approximate the lim inf and lim sup in the definition of Σ_{CL}, we reason as follows. Let $b(t)$ be a given diagonal element of the upper triangular transformed coefficient matrix B. Let $\lambda(t) = \frac{1}{t}\int_0^t b(s)ds$. Recall that

$$\lambda^+ = \lim_{\tau\to\infty} \sup_{t\geq\tau} \lambda(t) \quad \text{and} \quad \lambda^- = \lim_{\tau\to\infty} \inf_{t\geq\tau} \lambda(t)\,.$$

So, if we let

$$g(\tau) = \sup_{t\geq\tau} \frac{1}{t}\int_0^t b(s)ds \quad \text{and} \quad h(\tau) = \inf_{t\geq\tau} \frac{1}{t}\int_0^t b(s)ds\ ,$$

then for every $\epsilon > 0$ there exists $\tau(\epsilon)$ such that $\tau \geq \tau(\epsilon)$ implies $|g(\tau) - \lambda^+| < \epsilon$ and $|h(\tau) - \lambda^-| < \epsilon$. In our experiments, we mimic this definition on a finite interval. For given $T > 0$, we specify a value τ_0, $T > \tau_0 > 0$, and compute

$$g_f(T,\tau_0) = \sup_{T\geq t\geq\tau_0} \frac{1}{t}\int_0^t b(s)ds \quad \text{and} \quad h_f(T,\tau_0) = \inf_{T\geq t\geq\tau_0} \frac{1}{t}\int_0^t b(s)ds\,, \tag{8.12}$$

which will provide approximations to λ^+ and λ^-.

8.3.2. Approximating Σ_{ED}. Our approach to approximation of Σ_{ED} is motivated by the relationship between exponential dichotomy and integral separation as was seen in Example 6.2. We develop a procedure for approximating Σ_{ED} for diagonal systems, or for any system that is reducible to a diagonal system through a Lyapunov transformation. For example (see the proof of Theorem 5.1), our procedure applies to triangular systems whose diagonal is integrally separated.

So, consider $\dot{x} = D(t)x$, where $D = \mathrm{diag}(B_{jj}, j = 1,\dots,n)$ and where we may think of the B_{jj}'s as the diagonal entries of the upper triangular function B. For each $j = 1,\dots,n$, we consider the diagonal planar systems (cf. (6.13) and (6.16))

$$\dot{y}_j = \begin{pmatrix} \lambda & 0 \\ 0 & B_{jj}(t) \end{pmatrix} y_j \quad \text{and} \quad \dot{y}_j = \begin{pmatrix} B_{jj}(t) & 0 \\ 0 & \lambda \end{pmatrix} y_j. \tag{8.13}$$

Following the argument relating exponential dichotomy and integral separation in Example 6.2 (see (6.14) and (6.17)), we obtain the following result.

LEMMA 8.2. *Consider the diagonal system $\dot{x} = D(t)x$, $D = \mathrm{diag}(B_{jj}, j = 1,\dots,n)$. Then, for each $j = 1,\dots,n$, the Sacker–Sell spectrum corresponding to the jth diagonal element is given by the interval*

$$\Lambda_j = \{\lambda \in \mathbb{R} : (8.13) \textit{ are not integrally separated}\}\,. \tag{8.14}$$

As a consequence of Lemma 8.2, we have (cf. [16]) the following theorem.

THEOREM 8.3. *The Sacker–Sell spectrum of the diagonal system $\dot{x} = D(t)x$, $D = \mathrm{diag}(B_{jj}, j = 1,\dots,n)$, is given by*

$$\Sigma_{\mathrm{ED}} = \bigcup_{j=1}^{n} \Lambda_j\,, \tag{8.15}$$

where Λ_j is defined in (8.14)*, $j = 1,\dots,n$.*

To obtain a computational procedure for Σ_{ED} out of Theorem 8.3, we rely on Steklov functions. Indeed (recall Lemma 8.1), the systems in (8.13) are integrally separated if and only if for $H > 0$ sufficiently large the Steklov differences of λ and B_{jj}, respectively, B_{jj} and λ, are positive for all t.

Now, given any $H > 0$, for $j = 1,\dots,n$, consider

$$\alpha_j^H = \inf_t \frac{1}{H}\int_t^{t+H} B_{jj}(s)ds \quad \text{and} \quad \beta_j^H = \sup_t \frac{1}{H}\int_t^{t+H} B_{jj}(s)ds. \tag{8.16}$$

We will use $[\alpha_j^H, \beta_j^H]$ to approximate the jth spectral interval of Σ_{ED}, $j = 1, \ldots, n$. The following result justifies our approach on an infinite time interval.

THEOREM 8.4. *Consider $\dot{x} = D(t)x$, where $D = \text{diag}(B_{jj}, j = 1, \ldots, n)$. For $j = 1, \ldots, n$, let α_j^H and β_j^H be given as in* (8.16). *Let $H > 0$ be given. Then, for each $j = 1, \ldots, n$, $\Lambda_j \subseteq [\alpha_j^H, \beta_j^H]$. Moreover, for $H > 0$ sufficiently large, $[\alpha_j^H, \beta_j^H] \subseteq \Lambda_j$ and hence $[\alpha_j^H, \beta_j^H] = \Lambda_j$, $j = 1, \ldots, n$.*

Proof. First, assume that $H > 0$ is arbitrary and that $\lambda > \beta_j^H$ for some $j = 1, \ldots, n$. Then, there exists $a_j > 0$ such that

$$\int_t^{t+H} (\lambda - B_{jj}(\tau))d\tau \geq a_j H \quad \forall t\,. \tag{8.17}$$

We want to show that λ and B_{jj} are integrally separated functions. That is, we need to show that for all t, s, $t \geq s$, there exists $a > 0$ and $d \in \mathbb{R}$ such that

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau \geq a(t-s) - d. \tag{8.18}$$

We will verify (8.18) with $a = a_j$ and $d = d_j := 2H(|\lambda| + \max_t |B_{jj}(t)|)$. Because of (8.17), (8.18) holds for all t and s with $t = s + H$. Consider the case of t, s, with $t < s + H$. Then, rewrite

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau = \int_s^{s+H} (\lambda - B_{jj}(\tau))d\tau - \int_t^{s+H} (\lambda - B_{jj}(\tau))d\tau\,,$$

and thus $\int_t^{s+H}(\lambda - B_{jj}(\tau))d\tau \leq (|\lambda| + \max_t |B_{jj}(t)|)(s + H - t) \leq d_j$, so that

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau \geq a_j H - d_j \geq a_j(t-s) - d_j\,.$$

Next, let t, s, with $t > s + H$. Then, for some integer $k > 1$, we have $t = s + kH + \sigma$, $\sigma \in [0, H)$. Therefore,

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau = \sum_{j=0}^{k} \int_{s+jH}^{s+(j+1)H} (\lambda - B_{jj}(\tau))d\tau - \int_{s+kH+\sigma}^{s+(k+1)H} (\lambda - B_{jj}(\tau))d\tau\,,$$

and thus (using (8.17) and the previous argument used when $t < s + H$) we get

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau \geq a_j(k+1)H - d_j \geq a_j(t-s) - d_j\,,$$

and (8.18) follows. Therefore, λ and $B_{jj}(t)$ are integrally separated, and so $\lambda \notin \Lambda_j$. A similar proof for $\lambda < \alpha_j^H$ establishes that $\Lambda_j \subseteq [\alpha_j^H, \beta_j^H]$ for any given $H > 0$.

Assume now that $\lambda \notin \Lambda_j$. Then λ and $B_{jj}(t)$ and/or $B_{jj}(t)$ and λ are integrally separated. Suppose that λ and $B_{jj}(t)$ are integrally separated; the argument for $B_{jj}(t)$ and λ integrally separated is similar. Then there exists $a > 0$ and $d \in \mathbb{R}$ such that for all t, s, with $t \geq s$, we have

$$\int_s^t (\lambda - B_{jj}(\tau))d\tau \geq a(t-s) - d. \tag{8.19}$$

Choose $H > 0$ large enough so that $a - d/H > a/2$. Thus, for all t,

$$\frac{1}{H}\int_t^{t+H} (\lambda - B_{jj}(s))ds \geq a - d/H > a/2, \tag{8.20}$$

and so $\lambda > \beta_j^H$. A similar proof for $B_{jj}(t)$ and λ integrally separated implies that $\lambda < \alpha_j^H$, and thus $\lambda \notin [\alpha_j^H, \beta_j^H]$. Therefore, for $H > 0$ sufficiently large, $[\alpha_j^H, \beta_j^H] = \Lambda_j$. □

On a finite time interval, our computational procedure to approximate Σ_{ED} mimics Theorem 8.4. Given $H > 0$, and $T > t_0 > 0$, we let $b(t)$ be a diagonal element $B_{jj}(t)$, for some $j = 1, \ldots, n$, of the triangular coefficient matrix B, defined on the time interval $[0, T]$. We compute the Steklov averages of b with respect to the given H: $b_H(t) := \frac{1}{H}\int_t^{t+H} b(\tau)d\tau$, for $T - H \geq t \geq t_0$. Next, we compute

$$\overline{b}_H = \sup_{T-H\geq t\geq t_0} b_H(t) \quad \text{and} \quad \underline{b}_H = \inf_{T-H\geq t\geq t_0} b_H(t) \tag{8.21}$$

and use these as approximations to $[\alpha_j^H, \beta_j^H]$ in (8.16).

9. Examples and numerical results. We first consider a linear example for which the Lyapunov exponents do not exist as limits; in this case, we approximate the spectral intervals. Then, we approximate the Lyapunov exponents for two nonlinear systems, i.e., the exponents associated with linearization about computed trajectories; in both cases considered, the Lyapunov exponents appear to exist as limits. Thus, we attempt verifying integral separation of the diagonal of the transformed triangular problem in order to infer stability of the Lyapunov exponents: in one case we are successful, in another we are not.

In all examples below, integration for Q is carried out with `QRINT` (see [15]) using Jacobi rotations (the so-called θ-method in `QRINT`).

Example 9.1. Consider a planar linear problem $\dot{x} = A(t)x$ with continuous spectrum, where $A(t)$ is defined by

$$\begin{aligned}
A_{11}(t) &= (2\sin(\tau(t)) + \alpha)\cos^2(\theta(t)) + \cos(\tau(t)) - \sin(\tau(t)) - \alpha - \beta\cos(\theta(t))\sin(\theta(t)),\\
A_{12}(t) &= (2\sin(\tau(t)) + \alpha)\cos(\theta(t))\sin(\theta(t)) - \dot{\theta}(t) + \beta\cos^2(\theta(t)),\\
A_{21}(t) &= (2\sin(\tau(t)) + \alpha)\cos(\theta(t))\sin(\theta(t)) + \dot{\theta}(t) - \beta\sin^2(\theta(t)),\\
A_{22}(t) &= -(2\sin(\tau(t)) + \alpha)\cos^2(\theta(t)) + \cos(\tau(t)) + \sin(\tau(t)) + \beta\cos(\theta(t))\sin(\theta(t)),
\end{aligned}$$

and $\tau(t) = \ln(t + 1)$. This problem is designed so that the orthogonal change of variables Q and the upper triangular coefficient matrix function B are

$$Q(t) = \begin{pmatrix} \cos(\theta(t)) & -\sin(\theta(t)) \\ \sin(\theta(t)) & \cos(\theta(t)) \end{pmatrix} \quad \text{and} \quad B(t) = \begin{pmatrix} B_{11}(t) & \beta \\ 0 & B_{22}(t) \end{pmatrix},$$

where $B_{11}(t) = \cos(\tau(t)) + \sin(\tau(t))$ and $B_{22}(t) = \cos(\tau(t)) - \sin(\tau(t)) - \alpha$. It is not hard to explicitly obtain the spectral intervals (recall the result for Example 6.2): we have $\Sigma_{\mathrm{L}} = [-1, 1] \cup [-\alpha - 1, -\alpha + 1]$ and $\Sigma_{\mathrm{ED}} = [-\sqrt{2}, \sqrt{2}] \cup [-\alpha - \sqrt{2}, -\alpha + \sqrt{2}]$. For our experiments we choose $\beta = 1$, $\theta(t) = \omega t$, and $\alpha = 4$. Integration for Q was done with local error tolerance 10^{-5}.

In Table 9.1, we report on results of experiments to approximate Σ_{L}. We approximate all integrals in (8.12) with the composite trapezoidal rule on data sampled at

TABLE 9.1
Example 1. *Approximation of* Σ_{L}.

T	τ_0	$[\lambda_1^-, \lambda_1^+]$	$[\lambda_2^-, \lambda_2^+]$
1.E4	1.E2	[-1.0191, 1.0004]	[-4.9774, -2.9998]
1.E6	1.E2	[-1.0191,1.0004]	[-5.0002,-2.9998]
1.E6	1.E4	[-1,0.94871]	[-5.0002,-3]
1.E7	1.E4	[-1,1]	[-5.0002,-3]

TABLE 9.2
Example 1. *Approximation of* Σ_{ED} $(t_0 = 0)$.

T	H	$[\alpha_1^H, \beta_1^H]$	$[\alpha_2^H, \beta_2^H]$
1.E7	1.E4	$[-1.4063, 1.4142]$	$[-5.4142, -2.5861]$
1.E7	1.E5	$[-1.2576, 1.4127]$	$[-5.4141, -2.6191]$
1.E8	1.E4	$[-1.4142, 1.4142]$	$[-5.4142, -2.5858]$
1.E8	1.E5	$[-1.4142, 1.4127]$	$[-5.4141, -2.5858]$

integer times. In the table we specify the values T, τ_0, and report on the approximations (at 5 digits) for the two spectral intervals making up Σ_{L}. In spite of the crudeness of the quadrature rule, quite clearly Σ_{L} is approximated very well.

In Table 9.2 we report on calculations to approximate Σ_{ED}. In the table, we vary quantities in the procedure outlined in section 8.3; see (8.21). In particular, we vary the final time, T, and the length of the Steklov averages, H. The initial time for which the Steklov averages are maximized/minimized is fixed at $t_0 = 0$, and the approximations we obtain to $[\alpha_j^H, \beta_j^H]$ for $j = 1, 2$ are recorded to 5 digits. Our calculations are based upon data from the diagonal of B that we have sampled using a large step size of $h = 10$. The Steklov averages are approximated with the composite trapezoidal rule. The results point out the difficulty in finding an appropriate value for H: simultaneously, one would need H large enough so that the endpoints of the intervals in Σ_{ED} are approximated accurately, yet not so large with respect to the final time T that little data are sampled.

Example 9.2 (Lorenz equation). Our next example is the Lorenz equation

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} \sigma(y - x) \\ \rho x - xz - y \\ xy - \beta z \end{pmatrix}.$$

We consider the parameter values $\sigma = 16$, $\beta = 4.0$, and $\rho = 45.92$ and the initial condition $(x(0), y(0), z(0)) = (0, 1, 0)$. In Table 9.3, we summarize some results which are typical. Error control is done on the trajectory x, on Q, and on ν (see (8.8)). Apparently, the Lyapunov exponents exist as limits: the linearized problem appears to be regular.

Based upon the results in Table 9.3, we observe that (1) there is an obvious relation between the number of steps taken and the length of integration (recall that we are integrating with variable step size). This suggests that we are tracing "alike trajectories" on the Lorenz attractor; (2) with all the imperfections of finite precision arithmetic, the Lyapunov exponents are clearly converging towards $\lambda_1 \approx 1.5$, $\lambda_2 = 0$, and $\lambda_3 \approx -22.5$.

In order to infer stability of the exponents, we have verified if the linearized system enjoys integral separation. As far as we know, this is the first attempt of this type, on the Lorenz system or otherwise. We use the construction outlined in section 8.2 on the transformed, triangular problem. So, we have to check if the three functions

TABLE 9.3
Example 2. `TOLX=TOLQ=TOLL`$=1.E\text{-}6$.

t_{end}	`Steps`	λ_1	λ_2	λ_3
1.E2	8.6E3	1.415	3.E-2	-22.466
1.E3	8.6E4	1.4892	4.64E-3	-22.494
1.E4	8.6E5	1.499	4.64E-4	-22.499
1.E5	8.6E6	1.5027	4.07.E-5	-22.5027
1.E6	8.6E7	1.5024	7.6E-6	-22.5024

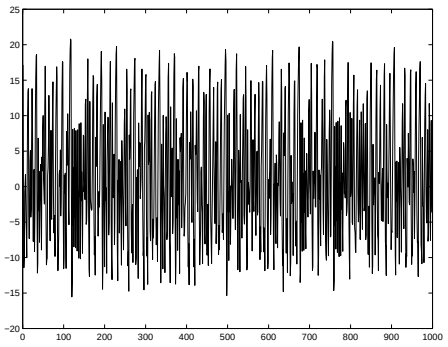


FIG. 9.1. $b_{11} - b_{22}$.

b_{11}, b_{22}, b_{33} are integrally separated. As it turns out, the first two functions are the hard ones (the third is more clearly integrally separated): Figure 9.1 shows $(b_{11} - b_{22})$ on $[0, 100]$, and clearly b_{11} and b_{22} are not separated. So, we check if (8.11) holds for H sufficiently large. In practice, to form b_{11}^H and b_{22}^H, we approximate the integral by the composite trapezoidal rule. We look for H in the range $[1, 20]$ and, for $t \in [0, 10000]$, the value $H = 20$ gave sufficient separation; see Figure 9.2. We conclude that, on the given interval, and subject to the limitations of finite precision computation, the diagonal of the transformed triangular system is integrally separated, and thus so is the linearized system, and the Lyapunov exponents are stable.

The next example highlights the difficulties in inferring integral separation of the diagonal of B for problems with close exponents, even when the Lyapunov exponents appear to exist as limits and to be stable.

Example 9.3. This example is adapted from one in [18] (used also in [13] and then [3]). We have a ring of oscillators with an external force proportional to the position component of the limit cycle of the van der Pol oscillator:

$$\begin{aligned} \ddot{y} + \alpha(y^2 - 1)\dot{y} + \omega^2 y &= 0, \\ \ddot{x}_i + d_i\dot{x}_i + \gamma[\Phi'(x_i - x_{i-1}) - \Phi'(x_{i+1} - x_i)] &= \sigma y \delta_{i1}\,, \qquad i = 1, \ldots, n\,. \end{aligned} \tag{9.1}$$

Above, $\Phi(x) = (x^2/2) + (x^4/4)$ is the single well Duffing potential, $\alpha, \omega, \gamma, \sigma$ are scalar parameters, x_i is the displacement of the ith particle, d_i is the damping coefficient, and we have periodic boundary conditions to be used in the expressions for Φ' ($x_0 = x_n$ and $x_{n+1} = x_1$). For our experiments, we set $n = 5$ and set $\alpha = 1$, $\omega = 1.82$, $\gamma = 1$, and $\sigma = 4$. We set $d_i = 0.25$ for i odd and $d_i = 0.15$ for i even. Initial conditions are taken as $y(0) = 0,\ \dot{y}(0) = -2,\ x_i(0) = \dot{x}_i(0) = 1,\ i = 1, \ldots, n$. Error control is performed on x, Q, and ν.

From the results summarized in Table 9.4, we observe good convergence for the four exponents on which we report. However, inferring integral separation, although

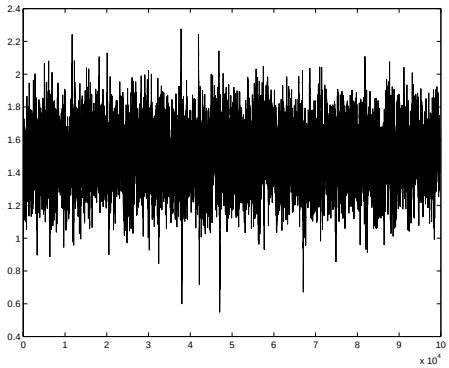


FIG. 9.2. $b^H_{11} - b^H_{22}$.

TABLE 9.4
Example 3. `TOLX=TOLQ=TOLL=TOL`.

t_{end}	`TOL`	`Steps`	λ_1	λ_2	λ_3	λ_4
1000	1.E-4	15214	1.8E-3	8.3E-4	-9.72E-2	-9.99E-2
5000	1.E-4	78599	4.9E-4	1.5E-4	-9.80E-2	-9.86E-2
10000	1.E-4	157372	2.1E-4	4.5E-5	-9.82E-2	-9.83E-2
1000	1.E-6	34911	1.7E-3	8.7E-4	-9.74E-2	-9.99E-2
5000	1.E-6	115135	4.2E-4	1.7E-4	-9.81E-2	-9.86E-2
10000	1.E-6	364206	1.4E-4	8.4E-5	-9.82E-2	-9.84E-2
1000	1.E-8	84584	1.7E-3	8.7E-4	-9.74E-2	-9.99E-2
5000	1.E-8	222556	4.2E-4	1.7E-4	-9.81E-2	-9.86E-2
10000	1.E-8	883292	1.4E-4	8.4E-5	-9.82E-2	-9.84E-2

perhaps possible, is quite difficult because of the clustering of the exponents. To illustrate, with $t_{\text{end}} = 1000$ and `TOL` $= 1.E-6$, at 2 digits the 12 approximate exponents are

$$\begin{array}{l} 1.7\text{E}-3,\ 8.7\text{E}-4,\ -9.7\text{E}-2,\ -1.0\text{E}-1,\ -1.0\text{E}-1,\ -1.0\text{E}-1, \\ -1.1\text{E}-1,\ -1.1\text{E}-1,\ -1.1\text{E}-1,\ -1.2\text{E}-1,\ -2.1\text{E}-1,\ -1.0\text{E}0. \end{array} \tag{9.2}$$

We attempted to verify if the linearized problem was integrally separated, but failed. To be precise, on the interval $[0, 1000]$, the value of $H = 100$ was sufficient to establish positivity of the Steklov differences $b^H_{22} - b^H_{33}$, $b^H_{10,10} - b^H_{11,11}$, and $b^H_{11,11} - b^H_{12,12}$, and hence integral separation of the respective diagonal entries of B, but all other Steklov differences were oscillating about 0 (even for larger values of H), therefore precluding us from inferring integral separation of the linearized problem. This highlights that, for a problem with close (or identical) exponents, it will be necessary to develop block analogues of QR techniques and associated criteria to infer integral separation (in a block sense).

10. Conclusions. In this paper we have blended theoretical studies on stability of Lyapunov exponents with computational techniques which target the Lyapunov exponents. Stability of the exponents is equivalent (in the case of distinct exponents) to having an integrally separated fundamental matrix solution. We have assumed that the system was integrally separated and further explored what conditions are needed to validate popular numerical methods, in particular those based on the QR and SVD of fundamental matrix solutions. We also explored the implications of integral

separation on approximation of three different spectra of linear systems: first, Σ_{ED}, of Sacker and Sell based upon exponential dichotomy; second, Σ_{L}, that naturally generalizes Lyapunov's upper and lower exponents to a spectrum; and third, Σ_{CL}, based on the diagonal elements of the upper triangular coefficient matrix B that is obtained through an orthogonal change of variables. In general, the Sacker–Sell spectrum is larger than the other two spectra, while under the assumption of integral separation of the diagonal of B we have that $\Sigma_{\mathrm{L}} = \Sigma_{\mathrm{CL}}$. We also showed how to approximate Σ_{ED} when the diagonal of B is integrally separated.

Future work will need to address several issues which we did not resolve in the present paper. In no particular order, we believe the following will be worthwhile investments:

1. Careful implementation and analysis of continuous SVD techniques.
2. Block analogues of QR and SVD techniques for the case of nondistinct exponents.
3. Refined implementation and study of techniques to approximate Σ_{ED} along the lines of the approach we laid down in section 8.3 and used in Example 9.1.
4. More thorough study of techniques to approximate Steklov averages, and further exploitation of the power of this tool.

REFERENCES

[1] L. Ya. Adrianova, *Introduction to Linear Systems of Differential Equations*, Transl. Math. Monographs 146, AMS, Providence, RI, 1995.

[2] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, *Lyapunov exponents for smooth dynamical systems and for Hamiltonian systems; A method for computing all of them. Part* 1: *Theory* and *Part* 2: *Numerical applications,* Meccanica, 15 (1980), pp. 9–30.

[3] T. Bridges and S. Reich, *Computing Lyapunov exponents on a Stiefel manifold,* Phys. D, 156 (2001), pp. 219–238.

[4] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N.K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function,* Numer. Math., 60 (1991), pp. 1–40.

[5] B.F. Bylov, R.E. Vinograd, D.M. Grobman, and V.V. Nemyckii, *The Theory of Lyapunov Exponents and Its Applications to Problems of Stability*, Nauka, Moscow, 1966 (in Russian).

[6] B.F. Bylov, *On the reduction of systems of linear equations to the diagonal form,* Math. Sb., 67 (1965), pp. 338–334.

[7] B.F. Bylov and N.A. Izobov, *Necessary and sufficient conditions for stability of characteristic exponents of a linear system,* Differ. Uravn., 5 (1969), pp. 1794–1903.

[8] F. Christiansen and H. H. Rugh, *Computing Lyapunov spectra with continuous Gram-Schmidt orthonormalization,* Nonlinearity, 10 (1997), pp. 1063–1072.

[9] W.A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, 1978.

[10] W.A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, Heath Math. Monographs, Heath & Co., Boston, 1965.

[11] L. Dieci and T. Eirola, *On smooth decompositions of matrices,* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 800–819.

[12] L. Dieci, R.D. Russell, and E.S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems,* SIAM J. Numer. Anal., 34 (1997), pp. 402–423.

[13] L. Dieci and E.S. Van Vleck, *Computation of a few Lyapunov exponents for continuous and discrete dynamical systems,* Appl. Numer. Math., 17 (1995), pp. 275–291.

[14] L. Dieci and E.S. Van Vleck,*Computation of orthonormal factors for fundamental solution matrices,* Numer. Math., 83 (1999), pp. 599–620.

[15] L. Dieci and E.S. Van Vleck, *Orthonormal integrators based on Householder and Givens transformations,* submitted.

[16] L. Dieci and E.S. Van Vleck, *Lyapunov and other spectra: A survey,* in Collected Lectures on the Preservation of Stability Under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002, pp. 197–218.
[17] S.P. Diliberto, *On systems of ordinary differential equations,* in Contributions to the Theory of Nonlinear Oscillations, Ann. of Math. Stud. 20, Princeton University Press, Princeton, NJ, 1950, pp. 1–38.
[18] U. Dressler, *Symmetry property of the Lyapunov spectra of a class of dissipative dynamical systems with viscous damping,* Phys. Rev. A, 38 (1988), pp. 2103–2109.
[19] J.M. Greene and J.-S. Kim, *The calculation of Lyapunov spectra,* Phys. D, 24 (1987), pp. 213–225.
[20] K. Geist, U. Parlitz, and W. Lauterborn, *Comparison of different methods for computing Lyapunov exponents,* Prog. Theoret. Phys., 83 (1990), pp. 875–893.
[21] I. Goldhirsch, P.L. Sulem, and S. A. Orszag, *Stability and Lyapunov stability of dynamical systems: A differential approach and a numerical method,* Phys. D, 27 (1987), pp. 311–337.
[22] A. Lyapunov, *Problém Géneral de la Stabilité du Mouvement*, Ann. of Math. Stud. 17, Princeton University Press, Princeton, NJ, 1949.
[23] H.D. Mayer, *Theory of the Lyapunov exponents of Hamiltonian systems and a numerical study of the transition from regular to irregular classical motion,* J. Chem. Phys., 84 (1986), pp. 3147–3161.
[24] V.M. Millionshchikov, *Systems with integral division are everywhere dense in the set of all linear systems of differential equations,* Differ. Uravn., 5 (1969), pp. 1167–1170.
[25] V.M. Millionshchikov, *Structurally stable properties of linear systems of differential equations,* Differ. Uravn., 5 (1969), pp. 1775–1784.
[26] V.V. Nemytskii and V.V. Stepanov, *Qualitative Theory of Differential Equations*, Dover, New York, 1989.
[27] V.I. Oseledec, *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems,* Trans. Moscow Math. Soc., 19 (1968), pp. 197–231.
[28] K. Palmer, *The structurally stable systems on the half-line are those with exponential dichotomy,* J. Differential Equations, 33 (1979), pp. 16–25.
[29] O. Perron, *Die Ordnungszahlen Linearer Differentialgleichungssysteme,* Math. Z., 31 (1930), pp. 748–766.
[30] R.J. Sacker and G.R. Sell, *A spectral theory for linear differential systems,* J. Differential Equations, 7 (1978), pp. 320–358.
[31] A. Wolf, J.B. Swift, H.L. Swinney, and J.A. Vastano, *Determining Lyapunov exponents from a time series,* Phys. D, 16 (1985), pp. 285–317.
[32] K. Wright, *Differential equations for the analytic singular value decomposition of a matrix,* Numer. Math., 63 (1992), pp 283–295.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 543–577

MULTISCALE ASYMPTOTIC ANALYSIS AND NUMERICAL SIMULATION FOR THE SECOND ORDER HELMHOLTZ EQUATIONS WITH RAPIDLY OSCILLATING COEFFICIENTS OVER GENERAL CONVEX DOMAINS*

LI-QUN CAO†, JUN-ZHI CUI†, AND DE-CHAO ZHU‡

Abstract. The multiscale asymptotic analysis and numerical simulation for the second order Helmholtz equations with rapidly oscillating coefficients over general convex domains are discussed in this paper. A multiscale asymptotic analysis formulation for this problem is presented by constructing properly the boundary layer. A multiscale numerical algorithm and a postprocessing technique are given. Finally, numerical results show that the method presented in this paper is effective and reliable.

Key words. multiscale asymptotic analysis, second order Helmholtz equation, rapidly oscillating coefficients, finite element method, postprocessing technique

AMS subject classifications. 65F10, 35P15

PII. S0036142900376110

1. Introduction.

In this paper, we analyze the spectral properties of a second order elliptic operator corresponding to a composite medium with a periodic microstructure. More precisely, we consider the following Helmholtz problem:

$$\begin{cases} \mathcal{L}_\varepsilon u^\varepsilon(x) \equiv -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial u^\varepsilon(x)}{\partial x_j}\right) + b\left(\dfrac{x}{\varepsilon}\right)u^\varepsilon(x) = \lambda^\varepsilon \rho\left(\dfrac{x}{\varepsilon}\right)u^\varepsilon(x) \quad \text{in} \quad \Omega, \\ \mathcal{B}_\varepsilon(u^\varepsilon) = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is an arbitrary bounded convex Lipschitz domain, B_ε is a boundary operator, i.e., it is either Dirichlet's or Neumann's one.

Let us make assumptions as follows:

(A_1) Let $\xi = \varepsilon^{-1}x$, $a_{ij}(\xi)$, $b(\xi)$, and $\rho(\xi)$ are 1-periodic in ξ;

(A_2)

$$\gamma_0 \sum_{i=1}^{n} \eta_i^2 \le \sum_{i,j=1}^{n} a_{ij}\left(\frac{x}{\varepsilon}\right)\eta_i\eta_j \le \gamma_1 \sum_{i=1}^{n} \eta_i^2 \tag{1.2}$$

$\exists \gamma_0 > 0$, $\gamma_1 > 0$, $\forall (\eta_1, \dots \eta_n) \in R^n$;

(A_3) $a_{ij}(\frac{x}{\varepsilon}) = a_{ji}(\frac{x}{\varepsilon})$, $\rho(\frac{x}{\varepsilon}) \ge \rho_0 = \text{const} > 0$, $b(\frac{x}{\varepsilon}) \ge 0$;

(A_4) $a_{ij}(\frac{x}{\varepsilon})$, $\rho(\frac{x}{\varepsilon})$, $b(\frac{x}{\varepsilon}) \in L^\infty(\Omega)$.

It is well known that spectral theories have extensive applications in physics, chemistry, mechanics, and vibration engineering. For example, we analyze the stability of a structure through calculating the natural frequencies and natural vibration

*Received by the editors August 2, 2000; accepted for publication (in revised form) November 13, 2001; published electronically May 29, 2002. This research was supported by the National Natural Science Foundation of China and Special Funds for Major State Basic Research Projects.

http://www.siam.org/journals/sinum/40-2/37611.html

†Institute of Computational Mathematics and Science/Engineering Computing, The Academy of Mathematics and System Sciences, The Chinese Academy of Sciences, P.O. Box 2719, 100080, Beijing, People's Republic of China (clq@lsec.cc.ac.cn, cjz@lsec.cc.ac.cn).

‡Solid Mechanics Research Center, Beijing University of Aeronautics and Astronautics, 100083, Beijing, People's Republic of China (dchaozhu@public.fhnet.cn.net).

modes. Also, in quantum physics, we investigate the electronic and magnetic properties of materials by computing eigenvalues and eigenfunctions of state Schrödinger equations associated with atomic structures. With the continuous emergence of some new materials such as composite materials, semiconductor superlattice, nanomaterials, and so forth, it is very necessary and pressing that the spectral properties of strongly heterogeneous materials are studied.

In mathematics, the above physical and mechanical problems can be represented as the Helmholtz equation of second order elliptic operator with rapidly oscillating coefficients in many cases. One of the main difficulties for analyzing these kinds of problems is that the computing amounts are too large to solve them numerically. For this purpose, a homogenization method was previously provided and analyzed; see [11, 13, 14, 16, 18]. The principal idea is to obtain the average field equation associated with the original problem by constructing a local smoothing operator and to discuss its convergence the on the basis of the compensated compactness theorem presented by L. Tartar (cf. [11]). Clearly, using classical numerical methods, we can numerically solve the homogenized Helmholtz equation in coarse meshes. Both theoretical analysis and numerical experiments show that the homogenization method is suitable for computing some macroscopic physical variables such as effective medium, natural frequencies, and so on. However, it is unable to accurately describe the local fluctuation of some physical variables such as natural vibration modes, stresses and strains, and temperature field, and so forth.

The crucial point of analyzing precisely the above problem is to find out the multiscale asymptotic expansion for the solution of considered problem (cf. [3, 4, 6]). Jikov, Zozlov, and Oleinik [11] and Oleinik, Shamaev, and Yosifian [18], investigated the Sturm–Liouville equation with oscillating coefficients in one dimension and obtained the complete asymptotic expansion of eigenvalues and eigenfunctions on the basis of analytic formulas of eigenvalues and eigenfunctions associated with the one-dimensional (1-D) Helmholtz problem with constant coefficients. Clearly, this method is not suitable for discussing the Helmholtz problems in higher-dimensional cases ($n \geq 2$). In [20], F. Santosa and M. Vogelius studied the first order corrections for a kind of eigenvalue problem associated with the vibration of periodic composite and analyzed the set of associated weak limit points in the case when Ω is a convex polygon, the sides of which all have a normal integer entries (the slopes are all rational or infinite).

S. Moskow and M. Vogelius (see [17]) gave a presentation formula for the set of first order corrections to a simple eigenvalue when the eigenvector is only in $H^{2+\omega}(\Omega)$ for some $\omega > 0$. In particular, this regularity assumption is sufficiently weak that it makes the representation formula valid when Ω is a convex, classical polygon. This is an important and excellent result.

T.Y. Hou, X.H. Wu, and Z. Cai [7] and T.Y. Hou and X.H. Wu [8] provided an interesting multiscale finite element method (FEM) based on the first order asymptotic expansion—the crucial idea is to find new finite element space; i.e., the set of basis functions consists of two parts, the first part being the set of piecewise polynomials and the second part the set of some oscillatory functions obtained by simultaneously solving locally partial differential equations in some subdomains.

Does there exist the multiscale asymptotic expansions of eigenvalues and eigenfunctions for second order Helmholtz equation with rapidly oscillating coefficients over general convex domains Ω in higher-dimensional cases ($n \geq 2$)? This is an interesting and difficult problem; refer to section 16 of [16]. One of the main results of this paper is that we answer the above problems in a sense and derive their rigorous verification.

The remainder of this paper is outlined as follows. In section 2 we introduce two examples associated with the themes of the paper derived from classical mechanics and Maxwell's equations. In section 3 the multiscale asymptotic analysis formulas of eigenvalues and eigenfunctions for the second order Helmholtz problem (1.1) with highly oscillatory coefficients over general convex domains are obtained, and the related regularity and error estimates are given. In section 4 the variation of eigenvalues and eigenfunctions is precisely analyzed due to the perturbation of the coefficients arising from numerically computing periodic solutions $N_{\alpha_1}(\xi),\ \alpha_1 = 1, 2, \dots n$. Section 5 is devoted to the finite element computations of the homogenized Helmholtz equation in the whole domain Ω and a boundary layer in a smaller computing scale. In section 6 the multiscale finite element algorithm and the postprocessing technique are presented, and the total error estimates are shown. Finally, some numerical results are reported, which provide a strong support for the effectiveness of the methods presented in this paper.

For convenience, we use throughout the paper the convention of summation upon repeated indices; C (with and without a subscript) denotes a generic positive constant, which is independent of ε.

2. The background of classical mechanics and Maxwell's equations. In this section, we concisely introduce some physical models of classical mechanics and Maxwell's equations associated with the themes of the paper.

2.1. Vibrations of the membrane with a periodic microstructure. As we know, the vibration of the membrane can be read as the following initial-boundary problem:

$$\begin{cases} \dfrac{\partial^2 V^\varepsilon(x,t)}{\partial t^2} - \dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial V^\varepsilon(x,t)}{\partial x_j}\right) = f(x,t) \quad \text{in} \quad \Omega \\ \text{subject to appropriate initial and boundary conditions,} \end{cases} \tag{2.1}$$

where $(a_{ij}(\frac{x}{\varepsilon}))$ is a symmetric, positive-definite matrix, and its elements are highly oscillatory coefficients with a small periodic parameter ε, and $f(x,t)$ is a given nonperiodic perturbation. Domain Ω and its basic configuration are shown in Figures 1 and 2.

By virtue of Fourier's transform, we know that

$$V^\varepsilon(x,t) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{+\infty} F(x,\omega)\chi^\varepsilon(x,\omega)e^{-i\omega t}d\omega, \tag{2.2}$$

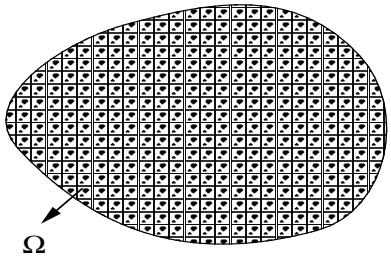


FIG. 1. *Periodic structure.*

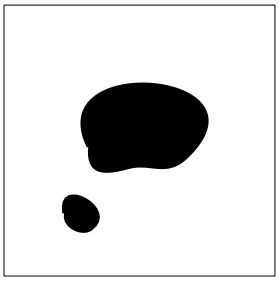


FIG. 2. *Basic configuration*

where

$$F(x,\omega) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{+\infty} f(x,t) e^{i\omega t} dt, \tag{2.3}$$

$$f(x,t) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{+\infty} F(x,\omega) e^{-i\omega t} d\omega, \tag{2.4}$$

and the admittance function

$$\chi^{\varepsilon}(x,\omega) = U^{\varepsilon}(x,\omega)/F(x,\omega), \tag{2.5}$$

where $U^{\varepsilon}(x,\omega)$ satisfies the following Helmholtz equation:

$$\begin{cases} -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{U^{\varepsilon}(x,\omega)}{\partial x_j}\right) = \omega^2 U^{\varepsilon}(x,\omega) \\ \text{subject to appropriate boundary condition.} \end{cases} \tag{2.6}$$

2.2. Propagation in a waveguide partially filled with anisotropic dielectric material. Consider Maxwell's equation without sources:

$$\begin{cases} \nabla \times \underline{E} + \dfrac{\partial \underline{B}}{\partial t} = \underline{0}, \\ \nabla \times \underline{H} - \dfrac{\partial \underline{D}}{\partial t} = \underline{0}, \\ \nabla \cdot \underline{D} = 0, \\ \nabla \cdot \underline{B} = 0, \\ \underline{D} = \epsilon \underline{E}, \\ \underline{B} = \mu \underline{H}, \end{cases} \tag{2.7}$$

where $\underline{E} = (E_x, E_y, E_z)^T, \underline{D} = (D_x, D_y, D_z)^T, \underline{H} = (H_x, H_y, H_z)^T, \underline{B} = (B_x, B_y, B_z)^T$ are electric field intensity, electric flux density, magnetic field strength, and magnetic flux density, respectively. ϵ, μ are the dielectric constant tensor and the magnetic permeability tensor of anisotropic media.

Suppose that the propagation of a wave is in the z-axis; then

$$\underline{E}(t,x,y,z) = \underline{E}(x,y) e^{i(\omega t - k_z z)}, \tag{2.8a}$$

$$\underline{H}(t,x,y,z) = \underline{H}(x,y) e^{i(\omega t - k_z z)}, \tag{2.8b}$$

where $\sqrt{-1} = i$ and ω, k_z are the angular frequency and the propagation constant, respectively.

If $\epsilon = (\epsilon_{ij})_{3\times 3}$ is a symmetric tensor, then we can always change it into a diagonal tensor by using an orthogonal transformation. Therefore assume here that ϵ is a diagonal tensor without loss of generality, i.e., $\epsilon = diag(\epsilon_x, \epsilon_y, \epsilon_z)$, $\mu = diag(\mu_x, \mu_y, \mu_z)$.

Substituting (2.8a) and (2.8b) into (2.7), one can obtain

$$\begin{cases} \dfrac{\partial E_z}{\partial y} = -ik_z E_y - i\omega\mu_x H_x, \\ \dfrac{\partial E_z}{\partial x} = -ik_z E_x + i\omega\mu_y H_y, \\ \dfrac{\partial E_y}{\partial x} - \dfrac{\partial E_x}{\partial y} = -i\omega\mu_z H_z, \\ \dfrac{\partial H_z}{\partial y} = -ik_z H_y + i\omega\epsilon_x E_x, \\ \dfrac{\partial H_z}{\partial x} = -ik_z H_x - i\omega\epsilon_y E_y, \\ \dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = i\omega\epsilon_z E_z. \end{cases} \tag{2.9}$$

From (2.9), we can easily see that, once E_z, H_z are determined, other variables can be calculated.

One can directly verify that $\Phi = (E_z, H_z)$ satisfies the following governed equation:

$$T\Phi = k_0^2 M\Phi, \tag{2.10}$$

where

(2.11)

$$T = \begin{pmatrix} -\dfrac{\partial}{\partial x}\left(\dfrac{\epsilon_x}{\kappa_x^2}\dfrac{\partial}{\partial x}\right) - \dfrac{\partial}{\partial y}\left(\dfrac{\epsilon_y}{\kappa_y^2}\dfrac{\partial}{\partial y}\right) & \dfrac{k_z}{\omega}\left[\dfrac{\partial}{\partial y}\left(\dfrac{1}{\kappa_y^2}\dfrac{\partial}{\partial x}\right) - \dfrac{\partial}{\partial x}\left(\dfrac{1}{\kappa_x^2}\dfrac{\partial}{\partial y}\right)\right] \\ \dfrac{k_z}{\omega}\left[\dfrac{\partial}{\partial x}\left(\dfrac{1}{\kappa_y^2}\dfrac{\partial}{\partial y}\right) - \dfrac{\partial}{\partial y}\left(\dfrac{1}{\kappa_x^2}\dfrac{\partial}{\partial x}\right)\right] & -\dfrac{\partial}{\partial x}\left(\dfrac{\mu_x}{\kappa_x^2}\dfrac{\partial}{\partial x}\right) - \dfrac{\partial}{\partial y}\left(\dfrac{\mu_y}{\kappa_y^2}\dfrac{\partial}{\partial y}\right) \end{pmatrix},$$

$$M = \begin{pmatrix} \epsilon_z & 0 \\ 0 & \mu_z \end{pmatrix}. \tag{2.12}$$

3. Multiscale asymptotic analysis for the second order Helmholtz equation on general convex domains. As one of the main results of this section, we will obtain multiscale asymptotic analysis formulas of eigenvalues and eigenfunctions of the Helmholtz problem (1.1).

Formally, set

$$u^\varepsilon(x) \cong \sum_{l=0}^{+\infty} \varepsilon^l \sum_{\alpha_1,\cdots,\alpha_l=1}^{n} N_{\alpha_1\cdots\alpha_l}(\xi) D^\alpha u^0(x), \tag{3.1}$$

$$\lambda^\varepsilon \cong \sum_{i=0}^{+\infty} \varepsilon^i \lambda^{(i)}(\varepsilon). \tag{3.2}$$

In contrast to the usual expression, for the sake of convenience, we use here the following notation:

$$D^\alpha v = \frac{\partial^l v}{\partial x_{\alpha_1}\cdots\partial x_{\alpha_l}}, \qquad \alpha = \{\alpha_1, \cdots \alpha_l\},\ \langle\alpha\rangle = l, \tag{3.3}$$

α_i takes the values $1, 2, \dots n$.

Substituting (3.1), (3.2) into (1.1), and taking into account that $\frac{\partial}{\partial x_i} \to \frac{\partial}{\partial x_i} + \frac{1}{\varepsilon}\frac{\partial}{\partial \xi_i}$, one can write

(3.4)
$$\begin{aligned} 0 &= \mathcal{L}_\varepsilon u^\varepsilon(x) - \lambda^\varepsilon \rho(\tfrac{x}{\varepsilon}) u^\varepsilon(x) \\ &= -\sum_{l=0}^{+\infty} \varepsilon^{l-2} \sum_{\alpha_1,\dots,\alpha_l=1}^{n} H_{\alpha_1\cdots\alpha_l}(\xi) D^\alpha u^0(x) + \sum_{l=0}^{+\infty} \varepsilon^l \sum_{\alpha_1,\dots,\alpha_l=1}^{n} b(\xi) N_{\alpha_1\cdots\alpha_l}(\xi) D^\alpha u^0(x) \\ &\quad - \sum_{s=0}^{+\infty} \varepsilon^s \sum_{i=0}^{s} \lambda^{(i)} \sum_{\alpha_1,\dots,\alpha_{s-i}=1}^{n} \rho(\xi) N_{\alpha_1\cdots\alpha_{s-i}}(\xi) D^\alpha u^0(x), \end{aligned}$$

where

$$H_0(\xi) = \frac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\frac{\partial N_0(\xi)}{\partial \xi_j}\right), \tag{3.5}$$

$$H_{\alpha_1}(\xi) = \frac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\frac{\partial N_{\alpha_1}(\xi)}{\partial \xi_j}\right) + \frac{\partial}{\partial \xi_i}\left(a_{i\alpha_1}(\xi) N_0(\xi)\right) + a_{\alpha_1 j}(\xi)\frac{\partial N_0(\xi)}{\partial \xi_j}. \tag{3.6}$$

For $\langle\alpha\rangle = l \geq 2$

$$\begin{aligned} H_{\alpha_1\cdots\alpha_l}(\xi) &= \frac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\frac{\partial N_{\alpha_1\cdots\alpha_l}(\xi)}{\partial \xi_j}\right) + \frac{\partial}{\partial \xi_i}(a_{i\alpha_1}(\xi) N_{\alpha_2\cdots\alpha_l}(\xi)) \\ &\quad + a_{\alpha_1 j}(\xi)\frac{\partial N_{\alpha_2\cdots\alpha_l}(\xi)}{\partial \xi_j} + a_{\alpha_1\alpha_2}(\xi) N_{\alpha_3\cdots\alpha_l}(\xi). \end{aligned} \tag{3.7}$$

In order to compare the coefficients of powers $\varepsilon^{-2}, \varepsilon^{-1}, \varepsilon^0, \varepsilon^1, \dots$ on both sides of (3.4) and to ensure that (3.4) is an identity equation, we assume that

$$\begin{cases} H_0(\xi) = 0, \\ H_{\alpha_1}(\xi) = 0, \\ H_{\alpha_1\alpha_2}(\xi) = \hat{a}_{\alpha_1\alpha_2}, \\ H_{\alpha_1\alpha_2\alpha_3}(\xi) = N_{\alpha_1}(\xi)\hat{a}_{\alpha_2\alpha_3}, \\ \cdots \\ H_{\alpha_1\alpha_2\cdots\alpha_l}(\xi) = N_{\alpha_1\cdots\alpha_{l-2}}(\xi)\hat{a}_{\alpha_{l-1}\alpha_l}, \quad l \geq 4. \end{cases} \tag{3.8}$$

Notice that $\hat{a}_{\alpha_1\alpha_2}$, $N_{\alpha_1\cdots\alpha_j}(\xi)$, $\alpha_j = 1, 2, \dots n$, $j = 0, 1, 2, \dots$, will be defined below.

Substituting (3.8) into (3.4), one obtains

(3.9)
$$\begin{aligned} 0 = \sum_{l=0}^{+\infty} \varepsilon^l \Bigg\{ \sum_{\alpha_1,\cdots\alpha_l=1}^{n} N_{\alpha_1,\dots,\alpha_l}(\xi) \frac{\partial^l}{\partial x_{\alpha_1}\cdots\partial x_{\alpha_l}} \Bigg[- \sum_{\alpha_{l+1},\alpha_{l+2}=1}^{n} \hat{a}_{\alpha_{l+1}\alpha_{l+2}} \frac{\partial^2 u^0(x)}{\partial x_{\alpha_{l+1}}\partial x_{\alpha_{l+2}}} \\ + b(\xi) u^0(x) \Bigg] - \sum_{i=0}^{l} \lambda^{(i)} \sum_{\alpha_1,\cdots,\alpha_{l-i}=1}^{n} \rho(\xi) N_{\alpha_1\cdots\alpha_{l-i}}(\xi) D^\alpha u^0(x) \Bigg\}. \end{aligned}$$

To begin with, consider the first equation of (3.8), and conclude that

$$\begin{cases} \dfrac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\dfrac{\partial N_0(\xi)}{\partial \xi_j}\right) = 0 \quad \text{in} \quad R^n, \\ N_0(\xi) \quad \text{is} \quad \text{1-periodic} \quad \text{in} \quad \xi. \end{cases} \tag{3.10}$$

Let us remark that the equation

$$\begin{cases} \mathcal{L}_1\phi(\xi) \equiv \dfrac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\dfrac{\partial \phi(\xi)}{\partial \xi_j}\right) = F(\xi) \quad \text{in} \quad Q, \\ \phi(\xi) \quad \text{is 1-periodic in} \quad \xi \end{cases} \tag{3.11}$$

admits a unique solution (up to an additive constant) iff

$$\int_Q F(\xi)d\xi = 0,$$

where the unit cube $Q = \{\xi \in R^n,\ 0 < \xi_j < 1,\ j = 1, \dots n\}$.

Combining (3.10) with (3.11), it is easy to see that $N_0(\xi) \equiv C$; here set $C = 1$ without loss of generality.

From the second equation of (3.8), we define $N_{\alpha_1}(\xi)$ in the following way:

$$\begin{cases} \dfrac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\dfrac{\partial N_{\alpha_1}(\xi)}{\partial \xi_j}\right) = -\dfrac{\partial}{\partial \xi_i}(a_{i\alpha_1}(\xi)) \quad \text{in} \quad Q, \\ N_{\alpha_1}(\xi) = 0 \quad \text{on} \quad \partial Q. \end{cases} \tag{3.12}$$

Integrating on both sides of the third equation of (3.8) in ξ over the unit cell Q, and taking into account that $N_{\alpha_1}(\xi), N_{\alpha_1\alpha_2}(\xi)$ are 1-periodic functions in ξ, one can conclude that

$$\hat{a}_{\alpha_1\alpha_2} = \int_Q \left(a_{\alpha_1\alpha_2}(\xi) + a_{\alpha_1 j}(\xi)\frac{\partial N_{\alpha_2}(\xi)}{\partial \xi_j}\right) d\xi. \tag{3.13}$$

From (3.8), we next define $N_{\alpha_1\alpha_2}(\xi), \dots N_{\alpha_1\cdots\alpha_l}(\xi)$ in the following ways:

$$\begin{cases} \dfrac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\dfrac{\partial N_{\alpha_1\alpha_2}(\xi)}{\partial \xi_j}\right) = -\dfrac{\partial}{\partial \xi_i}(a_{i\alpha_1}(\xi)N_{\alpha_2}(\xi)) \\ \qquad - a_{\alpha_1 j}(\xi)\dfrac{\partial N_{\alpha_2}(\xi)}{\partial \xi_j} - a_{\alpha_1\alpha_2}(\xi) + \hat{a}_{\alpha_1\alpha_2} \quad \text{in } Q, \\ N_{\alpha_1\alpha_2}(\xi) = 0 \quad \text{on} \quad \partial Q. \end{cases} \tag{3.14}$$

For $\langle\alpha\rangle = l \geq 3$

$$\begin{cases} \dfrac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\dfrac{\partial N_{\alpha_1\cdots\alpha_l}(\xi)}{\partial \xi_j}\right) = -\dfrac{\partial}{\partial \xi_i}(a_{i\alpha_1}(\xi)N_{\alpha_2\cdots\alpha_l}(\xi)) \ - a_{\alpha_1 j}(\xi)\dfrac{\partial N_{\alpha_2\cdots\alpha_l}(\xi)}{\partial \xi_j} \\ \qquad - a_{\alpha_1\alpha_2}(\xi)N_{\alpha_3\cdots\alpha_l}(\xi) + N_{\alpha_1\cdots\alpha_{l-2}}(\xi)\hat{a}_{\alpha_{l-1}\alpha_l} \quad \text{in } Q, \\ N_{\alpha_1\cdots\alpha_l}(\xi) = 0 \quad \text{on} \quad \partial Q. \end{cases} \tag{3.15}$$

Remark 3.1. Existence and uniqueness of the solutions $N_{\alpha_1}(\xi), \ldots, N_{\alpha_1\cdots\alpha_l}(\xi)$ for problems (3.12), (3.14), and (3.15), respectively, can be easily established by induction with respect to l due to the uniform elliptic condition (A_2)–(A_4), Poincaré–Friedrichs' inequality, and Lax–Milgram's lemma. Then they are extended to the whole R^n by the 1-periodicity.

Remark 3.2. It is worthwhile to notice that the periodic solutions $N_{\alpha_1}(\xi)$, $\alpha_1 = 1, \ldots n$, defined in this paper, generally speaking, are different from $\tilde{N}_{\alpha_1}(\xi)$ defined in classical homogenization books (see [2, 11, 16, 18]) due to the different boundary conditions on ∂Q. But we can prove that their homogenized matrices are the same; see Appendix A.

Next, let us identify the coefficients of powers $\varepsilon^0, \varepsilon^1, \varepsilon^2, \ldots$ on the both sides of (3.9) for $l = 0$:

$$-\frac{\partial}{\partial x_{\alpha_1}}\left(\hat{a}_{\alpha_1\alpha_2}\frac{\partial u^0(x)}{\partial x_{\alpha_2}}\right) + b(\xi)u^0(x) = \lambda^{(0)}\rho(\xi)u^0(x), \quad \text{a.e.} \quad \xi \in R^n. \tag{3.16}$$

For $l = 1$

$$N_{\alpha_1}(\xi)\frac{\partial}{\partial x_{\alpha_1}}\left\{-\frac{\partial}{\partial x_{\alpha_2}}\left(\hat{a}_{\alpha_2\alpha_3}\frac{\partial u^0(x)}{\partial x_{\alpha_3}}\right) + b(\xi)u^0(x) - \lambda^{(0)}\rho(\xi)u^0(x)\right\} - \lambda^{(1)}\rho(\xi)u^0(x) = 0. \tag{3.17}$$

From (3.16), we know

$$\lambda^{(1)}\langle\rho\rangle u^0(x) = 0,$$

where $\langle f\rangle = \frac{1}{|Q|}\int_Q f(\xi)d\xi$, $|Q|$ denotes the Lebesgue measure of the unit cube Q.

Since $\langle\rho\rangle \neq 0$, $\|u^0\|_{L^2(\Omega)} = 1$, then $\lambda^{(1)} = 0$ holds. For $l = 2$

$$\begin{aligned} &N_{\alpha_1\alpha_2}(\xi)\frac{\partial^2}{\partial x_{\alpha_1}\partial x_{\alpha_2}}\left\{-\frac{\partial}{\partial x_{\alpha_3}}\left(\hat{a}_{\alpha_3\alpha_4}\frac{\partial u^0(x)}{\partial x_{\alpha_4}}\right) + b(\xi)u^0(x) - \lambda^{(0)}\rho(\xi)u^0(x)\right\} \\ &-\lambda^{(1)}\rho(\xi)N_{\alpha_1}(\xi)\frac{\partial u^0(x)}{\partial x_{\alpha_1}} - \lambda^{(2)}\rho(\xi)u^0(x) = 0. \end{aligned} \tag{3.18}$$

Analogously, we can infer that $\lambda^{(2)} = 0$.

The remainder shall similarly be proven, i.e., $\lambda^{(i)} = 0$, $i \geq 3$.

On the other hand, from (3.16), we know

$$-\frac{\partial}{\partial x_{\alpha_1}}\left(\hat{a}_{\alpha_1\alpha_2}\frac{\partial u^0(x)}{\partial x_{\alpha_2}}\right) = (\lambda^{(0)}\rho(\xi) - b(\xi))u^0(x).$$

Since $u^0(x) \not\equiv 0, x \in \Omega$, then there are some points $x \in \Omega$ such that $u^0(x) \neq 0$, and

$$-\frac{1}{u^0(x)}\frac{\partial}{\partial x_{\alpha_1}}\left(\hat{a}_{\alpha_1\alpha_2}\frac{\partial u^0(x)}{\partial x_{\alpha_2}}\right) = \lambda^{(0)}\rho(\xi) - b(\xi) \equiv C. \tag{3.19}$$

Integrating on the both sides of (3.19) in Q, we have

$$C = \lambda^{(0)}\langle\rho\rangle - \langle b\rangle. \tag{3.20}$$

Therefore we can prove that (3.16) and $\widehat{\mathcal{B}}(u^0) = 0$ on $\partial\Omega$ are equivalent to the following homogenized Helmholtz equation associated with problem (1.1):

$$\begin{cases} \widehat{\mathcal{L}}u^0(x) \equiv -\dfrac{\partial}{\partial x_i}\left(\hat{a}_{ij}\dfrac{\partial u^0(x)}{\partial x_j}\right) + \langle b\rangle u^0(x) = \lambda^{(0)}\langle\rho\rangle u^0(x) \quad \text{in} \quad \Omega, \\ \widehat{\mathcal{B}}(u^0) = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{3.21}$$

where $\hat{a}_{ij}$ is as shown in (3.13), and

$$\widehat{\mathcal{B}}(v) = \begin{cases} v & \text{for Dirichlet's boundary condition,} \\ \nu_i \hat{a}_{ij}\dfrac{\partial v}{\partial x_j} & \text{for Neumann's boundary condition.} \end{cases}$$

Remark 3.3. One can check that $\widehat{\mathcal{L}}$ is a linear symmetric positive-definite operator; see [2, 11, 16, 18].

For an integer $M \geq 2$, set

$$u_k^{\varepsilon,M}(x) = \sum_{l=0}^{M}\varepsilon^l \sum_{\alpha_1,\cdots,\alpha_l=1}^{n} N_{\alpha_1\cdots\alpha_l}(\xi)D^\alpha u_k^0(x), \tag{3.22}$$

$$\lambda_k^{\varepsilon,M} \equiv \lambda_k^{(0)}, \quad k = 1, 2, \ldots, \tag{3.23}$$

$$\begin{aligned} &\mathcal{L}_\varepsilon u_k^{\varepsilon,M}(x) - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)u_k^{\varepsilon,M}(x) = \sum_{l=0}^{M}\varepsilon^{l-2}\sum_{\alpha_1,\ldots,\alpha_l=1}^{n} H_{\alpha_1\cdots\alpha_l}(\xi)D^\alpha u_k^0(x) \\ &+ \sum_{l=0}^{M-2}\varepsilon^l \sum_{\alpha_1,\ldots,\alpha_l=1}^{n} b(\xi)N_{\alpha_1\cdots\alpha_l}(\xi)D^\alpha u_k^0(x) \\ &- \sum_{l=0}^{M-2}\varepsilon^l \sum_{\alpha_1,\ldots,\alpha_l=1}^{n} \lambda_k^{(0)}\rho(\xi)N_{\alpha_1\cdots\alpha_l}(\xi)D^\alpha u_k^0(x) + \varepsilon^{M-1}F_0(x,\varepsilon) \\ &= \sum_{l=0}^{M-2}\varepsilon^l \sum_{\alpha_1\cdots\alpha_l=1}^{n} N_{\alpha_1\cdots\alpha_l}(\xi)\frac{\partial^l}{\partial x_{\alpha_1}\cdots\partial x_{\alpha_l}}\left[\hat{a}_{\alpha_{l+1}\alpha_{l+2}}\frac{\partial^2 u_k^0(x)}{\partial x_{\alpha_{l+1}}\partial x_{\alpha_{l+2}}}\right. \\ &\left. + b(\xi)u_k^0(x) - \lambda_k^{(0)}\rho(\xi)u_k^0(x)\right] + \varepsilon^{M-1}F_0(x,\varepsilon) \\ &= \varepsilon^{M-1}F_0(x,\varepsilon), \end{aligned} \tag{3.24}$$

where $F_0(x,\varepsilon)$ is a sum of terms having the form $\varepsilon^i\psi(\xi)D^l u_k^0(x)$, $2 \leq M$, $i \geq 0$, $\psi(\xi)$ is a bounded function, and $\|F_0(x,\varepsilon)\|_{L^2(\Omega)} \leq C$, C is a constant independent of ε, x.

Let $\Omega_0 = \bigcup_{z\in\widehat{T}_\varepsilon}\varepsilon(z+Q) \subset \Omega$ as shown in Figure 3, where the index set $\widehat{T}_\varepsilon = \{z = (z_1,\ldots,z_n) \in Z^n,\ \varepsilon(z+Q) \subset \Omega\}$, and the unit cube $Q = \{\xi \in R^n : 0 < \xi_j < 1,\ j = 1, 2, \ldots, n\}$.

For simplicity, we assume that $\varepsilon/2 \leq dist(\partial\Omega_0, \partial\Omega) \leq 2\varepsilon$, $\Omega_1 = \Omega \setminus \bar{\Omega}_0$,$\Gamma^* = \partial\Omega_0 \cap \partial\Omega_1$ as shown in Figure 4.

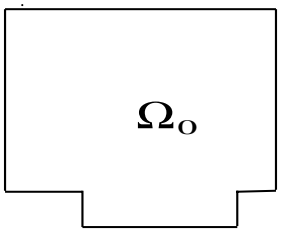


FIG. 3. *Subdomain* Ω_0.

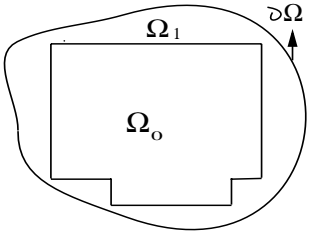


FIG. 4. *Boundary layer* Ω_1.

Now let us define the boundary layer in the following way:

$$\begin{cases} -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial W_k^\varepsilon(x)}{\partial x_j}\right)+\left(b\left(\dfrac{x}{\varepsilon}\right)-\lambda_k^{(0)}\rho\left(\dfrac{x}{\varepsilon}\right)\right)W_k^\varepsilon(x)=0, \quad x\in\Omega_1, \\ W_k^\varepsilon(x)=u_k^0(x), \quad x\in\partial\Omega_0, \\ \mathcal{B}_\varepsilon(W_k^\varepsilon)=0, \quad x\in\partial\Omega, \end{cases} \tag{3.25}$$

where $(\lambda_k^{(0)}, u_k^0(x))$, is the kth eigenpair of the homogenized Helmholtz problem (3.21), $k=1,2,\ldots$.

Remark 3.4. It is worthwhile to notice that $W_k^\varepsilon(x)$ is independent of ε along $\partial\Omega_0\cap\partial\Omega_1$. We will prove that it is correct in some cases later.

To begin with, let us discuss the corresponding homogeneous boundary value problem as follows:

$$\begin{cases} -\dfrac{\partial}{\partial x_i}\left(a_{ij}(\frac{x}{\varepsilon})\dfrac{\partial V_k^\varepsilon(x)}{\partial x_j}\right)+\left(b\left(\frac{x}{\varepsilon}\right)-\lambda_k^{(0)}\rho\left(\frac{x}{\varepsilon}\right)\right)V_k^\varepsilon(x)=0, \quad x\in\Omega_1, \\ V_k^\varepsilon(x)=0, \quad x\in\partial\Omega_0, \\ \mathcal{B}_\varepsilon(V_k^\varepsilon)=0, \quad x\in\partial\Omega. \end{cases} \tag{3.26}$$

Now we define the operator $\mathcal{K}_\varepsilon : L^2(\Omega_1)\to L^2(\Omega_1)$, set $\mathcal{K}_\varepsilon f^\varepsilon = v^\varepsilon$, where v^ε is the solution of the following problem:

$$\begin{cases} \mathcal{Q}_\varepsilon(v^\varepsilon)\equiv -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial v^\varepsilon}{\partial x_j}\right)+b\left(\dfrac{x}{\varepsilon}\right)v^\varepsilon=\rho\left(\dfrac{x}{\varepsilon}\right)f^\varepsilon \quad \text{in} \quad \Omega_1, \\ v^\varepsilon(x)=0 \quad \text{on} \quad \partial\Omega_0, \\ \mathcal{B}_\varepsilon(v^\varepsilon)=0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{3.27}$$

By virtue of conditions (A_2)–(A_3), we know that $\mathcal{Q}_\varepsilon : H^1(\Omega_1,\partial\Omega_0)\to L^2(\Omega_1)$ is a symmetric and positive-definite operator; therefore the inverse operator $\mathcal{K}_\varepsilon=\mathcal{Q}_\varepsilon^{-1}$: $L^2(\Omega_1)\to L^2(\Omega_1)$ is a bounded self-adjoint compact operator due to the compact imbedding $H^1(\Omega_1)\to L^2(\Omega_1)$. $\sigma_d(\mathcal{K}_\varepsilon)$ denotes the set of discrete spectra of operator $\mathcal{K}_\varepsilon$.

For simplicity, in this section we assume that

$$(\lambda_k^{(0)})^{-1}\notin\sigma_d(\mathcal{K}_\varepsilon). \tag{3.28}$$

Remark 3.5. We can prove that the first eigenvalue $(\lambda_1^{(0)})^{-1} \notin \sigma_d(\mathcal{K}_\varepsilon)$; see Appendix B.

It follows from Fredholm's alternative theorem that (3.26) admits a unique solution $V_k^\varepsilon(x) = 0$. Therefore (3.25) has one and only one solution.

For $M \geq 2$, set

(3.29)
$$U_k^{\varepsilon,M}(x) = \begin{cases} u_k^{\varepsilon,M}(x) = u_k^0(x) + \sum_{l=1}^{M} \varepsilon^l \sum_{\alpha_1,\ldots,\alpha_l=1}^{n} N_{\alpha_1\cdots\alpha_l}(\xi) D^\alpha u_k^0(x), & x \in \overline{\Omega}_0, \\ W_k^\varepsilon(x), \quad x \in \Omega_1 = \Omega \setminus \overline{\Omega}_0, \end{cases}$$

where $N_{\alpha_1\cdots\alpha_l}(\xi)$ is as defined in (3.12), (3.14), and (3.15).

Since $u_k^{\varepsilon,M}(x)|_{\partial\Omega_0\cap\partial\Omega_1} = u_k^0(x)|_{\partial\Omega_0\cap\partial\Omega_1} = W_k^\varepsilon(x)|_{\partial\Omega_0\cap\partial\Omega_1}$, then $U_k^{\varepsilon,M}(x) \in H^1(\Omega)$ holds. But, generally speaking, $[\frac{\partial U_k^{\varepsilon,M}}{\partial n}]|_{\partial\Omega_0\cap\partial\Omega_1} \neq 0$. To this end, we have to do it through some well-known regularizations.

At first, let us introduce a set of open covering $\{\mathcal{V}_l\}_{l=1}^3$ of the bounded closed set $\overline{\Omega} \subset R^n$:

(3.30)
$$\begin{aligned} \mathcal{V}_1 &= \left\{ x \in \Omega_0 : \quad dist(x, \partial\Omega_0) > \frac{\delta}{2} \right\}, \\ \mathcal{V}_2 &= \left\{ x \in (R^n \setminus \overline{\Omega}_0) : dist(x, \partial\Omega_0) > \frac{\delta}{2}, dist(x, \partial\Omega) < \delta \right\}, \\ \mathcal{V}_3 &= \{ x \in \Omega : \quad dist(x, \partial\Omega_0) < \delta \}, \end{aligned}$$

$\overline{\Omega} \subset \cup_{l=1}^3 \mathcal{V}_l$.

Using the partition of unity theorem, there exists a set of functions $\{\psi_l(x)\}_{l=1}^3$ such that

(1) $\psi_l(x) \in C_0^\infty(\mathcal{V}_l)$;

(2) $\sum_{l=1}^3 \psi_l(x) \equiv 1$, $\forall x \in \Omega$.

Set $\Omega_0'' = \Omega_0 \setminus \overline{\mathcal{V}}_3$, $\Omega_1'' = \Omega_1 \setminus \overline{\mathcal{V}}_3$, and choose a sufficiently small $\delta > 0$ such that $\delta \leq C \cdot \varepsilon^M$, $M \geq 2$.

Define

$$\tilde{U}_k^{\varepsilon,M}(x) = \psi_1(x) \cdot U_k^{\varepsilon,M}(x) + \psi_2(x) \cdot U_k^{\varepsilon,M}(x) + J_\delta * (\psi_3(x) \cdot U_k^{\varepsilon,M}(x)), \tag{3.31}$$

where the regularization $J_\delta * u$ is defined in section 2.17 of [1].

One can directly verify that $\tilde{U}_k^{\varepsilon,M}(x) \in H^1(\Omega)$ and $[\frac{\partial \tilde{U}_k^{\varepsilon,M}}{\partial n}]|_{\partial\Omega_0\cap\partial\Omega_1} = 0$.

LEMMA 3.1 (see [11, 18]). *Let A: $H \to H$ be a linear self-adjoint compact operator in a Hilbert space H. Let $\mu \in R^1$, and let $u \in H$ be such that $\|u\|_H = 1$ and*

$$\|Au - \mu u\|_H \leq \beta, \qquad \beta = \text{const} > 0. \tag{3.32}$$

Then there exists an eigenvalue μ_i of operator A such that

$$|\mu_i - \mu| \leq \beta. \tag{3.33}$$

Moreover, for any $d > \beta$ there exists a vector $\bar{u} \in H$ such that

$$\|u - \bar{u}\|_H \leq 2\beta d^{-1}, \qquad \|\bar{u}\|_H = 1, \tag{3.34}$$

and $\bar{u}(x)$ is a linear combination of the eigenvectors of operator A corresponding to the eigenvalues within the interval $[\mu - d, \mu + d]$.

THEOREM 3.1. *Let $(\lambda_k^\varepsilon, u_k^\varepsilon)$ be the kth eigenpair of the original Helmholtz problem* (1.1)*, set $\lambda_0^\varepsilon = 0$, and let $U_k^{\varepsilon,M}$, $\tilde{U}_k^{\varepsilon,M}$, $\lambda_k^{\varepsilon,M}$ be defined in the formulas* (3.29)*,* (3.31)*, and* (3.23)*, respectively. If $a_{ij}(\frac{x}{\varepsilon}) \in C(\overline{\Omega})$, $\nabla_\xi a_{ij}(\xi) \in L^\infty(\Omega)$, then it holds that*

$$|\lambda_k^\varepsilon - \lambda_k^{\varepsilon,M}| \le C_1(k)\varepsilon^{M-1}, \quad k = 1, 2, \ldots. \tag{3.35}$$

Moreover, if the multiplicity of the eigenvalues $\lambda_k^{(0)}$ is equal to t, i.e.,

$$\lambda_{k-1}^{(0)} < \lambda_k^{(0)} = \cdots = \lambda_{k+t-1}^{(0)} < \lambda_{k+t}^{(0)}, \quad \lambda_0^{(0)} = 0,$$

then

$$\|U_k^{\varepsilon,M} - \bar{u}_k^\varepsilon\|_{L^2(\Omega)} \le C_2(k)\varepsilon^{M-1}, \quad M \ge 2, \tag{3.36}$$

where $\bar{u}_k^\varepsilon$ is a linear combination of the eigenfunctions of problem (1.1) *corresponding to the eigenvalues $\lambda_k^\varepsilon \cdots \lambda_{k+t-1}^\varepsilon$.*

In particular, if $\lambda_k^{(0)}$ is a simple eigenvalue, then

$$\|U_k^{\varepsilon,M} - u_k^\varepsilon\|_{L^2(\Omega)} \le C_2(k)\varepsilon^{M-1}. \tag{3.37}$$

Proof. To begin, consider the auxiliary problem as follows:

$$\begin{cases} \mathcal{L}_\varepsilon w^\varepsilon(x) = f^\varepsilon(x) & \text{in } \Omega, \\ w^\varepsilon(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.38}$$

If $f^\varepsilon(x) \in L^2(\Omega)$, it follows from (A_2)–(A_4), Poincaré–Friedrichs' inequality, and Lax–Milgram's lemma that there exists a unique weak solution of problem (3.38). In other words, $\mathcal{L}_\varepsilon : H_0^1(\Omega) \to L^2(\Omega)$ is a homeomorphism mapping.

Set $\mathcal{N}_\varepsilon = \mathcal{L}_\varepsilon^{-1}$, one can check that $\mathcal{N}_\varepsilon : L^2(\Omega) \to L^2(\Omega)$ is a uniform bounded linear operator in ε, i.e., $\|\mathcal{N}_\varepsilon\|_{L^2(\Omega)\to L^2(\Omega)} \le C$, where C is a positive constant independent of ε.

If $x \in \overline{\Omega}_0$, from (3.24), then it holds in the sense of distributions that

$$\mathcal{L}_\varepsilon u_k^{\varepsilon,M}(x) - \lambda_k^{(0)}\rho\left(\frac{x}{\varepsilon}\right) u_k^{\varepsilon,M}(x) = \varepsilon^{M-1}F_0(x,\varepsilon). \tag{3.39}$$

If $x \in \Omega_1$, from (3.25), we have

$$\mathcal{L}_\varepsilon W_k^\varepsilon(x) - \lambda_k^{(0)}\rho\left(\frac{x}{\varepsilon}\right) W_k^\varepsilon(x) = 0. \tag{3.40}$$

From (3.39), (3.40), and (3.31), we obtain the following equation which holds in the sense of distributions:

$$\begin{cases} \mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M}(x) - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\overline{\varepsilon}}\right)\tilde{U}_k^{\varepsilon,M}(x) = \tilde{F}_0(x,\varepsilon), \quad x \in \Omega, \\ \mathcal{B}_\varepsilon \tilde{U}_k^{\varepsilon,M}(x) = 0, \quad x \in \partial\Omega. \end{cases} \tag{3.41}$$

For any $v(x) \in L^2(\Omega)$, we obtain

(3.42)

$$\begin{aligned} (\tilde{F}_0, v)_\Omega &= \left(\mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\tilde{U}_k^{\varepsilon,M}, v\right)_\Omega \\ &= \left(\mathcal{L}_\varepsilon u_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right) u_k^{\varepsilon,M}, v\right)_{\Omega_0''} + \left(\mathcal{L}_\varepsilon W_k^\varepsilon - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right) W_k^\varepsilon, v\right)_{\Omega_1''} \\ &+ \left(\mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\overline{\varepsilon}}\right)\tilde{U}_k^{\varepsilon,M}, v\right)_{\mathcal{V}_3\cap\Omega_0} + \left(\mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\overline{\varepsilon}}\right)\tilde{U}_k^{\varepsilon,M}, v\right)_{\mathcal{V}_3\cap\Omega_1}. \end{aligned}$$

On the other hand, we know that

(3.43a)
$$\left(\mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\tilde{U}_k^{\varepsilon,M}, v\right)_{\mathcal{V}_3\cap\Omega_0} = \left(\mathcal{L}_\varepsilon u_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)u_k^{\varepsilon,M}, v\right)_{\mathcal{V}_3\cap\Omega_0} + \left(\mathcal{L}_\varepsilon \Lambda(x) - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\Lambda(x), v\right)_{\mathcal{V}_3\cap\Omega_0},$$

where $\Lambda(x) = \psi_3(x)u_k^{\varepsilon,M}(x) - J_\delta * (\psi_3(x)u_k^{\varepsilon,M}(x))$.

Similarly, we have

(3.43b)
$$\left(\mathcal{L}_\varepsilon \tilde{U}_k^{\varepsilon,M} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\tilde{U}_k^{\varepsilon,M}, v\right)_{\mathcal{V}_3\cap\Omega_1} = \left(\mathcal{L}_\varepsilon W_k^{\varepsilon} - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)W_k^{\varepsilon}, v\right)_{\mathcal{V}_3\cap\Omega_1} + \left(\mathcal{L}_\varepsilon \Theta(x) - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\Theta(x), v\right)_{\mathcal{V}_3\cap\Omega_1},$$

where $\Theta(x) = \psi_3(x)W_k^{\varepsilon}(x) - J_\delta * (\psi_3(x)W_k^{\varepsilon}(x))$.

Assume that Ω is a bounded convex Lipschitz domain and $a_{ij}(\frac{x}{\varepsilon}) \in C(\overline{\Omega})$, $\nabla_\xi a_{ij}(\xi) \in L^\infty(\Omega)$. By virtue of a priori estimates of PDEs (see [9, 10, 12]), we can prove that $u_k^\varepsilon(x) \in H^2(\Omega)$, $u_k^{\varepsilon,M}(x) \in H^2(\Omega_0)$, $W^\varepsilon(x) \in W^{2,p}(\Omega_1), 1 < p \le p_0 < +\infty$; also see Appendix C.

Using Theorem 3.16 of [1], we know that

$$\|\Lambda\|_{2,\mathcal{V}_3\cap\Omega_0} \le \delta, \quad \|\Theta\|_{2,p,\mathcal{V}_3\cap\Omega_1} \le \delta. \tag{3.44}$$

From (3.42), (3.43a), (3.43b), (3.44), (3.39), and (3.40), one can obtain

(3.45a)
$$\begin{aligned}
&\|\tilde{F}_0\|_{0,\Omega}^2 = (\tilde{F}_0, \tilde{F}_0)_\Omega \\
&\le C\left\{\varepsilon^{M-1}\|\tilde{F}_0\|_{0,\Omega_0} + \varepsilon^{-1}\cdot\|\Lambda\|_{2,\mathcal{V}_3\cap\Omega_0}\cdot\|\tilde{F}_0\|_{0,\mathcal{V}_3\cap\Omega_0} + \varepsilon^{-1}\cdot\|\Theta\|_{2,p,\mathcal{V}_3\cap\Omega_1}\cdot\|\tilde{F}_0\|_{0,p',\mathcal{V}_3\cap\Omega_1}\right\} \\
&\le C\left\{\varepsilon^{M-1}\|\tilde{F}_0\|_{0,\Omega_0} + \varepsilon^{-1}\cdot\delta\cdot\|\tilde{F}_0\|_{0,\Omega_0} + \varepsilon^{-1}\cdot\delta\cdot\|\tilde{F}_0\|_{0,p',\Omega_1}\right\} \\
&\le C\left\{\varepsilon^{M-1}\|\tilde{F}_0\|_{0,\Omega_0} + \varepsilon^{M-1}\|\tilde{F}_0\|_{0,p',\Omega_1}\right\},
\end{aligned}$$

where $p' = \frac{p}{p-1} > 2$. Let $p' = 2(1-\theta) + \theta\tilde{p}, \quad 2 < \tilde{p} < +\infty, \quad 0 \le \theta \le 1$.

Using the interpolation theorem (see Theorem 1.3.7 of [19]) we know that

$$\|\tilde{F}_0\|_{0,p',\Omega_1} \le C\|\tilde{F}_0\|_{0,\Omega_1}^{1-\theta}\cdot\|\tilde{F}_0\|_{0,\tilde{p},\Omega_1}^{\theta}. \tag{3.45b}$$

If we assume $\tilde{F}_0 \in L^{\tilde{p}}(\Omega_1)$, $\tilde{p} >> 2$, then we have $1-\theta \approx 1$. For convenience, we say that

$$\|\tilde{F}_0\|_{0,p',\Omega_1} \le C\cdot\|\tilde{F}_0\|_{0,\Omega_1}. \tag{3.45c}$$

Combining (3.45c) with (3.45a), one obtains

$$\|\tilde{F}_0\|_{0,\Omega} \le C\cdot\varepsilon^{M-1}. \tag{3.46}$$

From (3.41), we have

$$\tilde{U}_k^{\varepsilon,M}(x) - \lambda_k^{\varepsilon,M}\rho\left(\frac{x}{\varepsilon}\right)\mathcal{N}_\varepsilon(\tilde{U}_k^{\varepsilon,M}) = \mathcal{N}_\varepsilon(\tilde{F}_0(x,\varepsilon)).$$

Set $u = (\|\tilde{U}_k^{\varepsilon,M}\|_{0,\Omega})^{-1}\tilde{U}_k^{\varepsilon,M}$, $A = \mathcal{N}_\varepsilon$, $\lambda = \lambda_k^{\varepsilon,M} \equiv \lambda_k^{(0)}$, $\beta = -\|\mathcal{N}_\varepsilon(\tilde{F}_0)\|_{0,\Omega}(\|\tilde{U}_k^{\varepsilon,M}\|_{0,\Omega})^{-1}$, $H = L^2(\Omega)$.

It follows from Lemma 3.1 that there exists an eigenvalue $(\lambda_{n(k)}^\varepsilon)^{-1}$ of operator $\mathcal{N}_\varepsilon$ such that

$$|(\lambda_k^{\varepsilon,M})^{-1} - (\lambda_{n(k)}^\varepsilon)^{-1}| = |(\lambda_k^{(0)})^{-1} - (\lambda_{n(k)}^\varepsilon)^{-1}| \le C\varepsilon^{M-1}, \quad M \ge 2.$$

In accordance with the proof procedure of Theorem 2.1 in section 2.1, Chapter III of [18] (also see [11]), we know that $\lambda_k^\varepsilon \to \lambda_k^{(0)}$, $\varepsilon \to 0$. Hence, we can conclude that there is a small neighborhood of the point $\lambda_k^{(0)}$ which contains a eigenvalue λ_k^ε such that $\lambda_{n(k)}^\varepsilon = \lambda_k^\varepsilon$. Therefore

$$|\lambda_k^\varepsilon - \lambda_k^{\varepsilon,M}| \le C_1(k)\varepsilon^{M-1}.$$

Using Lemma 3.1 again, one obtains

$$\|\tilde{U}_k^{\varepsilon,M} - \bar{u}_k^\varepsilon\|_{0,\Omega} \le C_2(k)\varepsilon^{M-1}.$$

Especially if the eigenvalue $\lambda_k^{(0)}$ of problem (3.21) is simple, then one can choose $\bar{u}_k^\varepsilon = c_0 u_k^\varepsilon$, $c_0 = \text{const}$, such that

$$\|\tilde{U}_k^{\varepsilon,M} - u_k^\varepsilon\|_{0,\Omega} \le C_2(k)\varepsilon^{M-1}.$$

By using Theorem 3.16 of [1] again, it is easy to prove that

$$\|\tilde{U}_k^{\varepsilon,M} - U_k^{\varepsilon,M}\|_{0,\Omega} \le \delta \le C\cdot\varepsilon^M.$$

The proof of Theorem 3.1 is complete. □

Remark 3.6. One can directly verify that $\|U_k^{\varepsilon,M} - \bar{u}_k^\varepsilon\|_{1,\Omega} \le C\cdot\varepsilon^{\frac{(M-1)}{2}}$. Therefore, if $\lambda_k^{(0)}$ is simple, then $\|u_k^\varepsilon - W_k^\varepsilon\|_{0,\Omega_1} \le C|u_k^\varepsilon - u_k^0|_{1/2,\partial\Omega_0\cap\partial\Omega_1} \le C\|U_k^{\varepsilon,M} - u_k^\varepsilon\|_{1,\Omega} \le C\cdot\varepsilon^{\frac{(M-1)}{2}}$ on the basis of the trace theorem.

Next let us give some regularity results about $W_k^\varepsilon(x)$, which is of great use for error estimates of numerical computation.

THEOREM 3.2. *Let $W_k^\varepsilon(x)$ be the weak solution of problem* (3.25). *If $a_{ij}(\frac{x}{\varepsilon})$, $b(\frac{x}{\varepsilon})$, $\rho(\frac{x}{\varepsilon})$ satisfy conditions* (A_2)–(A_4), $(\lambda_k^{(0)})^{-1} \notin \sigma_d(\mathcal{K}_\varepsilon)$, *then it holds that*

$$\|W_k^\varepsilon\|_{1,\Omega_1} \le C\|u_k^0\|_{1,\Omega}, \quad k = 1,2,\ldots, \tag{3.47}$$

where C is independent of $\varepsilon, W_k^\varepsilon, u_k^0$.

The proof of Theorem 3.2 refers to section 3, Chapter I of [2].

THEOREM 3.3. *Let $\Omega_1 = \Omega \setminus \overline{\Omega}_0 \subset R^2$, and let $W^\varepsilon(x)$ be the weak solution of problem* (3.25). *For the sake of convenience, we do omit the subscript k. Under the hypotheses of Theorem* 3.2, *if $a_{ij}(\frac{x}{\varepsilon}) \in C(\overline{\Omega})$, $\nabla_\xi a_{ij}(\xi) \in L^\infty(\Omega)$, then there exists $1 < p_0 < +\infty$ such that*

$$W^\varepsilon(x) \in W^{2,p}(\Omega_1), \quad 1 < p \le p_0, \tag{3.48}$$

$$\|W^\varepsilon\|_{2,p,\Omega_1} \le C\varepsilon^{-2}\|u^0\|_{2,p,\Omega}. \tag{3.49}$$

The proof of Theorem 3.3 refers to Appendix C.

4. Finite element computations of periodic solutions $N_\alpha(\xi)$ and their related problems.

4.1. Finite element computations of periodic solutions $N_\alpha(\xi)$. For simplicity, we discuss only two-dimensional (2-D) problems without loss of generality.

Let $\mathcal{J}^{h_0} = \{K\}$ be a family of regular triangulations of the square Q, $h_0 = \max_K\{h_K\}$. Define a piecewise linear finite element space

$$V_{h_0} = \{v \in C(\bar{Q}) : v|_K \in P_1(K), \quad v|_{\partial Q} = 0\} \subset H_0^1(Q). \tag{4.1}$$

PROPOSITION 4.1. *Let $N_{\alpha_1\cdots\alpha_l}(\xi)$, $\alpha_j = 1, 2, \ldots, n$, $j = 1, \ldots, l$, be the weak solutions of problems* (3.12), (3.14), *and* (3.15), *respectively, and $N^{h_0}_{\alpha_1\cdots\alpha_l}(\xi)$ be the corresponding finite element solutions. If $N_{\alpha_1\cdots\alpha_j}(\xi) \in H^2(Q), j = 1, 2, \ldots, l$, then it holds that*

$$\|N_{\alpha_1\cdots\alpha_l} - N^{h_0}_{\alpha_1\cdots\alpha_l}\|_{1,Q} \le Ch_0 \left(\sum_{j=1}^{l} \|N_{\alpha_1\cdots\alpha_j}\|_{2,Q}\right), \tag{4.2}$$

where $C > 0$ is independent of $h_0, \varepsilon, N_{\alpha_1\cdots\alpha_j}, j = 1, \ldots, l$.

4.2. Perturbation bounds for eigenvalues and eigenfunctions of the modified homogenized Helmholtz equation. In practice, we need to solve the following modified homogenized Helmholtz equation associated with (3.21):

$$\begin{cases} \widehat{\mathcal{L}}_{h_0}\tilde{u}^0(x) \overset{\text{def}}{\equiv} -\dfrac{\partial}{\partial x_i}\left(\hat{a}^{h_0}_{ij}\dfrac{\partial \tilde{u}^0(x)}{\partial x_j}\right) + \langle b\rangle \tilde{u}^0(x) = \tilde{\lambda}^{(0)}\langle\rho\rangle\tilde{u}^0(x) \quad \text{in} \quad \Omega, \\ \widehat{\mathcal{B}}_{h_0}(\tilde{u}^0) = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{4.3}$$

where

$$\hat{a}^{h_0}_{ij} = \int_Q \left(a_{ij}(\xi) + a_{ik}(\xi)\frac{\partial N^{h_0}_j(\xi)}{\partial \xi_k}\right) d\xi. \tag{4.4}$$

$N^{h_0}_j(\xi)$ are the finite element approximate solution of $N_j(\xi)$ defined in (3.12).

$$\widehat{\mathcal{B}}_{h_0}(v) = \begin{cases} v \quad \text{for Dirichlet's boundary condition,} \\ \nu_i \hat{a}^{h_0}_{ij}\dfrac{\partial v}{\partial x_j} \quad \text{for Neumann's boundary condition.} \end{cases} \tag{4.5}$$

One can verify the following.

PROPOSITION 4.2. *The partial differential operator $\hat{\mathcal{L}}_{h_0}$ defined by* (4.3) *satisfies the following properties:*

$$(1) \quad \hat{a}^{h_0}_{ij} = \hat{a}^{h_0}_{ji}, \tag{4.6}$$

$$(2) \quad \bar{\mu}_1\eta_i\eta_i \le \hat{a}^{h_0}_{ij}\eta_i\eta_j \le \bar{\mu}_2\eta_i\eta_i \quad \forall(\eta_1\cdots\eta_n) \in R^n, \tag{4.7}$$

where $\bar{\mu}_1, \bar{\mu}_2 > 0$ are constants independent of the mesh size h_0.

Next we will precisely analyze the influence on the eigenvalues and the eigenfunctions because of the perturbation of the coefficients arising from computing numerically $N_{\alpha_1}(\xi)$, $\alpha_1 = 1, 2, \ldots, n$.

THEOREM 4.1. *Suppose that* $(\lambda_k^{(0)}, u_k^0)$ *and* $(\tilde{\lambda}_k^{(0)}, \tilde{u}_k^0)$ *are the kth eigenpairs of problems* (3.21) *and* (4.3)*, respectively. Then it holds that*

$$|\tilde{\lambda}_k^{(0)} - \lambda_k^{(0)}| \le C_k h_0^2 \|N_i\|_{2,Q}^2. \tag{4.8}$$

Moreover, if the multiplicity of the eigenvalue $\lambda_k^{(0)}$ *is equal to* t*, i.e.,*

$$\lambda_{k-1}^{(0)} < \lambda_k^{(0)} = \cdots = \lambda_{k+t-1}^{(0)} < \lambda_{k+t}^{(0)}, \quad \lambda_0^{(0)} = 0,$$

then

$$\|u_k^0 - \bar{u}_k^0\|_{L^2(\Omega)} \le C_k h_0^2 \|N_i\|_{2,Q}^2, \tag{4.9}$$

where $\bar{u}_k^0$ *is a linear combination of eigenfunctions of problem* (4.3) *corresponding to the eigenvalues* $\tilde{\lambda}_k^{(0)}, \ldots, \tilde{\lambda}_{k+t-1}^{(0)}$.

For the proof of Theorem 4.1, please refer to Appendix D.

COROLLARY 4.1. *Under the assumptions of Theorem* 4.1*, it then holds that*

$$\|u_k^0 - \bar{u}_k^0\|_{1,\Omega} \le C_k h_0 \|N_j\|_{2,Q}. \tag{4.10}$$

By virtue of interior regularity estimates of PDEs, one can directly prove the following.

THEOREM 4.2. *Suppose that* $(\lambda_k^{(0)}, u_k^{(0)}(x))$, $(\tilde{\lambda}_k^{(0)}, \tilde{u}_k^{(0)}(x))$ *are the kth eigenpairs of* (3.21) *and* (4.3)*, respectively,* $k = 1, 2, \ldots$, $\|u_k^0\|_{0,\Omega} = 1$, $\|\tilde{u}_k^0\|_{0,\Omega} = 1$, Ω_0 *as shown in Figure* 3*, and* $\Omega_0 \subset\subset \Omega' \subset\subset \Omega$*. If* $u^0(x) \in H^{M+2}(\Omega')$*, then it holds that*

$$\|u_k^0(x) - \tilde{u}_k^0(x)\|_{s,\Omega_0} \le C h_0^2 \|N_i\|_{2,Q}^2 \|u^0\|_{M+2,\Omega'}, \tag{4.11}$$

where $s = 0, 1, \ldots, M$, $k = 1, 2, \ldots$.

5. Solve approximately the homogenized Helmholtz equation and the boundary layer.

5.1. Finite element computation of the homogenized Helmholtz equation. For the sake of convenience, here we numerically solve the homogenized Helmholtz equation with the Dirichlet's boundary conditions over the whole domain Ω in a coarse mesh, and we assume that $\langle b \rangle \equiv 0$, $\langle \rho \rangle \equiv 1$.

For simplicity, suppose that $\Omega \subset R^2$ is a bounded smooth domain. Let $J^h = \{e\}$ be a family of regular subdivisions of $\Omega, h = \max_e h_e$, and satisfying the following properties:

(F_1) The elements are uniform rectangles in the interior domain $\Omega_0 \subset\subset \Omega$.

(F_2) The elements are regular triangles in region $\Omega_1 = \Omega \setminus \overline{\Omega}_0$, and the elements are (curved) triangles near the boundary $\partial\Omega$.

(F_3) Any face of any element e_1 is either a subset of the boundary $\partial\Omega$ or a face of another element e_2 in the subdivision.

Define a finite element space, $r \ge 1$,

$$S_0^h(\Omega) = \{v \in C(\overline{\Omega}) : \quad v|_e \in \overline{P}_r(e), \quad v|_{\partial\Omega} = 0\} \subset H_0^1(\Omega), \tag{5.1}$$

where

$$\overline{P}_r = \begin{cases} Q_r, & e \quad \text{is a rectangle,} \\ P_r, & e \quad \text{is a triangle.} \end{cases}$$

For simplicity, we give only the error estimates for the first eigenvalue problem without loss of generality.

The discrete variational form for the modified Helmholtz equation (4.3) is

$$A(\tilde{u}^0_{1,h}, v_h) = \tilde{\lambda}^{(0)}_{1,h}(\tilde{u}^0_{1,h}, v_h) \quad \forall v_h \in S^h_0(\Omega), \tag{5.2}$$

where the bilinear form

$$A(u, v) = \int_\Omega \hat{a}^{h_0}_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx. \tag{5.3}$$

At first, let us introduce some notation. Set $\|w\|^2_A = A(w, w)$, where $A(\cdot, \cdot)$ is as stated in (5.3). Define a Ritz–Galerkin projection operator $R_h : H^1_0(\Omega) \to S^h_0(\Omega)$ such that

$$A(u - R_h u, v_h) = 0, \quad u \in H^1_0(\Omega), \quad \forall v_h \in S^h_0(\Omega). \tag{5.4}$$

THEOREM 5.1. *Suppose that* $(\tilde{\lambda}^{(0)}_1, \tilde{u}^0_1)$ *and* $(\tilde{\lambda}^{(0)}_{1,h}, \tilde{u}^0_{1,h})$ *are the first eigenpairs of problems* (4.3), (5.2), *respectively; it then holds that*

$$0 \le \frac{A(w, w)}{(w, w)} - \tilde{\lambda}^0_1 \le \frac{\|w - \tilde{u}^0_1\|^2_A}{(w, w)} \quad \forall w \in H^1_0(\Omega), \tag{5.5}$$

$$0 < \tilde{\lambda}^{(0)}_1 \le \frac{A(\tilde{u}^0_{1,h}, \tilde{u}^0_{1,h})}{(\tilde{u}^0_{1,h}, \tilde{u}^0_{1,h})} = \tilde{\lambda}^{(0)}_{1,h} \le \frac{A(v, v)}{(v, v)} \quad \forall v \in S^h_0(\Omega), \tag{5.6}$$

and

$$0 \le \tilde{\lambda}^{(0)}_{1,h} - \tilde{\lambda}^{(0)}_1 \le \frac{A(R_h\tilde{u}^0_1, R_h\tilde{u}^0_1)}{(R_h\tilde{u}^0_1, R_h\tilde{u}^0_1)} - \tilde{\lambda}^{(0)}_1 \le \frac{\|R_h\tilde{u}^0_1 - \tilde{u}^0_1\|^2_A}{(R_h\tilde{u}^0_1, R_h\tilde{u}^0_1)}. \tag{5.7}$$

Proof. It is easy to see that (5.6) is true, so let us turn to the proof of (5.5). One can directly check that

$$\|w\|^2_A = \tilde{\lambda}^{(0)}_1 (w, \tilde{u}^0_1)^2 + \|w - (w, \tilde{u}^0_1)\tilde{u}^0_1\|^2_A. \tag{5.8}$$

It follows from (5.8) that

$$\tilde{\lambda}^{(0)}_1 = \min_{w \in H^1_0(\Omega),\ w \ne 0} \frac{\|w\|^2_A}{\|w\|^2_0} \le \frac{\|w\|^2_A}{\|w\|^2_0} \le \tilde{\lambda}^{(0)}_1 + \frac{\|w - (w, \tilde{u}^0_1)\tilde{u}^0_1\|^2_A}{\|w\|^2_0}. \tag{5.9}$$

On the other hand, one can show that

$$A(w - (w, \tilde{u}^0_1)\tilde{u}^0_1, \alpha\tilde{u}^0_1) = 0 \quad \forall \alpha \in R.$$

Hence

$$\begin{aligned} \|w - \tilde{u}^0_1\|^2_A &= \|w - (w, \tilde{u}^0_1)\tilde{u}^0_1\|^2_A + \|(w, \tilde{u}^0_1)\tilde{u}^0_1 - \tilde{u}^0_1\|^2_A \\ &\ge \|w - (w, \tilde{u}^0_1)\tilde{u}^0_1\|^2_A. \end{aligned}$$

From (5.9), we have

$$0 \le \frac{A(w, w)}{(w, w)} - \tilde{\lambda}^{(0)}_1 \le \frac{\|w - (w, \tilde{u}^0_1)\tilde{u}^0_1\|^2_A}{\|w\|^2_0} \le \frac{\|w - \tilde{u}^0_1\|^2_A}{\|w\|^2_0}.$$

Setting $w = R_h\tilde{u}_1^0$ in (5.5) and using (5.6), one can obtain

$$0 \le \tilde{\lambda}_{1,h}^{(0)} - \tilde{\lambda}_1^{(0)} \le \frac{\|R_h\tilde{u}_1^0 - \tilde{u}_1^0\|_A^2}{(R_h\tilde{u}_1^0, R_h\tilde{u}_1^0)}.$$

The proof of Theorem 5.1 is complete. □

THEOREM 5.2. *Let $(\tilde{\lambda}_1^{(0)}, \tilde{u}_1^0)$ and $(\tilde{\lambda}_{1,h}^{(0)}, \tilde{u}_{1,h}^0)$ be the first eigenpairs of problems* (4.3) *and* (5.2)*, respectively. If $\tilde{u}_1^0 \in H^{r+1}(\Omega)$, $S_0^h(\Omega)$ as indicated in* (5.1)*, then it holds that*

$$0 \le \tilde{\lambda}_{1,h}^{(0)} - \tilde{\lambda}_1^{(0)} \le Ch^{2r}. \tag{5.10}$$

Proof. It is well known that

$$\|R_h\tilde{u}_1^0 - \tilde{u}_1^0\|_0 \le Ch^2\|\tilde{u}_1^0\|_2,$$

$$\|R_h\tilde{u}_1^0 - \tilde{u}_1^0\|_A^2 \le Ch^{2r}\|\tilde{u}_1^0\|_{r+1}^2.$$

Choosing a sufficiently small $h > 0$ such that

$$\|R_h\tilde{u}_1^0\|_0 \ge \|\tilde{u}_1^0\|_0 - Ch^2\|\tilde{u}_1^0\|_2 \ge \frac{1}{2},$$

it follows from (5.7) that

$$0 \le \tilde{\lambda}_{1,h}^{(0)} - \tilde{\lambda}_1^{(0)} \le C \cdot h^{2r}\|\tilde{u}_1^0\|_{r+1}^2.$$

Now let us give some superconvergence estimates for the first eigenfunctions. For simplicity, set $\lambda = \tilde{\lambda}_1^{(0)}$, $\lambda_h = \tilde{\lambda}_{1,h}^{(0)}$, $u_h = \tilde{u}_{1,h}^0$; H_λ is the eigenspace of the operator $\widehat{\mathcal{L}}_{h_0}$ with respect to eigenvalue $\lambda = \tilde{\lambda}_1^{(0)}$. Define a projection operator $P : L^2(\Omega) \to H_\lambda$ such that

$$Pu = \sum_{i=1}^{l}(u, u_i)u_i, \tag{5.11}$$

where u_i, $i = 1, \ldots, l$, form a set of orthonormal basis of H_λ.

Let K be the inverse operator of $\widehat{\mathcal{L}}_{h_0}$; then K is a bounded self-adjoint compact operator due to Proposition 4.2.

LEMMA 5.1 (see [15]). *Let $R_h : H_0^1(\Omega) \to S_0^h(\Omega)$ be the Ritz–Galerkin projection operator, for $q_0 > 2$, $1 < q < q_0$; then it holds that*

$$\|\lambda_h R_h K - \lambda K\|_\infty \to 0 \quad as \quad h \to 0, \tag{5.12}$$

$$\|R_h K\|_\infty \le C. \tag{5.13}$$

LEMMA 5.2 (see [15]). *Let $u = Pu_h \in H_\lambda \subset W^{r+1,q}(\Omega)$, $q > 2$; then it holds that*

$$\|u\|_{r+1,q} \le C, \tag{5.14}$$

$$\|R_h K(I - R_h)u\|_\infty \le Ch^{r+2} \quad (r \ge 2). \tag{5.15}$$

THEOREM 5.3. *Let (λ, H_λ) be the solution of problem (4.3) and let (λ_h, V_λ) be the solution of problem (5.2), $V_\lambda \subset S_0^h(\Omega)$; if $H_\lambda \subset W^{r+1,q} \cap H_0^1(\Omega)$, $q > 2$, for any given $u_h \in V_\lambda$, then there exists $u \in H_\lambda$ such that*

$$\|R_h u - u_h\|_{0,\infty} \le Ch^{r+2} \quad (r \ge 2). \tag{5.16}$$

Proof. Set $u = Pu_h = \sum_{j=1}^{l}(u_h, u_i)u_i$, $\bar{u} = R_h u - u_h - P(R_h u - u_h)$. It is obvious that $(\bar{u}, u) = 0$ for any $u \in H_\lambda$, thus $\bar{u} \in H_\lambda^\perp$.

It follows from Fredholm's alternative theorem that the operator $(I - \lambda K)$ has a bounded inverse operator. Thus there exists a constant $\delta_0 > 0$ such that

$$\begin{aligned}
&\delta_0\|\bar{u}\|_{0,\infty} \le \|(I-\lambda K)\bar{u}\|_{0,\infty} = \|(I-\lambda K)(R_h u - u_h)\|_{0,\infty} \\
&= \|\lambda R_h K(I - R_h)u + (\lambda_h R_h K - \lambda K)(R_h u - u_h) \\
&\qquad + (\lambda - \lambda_h)R_h K R_h u\|_{0,\infty} \quad (\text{since } u_h = \lambda_h R_h K u_h, \quad u = \lambda K u) \\
&\le \lambda\|R_h K(I - R_h)u\|_{0,\infty} \\
&\quad + \|\lambda_h R_h K - \lambda K\|_\infty \|R_h u - u_h\|_{0,\infty} + Ch^{2r}.
\end{aligned} \tag{5.17}$$

Since $Pu - Pu_h = 0$,

$$\begin{aligned}
&\|P(R_h u - u_h)\|_{0,\infty} = \|P(R_h u - u)\|_{0,\infty} = \left\|\sum_{j=1}^{l}(R_h u - u, u_j)u_j\right\|_{0,\infty} \\
&\le \sum_{j=1}^{l}|(R_h u - u, u_j)|\|u_j\|_{0,\infty} = \sum_{j=1}^{l}\frac{1}{\lambda}|A(R_h u - u, u_j - u_j^I)|\|u_j\|_{0,\infty} \\
&\le Ch^{2r}\sum_{j=1}^{l}\|u\|_{r+1}\|u_j\|_{r+1}\|u_j\|_{0,\infty} \le Ch^{2r}\|u\|_{r+1}.
\end{aligned}$$

Using the triangle inequality, we have

$$\|R_h u - u_h\|_{0,\infty} \le \|\bar{u}\|_{0,\infty} + \|P(R_h u - u_h)\|_{0,\infty} \le \|\bar{u}\|_{0,\infty} + Ch^{2r}. \tag{5.18}$$

Substituting (5.17) into (5.18), one obtains

$$\begin{aligned}
&(1 - \tfrac{1}{\delta_0}\|\lambda_h R_h K - \lambda K\|_\infty)\|R_h u - u_h\|_{0,\infty} \\
&\le \tfrac{1}{\delta_0}\|R_h K(I - R_h)u\|_{0,\infty} + Ch^{2r}\|u\|_{r+1}.
\end{aligned}$$

Using Lemmas 5.1 and 5.2 and choosing a sufficiently small $h > 0$, we have

$$\frac{1}{2}\|R_h u - u_h\|_{0,\infty} \le Ch^{r+2}\|u\|_{r+1,q}. \qquad \square$$

Now we apply the above superconvergence results to implement the postprocessing technique of $D^\alpha \tilde{u}_k^0(x)$, where $\tilde{u}_k^0$ is the kth eigenfunction associated with problem (4.3), $k = 1, 2, \ldots$.

Using the nodal values of the bi-r-th finite element solution, we construct a bi-2r-th interpolation function at a new larger element with respect to a coarse mesh, which is called as the interpolated FEM refer to [15], as shown in Figures 5 and 6. Denote by $\mathcal{I}_{2h}^{(2r)}$ the bi-2r-th order interpolation operator.

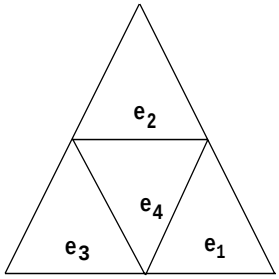


FIG. 5. *Triangular mesh.*

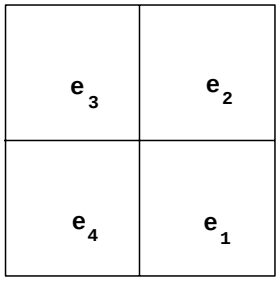


FIG. 6. *Rectangular mesh.*

LEMMA 5.3 (see [15]). *Let $\mathcal{I}_h : H^1(\Omega) \to S_0^h(\Omega)$ be a usual Lagrange interpolation operator; then the interpolation operators $\mathcal{I}_h$ and $\mathcal{I}_{2h}^{(2r)}$ satisfy the following properties:*

$$\|\mathcal{I}_{2h}^{(2r)}u\|_{m,p} \le C\|u\|_{m,p}, \quad 1 \le p \le \infty,\ m = 0, 1,\ \forall\, u \in S^h(\Omega_0), \tag{5.19}$$

where $C > 0$ does depend on r, p but does not depend on u, h;

$$(\mathcal{I}_{2h}^{(2r)})^2 = \mathcal{I}_{2h}^{(2r)}, \quad \mathcal{I}_{2h}^{(2r)}\mathcal{I}_h = \mathcal{I}_{2h}^{(2r)}, \quad \mathcal{I}_h\mathcal{I}_{2h}^{(2r)} = \mathcal{I}_h, \tag{5.20}$$

$$\forall\, P_i \in T_0^h,\ \mathcal{I}_{2h}^{(2r)}u(P_i) = \mathcal{I}_h u(P_i) = u(P_i), \quad u \in C(\bar{\Omega}_0), \tag{5.21}$$

where T_0^h is the set of nodal points of J^h restricted to $\overline{\Omega}_0$;

$$\|u - \mathcal{I}_{2h}^{(2r)}u\|_{m,p,E} \le Ch^{2r+1-m}\|u\|_{2r+1,p,E} \tag{5.22}$$

$$\forall u \in W^{2r+1,p}(E), \quad m = 0, 1,\ 1 \le p \le +\infty,\ \forall E \in J^{2h}|_{\Omega_0}.$$

THEOREM 5.4. *Assume that $(\tilde{\lambda}_k^{(0)}, \tilde{u}_k^0(x))$ is the kth eigenpair of problem (4.3) and $(\tilde{\lambda}_{k,h}^{(0)}, \tilde{u}_{k,h}^0(x))$ is its finite element approximate solution in $S_0^h(\Omega)$, and let $\Omega_0 \subset\subset \Omega' \subset\subset \Omega$ and let Ω' be covered by uniform rectangular meshes; then it holds that*

$$\|\tilde{u}_k^0(x) - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0(x)\|_{0,\Omega_0} + h\|\tilde{u}_k^0(x) - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0(x)\|_{1,\Omega_0} \le Ch^{r+2}, \tag{5.23}$$

where $C > 0$ is independent of $h, h_0, r \ge 2,\ k = 1, 2, \ldots$.

Proof. It follows from Lemma 5.3 that

$$\begin{aligned}
&\|\mathcal{I}_{2h}^{(2r)}\tilde{u}_k^0 - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0\|_{1,\Omega_0} = \|\mathcal{I}_{2h}^{(2r)}(\mathcal{I}_h\tilde{u}_k^0 - \tilde{u}_{k,h}^0)\|_{1,\Omega_0} \\
&\le C\|\mathcal{I}_h\tilde{u}_k^0 - \tilde{u}_{k,h}^0\|_{1,\Omega_0} \le Ch^{r+1}\|\tilde{u}_k^0\|_{r+2,\Omega_0} \\
&+C\|\tilde{u}_k^0 - \tilde{u}_{k,h}^0\|_{-s,\Omega'} \le Ch^{r+1}\|\tilde{u}_k^0\|_{r+2,\Omega'}.
\end{aligned}$$

Thus

$$\|\tilde{u}_k^0 - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0\|_{1,\Omega_0} \le \|\tilde{u}_k^0 - \mathcal{I}_{2h}^{(2r)}\tilde{u}_k^0\|_{1,\Omega_0} + \|\mathcal{I}_{2h}^{(2r)}\tilde{u}_k^0 - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0\|_{1,\Omega_0} \le Ch^{r+1}\|\tilde{u}_k^0\|_{r+2,\Omega'}.$$

The remainder can similarly be proved. ∎

5.2. Finite element computation of the boundary layer. In practice, we need to solve the modified boundary layer as follows:

(5.24)
$$\begin{cases} \mathcal{L}_\varepsilon \tilde{W}_k^\varepsilon(x) \equiv -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial \tilde{W}_k^\varepsilon}{\partial x_j}\right) + \left(b\left(\dfrac{x}{\varepsilon}\right) - \tilde{\lambda}_{k,h}^{(0)}\rho\left(\dfrac{x}{\varepsilon}\right)\right)\tilde{W}_k^\varepsilon(x) = 0, \quad x \in \Omega_1, \\ \mathcal{B}_\varepsilon \tilde{W}_k^\varepsilon(x) = 0, \quad x \in \partial\Omega, \\ \tilde{W}_k^\varepsilon(x) = \tilde{u}_{k,h}^0(x), \quad x \in \partial\Omega_0, \end{cases}$$

where $(\tilde{\lambda}_{k,h}^{(0)}, \tilde{u}_{k,h}^0(x))$ is the finite element solution associated with $(\tilde{\lambda}_k^{(0)}, \tilde{u}_k^0(x))$ in $S_0^h(\Omega)$.

For the sake of simplicity, here assume that $\mathcal{B}_\varepsilon \equiv I$ is a Dirichlet boundary operator.

Let $\mathcal{F}^{h_1} = \{e\}$ be a family of regular triangulations of subdomain $\Omega_1 = \Omega \setminus \overline{\Omega}_0$ as shown in Figure 4. Let $h_1 = \max_{e\in\mathcal{F}^{h_1}}\{h_e\}$, $\frac{h_1}{\varepsilon^2} << 1$.

Define a piecewise linear finite element space as

$$S_{h_1}^\varepsilon(\Omega_1) = \{v \in C(\overline{\Omega}_1) : \quad v|_e \in P_1(e), \quad v|_{\partial\Omega\cup\partial\Omega_0} = 0\}. \tag{5.25}$$

From (3.25) and (5.24), one can easily check that

$$\begin{cases} -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial (W_k^\varepsilon - \tilde{W}_k^\varepsilon)}{\partial x_j}\right) + \left(b\left(\dfrac{x}{\varepsilon}\right) - \tilde{\lambda}_k^{(0)}\rho\left(\dfrac{x}{\varepsilon}\right)\right)(W_k^\varepsilon - \tilde{W}_k^\varepsilon)(x) \\ \qquad = \rho\left(\dfrac{x}{\varepsilon}\right)\left(\lambda_k^{(0)} - \tilde{\lambda}_{k,h}^{(0)}\right)\tilde{W}_k^\varepsilon, \quad x \in \Omega_1, \\ W_k^\varepsilon - \tilde{W}_k^\varepsilon(x) = 0, \quad x \in \partial\Omega, \\ W_k^\varepsilon - \tilde{W}_k^\varepsilon(x) = u_k^0(x) - \tilde{u}_{k,h}^0(x), \quad x \in \partial\Omega_0. \end{cases} \tag{5.26}$$

THEOREM 5.5. *Let $W_k^\varepsilon(x)$, $\tilde{W}_k^\varepsilon(x)$ be the weak solutions of problems* (3.25) *and* (5.24)*, respectively, and let $\tilde{W}_{k,h_1}^\varepsilon(x)$ be the finite element solution of $\tilde{W}_k^\varepsilon(x)$ in $S_{h_1}^\varepsilon(\Omega_1)$; then it holds that*

$$\|W_k^\varepsilon(x) - \tilde{W}_{k,h_1}^\varepsilon(x)\|_{1,p,\Omega_1} \le C\left\{\left(\frac{h_1}{\varepsilon^2}\right) + h_0 + h^r\right\}, \tag{5.27}$$

where C is a constant independent of ε, h_0, h, h_1, $1 < p \le p_0 < +\infty$, and h_0, h, h_1 are the mesh sizes associated with Q, Ω, Ω_1, respectively.

Proof. It follows from Theorem 3.3 and Theorem 5.2 that

$$\begin{aligned} &\|\tilde{W}_k^\varepsilon(x) - \tilde{W}_{k,h_1}^\varepsilon(x)\|_{1,p,\Omega_1} \le Ch_1\|\tilde{W}_k^\varepsilon\|_{2,p,\Omega_1} \\ &\le C\cdot\frac{h_1}{\varepsilon^2}\|\tilde{u}_{k,h}^0\|_{2,p,\Omega} \le C\cdot\frac{h_1}{\varepsilon^2}\|\tilde{u}_k^0\|_{2,p,\Omega_1}. \end{aligned} \tag{5.28}$$

On the other hand, it follows from (5.26), Theorem 4.1, Corollary 4.1, Theorem 5.2, and Theorem 5.3 that

$$\begin{aligned} \|W_k^\varepsilon(x) - \tilde{W}_k^\varepsilon(x)\|_{1,p,\Omega_1} &\le C\{\|u_k^0(x) - \tilde{u}_{k,h}^0(x)\|_{1,p,\Omega_1} \\ &+|\lambda_k^{(0)} - \tilde{\lambda}_{k,h}^{(0)}|\|\tilde{W}_k^\varepsilon(x)\|_{0,p,\Omega_1}\} \le C\{h^r + h_0\}. \end{aligned} \tag{5.29}$$

Combining (5.28) with (5.29) and using the triangle inequality, one can easily obtain (5.27).

6. Multiscale finite element algorithm and the postprocessing technique. To begin with, let us introduce the first order difference quotient as follows:

$$\delta_{x_i}\tilde{u}^0_{k,h}(N_p)=\frac{1}{\tau(N_p)}\sum_{e\in\sigma(N_p)}\left[\frac{\partial\tilde{u}^0_{k,h}}{\partial x_i}\right]_e(N_p), \tag{6.1}$$

where $\sigma(N_p)$ is the set of elements with node N_p, $\tau(N_p)$ is the number of elements of $\sigma(N_p)$, $\tilde{u}^0_{k,h}(x)$ is the finite element solution of $\tilde{u}^0_k(x)$ in $S^h_0(\Omega)$, and $[\frac{\partial\tilde{u}^0_{k,h}}{\partial x_i}]_e(N_p)$ is the value of the derivative $\frac{\partial\tilde{u}^0_{k,h}}{\partial x_i}$ at node N_p relative to element e.

Analogously, define any higher order difference quotients as follows:

$$\delta^l_{x_{\alpha_1}\cdots x_{\alpha_l}}\tilde{u}^0_{k,h}(N_p)=\frac{1}{\tau(N_p)}\sum_{e\in\sigma(N_p)}\left[\sum_{j=1}^{d}\delta^{l-1}_{x_{\alpha_1}\cdots x_{\alpha_{l-1}}}\tilde{u}^0_{k,h}(P_j)\frac{\partial\psi_j}{\partial x_{\alpha_l}}\right]_e(N_p), \tag{6.2}$$

where d is the number of nodes in e, P_j are the nodes of e, $\psi_j(x)$ are Lagrange's shape functions, $j=1,2,\ldots,d$.

Now let us give the multiscale finite element computing scheme:

$$U^{\varepsilon,h_0,h}_{k,M,h_1}(N_p)=\begin{cases}\tilde{u}^0_{k,h}(N_p)+\displaystyle\sum_{l=1}^{M}\varepsilon^l\sum_{\langle\alpha\rangle=l}N^{h_0}_{\alpha}(\xi(N_p))\delta^l_{x_{\alpha_1}\cdots x_{\alpha_l}}\tilde{u}^0_{k,h}(N_p), & N_p\in\overline{\Omega}_0,\\ W^{\varepsilon}_{k,h_1}(N_p), & N_p\in\Omega_1,\end{cases} \tag{6.3}$$

where the integer $M\geq 2$ and h_0,h,h_1 are the mesh parameters of Q, Ω, Ω_1, respectively.

The postprocessing computing scheme is presented for obtaining high accuracy:

$$\mathcal{P}U^{\varepsilon,h_0,h}_{k,M,h_1}(x)=\begin{cases}\mathcal{I}^{(2r)}_{2h}\tilde{u}^0_{k,h}(x)+\displaystyle\sum_{l=1}^{M}\varepsilon^l\sum_{\langle\alpha\rangle=l}N^{h_0}_{\alpha_1\cdots\alpha_l}(\xi)\delta^l_{x_{\alpha_1}\cdots x_{\alpha_l}}\mathcal{I}^{(2r)}_{2h}\tilde{u}^0_{k,h}(x), & x\in\overline{\Omega}_0,\\ \tilde{W}^{\varepsilon}_{k,h_1}(x), & x\in\Omega_1.\end{cases} \tag{6.4}$$

Finally, we will give the total error estimations.

THEOREM 6.1. *Let Ω be either a bounded smooth or a convex polygonal domain. With all assumptions as indicated above, it then holds that*

$$\|u^{\varepsilon}_k(x)-\mathcal{P}U^{\varepsilon,h_0,h}_{k,M,h_1}(x)\|_{0,\Omega_0}\leq C(\varepsilon^{M-1}+h_0^2+h^{2r-M}), \tag{6.5}$$

$$\|u^{\varepsilon}_k(x)-\tilde{W}^{\varepsilon}_{k,h_1}(x)\|_{0,p,\Omega_1}\leq C\left\{\varepsilon^{M-1}+h_0^2+h+\left(\frac{h_1}{\varepsilon^2}\right)\right\}, \tag{6.6}$$

where $C>0$ is a constant independent of ε, h_0, h, h_1; $\Omega_0\subset\subset\Omega$ is the union of periodic cells, $\Omega_1=\Omega\setminus\overline{\Omega}_0$; r denotes the degree of piecewise polynomials in $S^h_0(\Omega)$; h_0, h, h_1 are the mesh parameters of Q, Ω, Ω_1, respectively; and $M\geq 2$, $2r\geq M+1$, $0<h_1<<\varepsilon^2$, $1<p\leq p_0<+\infty$.

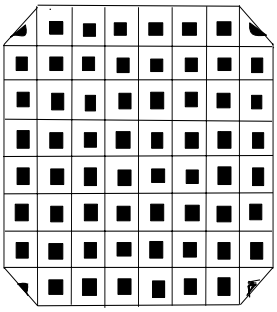


FIG. 7. *Domain* Ω.

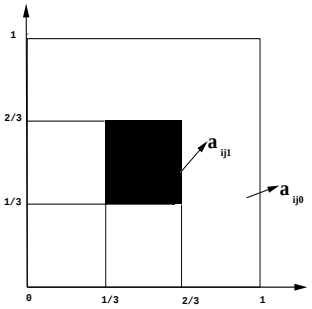


FIG. 8. *Unit cell* $Q = [0,1]^2$.

Proof. For $x \in \overline{\Omega}_0$, from (5.27), (3.25), (3.21), and (4.3), we obtain

$$u_k^\varepsilon(x) - \mathcal{P}U_{k,M,h_1}^{\varepsilon,h_0,h}(x) = u_k^\varepsilon(x) - u_k^{\varepsilon,M}(x) + u_k^{\varepsilon,M}(x) - \mathcal{P}U_{k,M,h_1}^{\varepsilon,h_0,h}(x)$$

$$= u_k^\varepsilon(x) - u_k^{\varepsilon,M}(x) + u_k^0(x) - \tilde{u}_k^0(x) + \tilde{u}_k^0(x) - \mathcal{I}_{2h}^{(2r)}\tilde{u}_{k,h}^0(x)$$

$$+ \sum_{l=1}^{M} \varepsilon^l \sum_{|\alpha|=l} (N_\alpha(\xi) - N_\alpha^{h_0}(\xi)) D^\alpha u_k^0(x) \tag{6.7}$$

$$+ \sum_{l=1}^{M} \varepsilon^l \sum_{|\alpha|=l} N_\alpha^{h_0}(\xi) D^\alpha (u_k^0(x) - \tilde{u}_k^0(x))$$

$$+ \textstyle\sum_{l=1}^{M} \varepsilon^l \sum_{|\alpha|=l} N_\alpha^{h_0}(\xi)(D^\alpha \tilde{u}_k^0(x) - \delta^l_{x_{\alpha_1}\cdots x_{\alpha_l}} \mathcal{I}_{2h}^{(2r)} \tilde{u}_{k,h}^0(x)).$$

It follows from Theorem 3.1, Theorem 4.1, Theorem 5.4, Proposition 4.1, and Theorem 4.2 that

$$\|u_k^\varepsilon(x) - \mathcal{P}U_{k,M,h_1}^{\varepsilon,h_0,h}(x)\|_{0,\Omega_0} \le C(\varepsilon^{M-1} + h_0^2 + h^{2r-M}).$$

On the other hand, for $x \in \Omega_1$, we have

$$u_k^\varepsilon(x) - \tilde{W}_{k,h_1}^\varepsilon(x) = u_k^\varepsilon(x) - W_k^\varepsilon(x) + W_k^\varepsilon(x) - \tilde{W}_{k,h_1}^\varepsilon(x). \tag{6.8}$$

It follows from Theorem 3.1 and Theorem 5.5 that

$$\|u_k^\varepsilon(x) - \tilde{W}_{k,h_1}^\varepsilon(x)\|_{0,p,\Omega_1} \le C\left\{\varepsilon^{M-1} + h_0^2 + h^r + \left(\frac{h_1}{\varepsilon^2}\right)\right\},$$

where $1 < p \le p_0 < +\infty$, $\Omega_1 = \Omega \setminus \overline{\Omega}_0$. ∎

7. Numerical results. In this section, we only consider the first eigenvalue and eigenfunction.

EXAMPLE 7.1. *We consider the following Helmholtz equation:*

$$\begin{cases} \mathcal{L}_\varepsilon u^\varepsilon \equiv -\dfrac{\partial}{\partial x_i}\left(a_{ij}\left(\dfrac{x}{\varepsilon}\right)\dfrac{\partial u^\varepsilon}{\partial x_j}\right) + b\left(\dfrac{x}{\varepsilon}\right)u^\varepsilon(x) = \lambda^\varepsilon \rho\left(\dfrac{x}{\varepsilon}\right)u^\varepsilon(x) \quad in \quad \Omega, \\ u^\varepsilon(x) = 0 \quad on \quad \partial\Omega, \end{cases} \tag{7.1}$$

where Ω *is as shown in Figure* 7*—note that* Ω *is not entire periodic domain. The unit cell* Q *is shown in Figure* 8, $\varepsilon = \frac{1}{8}$.

Case 1. $a_{ij0} = \delta_{ij}$, $a_{ij1} = 10.0\delta_{ij}$, $b(\frac{x}{\varepsilon}) = 30$, $\rho(\frac{x}{\varepsilon}) = 1$.
Case 2. $a_{ij0} = \delta_{ij}$, $a_{ij1} = \frac{1}{1000.0}\delta_{ij}$, $b(\frac{x}{\varepsilon}) = 30$, $\rho(\frac{x}{\varepsilon}) = 1$.
Case 3. $a_{ij0} = \delta_{ij}$, $a_{ij1} = \frac{1}{200.0}\delta_{ij}$, $b(\frac{x}{\varepsilon}) = 30$, $\rho(\frac{x}{\varepsilon}) = 1$.
Case 4. $a_{ij0} = \delta_{ij}$, $a_{ij1} = 200.0\delta_{ij}$, $b(\frac{x}{\varepsilon}) = 30$, $\rho(\frac{x}{\varepsilon}) = 1$.

Tables 1–3 show some numerical results. Where $e_0 = u^\varepsilon - u^0$, $e_1 = u^\varepsilon - U_1^\varepsilon$, $e_2 = u^\varepsilon - U_2^\varepsilon$, $u^0(x)$ is the finite element solution of the first eigenfunction for the homogenized Helmholtz equation, and $U_1^\varepsilon(x)$, $U_2^\varepsilon(x)$ are the first order and the second order multiscale finite element solutions calculated by multiscale finite element formulation (6.3), respectively. $u^\varepsilon(x)$, $U_2^\varepsilon(x)$, $U_1^\varepsilon(x)$, $e_2(x)$ are shown in Figure 9 and Figure 10.

TABLE 1
Compare with computational amount.

	Original equation	Unit cell	Homogenized equation	Boundary layer
Elements	17856	1296	4464	1872
Nodes	9097	1369	2317	1092

TABLE 2
Comparison of computation results, I. *Eigenvalues.*

	Original equation	Homogenized equation	$\frac{\lambda^\varepsilon - \lambda^0}{\lambda^\varepsilon}$
Case 1	54.3032	54.1539	2.749×10^{-3}
Case 2	41.6690	45.6433	-9.5378×10^{-2}
Case 3	45.5082	45.6782	-3.7356×10^{-3}
Case 4	55.6577	55.3892	4.824×10^{-3}

TABLE 3
Comparison of computation results, II. *Eigenfunctions.*

	$\frac{\|e_0\|_{L^2}}{\|u^0\|_{L^2}}$	$\frac{\|e_1\|_{L^2}}{\|U_1^\varepsilon\|_{L^2}}$	$\frac{\|e_2\|_{L^2}}{\|U_2^\varepsilon\|_{L^2}}$	$\frac{\|e_0\|_{H^1}}{\|u^0\|_{H^1}}$	$\frac{\|e_1\|_{H^1}}{\|U_1^\varepsilon\|_{H^1}}$	$\frac{\|e_2\|_{H^1}}{\|U_2^\varepsilon\|_{H^1}}$
Case 1	0.024956	0.009094	0.022298	0.074006	0.067905	0.173193
Case 2	0.871181	0.870530	0.175682	0.972558	2.892918	0.181331
Case 3	0.292920	0.296389	0.030354	0.334210	1.052346	0.132701
Case 4	0.032661	0.004042	0.809928	0.097027	0.061860	0.988662

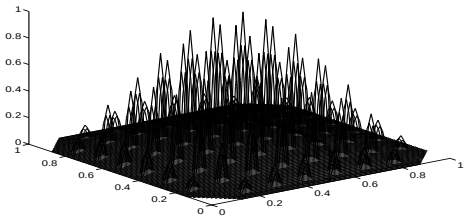


FIG. 9A. *Case* 2. *solution* u^ε.

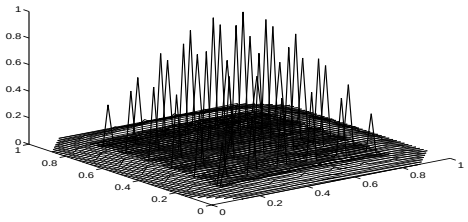


FIG. 9B. *Case* 2. *MFEM* U_2^ε.

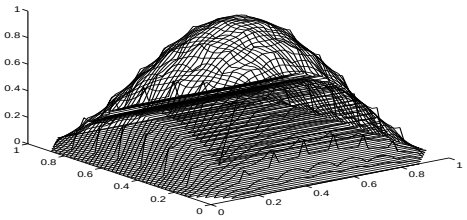


FIG. 9C. *Case* 2. *MFEM* U_1^ε.

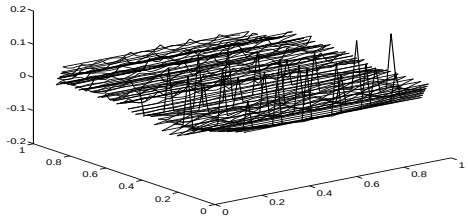


FIG. 9D. *Case* 2. $e_2(x) = u^\varepsilon(x) - U_2^\varepsilon$.

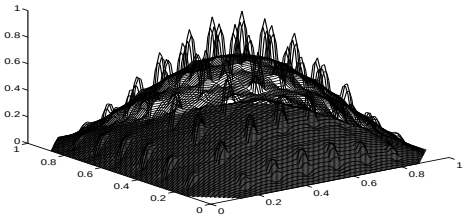


FIG. 10A. *Case* 3. *solution* u^ε.

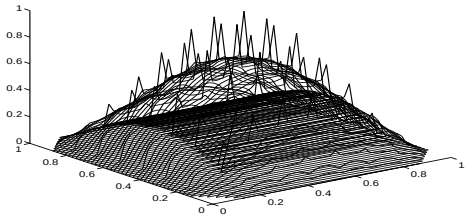


FIG. 10B. *Case* 3. *MFEM* U_2^ε.

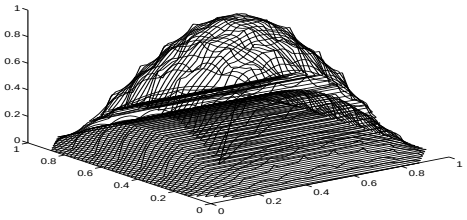


FIG. 10C. *Case* 3. *MFEM* U_1^ε.

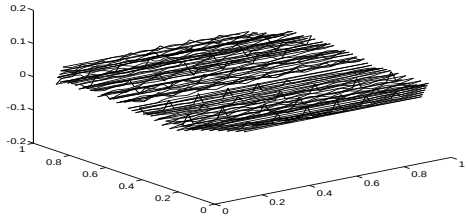


FIG. 10D. *Case* 3. $e_2(x) = u^\varepsilon(x) - U_2^\varepsilon(x)$.

Remark 7.1. It is worthwhile to notice that, in Case 1 and Case 4, errors e_2 for the second order method are larger than the error e_1 for the first order method, respectively. As we can judge, from the viewpoint of numerical computation, the reason is that if e_1 have had the better accuracy, then e_2 might become worse due to adding other items.

Concluding remarks. First of all, in this paper the main objectives are to obtain the multiscale asymptotic analysis formulas of eigenvalues and eigenfunctions of second order Helmholtz equation with rapidly oscillating coefficients over general bounded Lipschitz convex domains and to propose the multiscale finite element computing schemes and the postprocessing technique. Meanwhile, we derive their rigorous verifications.

Second, in solving the Helmholtz problems of composite media with a periodic structure by using multiscale numerical method, our work consists of the following parts:

(1) Compute the periodic solutions $N_{\alpha_1\cdots\alpha_l}(\xi)$, $l \geq 1$, in the unit cell Q.

(2) Solve the homogenized Helmholtz equation with constant coefficients over the whole domain Ω.

(3) Solve the boundary layer.

(4) Compute the higher order partial derivatives for eigenfunctions $u_k^0(x)$.

Third, from the viewpoint of mechanics, the work of this paper tell us some important facts:

(a) In the sense of given accuracy, the natural frequencies of the homogenized problem and the original linear problem are the same.

(b) According to the superposition priciple, the natural vibration modes consist of two parts: the natural vibration modes of the homogenized problem reflect the macroscopic behavior, and the periodic solutions $N_{\alpha_1\cdots\alpha_l}(\xi)$, $l \geq 1$, $\alpha_j = 1, \ldots, n$ describe the mesoscopic fluctuations of the natural vibration modes.

Finally, we would like to say that the method proposed in this paper can easily be extended into the elastic structures of composite materials with a small period. In the viewpoint of numerical computation, it is suitable for subdivided periodic structures and some random structures. We will discuss these problems in other papers.

Appendix A. The equivalence of two kinds of homogenization methods. In [2], we know that $\tilde{N}_{\alpha_1}(\xi)$ is defined in the following way:

$$\begin{cases} \dfrac{\partial}{\partial\xi_k}\left(a_{kj}(\xi)\dfrac{\partial\tilde{N}_{\alpha_1}(\xi)}{\partial\xi_j}\right) = -\dfrac{\partial}{\partial\xi_k}(a_{k\alpha_1}(\xi)) \quad \text{in} \quad R^n, \\ \tilde{N}_{\alpha_1}(\xi) \quad \text{is 1-periodic in } \xi, \\ \displaystyle\int_Q \tilde{N}_{\alpha_1}(\xi)d\xi = 0. \end{cases} \tag{A.1}$$

Define

$$V_{per} = \{v \in H^1(R^n) : \quad v(\xi) \quad \text{is 1-periodic in } \xi\}.$$

The variational form is the following:

$$\int_Q a_{kj}(\xi)\frac{\partial\tilde{N}_{\alpha_1}(\xi)}{\partial\xi_j}\frac{\partial\tilde{v}(\xi)}{\partial\xi_k}d\xi = -\int_Q a_{k\alpha_1}(\xi)\frac{\partial\tilde{v}(\xi)}{\partial\xi_k}d\xi,$$

i.e.,

$$\int_Q \left(a_{k\alpha_1}(\xi) + a_{kj}(\xi)\frac{\partial\tilde{N}_{\alpha_1}(\xi)}{\partial\xi_j}\right)\frac{\partial\tilde{v}(\xi)}{\partial\xi_k}d\xi = 0 \quad \forall \quad \tilde{v}(\xi) \in V_{per}. \tag{A.2}$$

On the other hand, from (3.12) we know that

$$\int_Q \left(a_{k\alpha_1}(\xi) + a_{kj}(\xi)\frac{\partial N_{\alpha_1}(\xi)}{\partial\xi_j}\right)\frac{\partial v(\xi)}{\partial\xi_k}d\xi = 0 \quad \forall \quad v(\xi) \in H_0^1(Q). \tag{A.3}$$

For the sake of simplicity, let $\tilde{v}(\xi) = e^{i2\pi m\cdot\xi}$ (in practice, we should choose $\tilde{v}(\xi) = \cos 2\pi m\cdot\xi, \sin 2\pi m\cdot\xi$), $v^{(k)}(\xi) = e^{i2\pi m\cdot\xi} - e^{i2\pi m\cdot\tilde{\xi}^{(k)}}$, where $m = (m_1, \ldots, m_k, \ldots, m_n) \in Z^n$, $\xi = (\xi_1, \ldots, \xi_k, \ldots, \xi_n)^T$, $\tilde{\xi}^{(k)} = (\xi_1, \ldots, \xi_{k-1}, 0, \xi_{k+1}, \ldots, \xi_n)^T$. One can directly check that $\tilde{v}(\xi) \in V_{per}$, $v^{(k)}(\xi) \in H_0^1(Q)$.

Set $\tilde{\Lambda}_{k\alpha_1}(\xi) = (a_{k\alpha_1}(\xi) + \sum_{j=1}^{n} a_{kj}(\xi)\frac{\partial \tilde{N}_{\alpha_1}(\xi)}{\partial \xi_j})$, $\Lambda_{k\alpha_1}(\xi) = (a_{k\alpha_1}(\xi) + \sum_{j=1}^{n} a_{kj}(\xi)\frac{\partial N_{\alpha_1}(\xi)}{\partial \xi_j}), \Theta_{k\alpha_1}(\xi) = \tilde{\Lambda}_{k\alpha_1}(\xi) - \Lambda_{k\alpha_1}(\xi)$.

Substituting $\tilde{v}(\xi), v^{(k)}(\xi)$ into (A.2), (A.3), respectively, one can obtain

$$\int_Q \tilde{\Lambda}_{k\alpha_1}(\xi) e^{i2\pi m\cdot\xi} d\xi = 0, \tag{A.4}$$

$$\int_Q \Lambda_{k\alpha_1}(\xi) e^{i2\pi m\cdot\xi} d\xi = 0, \tag{A.5}$$

i.e.,

$$\int_Q \Theta_{k\alpha_1}(\xi) e^{i2\pi m\cdot\xi} d\xi = 0 \quad \forall m \in Z^n. \tag{A.6}$$

Let

$$\Theta_{k\alpha_1}(\xi) = \sum_{q_1\cdots q_n=-\infty}^{+\infty} \widehat{\Theta}_{k\alpha_1}(q)\cdot e^{-i2\pi q\cdot\xi}. \tag{A.7}$$

Substituting (A.7) into (A.6), we obtain

$$\sum_{q_1,\cdots q_n=-\infty}^{+\infty} \widehat{\Theta}_{k\alpha_1}(q) \int_Q e^{-i2\pi(q-m)\cdot\xi} d\xi = 0. \tag{A.8}$$

Hence

$$\widehat{\Theta}_{k\alpha_1}(m) = 0 \quad \forall m \in Z^n.$$

This implies that

$$\Theta_{k\alpha_1}(\xi) = 0.$$

Therefore

$$\begin{aligned} \hat{a}_{k\alpha_1} &= \int_Q \left(a_{k\alpha_1}(\xi) + \sum_{j=1}^{n} a_{kj}(\xi)\frac{\partial N_{\alpha_1}(\xi)}{\partial \xi_j} \right) d\xi \\ &= \int_Q \left(a_{k\alpha_1}(\xi) + \sum_{j=1}^{n} a_{kj}(\xi)\frac{\partial \tilde{N}_{\alpha_1}(\xi)}{\partial \xi_j} \right) d\xi = \hat{\tilde{a}}_{k\alpha_1}, \quad k, \alpha_1 = 1, 2, \ldots, n. \end{aligned} \tag{A.9}$$

Appendix B. Two useful properties of the first eigenvalue and eigenfunction for the homogenized Helmholtz equation in some cases.

LEMMA B.1 (see [5]). *Assume that Ω is a bounded smooth domain, and let $(\lambda_1^{(0)}, u_1^0(x))$ be the first eigenpair of the homogenized Helmholtz problem (3.21); then it holds that $u_1^0(x)$ is a smooth function in Ω such that $u_1^0(x) \neq 0$ in Ω and $|\nabla u_1^0(x)| \neq 0$ in a neighborhood of $\partial\Omega$.*

PROPOSITION B.1. *Under the assumptions of Lemma B.1, one can prove that the first eigenvalue $\lambda_1^{(0)}$ of problem (3.21) is simple, and the corresponding eigenfunction $u_1^0(x)$ has a constant sign in Ω and is unique up to a constant factor.*

Proof. Set

$$V = \left\{ u \in H^1(\Omega), \langle \rho \rangle \int_\Omega |u(x)|^2 dx = 1 \right\},$$

$$D(u) = \int_\Omega \left(\hat{a}_{ij} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} + \langle b \rangle u^2 \right) dx. \tag{B.1}$$

Let $\lambda_1^{(0)} = \inf_{u \in V} D(u)$, $u_1^0(x)$ be a function giving the minimal value. Since $|u_1^{(0)}(x)| \in V$, if $u_1^0(x) \in V$, and $D(|u_1^0|) = D(u_1^0)$, then the function $|u_1^0(x)|$ is an eigenfunction corresponding to the eigenvalue $\lambda = \lambda_1^{(0)}$. It follows from Lemma B.1 that $|u_1^0(x)|$ does not vanish anywhere in Ω.

Next let $\bar{u}_1^0(x)$ be another eigenfunction corresponding to $\lambda_1^{(0)}$. Using Schmidt's technique and substituting $\bar{u}_1^0(x)$ by the function $\hat{u}_1^0(x) = \bar{u}_1^0(x) + \tau \cdot u_1^0(x)$, we can have that the function $\hat{u}_1^0(x)$ is orthogonal to $u_1^0(x)$; i.e., $\int_\Omega \hat{u}_1^0(x) \cdot u_1^0(x) dx = 0$ and $\langle \rho \rangle \int_\Omega |\hat{u}_1^0(x)|^2 dx = 1$. Since $u_1^0(x) > 0$ in Ω, the function $\hat{u}_1^0(x)$ changes its sign in Ω. However, $|\hat{u}_1^0(x)|$ is an eigenfunction with respect to $\lambda_1^{(0)}$, and $|\hat{u}_1^0(x)|$ vanishes at an inner point of Ω; it follows from Lemma B.1 that $\hat{u}_1^0(x) \equiv 0$.

The proof of Proposition B.1 is complete. ∎

PROPOSITION B.2. *Under the assumptions of Proposition* B.1*, if we consider the first eigenvalue* $\lambda_1^{(0)}$ *of* (3.21)*, then it holds that*

$$\left(\lambda_1^{(0)} \right)^{-1} \notin \sigma_d(\mathcal{K}_\varepsilon). \tag{B.2}$$

Proof. Given $\Omega_1 \subset\subset \Omega$ and $meas(\Omega \setminus \overline{\Omega}_1) = meas(\Omega_0) > 0$, denote by $\lambda_1^{(0)}(\Omega)$, $\tilde{\lambda}_1^{\varepsilon}(\Omega_1)$ the first eigenvalues associated with problem (3.21) and operator $\mathcal{Q}_\varepsilon$ of (3.27), respectively. The variational principle implies that $\lambda_1^{(0)}(\Omega) \le \tilde{\lambda}_1^{\varepsilon}(\Omega_1)$. Suppose that $\lambda_1^0(\Omega) = \tilde{\lambda}_1^{\varepsilon}(\Omega_1) = \lambda$. Then the eigenfunction corresponding to $\mathcal{Q}_\varepsilon$ with eigenvalue λ expanded by zero values on $\Omega \setminus \Omega_1$ is an eigenfunction in Ω. However, it vanishes at some points of Ω, contrary to the result of Lemma B.1.

The proof of Proposition B.2 is complete. ∎

Appendix C. Regularity results of the solution for the boundary layer. For the sake of simplicity, now we consider only 2-D problems; for the detailed discussion of three-dimensional problems, we refer to [10].

To begin with, consider the following boundary value problems over concave domain $\Omega_1 \subset R^2$, as shown in Figure 2:

$$\begin{cases} -\triangle u = f(x) & \text{in} \quad \Omega_1, \\ u(x) = 0 & \text{on} \quad \partial\Omega_1. \end{cases} \tag{C.1a}$$

Let $\{\sigma_j\}_{j=1}^N$ denote the angular points of Ω_1 and $\beta_j \pi$, $j = 1, \ldots, N$, are the corresponding internal angles, i.e.,

$$\beta_1 \le \beta_2 \le \cdots \le \beta_N, \quad \gamma_j = \frac{1}{\beta_j}.$$

It is obvious that $1 < \beta_N \le 2$, $\frac{1}{2} \le \gamma_N < 1$. Suppose that

$$V_j = \{x \in \Omega_1 : |x - \sigma_j| < r_j\}, \quad j = 1, \ldots, N, \tag{C.1b}$$

satisfy

$$V_i \cap V_j = \emptyset, \quad V_0 = \Omega_1 \setminus \cup_{j=1}^{N} \overline{V}_j. \tag{C.1c}$$

LEMMA C.1 (see [10]). *Suppose that u is the unique solution of problem* (C.1a); *if $f \in L^2(\Omega)$, then it holds that*

$$u(x) = \sum_{j=1}^{N} c_j(f) u_j + U(x), \tag{C.2}$$

where $U(x) \in H^2(\Omega_1) \cap H_0^1(\Omega_1)$, $\|U\|_2 \le C\|f\|_0$, and the constants $c_j(f)$ satisfy $|c_j(f)| \le C\|f\|_0$.

Note that u_j are some functions independent of f, u that satisfy the following conditions:

(B_1) *If $\gamma_j > 1$, then $u_j(x) \equiv 0$. In addition, $u_j(x) \equiv 0$ outside of V_j.*

(B_2) *If $\frac{1}{2} < \gamma_j < 1$, then there exists the following formula in a neighborhood of σ_j:*

$$u_j = \rho^{\gamma_j} \sin \gamma_j \theta \quad \text{if} \quad (\rho, \theta) \in V_j, \tag{C.3}$$

where $V_j = \{(\rho, \theta) : 0 < \rho < r_j, 0 < \theta < \beta_j \pi\}$.

Remark C.1. By virtue of (C.3), one can easily show that

$$|D^k u| \le C\rho^{\gamma_j - |k|} \tag{C.4}$$

in a neighborhood of σ_j.

Let us turn to the proof of Theorem 3.3. It follows from the finite covering theorem that there exist finite points $P_1, \ldots, P_s$ and the corresponding neighborhoods $\mathcal{O}_l$, $l = 1, \ldots, s$, such that

(i) $\cup_{l=1}^{s} \mathcal{O}_l \supset \overline{\Omega}_1$;

(ii) $diam(\mathcal{O}_l) \le \varepsilon R_0$, R_0 will be chosen later;

(iii) $\mathcal{I}_i = \{j : \mathcal{O}_j \cap \mathcal{O}_i \ne \emptyset, \}$, $\sigma(\mathcal{I}_i) \le s_0$, where $\sigma(\mathcal{I}_i)$ denote the number of elements in $\mathcal{I}_i$, $i = 1, \ldots, t$ and s_0 is a constant.

From the partition of unity theorem, there exist $\phi_l \in C_0^\infty(R^n)$, $l = 1, \ldots, s$, such that $0 \le \phi_l \le 1, \text{supp}\phi_l \subset \mathcal{O}_l$, and

$$\sum_{l=1}^{s} \phi_l \equiv 1 \quad \text{in} \quad \Omega_1.$$

Let $\mathcal{A}_\varepsilon \cdot = -\frac{\partial}{\partial x_i}(a_{ij}(\frac{x}{\varepsilon})\frac{\partial}{\partial x_j})\cdot$, $W^\varepsilon = \sum_{l=1}^{s} W_l^\varepsilon$, $W_l^\varepsilon = \phi_l \cdot W^\varepsilon$,

$$\mathcal{A}_\varepsilon W_l^\varepsilon = \phi_l \cdot \mathcal{A}_\varepsilon W^\varepsilon + \eta_l, \tag{C.5}$$

where

$$\begin{aligned} \eta_l = & -\frac{\partial \phi_l}{\partial x_j}\left[\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\right) W^\varepsilon + a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial W^\varepsilon}{\partial x_j}\right] \\ & -a_{ij}\left(\frac{x}{\varepsilon}\right)\left[\frac{\partial \phi_l}{\partial x_i}\frac{\partial W^\varepsilon}{\partial x_j} + W^\varepsilon \frac{\partial^2 \phi_l}{\partial x_i \partial x_j}\right] \end{aligned} \tag{C.6}$$

and

$$\|\eta_l\|_{0,p,\mathcal{O}_l\cap\Omega_1} \le C\frac{1}{\varepsilon^2}\|W^\varepsilon\|_{1,p,\mathcal{O}_l\cap\Omega_1}. \tag{C.7}$$

$\forall R>0$, let

$$\omega_\varepsilon(R)=\max_{i,j}\max_{|x-x'|<\varepsilon R}\left|a_{ij}\left(\frac{x}{\varepsilon}\right)-a_{ij}\left(\frac{x'}{\varepsilon}\right)\right|,\quad x,x'\in\Omega_1.$$

For any fixed $x_0\in\mathcal{O}_l\cap\Omega_1$, set $A^\varepsilon=(a_{ij}(\frac{x_0}{\varepsilon}))$; it follows from (A_3) that there exists a orthogonal matrix T such that

$$TA^\varepsilon T'=\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}=D,$$

where T' denotes the transpose of matrix T.

From (A_2), we know $\lambda_i\ge\sigma>0$, $i=1,2$, and let $B=\ D^{-1/2}T$; then $BA^\varepsilon B'=I$ and $\|B\|\le\sqrt{\|D^{-1}\|}\le\sqrt{\frac{1}{\sigma}}$,

$$\|B^{-1}\|=\|D^{1/2}\|\le\sum_{i=1}^{2}\lambda_i=\sum_i a_{ii}\left(\frac{x_0}{\varepsilon}\right)\le M_0,$$

where M_0 is a positive constant independent of ε.

If we let $\hat{\mathcal{O}}_l=B(\mathcal{O}_l\cap\Omega_1)$, then $\hat v(y)=v(B^{-1}y)\in W^{2,p}(\hat{\mathcal{O}}_l)$ for any $v\in W^{2,p}(\mathcal{O}_l\cap\Omega_1)$, where p will be stated below. We have that

$$C\|v\|_{2,p,\mathcal{O}_l\cap\Omega_1}\le\|\hat v\|_{2,p,\hat{\mathcal{O}}_l}\le C'\|v\|_{2,p,\mathcal{O}_l\cap\Omega_1}. \tag{C.8}$$

Let

$$\begin{aligned}g(x)&=-a_{ij}\left(\frac{x_0}{\varepsilon}\right)\frac{\partial^2 W_l^\varepsilon}{\partial x_i\partial x_j}=-\left(a_{ij}\left(\frac{x_0}{\varepsilon}\right)-a_{ij}\left(\frac{x}{\varepsilon}\right)\right)\frac{\partial^2 W_l^\varepsilon}{\partial x_i\partial x_j}-a_{ij}\left(\frac{x}{\varepsilon}\right)\frac{\partial^2 W_l^\varepsilon}{\partial x_i\partial x_j}\\ &=-\left(a_{ij}\left(\frac{x_0}{\varepsilon}\right)-a_{ij}\left(\frac{x}{\varepsilon}\right)\right)\frac{\partial^2 W_l^\varepsilon}{\partial x_i\partial x_j}+\phi_l(x)\left(\lambda^0\rho\left(\frac{x}{\varepsilon}\right)-b\left(\frac{x}{\varepsilon}\right)\right)\cdot W^\varepsilon(x)\\ &\quad+\eta_l(x)+\frac{\partial}{\partial x_i}\left(a_{ij}\left(\frac{x}{\varepsilon}\right)\right)\frac{\partial W_l^\varepsilon}{\partial x_j}.\end{aligned}$$

From (C.7), we obtain

$$\|g\|_{0,p,\mathcal{O}_l\cap\Omega_1}\le\omega_\varepsilon(R)\|W_l^\varepsilon\|_{2,p,\mathcal{O}_l\cap\Omega_1}+C\frac{1}{\varepsilon^2}\|W^\varepsilon\|_{1,p,\mathcal{O}_l\cap\Omega_1}. \tag{C.9}$$

On the other hand, set $\hat W_l^\varepsilon(y)=W_l^\varepsilon(B^{-1}y)$, $\quad\hat g(y)=g(B^{-1}y)$; then

$$a_{ij}\left(\frac{x_0}{\varepsilon}\right)\frac{\partial^2 W_l^\varepsilon}{\partial x_i\partial x_j}=\triangle\hat W_l^\varepsilon(y),\quad y=Bx.$$

Thus

$$\triangle\hat W_l^\varepsilon(y)=\hat g(y).$$

It follows from Lemma C.1 that

$$\|\hat{W}_l^\varepsilon\|_{2,p,\hat{\mathcal{O}}_l} \le C(p)\|\triangle \hat{W}_l^\varepsilon\|_{0,p,\hat{\mathcal{O}}_l} \le C(p)\|\hat{g}\|_{0,p,\hat{\mathcal{O}}_l}, \tag{C.10}$$

where $1 < p \le p_0 = \frac{2\beta_N}{2\beta_N - 1} < +\infty$ and β_N is the maximum internal angle of $B\Omega_1$.

From (C.8), (C.9), and (C.10), one obtains

$$\|W_l^\varepsilon\|_{2,p,\mathcal{O}_l\cap\Omega_1} \le C(p)\left\{\omega_\varepsilon(R)\|W_l^\varepsilon\|_{2,p,\mathcal{O}_l\cap\Omega_1} + \frac{1}{\varepsilon^2}\|W^\varepsilon\|_{1,p,\mathcal{O}_l\cap\Omega_1}\right\}.$$

Since $a_{ij}(\frac{x}{\varepsilon}) \in C(\overline{\Omega})$, $\nabla_\xi a_{ij}(\xi) \in L^\infty(\Omega)$; then there exists a constant $R_0 > 0$ such that

$$\omega_\varepsilon(R) < \frac{1}{3C(p)} \quad \text{for} \quad 0 < R < R_0.$$

Hence,

$$\begin{aligned}\|W_l^\varepsilon\|_{2,p,\mathcal{O}_l\cap\Omega_1} &\le C(p)\varepsilon^{-2}\{\|W_l^\varepsilon\|_{1,p,\mathcal{O}_l\cap\Omega_1} + \|u^0\|_{2,p,\mathcal{O}_l\cap\Omega_1}\}\\ &\le C(p)\varepsilon^{-2}\|u^0\|_{2,p,\mathcal{O}_l\cap\Omega_1}.\end{aligned}$$

Therefore

$$\begin{aligned}\|W^\varepsilon\|_{2,p,\Omega_1} &= \left\|\sum_{l=1}^s W_l^\varepsilon\right\|_{2,p,\Omega_1} \le \sum_{l=1}^s \|W_l^\varepsilon\|_{2,p,\mathcal{O}_l\cap\Omega_1}\\ &\le C(p)\varepsilon^{-2}\|u^0\|_{2,p,\Omega}.\end{aligned}$$

Appendix D. The difference between the eigenvalues and eigenfunctions of the homogenized Helmholtz equation (3.21) and those of the modified homogenized Helmholtz equation (4.3). Here we formulate some results in the spectral theory of linear abstract operators, which are useful for applications considered below.

Let $\mathcal{H}_\tau$, $0 < \tau \le 1$, be a family of Hilbert spaces with scalar products $(u, v)_{\mathcal{H}_\tau}$, and let $\mathcal{H}_0$ be a Hilbert space with a scalar product $(u, v)_{\mathcal{H}_0}$. Consider bounded linear operators $\mathcal{A}_\tau : \mathcal{H}_\tau \to \mathcal{H}_\tau$, $\mathcal{A}_0 : \mathcal{H}_0 \to \mathcal{H}_0$. We assume that spaces $\mathcal{H}_\tau$, $\mathcal{H}_0$ and operators $\mathcal{A}_\tau$, $\mathcal{A}_0$ are subject to the following conditions:

(I) There exist continuous linear operators $\mathcal{R}_\tau : \mathcal{H}_0 \to \mathcal{H}_\tau$ such that

$$\|\mathcal{R}_\tau u\|_{\mathcal{H}_\tau} \le c_0\|u\|_{\mathcal{H}_0} \quad \forall u \in \mathcal{H}_0, \tag{D.1}$$

where the constant c_0 is independent of τ; moreover,

$$\lim_{\tau\to 0}(u^\tau, v^\tau)_{\mathcal{H}_\tau} = (u^0, v^0)_{\mathcal{H}_0} \tag{D.2}$$

provided that

$$\lim_{\tau\to 0}\|u^\tau - \mathcal{R}_\tau u^0\|_{\mathcal{H}_\tau} = 0, \quad \lim_{\tau\to 0}\|v^\tau - \mathcal{R}_\tau v^0\|_{\mathcal{H}_\tau} = 0,$$

$$u^\tau, \quad v^\tau \in \mathcal{H}_\tau, \quad u^0, \quad v^0 \in \mathcal{H}_0.$$

(II) The operators $\mathcal{A}_\tau$, $\mathcal{A}_0$ are positive, compact, and self-adjoint, and the norms $\|\mathcal{A}_\tau\| = \|\mathcal{A}_\tau\|_{\mathcal{L}(\mathcal{H}_\tau)}$ are bounded by a constant independent of τ.

(III) If $f^\tau \in \mathcal{H}_\tau$, $f^0 \in \mathcal{H}_0$, and

$$\lim_{\tau\to 0} \|f^\tau - \mathcal{R}_\tau f^0\|_{\mathcal{H}_\tau} = 0, \tag{D.3}$$

then

$$\lim_{\tau\to 0} \|\mathcal{A}_\tau f^\varepsilon - \mathcal{R}_\tau \mathcal{A}_0 f^0\|_{\mathcal{H}_\tau} = 0. \tag{D.4}$$

(IV) For any sequence $f^\tau \in \mathcal{H}_\tau$ such that $\sup_\tau \|f^\tau\|_{\mathcal{H}_\tau} < \infty$, there exists a subsequence $f^{\tau'}$ and a vector $w^0 \in \mathcal{H}_0$ such that

$$\|\mathcal{A}_{\tau'} f^{\tau'} - \mathcal{R}_{\tau'} w^0\|_{\mathcal{H}_{\tau'}} \to 0 \quad \text{as} \quad \tau' \to 0. \tag{D.5}$$

Consider the spectral problems for the operators $\mathcal{A}_\tau$:

$$\begin{aligned} &u_\tau^k \in \mathcal{H}_\tau, \quad \mathcal{A}_\tau u_\tau^k = \mu_\tau^k u_\tau^k, \quad k = 1, 2, \dots, \\ &\mu_\tau^1 \ge \mu_\tau^2 \ge \cdots \ge \mu_\tau^k, \quad \mu_\tau^k > 0, \\ &(u_\tau^l, u_\tau^m) = \delta_{lm}, \end{aligned} \tag{D.6}$$

and consider the spectral problem for $\mathcal{A}_0$:

$$\begin{aligned} &u_0^k \in \mathcal{H}_0, \quad \mathcal{A}_0 u_0^k = \mu_0^k u_0^k, \quad k = 1, 2, \dots, \\ &\mu_0^1 \ge \mu_0^2 \ge \cdots \ge \mu_0^k, \quad \mu_0^k > 0, \\ &(u_0^l, u_0^m) = \delta_{lm}, \end{aligned} \tag{D.7}$$

where δ_{lm} is the Kronecker symbol.

LEMMA D.1 (see [11, 18]). *Let the space $\mathcal{H}_\tau, \mathcal{H}_0$ and operators $\mathcal{A}_\tau, \mathcal{A}_0$ satisfy conditions* (I)–(IV); *then for sufficiently small τ*

$$|\mu_\tau^k - \mu_0^k| \le 2 \sup_{u\in N(\mu_0^k, \mathcal{A}_0),\ \|u\|_{\mathcal{H}_0}=1} \|\mathcal{A}_\tau \mathcal{R}_\tau u - \mathcal{R}_\tau \mathcal{A}_0 u\|_{\mathcal{H}_\tau}, \quad k = 1, 2, \dots, \tag{D.8}$$

where μ_τ^k, μ_0^k are eigenvalues of problems (D.6) *and* (D.7), *respectively. $N(\mu_0^k, \mathcal{A}_0) = \{u \in \mathcal{H}_0, \mathcal{A}_0 u = \mu_0^k u\}$ is the eigenspace of operator $\mathcal{A}_0$ corresponding to the eigenvalue μ_0^k.*

LEMMA D.2 (see [11, 18]). *Assume that $k \ge 1$, $t \ge 1$ are integers, and*

$$\mu_0^{k-1} > \mu_0^k = \cdots = \mu_0^{k+t-1} > \mu_0^{k+t}, \tag{D.9}$$

i.e., the multiplicity of the eigenvalue μ_0^k is equal to t (here $\mu_0^0 = \infty$). Then for any $w \in N(\mu_0^k, \mathcal{A}_0)$, $\|w\|_{\mathcal{H}_0} = 1$, there exists a linear combination $\bar{u}_\tau$ of eigenvectors $u_\tau^k \cdots u_\tau^{k+t-1}$ of problem (D.6) *such that*

$$\|\bar{u}_\tau - \mathcal{R}_\tau w\|_{\mathcal{H}_\tau} \le M_k \|\mathcal{A}_\tau \mathcal{R}_\tau w - \mathcal{R}_\tau \mathcal{A}_0 w\|_{\mathcal{H}_\tau}, \tag{D.10}$$

where the constant M_k does not depend on τ.

Next let us turn to the proof of Theorem 4.1.

Proof. In Lemmas D.1 and D.2, choose $0 < \tau = h_0 << 1$, $\mathcal{H}_{h_0} = \mathcal{H}_0 = L^2(\Omega)$; $\mathcal{R}_{h_0} \equiv I$ is an identity operator.

For the sake of convenience, we prove that (4.8), (4.9) are valid for Dirichlet's boundary conditions, i.e., $\mathcal{B}_{h_0} \equiv I$.

Define the operators $\mathcal{A}_{h_0} : \mathcal{H}_{h_0} \to \mathcal{H}_{h_0}$ setting $\mathcal{A}_{h_0} f^{h_0} = w^{h_0}$, where w^{h_0} is the weak solution of the following problem:

$$\begin{cases} \widehat{\mathcal{L}}_{h_0} w^{h_0} = f^{h_0} & \text{in} \quad \Omega, \\ w^{h_0} = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{D.11}$$

where $w^{h_0} \in H^1(\Omega)$, $f^{h_0} \in L^2(\Omega)$.

It follows from Proposition 4.2 that the norm $\|\mathcal{A}_{h_0}\|$ is bounded. The compactness of the operator $\mathcal{A}_{h_0} : \mathcal{H}_{h_0} \to \mathcal{H}_{h_0}$ is due to the compact imbedding $H_0^1(\Omega) \subset L^2(\Omega)$. The fact that $\widehat{\mathcal{L}}_{h_0}$ is symmetric guarantees that $\mathcal{A}_{h_0}$ is a self-adjoint operator in $\mathcal{H}_{h_0}$, since

$$\begin{aligned} &(\mathcal{A}_{h_0} f^{h_0}, g^{h_0})_{\mathcal{H}_{h_0}} = (\mathcal{A}_{h_0} f^{h_0}, g^{h_0})_{L^2(\Omega)} = (w^{h_0}, g^{h_0})_{L^2(\Omega)} \\ &= (w^{h_0}, \widehat{\mathcal{L}}_{h_0} v^{h_0})_{L^2(\Omega)} = (\widehat{\mathcal{L}}_{h_0} w^{h_0}, v^{h_0})_{L^2(\Omega)} = (f^{h_0}, \mathcal{A}_{h_o} g^{h_0})_{L^2(\Omega)}, \end{aligned} \tag{D.12}$$

where $w^{h_0} = \mathcal{A}_{h_0} f^{h_0}$, $v^{h_0} = \mathcal{A}_{h_0} g^{h_0}$.

Below we need to verify that conditions (I)–(IV) are valid on purpose to use Lemmas D.1 and D.2.

It is easy to see that condition (I) is valid due to $\mathcal{R}_{h_0} \equiv I$.

In a similar way, we define the operator $\mathcal{A}_0 : \mathcal{H}_0 \to \mathcal{H}_0$ by $\mathcal{A}_0 f = w$, where w is the solution of the following Dirichlet problem:

$$\begin{cases} \widehat{\mathcal{L}} w = f & \text{in} \quad \Omega, \\ w = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{D.13}$$

where $w \in H_0^1(\Omega)$, $f \in L^2(\Omega)$.

Thus condition (II) has also been verified.

From (D.11) and (D.13), one obtains

$$-\frac{\partial}{\partial x_i}\left(\hat{a}_{ij}^{h_0} \frac{\partial (w^{h_0} - w)}{\partial x_j}\right) = -\frac{\partial}{\partial x_i}\left(\hat{r}_{ij} \frac{\partial w}{\partial x_j}\right) - \langle b \rangle (w^{h_0} - w) + f^{h_0}(x) - f(x), \tag{D.14}$$

where $\hat{r}_{ij} = \hat{a}_{ij}^{h_0} - \hat{a}_{ij}$.

Since $w^{h_0}(x) - w(x) \in H_0^1(\Omega)$, it follows from Proposition 4.2 and the Poincaré–Friedrichs inequality that

$$\begin{aligned} &\|w^{h_0} - w\|_{1,\Omega}^2 \le C a(w^{h_0} - w, w^{h_0} - w) \\ &\le C \int_\Omega \hat{r}_{ij} \frac{\partial w}{\partial x_j} \cdot \frac{\partial (w^{h_0} - w)}{\partial x_i} dx + C \int_\Omega (f^{h_0} - f) \cdot (w^{h_0} - w) dx \\ &\le C h_0^2 \|N_i\|_{2,Q} \|N_j\|_{2,Q} \|w\|_{1,\Omega} \|w^{h_0} - w\|_{1,\Omega} \\ &+ C \|f^{h_0} - f\|_{0,\Omega} \|w^{h_0} - w\|_{0,\Omega}. \end{aligned} \tag{D.15}$$

Thus

$$\|w^{h_0} - w\|_{1,\Omega} \le C h_0^2 \|N_i\|_{2,Q}^2 \|w\|_{1,\Omega} + C \|f^{h_0} - f\|_{0,\Omega}. \tag{D.16}$$

If $f^{h_0} \to f$ in $L^2(\Omega)$ as $h_0 \to 0$, by using (D.16) we have

$$w^{h_0} \to w \quad \text{in} \quad H_0^1(\Omega) \quad \text{as} \quad h_0 \to 0 \tag{D.17}$$

Therefore condition (III) holds, too.

Returning to condition (IV), for any sequence $f^{h_0} \in L^2(\Omega)$

$$\sup_{h_0} \|f^{h_0}\|_{L^2(\Omega)} < \infty.$$

Since $L^2(\Omega)$ is a reflexive Hilbert space, it follows from Eberlein's theorem that there exists a subsequence $f^{h_0'} \in L^2(\Omega)$ such that $f^{h_0'} \overset{w}{\rightharpoonup} f \in L^2(\Omega)$. Similarly to (D.15), one can obtain

$$\begin{aligned} &\|w^{h_0'} - w\|_{1,\Omega}^2 \le Ca(w^{h_0'} - w, w^{h_0'} - w) \\ &\le C \int_\Omega \hat{r}_{ij} \frac{\partial w}{\partial x_j} \frac{\partial (w^{h_0'} - w)}{\partial x_j} dx + C \int_\Omega (f^{h_0'} - f) \cdot (w^{h_0'} - w) dx \\ &= J_1 + J_2. \end{aligned} \tag{D.18}$$

$J_1 \to 0$ as $h_0 \to 0$ is due to the fact $\|\hat{r}_{ij}\|_F \le Ch_0^2 \|N_i\|_{2,Q} \|N_j\|_{2,Q}$. Since $f^{h_0'} \overset{w}{\rightharpoonup} f \in L^2(\Omega)$ as $h_0' \to 0$, then $J_2 \to 0$ as $h_0' \to 0$. Thus $w^{h_0'} \to w$ in $H^1(\Omega)$ as $h_0' \to 0$. Therefore condition (IV) is verified.

Setting $\mu_{h_0}^k = (\tilde{\lambda}_k^{(0)})^{-1}$, $\mu_0^k = (\lambda_k^{(0)})^{-1}$, it follows from Lemma D.2 that

$$|(\tilde{\lambda}_k^{(0)})^{-1} - (\lambda_k^{(0)})^{-1}| \le 2 \sup_{w \in N((\lambda_k^{(0)})^{-1}, \mathcal{A}_0), \quad \|w\|_{\mathcal{H}_0} = 1} \|\mathcal{A}_{h_0} w - \mathcal{A}_0 w\|_{\mathcal{H}_{h_0}}, \tag{D.19}$$

where $N((\lambda_k^{(0)})^{-1}, \mathcal{A}_0)$ as indicated in Lemma D.2.

For any $w \in N((\lambda_k^{(0)})^{-1}, \mathcal{A}_0)$, $\|w\|_{\mathcal{H}_0} = 1$, define

$$\begin{cases} \widehat{\mathcal{L}}_{h_0} v^{h_0} = w & \text{in} \quad \Omega, \\ v^{h_0} = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{D.20}$$

$$\begin{cases} \widehat{\mathcal{L}} v = w & \text{in} \quad \Omega, \\ v = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{D.21}$$

From the proof of Proposition 4.2, similarly, we can conclude that

$$\|v^{h_0} - v\|_{1,\Omega} \le Ch_0^2 \|N_i\|_{2,Q}^2. \tag{D.22}$$

Thus

$$\begin{aligned} &|(\tilde{\lambda}_k^{(0)})^{-1} - (\lambda_k^{(0)})^{-1}| \le 2 \sup_{w \in N((\lambda_k^{(0)})^{-1}, \mathcal{A}_0), \ \|w\|_{\mathcal{H}_0} = 1} \|\mathcal{A}_{h_0} w - \mathcal{A}_0 w\|_{\mathcal{H}_{h_0}} \\ &\le C \|v^{h_0} - v\|_{1,\Omega} \le Ch_0^2 \|N_i\|_{2,Q}^2, \end{aligned}$$

i.e.,

$$|\tilde{\lambda}_k^{(0)} - \lambda_k^{(0)}| \le C_k h_0^2 \|N_i\|_{2,Q}^2.$$

It follows from Lemma D.2 that

$$\|u_k^0 - \bar{u}_k^0\|_{0,\Omega} \le C_k h_0^2 \|N_i\|_{2,Q}^2. \qquad \square$$

Acknowledgments. The authors of this paper thank the State Key Laboratory of Science-Engineering Computing (China) and Prof. Z. S. Cai for their support. Also, we thank the referees for their constructive suggestions.

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.
[2] A. Bensoussan, J.L. Lions, and G. Papanicolaou, *Asymptotic Analysis of Periodic Structures*, North-Holland, Amsterdam, 1978.
[3] J.F. Bourgat, *Numerical experments to the homogenization method for operators with periodic coefficients*, in Computing Methods in Applied Sciences and Engineering, I, Lecture Notes in Math. 705, 1977, Springer-Verlag, 1979, pp. 330–356.
[4] L.Q. Cao, J.Z. Cui, and H. Yue, *Finite element computation for elastic structures of composite materials formed by entirely basic configuration*, Chinese J. Numer. Math. Appl., 20 (1998), pp. 25–37.
[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1953.
[6] J.Z. Cui and H.Y. Yang, *A dual coupled method for boundary value problems of PDE with coefficients of small period*, J. Comput. Math., 14 (1996), pp. 159–174.
[7] T.Y. Hou, X.H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[8] T.Y. Hou and X.H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[9] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.
[10] P. Grisvard, *Behavior of the solutions of an elliptic boundary value problem in a polygonal or polyhedral domain*, in Numerical Solution of Partial Differential Equations, III, B. Hubbard, ed., Academic Press, New York, 1976, pp. 207–274.
[11] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[12] H. Kardestuncer and D.H. Norrie, eds., *Finite Element Handbook*, McGraw-Hill, New York, 1987.
[13] S. Kesavan, *Homogenization of elliptic eigenvalue problems*. I, Appl. Math. Optim., 5 (1979), pp. 153–167.
[14] S. Kesavan, *Homogenization of elliptic eigenvalue problems*. II, Appl. Math. Optim., 5 (1979), pp. 197–216.
[15] Q. Lin and Q.D. Zhu, *The Preprocessing and Postprocessing for the Finite Element Method*, Shanghai Scientific and Technical, Shanghai, 1994 (in Chinese).
[16] J.L. Lions, *Some Methods for the Mathematical Analysis of Systems and Their Controls*, Science Press, Beijing, 1981.
[17] S. Moskow and M. Vogelius, *First-order corrections to the homogenised eigenvalues of a periodic composite medium, A convergence proof*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1263–1299.
[18] O.A. Oleinik, A.S. Shamaev, and G.A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, North-Holland, Amsterdam, 1992.
[19] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1997.
[20] F. Santosa and M. Vogelius, *First-order corrections to the homogenized eigenvalues of a periodic composite medium*, SIAM J. Appl. Math., 53 (1993), pp. 1636–1668.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 578–600

NUMERICAL SCHUBERT CALCULUS BY THE PIERI HOMOTOPY ALGORITHM*

T. Y. LI†, XIAOSHEN WANG‡, AND MENGNIEN WU§

Abstract. Based on Pieri's formula on Schubert varieties, the Pieri homotopy algorithm was first proposed by Huber, Sottile, and Sturmfels [*J. Symbolic Comput.*, 26 (1998), pp. 767–788] for numerical Schubert calculus to enumerate all p-planes in $\mathbb{C}^{m+p}$ that meet n given planes in general position. The algorithm has been improved by Huber and Verschelde [*SIAM J. Control Optim.*, 38 (2000), pp. 1265–1287] to be more intuitive and more suitable for computer implementations.

A different approach of employing the Pieri homotopy algorithm for numerical Schubert calculus is presented in this paper. A major advantage of our method is that the polynomial equations in the process are all square systems admitting the same number of equations and unknowns. Moreover, the degree of each polynomial equation is always 2, which warrants much better numerical stability when the solutions are being solved. Numerical results for a big variety of examples illustrate that a considerable advance in speed as well as much smaller storage requirements have been achieved by the resulting algorithm.

Key words. enumerative geometry, Schubert variety, Pieri formula, Pieri homotopy algorithm, Pieri poset

AMS subject classifications. 14N10, 14M15, 65H10, 68Q40

PII. S003614290139175X

1. Introduction.

With "l-planes" representing l dimensional linear subspaces, a general problem in enumerative geometry is

($\star$) Enumerate all p-planes in $\mathbb{C}^{m+p}$ that meet n given planes $L_1, \ldots, L_n$ in general position of dimension $m+1-k_i$ for $i = 1, \ldots, n$, with $k_1 + \cdots + k_n = mp$.

The condition that $k_1 + \cdots + k_n = mp$ guarantees a finite number of p-planes meeting those given planes.

Based on Pieri's formula, and following the new geometric proof of Pieri's formula established by Sottile [9], Huber, Sottile, and Sturmfels [3] proposed the *Pieri homotopy algorithm* to deal with this problem numerically. The homotopies in the algorithm have then been simplified by Huber and Verschelde [4] via the poset of localization patterns, making the algorithm more suitable for computer implementations.

In both of those works, each given plane L_i for $i = 1, \ldots, n$ with dimension $d_i = m + 1 - k_i$ is represented, as they were traditionally, by an $(m+p) \times d_i$ matrix consisting of d_i linearly independent vectors in $\mathbb{C}^{m+p}$. Let X be a p-plane that intersects all those given planes. Without loss, one may represent X by the $(m+p) \times p$

*Received by the editors July 2, 2001; accepted for publication (in revised form) January 17, 2002; published electronically June 12, 2002.

http://www.siam.org/journals/sinum/40-2/39175.html

†Department of Mathematics, Michigan Sate University, East Lansing, MI 48824 (li@math.msu.edu). The research of this author was supported in part by NSF under grant DMS-0104009.

‡Department of Mathematics and Statistics, University of Arkansas at Little Rock, Little Rock, AR 72204 (xxwang@ualr.edu). The research of this author was supported in part by the Visiting Scholar Foundation of Ministry of Education of China. Part of this author's work was done during a stay at the Key Laboratory for Symbolic Computation and Knowledge Engineering in China.

§Department of Mathematics, Tamkang University, Taipei, Taiwan (wu@math.tku.edu.tw).

matrix

$$\begin{bmatrix} 1 & & 0 \\ x_{11} & \ddots & \\ \vdots & \ddots & 1 \\ x_{m1} & & x_{1p} \\ & \ddots & \vdots \\ 0 & & x_{mp} \end{bmatrix}.$$

For $i = 1, \ldots, n$, let $[\alpha]_i(x)$, where $x = (x_{11}, \ldots, x_{mp})$, denote the maximal minor of the $(m+p) \times (p+d_i)$ matrix $[X|L_i]$ with row indices $\alpha = (\alpha_1, \ldots, \alpha_{p+d_i})$. Then the intersection conditions in problem $(\star)$ become, for each $i = 1, \ldots, n$,

$$X \text{ meets } L_i \iff [\alpha]_i(x) = 0 \quad \forall \text{ possible row indices } \alpha = (\alpha_1, \ldots, \alpha_{p+d_i}).$$

The backbone of the Pieri homotopy algorithm [3, 4] is to solve k_i more variables in $x = (x_{11}, \ldots, x_{mp})$ one at a time for $i = 1, \ldots, n$ successively to satisfy the intersection conditions with $L_1, \ldots, L_i$:

$$[\alpha]_l(x) = 0 \ \forall \text{ possible row indices } \alpha = (\alpha_1, \ldots, \alpha_{p+d_l}), \text{ for } l = 1, \ldots, i. \tag{1}$$

To solve the above systems successively for $i = 1, \ldots, n$, different homotopies based on Pieri's formula on Schubert varieties are constructed at each stage where the solutions of the system at the current stage taken as the solutions of the target system of the current homotopy are the solutions of the start system of the homotopy at the next stage. In the process, if $k_i = 1$, then $d_i = m + 1 - 1 = m$, making $[X|L_i]$ a square matrix and resulting in the increment of one more equation in one more unknown in (1) from $(i-1)$th stage to ith stage. However, when $k_i > 1$, then $d_i = m + 1 - k_i < m$, and consequently the number of all possible maximal minors in the $(m+p) \times (p+d_i)$ matrix $[X|L_i]$ equals

$$\binom{p+m}{p+d_i} = \binom{p+m}{p+m+1-k_i} = \binom{p+m}{k_i - 1} > k_i$$

since $k_i = m+1-d_i < m$. When this occurs, the system in (1) admits more equations than unknowns and constitutes an overdetermined system.

Solving an overdetermined system by the homotopy continuation method as proposed in [8], a square system is constructed by using random linear combinations of all equations in (1). This reduction to a square system destroys the geometric structure and creates many excess solution paths to follow, which may lead to a considerable inefficiency of the algorithm since the solution sets of the new square system may properly contain the original ones.

In this paper, we present a different approach. Most importantly, we will represent each given plane L_i, $i = 1, \ldots, n$, in general position by a set of $m+p-d_i = p+k_i-1$ linear equations which defines L_i. The collection of the normals of those equations forms a $(p+k_i-1) \times (m+p)$ matrix, denoted by K_i, and X meets L_i if and only if

$$K_i X \Lambda_i = 0 \quad \text{for some } \Lambda_i \in \mathbb{P}^{p-1}.$$

Employing the same strategy as in [3, 4], we will solve k_i more variables in $x = (x_{11}, \ldots, x_{mp})$ one at a time from $i = 1$ to $i = n$ by solving for each i the system

$$K_l X \Lambda_l = 0 \ \ \Lambda_l \in \mathbb{P}^{p-1} \qquad \text{for } \ l = 1, \ldots, i. \tag{2}$$

And different homotopies are constructed, also based on Pieri's formula, at different stages to connect the solutions of the systems in (2) for consecutive i's. For each fixed i, the system in (2) has $(p+k_1-1)+\cdots+(p+k_i-1)=(p-1)i+k_1+\cdots+k_i$ equations. On the other hand, since $\Lambda_l \in \mathbb{P}^{p-1}$, it admits only $p-1$ variables for each l; together with $k_1+\cdots+k_i$ variables in $x=(x_{11},\ldots,x_{mp})$, the system has $(p+k_1-1)+\cdots+(p+k_i-1)=(p-1)i+k_1+\cdots+k_i$ variables. We therefore deal with square systems throughout the process even when $k_i>1$ occurs and never have to undertake the disadvantages of solving overdetermined systems. Moreover, another important advantage of our approach is that the degree of each polynomial equation in (2) is always 2 while polynomial equations in the previous approaches in [3, 4] may reach quite higher degrees in many situations, which may severely affect the numerical stability when solutions of the systems are being solved.

The computational experiences of the resulting algorithm are listed at the end of the paper to illustrate the remarkable speed up of our method has achieved over the existing algorithm in [4] for a big variety of examples, and our algorithm is particularly valuable for general cases when $k_i>1$ appears.

While, in this paper, we only deal with given planes in general position, the input data of planes for applications may not be so general. An approach common to practitioners of homotopies to solve a given problem is to deform the solutions of the general problem to those of the special problem by applying cheater's homotopy [5] or coefficient-parameter polynomial continuation [6, 7].

In [4], new homotopies were presented to compute p-plane producing curves intersecting m-planes at prescribed interpolation points. A future project would be to investigate whether the improvements proposed in this paper also apply to those new homotopies developed in [4].

2. Preliminaries.

DEFINITION 1. *Let $A_1 \subsetneq A_2 \subsetneq \cdots \subsetneq A_p$ be a set of planes in $\mathbb{C}^{m+p}$ with $\dim(A_i)=a_i$. The set*

$$\Omega(A_1,\ldots,A_p) := \{\ p\text{-planes } X \text{ in } \mathbb{C}^{m+p} |\ \dim(X \bigcap A_i) \geq i, i=1,\ldots,p\ \}$$

is called a Schubert variety.

For planes $A_1 \subsetneq \cdots \subsetneq A_p$ and $B_1 \subsetneq \cdots \subsetneq B_p$ with $\dim(B_i)=\dim(A_i)=a_i$, for $i=1,\ldots,p$, a nonsingular linear transformation in $\mathbb{C}^{m+p}$ can be constructed to transform A_i to B_i for $i=1,\ldots,p$, and the induced transformation transforms $\Omega(A_1,\ldots,A_p)$ onto $\Omega(B_1,\ldots,B_p)$. For this reason, the notation $\Omega(a_1,\ldots,a_p)$ is frequently used without specifying the planes A_i where $\dim(A_i)=a_i, i=1,\ldots,p$.

Now consider planes $A_1 \subsetneq A_2 \subsetneq \cdots \subsetneq A_p$ and $B_1 \subsetneq \cdots \subsetneq B_p$ with $\dim(A_i)=a_i$ and $\dim(B_i)=b_i$, $i=1,\ldots,p$. When they are all in general position, we may assume

$$A_i = \langle e_1,\ldots,e_{a_i}\rangle \ \text{ and } \ B_i = \langle e_{m+p+1-b_i},\ldots,e_{m+p}\rangle, \quad i=1,\ldots,p,$$

where e_j is the unit vector in $\mathbb{C}^{m+p}$ with unit at the jth entry. Here, and from here on, $\langle v_1,\ldots,v_l\rangle$ denotes the plane spanned by $v_1,\ldots,v_l$. If $X \in \Omega(A_1,\ldots,A_p) \bigcap \Omega(B_1,\ldots,B_p)$, then $\dim(X \bigcap A_{p+1-i}) \geq p+1-i$ and $\dim(X \bigcap B_i) \geq i$. Thus, since $\dim(X)=p$ and both $X \bigcap A_{p+1-i}$ and $X \bigcap B_i$ are planes in X,

$$\begin{aligned} \dim(A_{p+1-i} \bigcap B_i) &\geq \dim((X \bigcap A_{p+1-i}) \bigcap (X \bigcap B_i)) \\ &\geq \dim(X \bigcap A_{p+1-i}) + \dim(X \bigcap B_i) - \dim(X) \\ &\geq p+1-i+i-p = 1. \end{aligned}$$

So, $a_{p+1-i}+b_i \geq m+p+1$. Conversely, if $a_{p+1-i}+b_i \geq m+p+1$, then $a_{p+1-i} \geq m+p+1-b_i$. Thus $e_{m+p+1-b_i} \in B_i \bigcap A_{p+1-i}$. Let $X = \langle e_{m+p+1-b_p}, e_{m+p+1-b_{p-1}}, \ldots, e_{m+p+1-b_1}\rangle$. Obviously, $X \bigcap B_i = \langle e_{m+p+1-b_1}, e_{m+p+1-b_2}, \ldots, e_{m+p+1-b_i}\rangle$ and $\dim(X \bigcap B_i) = i$. Furthermore, $X \bigcap A_{p+1-i} \supseteq \langle e_{m+p+1-b_p}, \ldots, e_{m+p+1-b_i}\rangle$ and hence $\dim(X \bigcap A_{p+1-i}) \geq p+1-i$. Thus $X \in \Omega(A_1, \ldots, A_p) \bigcap \Omega(B_1, \ldots, B_p)$. Therefore, we have the following proposition.

PROPOSITION 1 (Theorem I, p. 327 [1]). *When $A_1 \subsetneq A_2 \subsetneq \cdots \subsetneq A_p$ and $B_1 \subsetneq \cdots \subsetneq B_p$ are planes in general position in $\mathbb{C}^{m+p}$ with $\dim(A_i) = a_i$ and $\dim(B_i) = b_i$ for $i = 1, \ldots, p$, then $\Omega(a_1, \ldots, a_p)$ and $\Omega(b_1, \ldots, b_p)$ intersect if and only if*

$$a_{p+1-i} + b_i \geq m+p+1 \qquad \text{for} \quad i = 1, \ldots, p.$$

As a corollary, we have the following proposition.

PROPOSITION 2 (Corollary, p. 328 [1]). *$\Omega(a_1, \ldots, a_p) \bigcap \Omega(m+p+1-a_p, \ldots, m+p+1-a_1)$ consists of a unique p-plane for given $1 \leq a_1 < \cdots < a_p \leq m+p$.*

EXAMPLE 1. *Let $m = p = 2$, $A_1 = \langle e_1, e_2\rangle$, $A_2 = \langle e_1, e_2, e_3\rangle$, $B_1 = \langle e_3, e_4\rangle$, and $B_2 = \langle e_2, e_3, e_4\rangle$. Then $\Omega(A_1, A_2) \bigcap \Omega(B_1, B_2) = \langle e_2, e_3\rangle$.*

In the rest of the paper, when we write $\mathbf{a} = (a_1, \ldots, a_p)$, those coordinates will satisfy $1 \leq a_1 < \cdots < a_p \leq m+p$. Because of the importance of Proposition 2, $\mathbf{a}^* = (m+p+1-a_p, \ldots, m+p+1-a_1)$ is called the *dual* of $\mathbf{a} = (a_1, \cdots, a_p)$.

For $0 \leq h \leq m$, let $\sigma_h := \Omega(m+1-h, m+2, \ldots, m+p)$, the set of p-planes that meet a given $(m+1-h)$-plane. Since every p-plane will meet any $(m+1)$-plane, σ_0 is the collection of all p-planes.

For $(a_1, \ldots, a_p)$ and $(b_1, \ldots, b_p)$ with $a_{p+1-i} \geq m+p+1-b_i$, for $i = 1, \ldots, p$, let $A_1 \subsetneq A_2 \subsetneq \cdots \subsetneq A_p$ and $B_1 \subsetneq \cdots \subsetneq B_p$ be planes in $\mathbb{C}^{m+p}$ with $\dim(A_i) = a_i$ and $\dim(B_i) = b_i$. If $X \in \Omega(a_1, \ldots, a_p) \bigcap \Omega(b_1, \ldots, b_p)$, then X meets $A_{p+1-i} \bigcap B_i$ for $i = 1, \ldots, p$. Let D be the smallest plane containing $A_p \bigcap B_1, \ldots, A_1 \bigcap B_p$. Then $X \subset D$ and

$$\begin{aligned} \dim(D) &\leq \dim(A_p \bigcap B_1) + \cdots + \dim(A_1 \bigcap B_p) \\ &= a_p + b_1 - (m+p) + \cdots + a + 1 + b_p - (m+p) \\ &= \textstyle\sum_{i=1}^{p} (a_i + b_i) - (m+p)p. \end{aligned}$$

Let $h = \sum a_i + \sum b_i - (m+p+1)p$. Clearly,

$$\dim(D) = h + p \Longleftrightarrow a_{p-i} < m+p+1-b_i \leq a_{p-i+1} \quad \forall i = 1, \ldots, p.$$

When $\dim(D) = h+p$, let G_h be a generic $(m+1-h)$-plane. Representing D and G_h by matrices consisting of independent vectors in $\mathbb{C}^{m+p}$, the rank of the $(m+p)\times(m+p+1)$ matrix $[D|G_h]$ is $m+p$. Thus, up to a scalar factor, there is a unique nonzero vector $g \in G_h$, where $g = v_1 + \cdots + v_p$ with $v_i \in A_{p+1-i} \bigcap B_i$ for $i = 1, \ldots, p$. Let $X = \langle v_1, \ldots, v_p\rangle$; then $X \in \Omega(A_1, \ldots, A_p) \bigcap \Omega(B_1, \ldots, B_p)$ and meets G_h.

PROPOSITION 3 (Theorem III, p. 333 [1]). *Let $1 \leq h \leq m$. For $(a_1, \ldots, a_p)$ and $(b_1, \ldots, b_p)$ satisfying*

$$a_{p-i} < m+p+1-b_i \leq a_{p+1-i}, \quad h = \textstyle\sum a_i + \sum b_i - (m+p+1)p, \tag{3}$$

the intersection $\Omega(a_1, \ldots, a_p) \bigcap \Omega(b_1, \ldots, b_p) \bigcap \sigma_h$ consists of a unique p-plane.

EXAMPLE 2. *Let $m = p = 2$, $A_1 = \langle e_1, e_2\rangle$, $A_2 = \langle e_1, e_2, e_3\rangle$, $B_1 = \langle e_3, e_4\rangle$, $B_2 = \langle e_1, e_2, e_3, e_4\rangle$, and $\sigma_1 = \Omega(D_1, D_2)$, where D_1 is a generic 2-plane and $D_2 = \mathbb{C}^4$. Then $A_1 \bigcap B_2 = A_1$ and $A_2 \bigcap B_1 = \langle e_3\rangle$. Denote the 1-plane $D_1 \bigcap \langle A_1, e_3\rangle$ by D_1'.*

Let $u = (u_1, u_2, u_3, u_4) \in D_1'$ *and* $f_1 = (u_1, u_2, 0, 0)$. *Then* $\Omega(2,3) \bigcap \Omega(2,4) \bigcap \sigma_1 = \langle f_1, e_3 \rangle$.

For $\Omega(a_1, \ldots, a_p)$ and $\Omega(b_1, \ldots, b_p)$, $\Omega(a_1, \ldots, a_p) + \Omega(b_1, \ldots, b_p)$ denotes the class of p-planes X, where for planes $A_1 \subsetneq A_2 \subsetneq \cdots \subsetneq A_p$ and $B_1 \subsetneq \cdots \subsetneq B_p$ with $\dim(A_i) = a_i$ and $\dim(B_i) = b_i$ for $i = 1, \ldots, p$, $\dim(X \bigcap A_i) \geq i$ (or $\dim(X \bigcap B_i) \geq i$) for all $i = 1, \ldots, p$. We abbreviate $\Omega(a_1, \ldots, a_p) + \Omega(a_1, \ldots, a_p)$ by $2\Omega(a_1, \ldots, a_p)$ and in general

$$\sum_{i=1}^{d} \Omega(a_1, \ldots, a_p) := d\,\Omega(a_1, \ldots, a_p).$$

Furthermore, $\Omega(a_1, \ldots, a_p) \bullet \Omega(b_1, \ldots, b_p)$ represents the class of p-planes X, where $\dim(X \bigcap A_i) \geq i$ *and* $\dim(X \bigcap B_i) \geq i$ for all $i = 1, \ldots, p$.

For sets of p-planes A and B, we write $A \bullet B$ for $A \bigcap B$. We say A is *equivalent* to B, denoted by $A \sim B$, if whenever

$$A \bullet \Omega(\mathbf{c}) = k\,\Omega(1, \ldots, p)$$

for some $\mathbf{c} = (c_1, \ldots, c_p)$, we also have

$$B \bullet \Omega(\mathbf{c}) = k\,\Omega(1, \ldots, p).$$

Note that $\Omega(1, \ldots, p)$ represents a general p-plane. The following property [1, 2, 3] will be used repeatedly for the establishment of our algorithm:

$$A \sim B \Longrightarrow A \bullet \sigma_h \sim B \bullet \sigma_h.$$

Following Proposition 3, for fixed $a_1, \ldots, a_p$ and h, any $\mathbf{b} = (b_1, \ldots, b_p)$ satisfying (3) yields

$$\Omega(a_1, \ldots, a_p) \bullet \sigma_h \bullet \Omega(b_1, \ldots, b_p) = \Omega(1, \ldots, p).$$

On the other hand, for the dual $\mathbf{b}^* = (b_1^*, \ldots, b_p^*) = (m+p+1-b_p, \ldots, m+p+1-b_1)$ of $\mathbf{b}$, by Proposition 2,

$$\Omega(b_1^*, \ldots, b_p^*) \bullet \Omega(b_1, \ldots, b_p) = \Omega(1, \ldots, p).$$

Moreover, for $\bar{\mathbf{b}} = (\bar{b}_1, \ldots, \bar{b}_p)$ satisfying (3), but $\bar{\mathbf{b}} \neq \mathbf{b}^*$,

$$\Omega(\bar{b}_1, \ldots, \bar{b}_p) \bullet \Omega(b_1, \ldots, b_p) = \emptyset.$$

These observations lead to the following important formula.

Proposition 4 (Pieri's formula, p. 354 [1]).

$$\Omega(a_1, \ldots, a_p) \bullet \sigma_h \sim \sum_{\mathbf{b}=(b_1,\ldots,b_p)} \Omega(b_1, \ldots, b_p), \ \textit{where}$$

$$0 < b_1 \leq a_1 < b_2 \leq a_2 < \cdots \leq a_{p-1} < b_p \leq a_p \ \textit{with} \ \sum b_j = \sum a_j - h. \tag{4}$$

When we fix $\mathbf{a} = (a_1, \ldots, a_p)$ and h, those $\mathbf{b} = (b_1, \ldots, b_p)$ satisfying (4) together with $\mathbf{a}$ are called the *Pieri nodes*; the nodes $\mathbf{b}$ are *induced* Pieri nodes of node $\mathbf{a}$.

From here on, we will use $[a_1, \ldots, a_p]$ to denote a Pieri node or its dual. Recall that for $i = 1, \ldots, n$, p-planes that meet plane L_i with $\dim(L_i) = m + 1 - k_i$ belong to $\Omega(m+1-k_i, m+2, \ldots, m+p) = \sigma_{k_i}$, and the condition $k_1 + \cdots + k_n = mp$ warrants

$$\sigma_{k_1} \bullet \sigma_{k_2} \bullet \cdots \bullet \sigma_{k_n} = d\,\Omega(1, \ldots, p). \tag{5}$$

Thus problem ($\star$) introduced in section 1 can now be interpreted as follows: finding all d specific p-planes in $\sigma_{k_1} \bullet \sigma_{k_2} \bullet \cdots \bullet \sigma_{k_n}$ for given planes $L_1, \ldots, L_n$ in general position. To calculate $\sigma_{k_1} \bullet \sigma_{k_2} \bullet \cdots \bullet \sigma_{k_n}$ in (5), Pieri's formula in Proposition 4 will be used as a main tool. The Pieri nodes derived in the process constitute a *Pieri poset*, and the number d is called the *Pieri root count*.

EXAMPLE 3. *For $m=2$, $p=2$ and given planes L_1, L_2, L_3, L_4 in general position with $\dim(L_i) = 2$ and $k_i = m + 1 - d_i = 1$ for all $i = 1, \ldots, 4$,*

$$\begin{aligned} & \sigma_{k_1} \bullet \sigma_{k_2} \bullet \sigma_{k_3} \bullet \sigma_{k_4} \\ = \;& \Omega(2,4) \bullet \sigma_{k_2} \bullet \sigma_{k_3} \bullet \sigma_{k_4} \\ \sim \;& (\Omega(1,4) + \Omega(2,3)) \bullet \sigma_{k_3} \bullet \sigma_{k_4} \\ \sim \;& 2\,\Omega(1,3) \bullet \sigma_{k_4} \\ \sim \;& 2\,\Omega(1,2). \end{aligned}$$

The Pieri poset of all the Pieri nodes and the poset that consists of their duals are shown in Figure 1.

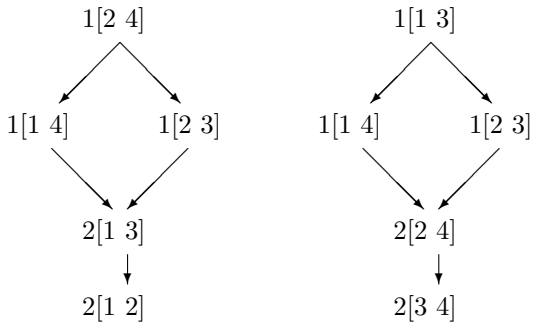


FIG. 1.

Now, any 2-plane X that meets L_1 must be in $\Omega(3,4) \bullet \sigma_1 \sim \Omega(2,4)$. Since $[2,4]^* = [1,3]$, there is a unique 2-plane in $\Omega(2,4) \bullet \Omega(1,3)$. So, if we let $A_1 = \langle e_1 \rangle$ and $A_2 = \langle e_1, e_2, e_3 \rangle$, there is a unique 2-plane in $\Omega(A_1, A_2)$ consisting of 2-planes of the form

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & u \\ 0 & 0 \end{bmatrix} := X_{[1,3]}$$

that meet L_1. We may determine this unique $X_{[1,3]}$ by finding u via its intersection condition with L_1.

Similarly, any 2-plane X that meets both L_1 and L_2 must lie in $\Omega(3,4) \bullet \sigma_1 \bullet \sigma_1 \sim \Omega(1,4) + \Omega(2,3)$ by Proposition 4. Since $[1,4]^* = [1,4]$ and $\Omega(2,3) \bullet \Omega(1,4) = \emptyset$, by

letting $A_1 = \langle e_1 \rangle$ *and* $A_2 = \langle e_1, e_2, e_3, e_4 \rangle$, *there is a unique* 2*-plane in* $\Omega(A_1, A_2)$ *consisting of* 2*-planes of the form*

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & v_1 \\ 0 & v_2 \end{bmatrix} := X_{[1,4]}$$

that meet both L_1 *and* L_2. *We may determine this unique* $X_{[1,4]}$ *by finding* v_1 *and* v_2 *via the intersection conditions of meeting* L_1 *and* L_2. *On the other hand, since* $[2,3]^* = [2,3]$ *and* $\Omega(1,4) \bullet \Omega(2,3) = \emptyset$, *there is a unique* 2*-plane in* $\Omega(A_1, A_2)$ *with* $A_1 = \langle e_1, e_2 \rangle$ *and* $A_2 = \langle e_1, e_2, e_3 \rangle$ *consisting of* 2*-planes of the form*

$$\begin{bmatrix} 1 & 0 \\ v_1' & 1 \\ 0 & v_2' \\ 0 & 0 \end{bmatrix} := X_{[2,3]}$$

that meet L_1 *and* L_2. *This* $X_{[2,3]}$ *is decided when* v_1' *and* v_2' *are found.*

Continuing the same pattern, since $\Omega(3,4) \bullet \sigma_1 \bullet \sigma_1 \bullet \sigma_1 \sim 2\Omega(1,3)$ *and* $[1,3]^* = [2,4]$, *there are two* 2*-planes in* $\Omega(A_1, A_2)$, *with* $A_1 = \langle e_1, e_2 \rangle$ *and* $A_2 = \langle e_1, e_2, e_3, e_4 \rangle$, *consisting of* 2*-planes of the form*

$$\begin{bmatrix} 1 & 0 \\ w_1 & 1 \\ 0 & w_2 \\ 0 & w_3 \end{bmatrix} := X_{[2,4]}$$

that meet L_1, L_2, *and* L_3. *And,* $\Omega(3,4) \bullet \sigma_1 \bullet \sigma_1 \bullet \sigma_1 \bullet \sigma_1 \sim 2\Omega(1,2)$ *as well as* $[1,2]^* = [3,4]$ *imply that the two* 2*-planes that meet all* $L_1, \ldots, L_4$ *can be found by solving two set of* y*'s of*

$$\begin{bmatrix} 1 & 0 \\ y_1 & 1 \\ y_3 & y_2 \\ 0 & y_4 \end{bmatrix} := X_{[3,4]}$$

in $\Omega(A_1, A_2)$ *with* $A_1 = \langle e_1, e_2, e_3 \rangle$ *and* $\langle e_1, e_2, e_3, e_4 \rangle$.

The theme of the so-called Pieri homotopy algorithm is as follows:

1. *Finding* u *in* $X_{[1,3]}$ *by the criteria of meeting* L_1.
2. (a) *Solving* $\{v_1, v_2\}$ *in* $X_{[1,4]}$ *by a homotopy with a starting point containing* $\{v_1 = u, v_2 = 0\}$.
   (b) *Solving* $\{v_1, v_2\}$ *in* $X_{[2,3]}$ *by a different homotopy with a starting point containing* $\{v_1' = 0, v_2' = u\}$.
3. *Solving two sets of* $\{w_1, w_2, w_3\}$ *in* $X_{[2,4]}$ *by a homotopy with two starting points containing* $\{w_1 = 0, w_2 = v_1, w_3 = v_2\}$ *and* $\{w_1 = v_1', w_2 = v_2' w_3 = 0\}$, *respectively.*
4. *Solving two sets of* $\{y_1, y_2, y_3, y_4\}$ *in* $X_{[3,4]}$ *by a homotopy with two starting points containing* $\{y_1 = w_1, y_2 = w_2, y_3 = w_3, y_4 = 0\}$ *with two sets of values of* $\{w_1, w_2, w_3\}$ *obtained at the last step.*

The details of those homotopies of our approach will be elaborated in the next section.

From the process in the above example, we solve the ultimate solutions in $X_{[3,4]}$ by following the cascade of solving

(6)
$$\begin{array}{ccc} & X_{[1\,3]} & \\ \swarrow & & \searrow \\ X_{[2\,3]} & & X_{[1\,4]} \\ \searrow & & \swarrow \\ & 2X_{[2\,4]} & \\ & \downarrow & \\ & 2X_{[3\,4]}. & \end{array}$$

For $\mathbf{a} = [a_1, \ldots, a_p]$, write

$$X_{\mathbf{a}} = X_{[a_1,\ldots,a_p]} := \begin{bmatrix} 1 & & 0 \\ x_{1,1} & \ddots & \\ \vdots & \ddots & 1 \\ x_{(a_1-1),1} & & x_{1,p} \\ & \ddots & \vdots \\ & & x_{(a_p-p),p} \\ & & 0 \\ 0 & & \vdots \\ & & 0 \end{bmatrix}.$$

Those $\mathbf{a}$'s in (6) actually follow the duals of the Pieri poset in Figure 1. For nodes $\mathbf{a}$ and $\mathbf{b}$,

$$\mathbf{a} \to \cdots \to \cdots \to \mathbf{b}$$

is called a *chain* joining $\mathbf{a}$ and $\mathbf{b}$. A chain joining $\mathbf{a} = [m+1, m+2, \ldots, m+p]$ and $\mathbf{b} = [1, \ldots, p]$ is called a *complete chain*. The Pieri homotopy algorithms in general are constructed based on the duals of the Pieri poset consisting of all the derived Pieri nodes.

3. Algorithms. For given planes $L_1, \ldots, L_n$ in $\mathbb{C}^{m+p}$ in general position with $\dim(L_i) = m + 1 - k_i$ for $i = 1, \ldots, n$, all derived Pieri nodes in

$$\sigma_{k_1} \bullet \sigma_{k_2} \bullet \cdots \bullet \sigma_{k_n}$$

form a poset. Unless otherwise indicated, we shall use the term "Pieri poset" for the poset of duals of all those Pieri nodes. As mentioned in the introduction, we shall represent each L_i by a $(p + k_i - 1) \times (m + p)$ matrix K_i whose rows consist of all the normals of the linear equations that define L_i.

(a) Hypersurface intersection conditions, where $k_i = 1$ for all $i = 1, \ldots, n$. Letting $\mathbf{a}^0 = [1, 2, \ldots, p]$ sit on top of the Pieri poset, we may write

$$\mathbf{a}^0 \;\to\; \mathbf{a}^1 \to \;\cdots\; \to \;\mathbf{a}^n \tag{7}$$

for a complete chain in the poset, and it is obvious that the coordinates of consecutive nodes $\mathbf{a}^j$ and $\mathbf{a}^{j+1}$ in the chain can differ by 1 on only one component. We shall use

$$\mathbf{a}^j \;\stackrel{\mu_{j+1}}{\longrightarrow}\; \mathbf{a}^{j+1}$$

to denote that the μ_{j+1}th component of $\mathbf{a}^j$ is increased by 1 to reach $\mathbf{a}^{j+1}$. We may therefore write

$$\mathbf{a}^0 \;\stackrel{\mu_1}{\longrightarrow}\; \mathbf{a}^1 \;\stackrel{\mu_2}{\longrightarrow}\; \cdots \;\stackrel{\mu_n}{\longrightarrow}\; \mathbf{a}^n$$

for a complete chain. Recall that for $\mathbf{a} = [a_1, \ldots, a_p]$

$$X_{\mathbf{a}} = \begin{bmatrix} 1 & & 0 \\ x_{1,1} & \ddots & \\ \vdots & \ddots & 1 \\ x_{(a_1-1),1} & & x_{1,p} \\ & \ddots & \vdots \\ & & x_{(a_p-p),p} \\ & & 0 \\ 0 & & \vdots \\ & & 0 \end{bmatrix}.$$

For $\mathbf{a}^0 \;\stackrel{\mu_1}{\longrightarrow}\; \mathbf{a}^1$, the only unknown in $X_{\mathbf{a}^1}$ can be determined by

$$K_1 X_{\mathbf{a}^1} \Lambda_1^1 = 0,$$

where $\Lambda_1^1 = e_{\mu_1} \in \mathbb{C}^p$. Now, suppose we have proceeded up to

$$\mathbf{a}^0 \;\stackrel{\mu_1}{\longrightarrow}\; \mathbf{a}^1 \;\stackrel{\mu_2}{\longrightarrow}\; \cdots \;\stackrel{\mu_j}{\longrightarrow}\; \mathbf{a}^j$$

in the chain. This means that we have solved all the variables in $X_{\mathbf{a}^j}$ and found $\Lambda_1^j, \ldots, \Lambda_j^j \in \mathbb{P}^{p-1}$ such that

$$K_l X_{\mathbf{a}^j} \Lambda_l^j = 0 \;\text{ for }\; l = 1, \ldots, j.$$

Namely, a p-plane in the form $X_{\mathbf{a}^j}$ that meets planes $L_1, \ldots, L_j$ has been determined. To proceed one step further in the chain, for

$$\mathbf{a}^j \;\stackrel{\mu_{j+1}}{\longrightarrow}\; \mathbf{a}^{j+1}, \quad \text{where } \mathbf{a}^{j+1} = [a_1^{(j+1)}, \ldots, a_p^{(j+1)}],$$

consider the homotopy

$$H(t, X_{\mathbf{a}^{j+1}}, \Lambda^{j+1}) = \begin{cases} K_1 X_{\mathbf{a}^{j+1}} \Lambda_1^{j+1} = 0, \\ \vdots \\ K_j X_{\mathbf{a}^{j+1}} \Lambda_j^{j+1} = 0, \\ [(1-t)\hat{K}_{\mathbf{a}^{j+1}} + tK_{j+1}] X_{\mathbf{a}^{j+1}} \Lambda_{j+1}^{j+1} = 0, \end{cases} \tag{8}$$

where the μ_lth component of Λ_l^{j+1} is 1 for $l = 1, \ldots, j+1$, and $\widehat{K}_{\mathbf{a}^{j+1}}$ is the matrix $[e_{a_1^{(j+1)}}, \ldots, e_{a_p^{(j+1)}}]^T$. For each $t \in [0,1]$, the system admits $p-1$ variables in Λ_l^{j+1} for each $l = 1, \ldots, j+1$ and $j+1$ variables in $X_{\mathbf{a}^{j+1}}$; it admits, in total, $(p-1)(j+1)+(j+1) = p(j+1)$ variables. It is clear that the total number of equations is also $p(j+1)$, making the system a square system. When $t = 0$,

$$\begin{array}{lcll} X_{\mathbf{a}^{j+1}} & = & X_{\mathbf{a}^j}, & \\ \Lambda_l^{j+1} & = & \Lambda_l^j, & l = 1, \ldots, j, \\ \Lambda_{j+1}^{j+1} & = & e_{\mu_{j+1}} (\in \mathbb{C}^p) & \end{array}$$

is a solution of the system $H(0, X_{\mathbf{a}^{j+1}}, \Lambda^{j+1}) = 0$ in (8). Following the homotopy path of $H(t, X_{\mathbf{a}^{j+1}}, \Lambda^{j+1}) = 0$ emanating from this solution, we obtain a solution of $X_{\mathbf{a}^{j+1}}$ and Λ_l^{j+1} for $l = 1, \ldots, j+1$ at $t = 1$ that satisfies

$$K_l X_{\mathbf{a}^{j+1}} \Lambda_l^{j+1} = 0 \quad \text{for } l = 1, \ldots, j+1.$$

A p-plane that meets $L_1, \ldots, L_{j+1}$ in the form of $X_{\mathbf{a}^{j+1}}$ is then found and the chain has been extended one step further; namely, we have proceeded along the chain up to

$$\mathbf{a}^0 \xrightarrow{\mu_1} \mathbf{a}^1 \xrightarrow{\mu_2} \cdots \xrightarrow{\mu_{j+1}} \mathbf{a}^{j+1}.$$

When we proceed further along the chain and arrive at $\mathbf{a}^n$, a p-plane that meets all L_i, $i = 1, \ldots, n$, becomes available.

EXAMPLE 4. *In Example* 3*, there are two chains in the dual poset:*

$$\textit{chain } 1\textit{:}\ [1\,2] \xrightarrow{2} [1\,3] \xrightarrow{2} [1\,4] \xrightarrow{1} [2\,4] \xrightarrow{1} [3\,4],$$
$$\textit{chain } 2\textit{:}\ [1\,2] \xrightarrow{2} [1\,3] \xrightarrow{1} [2\,3] \xrightarrow{2} [2\,4] \xrightarrow{1} [3\,4],$$

and the corresponding homotopies are

$$\begin{array}{l} [1\,2] \\ \downarrow \\ [1\,3]\text{:} \left[\begin{array}{l} K_1 X_{[1\,3]} \Lambda_1^1 = 0, \Lambda_1^1 = \left[\begin{array}{c} 0 \\ 1 \end{array} \right] \end{array} \right. \\ \downarrow \\ [1\,4]\text{:} \left[\begin{array}{r} K_1 X_{[1\,4]} \Lambda_1^2 = 0 \\ \{(1-t)\underline{[e_1, e_4]^T} + tK_2\} X_{[1\,4]} \Lambda_2^2 = 0 \end{array} \right. \\ \downarrow \\ [2\,4]\text{:} \left[\begin{array}{r} K_1 X_{[2\,4]} \Lambda_1^3 = 0 \\ K_2 X_{[2\,4]} \Lambda_2^3 = 0 \\ \{(1-t)[e_2, e_4]^T + tK_3\} X_{[2\,4]} \Lambda_3^3 = 0 \end{array} \right. \\ \downarrow \\ [3\,4]\text{:} \left[\begin{array}{r} K_1 X_{[3\,4]} \Lambda_1^4 = 0 \\ K_2 X_{[3\,4]} \Lambda_2^4 = 0 \\ K_3 X_{[3\,4]} \Lambda_3^4 = 0 \\ \{(1-t)[e_3, e_4]^T + tK_4\} X_{[3\,4]} \Lambda_4^4 = 0 \end{array} \right. \end{array} \qquad \begin{array}{l} [1\,2] \\ \downarrow \\ [1\,3]\text{:} \left[\begin{array}{l} K_1 X_{[1\,3]} \Lambda_1^1 = 0, \Lambda_1^1 = \left[\begin{array}{c} 0 \\ 1 \end{array} \right] \end{array} \right. \\ \downarrow \\ [2\,3]\text{:} \left[\begin{array}{r} K_1 X_{[2\,3]} \Lambda_1^2 = 0 \\ \{(1-t)\underline{[e_2, e_3]^T} + tK_2\} X_{[2\,3]} \Lambda_2^2 = 0 \end{array} \right. \\ \downarrow \\ [2\,4]\text{:} \left[\begin{array}{r} K_1 X_{[2\,4]} \Lambda_1^3 = 0 \\ K_2 X_{[2\,4]} \Lambda_2^3 = 0 \\ \{(1-t)[e_2, e_4]^T + tK_3\} X_{[2\,4]} \Lambda_3^3 = 0 \end{array} \right. \\ \downarrow \\ [3\,4]\text{:} \left[\begin{array}{r} K_1 X_{[3\,4]} \Lambda_1^4 = 0 \\ K_2 X_{[3\,4]} \Lambda_2^4 = 0 \\ K_3 X_{[3\,4]} \Lambda_3^4 = 0 \\ \{(1-t)[e_3, e_4]^T + tK_4\} X_{[3\,4]} \Lambda_4^4 = 0. \end{array} \right. \end{array}$$

For chain 1, $\Lambda_k^l = \left[\begin{smallmatrix} * \\ 1 \end{smallmatrix} \right]$, $k = 1, 2$, $l = 1, \ldots, k$, $\Lambda_k^l = \left[\begin{smallmatrix} 1 \\ * \end{smallmatrix} \right]$, $k = 3, 4$, $l = 1, \ldots, k$; *for chain* 2, $\Lambda_k^l = \left[\begin{smallmatrix} * \\ 1 \end{smallmatrix} \right]$, $k = 1, 3$, $l = 1, \ldots, k$, $\Lambda_k^l = \left[\begin{smallmatrix} 1 \\ * \end{smallmatrix} \right]$, $k = 2, 4$, $l = 1, \ldots, k$. *Between chain* 1 *and* 2*, the only distinct homotopies are* $[1\,3] \xrightarrow{2} [1\,4]$ *in chain* 1 *and* $[1\,3] \xrightarrow{1} [2\,3]$ *in chain* 2. □

EXAMPLE 5. *For* $m = 3$, $p = 2$, *let* $L_1, \ldots, L_6$ *be planes with* $\dim(L_i) = 3$ *for* $i = 1, \ldots, 6$. *The Pieri poset is shown in Figure* 2. *For the complete chain*

$$[1\,2] \xrightarrow{2} [1\,3] \xrightarrow{2} [1\,4] \xrightarrow{1} [2\,4] \xrightarrow{2} [2\,5] \xrightarrow{1} [3\,5] \xrightarrow{1} [4\,5],$$

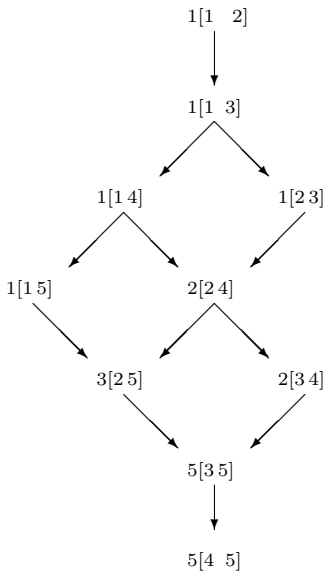


FIG. 2.

the homotopies are

$$
\begin{array}{ll}
[1\,2] & \\
\downarrow & \\
[1\,3] & : \left[\begin{array}{r} K_1 X_{[1\,3]} \Lambda_1^1, = 0, \Lambda_1^1 = \left[\begin{array}{c} 0 \\ 1 \end{array} \right] \end{array} \right. \\
\downarrow & \\
[1\,4] & : \left[\begin{array}{r} K_1 X_{[1\,4]} \Lambda_1^2 = 0, \\ \left\{ (1-t)[e_1, e_4]^T + tK_2 \right\} X_{[1\,4]} \Lambda_2^2 = 0, \end{array} \right. \\
\downarrow & \\
[2\,4] & : \left[\begin{array}{r} K_l X_{[2\,4]} \Lambda_l^3 = 0, \quad l = 1, 2, \\ \left\{ (1-t)[e_2, e_4]^T + tK_3 \right\} X_{[2\,4]} \Lambda_3^3 = 0, \end{array} \right. \\
\downarrow & \\
[2\,5] & : \left[\begin{array}{r} K_l X_{[2\,5]} \Lambda_l^4 = 0, \quad l = 1, 2, 3, \\ \left\{ (1-t)[e_2, e_5]^T + tK_4 \right\} X_{[2\,5]} \Lambda_4^4 = 0, \end{array} \right. \\
\downarrow & \\
[3\,5] & : \left[\begin{array}{r} K_l X_{[3\,5]} \Lambda_l^5 = 0, \quad l = 1, 2, 3, 4, \\ \left\{ (1-t)[e_3, e_5]^T + tK_5 \right\} X_{[3\,5]} \Lambda_5^5 = 0, \end{array} \right. \\
\downarrow & \\
[4\,5] & : \left[\begin{array}{r} K_l X_{[4\,5]} \Lambda_l^6 = 0, \quad l = 1, 2, 3, 4, 5, \\ \left\{ (1-t)[e_4, e_5]^T + tK_6 \right\} X_{[4\,5]} \Lambda_6^6 = 0, \end{array} \right.
\end{array}
$$

where $\Lambda_1^l = \left[\begin{smallmatrix} * \\ 1 \end{smallmatrix} \right]$, $\Lambda_2^l = \left[\begin{smallmatrix} * \\ 1 \end{smallmatrix} \right]$, $\Lambda_3^l = \left[\begin{smallmatrix} 1 \\ * \end{smallmatrix} \right]$, $\Lambda_4^l = \left[\begin{smallmatrix} * \\ 1 \end{smallmatrix} \right]$, $\Lambda_5^l = \left[\begin{smallmatrix} 1 \\ * \end{smallmatrix} \right]$, *and* $\Lambda_6^l = \left[\begin{smallmatrix} 1 \\ * \end{smallmatrix} \right]$. □

Remark 1. Let $\mathbf{a}$ be a node shared by k different complete chains

$$\mathbf{a}_l^0 \xrightarrow{\mu_{1l}} \mathbf{a}_l^1 \xrightarrow{\mu_{2l}} \cdots \xrightarrow{\mu_{nl}} \mathbf{a}_l^n, \quad l = 1, \ldots, k.$$

Say $\mathbf{a}_l^{j+1} = \mathbf{a}$, for $l = 1, \ldots, k$. This means $\sigma_{k_1} \bullet \sigma_{k_2} \bullet \cdots \bullet \sigma_{k_j} \bullet \Omega(\mathbf{a}) = k\,\Omega(1, \ldots, p)$, where $k_i = 1$ for $i = 1, \ldots, j$. In this situation, the homotopies for the extensions $\mathbf{a}_l^j \xrightarrow{\mu_{(j+1)l}} \mathbf{a}_l^{j+1} = \mathbf{a}$ in (8) are the same for all l. It is critically important that those

k paths that emanate from k different starting points

$$X_{\mathbf{a}} = X_{\mathbf{a}_l^j}, \quad \Lambda_{il}^{j+1} = \Lambda_{il}^j \text{ for } i = 1, \ldots, j, \quad \Lambda_{(j+1)l}^{j+1} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \mu_{(j+1)l}\text{th}$$

will reach different solutions at $t = 1$. This assertion is warranted by the following observations. Since for each $t \in [0,1]$, $(1-t)\widehat{K}_{\mathbf{a}} + tK_{j+1}$ represents an m-plane $L_{j+1}(t)$, and those m-planes $L_{j+1}(t)$ are in general position for $0 < t \leq 1$, it follows that for each $t \in (0,1]$ the system

$$\begin{aligned} K_1 X_{\mathbf{a}} \Lambda_1^{j+1} &= 0, \\ &\vdots \\ K_j X_{\mathbf{a}} \Lambda_j^{j+1} &= 0, \\ [(1-t)\widehat{K}_{\mathbf{a}} + tK_{j+1}] X_{\mathbf{a}} \Lambda_{j+1}^{j+1} &= 0 \end{aligned}$$

has k solutions and all of them are nonsingular. Since at $t = 0$ those k solutions are also nonsingular, those k different paths of the same homotopy will lead to k different solutions at $t = 1$.

(b) General intersection conditions, where $k_i > 1$ for certain $1 \leq i \leq n$. The Pieri poset in this case is somewhat more complicated. For $k_i > 1$, let $\mathbf{a}^i$ be a derived node of $\mathbf{a}^{i-1}$. The coordinates of nodes $\mathbf{a}^{i-1}$ and $\mathbf{a}^i$ may have several different components and their differences may not simply differ by just 1. Moreover, as the following example shows, not all the nodes can be proceeded to reach final node $\mathbf{a}^n$ to be part of a complete chain.

EXAMPLE 6. *For $m = 5$, $p = 3$, and given planes $L_1, \ldots, L_5$ in general position with $\dim(L_i) = 3$, for all $i = 1, \ldots, 5$, $\sum_{i=1}^5 k_i = 15 = mp$. Furthermore,*

$$\begin{aligned} & \sigma_0 \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \\ \sim\ & [\Omega(6,7,8) \bullet \sigma_3] \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \\ \sim\ & [\Omega(3,7,8) \bullet \sigma_3] \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \\ \sim\ & [\Omega(1,6,8) + \Omega(2,5,8) + \Omega(3,4,8)] \bullet \sigma_3 \bullet \sigma_3 \bullet \sigma_3 \\ \sim\ & [2\Omega(1,3,8) + 3\Omega(1,4,7) + 2\Omega(2,4,6) + \Omega(1,5,6) + \Omega(3,4,5)] \bullet \sigma_3 \bullet \sigma_3 \\ \sim\ & [7\Omega(1,3,5) + 6\Omega(1,2,6)] \bullet \sigma_3 \\ \sim\ & 6\Omega(1,2,3). \end{aligned}$$

The Pieri poset in this case is shown in Figure 3*, and the poset consisting of complete chains is shown in Figure* 4.

Of course, only complete chains in the Pieri poset are meaningful in computing our solutions. For $\mathbf{a}^0 = (1, \ldots, p)$, let

$$\mathbf{a}^0 \longrightarrow \mathbf{a}^1 \longrightarrow \cdots \longrightarrow \mathbf{a}^n$$

be a complete chain, where $\mathbf{a}^{j+1}$ is derived from $\mathbf{a}^j$ via $\sigma_{k_{j+1}}$ for $j = 1, \ldots, n-1$. Namely, $\Omega(\mathbf{a}^j) \subset \sigma_0 \bullet \sigma_{k_1} \bullet \cdots \sigma_{k_j}$ and $\Omega(\mathbf{a}^{j+1}) \subset \sigma_0 \bullet \sigma_{k_1} \bullet \cdots \sigma_{k_{j+1}}$. When $k_i > 1$ for certain $i \in \{1, \ldots, n\}$, we will insert artificial *intermediate* nodes between nodes $\mathbf{a}^{i-1}$ and $\mathbf{a}^i$ for our algorithm as follows:

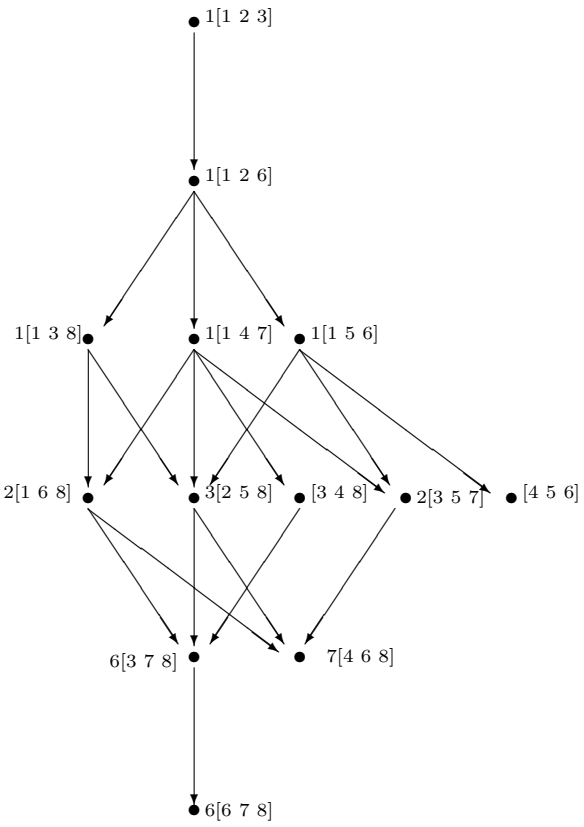


FIG. 3.

Writing

$$\mathbf{a}^{i-1} = [a_1^{(i-1)}, \ldots, a_p^{(i-1)}] \quad \text{and} \quad \mathbf{a}^{i} = [a_1^{(i)}, \ldots, a_p^{(i)}],$$

we let $l_1 = \min\{j \mid a_j^{(i-1)} < a_j^{(i)}\}$ and $\mathbf{b}^1 = (b_1^{(1)}, \ldots, b_p^{(1)})$, where

$$b_j^{(1)} = \begin{cases} a_j^{(i-1)} & \text{for } j = 1, \ldots, l_1 - 1, \\ a_{l_1}^{(i-1)} + 1 & \text{for } j = l_1 \\ a_j^{(i)} & \text{for } j = l_1 + 1, \ldots, p. \end{cases}$$

Inductively, when $\mathbf{b}^s = (b_1^{(s)}, \ldots, b_p^{(s)})$ is defined for $s < k_i - 1$, let $l_{s+1} = \min\{j \mid b_j^{(s)} < a_j^{(i)}\}$ and $\mathbf{b}^{s+1} = (b_1^{(s+1)}, \ldots, b_p^{(s+1)})$, where

$$b_j^{(s+1)} = \begin{cases} b_j^{(s)} & \text{for } j = 1, \ldots, l_{s+1} - 1, \\ b_{l_{s+1}}^{(s)} + 1 & \text{for } j = l_{s+1} \\ a_j^{(i)} & \text{for } j = l_{s+1} + 1, \ldots, p. \end{cases}$$

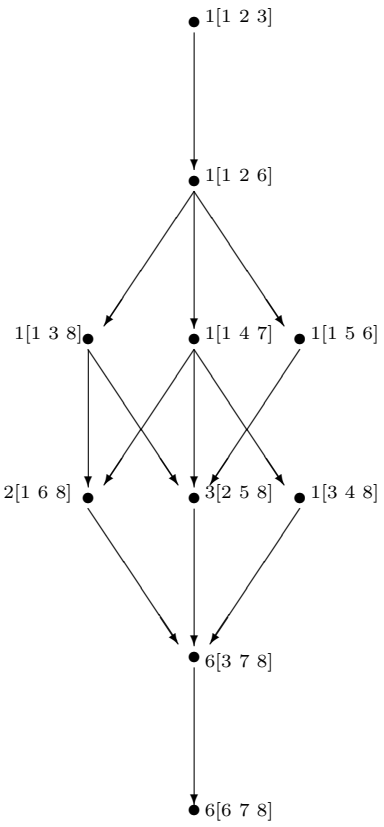


FIG. 4.

We insert those nodes $\mathbf{b}^1, \ldots, \mathbf{b}^{k_i-1}$ defined above between $\mathbf{a}^{i-1}$ and $\mathbf{a}^i$. Obviously, the coordinates of any two consecutive nodes among them can differ by 1 on only one coordinate. Therefore, we may write

$$\mathbf{a}^{i-1} := \mathbf{b}^0 \xrightarrow{\mu_1} \mathbf{b}^1 \xrightarrow{\mu_2} \cdots \xrightarrow{\mu_{k_i-1}} \mathbf{b}^{k_i} := \mathbf{a}^i,$$

where μ_j in $\mathbf{b}^{j-1} \xrightarrow{\mu_j} \mathbf{b}^j$ represents the coordinate where $\mathbf{b}^{j-1}$ and $\mathbf{b}^j$ differ.

EXAMPLE 7. *For instance, the node insertion between* [1 4 7] *and* [1 6 8] *on Figure* 4 *of Example* 6 *is*

$$[1\ 4\ 7] \xrightarrow{2} (1\ 5\ 7) \xrightarrow{2} (1\ 6\ 7) \xrightarrow{3} [1\ 6\ 8],$$

and when all intermediate nodes are inserted the poset with complete chains is shown in Figure 5.

For consecutive nodes $\mathbf{a}^{i-1}$ and $\mathbf{a}^i$ with $k_i > 1$ and the chain joining the intermediate nodes between them,

$$\mathbf{a}^{i-1} = \mathbf{b}^0 \xrightarrow{\mu_1} \mathbf{b}^1 \xrightarrow{\mu_2} \cdots \xrightarrow{\mu_{k_i-1}} \mathbf{b}^{k_i} = \mathbf{a}^i, \tag{9}$$

suppose we have solved all the variables in $X_{\mathbf{a}^{i-1}}$ as well as $\Lambda_l^{i-1} \in \mathbb{P}^{p-1}$ for $l = 1, \ldots, i-1$ for which

$$K_l X_{\mathbf{a}^{i-1}} \Lambda_l^{i-1} = 0 \qquad \text{for} \quad l = 1, \ldots, i-1. \tag{10}$$

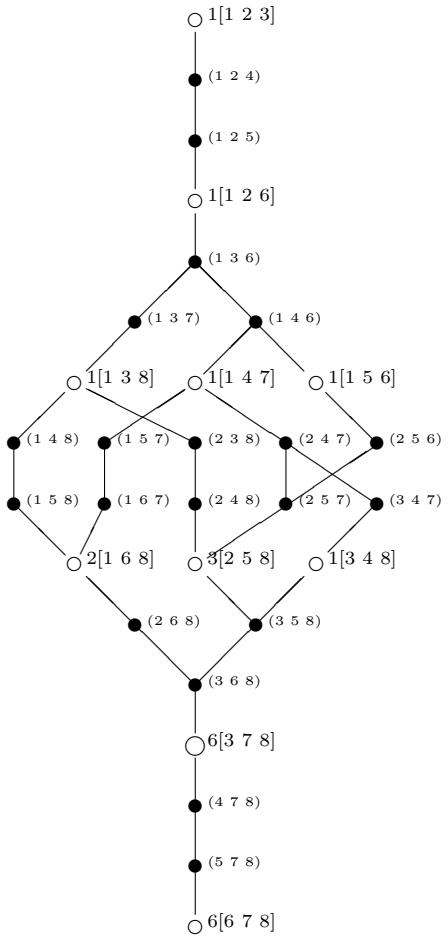


FIG. 5.

Recall that the $(p+k_j-1)\times(m+p)$ matrix K_l is a representation of the plane L_l. Let $K_i=[v_1,\ldots,v_{p+k_i-1}]^T$, where v_s for $s=1,\ldots,p+k_i-1$ are linearly independent vectors in $\mathbb{C}^{m+p}$. For

$$\mathbf{a}^{i-1}=\mathbf{b}^0 \ \xrightarrow{\mu_1}\ \mathbf{b}^1$$

consider the homotopy

$$\begin{aligned} K_1X_{\mathbf{b}^1}\Lambda_1^{i+k_i-1} &= 0,\\ &\vdots\\ K_{i-1}X_{\mathbf{b}^1}\Lambda_{i-1}^{i+k_i-1} &= 0,\\ [(1-t)\hat{K}_1^0+t\hat{K}_1^1]X_{\mathbf{b}^1}\Lambda_{i+k_i-1}^{i+k_i-1} &= 0, \end{aligned} \tag{11}$$

where for $\mathbf{b}^1=(b_1^{(1)},\ldots,b_p^{(1)})$

$$\begin{aligned} &\hat{K}_1^0:=[e_{b_1^{(1)}},\ldots,e_{b_p^{(1)}}]^T\\ \text{and}\quad &\hat{K}_1^1:=[e_{b_1^{(1)}},\ldots,e_{b_{\mu_1-1}^{(1)}},v_1,e_{b_{\mu_1+1}^{(1)}},\ldots,e_{b_p^{(1)}}]^T. \end{aligned}$$

Moreover, the μ_1th coordinate of $\Lambda_{i+k_i-1}^{i+k_i-1} \in \mathbb{P}^{p-1}$ is set to be 1, and for $l = 1, \ldots, i-1$ the coordinate of $\Lambda_l^{i+k_i-1} \in \mathbb{P}^{p-1}$ is set to be 1 if the same coordinate of $\Lambda_l^{i-1} \in \mathbb{P}^{p-1}$ is 1.

This homotopy is a deformation of square systems of size

$$p + \sum_{l=1}^{i-1}(p + k_l - 1).$$

Clearly, when $t = 0$ any solution $X_{\mathbf{a}^{i-1}}$, $\Lambda_l^{i+k_i-1}$ for $l = 1, \ldots, i-1$ of (10) coupled with $\Lambda_{i+k_i-1}^{i+k_i-1} = e_{\mu_1}$ is a solution of (11). The solutions we obtain at $t = 1$ by following the paths of the homotopy in (11) emanating from those solutions will be established as solutions of the start system of the homotopy constructed for the next step.

Inductively, write $\mathbf{b}^l = (b_1^{(l)}, \ldots, b_p^{(l)})$ for $l = 0, \ldots, k_i$ and suppose for $2 \le j \le k_i$ the system

$$\begin{aligned} K_1 X_{\mathbf{b}^{j-1}} \Lambda_1^{i+k_i-(j-1)} &= 0, \\ &\vdots \\ K_{i-1} X_{\mathbf{b}^{j-1}} \Lambda_{i-1}^{i+k_i-(j-1)} &= 0, \\ \hat{K}_{j-1}^1 X_{\mathbf{b}^{j-1}} \Lambda_{i+k_i-(j-1)}^{i+k_i-(j-1)} &= 0, \end{aligned} \tag{12}$$

where

$$\hat{K}_{j-1}^1 := [e_{b_1^{(j-1)}}, \ldots, e_{b_{\mu_{j-1}-1}^{(j-1)}}, v_{j-1}, e_{b_{\mu_{j-1}+1}^{(j-1)}}, \ldots, e_{b_p^{(j-1)}}, v_1, \ldots, v_{j-2}]^T$$

has been solved. For

$$\mathbf{b}^{j-1} \xrightarrow{\mu_j} \mathbf{b}^j$$

consider the homotopy

$$\begin{aligned} K_1 X_{\mathbf{b}^j} \Lambda_1^{i+k_i-j} &= 0, \\ &\vdots \\ K_{i-1} X_{\mathbf{b}^j} \Lambda_{i-1}^{i+k_i-j} &= 0, \\ [(1-t)\hat{K}_j^0 + t\hat{K}_j^1] X_{\mathbf{b}^j} \Lambda_{i+k_i-j}^{i+k_i-j} &= 0, \end{aligned} \tag{13}$$

where

$$\begin{aligned} &\hat{K}_1^0 = [e_{b_1^{(j)}}, \ldots, e_{b_p^{(j)}}, v_1, \ldots, v_{j-1}]^T \\ \text{and} \quad &\hat{K}_j^1 = [e_{b_1^{(j)}}, \ldots, e_{b_{\mu_j-1}^{(j)}}, v_j, e_{b_{\mu_j+1}^{(j)}}, \ldots, e_{b_p^{(j)}}, v_1, \ldots, v_{j-1}]^T. \end{aligned}$$

And, as in (11), the μ_jth coordinate of $\Lambda_{i+k_i-j}^{i+k_i-j} \in \mathbb{P}^{p-1}$ is set to be 1, and for $l = 1, \ldots, i-1$, the coordinate of $\Lambda_l^{i+k_i-j}$ is set to be 1 if the same coordinate of $\Lambda_l^{i+k_i-(j-1)}$ is 1.

This homotopy is a deformation of square system of size

$$p + j - 1 + \sum_{l=1}^{i-1}(p + k_l - 1),$$

and it is straightforward that any solution of the system in (12) induces a solution of (13) when $t = 0$. Those paths of the homotopy in (13) emanating from those solutions lead to, at $t = 1$, a set of solutions of

$$\begin{aligned} K_1 X_{\mathbf{b}^j} \Lambda_1^{i+k_i-j} &= 0, \\ &\vdots \\ K_{i-1} X_{\mathbf{b}^j} \Lambda_{i-1}^{i+k_i-j} &= 0, \\ \hat{K}_j^1 X_{\mathbf{b}^j} \Lambda_{i+k_i-j}^{i+k_i-j} &= 0. \end{aligned} \tag{14}$$

Continuing those steps successively from $j = 2$, when we reach $j = k_i$, the solutions at $t = 1$ provide a set of p-planes in the form $X_{\mathbf{a}^i}$ that meet $L_1, \ldots, L_i$.

EXAMPLE 8. *In Example* 7*, the homotopies of the chain*

$$\underbrace{[1\,2\,3] \xrightarrow{3} (1\,2\,4) \xrightarrow{3} (1\,2\,5) \xrightarrow{3} [1\,2\,6]}_{\text{level 1}}$$

$$\underbrace{\xrightarrow{2} (1\,3\,6) \xrightarrow{2} (1\,4\,6) \xrightarrow{3} [1\,4\,7]}_{\text{level 2}}$$

$$\underbrace{\xrightarrow{1} (2\,4\,7) \xrightarrow{2} (2\,5\,7) \xrightarrow{3} [2\,5\,8]}_{\text{level 3}} \longrightarrow \cdots$$

at the third level with $K_3 := [v_1, v_2, v_3, v_4, v_5]^T$ *are*

$$\begin{array}{l} [1\,4\,7] \\ \downarrow \\ (2\,4\,7) \\ \qquad : \left[\begin{array}{r} K_1 X_{(2,4,7)} \Lambda_1^5 = 0, \\ K_2 X_{(2,4,7)} \Lambda_2^5 = 0, \\ \{(1-t)[e_2, e_4, e_7]^T + t[e_2, v_5, e_7]^T\} X_{(2,4,7)} \Lambda_5^5 = 0, \end{array} \right. \\ \downarrow \\ (2\,5\,7) \\ \qquad : \left[\begin{array}{r} K_1 X_{(2,5,7)} \Lambda_1^4 = 0, \\ K_2 X_{(2,5,7)} \Lambda_2^4 = 0, \\ \{(1-t)[e_2, e_5, e_7, v_5]^T + t[e_2, e_5, v_4, v_5]^T\} X_{(2,5,7)} \Lambda_4^4 = 0, \end{array} \right. \\ \downarrow \\ [2\,5\,8] \\ \qquad : \left[\begin{array}{r} K_1 X_{[2,5,8]} \Lambda_1^3 = 0, \\ K_2 X_{[2,5,8]} \Lambda_2^3 = 0, \\ \{(1-t)[e_2, e_5, e_8, v_4, v_5] + t[v_1, v_2, v_3, v_4, v_5]\} X_{[2,5,8]} \Lambda_3^3 = 0, \end{array} \right. \end{array}$$

where

$$\Lambda_1^5, \Lambda_1^4, \Lambda_1^3 = \begin{bmatrix} * \\ * \\ 1 \end{bmatrix}, \quad \Lambda_2^5, \Lambda_2^4, \Lambda_2^3 = \begin{bmatrix} * \\ * \\ 1 \end{bmatrix},$$

$$\textit{and} \quad \Lambda_5^5 = \begin{bmatrix} 1 \\ * \\ * \end{bmatrix}, \quad \Lambda_4^4 = \begin{bmatrix} * \\ 1 \\ * \end{bmatrix}, \quad \Lambda_3^3 = \begin{bmatrix} * \\ * \\ 1 \end{bmatrix}.$$

Write $\widehat{K}_1^t = (1-t)[e_2, e_4, e_7]^T + t[e_2, v_5, e_7]^T$. *Then,*

$$\widehat{K}_1^t X_{(247)}\Lambda_5^5 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ x_1 & 1 & 0 \\ & x_2 & 1 \\ & x_3 & x_4 \\ & & x_5 \\ 0 & & \\ & & x_6 \\ & & x_7 \\ & & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & 1 & 0 \\ \star & \star & \star \\ 0 & 0 & x_7 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Obviously, $\lambda_3 = 0$ *for all* $t \in [0,1]$, *and, as assigned,* $\lambda_1 = 1$ *for all* $t \in [0,1]$. *Let* $x_1^{(1)}, \ldots, x_7^{(1)}, \lambda_1^{(1)}(=1), \lambda_2^{(1)}, \lambda_3^{(1)}(=0)$ *be a solution of* $\widehat{K}_1^1 X_{(247)}\Lambda_5^5 = 0$. *Now, for* $\widehat{K}_2^t := (1-t)[e_2, e_5, e_7, v_5]^T + t[e_2, e_5, v_4, v_5]$ *at* $t = 0$,

$$\widehat{K}_2^0 X_{(257)}\Lambda_4^4 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ * & * & * & * & * & * & * & * \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ x_1 & 1 & 0 \\ & x_2 & 1 \\ & x_3 & x_4 \\ & y & x_5 \\ 0 & & \\ & & x_6 \\ & & x_7 \\ & & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & 1 & 0 \\ 0 & y & x_5 \\ 0 & 0 & x_7 \\ \star & \star & \star \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

As assigned, $\lambda_2 = 1$ *and obviously* $\lambda_3 = 0$, *and the new variable* y *must be zero. And, since*

$$x_1\lambda_1 + 1 = 0,$$

$x_l = x_l^{(1)}$ *for* $l = 1, \ldots, 7$, *along with* $\lambda_1 = -\frac{1}{x_1^{(1)}}$, $\lambda_2 = 1$, $\lambda_3 = 0$, *is a solution of* $\widehat{K}_2^0 X_{(257)}\Lambda_4^4 = 0$. *Similarly, with* $\widehat{K}_2^1 = [e_2, e_5, v_4, v_5]^T$, *let* $x_1^{(2)}, \ldots, x_7^{(2)}, y^{(2)}, \lambda_1^{(2)}$,

$\lambda_2^{(2)}(=1), \lambda_3^{(2)}$ *be a solution of*

$$\widehat{K}_2^1 X_{(257)} \Lambda_4^4 = \begin{bmatrix} 0&1&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ *&*&*&*&*&*&*&*\\ *&*&*&*&*&*&*&* \end{bmatrix} \begin{bmatrix} 1&0&0\\ x_1&1&0\\ &x_2&1\\ &x_3&x_4\\ 0&y&x_5\\ &&x_6\\ &&x_7\\ &&0 \end{bmatrix} \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1&1&0\\ 0&y&x_5\\ \star&\star&\star\\ \star&\star&\star \end{bmatrix} \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 0\\0\\0\\0 \end{bmatrix}.$$

Then for $\widehat{K}_3^t := (1-t)[e_2, e_5, e_8, v_4, v_5]^T + t[v_1, v_2, v_3, v_4, v_5]$ *at* $t = 0$,

$$\widehat{K}_3^0 X_{[258]} \Lambda_3^3 = \begin{bmatrix} 0&1&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&1\\ *&*&*&*&*&*&*&*\\ *&*&*&*&*&*&*&* \end{bmatrix} \begin{bmatrix} 1&0&0\\ x_1&1&0\\ &x_2&1\\ &x_3&x_4\\ &y&x_5\\ 0&&x_6\\ &&x_7\\ &&z \end{bmatrix} \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1&1&0\\ 0&y&x_5\\ 0&0&z\\ \star&\star&\star\\ \star&\star&\star \end{bmatrix} \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 0\\0\\0\\0\\0 \end{bmatrix}.$$

Then $\lambda_3 = 1$, *as assigned, implies* $z = 0$. *On the other hand, since*

$$y\lambda_2 + x_5 = 0 \text{ and } x_1\lambda_1 + \lambda_2 = 0,$$

$x_l = x_l^{(2)}$ *for* $l = 1, \ldots, 7$, $y = y^{(2)}$, $\lambda_2 = -\frac{x_5^{(2)}}{y^{(2)}}$, *and* $\lambda_1 = -\frac{\lambda_2}{x_1^{(2)}}$ *is a solution of* $\widehat{K}_3^0 X_{[258]} \Lambda_3^3 = 0$. ∎

Remark 2. In Example 8, $\widehat{K}_1^1$ defines a 5-plane L_3^1 containing L_3, $\widehat{K}_2^1$ defines a 4-plane L_3^2 containing L_3, and $\widehat{K}_3^1 = K_3$ represents L_3. So the strategy behind the homotopies we construct between intermediate nodes is the following. To find the 3-planes in the form of $X_{[258]}$ which meet L_1, L_2, L_3 (those $X_{[258]}$ are in $\sigma_3 \bullet \sigma_3 \bullet \sigma_3 \bullet \Omega(2,5,8)$), we first find the 3-planes $X_{(247)}$ which meet L_1, L_2, L_3^1 (those $X_{(247)}$ are in $\sigma_3 \bullet \sigma_3 \bullet \sigma_1 \bullet \Omega(2,4,7)$). Then we find the 3-planes $X_{(257)}$ which meet L_1, L_2, L_3^2 (those X's are in $\sigma_3 \bullet \sigma_3 \bullet \sigma_2 \bullet \Omega(2,5,7)$). Ultimately, we find the 3-planes $X_{[258]}$ which meet L_1, L_2, L_3.

4. Numerical results. An implementation of a previous version of the Pieri homotopy algorithm for numerical Schubert calculus [4] exists in the module of the extended version of PHCpack in [10] that also provides the SAGBI homotopy proposed in [3] for solving general problems in enumerative geometry numerically. In general, as reported in [4], the Pieri homotopy algorithms are much superior in speed as well as the range of applications than the SAGBI homotopies. We therefore compare only the results of the implementation of our algorithm with those of the code in PHCpack. All computations were carried out on a 400 MHz Intel Pentium II CPU with 256 MB of RAM, running on SunOS 5.6. In all the tables below #hty represents the total number of homotopies we followed in the corresponding cases, and **Wu** is the symbol representing our code.

1. **$k_i = 1$ for all i, as shown in Table 1.**

TABLE 1

m	p	**#soln**	#hty	**Wu**	**PHC**
3	2	**5**	21	220ms	720ms
4	2	**14**	63	1s270ms	5s50ms
3	3	**42**	183	13s420ms	41s480ms
5	2	**42**	195	8s870ms	39s870ms
6	2	**132**	360	13s330ms	1m13s
7	2	**429**	1196	1m16s	8m14s
4	3	**462**	1110	1m58s	8m59s
8	2	**1,430**	4,056	7m44s	53m2s
9	2	**4,862**	13,988	38m11s	6h29m1s
5	3	**6,006**	14,683	57m59s	7h53m28s
6	3	**87,516**	217,276	28h44m13s	-

The code in PHCpack requires a much bigger RAM than our code in all of the cases. For instance, for $(m, p) = (4, 3)$ above, PHC needs more than 7,044KB whereas **Wu** only needs 996KB.

2. **$k_i > 1$ for certain i's, as shown in Tables 2–10.** The first column of each table shows the numbers of all those k_i's.

TABLE 2
$(m, p) = (3, 2)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
321	**1**	6	40ms	470ms
222	**1**	6	60ms	550ms
2211	**2**	9	80ms	1s30ms
21111	**3**	13	130ms	2s290ms

TABLE 3
$(m, p) = (3, 3)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
333	**1**	9	160ms	2s250ms
3222	**1**	9	250ms	5s70ms
33111	**1**	9	200ms	4s420ms
32211	**2**	13	490ms	8s120ms
22221	**3**	21	870ms	10s480ms
222111	**6**	32	1s160ms	20s670ms
2211111	**11**	50	3s310ms	42s190ms
21111111	**21**	92	7s80ms	1m10s830ms

TABLE 4
$(m, p) = (4, 2)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
2222	**3**	20	220ms	6s750ms
3311	**2**	12	150ms	4s730ms
4211	**1**	8	60ms	3s70ms
32111	**3**	16	240ms	6s850ms
41111	**1**	8	80ms	2m510ms
221111	**6**	30	680ms	14s240ms
311111	**4**	20	380ms	8s970ms
2111111	**9**	41	910ms	16s460ms

TABLE 5
$(m, p) = (4, 3)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
44211	**1**	12	330ms	25s550ms
43311	**2**	16	750ms	53s700ms
43221	**2**	17	730ms	1m9s320ms
33222	**4**	29	1s730ms	1m39s800ms
222222	**16**	120	7s340ms	3m57s880ms
2222211	**26**	166	15s510ms	9m26s530ms
22221111	**45**	226	25s300ms	15m20s820ms
222111111	**79**	360	49s840ms	25m8s680ms
2211111111	**140**	622	1m35s740ms	41m42s700ms
21111111111	**252**	1,112	3m34s200ms	1h16m48s270ms

TABLE 6
$(m, p) = (5, 3)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
54321	**2**	20	1s240ms	6m0s660ms
44421	**3**	30	2s50ms	8m12s730ms
44322	**4**	37	3s340ms	9m23s770ms
43332	**5**	49	4s230ms	11m30s110ms
33333	**6**	65	5s80ms	10m21s130ms
543111	**3**	23	1s660ms	6m46s850ms
5421111	**4**	28	1s970ms	9m55s900ms
333321	**14**	118	12s540ms	24m29s30ms
3222222	**60**	451	1m2s180ms	1h14m22s370ms

TABLE 7
$(m, p) = (5, 2)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
4222	**2**	16	270ms	24s560ms
5311	**1**	10	210ms	10s280ms
3322	**3**	23	360ms	35s40ms
22222	**6**	44	1s20ms	1m14s520ms

TABLE 8
$(m, p) = (5, 4)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
44444	**1**	20	2s490ms	49m2s760ms
553322	**3**	33	6s480ms	1h48m10s470ms
443333	**9**	102	22s170ms	2h50m18s640ms
544322	**4**	42	7s950ms	2h14m59s90ms
4443221	**18**	145	45s710ms	-
4433222	**32**	261	1m25s430ms	-
2222222222	**3,396**	25,938	5h4m39s444ms	-

TABLE 9
$(m, p) = (6, 3)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
333333	**40**	413	1m6s560ms	3h22m45s430ms
443322	**24**	208	30s920ms	-
433332	**30**	286	40s650ms	-
3333222	**104**	830	2m20s260ms	-
222222222	**876**	6,547	30m22s470ms	-

TABLE 10
$(m, p) = (6, 4)$

$[k_1, \ldots, k_n]$	**#soln**	#hty	**Wu**	**PHC**
664422	**3**	37	10s0ms	-
654333	**6**	70	22s850ms	-
554433	**10**	123	45s870ms	-
444444	**15**	220	1m10s440ms	22h58m54s70ms
33333333	**790**	8,413	1h15m45s778ms	>148.5h

As we can see from the results above, our novel approach of employing the Pieri homotopy algorithm for the numerical Schubert calculus has made a considerable advance in speed. And, in all the cases we have tried, the storage requirement for our code is much smaller than that of the existing code. The algorithm is particularly valuable when $k_i > 1$ appears.

REFERENCES

[1] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. II., Cambridge University Press, Cambridge, UK, 1968.

[2] S. KLEIMAN AND D. LAKSOV, *Schubert calculus*, Amer. Math. Monthly, 79 (1974), pp. 1061–1082.

[3] B. HUBER, F. SOTTILE, AND B. STURMFELS, *Numerical Schubert calculus*, J. Symbolic Comput., 26 (1998), pp. 767–788.

[4] B. HUBER AND J. VERSCHELDE, *Pieri homotopies for problems in enumerative geometry applied to pole placement in linear systems control*, SIAM J. Control Optim., 38 (2000), pp. 1265–1287.

[5] T. Y. LI, T. SAUER, AND J. A. YORKE, *The cheater's homotopy: An efficient procedure for solving systems of polynomial equations*, SIAM J. Numer. Anal., 26 (1989), pp. 1241–1251.

[6] A. P. MORGAN AND A. J. SOMMESE, *Coefficient-parameter polynomial continuation*, Appl. Math. Comput., 29 (1989), pp. 123–160.

[7] A. P. MORGAN AND A. J. SOMMESE, *Errata: "Coefficient-parameter polynomial continuation"* [Appl. Math. Comput. 29 (1989), no. 2, part II, 123–160], Appl. Math. Comput., 51 (1992), p. 207.

[8] A. J. SOMMESE AND C. W. WAMPLER, *Numerical algebraic geometry*, in The Mathematics

of Numerical Analysis, Park City, Utah, 1995, J. Renegar, M. Shub, and S. Smale, eds., Lectures in Appl. Math. 32, AMS, Providence, RI, 1996, pp. 749–763.

[9] F. Sottile, *Pieri's formula via explicit rational equivalence*, Canad. J. Math., 49 (1997), pp. 1281–1298.

[10] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 601–628

THE MORTAR ELEMENT METHOD WITH OVERLAPPING SUBDOMAINS*

YVES ACHDOU† AND YVON MADAY‡

Abstract. In this paper, we discuss an extension of the mortar element method to overlapping subdomains and nonmatching grids in the overlapped zones. We discuss in particular the case where more than two subdomains overlap. The method is described and analyzed, as are some preconditioned iterative methods for solving the linear systems arising from this discretization.

Key words. overlapping domain decomposition method, nonconforming method

AMS subject classifications. 65N30, 65N15, 65N55

PII. S0036142900375256

1. Introduction. Mortar element methods were introduced in [4] for nonoverlapping domain decompositions in order to couple different variational approximations in different subdomains. In the finite element context, one important advantage of the mortar element methods is that they allow for using structured grids in subdomains and thus are fast solvers [1]. The resulting methods are nonconforming but still yield optimal approximations. The literature on the mortar element methods is growing numerous; see [2] and references therein.

In this paper, we shall discuss the case of overlapping subdomains, with meshes constructed in an independent manner in each subdomain. As pointed out by F. Hecht, J. L. Lions, and O. Pironneau [12] and J.-L. Lions and O. Pironneau [15], such a situation can occur if the domain of computation is a scene constructed by constructive solid geometry in image synthesis and virtual reality: each object of the scene is described by set operations on primitive shapes like cubes, cylinders, spheres, and cones. With VRML (the language of virtual reality), the objects may be described as unions of more elementary objects with primitive shapes, which are never intersected, so it is not possible to construct a global mesh. Each simple object must have its individual mesh. In [12, 15], many algorithms (including algorithms from control theory) for this situation are proposed and cover cases more general than overlapping subdomains (domain with holes, for example).

We also note that independent of the development of the mortar methods, overlapping domain decomposition with nonmatching grids has been used for finite difference discretizations in the engineering community: these methods are often referred to as the *chimera methods*; see [6, 18]. We refer to [16] and the references therein for the numerical analysis of these methods.

To our knowledge, mortar methods with overlapping subdomains have been proposed first by Y. Kuznetsov [13], who focused on iterative solvers with Lagrange multipliers. For two overlapping subdomains, the mortar method has been analyzed by X. C. Cai, M. Dryja, and M. Sarkis [5] in two dimensions. They have considered two subdomains, with nonmatching grids and piecewise linear Lagrange finite ele-

*Received by the editors July 19, 2000; accepted for publication (in revised form) August 20, 2001; published electronically June 12, 2002.
http://www.siam.org/journals/sinum/40-2/37525.html
†UFR Mathématiques, Université Paris 7, Case 7012, 75251 Paris Cedex 05, France and Laboratoire d'Analyse Numérique, Université Paris 6, Paris, France (achdou@math.jussieu.fr).
‡Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Boîte courrier 187, 75252 Paris Cedex 05, France (maday@ann.jussieu.fr).

ments. In particular, they have considered the case when the overlapping parameter is 0 (two rectangular subdomains for an L shaped domain). They have also proposed iterative solvers and preconditioners for the linear systems arising from the mortar discretization.

In this paper, we generalize their method in two dimensions, with more than two subdomains. We shall see that technical difficulties arise when the boundaries of two subdomains cross each other. For simplicity, we consider the Laplace equation and rule out the case when the overlap may vanish. For such situations, one should mix the method described in [5] and the one below. Also, we deal with first order Lagrange elements, but, with some effort, the ideas below may be generalized to higher order finite elements.

The paper is organized as follows. In section 2, we propose some mortar discretizations, and we study the ellipticity of the discrete problems. Section 3 is devoted to an error analysis following the lines of [4]. In section 4, we propose an alternative matching condition. In section 5, we study additive Schwarz preconditioners (see [14, 17] for reviews) for the linear system arising from the mortar discretizations: we generalize one of the preconditioners proposed in [5] which is not optimal but fairly easy to implement, and we propose another optimal preconditioner.

2. The discretization.

2.1. First definitions. In all of what follows, c or C will stand for various positive constants, independent of the geometric parameters.

We consider a polygonal domain Ω of $\mathbb{R}^2$ and the model boundary value problem in Ω:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

We consider first a family of overlapping subdomains $(\Omega_k)_{k\in\{1,\dots,K\}}$ with polygonal shapes covering Ω:

$$\Omega = \bigcup_{k=1}^{K} \Omega_k. \tag{2.2}$$

We denote by $(\Gamma_k^l)_{1\le l\le E_k}$ the sides of the polygonal boundary $\partial\Omega_k$.

We denote by H_k the diameter of Ω_k and by H the maximal diameter $H = \max_{1\le k\le K} H_k$. We assume that there exists a constant c such that for any k, $1 \le k \le K$, $cH \le H_k \le H$. We also suppose that there exists a positive constant τ such that any subdomain Ω_k contains a ball of diameter greater than τH.

For any subdomain Ω_k, we denote by δ_k the minimum distance of overlap between Ω_k and $\cup_{i\neq k}\Omega_i$:

$$\delta_k = \inf_{x\in\Omega_k\setminus\cup_{i\neq k}\Omega_i} \inf_{y\in\cup_{i\neq k}\Omega_i\setminus\Omega_k} |x-y|.$$

We also define $\delta \equiv \min_k \delta_k$.

Assumption 1. We assume that the intersection of two subdomains' boundaries can only be isolated points, called crosspoints. We assume that there exists a constant α, $0 < \alpha \le \frac{\pi}{2}$, such that the angles (taken not greater than $\frac{\pi}{2}$) between two subdomains' boundaries crossing each other are all greater than α. For simplicity, we assume also that a given crosspoint is neither the intersection of more than two subdomains' boundaries nor the vertex of a subdomain.

Assumption 2. We assume that there exists a constant N_1 such that, for any ball B of diameter H, $B \cap \Omega$ is covered by at most N_1 subdomains.

This assumption yields two important consequences.

PROPERTY 1. *We denote by ω_k the union of the subdomains intersecting Ω_k and by $\mathcal{I}_k$ the set of the integers i such that $\Omega_i \subset \omega_k$. There exists a constant $n_1(N_1)$ such that, for any k, $1 \le k \le K$, cardinal$(\mathcal{I}_k) \le n_1(N_1)$.*

PROPERTY 2. *There exists a constant $n_2(N_1)$ such that the number of subdomains containing a given point in Ω is bounded by $n_2(N_1)$.*

We also make the following assumption.

Assumption 3. There exists a constant N_2 such that the number of sides E_k of a given subdomain Ω_k is smaller than N_2.

As a consequence, we have the following property.

PROPERTY 3. *The number of crosspoints lying on $\partial\Omega_k$ is bounded by a constant.*

On each subdomain Ω_k, we have a family of triangular meshes $\mathcal{T}_{k,h_k}$ whose triangles have maximal diameters h_k. The meshes are constructed in an independent manner. The mesh nodes on $\partial\Omega_k$ need not match with the mesh nodes in the adjacent subdomains. We assume that the families $(\mathcal{T}_{k,h_k})_{h_k}$ are shape regular and quasi uniform; see [7]. We agree to simplify the notations by replacing $\mathcal{T}_{k,h_k}$ with $\mathcal{T}_k$.

Assumption 4. We call $h = \max_k h_k$, and we assume that, for a positive constant C,

$$h < C\delta. \tag{2.3}$$

Associated with the mesh $\mathcal{T}_k$, we consider the spaces Z_k and X_k of piecewise linear Lagrange finite elements:

$$Z_k \equiv \left\{ \begin{array}{l} u_k \text{ is continuous in } \overline{\Omega}_k, \\ \forall t \in \mathcal{T}_k, u_k|_t \text{ is linear} \end{array} \right\}, \quad X_k \equiv \{u_k \in Z_k, u_k = 0 \text{ on } \partial\Omega \cap \partial\Omega_k\}.$$

Each space X_k and Z_k is supplied with its usual nodal basis functions. We define $X = \Pi_{k=0}^{K} X_k$. The vectors $u = (u_k)_{k\in\{1,\dots,K\}}$ of X are collections of functions defined in the subdomains, but no continuity constraint is imposed at the subdomains boundaries. The nodal basis of X can be found by taking the K-tuple of the nodal bases of the spaces X_k.

For a side Γ_k^l of $\partial\Omega_k$, we denote by Z_k^l the space of functions obtained by taking the trace on Γ_k^l of the functions of Z_k and by $\mathcal{T}_k^l$ the trace of the mesh $\mathcal{T}_k$ on Γ_k^l. Thus $\mathcal{T}_k^l$ is composed of elementary segments that are the sides of some triangles of $\mathcal{T}_k$. The space Z_k^l is the space of piecewise linear Lagrange finite elements on $\mathcal{T}_k^l$.

REMARK 1. *All the assumptions on the domain decomposition are not too stringent. Indeed, in the cases of interest described in the introduction, the decomposition is done a priori before constructing the mesh and therefore not through an automatic mesh partitioner. In addition, two other assumptions on the meshes will be made below.*

2.2. The matching condition. In order to discretize (2.1), we need to define a subspace Y of X by imposing weak continuity constraints at the subdomain boundaries $\partial\Omega_k$, $1 \le k \le K$.

For a side Γ_k^l of $\partial\Omega_k \backslash \partial\Omega$, we denote by $\widetilde{W_k^l}$ the subspace of Z_k^l of the functions whose restrictions to the extreme elementary segments (the first and the last) of $\mathcal{T}_k^l$ are constant; see Figure 2.1. Such spaces are used as mortar spaces for the nonoverlapping case (see [4]). Here, we will have to additionally modify them locally, near the crosspoints.

FIG. 2.1. *The nodal bases of the spaces Z_k^l and $\widetilde{W_k^l}$; here, the dimension of Z_k^l is five.*

Let $(j_i)_{i\in\{1,\dots,n_k^l\}}$ be the family of the indices such that $|\Gamma_k^l \cap \Omega_{j_i}| > 0$ and $j_i \neq k$. The number n_k^l is the number of subdomains covering Γ_k^l. Note that, from Property 1, n_k^l is bounded by $n_1(N_1)$. For $i \in \{1,\dots,n_k^l\}$, we define $\Gamma_k^{l,i} = \Gamma_k^l \cap \Omega_{j_i}$. From (2.2), we have the following overlapping decomposition:

$$\Gamma_k^l = \bigcup_{i=1}^{n_k^l} \Gamma_k^{l,i}.$$

Call $p_k^l(x)$ the piecewise constant counting function defined on Γ_k^l by

$$p_k^l(x) = \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}}(x), \tag{2.4}$$

which represents the number of subdomains covering x. Here, $1_{\Gamma_k^{l,i}}$ is the characteristic function of $\Gamma_k^{l,i}$ (and is equal to one if $x \in \Gamma_k^{l,i}$ and is zero otherwise). From (2.2) and Property 2, p_k^l is greater than or equal to one and bounded from above by a constant.

Given W_k^l, a space of test functions defined on Γ_k^l, the first possible matching condition on Γ_k^l is of the form

$$\forall w \in W_k^l, \quad \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} \left(u_k(x) - \frac{1}{p_k^l(x)} \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) \right) w(x)dx = 0. \tag{2.5}$$

Basically, the space W_k^l will be a subspace of $\widetilde{W_k^l}$, and the spaces will differ essentially due to the presence of crosspoints.

It now remains to define the space W_k^l. Suppose that for $i \in \{1,\dots,n_k^l\}$, $\Gamma_k^l \cap \partial\Omega_{j_i} \neq \emptyset$.[1] Then, from Assumption 1, we know that the intersections do not take place at a vertex of $\partial\Omega_{j_i}$ and let $\Gamma_{j_i}^{l'}$ be the side of $\partial\Omega_{j_i}$ such that $\Gamma_k^l \cap \Gamma_{j_i}^{l'}$ is a point denoted by x^*. If no special care is taken for the choices of W_k^l and $W_{j_i}^{l'}$, then the matching condition (2.5) on Γ_k^l and $\Gamma_{j_i}^{l'}$ will strongly couple the degrees of freedom (d.o.f.) of u_k and u_{j_i} near the crosspoint x^*, and there might be cases when these conditions are too restrictive; i.e., the functions u_k and u_{j_i} must be constant—even zero—near x^*. To avoid such a situation, and also in order to get a solver with good parallel properties (see section 5), we have to relax the weak continuity condition near x^*.

We call $(x_m)_{m\in\{1,\dots,M_k^l\}}$ the nodes of $\mathcal{T}_k^l$ different from the endpoints of Γ_k^l and $(\phi_m)_{m\in\{1,\dots,M_k^l\}}$ (resp., $(\psi_m)_{m\in\{0,\dots,M_k^l+1\}}$) the nodal basis functions of $\widetilde{W_k^l}$ (resp., of Z_k^l). Note that $\phi_m = \psi_m$ for $2 \le m \le M_k^l - 1$ and $\phi_1 = \psi_0 + \psi_1$ and $\phi_{M_k^l} = \psi_{M_k^l} + \psi_{M_k^l+1}$.

[1] The situation $\Gamma_k^l \cap \partial\Omega_{j_i} = \emptyset$ corresponds to $\Gamma_k^l \subset \Omega_{j_i}$.

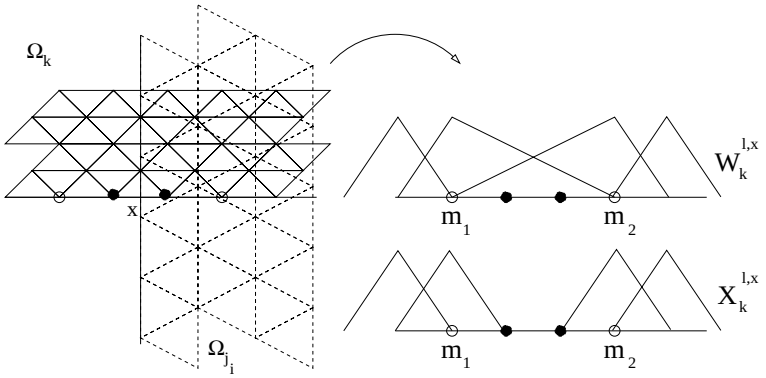


FIG. 2.2. *The nodal bases of the spaces W_k^{l,x^*} and X_k^{l,x^*}. (Only two subdomains have been represented.)*

We select the nodes of $\mathcal{T}_k^l$ for which the support of the associated basis function of X_k does not intersect $\Gamma_{j_i}^{l'}$: we obtain the set of nodes $(x_m)_{m\in\{1,\ldots,m_1\}\cup\{m_2,\ldots,M_k^l\}}$. We call $\widetilde{\phi}_{m_1}$ the continuous function vanishing outside (x_{m_1-1}, x_{m_2}), linear on (x_{m_1-1}, x_{m_1}) and on (x_{m_1}, x_{m_2}), such that $\widetilde{\phi}_{m_1}(x_{m_1}) = 1$. Likewise, $\widetilde{\phi}_{m_2}$ is the continuous function vanishing outside (x_{m_1}, x_{m_2+1}), linear on (x_{m_1}, x_{m_2}) and on (x_{m_2}, x_{m_2+1}), such that $\widetilde{\phi}_{m_2}(x_{m_2}) = 1$. The space W_k^{l,x^*} is defined by

$$W_k^{l,x^*} \equiv span(\phi_1, \ldots, \phi_{m_1-1}, \widetilde{\phi}_{m_1}, \widetilde{\phi}_{m_2}, \phi_{m_2+1}, \ldots, \phi_{M_k^l}). \tag{2.6}$$

The space W_k^{l,x^*} is displayed in Figure 2.2. For what follows, we also define the space

$$\begin{aligned} X_k^{l,x^*} &\equiv \{u \in Z_k^l, u = 0 \text{ at the endpoints of } \Gamma_k^l \text{ and } x_{m_1+1}, \ldots, x_{m_2-1}\} \\ &= span(\psi_1, \ldots, \psi_{m_1}, \psi_{m_2}, \ldots, \psi_{M_k^l}). \end{aligned} \tag{2.7}$$

DEFINITION 2.1. *For the crosspoint x^*, we define the* zone of influence *of x^* on Γ_k^l as the interval (x_{m_1-1}, x_{m_2+1}). We also define the* zone of influence *of a vertex x^* of Ω_k on Γ_k^l as the union of the two elements of $\mathcal{T}_k^l$ next to x^*. From Assumption* 1*, the zone of influence of a crosspoint has a size smaller than Ch.*

Assumption 5. The zones of influence of two crosspoints on Γ_k^l are disjoint. Moreover, the zones of influence on Γ_k^l of a crosspoint and a vertex of Ω_k are disjoint.

Finally, we define $\mathcal{X}_k^l$ as the set of crosspoints on Γ_k^l, and we set

$$W_k^l \equiv \bigcap_{x^*\in\mathcal{X}_k^l} W_k^{l,x^*} \tag{2.8}$$

and, likewise,

$$X_k^l \equiv \bigcap_{x^*\in\mathcal{X}_k^l} X_k^{l,x^*}, \tag{2.9}$$

and Y is the subspace of X defined by

$$Y \equiv \{u \in X;\ \forall k \in \{1, \ldots, K\},\ \forall l \in \{1, \ldots, E_k\}, u \text{ satisfies } (2.5)\} \tag{2.10}$$

for W_k^l defined by (2.8) and (2.6).

REMARK 2. *The functions in W_k^l will resemble those of $\widetilde{W_k^l}$ except at a few nodes near crosspoints. Furthermore, from Assumption* 5*, these exceptional regions around crosspoints are disjoint.*

REMARK 3. *The spaces W_k^l and X_k^l have the same dimension.*

Let $\mathcal{V}_k$ be the set of the nodes containing

1. the vertices of $\partial\Omega_k$,
2. all the other nodes of $\mathcal{T}_k$ on $\partial\Omega_k$ such that the support of the associated nodal basis function of X_k intersects another subdomain's boundary.

Let u be a function in $L^2(\Gamma_k^l)$. Under a technical but reasonable assumption on the mesh, the following problem is well posed: find $u_k^l \in Z_k^l$ such that

$$\begin{aligned} &u_k^l \text{ is given at the nodes of } \Gamma_k^l \cap \mathcal{V}_k, \\ &\forall w_k^l \in W_k^l, \quad \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} u_k^l(x) w_k^l(x) dx = \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} u(x) w_k^l(x) dx. \end{aligned} \tag{2.11}$$

This is a corollary of the following lemma.

LEMMA 2.2. *For a given crosspoint x^* on Γ_k^l, let $(x_m)_{m\in\{1,\dots,m_1\}\cup\{m_2,\dots,M_k^l\}}$ be the nodes of $\mathcal{T}_k^l$ involved in the above construction of W_k^{l,x^*}. Let δ^{--}, δ^-, δ^+, and δ^{++} be defined by $\delta^{--} = \frac{x_{m_1}-x_{m_1-1}}{x_{m_2}-x_{m_1}}$, $\delta^- = \frac{x_{m_1+1}-x_{m_1}}{x_{m_2}-x_{m_1}} < 1$, $\delta^+ = \frac{x_{m_2}-x_{m_2-1}}{x_{m_2}-x_{m_1}} < 1$, and $\delta^{++} = \frac{x_{m_2+1}-x_{m_2}}{x_{m_2}-x_{m_1}}$. Assume that there exists a constant c such that for all crosspoints x^*,*

$$\begin{aligned} &\tfrac{3}{2}\delta^- + \delta^{--} - (\delta^+)^2 \geq c, \\ &\tfrac{3}{2}\delta^+ + \delta^{++} - (\delta^-)^2 \geq c, \end{aligned} \tag{2.12}$$

and there exists a constant C independent of h such that

$$\inf_{u\in X_k^l} \sup_{\substack{w\in W_k^l \\ w\neq 0}} \frac{\int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} u(x) w(x) dx}{\|w\|_{L^2(\Gamma_k^l)}} \geq C \|u\|_{L^2(\Gamma_k^l)}. \tag{2.13}$$

Proof. If there is no crosspoint on Γ_k^l, then the result can be found in [2, Lemma 1] and its proof. In the opposite case, consider the tridiagonal matrix M of the bilinear form in the nodal bases of $X_k^l \times W_k^l$. We wish to prove that there exists a constant C such that the symmetrized matrix $\frac{1}{2}(M + M^T)$ has all its eigenvalues not smaller than Ch. This will be a consequence of the stronger property: for any i,

$$M_{i,i} - \frac{1}{2}(M_{i,i+1} + M_{i+1,i} + M_{i,i-1} + M_{i-1,i}) \geq Ch. \tag{2.14}$$

The estimate (2.14) is true if the node corresponding to the ith nodal basis function of X_k^l and W_k^l is not contained in the region of influence of a crosspoint; see the proof of [2, Lemma 1]. In the opposite case, there is exactly one crosspoint x^* in the support of the ith nodal basis function of W_k^l; assume with no restrictions that in the neighborhood of x^*, $p_k^l(s) = r+1$ for $s < x^*$, and $p_k^l(s) = r$ for $s > x^*$. We have $W_k^{l,x^*} = span(\phi_1, \dots, \phi_{m_1-1}, \widetilde{\phi}_{m_1}, \widetilde{\phi}_{m_2}, \phi_{m_2+1}, \dots, \phi_{M_k^l})$ and $X_k^{l,x^*} = span(\psi_1, \dots, \psi_{m_1}, \psi_{m_2}, \dots, \psi_{M_k^l})$. Without restriction, we can assume that

the ith (resp., $(i+1)$th) nodal basis function of W_k^l is $\widetilde{\phi}_{m_1}$ (resp., $\widetilde{\phi}_{m_2}$). We have

$$
\begin{aligned}
&M_{i,i-1} = M_{i-1,i} = \tfrac{x_{m_1}-x_{m_1-1}}{6(r+1)},\\
&M_{i,i} = \tfrac{x_{m_1}-x_{m_1-1}}{3(r+1)} + (x_{m_1+1}-x_{m_1})\left(\tfrac{1}{2(r+1)} - \tfrac{x_{m_1+1}-x_{m_1}}{6(r+1)(x_{m_2}-x_{m_1})}\right),\\
&M_{i,i+1} = \tfrac{1}{6r}\tfrac{(x_{m_2}-x_{m_2-1})^2}{x_{m_2}-x_{m_1}}
\end{aligned}
$$

and

$$
\begin{aligned}
&M_{i+1,i+2} = M_{i+2,i+1} = \tfrac{x_{m_2+1}-x_{m_2}}{6r},\\
&M_{i+1,i+1} = \tfrac{x_{m_2+1}-x_{m_2}}{3r} + (x_{m_2}-x_{m_2-1})\left(\tfrac{1}{2r} - \tfrac{x_{m_2}-x_{m_2-1}}{6r(x_{m_2}-x_{m_1})}\right),\\
&M_{i+1,i} = \tfrac{1}{6(r+1)}\tfrac{(x_{m_1+1}-x_{m_1})^2}{x_{m_2}-x_{m_1}}.
\end{aligned}
$$

Then (2.14) for i and $i+1$ is equivalent to

(2.15)
$$\frac{x_{m_1}-x_{m_1-1}}{6(r+1)} + \frac{x_{m_1+1}-x_{m_1}}{2(r+1)}\left(1-\frac{1}{2}\frac{x_{m_1+1}-x_{m_1}}{(x_{m_2}-x_{m_1})}\right) - \frac{1}{12r}\frac{(x_{m_2}-x_{m_2-1})^2}{x_{m_2}-x_{m_1}} \geq Ch$$

and

(2.16)
$$\frac{x_{m_2+1}-x_{m_2}}{6r} + \frac{x_{m_2}-x_{m_2-1}}{2r}\left(1-\frac{1}{2}\frac{x_{m_2}-x_{m_2-1}}{(x_{m_2}-x_{m_1})}\right) - \frac{1}{12(r+1)}\frac{(x_{m_1+1}-x_{m_1})^2}{x_{m_2}-x_{m_1}} \geq Ch.$$

From the quasi-uniformity assumption, and from Assumption 5, this is equivalent to

$$
\begin{aligned}
&\tfrac{\delta^{--}}{3} + \delta^-(1-\tfrac{1}{2}\delta^-) - \tfrac{r+1}{6r}(\delta^+)^2 \geq C,\\
&\tfrac{\delta^{++}}{3} + \delta^+(1-\tfrac{1}{2}\delta^+) - \tfrac{r}{6(r+1)}(\delta^-)^2 \geq C.
\end{aligned}
\tag{2.17}
$$

But (2.17) is a consequence of

$$
\begin{aligned}
&\tfrac{3}{2}\delta^- + \delta^{--} - (\delta^+)^2 \geq C,\\
&\tfrac{3}{2}\delta^+ + \delta^{++} - \tfrac{1}{2}(\delta^-)^2 \geq C,
\end{aligned}
\tag{2.18}
$$

because $r \geq 1$. Suppose now that assumption (2.12) is satisfied. Then, if (2.15) and (2.16) are satisfied and the matrix $\frac{1}{2}(M+M^T)$ is positive definite with its eigenvalues not less than Ch, for any vector U, $(MU,U) \geq Ch(U,U)$, which yields (2.13). □

REMARK 4. *From the proof of Lemma* 2.2, (2.11) *is a linear system with a square matrix* M *of size* $\dim(W_k^l)$, *nonsymmetric, but such that* $(MU,U) \geq ch(U,U)$. *In other words, the matrix* M *resembles a mass matrix.*

From the previous lemma, we deduce the following.

COROLLARY 2.3. *Under assumption* (2.12), *the problem* (2.11) *has a unique solution. Furthermore, if we impose that* $u_k^l = 0$ *at the nodes in* $\Gamma_k^l \cap \mathcal{V}_k$, *then we have*

$$\left\|u_k^l\right\|_{L^2(\Gamma_k^l)} \leq C\|u\|_{L^2(\Gamma_k^l)}. \tag{2.19}$$

Likewise, let x_i be a given node in $\Gamma_k^l \cap \mathcal{V}_k$. Under assumptions (2.12)*, the solution of the problem, find $\widetilde{\psi}_i \in Z_k^l$ such that*

$$\begin{aligned} &\widetilde{\psi}_i(x_i) = 1, \\ &\widetilde{\psi}_i = 0 \text{ at the other nodes of } \Gamma_k^l \cap \mathcal{V}_k, \\ &\forall w_k^l \in W_k^l, \quad \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} \widetilde{\psi}_i(x) w_k^l(x) dx = 0, \end{aligned} \tag{2.20}$$

satisfies

$$\left\| \widetilde{\psi}_i \right\|_{L^2(\Gamma_k^l)} \leq C h^{\frac{1}{2}}. \tag{2.21}$$

REMARK 5. *Consider $u = (u_k)_k \in Y$. Then it is clear from* (2.5) *that all the nodal values of u_k located on $\Gamma_k^l \backslash \mathcal{V}_k$ can be found from the d.o.f. in the adjacent subdomains and from the d.o.f. on $\mathcal{V}_k$ by solving a system with a one-dimensional mass matrix. The nodal values of u_k located on $\partial\Omega_k \backslash \mathcal{V}_k$ can thus be seen as slave nodal values.*

2.3. The discrete problem. From now on, we shall assume that the conditions (2.12) are satisfied.

Let σ be the counting function $\sigma(x) = \sum_{k=1}^K 1_{\Omega_k}(x)$. The quantity $\sigma(x)$ represents the number of subdomains covering x. From Property 2, σ is bounded from above by a constant, and $\sigma \geq 1$. Consider the discrete problem: find $u \in Y$ such that for all $v \in Y$,

$$\sum_{k=1}^K \int_{\Omega_k} \frac{1}{\sigma} \nabla u_k \cdot \nabla v_k = \sum_{k=1}^K \int_{\Omega_k} \frac{1}{\sigma} f v_k. \tag{2.22}$$

Call a the symmetric bilinear form on Y:

$$a(u,v) = \sum_{k=1}^K \int_{\Omega_k} \frac{1}{\sigma} \nabla u_k \cdot \nabla v_k. \tag{2.23}$$

The following lemma allows us to state that problem (2.22) has a unique solution for Y defined by (2.10).

LEMMA 2.4. *The symmetric bilinear form a is positive definite.*

Proof. The bilinear form a is clearly positive semidefinite.

Let $u \in Y$ satisfy $a(u,u) = 0$. Then u_k is a constant ξ_k for $1 \leq k \leq K$. Let $\Xi = (\xi_1, \ldots, \xi_K)^T$. Taking $w = 1$ in (2.5) and summing up over all the sides of $\partial\Omega_k$, we obtain for each subdomain a linear equation for ξ. The system writes $S\Xi = 0$, where S is an irreducible $K \times K$ matrix such that

- for all $k \in \{1, \ldots, K\}$, $S_{kk} > 0$,
- for all $k \in \{1, \ldots, K\}$, for all $l \neq k$, $S_{kl} \leq 0$,
- for all $k \in \{1, \ldots, K\}$, $\sum_{l=1}^K S_{kl} = 0$.

From these properties, it is clear that there exists ξ such that all the ξ_k equal ξ. Now taking a subdomain Ω_k such that $meas(\partial\Omega_k \cap \partial\Omega) > 0$, we have that $u_k = 0$, which yields $\xi = 0$. ∎

Now, we wish to obtain an estimate on the ellipticity constant under typical but not necessarily optimal assumptions. For that, we recall that we have denoted by ω_k

the union of the subdomains intersecting Ω_k and by $\mathcal{I}_k$ the set of the integers i such that $\Omega_i \subset \omega_k$. Let us make the following assumption.

Assumption 6. Let Ω_k be a subdomain. We assume that for a positive constant C, for each $i \neq k \in \mathcal{I}_k$, there exists a side Γ_i^e and a subinterval γ_i of Γ_i^e such that

- $\gamma_i \subset \Omega_k$,
- $|\gamma_i| > CH$,
- γ_i is the union of elements of $\mathcal{T}_i^e$.

LEMMA 2.5. *Suppose that Assumptions* 1 *to* 6 *are satisfied. Then for* $\frac{h}{H}$ *small enough, there exists a constant* C *independent of* h, H, *and the overlap such that for any* u *in* Y,

$$\sum_{l\in\mathcal{I}_k}(\langle u_k\rangle - \langle u_l\rangle)^2 \leq C\sum_{l\in\mathcal{I}_k}\int_{\Omega_l}|\nabla u_l|^2, \tag{2.24}$$

where

$$\langle u_k\rangle = \frac{1}{|\Omega_k|}\int_{\Omega_k} u_k.$$

Proof. For $i \in \mathcal{I}_k$ and Γ_i^e as in Assumption 6, the subdomains intersecting Γ_i^e are (Ω_{j_m}), $j_m = 1, \ldots, n_i^e$. We wish to bound

$$\left|\int_{\gamma_i}\frac{1}{p_i^e(x)+1}\left(\langle u_i\rangle - \frac{1}{p_i^e(x)}\sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)\langle u_{j_m}\rangle\right)dx\right|$$

by using (2.5). The characteristic function 1_{γ_i} does not belong to W_i^e, so it is not a test function for (2.5). We thus take the test function in W_i^e supported in γ_i, equal to one at all the slave nodes of Ω_i whose related shape function is supported in γ_i. We call it w_i. We call $\overset{\circ}{\gamma}_i$ the region where $w_i = 1$. Clearly, from Property 3 and Assumption 1, the measure $meas(\gamma_i \setminus \overset{\circ}{\gamma}_i)$ is bounded by ch. From a scaled Poincaré–Wiertinger inequality, we have

$$\left|\int_{\gamma_i}\frac{w_i(x)}{p_i^e(x)+1}\left((\langle u_i\rangle - u_i(x)) - \frac{1}{p_i^e(x)}\sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)(\langle u_{j_m}\rangle - u_{j_m}(x))\right)dx\right|$$
$$\leq CH\left(\sum_{m\in\mathcal{I}_k}\int_{\Omega_m}|\nabla u_m|^2\right)^{\frac{1}{2}},$$

which yields, thanks to (2.5),

$$\left|\int_{\gamma_i}\frac{1}{p_i^e(x)+1}\left(\langle u_i\rangle - \frac{1}{p_i^e(x)}\sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)\langle u_{j_m}\rangle\right)w_i(x)dx\right|$$
$$\leq CH\left(\sum_{m\in\mathcal{I}_k}\int_{\Omega_m}|\nabla u_m|^2\right)^{\frac{1}{2}}.$$

Therefore,

$$\left| \int_{\gamma_i} \frac{1}{p_i^e(x)+1} \left(\langle u_i \rangle - \langle u_k \rangle - \frac{1}{p_i^e(x)} \sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)(\langle u_{j_m} \rangle - \langle u_k \rangle) \right) dx \right|$$
$$- \left| \int_{\gamma_i} \frac{1-w_i(x)}{p_i^e(x)+1} \left(\langle u_i \rangle - \langle u_k \rangle - \frac{1}{p_i^e(x)} \sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)(\langle u_{j_m} \rangle - \langle u_k \rangle) \right) \right|$$
$$\leq CH \left(\sum_{m \in \mathcal{I}_k} \int_{\Omega_m} |\nabla u_m|^2 \right)^{\frac{1}{2}}.$$

This yields the estimate

$$\left| \sum_{l \in \mathcal{I}_k} a_{li}(\langle u_l \rangle - \langle u_k \rangle) \right| - \frac{h_i}{H} \left| \sum_{l \in \mathcal{I}_k} b_{li}(\langle u_l \rangle - \langle u_k \rangle) \right| \leq C \left(\sum_{i \in \mathcal{I}_k} \int_{\Omega_i} |\nabla u_i|^2 \right)^{\frac{1}{2}}, \tag{2.25}$$

where it can be checked, thanks to Assumption 6, that

- $a_{ii} > C$, $|a_{ki}| > C$,
- $a_{li} \leq 0$ if $l \neq i$,
- $\sum_{l \in \mathcal{I}_k} a_{li} = 0$,
- $|b_{li}| < C$.

Thus, the matrix $(a_{ij})_{i \neq k, j \neq k}$ is a square M-matrix (in particular, $\sum_{k \neq l \in \mathcal{I}_k} a_{li} > C$). Therefore it is invertible and its inverse has a l^∞-norm bounded independently on h and H, and, thanks to Property 1, we have the desired result for h small enough. ∎

REMARK 6. *If Assumption* 6 *is not satisfied, then, thanks to the quasi-uniformity assumption on the grids inside the subdomains, we would obtain instead of* (2.24) *the estimate*

$$\sum_{l \in \mathcal{I}_k} (\langle u_k \rangle - \langle u_l \rangle)^2 \leq C \max_{l \in \mathcal{I}_k} \left(1 + \log \frac{H}{h_l} \right) \sum_{l \in \mathcal{I}_k} \int_{\Omega_l} |\nabla u_l|^2. \tag{2.26}$$

To prove this estimate, one has to take for each $i \neq k$ an index e such that $\Gamma_i^e \cap \Omega_k$ contains at least the support of a nodal basis function w_i of W_i^e. We have from the quasi-uniformity assumption the well-known estimate

$$\|u_m - \langle u_m \rangle\|_{L^\infty(\Omega_m)} \leq C \left(1 + \log \frac{H}{h_m} \right)^{\frac{1}{2}} \|\nabla u_m\|_{L^2(\Omega_m)}.$$

Thus

$$\left| \int_{\Gamma_i^e} \frac{w_i(x)}{p_i^e(x)+1} \left((\langle u_i \rangle - u_i(x)) - \frac{1}{p_i^e(x)} \sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x)(\langle u_{j_m} \rangle - u_{j_m}(x)) \right) dx \right|$$
$$\leq Ch_i \left(\sum_{m \in \mathcal{I}_k} \left(1 + \log \frac{H}{h_m} \right) \int_{\Omega_m} |\nabla u_m|^2 \right)^{\frac{1}{2}},$$

which yields, thanks to (2.5),

$$\left| \int_{\Gamma_i^e} \frac{1}{p_i^e(x)+1} \left(\langle u_i \rangle - \frac{1}{p_i^e(x)} \sum_{m=1}^{n_i^e} 1_{\Gamma_i^{e,m}}(x) \langle u_{j_m} \rangle \right) w_i(x) dx \right|$$
$$\leq C h_i \left(\sum_{m \in \mathcal{I}_k} \left(1 + \log \frac{H}{h_m} \right) \int_{\Omega_m} |\nabla u_m|^2 \right)^{\frac{1}{2}}.$$

This yields the estimate

$$\left| \sum_{l \in \mathcal{I}_k} a_{li} (\langle u_l \rangle - \langle u_k \rangle) \right| \leq C \left(\sum_{i \in \mathcal{I}_k} \left(1 + \log \frac{H}{h_m} \right) \int_{\Omega_i} |\nabla u_i|^2 \right)^{\frac{1}{2}},$$

where $a_{ii} > C$, $|a_{ki}| > C$, $a_{li} \leq 0$ *if* $l \neq i$, $\sum_{l \in \mathcal{I}_k} a_{li} = 0$, *and therefore* (2.26).

REMARK 7. *Lemma* 2.5 *(or Remark* 6*) will be used both for the ellipticity analysis and for the consistency error analysis.*

LEMMA 2.6. *Under Assumptions* 1 *to* 6 *and* (2.12), *there exists a constant* C *such that for all* $u \in Y$,

$$\sum_{k=0}^{K} H^2 \langle u_k \rangle^2 \leq C \sum_{k=0}^{K} \int_{\Omega_k} |\nabla u_k|^2, \tag{2.27}$$

and there exists a constant C *independent of* h *and* H *such that for all* $u \in Y$,

$$\sum_{k=0}^{K} ||u_k||^2_{H^1(\Omega_k)} \leq C \sum_{k=0}^{K} \int_{\Omega_k} |\nabla u_k|^2. \tag{2.28}$$

Proof. Choose the coordinates so that Ω is contained in the square $(0, diam(\Omega))^2$. From the quasi-uniformity assumption on the subdomains, there exists a constant δ independent of h and H, $0 < \delta \leq 1$, such that the straight lines Δ_j of equations $x_2 = j\delta H$, $1 \leq j \leq J = \frac{diam(\Omega)}{\delta H}$, cross each subdomain Ω_k at least once. For a subdomain Ω_k, we denote by $\zeta_{k,j}$

$$\zeta_{k,j} = \sup_{x \in \Omega_k \cap \Delta_j} x_1 \text{ if } meas(\Omega_k \cap \Delta_j) > 0, \quad \zeta_{k,j} = -\infty \text{ if } meas(\Omega_k \cap \Delta_j) = 0.$$

Let $(k_i^j)_{1 \leq i \leq I_j}$ be the indices such that the line Δ_j crosses $\Omega_{k_i^j}$, numbered in such a way that $\zeta_{k_i^j,j}$ is an increasing sequence. Necessarily, we have $meas(\Omega_{k_i^j} \cap \Omega_{k_{i+1}^j}) > 0$. From Assumption 2, we know that there exists a constant C such that $I_j \leq \frac{C}{H}$. We

have

$$\sum_{i=1}^{I_j}\langle u_{k_i^j}\rangle^2 \le \sum_{i=1}^{I_j}\left(\langle u_{k_1^j}\rangle + \sum_{1\le l\le i-1}\langle u_{k_{l+1}^j}\rangle - \langle u_{k_l^j}\rangle\right)^2$$
$$\le \sum_{i=1}^{I_j}\left(2\langle u_{k_1^j}\rangle^2 + C(i-1)\sum_{1\le l\le i-1}\sum_{m\in\mathcal{I}_{k_l^j}}\int_{\Omega_m}|\nabla u_m|^2\right)$$
$$\le \frac{C}{H}\int_{\Omega_{k_1^j}}|\nabla u_{k_1^j}|^2 + \frac{C}{H^2}\sum_{i=1}^{I_j-1}\sum_{m\in\mathcal{I}_{k_i^j}}\int_{\Omega_m}|\nabla u_m|^2$$
$$\le \frac{C}{H^2}\sum_{i=0}^{I_j-1}\sum_{m\in\mathcal{I}_{k_i^j}}\int_{\Omega_m}|\nabla u_m|^2.$$

Finally, we obtain

$$\sum_{k=1}^{K}\langle u_k\rangle^2 \le \frac{C}{H^2}\sum_{j=1}^{J}\sum_{i=1}^{I_j-1}\sum_{m\in\mathcal{I}_{k_i^j}}\int_{\Omega_m}|\nabla u_m|^2$$
$$\le \frac{C}{H^2}\sum_{k=1}^{K}\int_{\Omega_k}|\nabla u_k|^2,$$

since each subdomain appears a finite number of times in the triple sum above.

Then (2.28) is a direct application of the Poincaré–Wiertinger inequality. ∎

REMARK 8. *In fact, for nonconforming methods, the ellipticity analysis can be postponed after the consistency analysis and simplified by using a clever duality argument* [9]. *However, this would not lead to an optimal result as in Lemma* 2.6 *because the consistency estimate would not be completely optimal. (There are, in any case, logarithmic factors in* h/H*.) Furthermore, we have seen in Remark* 7 *that the consistency error analysis uses* (2.26) *and Lemma* 2.5.

Finally, we obtain the following.

COROLLARY 2.7. *Under Assumptions* 1 *to* 6 *and* (2.12)*, there exists a constant* C_e *independent on the mesh parameters such that*

$$\forall u\in Y,\quad a(u,u)\ge C_e\sum_{k=1}^{K}\int_{\Omega_k}\left(|\nabla u_k|^2+u_k^2\right). \tag{2.29}$$

If only Assumptions 1 *to* 5 *and* (2.12) *are satisfied, we have* (2.29)*, but we know only that there exists a constant* C *independent on the mesh parameters such that*

$$C_e\le C\frac{1}{\max_l(1+\log\frac{H}{h_l})}. \tag{2.30}$$

3. Error analysis. Let us introduce the *broken* seminorm and norm on $\prod_{k=1}^{K}H^1(\Omega_k)$:

$$\forall u\in\prod_{k=1}^{K}H^1(\Omega_k),\quad |u|_*^2\equiv\sum_{k=1}^{K}\int_{\Omega_k}|\nabla u_k|^2,\quad \|u\|_*^2\equiv\sum_{k=1}^{K}\int_{\Omega_k}\left(|\nabla u_k|^2+u_k^2\right). \tag{3.1}$$

Let u^* be the solution of the continuous problem (2.1). With an abuse of notations, we still call u^* the vector of $\prod_{k=1}^K H^1(\Omega_k)$, defined by $u_k^* = u^*|_{\Omega_k}$. By the Berger–Scott–Strang lemma (see [3, 19]), we know that the error of the method is the sum of a consistency error and of a best approximation error:

$$\|u - u^*\|_* \leq \frac{1}{C_e}\left(\inf_{v\in Y}|u^* - v|_* + \sup_{0\neq v\in Y}\frac{|a(u^*,v) - \sum_{k=1}^K \int_{\Omega_k}\frac{1}{\sigma}fv_k|}{|v|_*}\right), \tag{3.2}$$

where C_e is the ellipticity constant. We have seen above that under Assumptions 1 to 6, C_e is bounded independently of the mesh parameters and the subdomains diameters.

3.1. Consistency error. For the consistency error, we first need to study a weighted L^2 projection operator: call $\tilde{\pi}_k^l$ the projection operator on W_k^l defined by

$$\forall w_k^l \in W_k^l, \quad \int_{\Gamma_k^l}\frac{1}{p_k^l(x)+1}(\tilde{\pi}_k^l u)(x)w_k^l(x)dx = \int_{\Gamma_k^l}\frac{1}{p_k^l(x)+1}u(x)w_k^l(x)dx. \tag{3.3}$$

It can be checked that the functional

$$|||u||| \equiv \sup_{0\neq v\in H^{\frac{1}{2}}(\Gamma_k^l)}\frac{\left|\int_{\Gamma_k^l}\frac{1}{p_k^l(x)+1}u(x)v(x)dx\right|}{\|v\|_{H^{\frac{1}{2}}(\Gamma_k^l)}}$$

defines a norm on $C^\infty(\Gamma_k^l)$. This is due in particular to the fact that $\frac{1}{p_k^l(x)+1}$ is a positive weight function in L^∞ bounded away from 0. We call $(H^{\frac{1}{2}})'(\Gamma_k^l)$ the completion of $C^\infty(\Gamma_k^l)$ for this norm. For $u \in (H^{\frac{1}{2}})'(\Gamma_k^l)$, we define $\|u\|_{(H^{\frac{1}{2}})'(\Gamma_k^l)} = |||u|||$.

The properties of the operator $\tilde{\pi}_k^l$ are given by the following lemma.

LEMMA 3.1. *For any real number r, $0 \leq r \leq 1$, we have the following estimate for any function $u \in H^r(\Gamma_k^l)$:*

$$\|u - \tilde{\pi}_k^l u\|_{L^2(\Gamma_k^l)} + h_k^{-\frac{1}{2}}\|u - \tilde{\pi}_k^l u\|_{(H^{\frac{1}{2}})'(\Gamma_k^l)} \leq Ch_k^r|u|_{H^r(\Gamma_k^l)}. \tag{3.4}$$

Proof. The proof is analogous to the one in [4]. ∎

LEMMA 3.2. *Assume that the solution u^* of (2.1) is such that $u^*|_{\Omega_k}$ belongs to $H^{\sigma_k}(\Omega_k)$ with $\sigma_k > \frac{3}{2}$. Then the consistency error is bounded by*

$$C\left(1 + \max_k \log\frac{H}{h_k}\right)\left(\sum_{k=1}^K \max_{i\in\mathcal{I}_k}\left(1 + \sqrt{\frac{h_i}{h_k}}\right)^2 h_k^{2(\sigma_k-1)}|u^*|^2_{H^{\sigma_k}(\Omega_k)}\right)^{\frac{1}{2}}.$$

Proof. We apply Green's formula in each subdomain and obtain that

$$\begin{aligned} &a(u^*,v) - \sum_{k=1}^K \int_{\Omega_k}\frac{1}{\sigma}fv_k \\ &= \sum_{k=1}^K \int_{\Omega_k}\frac{1}{\sigma}(-\Delta u^* - f)v_k + \sum_{k=1}^K \int_{\partial\Omega_k}\frac{1}{\sigma^-}\frac{\partial u^*}{\partial n_k}v_k - \sum_{k=1}^K\sum_{l=1}^{E_k}\sum_{i=1}^{n_k^l}\int_{\Gamma_k^{l,i}}\left[\frac{1}{\sigma}\right]\frac{\partial u^*}{\partial n_k}v_{j_i}, \end{aligned} \tag{3.5}$$

where n_k is the unit vector normal to $\partial\Omega_k$ outgoing from Ω_k and σ^- is the trace of $\sigma|_{\Omega_k}$ on $\partial\Omega_k$. We denote by σ^+ the trace of $\sigma|_{\Omega\setminus\Omega_k}$ on $\partial\Omega_k$. Note that $\sigma^+ = \sigma^- - 1$. The function $[\frac{1}{\sigma}]$ is the jump $\frac{1}{\sigma^+} - \frac{1}{\sigma^-} = \frac{1}{\sigma^-(\sigma^- - 1)}$.

Thus

$$a(u^*, v) - \sum_{k=1}^{K} \int_{\Omega_k} \frac{1}{\sigma} f v_k = \sum_{k=1}^{K} \sum_{l=1}^{E_k} \int_{\Gamma_k^l} \frac{1}{\sigma^-} \frac{\partial u^*}{\partial n_k} \left(v_k - \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{\sigma^- - 1} v_{j_i} \right), \tag{3.6}$$

but on Γ_k^l we have that $p_k^l = \sigma^- - 1$, so by using (2.5),

$$\begin{aligned} a(u^*, v) - \sum_{k=1}^{K} \int_{\Omega_k} \frac{1}{\sigma} f v_k &= \sum_{k=1}^{K} \sum_{l=1}^{E_k} \int_{\Gamma_k^l} \frac{1}{p_k^l + 1} \frac{\partial u^*}{\partial n_k} \left(v_k - \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{p_k^l} v_{j_i} \right) \\ &= \sum_{k=1}^{K} \sum_{l=1}^{E_k} \int_{\Gamma_k^l} \frac{1}{p_k^l + 1} \left(\frac{\partial u^*}{\partial n_k} - w_k^l \right) \left(v_k - \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{p_k^l} v_{j_i} \right) \quad \forall w_k^l \in W_k^l. \end{aligned} \tag{3.7}$$

Therefore

$$\begin{aligned} &a(u^*, v) - \sum_{k=1}^{K} \int_{\Omega_k} \frac{1}{\sigma} f v_k \\ &= \sum_{k=1}^{K} \sum_{l=1}^{E_k} \int_{\Gamma_k^l} \frac{1}{p_k^l + 1} \left(\frac{\partial u^*}{\partial n_k} - w_k^l \right) \left(v_k - \langle v_k \rangle - \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{p_k^l} (v_{j_i} - \langle v_k \rangle) \right) \\ &= \sum_{k=1}^{K} \sum_{l=1}^{E_k} \int_{\Gamma_k^l} \frac{1}{p_k^l + 1} \left(\frac{\partial u^*}{\partial n_k} - w_k^l \right) \left(\begin{array}{l} v_k - \langle v_k \rangle - \displaystyle\sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{p_k^l} (v_{j_i} - \langle v_{j_i} \rangle) \\ + \displaystyle\sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}} \frac{1}{p_k^l} (\langle v_k \rangle - \langle v_{j_i} \rangle) \end{array} \right). \end{aligned} \tag{3.8}$$

Let us introduce some useful tools before choosing w_k^l. Call $\widetilde{v_{j_i}}$ and $\overline{v_{j_i}}$ the piecewise linear and continuous functions on Ω_{j_i}, respectively, obtained from $v_{j_i} - \langle v_{j_i} \rangle$ and $\langle v_k \rangle - \langle v_{j_i} \rangle$ in the following way: $\widetilde{v_{j_i}} = v_{j_i} - \langle v_{j_i} \rangle$, $\overline{v_{j_i}} = \langle v_k \rangle - \langle v_{j_i} \rangle$ at all the nodes of $\mathcal{T}_{j_i}$, except those on $\partial\Omega_{j_i}$ or in the balls of radius h centered at the crosspoints, where $\widetilde{v_{j_i}} = \overline{v_{j_i}} = 0$; see Figure 3.1. Then it is possible to extend $\widetilde{v_{j_i}}|_{\Gamma_k^{l,i}}$ and $\overline{v_{j_i}}$ by 0 on Γ_k^l. We still call these extensions $\widetilde{v_{j_i}}$ and $\overline{v_{j_i}}$: these functions belong to $H^{\frac{1}{2}}(\Gamma_k^l)$. We have

$$|\widetilde{v_{j_i}}|^2_{H^{\frac{1}{2}}(\Gamma_k^l)} \leq C \left(1 + \log \frac{H}{h_{j_i}} \right) |v_{j_i}|^2_{H^1(\Omega_{j_i})},$$

$$\|\widetilde{v_{j_i}}\|^2_{L^\infty(\Gamma_k^l)} \leq \|v_{j_i} - \langle v_{j_i} \rangle\|^2_{L^\infty(\Omega_{j_i})} \leq C \left(1 + \log \frac{H}{h_{j_i}} \right) |v_{j_i}|^2_{H^1(\Omega_{j_i})},$$

and

$$\|\widetilde{v_{j_i}} - v_{j_i} + \langle v_{j_i} \rangle\|^2_{L^2(\Gamma_k^{l,i})} \leq C h_{j_i} \left(1 + \log \frac{H}{h_{j_i}} \right) |v_{j_i}|^2_{H^1(\Omega_{j_i})}.$$

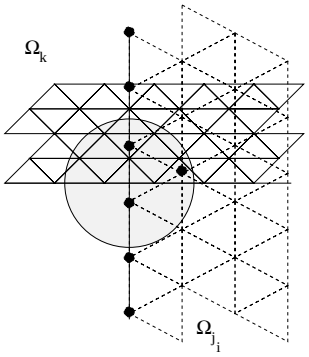


FIG. 3.1. *In black are the nodes where the functions* $\widetilde{v_{j_i}}$ *and* $\overline{v_{j_i}}$ *are set to* 0.

We also have that

$$|\overline{v_{j_i}}|^2_{H^{\frac{1}{2}}(\Gamma_k^l)} \leq C|\langle v_{j_i}\rangle - \langle v_k\rangle|^2$$

and

$$\|\overline{v_{j_i}} - (\langle v_{j_i}\rangle - \langle v_k\rangle)\|^2_{L^2(\Gamma_k^{l,i})} \leq Ch_{j_i}|\langle v_{j_i}\rangle - \langle v_k\rangle|^2.$$

Let us go back to the analysis of (3.8); we choose $w_k^l = \tilde{\pi}_k^l(\frac{\partial u^*}{\partial n_k})$. We decompose the integral

$$I_k^l = \int_{\Gamma_k^l} \frac{1}{p_k^l + 1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}(v_{j_i} - \langle v_{j_i}\rangle)\right)$$

into the sum

$$\int_{\Gamma_k^l} \frac{1}{p_k^l + 1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}\widetilde{v_{j_i}}\right)$$
$$+\int_{\Gamma_k^l} \frac{1}{p_k^l + 1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}(v_{j_i} - \langle v_{j_i}\rangle - \widetilde{v_{j_i}})\right).$$

The function $1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}\widetilde{v_{j_i}} = \frac{1}{p_k^l}\widetilde{v_{j_i}}$ belongs to $H^{\frac{1}{2}}(\Gamma_k^l)$ because $\widetilde{v_{j_i}}$ vanishes near the points where p_k^l jumps, and its $H^{\frac{1}{2}}$-norm is bounded by

$$C\left(\left(1 + \log\frac{H}{h_{j_i}}\right)^{\frac{1}{2}}\|\widetilde{v_{j_i}}\|_{L^\infty(\Gamma_k^l)} + \|\widetilde{v_{j_i}}\|_{H^{\frac{1}{2}}(\Gamma_k^l)}\right).$$

From (3.4), we have

$$|I_k^l| \leq Ch_k^{\sigma_k - \frac{3}{2}}|u^*|_{H^{\sigma_k}(\Omega_k)}\left(\begin{array}{l} h_k^{\frac{1}{2}}\left(\left(1 + \log\frac{H}{h_{j_i}}\right)^{\frac{1}{2}}\|\widetilde{v_{j_i}}\|_{L^\infty(\Gamma_k^l)} + |\widetilde{v_{j_i}}|_{H^{\frac{1}{2}}(\Gamma_k^l)}\right) \\ + \|\widetilde{v_{j_i}} - v_{j_i}\|_{L^2(\Gamma_k^{l,i})}\end{array}\right)$$

$$\leq Ch_k^{\sigma_k - 1}|u^*|_{H^{\sigma_k}(\Omega_k)}\left(1 + \log\frac{H}{h_{j_i}}\right)\left(1 + \sqrt{\frac{h_{j_i}}{h_k}}\right)|v_{j_i}|_{H^1(\Omega_{j_i})}.$$

We decompose the integral

$$J_k^l = \int_{\Gamma_k^l} \frac{1}{p_k^l+1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}(\langle v_k\rangle - \langle v_{j_i}\rangle)\right)$$

into the sum

$$\int_{\Gamma_k^l} \frac{1}{p_k^l+1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}\overline{v_{j_i}}\right)$$
$$+\int_{\Gamma_k^l} \frac{1}{p_k^l+1}\left(\frac{\partial u^*}{\partial n_k} - \tilde{\pi}_k^l\left(\frac{\partial u^*}{\partial n_k}\right)\right)\left(1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}(\langle v_k\rangle - \langle v_{j_i}\rangle - \overline{v_{j_i}})\right).$$

The function $1_{\Gamma_k^{l,i}}\frac{1}{p_k^l}\overline{v_{j_i}}$ belongs to $H^{\frac{1}{2}}(\Gamma_k^l)$, and its $H^{\frac{1}{2}}$-norm is bounded by

$$C\left(1+\log\frac{H}{h_{j_i}}\right)^{\frac{1}{2}}|\langle v_k\rangle - \langle v_{j_i}\rangle|.$$

Exactly as above, we obtain that

$$|J_k^l| \le Ch_k^{\sigma_k-1}|u^*|_{H^{\sigma_k}(\Omega_k)}\left(1+\log\frac{H}{h_{j_i}}\right)^{\frac{1}{2}}\left(1+\sqrt{\frac{h_{j_i}}{h_k}}\right)|\langle v_{j_i}\rangle - \langle v_k\rangle|.$$

We have seen in Remark 6 that, at worst,

$$|\langle v_{j_i}\rangle - \langle v_k\rangle|^2 \le C\left(1+\max_{j\in\mathcal{I}_k}\log\frac{H}{h_j}\right)\sum_{j\in\mathcal{I}_k}\int_{\Omega_j}|\nabla v_j|^2.$$

We conclude by summing up all the contributions. We get

$$\sup_{0\neq v\in Y}\frac{|a(u^*,v) - \sum_{k=1}^K\int_{\Omega_k}\frac{1}{\sigma}fv_k|}{|v|_*}$$
$$\le C\left(1+\max_k\log\frac{H}{h_k}\right)\left(\sum_{k=1}^K\max_{i\in\mathcal{I}_k}\left(1+\sqrt{\frac{h_i}{h_k}}\right)^2 h_k^{2(\sigma_k-1)}|u^*|^2_{H^{\sigma_k}(\Omega_k)}\right)^{\frac{1}{2}}. \qquad \square$$

REMARK 9. *It may be possible to prove a better result than Lemma* 3.2. *This question will be addressed in a forthcoming work.*

3.2. Best approximation error. Call i_k the interpolation operator onto X_k; we have for $u_k^* \in H^{1+\sigma_k}(\Omega_k)$, with $\sigma_k > 0$,

$$\sum_{k=1}^K h_k^2|u_k^* - i_ku_k^*|^2_{H^1(\Omega_k)} + \|u_k^* - i_ku_k^*\|^2_{L^2(\Omega_k)} \le C\sum_{k=1}^K h_k^{2+2\sigma_k}|u_k^*|^2_{H^{1+\sigma_k}(\Omega_k)}.$$

However, the vector $(i_ku_k^*)_{k=1,\dots,K}$ does not belong to Y, so we have to modify it in such a way that the matching conditions are satisfied at the subdomains' boundaries.

The function $i_ku_k^*$ will first be corrected on $\partial\Omega_k$; then the correction will be extended in the whole domain Ω_k by means of the trivial lifting operator L_k, which consists of setting the values at the nodes contained in Ω_k to 0. The properties of this standard lifting operator are given by the following lemma.

LEMMA 3.3. *There exists a constant $C > 0$ such that for any $u \in H^{\frac{1}{2}}(\partial\Omega_k)$, such that for all $l \in \{1, \ldots, E_k\}$, $u|_{\Gamma_k^l} \in X_k^l$, we have*

$$|L_k u|_{H^1(\Omega_k)} \leq C h_k^{-\frac{1}{2}} \|u\|_{L^2(\partial\Omega_k)} \tag{3.9}$$

and

$$\|L_k u\|_{L^2(\Omega_k)} \leq C h_k^{\frac{1}{2}} \|u\|_{L^2(\partial\Omega_k)}. \tag{3.10}$$

Proof. It is classical that

$$\|L_k u\|_{L^2(\Omega_k)} \leq C h_k^{\frac{1}{2}} \|u\|_{L^2(\partial\Omega_k)}.$$

Then (3.9) is obtained by an inverse inequality. ∎

Consider the side Γ_k^l, and let π^* be the operator from $\prod_{i=1}^{n_k^l} L^2(\Gamma_k^{l,i})$ into X_k^l defined by

$$\begin{aligned} \forall w_k^l \in W_k^l, \qquad & \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} (\pi^*(u_{j_i})_{i=1,\ldots,n_k^l})(x) w_k^l(x) dx \\ & = \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} \sum_{i=1}^{n_k^l} \frac{1}{p_k^l(x)} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) w_k^l(x) dx; \end{aligned} \tag{3.11}$$

then we have the following lemma.

LEMMA 3.4. *There exists a constant $C > 0$ such that for any $(u_{j_i})_{i=1,\ldots,n_k^l} \in \prod_{i=1}^{n_k^l} L^2(\Gamma_k^{l,i})$,*

$$\|\pi^*(u_{j_i})_{i=1,\ldots,n_k^l}\|_{L^2(\Gamma_k^l)} \leq C \sum_{i=1}^{n_k^l} \|u_{j_i}\|_{L^2(\Gamma_k^{l,i})}. \tag{3.12}$$

Proof. Consider first the operator $\tilde{\pi}$ from $\prod_{i=1}^{n_k^l} L^2(\Gamma_k^{l,i})$ into W_k^l defined by

$$\begin{aligned} \forall w_k^l \in W_k^l, \qquad & \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} (\tilde{\pi}(u_{j_i})_{i=1,\ldots,n_k^l})(x) w_k^l(x) dx \\ & = \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} \sum_{i=1}^{n_k^l} \frac{1}{p_k^l(x)} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) w_k^l(x) dx. \end{aligned} \tag{3.13}$$

Then it is clear that

$$\|\tilde{\pi}(u_{j_i})_{i=1,\ldots,n_k^l}\|_{L^2(\Gamma_k^l)} \leq C \sum_{i=1}^{n_k^l} \|u_{j_i}\|_{L^2(\Gamma_k^{l,i})}. \tag{3.14}$$

Now $\pi^*(u_i)_{i=1,\ldots,n_k^l}$ is obtained by correcting $\tilde{\pi}(u_i)_{i=1,\ldots,n_k^l}$ as in [4, Lemma 4.3] by using the functions defined in (2.20), and we obtain the desired estimate (3.12) from (2.21). ∎

We are now ready to establish a best approximation estimate.

We correct $i_k(u_k^*)$ in subdomain Ω_k in order to satisfy the matching condition on $\partial\Omega_k$. To do so, we add to $i_k(u_k^*)$ the function

$$e_k \equiv \sum_{l=1}^{E_k} L_k \left(\pi^*(i_{j_i}(u_{j_i}^*) - i_k(u_k^*))_{i=1,\ldots,n_k^l} \right).$$

The vector $(i_k(u_k^*)+e_k)_{1\le k\le K}$ belongs to Y. Indeed, $L_k(\pi^*(i_{j_i}(u_{j_i}^*)-i_k(u_k^*))_{i=1,\dots,n_k^l})$ vanishes in the triangles strictly embedded in Ω_k, so it does not play any role in the matching conditions on other subdomains' boundaries. The matching conditions have been designed for that.

We wish to estimate $|u_k^* - i_k(u_k^*) - e_k|_{H^1(\Omega_k)} \le |u_k^* - i_k(u_k^*)|_{H^1(\Omega_k)} + |e_k|_{H^1(\Omega_k)}$. From Lemma 3.3, we need to give a bound on $\|\pi^*(i_{j_i}(u_{j_i}^*) - i_k(u_k^*))_{i=1,\dots,n_k^l}\|_{L^2(\Gamma_k^l)}$. The bound is obtained from the stability of the operator π^* (see (3.12)) and from the estimate

$$\|i_{j_i}(u_{j_i}^*) - i_k(u_k^*)\|_{L^2(\Gamma_k^{l,i})} \le C(h^{\sigma_k-\frac{1}{2}}|u^*|_{H^{\sigma_k}(\Omega_k)} + h^{\sigma_{j_i}-\frac{1}{2}}|u^*|_{H^{\sigma_{j_i}}(\Omega_{j_i})}). \tag{3.15}$$

In the same manner, we can bound $\|u_k^* - i_k(u_k^*) - e_k\|_{L^2(\Omega_k)}$ so we obtain the best fit estimate.

LEMMA 3.5. *Let $u^* \in H^1(\Omega)$ be such that for $1 \le k \le K$, $u^*|_{\Omega_k} \in H^{\sigma_k}(\Omega_k)$ with $2 \ge \sigma_k > 1$. Then there exists $u \in Y$ such that*

$$\sum_{k=1}^K \frac{1}{h_k}\|u_k^* - u_k\|_{L^2(\Omega_k)} + |u_k^* - u_k|_{H^1(\Omega_k)} \le C\sum_{k=1}^K h_k^{\sigma_k-1}|u_k^*|_{H^{\sigma_k}(\Omega_k)}. \tag{3.16}$$

Then the error estimate is given by the following theorem.

THEOREM 3.6. *Assume that the solution u^* of (2.1) is such that for $1 \le k \le K$, $u^*|_{\Omega_k} \in H^{\sigma_k}(\Omega_k)$ with $2 \ge \sigma_k > \frac{3}{2}$. Then there exists a constant C such that, if $u \in Y$ is the solution of (2.22),*

$$\begin{aligned}&\sum_{k=1}^K \|u_k^* - u_k\|_{H^1(\Omega_k)}\\ &\le \frac{C}{C_e}\left(1+\max_k \log\frac{H}{h_k}\right)\left(\sum_{k=1}^K \max_{i\in\mathcal{I}_k}\left(1+\sqrt{\frac{h_i}{h_k}}\right)^2 h_k^{2(\sigma_k-1)}|u^*|^2_{H^{\sigma_k}(\Omega_k)}\right)^{\frac{1}{2}},\end{aligned} \tag{3.17}$$

where C_e is the ellipticity constant.

4. A strengthened matching condition. We give below an example of stronger matching conditions in the neighborhood of crosspoints.

4.1. An example of a strengthened matching condition. With the notations introduced in section 2.2, it is possible to strengthen the previous matching condition by supplementing the previous test function space $W_k^{l,0} \equiv W_k^l$ with Q supplementary spaces $(W_k^{l,q})_{1\le q\le Q}$ (to be defined below) such that $\dim(W_k^l) + \sum_{q=1}^Q \dim(W_k^{l,q}) \le \dim(\widetilde{W_k^l})$. Typically, each new space will correspond to a crosspoint on Γ_k^l. We define the direct sum as $\overline{W_k^l} = \bigoplus_{q=0}^Q W_k^{l,q}$, and we introduce a family of coefficients $\lambda_{0i} = 1$ for $1 \le i \le n_k^l$ and $\lambda_{qi} \in \{0,1\}$ for $1 \le q \le Q$ and $1 \le i \le n_k^l$ (these coefficients will be defined below), and we call p_q the function defined on Γ_k^l by

$$p_q(x) = \sum_{i=1}^{n_k^l} \lambda_{qi} 1_{\Gamma_k^{l,i}}(x). \tag{4.1}$$

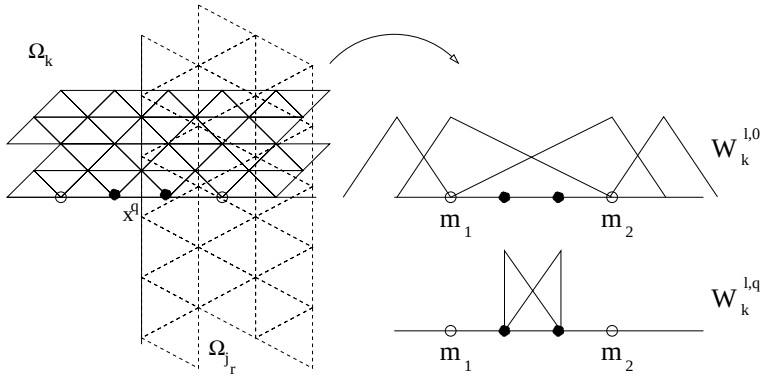


FIG. 4.1. *The spaces $W_k^{l,0}$ and $W_k^{l,q}$. (Only two subdomains have been represented.) In the case presented here, the dimension of $W_k^{l,q}$ is two.*

Then the strengthened matching condition reads as

$$\forall w \in W_k^{l,0}, \quad \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} \left(u_k(x) - \frac{1}{p_0(x)} \sum_{i=1}^{n_k^l} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) \right) w(x)dx = 0, \tag{4.2}$$

$$\forall q \in \{1,\dots,Q\},\ \forall w \in W_k^{l,q},$$
$$\int_{\Gamma_k^l} \left(u_k(x) - \frac{1}{p_q(x)} \sum_{i=1}^{n_k^l} \lambda_{qi} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) \right) w(x)dx = 0. \tag{4.3}$$

REMARK 10. *Conditions* (4.2), (4.3) *are stronger than* (2.5), *since* $W_k^{l,0} = W_k^l$.

We have to specify the spaces $W_k^{l,q}$ for $q \geq 1$. Call $(x^q)_{1\leq q\leq Q}$ the crosspoints $x^q \in \mathcal{X}_k^l$. For a crosspoint x^q (assume that $\{x^q\} = \Gamma_k^l \cap \Gamma_{j_r}^{l'}$) (see Figure 4.1), we call $\{x_{m_1+1},\dots,x_{m_2-1}\}$ the nodes of $\mathcal{T}_k^l$ for which the support of the corresponding basis function of X_k intersects the side $\Gamma_{j_r}^{l'}$. We call $\widetilde{\phi}_{m_1+1}$ the piecewise linear and continuous (except at x_{m_1+1}) function, vanishing outside $[x_{m_1+1}, x_{m_1+2})$, linear on $[x_{m_1+1}, x_{m_1+2})$, and equal to 1 at x_{m_1+1} and 0 at x_{m_1+2}. Likewise, we call $\widetilde{\phi}_{m_2-1}$ the piecewise linear and continuous (except at x_{m_2-1}) function, vanishing outside $(x_{m_2-2}, x_{m_2-1}]$, linear on $(x_{m_2-2}, x_{m_2-1}]$, and equal to 1 at x_{m_2-1} and 0 at x_{m_2-2}.

We define $W_k^{l,q} \equiv span(\widetilde{\phi}_{m_1+1}, \phi_{m_1+2}, \dots, \phi_{m_2-2}, \widetilde{\phi}_{m_2-1})$. The spaces $W_k^{l,0}$ and $W_k^{l,q}$ are displayed on Figure 4.1. Note that with this choice of $W_k^{l,q}$, the supports of the functions in $W_k^{l,q}$ do not intersect the supports of the functions of X_k^{l,x_q}.

Obviously we have

$$\dim(\widetilde{W_k^l}) = \dim\left(\bigoplus_{q=0}^{Q} W_k^{l,q} \right).$$

Now we need to define the coefficients λ_{qi}. We set $\lambda_{0i} = 1$ for all $1 \leq i \leq n_k^l$. For $k \geq 1$, assume that $\{x_q\} = \Gamma_k^l \cap \Gamma_{j_r}^{l'}$. Then we set $\lambda_{qr} = 0$ and $\lambda_{qi} = 1$ for all $1 \leq i \leq n_k^l$, $i \neq r$.

Then Y is the subspace of X defined by

$$Y \equiv \{u \in X;\ \forall k \in \{1,\dots,K\},\ \forall l \in \{1,\dots,E_k\},\ u \text{ satisfies (4.2), (4.3)}\} \tag{4.4}$$

for $W_k^{l,q}$ and λ_{qi} defined as above.

REMARK 11. *Let $u = (u_k) \in Y$. Then it is very clear from* (4.2), (4.3) *that all the nodal values of u_k located on $\partial\Omega_k$, except at the vertices of $\partial\Omega_k$, can be found from the d.o.f. in the adjacent subdomains and from the d.o.f. located at the vertices of $\partial\Omega_k$. With this matching condition, all the nodal values located on $\partial\Omega_k$, except at the vertices of $\partial\Omega_k$, are slave nodal values.*

REMARK 12. *Finding the slave nodal values can be achieved in two steps:*

1. *Find the unknown located at the black nodes on Figure* 4.1 *by taking the test functions in the spaces $W_k^{l,q}, q > 0$. This corresponds to solving a small linear system with a mass matrix for each crosspoint on Γ_k^l.*
2. *Find the remaining nodal values (located on $\Gamma_k^l \backslash \mathcal{V}_k$) by solving a problem of the type* (2.11). *We have seen above that this problem has a unique solution under conditions* (2.12).

4.2. Error estimate. Since the space $\overline{W_k^l}$ contains the space W_k^l, and (2.5) is equivalent to (4.2), the consistency error is bounded from above by the consistency error studied in section 3.1. Therefore we have the following result.

LEMMA 4.1. *Assume that $u^*|_{\Omega_k}$ belongs to $H^{\sigma_k}(\Omega_k)$, with $\sigma_k > \frac{3}{2}$. Then the consistency error is bounded by*

$$C\left(1 + \max_k \log \frac{H}{h_k}\right)\left(\sum_{k=1}^{K} \max_{i \in \mathcal{I}_k}\left(1 + \sqrt{\frac{h_i}{h_k}}\right)^2 h_k^{2(\sigma_k - 1)} |u|^2_{H^{\sigma_k}(\Omega_k)}\right)^{\frac{1}{2}}.$$

For the best fit error, we proceed as above. We correct $i_k(u_k^*)$ in subdomain Ω_k in order to satisfy the matching conditions. The correction is done first on $\partial\Omega_k$; then the same lifting operator L_k as above is used to correct u_k in Ω_k. To correct the trace of u_k on $\partial\Omega_k$, we proceed in two steps.

First step. Let us focus on the side Γ_k^l and keep the notations used for the definition of the space $W_k^{l,q}$ in section 4.1. We call $Z_k^{l,q}$ the space spanned by $(\widetilde{\psi}_{m_1+1}, \dots, \widetilde{\psi}_{m_2-1})$, where $\widetilde{\psi}_i$ have been defined by (2.20).

REMARK 13. *From the geometric assumptions, the dimensions of $W_k^{l,q}$ and $Z_k^{l,q}$ are equal ($q \geq 1$) and bounded by a constant independent of h.*

For $q \geq 1$, call $\mathcal{I}_k^{l,q}$ the set of indices i, such that $\lambda_{qi} = 1$ (see section 4.1), and $\gamma_k^{l,q}$ the support of the functions in $W_k^{l,q}$. ($\gamma_k^{l,q}$ is the interval located between the two black nodes in Figure 4.1.) Consider the projector $\pi_k^{l,q}$ from $\prod_{i \in \mathcal{I}_k^{l,q}} L^2(\Gamma_k^{l,i})$ into $Z_k^{l,q}$ defined by

$$\forall w \in W_k^{l,q}, \quad \int_{\Gamma_k^l} \pi_k^{l,q}(u_i)_{i=1,\dots,n_k^l}(x) w(x) dx = \int_{\Gamma_k^l} \frac{1}{p_q(x)} \sum_{i=1}^{n_k^l} \lambda_{qi} 1_{\Gamma_k^{l,i}}(x) u_{j_i}(x) w(x) dx. \tag{4.5}$$

This projector has exactly the same matrix as the $L^2(\gamma_k^{l,q})$ projector on $W_k^{l,q}$. So we have the estimate

$$\|\pi_k^{l,q}(u_i)_{i=1,\dots,n_k^l}\|_{L^2(\gamma_k^{l,q})} \leq C \sum_{i \in \mathcal{I}_k^{l,q}} \|u_{j_i}\|_{L^2(\Gamma_k^{l,i})}. \tag{4.6}$$

However, from (2.21) and from Remark 13, if $z \in Z_k^{l,q}$, then

$$\|z\|_{L^2(\Gamma_k^l)} \leq Ch^{\frac{1}{2}} \|z\|_{L^\infty(\gamma_k^{l,q})} \leq C\|z\|_{L^2(\gamma_k^{l,q})}. \tag{4.7}$$

Thus we deduce from (4.6) and (4.7) that

$$\|\pi_k^{l,q}(u_i)_{i=1,\ldots,n_k^l}\|_{L^2(\Gamma_k^l)} \leq C \sum_{i \in \mathcal{I}_k^{l,q}} \|u_{j_i}\|_{L^2(\Gamma_k^{l,i})}. \tag{4.8}$$

Consider the function defined on Γ_k^l by $z_k^l = \sum_{q=1}^Q \pi_k^{l,q}(i_{j_i}(u_{j_i}^*) - i_k(u_k^*))$ and the function z_k defined on $\partial\Omega_k$ by assembling the functions z_k^l. Now we have to correct z_k in order to satisfy (4.2).

Second step. The second correction consists of taking the function $z_k + e_k$, where e_k is obtained by assembling $e_k^l \equiv (\pi^*(i_{j_i}(u_{j_i}^*) - i_k(u_k^*))_{i=1,\ldots,n_k^l})$. Finally we set

$$u_k \equiv i_k(u_k^*) + L_k(z_k + e_k).$$

The vector $(u_k)_{k=1,\ldots,K}$ belongs to Y.

It is possible to check the following estimate.

LEMMA 4.2. *Let $u^* \in H^1(\Omega)$ such that for $1 \leq k \leq K$, $u^*|_{\Omega_k} \in H^{\sigma_k}(\Omega_k)$ with $2 \geq \sigma_k > 1$. Then the vector $u \in Y$ constructed above satisfies*

$$\sum_{k=1}^K \frac{1}{h_k} |u_k^* - u_k|_{L^2(\Omega_k)} + |u_k^* - u_k|_{H^1(\Omega_k)} \leq C \sum_{k=1}^K h_k^{\sigma_k - 1} |u_k^*|_{H^{\sigma_k}(\Omega_k)}. \tag{4.9}$$

Finally, the conclusions of Theorem 3.6 hold for the strengthened matching condition.

5. Additive Schwarz preconditioners. The linear system corresponding to the solution of (2.22) is sparse, symmetric positive definite, and usually ill-conditioned. It is often mandatory to use a preconditioned iterative method to solve it. We propose hereafter two additive Schwarz preconditioners. These methods are based on finding a decomposition

$$Y = \sum_{k=0}^K Y_k,$$

where for $k = 1, \ldots, K$ the subspace Y_k will contain multivalued functions localized, roughly speaking, around the subdomain Ω_k, and the subspace Y_0 (called coarse space) will contain slowly varying functions, the latter space being chosen for the speed of convergence to be independent of K.

5.1. A coarse space Y_0. We assume that we can find a quasi-uniform triangular mesh $\mathcal{T}_H$ of Ω, whose elements have diameters of order H, and we define $\mathcal{N}_H$ as the set of the nodes of $\mathcal{T}_H$ lying inside Ω. We call $\hat{Y}_0$ the space of the continuous functions on $\overline{\Omega}$ vanishing on $\partial\Omega$ and whose restriction to any element of $\mathcal{T}_H$ is linear. For each node x in $\mathcal{N}_H$, let ζ_x be the nodal shape function corresponding to the node x.

The space $\hat{Y}_0$ is not a subspace of Y. However, for a given function $\hat{u}_0 \in \hat{Y}_0$, it is possible to obtain an element of Y by applying the same process to $\hat{u}_0$ as that used for the best approximation estimate, i.e.,

1. interpolate $\hat{u}_0$ in the subdomains,

2. correct on the boundary of subdomains by mortar projections,
3. add the trivial lifting of these corrections to the interpolated functions.

We call $I_0\hat{u}_0$ the element of Y obtained in this way. We have the following easy lemma.

LEMMA 5.1. *There exists a positive constant C independent of h and H, such that for $\hat{u}_0 \in Y_0$,*

$$|I_0\hat{u}_0|_* \leq C|\hat{u}_0|_{H^1(\Omega)}. \tag{5.1}$$

Proof. The result follows from (3.16) and from the inverse inequality

$$\forall \sigma,\ 1 < \sigma < \frac{3}{2}, \quad |\hat{u}_0|_{H^\sigma(\Omega)} \leq CH^{-\sigma+1}|\hat{u}_0|_{H^1(\Omega)}. \qquad \square$$

We define

$$Y_0 = I_0\hat{Y}_0. \tag{5.2}$$

We also define the bilinear form b_0 on $\hat{Y}_0$ by

$$b_0(\hat{u}_0, \hat{v}_0) = \int_\Omega \nabla\hat{u}_0 \cdot \nabla\hat{v}_0 \tag{5.3}$$

and the projection operator $\hat{T}_0$ from Y into $\hat{Y}_0$ by

$$\forall u \in Y, \quad \hat{T}_0 u \in \hat{Y}_0, \quad \text{and} \quad b_0(\hat{T}_0 u, \hat{v}_0) = a(u, I_0\hat{v}_0) \quad \forall \hat{v}_0 \in \hat{Y}_0. \tag{5.4}$$

The operator $\hat{T}_0$ is clearly continuous from Y into $\hat{Y}_0$. We also define the operator $T_0 = I_0 \circ \hat{T}_0$ from Y into Y_0.

For the following, we need to find a Clément-like interpolation operator (see [8]) from Y to $\hat{Y}_0$. For that, we associate with each node x in $\mathcal{N}_H$ a subdomain $\Omega_{k(x)}$ such that $x \in \Omega_{k(x)}$. For $u \in Y$, we define Ξu by

$$\Xi u = \sum_{x \in \mathcal{N}_H} \langle u_{k(x)} \rangle \zeta_x. \tag{5.5}$$

LEMMA 5.2. *Under Assumptions* 1 *to* 6, *there exists a constant C such that, for $u \in Y$,*

$$b_0(\Xi u, \Xi u) \leq Ca(u, u). \tag{5.6}$$

Proof. From the quasi uniformity of the mesh $\mathcal{T}_H$, there exists a constant C such that

$$\begin{aligned} b_0(\Xi u, \Xi u) &\leq C \sum_{t \in \mathcal{T}_H} \sum_{x,y \in \mathcal{N}_H, x,y \in \partial t} (\Xi u(x) - \Xi u(y))^2 \\ &= C \sum_{t \in \mathcal{T}_H} \sum_{x,y \in \mathcal{N}_H, x,y \in \partial t} (\langle u_{k(x)} \rangle - \langle u_{k(y)} \rangle)^2. \end{aligned}$$

The desired estimate is a consequence of Lemma 2.5. $\square$

LEMMA 5.3. *Under Assumptions* 1 *to* 6, *there exists a constant C such that for all $u \in Y$,*

$$\sum_{k=1}^K |(I_0\Xi u)_k|^2_{H^1(\Omega_k)} + \frac{1}{H^2} \sum_{k=1}^K \|u_k - (I_0\Xi u)_k\|^2_{L^2(\Omega_k)} \leq Ca(u, u). \tag{5.7}$$

Proof. We have

$$\int_{\Omega_k} |u_k - (I_0\Xi u)_k|^2$$
$$\leq 3\int_{\Omega_k} |u_k - \langle u_k\rangle|^2 + 3\int_{\Omega_k} |\langle u_k\rangle - \Xi u|^2 + 3\int_{\Omega_k} |(\Xi u - (I_0\Xi u)_k|^2.$$

We have from the Poincaré–Wiertinger inequality that

$$\int_{\Omega_k} |u_k - \langle u_k\rangle|^2 \leq CH^2 \int_{\Omega_k} |\nabla u_k|^2.$$

We also have that

$$\int_{\Omega_k} |\langle u_k\rangle - \Xi u|^2 \leq CH^2 \sum_{t\in\mathcal{T}_H, |t\cap\Omega_k|>0} \sum_{x \text{ vertex of } t} (\langle u_{k(x)}\rangle - \langle u_k\rangle)^2$$
$$\leq CH^2 \sum_{t\in\mathcal{T}_H, |t\cap\Omega_k|>0} \sum_{l:|\Omega_l\cap t|>0} \int_{\Omega_l} |\nabla u_l|^2,$$

as in the proof of Lemma 2.5. Finally, as in Lemma 3.5 and from the quasi uniformity of the coarse mesh $\mathcal{T}_H$, we have that

$$\int_{\Omega_k} |\Xi u - (I_0\Xi u)_k|^2 \leq CH^2 \int_{\omega_k} |\nabla \Xi u|^2.$$

Thanks to Lemma 5.2, summing over k yields the desired estimate on $\sum_{k=1}^K \|u_k - (I_0\Xi u)_k\|^2_{L^2(\Omega_k)}$. The estimate on $\sum_{k=1}^K |(I_0\Xi u)_k|^2_{H^1(\Omega_k)}$ comes from Lemma 5.2 and from the stability of I_0. ∎

5.2. A nonoptimal Schwarz preconditioner. The first Schwarz preconditioner is very much inspired from that proposed in [5] for the case of two subdomains. The idea is to define the spaces Y_k by means of the trivial lifting operators introduced above.

We call R_k the operator in X_k defined by the following: for any $v \in X_k$, the master d.o.f. of $R_k v$ are those of v and

$$\forall w \in W_k^l, \quad \int_{\Gamma_k^l} \frac{1}{p_k^l(x)+1} R_k v(x) w(x) dx = 0.$$

We have a similar result to Lemma 3.3.

LEMMA 5.4. *There exists a positive constant C such that, for any $u \in X_k$,*

$$\|u - R_k u\|_{L^2(\Omega_k)} + h_k |u - R_k u|_{H^1(\Omega_k)} \leq C h_k^{\frac{1}{2}} \|u\|_{L^2(\partial\Omega_k)}. \tag{5.8}$$

The range of R_k is called $\hat{Y}_k$. It is an easy matter to map any element of $\hat{Y}_k$ to an element of Y by setting all the extra master d.o.f. to zero. We call I_k this trivial extension operator and Y_k the range of I_k. We clearly have that $Y = \bigoplus_{k=1}^K Y_k$ because $u = \sum_{k=1}^K I_k R_k u_k$.

We also define the bilinear form b_k on $\hat{Y}_k$ by

$$b_k(\hat{u}_k, \hat{v}_k) = \left(1 + \max_{k\neq l\in\mathcal{I}_k} \frac{h_k}{h_l}\right) \int_{\Omega_k} \nabla\hat{u}_k \cdot \nabla\hat{v}_k + \sum_{l\in\mathcal{I}_k} \sum_{x\in\mathcal{M}_k^l} \frac{h_k}{h_l} \hat{u}_k(x)\hat{v}_k(x), \tag{5.9}$$

where $\mathcal{M}_k^l$ is the set of the nodes of $\mathcal{T}_k$ for which the related d.o.f. influence some slave nodal values at the nodes of $\mathcal{T}_l$, and the projection operator $\hat{T}_k$ from Y into $\hat{Y}_k$ by

$$\forall u \in Y, \quad \hat{T}_k u \in \hat{Y}_k \quad \text{and} \quad b_k(\hat{T}_k u, \hat{v}) = a(u, I_k \hat{v}) \quad \forall \hat{v} \in \hat{Y}_k. \tag{5.10}$$

We are now ready to construct the preconditioned operator: we define the operator $T : Y \to Y$ by

$$T = \sum_{k=0}^{K} T_k = \sum_{k=0}^{K} I_k \hat{T}_k. \tag{5.11}$$

The condition number of T is analyzed in a now classical way by following the abstract additive Schwarz method theory of [17, 14, 10, 11].

LEMMA 5.5. *Let us define three relevant parameters:*

1. *Let $C_0(H,h)$ be the minimum real number such that for all $u \in Y$ there exists a sum $u = \sum_{k=0}^{K} I_k \hat{u}_k$ with $\hat{u}_k \in \hat{Y}_k$, and*

$$\sum_{k=0}^{K} b_k(\hat{u}_k, \hat{u}_k) \leq C_0^2(H,h) a(u,u).$$

2. *For $k, l \in \{1, \ldots, K\}$, let $\mathcal{E}_{kl}$ be the best constants such that for $\hat{u}_k \in \hat{Y}_k$ and $\hat{u}_l \in \hat{Y}_l$,*

$$a(I_k \hat{u}_k, I_l \hat{u}_l) \leq \mathcal{E}_{kl} a(I_k \hat{u}_k, I_k \hat{u}_k)^{\frac{1}{2}} a(I_l \hat{u}_l, I_l \hat{u}_l)^{\frac{1}{2}}.$$

   This estimate is called the strengthened Cauchy–Schwarz inequality. Let $\rho(\mathcal{E})$ be the spectral radius of $\mathcal{E} = (\mathcal{E}_{kl})_{1 \leq k,l \leq K}$.

3. *Let ω be the minimum number such that for $k \in \{0, \ldots, K\}$, for $\hat{u}_k \in \hat{Y}_k$,*

$$a(I_k \hat{u}_k, I_k \hat{u}_k) \leq \omega b_k(\hat{u}_k, \hat{u}_k).$$

Then T is invertible and symmetric with respect to the scalar product $a(\cdot,\cdot)$, and, for $u \in Y$,

$$\frac{1}{C_0^2} a(u,u) \leq a(Tu,u) \leq \omega(1 + \rho(\mathcal{E})) a(u,u). \tag{5.12}$$

In the following lemma, we prove that C_0 can be chosen as $C \max_{1 \leq k \leq K} \frac{H}{h_k}$.

LEMMA 5.6. *Under Assumptions 1 to 6 there exists a constant C such that for $u \in Y$ there exists $(\hat{u}_k)_{k \in \{0,\ldots,K\}}$ such that $u = \sum_{k=0}^{K} I_k \hat{u}_k$ and*

$$\sum_{k=0}^{K} b_k(\hat{u}_k, \hat{u}_k) \leq C \max_{1 \leq k \leq K} \frac{H}{h_k} a(u,u). \tag{5.13}$$

Proof. We take $\hat{u}_0 = \Xi u$ and $\hat{u}_k = (R_k)((u - I_0 \hat{u}_0)|_k)$. From Lemma 5.2, we already have (5.6). Let us focus on $b_k(\hat{u}_k, \hat{u}_k)$ for $k > 0$. We have $b_k(\hat{u}_k, \hat{u}_k) = (1 + \max_{l \in \mathcal{I}_k} \frac{h_k}{h_l}) A_k + B_k$, where $A_k = \int_{\Omega_k} |\nabla \hat{u}_k|^2$ and $B_k = \sum_{l \in \mathcal{I}_k \setminus \{l\}} \sum_{x \in \mathcal{M}_k^l} \frac{h_k}{h_l} \hat{u}_k^2(x)$.

$$\begin{aligned} A_k &= \int_{\Omega_k} |\nabla (u - I_0 \hat{u}_0)|_k - \nabla (I - R_k)(u - I_0 \hat{u}_0)|_k|^2 \\ &\leq C \left(\int_{\Omega_k} |\nabla (u - I_0 \hat{u}_0)|_k|^2 + \frac{1}{h_k} \int_{\partial \Omega_k} |(u - I_0 \hat{u}_0)|_k|^2 \right) \end{aligned}$$

by Lemma 5.4. But a very crude argument and Lemmas 5.1 and 5.2 lead to

$$\sum_{k=1}^{K}\left(1+\max_{l\in\mathcal{I}_k}\frac{h_k}{h_l}\right)\int_{\Omega_k}|\nabla(I_0\hat{u}_0)|_k|^2\le C\max_k\frac{H}{h_k}\int_{\Omega}|\nabla\hat{u}_0|^2\le C\max_k\frac{H}{h_k}a(u,u).$$

It remains for us to study $\frac{1}{h_k}\int_{\partial\Omega_k}|(u-I_0\hat{u}_0)|_k|^2$, which can be bounded exactly as in the proof of Lemma 2.6. We have

$$\begin{aligned}&\int_{\partial\Omega_k}|u-I_0\hat{u}_0|_k|^2\\&\le 3\int_{\partial\Omega_k}|u_k-\langle u_k\rangle|^2+3\int_{\partial\Omega_k}|\langle u_k\rangle-\hat{u}_0|^2+3\int_{\partial\Omega_k}|(\hat{u}_0-I_0\hat{u}_0)|_k|^2.\end{aligned}$$

The estimate $\int_{\partial\Omega_k}|u_k-\langle u_k\rangle|^2\le CH\int_{\Omega_k}|\nabla u_k|^2$ is just a trace inequality. We have that

$$\begin{aligned}\int_{\partial\Omega_k}|\langle u_k\rangle-\hat{u}_0|^2&\le CH\sum_{t\in\mathcal{T}_H,|t\cap\Omega_k|>0}\ \sum_{x\text{ vertex of }t}(\langle u_{k(x)}\rangle-\langle u_k\rangle)^2\\&\le CH\sum_{t\in\mathcal{T}_H,|t\cap\Omega_k|>0}\ \sum_{l:|\Omega_l\cap t|>0}\int_{\Omega_l}|\nabla u_l|^2,\end{aligned}$$

as for the proof of Lemma 2.5. From the best-fit estimate (3.16) and the quasi uniformity of the coarse mesh, we obtain that

$$\begin{aligned}&\sum_{k=1}^{K}\left(1+\max_{l\in\mathcal{I}_k}\frac{h_k}{h_l}\right)\frac{1}{h_k}\int_{\partial\Omega_k}|(\hat{u}_0-I_0\hat{u}_0)|_k|^2\\&\le C\sum_{k=1}^{K}\left(1+\max_{l\in\mathcal{I}_k}\frac{h_k}{h_l}\right)\left(\frac{H}{h_k}\int_{\Omega_k}|\nabla(\hat{u}_0-I_0\hat{u}_0)|_k|^2+\frac{1}{Hh_k}\int_{\Omega_k}|(\hat{u}_0-I_0\hat{u}_0)|_k|^2\right)\\&\le C\max_k\frac{H}{h_k}\int_{\Omega}|\nabla\hat{u}_0|^2\le C\max_k\frac{H}{h_k}a(u,u).\end{aligned}$$

Thus we have proved that

$$\sum_{k=1}^{K}\left(1+\max_{l\in\mathcal{I}_k}\frac{h_k}{h_l}\right)A_k\le C\max_k\frac{H}{h_k}a(u,u). \tag{5.14}$$

It remains for us to study $\sum_{k=1}^{K}B_k$. We have that

$$\sum_{x\in\mathcal{M}_k^l}\frac{h_k}{h_l}|(u_k-I_0\hat{u}_0)|_k(x)|^2\le C\frac{1}{h_l}\left(H\int_{\Omega_k}|\nabla(u_k-I_0\hat{u}_0)|_k|^2+\frac{1}{H}\int_{\Omega_k}(u_k-I_0\hat{u}_0)|_k^2\right),$$

and we conclude the proof exactly as above. □

The constant ω in Lemma 5.5 can be chosen independently of the mesh and domain decomposition parameters, as stated in the following lemma.

LEMMA 5.7. *There exists a constant ω independent on the mesh parameters such that, for $k\in\{0,\dots,K\}$ for $\hat{u}_k\in\hat{Y}_k$,*

$$a(I_k\hat{u}_k,I_k\hat{u}_k)\le\omega b_k(\hat{u}_k,\hat{u}_k). \tag{5.15}$$

Proof. For $k=0$, (5.15) is obtained by using Lemma 5.1. For $k>0$, we have

$$\begin{aligned} a(I_k\hat{u}_k, I_k\hat{u}_k) &= \sum_{l\in\mathcal{I}_k}\int_{\Omega_l}|\nabla(I_k\hat{u}_k)_l|^2 \\ &= \int_{\Omega_k}|\nabla\hat{u}_k|^2 + \sum_{k\neq l\in\mathcal{I}_k}\int_{\Omega_l}|\nabla(I_k\hat{u}_k)_l|^2 \\ &\le \int_{\Omega_k}|\nabla\hat{u}_k|^2 + C\sum_{k\neq l\in\mathcal{I}_k}\frac{1}{h_l}\int_{\partial\Omega_l}(I_k\hat{u}_k)_l|^2_{\partial\Omega_l} \\ &\le \int_{\Omega_k}|\nabla\hat{u}_k|^2 + C\sum_{k\neq l\in\mathcal{I}_k}\frac{1}{h_l}\int_{\partial\Omega_l}\hat{u}_k|^2_{\partial\Omega_l} \\ &\le \int_{\Omega_k}|\nabla\hat{u}_k|^2 + C\sum_{k\neq l\in\mathcal{I}_k}\frac{h_k}{h_l}\sum_{x\in\mathcal{M}_k^l}\hat{u}_k(x)^2 \\ &\le \omega b_k(\hat{u}_k,\hat{u}_k). \qquad \square \end{aligned}$$

The strengthened Cauchy–Schwarz inequality comes from the fact that if $\mathcal{I}_l\cap\mathcal{I}_k=\emptyset$, then $a(I_k\hat{u}_k, I_l\hat{u}_l)=0$, from the Cauchy–Schwarz inequality and from Property 1. From the geometric assumptions, we have the following lemma.

LEMMA 5.8. *There exists a constant C such that $\rho(\mathcal{E})\le C$.*

Finally, we have the following result.

THEOREM 5.9. *Under Assumptions 1 to 6, there exists a constant C such that*

$$cond_a(T) \le C\max_k\frac{H}{h_k}. \tag{5.16}$$

REMARK 14. *Remember that the condition number of the original problem scales like* $\max_k \frac{1}{h_k^2}$.

REMARK 15. *It is also possible to generalize the first preconditioner proposed in* [5] *which is based on some harmonic lifting operator rather than trivial lifting operators. However, this preconditioner seems to be too difficult to implement and too costly. For that reason, we shall not discuss this method.*

5.3. An optimal Schwarz preconditioner. We introduce new spaces for an additive decomposition of Y. Let $\widetilde{Y}_k$ be the subspace of Y defined by

$$\widetilde{Y}_k = \left\{ \begin{array}{l} u\in Y; \\ \text{the d.o.f. of } u \\ \text{ associated with nodes not located in } \Omega_k \text{ are zero} \end{array} \right\}. \tag{5.17}$$

REMARK 16. *Note that if $|\Omega_k\cap\Omega_l|>0$, then an element of $\widetilde{Y}_k$ has d.o.f. corresponding to some mesh nodes of $\mathcal{T}_l$, which is not the case for the elements of $\hat{Y}_k$.*

REMARK 17. *An element of $\widetilde{Y}_k$ is not necessarily made of functions supported in Ω_k because the slave nodal values depend nonlocally on the d.o.f. However, the elements of $\widetilde{Y}_k$ are made of functions supported in ω_k.*

Now the mapping I_k is the canonical injection from $\widetilde{Y}_k$ into Y, and we define $b_k(\widetilde{u}_k,\widetilde{v}_k)=a(I_k\widetilde{u}_k, I_k\widetilde{v}_k)$. The projection operator $\widetilde{T}_k$ from Y onto $\widetilde{Y}_k$ is defined by

$$\forall u\in Y,\ \widetilde{T}_k u\in\widetilde{Y}_k,\quad b_k(\widetilde{T}_k u, v)=a(u, I_k v)\quad \forall v\in\widetilde{Y}_k, \tag{5.18}$$

and the preconditioned operator $T : Y \to Y$ is given by

$$T = \sum_{k=0}^{K} T_k = \sum_{k=0}^{K} I_k \widetilde{T_k}, \tag{5.19}$$

where $\widetilde{T_0} \equiv \hat{T}_0$ has been introduced in (5.4) and I_0 has been defined in section 5.1.

Again, this preconditioner is analyzed thanks to Lemma 5.5. The only noticeable difference from the previous preconditioner is the following lemma.

LEMMA 5.10. *Under Assumptions* 1 *to* 6, *there exists a constant* C *such that for* $u \in Y$ *there exists a sum* $u = \sum_{k=0}^{K} I_k \tilde{u}_k$, $\tilde{u}_k \in \widetilde{Y}_k$, *such that*

$$\sum_{k=0}^{K} b_k(\tilde{u}_k, \tilde{u}_k) \leq C \left(1 + \frac{H}{\delta}\right) a(u,u). \tag{5.20}$$

Proof. For the sake of brevity, we just give a sketch of the proof.

We take $\tilde{u}_0 = \Xi u$, where Ξu is defined in (5.5). For constructing $\tilde{u}_k$, we introduce a smooth partition of unity $(\theta_k)_{k\in\{1,\ldots,K\}}$ such that $\theta_k(x) = 0$ if $x \notin \Omega_k$ and $\|\nabla\theta_k\|_\infty \leq C\frac{1}{\delta}$. For $k \in \{1,\ldots,K\}$ and $u \in Y$, let $v^k \in X$ be given by $v_l^k = i_l(\theta_k u_l)$ and $\widetilde{v}^k \in Y$ be obtained by local correction of v^k on the boundary of the subdomains by mortar projections, as in the proof of the best approximation estimate. In fact, we have $\widetilde{v}^k \in \widetilde{Y}_k$, and it is possible to prove exactly as in [17, 14] that

$$\sum_{l\in\mathcal{I}_k} |\widetilde{v}_l^k|^2_{H^1(\Omega_l)} \leq C \sum_{l\in\mathcal{I}_k} \left(|u_l|^2_{H^1(\Omega_k)} + \frac{1}{\delta^2}\|u_l\|^2_{L^2(\Omega_k)} \right). \tag{5.21}$$

We call $\widetilde{R}_k$ the mapping from Y onto $\widetilde{Y}_k$: $\widetilde{R}_k(u) = \widetilde{v}^k$.

For $u \in Y$, we can check that $u = \sum_{k=1}^{K} \widetilde{R}_k(u)$, and therefore

$$u = I_0 \Xi u + \sum_{k=1}^{K} \widetilde{R}_k(u - I_0 \Xi u). \tag{5.22}$$

Thanks to (5.21) and Lemmas 5.2 and 5.3, we obtain the desired result. ∎

Finally, we have the following result.

THEOREM 5.11. *Under Assumptions* 1 *to* 6, *there exists a constant* C *such that*

$$cond_a(T) \leq C \left(1 + \frac{H}{\delta}\right). \tag{5.23}$$

REFERENCES

[1] Y. ACHDOU, G. ABDULAIEV, J. C. HONTAND, Y. KUZNETSOV, O. PIRONNEAU, AND C. PRUD'HOMME, *Non matching grids for fluids*, in Domain Decomposition Methods 10, Contemp. Math. 218, X. C. Cai, J. Mandel, and C. Farhat, eds., AMS, Providence, RI, 1998, pp. 3–22.

[2] Y. ACHDOU, Y. MADAY, AND O. B. WIDLUND, *Iterative substructuring preconditioners for the mortar method in two dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 551–580.

[3] A. BERGER, R. SCOTT, AND G. STRANG, *Approximate boundary conditions in the finite element method*, in Sympos. Math. X, Academic Press, London, 1972, pp. 295–313.

[4] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new non conforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar 11, H. Brezis and J.-L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.

[5] X.-C. Cai, M. Dryja, and M. Sarkis, *Overlapping nonmatching grid mortar element method for elliptic problems*, SIAM J. Numer. Anal., 36 (1999), pp. 581–606.
[6] G. Chesshire and W. Henshaw, *Composite overlapping meshes for the solution of partial differential equations*, J. Comput. Phys., 90 (1990), pp. 1–64.
[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[8] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Num., 9 (1975), pp. 77–84.
[9] M. Crouzeix, *personal communication*.
[10] M. Dryja and O. B. Widlund, *Towards a unified theory of domain decomposition algorithms for elliptic problems*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, 1990, pp. 3–21.
[11] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.
[12] F. Hecht, J. L. Lions, and O. Pironneau, *Domain decomposition algorithm for computer aided design*, in Applied Nonlinear Analysis, Kluwer/Plenum, New York, 1999, pp. 185–198.
[13] Y. A. Kuznetsov, *Overlapping domain decomposition with non matching grids*, in Domain Decomposition Methods in Sciences and Engineering, P. E. Bjørstad, M. Espedal, and D. Keyes, eds., John Wiley and Sons, New York, 1997.
[14] P. Le Tallec, *Domain decomposition methods in computational mechanics*, Comput. Mech. Adv., 1 (1994), pp. 121–220.
[15] J.-L. Lions and O. Pironneau, *Domain decomposition methods for CAD*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 73–80.
[16] S. V. Parter, *On the overlapping grid method for elliptic boundary value problems*, SIAM J. Numer. Anal., 36 (1999), pp. 819–852.
[17] B. F. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[18] J. Steger and J. Benek, *On the use of composite grid schemes in computational aerodynamics*, Comput. Methods. Appl. Mech. Engrg., 64 (1997), pp. 301–320.
[19] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 629–650

ON STABILITY OF LMS METHODS AND CHARACTERISTIC ROOTS OF DELAY DIFFERENTIAL EQUATIONS*

K. ENGELBORGHS† AND D. ROOSE†

Abstract. We investigate the use of linear multistep (LMS) methods for computing characteristic roots of systems of (linear) delay differential equations (DDEs) with multiple fixed discrete delays. These roots are important in the context of stability and bifurcation analysis. We prove convergence orders for the characteristic root approximations and analyze under what condition for the steplength the discrete integration scheme retains certain delay-independent stability properties of the original equations. Unlike existing results, we concentrate on the recovery of both stability and instability. We illustrate our findings with a number of numerical test results.

Key words. delay differential equations, stability analysis, LMS methods

AMS subject classifications. 65L06, 65L07, 65Q05

PII. S003614290037472X

1. Introduction.

In this paper we study the (asymptotic) stability of the zero solution of a system of linear (or linearized) delay differential equations (DDEs),

$$\dot{x}(t) = A_0 x(t) + \sum_{i=1}^{m} A_i x(t-\tau_i), \quad \text{where } A_i \in \mathbb{R}^{n\times n},\ i = 0,\dots,m. \tag{1.1}$$

Using $\Delta(\lambda) := \lambda I - A_0 - \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda\tau_i}$, the characteristic equation for (1.1) reads

$$\det(\Delta(\lambda)) = 0. \tag{1.2}$$

The zero solution of (1.1) is (asymptotically) stable provided all the roots $\lambda \in \mathbb{C}$ of (1.2) have (strict) negative real parts. In correspondence with the infinite-dimensional nature of the DDE, there exists an infinite number of characteristic roots λ of (1.2). However, only a finite number have real parts greater than a given constant, $\Re(\lambda) > \gamma$, $\gamma \in \mathbb{R}$ [9, Lem. I.4.1]. Hence, a numerical method that automatically computes the rightmost roots of (1.2) would be of interest.

Equation (1.2) expresses a nonstandard, nonlinear eigenvalue problem, as the matrix Δ depends nonlinearly on λ. Individual roots can be computed efficiently using a Newton–Raphson iteration with a suitable starting value. However, even if a fine grid of starting values is used, there is no guarantee of finding all roots with a given property. For the purpose of bifurcation analysis, the rightmost roots, with leading real parts, are of interest in the determination of stability properties and in the detection of bifurcations.

In [4] a method is proposed to compute the rightmost characteristic roots based on an approximation of the time integration operator associated with the DDE. Indeed,

*Received by the editors July 5, 2000; accepted for publication (in revised form) November 8, 2001; published electronically June 12, 2002. This paper presents results of the research project OT/98/16 funded by the Research Council K.U. Leuven; of the research project G.0270.00 funded by the Fund for Scientific Research, Flanders (Belgium); and of the research project IUAP P4/02 funded by the programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology, and Culture. The first author is a Postdoctoral Fellow of the Fund for Scientific Research, Flanders (Belgium).

http://www.siam.org/journals/sinum/40-2/37472.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium (koen.engelborghs@cs.kuleuven.ac.be, dirk.roose@cs.kuleuven.ac.be).

the characteristic roots of, e.g., a linear multistep (LMS) or Runge–Kutta approximation to (1.1), can be computed from a large (but standard) eigenvalue problem. Established and efficient numerical algorithms exist to compute selected (e.g., dominant or rightmost) eigenvalues of possibly very large matrices; see [23, 21].

Computational issues of this approach are discussed in [4]. In this paper we investigate the effect of the discretization on the stability. In particular, we investigate the correspondence between the characteristic roots of (1.1) and of the related discrete map obtained from an LMS method applied to (1.1). We prove convergence and investigate under what conditions the approximation retains certain delay-independent stability properties of the original equations. From these results we obtain a steplength heuristic that is used in the package DDE-BIFTOOL [3].

Part of our results were already obtained in the literature on time integration of DDEs; see, e.g., [27, 28, 10, 11]. All of these results concentrate on recovery of delay-independent stability. In the context of bifurcation analysis, the recovery of delay-independent instability, as we analyze here, is likewise important. Furthermore, we prove convergence orders for the characteristic root approximations and comment on the issue of stiffness in DDEs. A number of similar results were also proven for Runge–Kutta methods applied to DDEs using different types of interpolation for the past terms; see, e.g., [19, 14, 11, 12]. For a special type of delay equation (scalar pure-delay DDE with one delay), convergence of the roots of the characteristic equation of the forward Euler method to simple roots of the characteristic equation of the DDE is proven in [30]. Related results were also proven for Hopf bifurcations in the numerical approximation of DDEs; cf. [5, 29, 6].

The paper is structured as follows. First, we discuss stability properties of DDEs in section 3. Then we compare and extend this analysis to LMS methods applied to DDEs in section 4. We illustrate our findings with numerical results in section 5, comment on the issue of stiffness for DDEs in section 6, and conclude in section 7.

2. Illustrative example. Consider the scalar one-delay equation,

$$\dot{x}(t) = ax(t) + bx(t-\tau), \tag{2.1}$$

and its corresponding characteristic equation

$$\lambda = a + b\mathrm{e}^{-\lambda\tau}. \tag{2.2}$$

The right half-plane (RHP), $\Re(\lambda) \geq 0$, is mapped under the right-hand side of (2.2) onto a circle centered at a with radius $|b|$. If this circle lies completely in the open left half-plane (LHP), it is clear that (2.2) can have no solutions in the RHP, as the latter is mapped onto disjunct regions by the left- and right-hand sides of (2.2). This occurs whenever

$$\Re(a) + |b| < 0. \tag{2.3}$$

Hence (2.3) is a sufficient condition for stability of the zero solution of (2.1) which we could term RHP-stability of (2.1). In fact, this stability is independent of the delay τ (through (2.3)) and can be shown to be nearly equivalent to delay-independent stability (i.e., stability for all values of $\tau \geq 0$) [16].

When applying an LMS method to (2.1) it can similarly be proven that if the same circle, scaled with the steplength h of the LMS method (that is, the circle centered at ha with radius $|hb|$) is part of the stability region of the LMS method, then the zero solution of the difference equation defined by applying the LMS method to (2.1) is

stable. In other words, if h is small enough, the LMS method will capture the RHP-stability of the original DDE (2.1). As such, this condition is a direct generalization of the well-known result for ordinary differential equations (where $b = 0$ and ha should lie in the stability region of the LMS method).

The purpose of this paper is to generalize these properties to systems of equations with multiple delays. In this more general case, the RHP is mapped to more complicated regions than the circle considered above. We also consider RHP-instability. Namely, when the above circle lies completely in the RHP, this can be proven to be a sufficient condition for instability. Based on these results, our analysis leads to a steplength heuristic used to compute the correct stability and bifurcations of a given DDE system.

3. Sufficient conditions for DDE stability. Rewrite the characteristic equation (1.2) as

$$\lambda \in \sigma\left(A_0 + \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda\tau_i}\right), \tag{3.1}$$

where $\sigma(B)$ denotes the spectrum of a given matrix B. The key idea is to study the mapping of the RHP under the left- and right-hand sides of (3.1). If the right-hand side of (3.1) maps the closed RHP into the open LHP, then it is clear that there can be no solutions λ in the closed RHP, and hence there can be no unstable characteristic roots.

First we introduce some necessary notation. Let $\mathbb{C}_0^+$, $\mathbb{C}^+$ denote the open, respectively, closed, RHP,

$$\mathbb{C}_0^+ = \{\lambda \in \mathbb{C} \,|\, \Re(\lambda) > 0\} \quad \text{and} \quad \mathbb{C}^+ = \{\lambda \in \mathbb{C} \,|\, \Re(\lambda) \geq 0\},$$

and corresponding definitions for the open, respectively, closed, LHP, $\mathbb{C}_0^-$, respectively, $\mathbb{C}^-$. When studying the mapping of the right-hand side of (3.1) it is necessary, for reasons that will be clear later on, to avoid dependency on the delays. Therefore, we define the following set-valued function $\Sigma(\cdot)$ as

$$\Sigma(C) = \bigcup_{(\lambda_1,\ldots,\lambda_m)\in C\times C\times\cdots\times C} \sigma\left(A_0 + \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda_i}\right), \tag{3.2}$$

where $C \subset \mathbb{C}$. Note that we have replaced each $\lambda\tau_i$ in (3.1) by a separate λ_i in (3.2).

The mapping of the RHP under the right-hand side of (3.1) is included in the mapping of the RHP under Σ, that is,

$$\bigcup_{\lambda\in\mathbb{C}^+} \sigma\left(A_0 + \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda\tau_i}\right) \subseteq \Sigma(\mathbb{C}^+), \tag{3.3}$$

where the equality holds when there is only one nonzero delay. Note $\sigma(A_0+\sum_{i=1}^m A_i\mathrm{e}^{-\lambda_i})$; also note that the region (or regions) defined by $\Sigma(\mathbb{C}^+)$ form a bounded subset of $\mathbb{C}$ because

$$\begin{array}{rcl} \lambda \in \Sigma(\mathbb{C}^+) & \Rightarrow & |\lambda| \leq \|A_0 + \sum_{i=1}^m A_i \mathrm{e}^{-\lambda_i}\| \text{ with } \Re(\lambda_i) \geq 0, \\ & \Rightarrow & |\lambda| \leq \|A_0\| + \sum_{i=1}^m \|A_i\| |\mathrm{e}^{-\lambda_i}| \text{ with } \Re(\lambda_i) \geq 0, \\ & \Rightarrow & |\lambda| \leq \sum_{i=0}^m \|A_i\|. \end{array} \tag{3.4}$$

We distinguish three different cases, depending on the location of $\Sigma(\mathbb{C}^+)$.

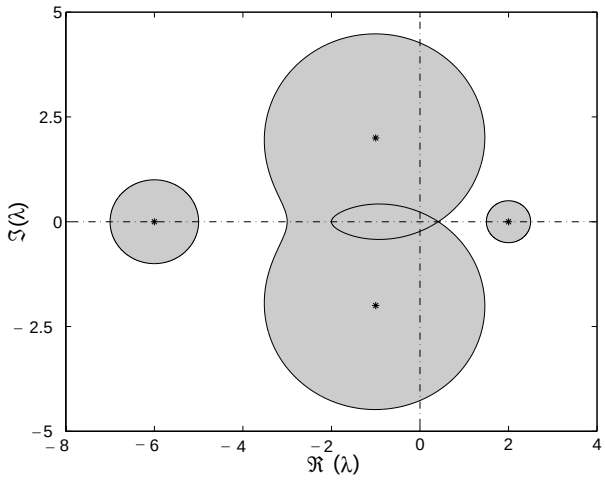


FIG. 3.1. *The mapping of the RHP under* Σ *(gray areas), the mapping of the imaginary axis under* Σ *(solid lines), and the eigenvalues of* A_0 $(*)$ *for the one-delay case with system matrices given by* (3.5).

DEFINITION 3.1. *Definition of stability of* Σ.

- *If* $\Sigma(\mathbb{C}^+) \subset \mathbb{C}_0^-$, *then we call* Σ *stable.*
- *If* $\Sigma(\mathbb{C}^+) \subset \mathbb{C}_0^+$, *then we call* Σ *unstable.*
- *If* $\exists \xi_0 \in \mathbb{R}_0 = \mathbb{R} \setminus \{0\} : \mathrm{i}\xi_0 \in \Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}_0\})$, *then we call* Σ *Hopf-like.*

An illustration of this is given in Figure 3.1 using the matrices

$$A_0 = \begin{bmatrix} -1 & 2 & 0 & 0 \\ -2 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -6 \end{bmatrix} \text{ and } A_1 = \begin{bmatrix} 2 & 2 & 2 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{3.5}$$

To check for stability of Σ it would be convenient if we could look only at the closed curve(s) described by $\Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}\})$ (cf., the lines in Figure 3.1). Because of the periodicity of $\mathrm{e}^{-\mathrm{i}\xi}$, it suffices to examine the latter only for $\xi \in [0, 2\pi]$, i.e., $\Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}\}) = \Sigma(\{\mathrm{i}\xi \,|\, \xi \in [0, 2\pi]\})$. For general matrices A_i, $i = 0, \ldots, m$, it is, however, not immediately clear whether these curves really form the boundary of $\Sigma(\mathbb{C}^+)$ in the complex plane. We now prove two theorems in this regard.

THEOREM 3.2. *The following statements hold:*

$$\begin{array}{c} \sigma(A_0 + A_1\mathrm{e}^{-\mathrm{i}\xi}) \subset \mathbb{C}_0^- \ \forall\, \xi \in [0, 2\pi], \\ \Updownarrow \\ \sigma(A_0 + A_1\mathrm{e}^{-\lambda}) \subset \mathbb{C}_0^- \ \forall\, \lambda \in \mathbb{C}^+, \end{array} \tag{3.6}$$

and

$$\begin{array}{c} \sigma(A_0 + A_1\mathrm{e}^{-\mathrm{i}\xi}) \subset \mathbb{C}_0^+ \ \forall\, \xi \in [0, 2\pi], \\ \Updownarrow \\ \sigma(A_0 + A_1\mathrm{e}^{-\lambda}) \subset \mathbb{C}_0^+ \ \forall\, \lambda \in \mathbb{C}^+. \end{array} \tag{3.7}$$

Proof. We prove only (3.6). The proof of (3.7) is analogous. It is clear that we need to prove only the downward implication.

We first treat the case of nonsingular A_1. We start from

$$\sigma(A_0 + A_1 \mathrm{e}^{-\mathrm{i}\xi}) \subset \mathbb{C}_0^-, \qquad \xi \in [0, 2\pi], \tag{3.8}$$

which we rewrite as

$$\det(\lambda I - A_0 - A_1 \mathrm{e}^{-\mathrm{i}\xi}) \neq 0,\ \xi \in [0, 2\pi],\ \Re(\lambda) \geq 0,$$

which, in turn, is equivalent to

$$\begin{array}{ll} \text{(a)} & \det(I - (\lambda I - A_0)^{-1} A_1 \mathrm{e}^{-\mathrm{i}\xi}) \neq 0,\ \xi \in [0, 2\pi],\ \Re(\lambda) \geq 0,\ \lambda \notin \sigma(A_0), \\ \text{(b)} & \det(\lambda I - A_0 - A_1 \mathrm{e}^{-\mathrm{i}\xi}) \neq 0,\ \xi \in [0, 2\pi],\ \Re(\lambda) \geq 0,\ \lambda \in \sigma(A_0). \end{array} \tag{3.9}$$

Part (a) of (3.9) is equivalent to saying that $(\lambda I - A_0)^{-1} A_1$ should have no eigenvalues on the unit circle,

$$|\sigma((\lambda I - A_0)^{-1} A_1)| \neq 1,\ \Re(\lambda) \geq 0,\ \lambda \notin \sigma(A_0),$$

(where the inequality holds over all elements of $\sigma(\cdot)$) and, because

$$\lim_{\Re(\lambda) \to +\infty} \|(\lambda I - A_0)^{-1} A_1\| = 0,$$

it is equivalent to

$$|\sigma((\lambda I - A_0)^{-1} A_1)| < 1,\ \Re(\lambda) \geq 0,\ \lambda \notin \sigma(A_0). \tag{3.10}$$

Suppose $\lambda_0 \in \sigma(A_0)$ with $\Re(\lambda_0) \geq 0$. For $\lambda \notin \sigma(A_0)$, we have that the matrix $A_1^{-1}(\lambda I - A_0)$ is regular and becomes singular in the limit $\lambda \to \lambda_0$, $\Re(\lambda) \geq 0$. By continuity of the eigenvalues, $A_1^{-1}(\lambda I - A_0)$ has an eigenvalue converging to zero and, correspondingly, $(\lambda I - A_0)^{-1} A_1$ has an eigenvalue converging to infinity (in modulus). The latter contradicts (3.10). Hence, from (3.9) it follows that $\sigma(A_0) \subset \mathbb{C}_0^-$. Therefore, we can conclude that (3.9), and thus also (3.8), is equivalent to

$$\sigma(A_0) \subset \mathbb{C}_0^-, \quad |\sigma((\lambda I - A_0)^{-1} A_1)| < 1,\ \Re(\lambda) \geq 0.$$

Note that the above equivalence still holds if A_1 is replaced by $A_1 \mathrm{e}^{-r}$ for some fixed $r \in \mathbb{R}^+$. Using this, we conclude that

$$\begin{array}{c} \sigma(A_0 + A_1 \mathrm{e}^{-\mathrm{i}\xi}) \subset \mathbb{C}_0^-,\ \xi \in [0, 2\pi], \\ \Updownarrow \\ \sigma(A_0) \subset \mathbb{C}_0^-, \quad |\sigma((\lambda I - A_0)^{-1} A_1)| < 1,\ \Re(\lambda) \geq 0, \\ \Downarrow \\ \sigma(A_0) \subset \mathbb{C}_0^-, \quad |\sigma((\lambda I - A_0)^{-1} A_1 \mathrm{e}^{-r})| < 1,\ \Re(\lambda) \geq 0,\ r \geq 0, \\ \Updownarrow \\ \sigma(A_0 + A_1 \mathrm{e}^{-(r+\mathrm{i}\xi)}) \subset \mathbb{C}_0^-,\ \xi \in [0, 2\pi],\ r \geq 0, \\ \Updownarrow \\ \sigma(A_0 + A_1 \mathrm{e}^{-\lambda}) \subset \mathbb{C}_0^-,\ \lambda \in \mathbb{C}^+. \end{array}$$

If A_1 is singular, then there exists an arbitrary small perturbation to a regular matrix. By continuity of the eigenvalues and the above statement, it follows that, in this case,

$$\begin{array}{c} \sigma(A_0 + A_1 \mathrm{e}^{-\mathrm{i}\xi}) \subset \mathbb{C}_0^-, \qquad \xi \in [0, 2\pi], \\ \Downarrow \\ \sigma(A_0 + A_1 \mathrm{e}^{-\lambda}) \subset \mathbb{C}^-, \qquad \lambda \in \mathbb{C}^+. \end{array}$$

However, since $\sigma(A_0 + A_1 e^{-i\xi}),\ \xi \in [0, 2\pi]$, forms a set of bounded and continuous curves, there exists an $s > 0$ such that the following holds:

$$\begin{gathered}
\sigma(A_0 + A_1 e^{-i\xi}) \subset \mathbb{C}_0^-,\ \xi \in [0, 2\pi], \\
\Downarrow \\
\exists s > 0,\ \sigma(A_0 + sI + A_1 e^{-i\xi}) \subset \mathbb{C}_0^-,\ \xi \in [0, 2\pi], \\
\Downarrow \\
\exists s > 0,\ \sigma(A_0 + sI + A_1 e^{-\lambda}) \subset \mathbb{C}^-,\ \lambda \in \mathbb{C}^+, \\
\Downarrow \\
\sigma(A_0 + A_1 e^{-\lambda}) \subset \mathbb{C}_0^-,\ \lambda \in \mathbb{C}^+,
\end{gathered}$$

which ends the proof. ∎

With the above proof for the one-delay case, it is rather straightforward to extend this result to the multiple delay situation. We will use the vector notation $\vec{\tau} = (\tau_1, \ldots, \tau_m)$, where inequalities hold componentwise, e.g., $\vec{\tau} \geq 0$, and we will use the notation C^m for $C \times C \times \cdots \times C$, where $C \subset \mathbb{C}$.

THEOREM 3.3. *The following statements hold:*

$$\begin{gathered}
\sigma\left(A_0 + \sum_{i=1}^{m} A_i e^{-i\xi_i}\right) \subset \mathbb{C}_0^-\ \forall \vec{\xi} \in [0, 2\pi]^m, \\
\Updownarrow \\
\sigma\left(A_0 + \sum_{i=1}^{m} A_i e^{-\lambda_i}\right) \subset \mathbb{C}_0^-\ \forall \vec{\lambda} \in (\mathbb{C}^+)^m,
\end{gathered} \tag{3.11}$$

and

$$\begin{gathered}
\sigma\left(A_0 + \sum_{i=1}^{m} A_i e^{-i\xi_i}\right) \subset \mathbb{C}_0^+\ \forall \vec{\xi} \in [0, 2\pi]^m, \\
\Updownarrow \\
\sigma\left(A_0 + \sum_{i=1}^{m} A_i e^{-\lambda_i}\right) \subset \mathbb{C}_0^+\ \forall \vec{\lambda} \in (\mathbb{C}^+)^m.
\end{gathered} \tag{3.12}$$

Proof. We prove the first statement. Applying Theorem 3.2, it follows that

$$\sigma\left(A_0 + \sum_{i=1}^{m} A_i e^{-i\xi_i}\right) \subset \mathbb{C}_0^-\ \forall \xi_1 \in [0, 2\pi]$$

is equivalent to

$$\sigma\left(A_0 + A_1 e^{-\lambda_1} + \sum_{i=2}^{m} A_i e^{-i\xi_i}\right) \subset \mathbb{C}_0^-\ \forall \lambda_1 \in \mathbb{C}^+$$

for fixed values of $\xi_2, \ldots, \xi_m$ (consider $A_0 + \sum_{i=2}^{m} A_i e^{-i\xi_i}$ as a new and fixed A_0 in Theorem 3.2). Applying Theorem 3.2 recursively on ξ_2 until ξ_m proves the above statement. ∎

Hence, for analyzing the stability of Σ, it suffices to examine the mapping of the imaginary axis under Σ. Note that the proof of Theorem 3.3 relies on the independent variation of the λ_i as introduced in the definition of Σ (3.2).

We are now ready to state the following theorem.

THEOREM 3.4 (relation of the stability of Σ to the stability of the zero solution of (1.1)).

(i) *If Σ is stable, then the zero solution of* (1.1) *is stable for all $\vec{\tau} \geq 0$.*

(ii) *If Σ is unstable, then the zero solution of* (1.1) *is unstable for all $\vec{\tau} \geq 0$.*

(iii) *The characteristic equation of* (1.1) *has a purely imaginary root $\lambda = \mathrm{i}\xi_0 \neq 0$ for some $\vec{\tau} \geq 0$ if and only if Σ's stability is Hopf-like.*

Proof.

(i) Since the mapping of the closed RHP under the right-hand side of the characteristic equation (3.1) is part of $\Sigma(\mathbb{C}^+)$ (cf., (3.3)), and the latter is, by assumption, mapped into the open LHP, then there can be no characteristic roots with $\Re(\lambda) \geq 0$. Since there exists only a finite number of roots with real part greater than any $\gamma < 0$, it follows that

$$\sup_{\det(\Delta(\lambda))=0} \Re(\lambda) < 0.$$

Because the stability of Σ does not depend on the values of $\vec{\tau}$, it follows that the zero solution of (1.1) is asymptotically stable for all $\vec{\tau} \geq 0$.

(ii) By continuity of the eigenvalues of $A_0 + \sum_{i=1}^m A_i \mathrm{e}^{-\lambda\tau_i}$ as a function of λ, it is possible to decompose $\sigma(A_0 + \sum_{i=1}^m A_i \mathrm{e}^{-\lambda\tau_i})$ into n continuous functions $\sigma_l(\lambda)$, $l = 1, \ldots, n$, such that

$$\sigma\left(A_0 + \sum_{i=1}^m A_i \mathrm{e}^{-\lambda\tau_i}\right) \equiv \{\sigma_1(\lambda), \ldots, \sigma_n(\lambda)\},$$

with correct multiplicity. Note that this decomposition may not be unique. Let G denote the closed convex hull of $\Sigma(\mathbb{C}^+)$. By assumption we have $G \subset \mathbb{C}_0^+$ and thus also

$$\bigcup_{\lambda\in G} \sigma_l(\lambda) \subset \bigcup_{\lambda\in\mathbb{C}^+} \sigma_l(\lambda) \subset \Sigma(\mathbb{C}^+) \subset G.$$

By Brouwer's fixed point theorem [13, Thm. VIII.1.1], this implies that $\sigma_l(\cdot)$ has a fixed point λ in G. This point is a solution of (3.1) with positive real part. Hence, since the stability of Σ does not depend on $\vec{\tau}$, the zero solution of (1.1) is unstable for all $\vec{\tau} \geq 0$.

(iii) From

$$\det\left(\mathrm{i}\xi_0 I - A_0 - \sum_{i=1}^m A_i \mathrm{e}^{-\mathrm{i}\xi_0\tau_i}\right) = 0$$

with $\xi_0 \neq 0$ it follows that $\mathrm{i}\xi_0 \in \Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}_0\})$ by taking $\lambda_i = \mathrm{i}\xi_0\tau_i$ if $\tau_i > 0$ and $\lambda_i = \mathrm{i}2\pi$ otherwise. On the other hand, if $\mathrm{i}\xi_0 \in \Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}_0\})$, then there exist $\lambda_i = \mathrm{i}\xi_i$, $\xi_i \neq 0$, such that $\mathrm{i}\xi_0 \in \sigma(A_0 + \sum_{i=1}^m \mathrm{e}^{-\lambda_i})$. Hence $\mathrm{i}\xi_0$ is a characteristic root for delays chosen such that $\tau_i = (\xi_i + 2\pi k)/\xi_0$, where $k \in \mathbb{Z}$ is chosen such that $\tau_i \geq 0$. ∎

As a result of this theorem we conclude that the stability of Σ is almost equivalent to delay-independent stability of the DDE (1.1). The "almost" refers to the fact that some cases are excluded from the definition of stability of Σ. Namely, as $\Sigma(\mathbb{C}^+)$ is not necessarily a connected region (rather, it consists of at most n connected components), it can map in both the RHP and the LHP without mapping onto the imaginary

axis. In this case, part (ii) of Theorem 3.4 still holds because one of the σ_l, $1 \leq l \leq n$ then maps into the open RHP (giving rise to a fixed point, and hence, to an unstable characteristic root). Thus, from all cases, only one degenerate case is excluded, namely, when $\Sigma(\mathbb{C}^+)$ maps onto the imaginary axis but only at the origin. This is in correspondence with the theorem on delay-independent stability given in [16] for the one-delay case.

In the next section we will see when an LMS method applied to (1.1) captures the stability properties of Σ.

4. LMS methods applied to DDEs. Consider the linear k-step formula [7]

$$\sum_{j=0}^{k} \alpha_j y_{s+j} = h \sum_{j=0}^{k} \beta_j f_{s+j} \tag{4.1}$$

applied to (1.1). Here, h is a (fixed) step size, $\alpha_k = 1$, and y_j and f_j present numerical approximations of $y(t)$, respectively, $A_0 y(t) + \sum_{i=1}^{m} A_i y(t - \tau_i)$, at the mesh point $t_j = jh$. During time integration, (4.1) is solved for y_{s+k} (for successive values of s) based on the previously computed mesh points y_{s+j}, $j < k$, and the initial condition.

The right-hand side approximation is chosen as

$$f_j := A_0 y_j + \sum_{i=1}^{m} A_i \tilde{y}(t_j - \tau_i), \tag{4.2}$$

where $\tilde{y}(t_j - \tau_i)$ presents an approximation for $y(t_j - \tau_i)$ obtained from the previously computed mesh points y_l, $l < j$. In particular, the use of so-called *Nordsieck interpolation* leads to

$$\tilde{y}(t_i + \epsilon h) = \sum_{l=-s_-}^{s_+} \psi_l(\epsilon) y_{i+l}, \qquad \epsilon \in [0, 1), \tag{4.3}$$

where the ψ_l are the Lagrange interpolation polynomials,

$$\psi_l(\epsilon) := \prod_{o=-s_-,\, o \neq l}^{s_+} \frac{\epsilon - o}{l - o}.$$

Applying (4.3) to (4.2) for the linear multiple delay DDE (1.1) we get

$$\sum_{j=0}^{k} \alpha_j y_{s+j} = h \sum_{j=0}^{k} \beta_j \left(A_0 y_{s+j} + \sum_{i=1}^{m} A_i \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i) y_{s+l+j-L_i} \right), \tag{4.4}$$

where $L_i := \lceil \tau_i / h \rceil$, $\epsilon_i := L_i - \tau_i / h \in [0, 1)$ (and $\lceil r \rceil$ is the smallest integer greater than or equal to $r \in \mathbb{R}$). In order to avoid the use of future mesh points while evaluating the past terms, we require that $L_i \geq s_+$ or, that,

$$\tau_i \geq (s_+ - \epsilon_i) h, \qquad i = 1, \ldots, m. \tag{4.5}$$

When there is only one delay or, more generally, when the delays are *commensurable*, that is, when all delays are an integer multiple of a single delay,

$$\tau_i = n_i \tau_0, \ n_i \in \mathbb{N}, \ i = 1, \ldots, m, \tag{4.6}$$

then, by choosing $h = \tau_0/L$, $L \in \mathbb{N}_0$, delayed mesh points are mapped onto mesh points in the past,

$$t_j - \tau_i = t_j - n_i L h = t_{j-n_i L}. \tag{4.7}$$

In this situation the mesh is called *constrained* and interpolation is no longer needed,

$$f_j = A_0 y_j + \sum_{i=1}^{m} A_i y_{j-n_i L}.$$

The LMS method is *explicit* whenever $\beta_k = 0$, and y_{s+k} can be directly computed from (4.1) by evaluating

$$y_{s+k} = -\sum_{j=0}^{k-1} \alpha_j y_{s+j} + h \sum_{j=0}^{k-1} \beta_j f_{s+j},$$

whose right-hand side depends only on y_{s+j}, $j < k$. Otherwise, the method is called *implicit*. Note that, if the (present) point y_{s+k} occurs due to a small delay, $\tau_i \le s_+ h$, $1 \le i \le n$, an explicit LMS method ($\beta_k = 0$) becomes implicit. The LMS method is of *order* p if

$$\sum_{j=0}^{k} \alpha_j = 0 \text{ and } \sum_{j=0}^{k} j^q \alpha_j = q \sum_{j=0}^{k} j^{q-1} \beta_j \text{ for } q = 1, \dots, p. \tag{4.8}$$

The method is called *consistent* if it is at least of order 1.

4.1. Stability of the LMS difference scheme. The stability of the difference scheme (4.4) can be obtained from a large but standard eigenvalue problem. In particular, the characteristic equation for the difference scheme (4.4) looks like

$$\det\left(\left(\sum_{j=0}^{k} \alpha_j \mu^{s+j}\right) I - h \left(\sum_{j=0}^{k} \beta_j \mu^{s+j}\right) \left(A_0 + \sum_{i=1}^{m} A_i \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i) \mu^{l-L_i}\right)\right) = 0. \tag{4.9}$$

To avoid spurious solutions μ (not influenced by the system matrices A_0, A_1) we require that the polynomials in ζ,

$$\sum_{j=0}^{k} \alpha_j \zeta^j \text{ and } \sum_{j=0}^{k} \beta_j \zeta^j, \tag{4.10}$$

have no roots in common. The corresponding LMS method is then called *irreducible* [7, section III.2.4]. To compare the roots μ with the solutions of the characteristic equation, we substitute μ for λ, using the relation

$$\mu = \exp(\lambda h). \tag{4.11}$$

Then, after dividing (4.9) by $h \sum_{j=0}^{k} \beta_j \mu^{s+j}$, we obtain

$$\det\left(\frac{1}{h}\left(\frac{\sum_{j=0}^{k} \alpha_j \mathrm{e}^{\lambda j h}}{\sum_{j=0}^{k} \beta_j \mathrm{e}^{\lambda j h}}\right) I - \left(A_0 + \sum_{i=1}^{m} A_i \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i) \mathrm{e}^{-\lambda(L_i - l)h}\right)\right) = 0. \tag{4.12}$$

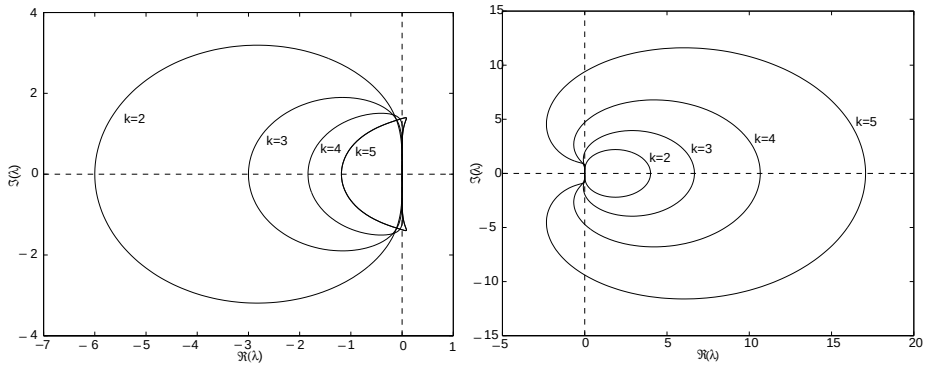


FIG. 4.1. *Boundaries of the stability regions for the Adams–Moulton (AM) (left), respectively, backwards-differentiation (BDF) (right) LMS methods of orders $k = 2, 3, 4, 5$. For the AM methods, the bounded stability region is the interior of curves depicted (left). For the BDF methods, the unbounded stability region is the exterior of the curves depicted (right).*

By the requirement of irreducibility and the fact that solutions $\mu = 0$ do not influence stability, it follows that the stability of the LMS method is now determined by the real parts of the (finite) solutions λ of (4.12) just as for the characteristic roots of (1.2).

To analyze the characteristic equation (4.12), set

$$\mathrm{LMS}(\lambda) := \frac{\sum_{j=0}^{k} \alpha_j \mathrm{e}^{\lambda j}}{\sum_{j=0}^{k} \beta_j \mathrm{e}^{\lambda j}}.$$

We will call the region $\mathbb{C} \setminus \mathrm{LMS}(\mathbb{C}^+)$ the *stability region* of the corresponding LMS method (for a more precise definition see [8, section V.1.1]). It is the region into which $\mathrm{LMS}(\cdot)$ does not map any unstable λ, $\Re(\lambda) \geq 0$. We further require that the *boundary locus curve* (which is the mapping of the imaginary axis under $\mathrm{LMS}(\cdot)$) describes the boundary of this region (this property is sometimes referred to as *property C* [8, section V.4.5]). A number of such regions are depicted in Figure 4.1.

The characteristic equation (4.12) is equivalent to

$$\frac{1}{h}\mathrm{LMS}(\lambda h) \in \sigma\left(A_0 + \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda\tau_i} \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i)\mathrm{e}^{\lambda(l-\epsilon_i)h}\right). \tag{4.13}$$

First we prove that the mapping of the RHP under the right-hand side of (4.13) is a subset of $\Sigma(\mathbb{C}^+)$. For this we need the following lemma.

LEMMA 4.1. *The polynomial in z,*

$$\sum_{l=-s_-}^{s_+} \psi_l(\epsilon) z^{s_+-l},$$

maps the (closed) unit circle into itself whenever $\epsilon \in [0, 1]$ and $s_- \leq s_+ \leq s_- + 2$. That is,

$$|z| \leq 1 \Rightarrow \left|\sum_{l=-s_-}^{s_+} \psi_l(\epsilon) z^{s_+-l}\right| \leq 1.$$

Proof. The proof of this lemma can be found in [17] and [26]. ∎

Under the conditions of this lemma, we obtain the following result.

LEMMA 4.2. *The mapping of the closed RHP under the right-hand side of* (4.13) *is a subset of* $\Sigma(\mathbb{C}^+)$.

Proof. For $\lambda \in \mathbb{C}^+$ we have

$$\begin{aligned} \mathrm{e}^{-\lambda\tau_i} \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i)\mathrm{e}^{\lambda(l-\epsilon_i)h} &= \mathrm{e}^{\lambda(-\tau_i-\epsilon_i h+s_+h)} \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i)\mathrm{e}^{-\lambda(s_+-l)h} \\ &= z_1 \sum_{l=-s_-}^{s_+} \psi_l(\epsilon_i) z_2^{s_+-l} \\ &= z_1 z_3, \end{aligned}$$

where $|z_1| \leq 1$ because $\Re(\lambda) \geq 0$ and due to (4.5), and where $|z_2| \leq 1$ because $\Re(\lambda) \geq 0$. Hence, by Lemma 4.1, $|z_3| \leq 1$. As a consequence, the above term can be replaced by an $\mathrm{e}^{-\lambda_i}$ with $\lambda_i \in \mathbb{C}^+$ which proves the lemma. ∎

The stability of the LMS method is related to the stability of Σ in the following way.

THEOREM 4.3 (dependence of the stability of the LMS-approximation on h and $\vec{\tau}$). *Suppose the LMS method is irreducible, consistent,* $\mathrm{LMS}(\mathbb{C}^+) \cap \mathrm{LMS}(\mathbb{C}_0^-) = \emptyset$, *and Nordsieck interpolation is used with* $s_- \leq s_+ \leq s_- + 2$. *Then the following statements hold:*

(i) *If* $h\Sigma(\mathbb{C}^+) \subset \mathrm{LMS}(\mathbb{C}_0^-)$ *for* $h \in (0, h^*]$, *then the zero solution of* (4.4) *is delay-independently stable* (*stable for all* $\vec{\tau} \geq 0$) *for* $h \in (0, h^*]$.

(ii) *If* $h\Sigma(\mathbb{C}^+) \subset \mathrm{LMS}(\mathbb{C}_0^+)$ *for* $h \in (0, h^*]$, *then the zero solution of* (4.4) *is delay-independently unstable* (*unstable for all* $\vec{\tau} \geq 0$) *for* $h \in (0, h^*]$.

(iii) *If the characteristic equation of* (1.1) *has a root* $\lambda \in \mathbb{C}$ *for some (fixed)* $\vec{\tau} \geq 0$ *with multiplicity* ν, *then there exists an* $h^* > 0$ *such that* (4.12) *has exactly* ν *roots* $\lambda_{h,i}$, $i = 1, \ldots, \nu$ *(taking into account multiplicities) with* $\max_{1\leq i\leq\nu} |\lambda - \lambda_{h,i}| = \mathcal{O}(h^{\frac{1}{\nu}\min\{p, s_-+s_++1\}})$ *for* $h \in (0, h^*]$ *and with* p *the order of the LMS method.*

Proof.

(i) By Lemma 4.2 we know that the mapping of the closed RHP under the right-hand side of the characteristic equation (4.13) maps into $\Sigma(\mathbb{C}^+)$, which, by assumption, maps into $\frac{1}{h}\mathrm{LMS}(\mathbb{C}_0^-)$. Hence, (4.13) can have no unstable roots because these are mapped into disjunct regions under the left- and right-hand sides of (4.13). This implies the asymptotic stability of the zero solution of the LMS method for $h \in (0, h^*]$ and for all $\vec{\tau} \geq 0$.

(ii) Formally rewrite (4.13) as

$$\lambda = \frac{1}{h}\mathrm{LMS}^{-1}\left(h\sigma_{h,l}\left(A_0 + \sum_{i=1}^{m} A_i \mathrm{e}^{-\lambda\tau_i} \sum_{q=-s_-}^{s_+} \psi_q(\epsilon_i)\mathrm{e}^{\lambda(q-\epsilon_i)h}\right)\right),$$

where $\sigma_{h,l}$ is similarly defined as in Theorem 3.4, part (ii). Now choose $0 < h < \hat{h} < h^*$ small enough such that the inverse LMS^{-1} uniquely exists for all arguments mapped onto by $h\sigma_{h,l}(\cdot)$ from $\lambda \in \mathbb{C}_0^+$. By the same reasoning as in the proof of Theorem 3.4, part (ii), this mapping has a fixed point $\lambda \in \mathbb{C}_0^+$ which is an unstable solution of (4.13). This root moves continuously in the function of h. For $h \in (0, h^*]$, (4.13) can have no pure imaginary

solutions λ (because $\Sigma(\{\mathrm{i}\xi \,|\, \xi \in \mathbb{R}\}) \subset \frac{1}{h}\mathrm{LMS}(\mathbb{C}_0^+)$). Hence for $h \in (0, h^*]$ this root cannot change sign and the zero solution of the LMS method is unstable for $h \in (0, h^*]$.

(iii) Denote the characteristic equation (4.9) after the substitution (4.11) by $P_h(\lambda) = 0$.

First, observe that

$$\sum_{l=-s_-}^{s_+} \psi_l(\epsilon)\mathrm{e}^{\lambda l h} = e^{\lambda\epsilon h} + \mathcal{O}(h^{s_-+s_++1})$$

uniformly in $\epsilon \in [0,1]$. We fix $\vec{\tau}$ and expand $\exp(\lambda h)$ into a Taylor series to obtain that

$$\begin{aligned} P_h(\lambda) \;=\; & \det\left(\lambda\left(\alpha_0 + \sum_{q=0}^{p}\sum_{j=1}^{k}\alpha_j\frac{(\lambda j h)^q}{q!}\right)I \right. \\ & - h\left(\beta_0 + \sum_{q=0}^{p-1}\sum_{j=1}^{k}\beta_j(\lambda j h)^q\right) \\ & \left.\cdot\left(A_0 + \sum_{i=1}^{m}A_i\mathrm{e}^{-\lambda\tau_i} + \mathcal{O}(h^{s_-+s_++1})\right) + \mathcal{O}(h^{p+1})\right). \end{aligned}$$

Using the order conditions (4.8), the latter simplifies to

$$\begin{aligned} P_h(\lambda) = \det & \left(h\left(\lambda I - \left(A_0 + \sum_{i=1}A_i\mathrm{e}^{-\lambda\tau_i} + \mathcal{O}(h^{s_-+s_++1})\right)\right)\right. \\ & \left.\cdot\left(\beta_0 + \sum_{q=0}^{p-1}\frac{(h\lambda)^q}{q!}\sum_{j=1}^{k}\beta_j j^q\right) + \mathcal{O}(h^{p+1})\right). \end{aligned}$$

From this it follows that

$$\frac{1}{h}P_h(\lambda) = P(\lambda)Q_h(\lambda) + \mathcal{O}(h^{\min\{p,s_-+s_++1\}}), \tag{4.14}$$

where $P(\lambda)$ denotes the characteristic equation (1.2) and

$$\lim_{h\to 0} Q_h(\lambda) = \left(\sum_{j=0}^{k}\beta_j\right)^n \neq 0, \tag{4.15}$$

where the latter inequality follows from the irreducibility (4.10) and the consistency (4.8). Moreover, it is clear that (4.14) holds uniform on bounded regions of λ in the complex plane.

The following statements follow the lines of proof of Hurwitz's theorem [2, section VII.2.5]. First, note that $P(\lambda)$ is an analytic function and that $P_h(\lambda)$ and $Q_h(\lambda)$ are analytic in bounded regions of the complex plane when h is small enough. Since λ^* is an isolated zero of $P(\lambda)$ of multiplicity ν, there exists an $R > 0$ and a $K_0 > 0$ such that

$$\inf_{\theta\in[0,2\pi]}\left|\left(\sum_{j=0}^{k}\beta_j\right)^n P(\lambda^* + r\mathrm{e}^{\mathrm{i}\theta})\right| > K_0 r^\nu \text{ for } r \in (0, R].$$

From (4.15), there exists an $h_0^* > 0$ and $0 < K_1 < 1$ such that

$$\inf_{|\lambda-\lambda^*|\le R} |Q_h(\lambda)| > K_1 \left| \sum_{j=0}^{k} \beta_j \right|^n \quad \text{for } h \in (0, h_0^*].$$

Hence,

$$\inf_{\theta\in[0,2\pi]} \left| Q_h(\lambda^* + re^{i\theta}) P(\lambda^* + re^{i\theta}) \right| > K_0 K_1 r^\nu \text{ for } r \in (0, R],\ h \in (0, h_0^*].$$

From (4.14), there exists a $0 < h_1^* < h_0^*$ and a $K_2 > 0$ such that (using $p^* = \min\{p, s_- + s_+ + 1\}$)

$$\sup_{|\lambda-\lambda^*|\le R} \left| \frac{1}{h} P_h(\lambda) - Q_h(\lambda) P(\lambda) \right| < K_2 h^{p^*} \text{ for } h \in (0, h_1^*].$$

Set $\kappa := (2\frac{K_2}{K_0 K_1})^{1/\nu}$; then, for each fixed $h < h^* := \min\{h_1^*, (R/\kappa)^{\nu/p^*}\}$ and corresponding $r = \kappa h^{p^*/\nu}$, we obtain

$$\begin{aligned} &\left| \frac{1}{h} P_h(\lambda^* + re^{i\theta}) - Q_h(\lambda^* + re^{i\theta}) P(\lambda^* + re^{i\theta}) \right| \\ &\qquad < K_2 h^{p^*} = K_2 \left(\frac{r}{\kappa}\right)^\nu = \frac{1}{2} K_0 K_1 r^\nu \\ &\qquad < \left| Q_h(\lambda^* + re^{i\theta}) P(\lambda^* + re^{i\theta}) \right| \end{aligned}$$

for $\theta \in [0, 2\pi]$. Hence, by Rouché's theorem [2, section V.3.8], $Q_h(\lambda)P(\lambda)$ and $P_h(\lambda)$, or, equivalently, $P(\lambda)$ and $\frac{1}{h}P_h(\lambda)$, have the same number of roots (counting multiplicities) in a circle of radius $r = \kappa h^{p^*/\nu}$ around λ^* for all $h \in (0, h^*]$. ∎

The above theorem provides conditions under which a given LMS method captures the stability properties of Σ, which, in turn, are related to the stability properties of the DDE. In the next section we exploit these results to obtain a steplength heuristic for h.

4.2. An LMS steplength heuristic. By comparing the first two parts of Theorems 3.4 and 4.3, it is clear that the stability region of the LMS method scaled with $1/h$ should mimic the LHP when compared to the regions $\Sigma(\mathbb{C}^+)$ in order to obtain a good correspondence between the stability of the LMS scheme and the original DDE (see also Figure 4.2 (right)).

For this reason we define a *safety radius* $\rho_{\mathrm{LMS},\epsilon}$ of the LMS stability region as follows:

$$\rho_{\mathrm{LMS},\epsilon} := \min\{\rho^-_{\mathrm{LMS},\epsilon}, \rho^+_{\mathrm{LMS},\epsilon}\}$$

with

$$\rho^-_{\mathrm{LMS},\epsilon} := \sup\{\rho > \epsilon \mid |\mathrm{LMS}(\lambda)| < \rho \text{ and } \Re(\mathrm{LMS}(\lambda)) < -\epsilon \Rightarrow \Re(\lambda) < 0\} \tag{4.16}$$

and

$$\rho^+_{\mathrm{LMS},\epsilon} := \sup\{\rho > \epsilon \mid |\mathrm{LMS}(\lambda)| < \rho \text{ and } \Re(\mathrm{LMS}(\lambda)) > \epsilon \Rightarrow \Re(\lambda) > 0\}, \tag{4.17}$$

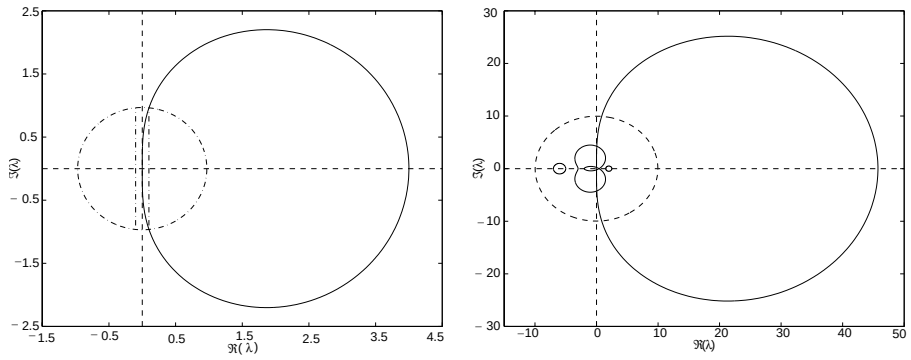


FIG. 4.2. *Left: Circle with radius $\rho_{\mathrm{LMS},0.1}$ for the stability region (—) of the second order BDF method. Right: Illustration of Heuristic* 4.2 *using $\rho_{\mathrm{LMS},0.1}$ shown left, $r = 0$, and the matrices* (3.5). *The $\Sigma(\mathbb{C}^+)$ regions of* (3.5) *are inside a circle with radius $\|A_0\| + \|A_1\|$ (−−). The stability region of the LMS method (−) is scaled with $1/h$, h obtained from* (4.21).

where we assume ϵ small enough such that the sets used in (4.16) and (4.17) are nonempty. In other words, the safety radius is the size of the circle in which the stability region of the LMS method mimics the LHP up to some accuracy ϵ (i.e., it has the correct stability except for a region of size 2ϵ around the imaginary axis; cf. Figure 4.2 (left)).

We can now use the safety radius $\rho_{\mathrm{LMS},\epsilon}$ and the bound on $\Sigma(\mathbb{C}^+)$ (see (3.4)) to obtain the following heuristic.

HEURISTIC 4.1. *The steplength h chosen as*

$$h = 0.9 \frac{\rho_{\mathrm{LMS},\epsilon}}{\sum_{i=0}^{m} \|A_i\|} \tag{4.18}$$

can be used to obtain an LMS scheme which approximates the (delay-independent) stability of the DDE up to some accuracy ϵ.

Indeed, if $\Sigma(\mathbb{C}^+)$ maps into $\Re(\lambda) < -\epsilon/h$, then there is delay-independent stability for the DDE, and this property is recovered by the LMS method (we assume ϵ is small enough such that $\epsilon/h \ll 1$). Similarly, if $\Sigma(\mathbb{C}^+)$ maps into $\Re(\lambda) > \epsilon/h$, then there is delay-independent instability for the DDE, and this property is recovered by the LMS method.

By virtue of the third parts of Theorems 3.4 and 4.3, we know that Hopf bifurcations as a function of the delay(s) (and hence delay-dependent stability) are associated with the crossings of the boundary of $\Sigma(\mathbb{C}^+)$ with the imaginary axis. These crossings are captured well when the boundary of the stability region of the LMS method mimics the imaginary axis when compared to $\Sigma(\mathbb{C}^+)$. This is the heuristic reason why the above steplength choice works well to capture the (more important) property of delay-dependent stability.

The previous analysis concentrates on recovering stability and hence on roots of the characteristic equations in their relation to the imaginary axis. The characteristic equation of (1.1) has a root $\lambda = r + \mathrm{i}\xi$ if and only if the system of DDEs,

$$\dot{x}(t) = (A_0 - rI)x(t) + \sum_{i=1}^{m} (A_i \mathrm{e}^{-r\tau_i})x(t - \tau_i), \tag{4.19}$$

has a root $\lambda = \mathrm{i}\xi$. Hence $\Sigma_r(\mathbb{C}^+)$ can be defined as before, in terms of the system matrices,

$$A_0^r = A_0 - rI \text{ and } A_i^r = A_i \mathrm{e}^{-r\tau_i},\ i = 1, \ldots, m, \tag{4.20}$$

to study the *r-stability* of the continuous problem and its dependency on the delay(s). This leads to the following adaptation of Heuristic 4.1.

HEURISTIC 4.2. *The heuristic choice of the steplength h,*

$$h = 0.9 \frac{\rho_{\mathrm{LMS},\epsilon}}{\|A_0\| + |r| + \sum_{i=1}^m \|A_i\| \mathrm{e}^{-r\tau_i}}, \tag{4.21}$$

can be used to approximate the roots with real parts greater than $r < 0$, $\Re(\lambda) \geq r$.

This heuristic choice of h is implemented in the package DDE-BIFTOOL [3], where it is applied to the original matrices $A_0, \ldots, A_m$ without the shift (4.20). The latter is done to avoid deterioration in the approximation to the most interesting part of the spectrum (near the imaginary axis).

Numerical results (see further) indicate that Heuristics 4.1 and 4.2 are quite effective.

5. Numerical results. In this section we illustrate our findings with numerical results.

In order to compute the characteristic roots of the LMS method (4.12) we compute the eigenvalues μ of the matrix M that maps $[y_{s-L}\ y_{s-L+1}\ \cdots\ y_{s+k-1}]$ onto $[y_{s-L+1}\ y_{s-L+2}\ \cdots\ y_{s+k}]$ using the LMS method for y_{s+k} and a shift for all other values (and where $L = \max_i L_i + s_-$). Then, we use the relation (4.11) to obtain

$$\Re(\lambda) = \frac{1}{h} \log(|\mu|)$$

and

$$\Im(\lambda) = \frac{1}{h} \arcsin\left(\frac{\Im(\mu)}{|\mu|}\right) \bmod \frac{\pi}{h}.$$

In our tests we use the explicit Adams–Bashforth (AB), implicit AM, and implicit BDF LMS methods with different number of steps k (see, e.g., [7, section III.1.1-2,6]).

Consider the well-known delayed logistic equation,

$$\dot{x}(t) = (\alpha - x(t-1))x(t), \tag{5.1}$$

cf., e.g., [18, section I.5.1] and [15]. The rightmost roots of the characteristic equation of the steady state solution $x^* = \alpha$ of (5.1) for $\alpha = 2$ are depicted in Figure 5.1 (left). The convergence towards the rightmost root by the corresponding root of the LMS method is shown in Figure 5.1 (right) for varying h. Here, the steplength h was chosen as an integer fraction of the delay, $h = 1/L$. Table 5.1 gives the numerical approximation of the orders of convergence based on a least squares approximation of the results for $L = 10, \ldots, 100$ using the AB, AM, and BDF methods with $k = 2, 3, 4, 5$ steps. The results for the AM method are shown in Figure 5.1 (right). The $\mathcal{O}(h^p)$ convergence, $p = k + 1$ for AM, $p = k$ for AB and BDF methods [7, Table III.2.1], is clearly apparent.

The following system of DDEs, taken from [24], models two coupled neurons with time delayed connections:

$$\begin{cases} \dot{x_1}(t) = -\kappa x_1(t) + \beta \tanh(x_1(t-\tau_s)) + a_{12} \tanh(x_2(t-\tau_2)), \\ \dot{x_2}(t) = -\kappa x_2(t) + \beta \tanh(x_2(t-\tau_s)) + a_{21} \tanh(x_1(t-\tau_1)). \end{cases} \tag{5.2}$$

We set $\kappa = 0.5$, $\beta = -1$, $a_{12} = 1$, $a_{21} = 2.34$, $\tau_1 = \tau_2 = 0.2$, and $\tau_s = 1.57$ and investigate the steady state solution $(x_1^*, x_2^*) = (0, 0)$. The convergence towards a rightmost root by the corresponding root of the LMS method is shown in Figure 5.2 (right) for varying h. Here, the steplength h is allowed to vary continuously, and appropriate interpolation is used for the past terms.

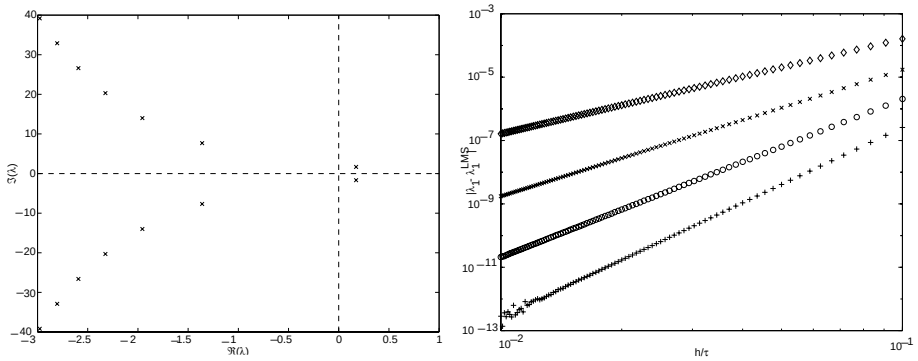


FIG. 5.1. *Left: The rightmost roots of the characteristic equation* (5.1) *for* $\alpha = 2$. *Right: Convergence of the corresponding LMS root to the rightmost root of* (5.1) *for* $\alpha = 2$, *for varying* $h = 1/L$, $L = 10, \ldots, 100$. *Here, AM methods were used with* $k = 2$ ($\diamond$), $k = 3$ ($\times$), $k = 4$ ($\circ$), *and* $k = 5$ ($+$).

TABLE 5.1

Numerically observed orders of convergence while approximating the rightmost root of the steady state $x^* = \alpha$ *of* (5.1) *at* $\alpha = 2$ *using LMS methods with* $k = 2, 3, 4, 5$ *steps.*

k	AB	AM	BDF
2	1.994	2.998	1.994
3	2.993	3.995	2.993
4	3.990	4.993	3.990
5	4.987	6.022	4.987

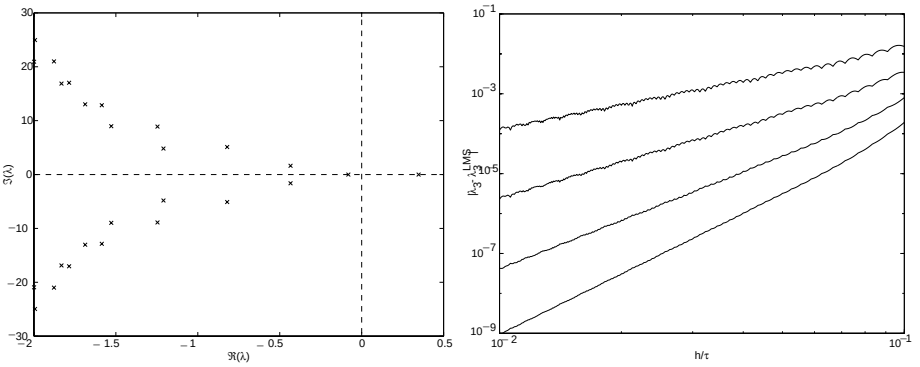


FIG. 5.2. *Left: The rightmost roots of the characteristic equation of the zero steady state solution of* (5.2) *for* $(\kappa, \beta, a_{12}, a_{21}, \tau_1, \tau_2, \tau_s) = (0.5, -1, 1, 2.34, 0.2, 0.2, 1.57)$. *Right: Convergence of the corresponding LMS root to the third rightmost root shown left for varying* h, $h/\tau \in [0.01, 0.1]$. *Here, BDF methods were used with* $k = 2, 3, 4, 5$ *from top to bottom and corresponding Nordsieck interpolation with* $s_- = \lfloor \frac{k-1}{2} \rfloor$ *and* $s_+ = \lceil \frac{k-1}{2} \rceil$.

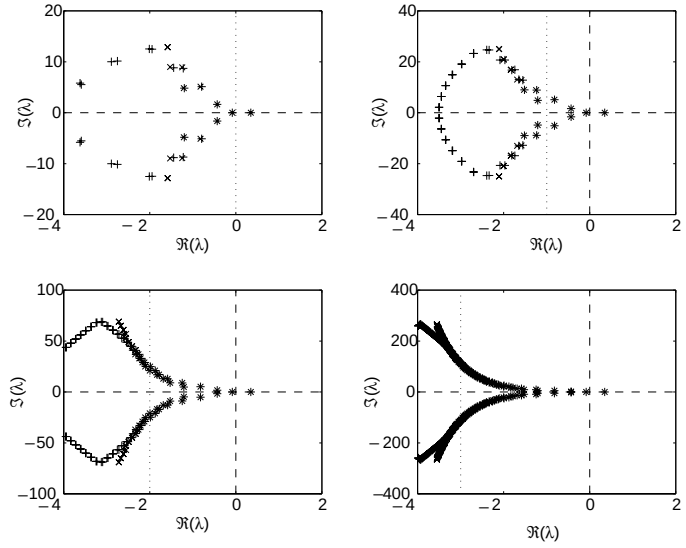


FIG. 5.3. *Illustration of Heuristic* 4.2 *for the zero steady state solution of* (5.2) *with* $(\kappa, \beta, a_{12}, a_{21}, \tau_1, \tau_2, \tau_s) = (0.5, -1, 1, 2.34, 0.2, 0.2, 1.57)$ *and varying* r. *Approximations* (+) *of roots obtained from the BDF method with* $k = 4$, $s_- = 1$, $s_+ = 2$, *and* h *from Table* 5.3 *versus their corrections* (×). *Approximations and corrections start to differ drastically only for* $\Re(\lambda) < r$ *(indicated by the dotted line).*

The scalar DDE,

$$\dot{x}(t) = 2x(t) - \mathrm{e}x(t-1), \tag{5.3}$$

with $\mathrm{e} = \exp(1)$ has a double characteristic root at $\lambda = 1$. Table 5.2 gives the numerical approximation of the orders of convergence towards this root based on a least squares approximation of the results for $h = 1/L$, $L = 10, \ldots, 100$ using the AB, AM, and BDF methods with $k = 2, 3, 4, 5$ steps. The $\mathcal{O}(|h|^{p/\nu})$ convergence, $\nu = 2$, is clearly apparent.

Finally, Table 5.3 and Figure 5.3 illustrate the results, respectively, the effectiveness, of Heuristic 4.2 for different values of r.

TABLE 5.2

Numerically observed orders of convergence while approximating the double characteristic root $\lambda = 1$ *of* (5.3) *using LMS methods with* $k = 2, 3, 4, 5$ *steps.*

k	AB	AM	BDF
2	0.99	1.50	1.03
3	1.48	1.99	1.49
4	1.98	2.48	1.98
5	2.47	2.94	2.47

TABLE 5.3

Values h *of the steplength Heuristic* 4.2 *for system* (5.2) *and varying* r *using the BDF method with* $k = 4$ *and* $\rho_{\mathrm{LMS}}, 0.01 \approx 0.57$. *Roots of the resulting LMS-approximations are shown in* 5.3.

r	0	-1	-2	-3
h	7.5e-02	3.8e-02	1.4e-02	3.8e-03

6. On stiffness in DDEs. Stiffness in the numerical solution of initial-value problems has been used with quite different meanings by a number of authors. For an overview in the context of ordinary differential equations, see [25]. The situation for DDEs is even less clear. In [1] Dahlquist is quoted as follows:

> *A "stiff" problem is characterized by the property that there are processes in the physical system, described by a system of ordinary differential equations, with significantly different time scales ....*

When applying this definition to DDEs one might conclude, observing the existence of sequences of characteristic roots with decreasing real parts (see Figure 5.3), that all DDEs are stiff. This is contradicted by the fact that almost all existing software for simulation of DDEs is based on explicit Runge–Kutta methods. The latter gives rise to a more practical observation of stiffness. We quote from Hairer and Wanner [8] as follows:

> *Stiff equations are problems for which explicit methods don't work.*

We now illustrate how visualization of the regions $\Sigma(\mathbb{C}^+)$ allows us to distinguish between stiff and nonstiff problems in the latter sense.

In [22] a model is introduced to describe the flow of a viscoelastic fluid (i.e., a fluid with a fading memory). This models consists of a partial differential equation with an infinite, distributed delay term. The classical method used to study this integrodifferential equation is the conversion into a pair of coupled partial differential equations. In [20], however, the distributed delay is replaced by two point delays. Here, we study a one-delay version,

$$\frac{\partial u}{\partial t} = \frac{1-\delta}{\nu}(u_{xx}(x,t) + u_{xx}(x,t-\tau)) + \delta u_{xx}(x,t) + Ru(x,t) - u^3(x,t), \tag{6.1}$$

on $x \in [0,\pi]$ with boundary conditions $u(0,t) = u(\pi,t) = 0$. $u(x,t)$ is the velocity of the fluid and ν, δ, τ, and R are parameters of the problem.

We use a classical second order central difference scheme in space to approximate (6.1) by a system of n DDEs,

$$\begin{aligned}\frac{\mathrm{d}u_i}{\mathrm{d}t} &= \left(\frac{1-\delta}{\nu} + \delta\right)\left(\frac{u_{i-1}(t) - 2u_i(t) + u_{i+1}(t)}{\Delta x^2}\right) \\ &\quad + \frac{1-\delta}{\nu}\left(\frac{u_{i-1}(t-\tau) - 2u_i(t-\tau) + u_{i+1}(t-\tau)}{\Delta x^2}\right) + Ru_i(t) - u_i^3(t),\end{aligned} \tag{6.2}$$

$i = 1, \ldots, n,$

where $u_0(t) \equiv u_{n+1}(t) \equiv 0$, $\Delta x = \pi/(n+1)$. We fix $\nu = 2$, $\delta = 0.1$, $R = 0.08$, and $\tau = 3$.

The regions $\Sigma(\mathbb{C}^+)$ of the linearization of (6.2) around its zero steady state solution are shown in Figure 6.1 (left) for $n = 8$. The figure indicates that this example is in fact mildly stiff. Indeed, the regions $\Sigma(\mathbb{C}^+)$ are such that an implicit method with unbounded stability region can capture the stability for a smaller h more efficiently than an explicit method can. This is illustrated in Figure 6.2 using the scaled stability regions of the BDF, respectively, AM, methods using, for both methods, $k = 2$ and $h = 0.3$. Both methods capture the stability of the regions $\Sigma(\mathbb{C}^+)$ near the origin, but the AM method fails to capture the "tail" of the regions $\Sigma(\mathbb{C}^+)$ at its leftmost end.

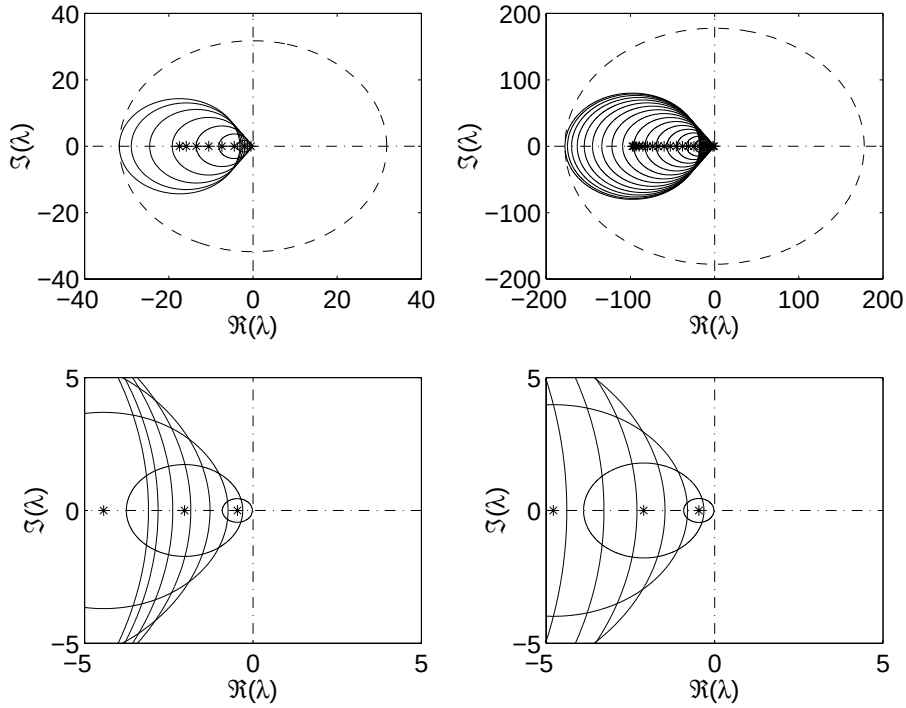


FIG. 6.1. *Stability regions of the zero steady state solution of* (6.2) *using* $n = 8$ *(left) and* $n = 20$ *(right). Full view (top) and a blow up near the origin (bottom).*

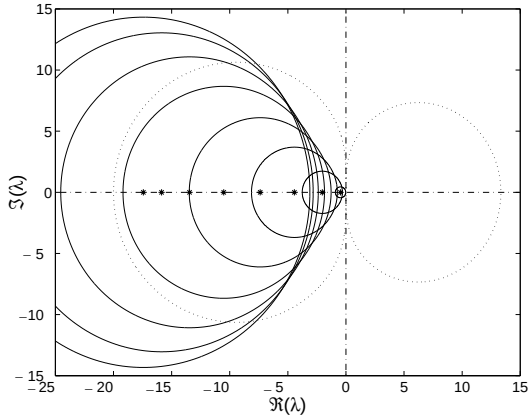


FIG. 6.2. *Stability regions of the zero steady state solution of* (6.2) *using* $n = 8$ *and stability regions of the BDF and AM methods for* $k = 2$ *scaled with* $1/h$ *using* $h = 0.3$.

The stability of the AM method for our case, $\tau_0 = 3$, is indeed wrong; see Figure 6.3 (upper right).

If system (6.2) is considered for a larger value of n, then stiffness increases and the situation is even more clear. Figure 6.1 shows the regions $\Sigma(\mathbb{C}^+)$ for a system of size $n = 20$. Taking a finer discretization enlarges the stability regions to the left, while the situation near the origin remains approximately the same. Hence, this effect is

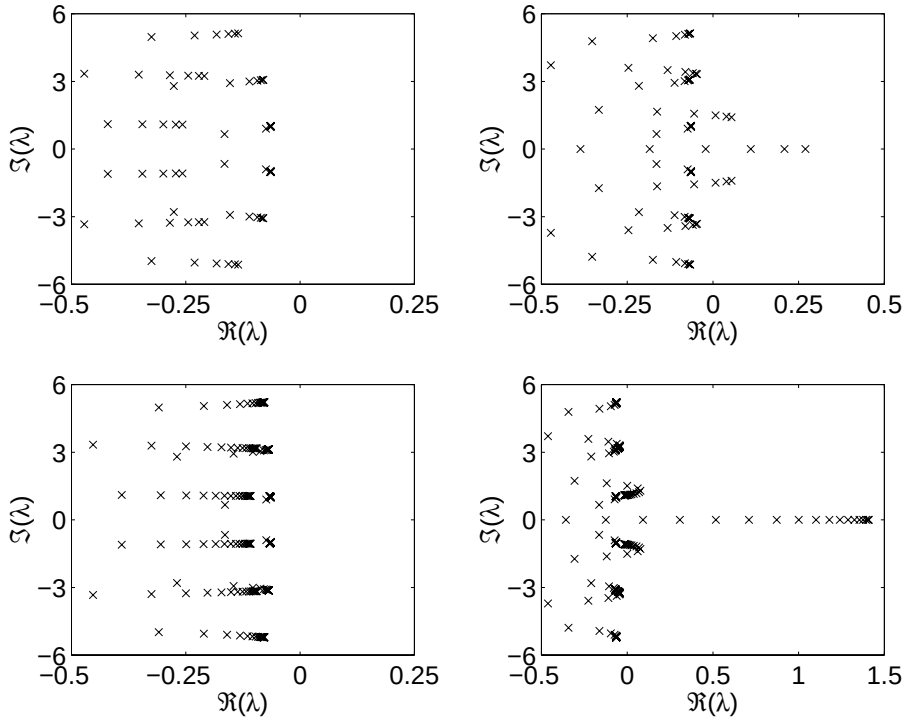


FIG. 6.3. *LMS approximations of the rightmost characteristic roots of the stable zero solution of* (6.2) *using* $h = 0.3$. *Left: BDF method,* $k = 2$. *Right: AM method,* $k = 2$. *Top:* $n = 8$. *Bottom:* $n = 20$. *Note that only some rightmost roots of the BDF method are close to actual correct roots of the system (not shown).*

not felt by the BDF method which produces (in contrast to the AM method) similar results for $n = 20$ using the same steplength $h = 0.3$; see Figure 6.3 (lower left).

As a last remark, we note that the above characterization of stiffness does not depend on the size of the delay. Hence one could say that a problem with the same system matrices but a larger delay is relatively more difficult (because the discretized state space grows with the size of the delay for fixed h) but is not more stiff.

7. Conclusion. DDEs are used to model systems with some form of memory or delayed feedback and arise in a growing number of applications.

The stability of a steady state solution of a DDE is governed by the roots of a characteristic equation which expresses a nonlinear, nonstandard eigenvalue problem. For stability and bifurcation analysis it is of interest to compute the stability determining, i.e., the rightmost characteristic, roots. Established numerical algorithms exist to compute selected eigenvalues of possibly very large matrices. Therefore, selected characteristic roots can be approximated by computing the eigenvalues of the map defined by a discrete numerical time integration approximation to the given DDE [4].

In this paper we investigate the correspondence between the characteristic roots of a DDE and the eigenvalues obtained from an LMS method approximation. In particular, we investigate under what conditions for the steplength the approximation retains certain delay-independent stability properties of the original system. We

concentrate on the recovery of both stability and instability and prove convergence orders of the approximate characteristic roots. This analysis allows us to obtain a steplength heuristic for the computation of the characteristic roots with real parts greater than a given constant, as used in the package DDE-BIFTOOL [3].

We illustrate the results using numerical experiments. In particular, the analysis of an example system of DDEs arising from a partial differential equation with delay discretized in space shows how the investigated stability issues allow us to interpret stiffness for DDEs.

Acknowledgment. The authors thank Professor J. Quaegebeur for helpful comments concerning Theorems 3.2 and 3.3.

REFERENCES

[1] C. T. H. Baker, C. A. H. Paul, and D. R. Willé, *Issues in the numerical solution of evolutionary delay differential equations,* Adv. Comput. Math., 3 (1995), pp. 171–196.

[2] J. B. Conway, *Functions of One Complex Variable*, 2nd Edition, Graduate Texts in Mathematics 11, Springer-Verlag, Berlin, 1978.

[3] K. Engelborghs, *DDE-BIFTOOL: A Matlab Package for Bifurcation Analysis of Delay Differential Equations,* Technical Report TW-305, Department of Computer Science, Katholique Universiteit Leuven, Leuven, Belgium, 2000. Available online at http://www.cs.kuleuven.ac.be/˜koen/delay/ddebiftool.shtml.

[4] K. Engelborghs and D. Roose, *Numerical computation of stability and detection of Hopf bifurcations of steady state solutions of delay differential equations,* Adv. Comput. Math., 10 (1999), pp. 271–289.

[5] N. J. Ford and V. Wulf, *The use of boundary locus plots in the identification of bifurcation points in numerical approximation of delay differential equations,* J. Comput. Appl. Math., 111 (1999), pp. 153–162.

[6] N. J. Ford and V. Wulf, *How do numerical methods perform for delay differential equations undergoing a Hopf bifurcation?* J. Comput. Appl. Math., 125 (2000), pp. 277–285.

[7] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations.* I. *Nonstiff Problems*, 2nd Edition, Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, 1993.

[8] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-Algebraic Problems*, 2nd Edition, Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, 1996.

[9] J. K. Hale, *Theory of Functional Differential Equations*, Appl. Math. Sci. 3, Springer-Verlag, Berlin, 1977.

[10] T. Hong-Jiong and K. Jiao-Xun, *The numerical stability of linear multistep methods for delay differential equations with many delays,* SIAM J. Numer. Anal., 33 (1996), pp. 883–889.

[11] G.-D. Hu, G.-D. Hu, and M.-Z. Liu, *Estimation of numerically stable step-size for neutral delay-differential equations via spectral radius,* J. Comput. Appl. Math., 78 (1997), pp. 311–316.

[12] G.-D. Hu, G.-D. Hu, and S. A. Meguid, *Stability of Runge-Kutta methods for delay differential systems with multiple delays,* IMA J. Numer. Anal., 19 (1999), pp. 349–356.

[13] V. Hutson and J. S. Pym, *Applications of functional analysis and operator theory*, Math. Sci. Engrg. 146, Academic Press, New York, 1980.

[14] K. J. in 't Hout, *On the stability of adaptations of Runge-Kutta methods to systems of delay differential equations,* Appl. Numer. Math., 22 (1996), pp. 237–250.

[15] K. J. in 't Hout and Ch. Lubich, *Periodic orbits of delay differential equations under discretization,* BIT, 38 (1998), pp. 72–91.

[16] K. J. in 't Hout, *The stability of θ-methods for systems of delay differential equations,* Ann. of Numer. Math., 1 (1994), pp. 323–334.

[17] A. Iserles and G. Strang, *The optimal accuracy of difference schemes,* Trans. Amer. Math. Soc., 277 (1983), pp. 779–803.

[18] V. Kolmanovskii and A. Myshkis, *Applied Theory of Functional Differential Equations*, Math. Appl. 85, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[19] T. Koto, *A stability property of A-stable natural Runge-Kutta methods for systems of delay differential equations,* BIT, 34 (1994), pp. 262–267.

[20] T. Luzyanina and D. Roose, *Numerical stability analysis and computation of Hopf bifurcation points for delay differential equations,* J. Comput. Appl. Math., 72 (1996), pp. 379–392.

[21] K. Meerbergen and D. Roose, *Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems,* IMA J. Numer. Anal., 16 (1996), pp. 297–346.

[22] W. E. Olmstead, S. H. Davis, S. Rosenblat, and W. L. Kath, *Bifurcation with memory,* SIAM J. Appl. Math., 46 (1986), pp. 171–188.

[23] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester; Halsted Press [John Wiley & Sons, Inc.], New York, 1992.

[24] L. P. Shayer and S. A. Campbell, *Stability, bifurcation and multistability in a system of two coupled neurons with multiple time delays,* SIAM J. Appl. Math., 61 (2000), pp. 673–700.

[25] M. N. Spijker, *Stiffness in numerical initial-value problems,* J. Comput. Appl. Math., 72 (1996), pp. 393–406.

[26] G. Strang, *Trigonometric polynomials and difference methods of maximum accuracy,* J. Math. Phys., 41 (1962), pp. 147–154.

[27] P. J. van der Houwen and B. P. Sommeijer, *Stability in linear multistep methods for pure delay equations,* J. Comput. Appl. Math., 10 (1984), pp. 55–63.

[28] D. S. Watanabe and M. G. Roth, *The stability of difference formulas for delay differential equations,* SIAM J. Numer. Anal., 22 (1985), pp. 132–145.

[29] V. Wulf and N. J. Ford, *Hopf bifurcations for numerical approximations to the delay logistic equation,* Internat. J. Appl. Sci. Comput., 6 (1999), pp. 167–172.

[30] V. Wulf and N. J. Ford, *Numerical Hopf bifurcation for a class of delay differential equations,* J. Comput. Appl. Math., 115 (2000), pp. 601–616.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 651–682

VARIATIONAL BARRIER METHOD OF ADAPTIVE GRID GENERATION IN HYPERBOLIC PROBLEMS OF GAS DYNAMICS*

BORIS N. AZARENOK†

Abstract. Application of the harmonic mapping using a variational approach to generate moving adaptive grids in the hyperbolic problems of gas dynamics is considered. Using the example of a three-point model of adaptation, the possibility to generate an unfolded mesh with strong grid lines condensing in the vicinity of discontinuities of the control/(monitor) function is demonstrated. The algorithm of redistributing the boundary nodes is suggested and consists of using constrained minimization of the discrete harmonic functional when constraints define the boundary of the domain. In real computations due to mesh adaptation it is possible to reduce the errors, caused by shock waves smearing over the cells, by many factors of ten. Modeling of the two-dimensional (2-D) supersonic gas flow in the channel has shown that the same accuracy on the adaptive grid with the same structure as the quasi-uniform mesh can be acheived while requiring less CPU memory by a factor of 25 and less running time by a factor of 50 to 60. Computational tests of the steady transonic and supersonic flow over an airfoil demonstrate the ability of the method to control mesh sizes across shocks.

Key words. moving adaptive grids, harmonic mapping, shock waves, constrained optimization

AMS subject classifications. 65M50, 76-08

PII. S0036142900382727

1. Introduction.

Moving adaptive grid technology has important applications in the problems of fluid dynamics. The essence of such an approach, referred to as r-refinement, is in adjusting redistribution of the grid nodes in such a manner to catch particularities in the solution of interest with fixed computer cost. The constructive way is that we try to position more grid nodes in the domains of sharp change in the solution being the regions of high gradients while the mesh retains the regular structure that makes the course of computation more simple. Examples of some r-refinement based methods can be found in Carey [10], Charakhch'yan and Ivanenko [12], Hawken, Gottlieb, and Hansen [21], Ivanenko [24], Jacquotte [26], Li, Tang, and Zhang [28], Tang and Tang [38], Thompson [39], and Zegeling [41]. In Liu, Ji, and Liao [30] the deformation method is suggested, where the grid velocities are determined by solving a scalar Poisson equation. An approach based on the moving mesh partial differential equations method has been considered in Cao, Huang, and Russell [8, 9] and Huang [22]. Here the Euler–Lagrange equations, both stationary and with the time dependent left part, to the variational functional are solved and mesh adaptation is performed using a class of monitor functions that are the symmetric positive definite matrices.

When constructing an adaptive moving grid the main difficulty is in maintaining its nondegeneracy. Probably, an approach should consist of extending the equidistribution principle, providing a one-to-one mapping of logic space onto the physical domain in a discrete approach, from the widely used one-dimensional (1-D) case to

*Received by the editors December 20, 2000; accepted for publication (in revised form) January 12, 2002; published electronically June 26, 2002. The paper appeared in a preliminary form in Adaptive moving grids in supersonic flow stimulation, numerical grid generation in computation field simulations, *Proceedings of the 7th International Conference on Numerical Grid Generation in Computational Field Simulations*, Whistler, British Columbia, Canada, 2000, ISGG. This work was supported by the Russian Fund of Fundamental Research (project code 02-01-00236).

http://www.siam.org/journals/sinum/40-2/38272.html

†Computing Center of Russian Academy of Sciences, Vavilov street 40, Moscow, 119991, Russia (azarenok@ccas.ru).

the two-dimensional (2-D) one. Some ways to define equidistribution-like methods in 2 dimensions can be found in Baines [6], Cao, Huang, and Russell [8, 9], Dwyer, Sanders, and Raiszadek [14], Huang [22], and Zegeling [41]. In Ivanenko [25] it was suggested to use a variational functional depending on derivatives of the functions sought and on variable coefficients, which are the elements of some symmetric positive defined matrix, and in a discrete approach the following variational principle is proved: a mapping, minimizing that functional, is one-to-one.

One of the mesh generation techniques is in using a harmonic mapping, first applied in Winslow [40], and in particular a variational approach. Here, when minimizing the Dirichlet (or harmonic) functional of smoothness, we ensure the grid lines are as smooth as possible; see Brackbill and Saltzman [7]. To include adaptivity in this process it has been suggested to write the Dirichlet functional on the surface of the graph of the control/(monitor) function, which is the solution of the basic problem or somehow connected with it, and to perform an adaptation by solving the Euler–Lagrange equations to the harmonic functional; see Liseikin [31, 32]. Application of such monitor surfaces has also been considered in Dwyer, Sanders, and Raiszadek [14], Eisman [15], and Spekreijse [36]. Note that Dvinsky [13] was the first to use the harmonic mapping for mesh adaptation where the gradient of the monitor function is utilized to perform grid lines clustering.

Numerical solution of the discrete Euler–Lagrange equations cannot always provide generation of the unfolded mesh even in domains without adaptation; see examples of grids to the backstep in Knupp and Steinberg [27] and Ivanenko [24]. It means, though, in a continuous approach that there exists a unique harmonic mapping, being one-to-one, and when discretizing the one-to-one property of the mapping can be lost. The discrete mapping can also be nonunique; see Garanzha and Kaporin [18].

To provide a one-to-one harmonic mapping at a discrete level Charakhch'yan and Ivanenko [11] have suggested a variational barrier method of grid generation in a physical domain without adaptation when the mapping is constructed by minimizing the harmonic functional. The functional is approximated in such a manner that there is an infinite barrier ensuring all grid cells to be convex quadrilaterals. This approach has been extended to adaptive grid generation when the harmonic functional is written on the surface of the control function; see Ivanenko [23, 24], Charakhch'yan and Ivanenko [12]. In Azarenok [2] and Azarenok and Ivanenko [4, 5] this approach has been applied in 2-D unsteady problems of gas dynamics when as a control function it used one of the flow parameters or the superposition of several parameters.

The purpose of the present work is to show some theoretical aspects and practical possibilities of using the variational barrier method of constructing structured adaptive grids in hyperbolic problems of gas dynamics with discontinuous solutions. In the 1-D case based on solving the nonlinear advection equation it is shown that the regularized discrete functional is convex and its minimum is attained on the mesh, which can be strongly condensed in the vicinity of discontinuities of the control function; meanwhile, this functional keeps the infinite barrier preventing the grid cells from collapsing. In the 2-D case it is shown that at strong grid lines condensing the infinite barrier disappears; nevertheless, when using the iterative procedure of mesh generation, we can guarantee the grid to be unfolded. The algorithm of redistributing the boundary nodes, consisting of using constrained minimization of the functional when constraints define the boundary of the domain, is suggested. Such an approach allows us to perform consistent redistribution of grid nodes inside the domain and on its boundary that increases the reliability of grid generation and the modeling of

the flow problem. Modeling of the 2-D supersonic gas flow in the channel has shown that the same accuracy on the adaptive grid with the same structure as on the quasi-uniform mesh can be achieved while requiring less CPU memory by a factor of 25 and less running time by a factor of 50 to 60. Computational tests of the steady transonic and supersonic flow over an airfoil demonstrate the possibility to control mesh sizes across shocks.

2. Problem formulation. The theory of harmonic mappings is useful for formulating a well-posed variational grid generation problem. The energy of a mapping $\phi : (M, g) \to (N, h)$ between two n-dimensional Riemannian manifolds M and N with metric tensors g_{ij} and h_{ij} is the function $e(\phi) : M \to \mathcal{R}(\geq 0)$, defined in some local coordinates ξ^i, μ^i as

$$e(\phi) = \frac{1}{2} g^{ij}(\xi) \frac{\partial \mu^k}{\partial \xi^i} \frac{\partial \mu^l}{\partial \xi^j} h_{kl}(\mu),$$

where g^{ij} is the inverse metric. The energy functional (or total energy) of the mapping ϕ is defined as (see Eells and Lemaire [16] and Dvinsky [13])

$$E(\phi) = \int_M e(\phi) d\xi, \qquad \text{where } d\xi = \sqrt{\det(g)} d\xi^1 \cdots \xi^n. \tag{2.1}$$

A smooth mapping $\phi : (M, g) \to (N, h)$ is harmonic if it is a critical point of the energy functional E.

Sampson [34] and Schoen and Yau [35] have shown that the harmonic mapping $M \to N$ when $\dim N = 2$ is always a homeomorphism (one-to-one) provided that the curvature of N is nonpositive and the boundary ∂N is convex (with respect to the metric h_{ij}). But it is sometimes not true when $\dim N > 2$ [17].

Now we turn to the problem formulation of the surface grid generation utilized for constructing the adapted mesh [12, 24, 32]. Let a simply connected domain $\Omega \subset \Re^2$ with a smooth boundary in plane x-y be given. Consider the surface S^{r2} of the graph of the control/(monitor) function $z = f(x, y)$; see Figure 1. The surface S^{r2} can be presented in Euclidean space $\mathcal{R}^3$ by a parametrization $\boldsymbol{r}(\xi, \eta) = (x(\xi, \eta), y(\xi, \eta), f[x(\xi, \eta), y(\xi, \eta)])$, where ξ, η are the local coordinates on S^{r2}. We seek a mapping of the unit square in parametric plane ξ-η onto the domain Ω, when a mapping of the boundary of a square onto the boundary of Ω is given, so that the mapping of the surface S^{r2} onto the parametric square is harmonic. The conditions of the above Sampson, Shoen and Yau's statement are obviously satisfied for the mapping of the physical domain Ω onto the parametric domain Ω_p with Euclidean metric $h_{ij} = \delta_{ij}$ (here δ_{ij} is the Kronecker symbol), which is the square in the parametric plane ξ-η. Hence, the nondegenerate harmonic coordinates ξ, η may be constructed on the surface S^{r2}. The coordinate lines ξ, η are then projected on the physical domain Ω and the result is an adaptive-harmonic grid.

To construct the harmonic mapping we use the variational formulation of the Dirichlet functional, which is the measure of closeness of the mapping to conformal. The task at hand is to minimize the Dirichlet functional written on the surface S^{r2}. This functional of smoothness can be expressed through the invariants $\mathcal{I}_1$, $\mathcal{I}_2$ of the orthogonal transforms of the covariant metric tensor g_{ij} [31, 32], namely,

$$I = \int_{S^{r2}} \frac{\mathcal{I}_1}{\mathcal{I}_2} dS^{r2}.$$

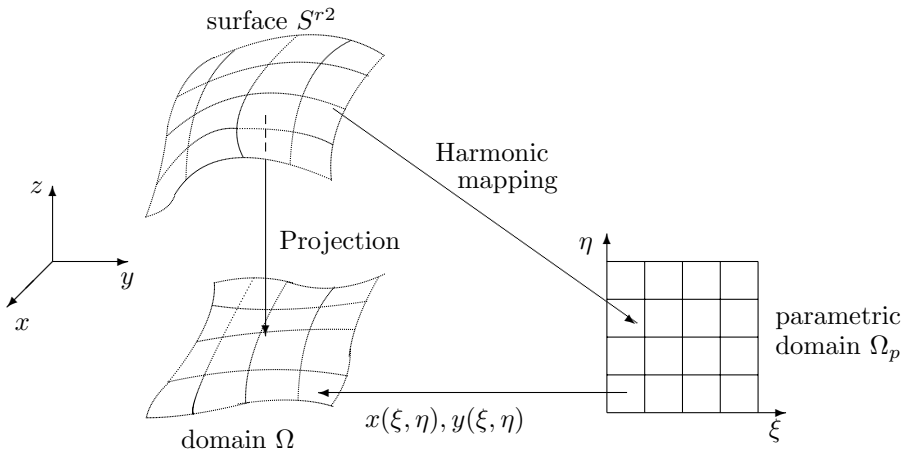


FIG. 1. *Adaptive grid generation.*

The invariant $\mathcal{I}_2 = \det(g_{ij}) = g_{11}g_{22} - (g_{12})^2$ and it equals the area squared of the cell on the surface S^{r2}; the invariant $\mathcal{I}_1 = g_{11} + g_{22}$ means the sum of the length squared of the cell sides. Noting that $dS^{r2} = \sqrt{g_{11}g_{22} - (g_{12})^2}d\xi d\eta$ we get

$$I = \int_0^1 \int_0^1 \frac{g_{11} + g_{22}}{\sqrt{g_{11}g_{22} - (g_{12})^2}} d\xi d\eta. \tag{2.2}$$

The components of the tensor g_{ij} are defined on the surface S^{r2} with the local coordinates ξ, η in Euclidean space $\mathcal{R}^3$ as follows:

$$g_{11} = \boldsymbol{r}_\xi^2 = x_\xi^2 + y_\xi^2 + f_\xi^2 = x_\xi^2 + y_\xi^2 + (f_x x_\xi + f_y y_\xi)^2,$$
$$g_{12} = (\boldsymbol{r}_\xi \cdot \boldsymbol{r}_\eta) = x_\xi x_\eta + y_\xi y_\eta + f_\xi f_\eta = x_\xi x_\eta + y_\xi y_\eta + (f_x x_\xi + f_y y_\xi)(f_x x_\eta + f_y y_\eta),$$
$$g_{22} = \boldsymbol{r}_\eta^2 = x_\eta^2 + y_\eta^2 + f_\eta^2 = x_\eta^2 + y_\eta^2 + (f_x x_\eta + f_y y_\eta)^2.$$

Substituting them in (2.2) we get the following form of the functional:

$$I = \int_0^1 \int_0^1 \frac{(x_\xi^2 + x_\eta^2)(1 + f_x^2) + (y_\xi^2 + y_\eta^2)(1 + f_y^2) + 2f_x f_y(x_\xi y_\xi + x_\eta y_\eta)}{(x_\xi y_\eta - x_\eta y_\xi)\sqrt{1 + f_x^2 + f_y^2}} d\xi d\eta. \tag{2.3}$$

This form of the harmonic functional has been derived in [12, 24] from the energy functional (2.1) (if multiplied by 2) written for the surface S^{r2}

$$E = \int_0^1 \int_0^1 Tr(g_{ij}^{-1})\sqrt{\det(g_{ij})} d\xi d\eta,$$

where $Tr(g_{ij}^{-1}) = g^{ii}$.

With the purpose of controlling the degree of coordinate lines condensing in the domains of high gradients, it is convenient to use $c_a f$ instead of the control function f, where c_a is a coefficient of adaptation [24] which can depend on variables x, y. Thus, we work with the control function multiplied by some coefficient c_a in order to increase or decrease adaptation.

In the 1-D case, to generate the inverse harmonic mapping of the graph of f onto the unit segment in parametric space ξ requires us to minimize the following

functional (see [31, 32]):

$$I = \int_0^1 \frac{1}{x_\xi \sqrt{1 + c_a^2 f_x^2}} d\xi. \tag{2.4}$$

Here at once we use $c_a f$ instead of f, and it means we seek the grid points arc-length equidistribution in the metric of the curve $c_a f$.

For the direct mapping of the unit segment in parametric space ξ to the curve $c_a f$, the Dirichlet functional has the form

$$I = \int_0^1 x_\xi^2 (1 + c_a^2 f_x^2) d\xi. \tag{2.5}$$

The Euler–Lagrange equations to the functionals (2.4) and (2.5) are similar, and

$$x_\xi \sqrt{1 + c_a^2 f_x^2} = \text{const}. \tag{2.6}$$

3. Approximation of a functional. We consider a piecewise bilinear mapping of the unit square $i \le \xi \le i+1$, $j \le \eta \le j+1$ in the parametric plane ξ-η onto a quadrilateral grid cell in plane x-y formed by nodes with coordinates $(x, y)_{i,j}$, $(x, y)_{i+1,j}$, $(x, y)_{i+1,j+1}$, $(x, y)_{i,j+1}$ numbered from 1 to 4 in an anticlockwise manner, where i, j are positive integers. The functional (2.3) is approximated in such a way that its minimum is attained on a grid of convex quadrilaterals, referred to as a convex grid [12, 24],

$$I^h = \sum_{i=1}^{i_{\max}} \sum_{k=1}^{4} \frac{1}{4} \left[F_k\right]_i, \tag{3.1}$$

where F_k is the integrand evaluated in the kth corner of the ith cell. If the set of convex meshes is not empty, the system of algebraic equations written to every interior node (here i is a global node number)

$$R_x = \frac{\partial I^h}{\partial x_i} = 0, \quad R_y = \frac{\partial I^h}{\partial y_i} = 0, \tag{3.2}$$

has at least one solution that is a convex mesh. To find it one should have an initial convex mesh and then use a method of unconstrained minimization [4, 12, 24]. It has been shown [11, 12, 24] when generating a curvilinear mesh without adaptation in an arbitrary simply connected domain Ω in plane x-y that the discrete functional (3.1) (if no adaptation, then in (2.3) we define $f_x = f_y = 0$) has an infinite barrier on the boundary of the set of convex grids; see Figure 2. This is caused by the condition of positiveness to the Jacobian of the mapping $J = x_\xi y_\eta - x_\eta y_\xi$ in (2.3). Should the vertexes of some cell be displaced so that the cell becomes nearly nonconvex (see Figure 2(b)), then one of four triangles, into which the cell is divided by its two diagonals, degenerates; its area, equal to $0.5J$, tends to zero and, therefore, $I^h \to \infty$. The infinite barrier holds in the case of adaptive grid generation as well as when the control function $f \in C^1(\Omega)$ since the values f_x, f_y under the square root in (2.3) are bounded everywhere in Ω. When starting from an initial convex mesh, due to the infinite barrier every step of the minimization procedure can be chosen so that the mesh always remains convex. However, if f is of the class of discontinuous functions, what we generally have in hyperbolic problems of gas dynamics, in the vicinity of a

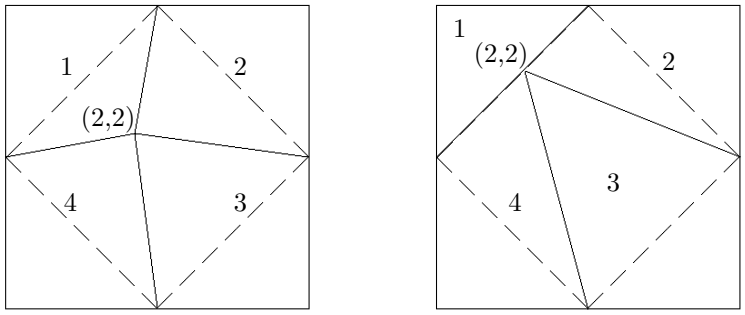


FIG. 2. *When minimizing the discrete functional I^h there is an infinite barrier preventing the cells from folding. The mesh shown consists of four cells. At minimization only the node $(2,2)$ moves. Dashed lines indicate the domain inside which the node $(2,2)$ shall be placed so that all four cells are convex* (a). *If this node approaches the dashed line* (b), *one* (*right*) *of two triangles, into which cell* 1 *is divided by the diagonal, degenerates, its area tends to zero, and, therefore, $I^h \to \infty$.*

discontinuity the derivatives f_x, f_y become unbounded, is that the the infinite barrier disappears and this causes some grid cells to fold and the modeling to break. In order to prevent grid lines overlapping, we use the procedure of regularization to the discrete functional described in section 5.

We make use of the rectangular rule to compute the 1-D functional (2.4)

$$I^h = \sum_{i=1}^{i_{\max}} \frac{\triangle\xi}{(x_\xi)_{i+1/2}\sqrt{1 + c_a^2 (f_x)_{i+1/2}^2}}, \tag{3.3}$$

where $i_{\max}$ is the number of spacings and the derivatives are computed via

$$(x_\xi)_{i+1/2} = (x_{i+1} - x_i)/\triangle\xi, \quad (f_x)_{i+1/2} = (f_{i+1} - f_i)/(x_{i+1} - x_i). \tag{3.4}$$

4. Three-point model of adaptation. In this section we demonstrate that when solving the Cauchy problem for the nonlinear advection equation under some conditions the values of the discrete function in the cells of the moving mesh remain invariable. Then the problem of constructing an adaptive mesh can be considered separately from the basic problem as if using an analytical control function. This allows us to analyze some properties of the discrete functional which also hold in the general case of computing a real flow with discontinuities.

Consider a 1-D adaptive mesh generation when solving the IVP to the nonlinear advection equation with discontinuous initial data when the shock moves from the left to the right:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \quad u(x,0) = \begin{cases} u_l & \text{if} \quad x < 0, \\ u_r & \text{if} \quad x \geq 0, \end{cases} \qquad u_l > u_r > 0. \tag{4.1}$$

We use the integral equation

$$\oint_C u\,dx - \frac{1}{2}u^2 dt = 0, \tag{4.2}$$

which in case of a smooth solution is equivalent to the above differential equation, and a discontinuous solution is governed by it as well. Here the contour C is the boundary of an arbitrary domain in plane x-t.

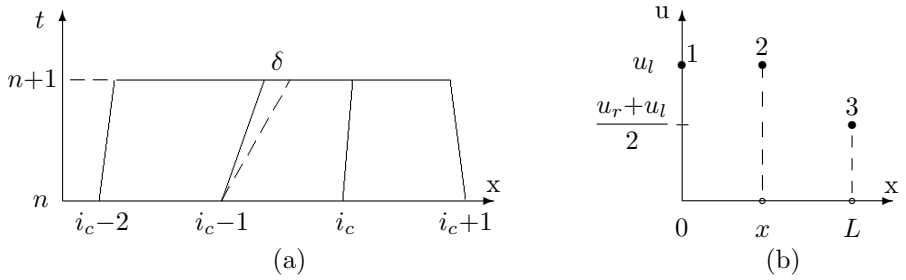


FIG. 3. *Moving grid* (a); i_c *is a center of the shock smeared. If the mesh moves with the speed of shock* $w = 0.5(u_r + u_l)$ *and the number of intervals* $i_{\max}$ *is even, then the updated value* $u^{i+1/2}$ *does not change. Three-point model of adaptation* (b); *in the coordinate system moving with the speed of shock* w, $x_1 = 0$ *and* $x_3 = L$ *are the coordinates of the fixed boundary nodes,* $x_2 = x$ *is a moving node.*

In plane x-t we introduce the moving grid (see Figure 3(a)) with spacings at the nth time level $h_{i+1/2} = x^n_{i+1} - x^n_i$ and at the $n+1$st level $h^{i+1/2} = x^{n+1}_{i+1} - x^{n+1}_i$ and time step $\triangle t = t^{n+1} - t^n$, where n, i are integers. Let at t^n the cell-average values of the discrete function $u_{i+1/2}$ be defined in the center of intervals. To update the values $u^{i+1/2}$ at time t^{n+1} we use the Godunov scheme on the moving grid [20, 19]. This scheme can be obtained if we integrate (4.2) along the contour C being the boundary of the computing cell

$$u^{i+1/2}h^{i+1/2} - u_{i+1/2}h_{i+1/2} - u_{i+1}h_{i+1} + h_i u_i + \frac{\triangle t}{2}\left[(u_{i+1})^2 - (u_i)^2\right] = 0, \tag{4.3}$$

where $h_i = x^{n+1}_i - x^n_i$. In order to determine a flux through the intercell boundary $[x^n_i, x^{n+1}_i]$ we need the value u_i on it, which is defined via the condition across the discontinuity

$$u_i = \begin{cases} u_{i-1/2} & \text{if } h_i/\triangle t < 0.5(u_{i-1/2} + u_{i+1/2}), \\ u_{i+1/2} & \text{otherwise.} \end{cases}$$

The above condition takes into account the inclinations of the straight-line characteristic $dx/dt = 0.5(u_{i-1/2} + u_{i+1/2})$ at the point x^n_i and the slanted boundary of the cell $[x^n_i, x^{n+1}_i]$ defined as $h_i/\triangle t$. If the ith node moves faster than perturbations from this node, then we set $u_i = u_{i+1/2}$, and if vice versa, then $u_i = u_{i-1/2}$.

From (4.3) we get the values at t^{n+1}

$$u^{i+1/2} = \frac{1}{h^{i+1/2}}\left\{u_{i+1/2}h_{i+1/2} - \frac{\triangle t}{2}\left[(u_{i+1})^2 - (u_i)^2\right] - u_i h_i + u_{i+1}h_{i+1}\right\}. \tag{4.4}$$

THEOREM 4.1. *Let, without adaptation, the grid nodes move with the speed of shock* $w = 0.5(u_r + u_l)$. *Suppose at time* t^n *the values* $u_{i+1/2} = u_l$ *if* $i < i_c$ *and* $u_{i+1/2} = u_r$ *if* $i \geq i_c$, *where* i_c *is a node that is the center of the shock smeared and the midmesh node, i.e.,* $i_c = 0.5i_{\max} + 1$, *where* $i_{\max}$ *is an even number of intervals. Then the updated values* $u^{i+1/2}$ *do not change.*

Proof. First, note that the problem is symmetrical about the i_cth node. This node moves with the velocity w. Consider the interval $(i_c - 1, i_c)$; see Figure 3(a). Let at time t^{n+1} the $i_c - 1$st node shift by δ from the position

$$x^{n+1}_{i_c-1} = x^n_{i_c-1} + w\triangle t,$$

where it would be if there were no adaptation. Denote $h_{i_c-1/2} = h$; then $h^{i_c-1/2} = h - \delta$. Observing that projections of the cell lateral edges onto x are $h_{i_c} = w\triangle t$, $h_{i_c-1} = w\triangle t + \delta$, and defining the intercell values from the condition across the discontinuity as $u_{i_c-1} = u_l$ and $u_{i_c} = u_r$, we get from (4.4) that

$$u^{i_c-1/2} = \frac{1}{h-\delta}\left[u_l h - \frac{\triangle t}{2}\left(u_r^2 - u_l^2\right) - (w\triangle t + \delta)u_l + u_r w\triangle t\right]$$
$$= \frac{1}{h-\delta}\left[u_l h - \frac{\triangle t}{2}\left(u_r + u_l\right)\left(u_r - u_l\right) + w\triangle t\left(u_r - u_l\right) - \delta u_l\right] = u_l.$$

A similar result holds if $u_{i_c} = u_l$. □

Consequently, when constructing the adaptive mesh we need not solve the IVP (4.1) and can merely set the values equal u_l in the cells to the left of the i_cth point and u_r to the right.

We shall construct the adaptive mesh minimizing the functional (3.3) and using u as a control function. We can simplify the model considering the left half of the mesh consisting of three points, i.e., when $i_c = 3$. Such an assumption does not change the mesh structure. The general case can be easily obtained from this three-point model. We pass into the new coordinate system moving with the velocity w so that $x_1 = 0$; see Figure 3(b). Then, when adapting, 1 and 3 are the fixed boundary nodes and $x_3 = L$, where $L = 2h$, h is a spacing of the initial uniform mesh, and coordinate x_2 is variable, referred to further as x. We also have $u_1 = u_2 = u_l$, $u_3 = (u_l + u_r)/2$.

5. Properties of the discrete functional. In this section we consider properties of the discrete functional in the 1-D and 2-D cases within the framework of the three-point model of adaptation. It will be shown that if the coefficient c_a is larger than some critical value, minimization of the functional leads the right cell of the two-cell grid to collapse. In order to get a convex functional in the 1-D case, i.e., to provide the unique solution of the minimization problem, we use a regularized functional. The improved functional holds an infinite barrier preventing the right cell from collapsing to any extent of grid nodes condensation. In the 2-D case, the meaning of the minimization procedure to the functional can be lost due to absence of a solution in the minimization problem. Nevertheless, we shall demonstrate the iterative procedure allows us to condense significantly the grid lines towards the discontinuity and guarantee the grid is unfolded.

First, consider the 1-D case. In assumptions of section 4 the approximation (3.3) to the functional on the two-cell grid reads (we set $\triangle\xi = 1$) as

$$I^h = \frac{1}{x} + \frac{1}{(L-x)\sqrt{1 + c_a^2\triangle u^2/(L-x)^2}}, \tag{5.1}$$

where $\triangle u = |u_3 - u_2|/2 = |u_r - u_l|/2$. To minimize the one-parametric functional I^h we apply the Newton method

$$x^{p+1} = x^p - \tau\frac{\partial I^h}{\partial x}\left[\frac{\partial^2 I^h}{\partial x^2}\right]^{-1}, \tag{5.2}$$

where the iterative parameter $\tau \le 1$, the first derivatives is

$$\frac{\partial I^h}{\partial x} = -\frac{1}{x^2} + \frac{1}{(L-x)^2\left[1 + c_a^2\triangle u^2/(L-x)^2\right]^{3/2}},$$

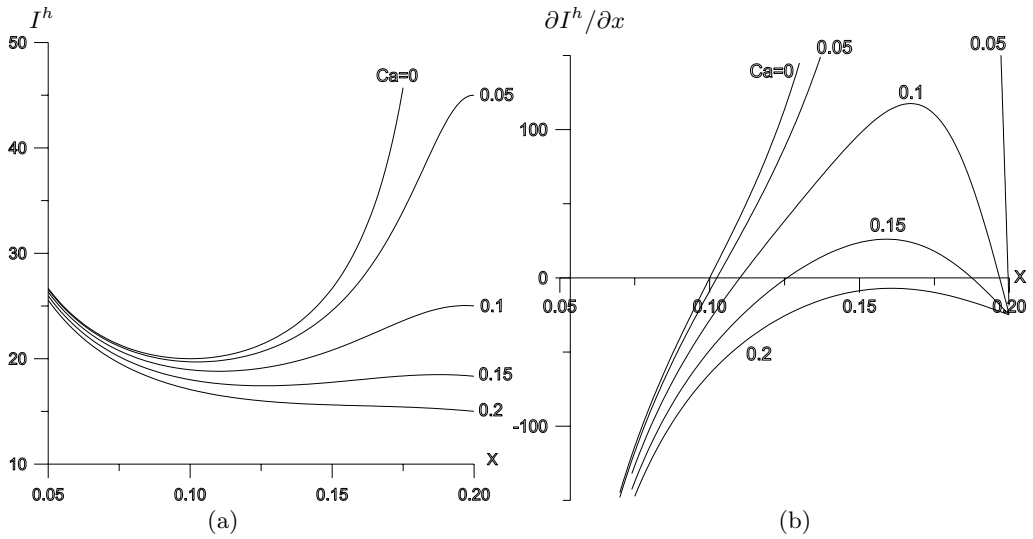


FIG. 4. *Dependence of distribution for the functional* I^h (a) *and* $\partial I^h/\partial x$ (b) *on* c_a *within the interval* $(0, L)$, *where* $L = 2h = 0.2$. *Here the part of the interval* $(0, 0.05)$ *is cut off since it is out of interest. At* $c_a > 0.185$ *there is no minimum of the functional* (*see the curve* $c_a = 0.2$), *and minimization of* I^h *causes the right cell to collapse.*

and the second derivative is derived from the first one.

In Figure 4 the distributions of I^h and $\partial I^h/\partial x$ are presented for several values of c_a at $u_l = 2, u_r = 1$, and initial uniform spacing $h = 0.1$. One can see at $c_a = 0$ the functional I^h has the minimum at $x = 0.1$ which corresponds to the point's equidistribution or a uniform mesh. When $c_a > 0$ the functional I^h loses convexity in the interval $(0, L)$ and besides the minimum there appears a maximum to the left of point 3. Therefore, the solution of the problem on finding its extremum becomes nonunique. When increasing c_a on one hand the value $x_{\min}$, where I^h reaches the minimum in $(0, L)$, shifts to the right corresponding to point 2 moving towards point 3 or grid clustering (see Figure 4(b)). On the other hand, $x_{\max}$, where I^h reaches the maximum in $(0, L)$, shifts to the left from point 3 causing grid rarefaction. In 2-D problems it can cause harsh displacements of the mesh nodes due to jumps of the solution from the minimum to maximum and vice versa during iterations, i.e., grid lines overlap and instability in the solution of the flow problem. Furthermore, for some critical value c_a^{crt}, both extrema merge at the point x^{crt}. To find c_a^{crt} it is required that we solve the system of two equations $\partial I^h/\partial x = 0$ and $\partial^2 I^h/\partial x^2 = 0$ about c_a^{crt} and x^{crt}. For the above parameters $c_a^{crt} \simeq 0.185$, $x^{crt} \simeq 0.159$. When $c_a > c_a^{crt}$ there is no extremum and minimization of I^h causes the right cell to collapse; see the curve $c_a = 0.2$ in Figure 4(b). Consequently, significant mesh clustering is impossible.

Let us write the discrete functional to the left half of the mesh including $0.5 i_{\max}$ intervals

$$I^h = \frac{(0.5 i_{\max} - 1)^2}{x} + \frac{1}{(L - x)\sqrt{1 + c_a^2 \triangle u^2/(L - x)^2}}; \tag{5.3}$$

here $L = 0.5 i_{\max} h$. The functional (5.3) is also one-parametric since all intervals except that on the right have the same length. Note the larger $i_{\max}$ is the smaller c_a^{crt} is. For example, at $i_{\max} = 100$ the value $c_a^{crt} = 0.126$, which reduces the possibility of condensing the mesh towards discontinuity.

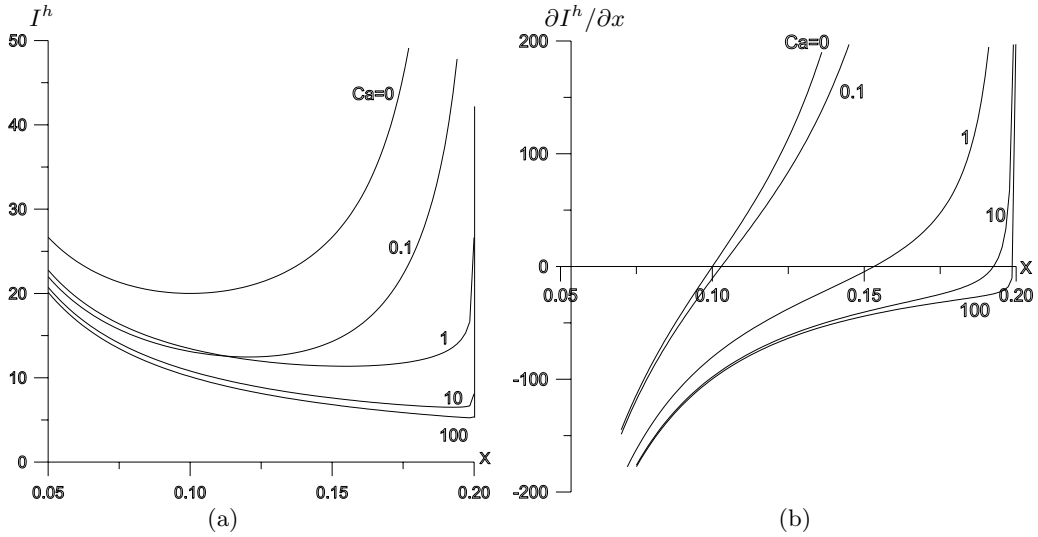


FIG. 5. *Dependence of distribution for* I_1^h (a) *and* $\partial I_1^h/\partial x$ (b) *on* c_a. *Functional* I_1^h *is convex within the interval* $(0, L)$ *for any* c_a *that guarantees existence of a unique solution of the minimization problem. The infinite barrier prevents the right cell from collapsing.*

In order to preserve convexity of the discrete functional we assume, when varying x, the derivative of the control function f_x in (3.3) (i.e., u_x in our case) remains fixed (invariable metric) as it was used in [24]. Assuming in (5.1) the derivative $(u_x)_{i+1/2}$ does not depend on x, we obtain the derivative of a new functional I_1^h:

$$\frac{\partial I_1^h}{\partial x} = -\frac{1}{x^2} + \frac{1}{(L-x)^2\sqrt{1 + c_a^2 \triangle u^2/(L-x)^2}}. \tag{5.4}$$

To obtain an explicit expression for I_1^h, referred to as the regularized functional, we integrate (5.4) and get

$$I_1^h = \frac{1}{x} + \frac{1}{c_a \triangle u} \ln\left[\frac{c_a \triangle u}{L-x} + \sqrt{1 + \frac{c_a^2 \triangle u^2}{(L-x)^2}}\right]. \tag{5.5}$$

Distributions of I_1^h and $\partial I_1^h/\partial x$, presented in Figure 5, illustrate three important properties of this new class of functionals:

1. *The functional* I_1^h *is convex within the interval* $(0, L)$ *for any* c_a.
2. *When* $c_a \to \infty$ *the position of* $x_{\min}$ *for* I_1^h *tends to* L *from the left.*
3. *There is an infinite barrier preventing the right cell from collapsing.*

The first property guarantees existence and uniqueness of the solution of the minimization problem within the interval $(0, L)$. According to the second one, point 2 can approach point 3 to within any small distance. The third one states that the infinite barrier keeps the mesh nondegenerate. The infinite barrier allows the clustering of the mesh in the vicinity of the discontinuity up to any small size. When modeling a flow with shock waves it allows the accuracy to increase significantly, since the error caused by shock wave smearing, in the integral norms L_1 or L_2, is proportional to the shock thickness. A corollary is that two adjacent cells, one located in the shock zone and the other in the domain of smooth flow, with sizes differing by

orders of magnitude, do not degrade the accuracy of the solution. Thus, the mesh can be sharply clustered within one cell towards discontinuity. Those properties hold in the case of a mesh with any number of intervals $i_{\max}$ as well.

Note, when adapting with using the regularized functional (5.5), the velocity of grid nodes condensing, defined by dependence of the node 2 position on c_a at fixed L and $\triangle u$, is reduced in comparison with the case of the functional (5.1). It can be shown that at small c_a the ratio of the velocities to the functionals (5.1) and (5.5) equal 3. It leads one to have to perform a greater number of mesh iterations to obtain necessary degree of nodes condensing. We can increase the velocity of condensing by raising the expression under the square root in (5.4) to the power of $1 - 1/2^m$ (here substitution $m = 1$ gives (5.4)), and such a functional will be referred to as I^h_m. In practice $m = 2$ can be used as well.

Now we turn to a special case of constructing the 2-D adaptive mesh when the discrete functional can be reduced to be one-parametric. Suppose the control function f in (2.3) depends only on the variable x. Therefore, when seeking the mapping of the parametric square onto the domain Ω we have $x = x(\xi)$, $y = a\eta$ ($a = \text{const}$), and $f_y = x_\eta = y_\xi = 0$. Then the harmonic functional (2.3) reads (we set $a = 1$) as

$$I = \int_0^1 \int_0^1 \frac{x_\xi^2(1 + c_a^2 f_x^2) + 1}{x_\xi \sqrt{1 + c_a^2 f_x^2}} d\xi d\eta = \int_0^1 \frac{x_\xi^2(1 + c_a^2 f_x^2) + 1}{x_\xi \sqrt{1 + c_a^2 f_x^2}} d\xi. \tag{5.6}$$

The functional (5.6) differs from (2.4) by the additional term $\int x_\xi \sqrt{1 + c_a^2 f_x^2} d\xi$ expressing the curve $c_a f$ length in a cross-section $y = \text{const}$. This term defines the difference in properties of the 1-D and 2-D regularized discrete functionals.

We follow assumptions of section 4 when approximating (5.6) on the two-cell grid

$$I^h = x + \frac{1}{x} + (L - x)\sqrt{1 + c_a^2 \triangle u^2/(L - x)^2} + \frac{1}{(L - x)\sqrt{1 + c_a^2 \triangle u^2/(L - x)^2}}.$$

This discrete functional possesses properties similar to (5.1). Within the interval $(0, L)$ at $c_a < c_a^{crt}$ there are a maximum and minimum of I^h which disappear at $c_a > c_a^{crt}$. Here the value of c_a^{crt} differs a bit from the one in the 1-D case.

To regularize I^h we again fix the metric when deriving the first derivative. Derivative of a new functional reads as

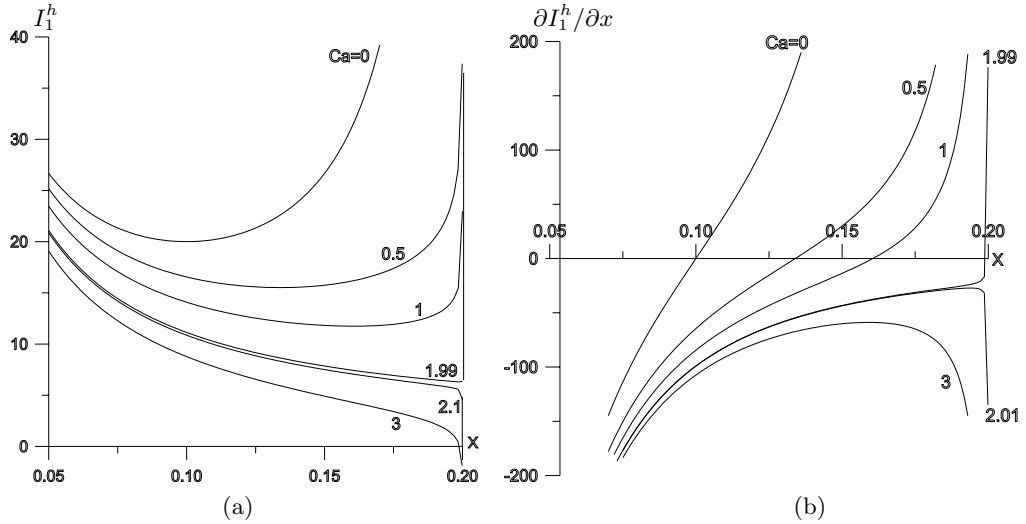
$$\frac{\partial I_1^h}{\partial x} = 1 - \frac{1}{x^2} - \sqrt{1 + c_a^2 \triangle u^2/(L - x)^2} + \frac{1}{(L - x)^2 \sqrt{1 + c_a^2 \triangle u^2/(L - x)^2}}.$$

Integrating it we get the regularized functional

$$I_1^h = \frac{1}{x} + x + \sqrt{(L - x)^2 + c_a^2 \triangle u^2} + \frac{1}{c_a \triangle u}(1 - c_a^2 \triangle u^2) \ln \left[\frac{c_a \triangle u}{L - x} + \sqrt{1 + \frac{c_a^2 \triangle u^2}{(L - x)^2}} \right].$$

Distributions of I_1^h and $\partial I_1^h / \partial x$ are presented in Figure 6. We see as soon as the term $1 - c_a^2 \triangle u^2$ becomes negative, in the above case at $c_a > 1/\triangle u = 2$, that the functional loses convexity. First, it seems to lead the right cell to collapse when finding the minimum of I_1^h via the iterative procedure (5.2). In practice it does not happen for the following reason. Derive the second derivative of I_1^h:

$$\frac{\partial^2 I_1^h}{\partial x^2} = \frac{1}{x^3} + \frac{2(L - x)^2 + c_a^2 \triangle u^2}{(L - x)^2 \sqrt{(L - x)^2 + c_a^2 \triangle u^2}} \left[\frac{1}{(L - x)^2 + c_a^2 \triangle u^2} + 1 \right].$$



(a) (b)

FIG. 6. 1-*D approach to* 2-*D functional* (2.3). *Dependence of distribution for* I_1^h (a) *and* $\partial I_1^h/\partial x$ (b) *on* c_a. *At* $c_a > 2$ *the functional loses convexity.*

When point 2 tends to point 3, i.e., x to L, and consequently we have $L - x \ll c_a \triangle u$, the ratio of the derivatives in (5.2) gives

$$\frac{\partial I_1^h}{\partial x} : \frac{\partial^2 I_1^h}{\partial x^2} \approx (L - x^p) \frac{1 - c_a^2 \triangle u^2}{1 + c_a^2 \triangle u^2}.$$

Since $1 - c_a^2 \triangle u^2 < 0$ at every iteration x^p gets an increment being smaller than the distance to the right node 3. The length of the right cell remains greater than zero within the truncation error or prescribed accuracy of calculation at any c_a. Thus, although it can turn out beginning from some value of c_a that the regularized functional will not have a minimum, nevertheless the iterative procedure (5.2) allows us to condense significantly the grid lines towards the discontinuity and to guarantee the grid to be unfolded.

The approach of the three-point model can also be applied in the general case of computing real 2-D flows with shocks. In the neighborhood of any point of discontinuity we introduce a local Cartesian system of coordinates with axis y directed along the tangent line towards the discontinuity at the point considered. Then, to a first approximation, we assume $x = x(\xi)$, $y = a\eta$, neglect the small terms in (2.3), i.e., f_x, x_η, y_ξ, and get (5.6). On the other hand, the presence of these small terms in the functional causes the nodes "to wander" permanently along some trajectory about an average position keeping strong grid lines condensing in the vicinity of shocks. Further, for the sake of simplicity, we shall refer the iterative procedure (quasi-Newton method in a real 2-D case, see section 7.2) as a minimization of the discrete functional independent of whether there is a solution of the minimization problem or not.

Performed analysis of the properties to the 1-D and 2-D functionals shows that these functionals are inconsistent, i.e., nodes clustering towards the discontinuity are performed in a different manner inside the domain Ω via minimization of the 2-D functional and on the boundary $\partial\Omega$ via the 1-D functional. The necessity arises in consistent redistribution of the grid nodes in Ω and on $\partial\Omega$. This matter will be considered in section 8.

6. Another method of adaptation. Let us see what the use of the functional (2.5) to the direct mapping or Euler–Lagrange equation (2.6) gives to constructing the adaptive mesh.

Consider the problem of minimizing the Dirichlet functional to the direct mapping of the domain $\Omega(x)$ onto ξ. Approximation of (2.5) on the two-cell mesh gives the discrete functional

$$I^h = x^2 + (L-x)^2 + c_a^2 \triangle u^2.$$

Its derivative

$$\frac{\partial I^h}{\partial x} = 4x - 2L$$

vanishes at $x = L/2$. Thus, independently on c_a the minimum of I^h reaches on the set of uniform meshes and grid clustering is not performed towards discontinuity.

Approximation of the Euler–Lagrange equation (2.6) is given

$$x = (L-x)\sqrt{1 + c_a^2 \triangle u^2/(L-x)^2}$$

and from it we obtain

$$x_{\min} = \frac{L}{2} + \frac{c_a^2 \triangle u^2}{2L}.$$

Consequently, at $c_a \triangle u = L$ the right cell collapses. In this case there is no barrier and in multidimensional problems, when grid lines are strongly bent near the shock wave, the derivative of a control function f towards the normal of the shock changes from one cell to another. In some cells we will not get enough condensation of grid lines and the other cells will have already been folded.

7. Optimization method and coupled algorithm.

7.1. 1-D case. We use the approximation (3.3) of the functional (2.4) and seek the minimum of I_m^h $(m = 1, 2)$ applying the Newton method (5.2) when $x = (x_1, \ldots, x_{i_{\max}+1})$. Using the approximations (3.4) to $(x_\xi)_{i+1/2}$ and $(f_x)_{i+1/2}$, fixing $(f_x)_{i+1/2}$ and setting $\triangle\xi = 1$ we get the first derivative

$$\frac{\partial I_m^h}{\partial x_i} = \frac{-1}{(x_i - x_{i-1})^2 \left[1 + c_a^2 (f_x)_{i-1/2}^2\right]^{1-2^{-m}}} + \frac{1}{(x_{i+1} - x_i)^2 \left[1 + c_a^2 (f_x)_{i+1/2}^2\right]^{1-2^{-m}}}.$$

The Hessian is a diagonal matrix with components $\mathcal{H}_{ii} = \partial^2 I_m^h / \partial x_i^2$. As above we fix $(f_x)_{i+1/2}$ while deriving the second derivative

$$\frac{\partial^2 I_m^h}{\partial x_i^2} = \frac{2}{(x_i - x_{i-1})^3 \left[1 + c_a^2 (f_x)_{i-1/2}^2\right]^{1-2^{-m}}} + \frac{2}{(x_{i+1} - x_i)^3 \left[1 + c_a^2 (f_x)_{i+1/2}^2\right]^{1-2^{-m}}}.$$

Here the case $m = 2$ is used to increase the velocity of grid nodes condensing.

If we need to perform adaptation along the curve, e.g., boundary $\partial\Omega$ of the domain Ω, the 1-D functional (2.4) can be written in the parametric form [31, 32]

$$I = \int_0^1 \frac{1}{t_\xi \sqrt{1 + c_a^2 f_t^2}} d\xi, \tag{7.1}$$

where the control function $f = f(t)$, parameter t defines the length of the boundary $\partial\Omega$. Then the Newton method (5.2) of minimization can be applied.

7.2. 2-D case. We use the optimization algorithm suggested in [12, 24]. Approximation of the functional (2.3) is performed on the mesh of quadrilaterals and is given by (3.1). If the set of convex grids is not empty, the system of algebraic equations (3.2) has at least one solution which is the convex mesh. Assuming the grid to be convex at the pth step of the iterative procedure we find the coordinates of the ith node at the $p+1$st step using the quasi-Newton method in the sense that the Hessian is a diagonal matrix (see [4, 12, 24] for details):

$$
\begin{aligned}
x_i^{p+1} &= x_i^p - \tau \left(R_x \frac{\partial R_y}{\partial y_i} - R_y \frac{\partial R_x}{\partial y_i} \right) \left(\frac{\partial R_x}{\partial x_i}\frac{\partial R_y}{\partial y_i} - \frac{\partial R_y}{\partial x_i}\frac{\partial R_x}{\partial y_i} \right)^{-1}, \\
y_i^{p+1} &= y_i^p - \tau \left(R_y \frac{\partial R_x}{\partial x_i} - R_x \frac{\partial R_y}{\partial x_i} \right) \left(\frac{\partial R_x}{\partial x_i}\frac{\partial R_y}{\partial y_i} - \frac{\partial R_y}{\partial x_i}\frac{\partial R_x}{\partial y_i} \right)^{-1};
\end{aligned}
\tag{7.2}
$$

here τ is the iterative parameter.

Note when finding the first and second derivatives of the functional (3.1) we fix the metric (derivatives f_x and f_y in (2.3)) and it is referred to as I_1^h.

To increase the velocity of grid nodes condensing in the discrete functional instead of $\sqrt{1+f_x^2+f_y^2}$ we use the term

$$
\left[1+(f_x)^2+(f_y)^2\right]^{1-2^{-m}},
$$

where $m = 1, 2$, and as in the 1-D case such a discrete functional will be referred to as I_m^h.

If a flow solver gives the values of the control function in the cell's center, it is required that we update them to the nodes. It can be done by interpolation and it is sufficient to use a first-order interpolation formula, e.g., to the ith node (except boundary nodes) which is surrounded by four cells; we have

$$
f_i = c_a \sum_{l=1}^{4} f_l^c A_l \Bigg/ \sum_{l=1}^{4} A_l,
$$

where f_l^c is the value in the lth cell center, A_l is the area of the triangle, one vertex of which is the ith node and two others are adjacent vertexes of the lth cell.

The coefficient of adaptation can depend on the node position, i.e., $c_a = c_a(x, y)$.

7.3. Coupled algorithm. One time step to solve the 1-D or 2-D equations of gas dynamics with grid adaptation contains the following steps:

1. Generate the mesh at the next time level $n+1$.
2. Compute the gas dynamics values at time t^{n+1}.
3. Make one iteration step and compute the new values of $(x, y)_i$ at t^{n+1} by formulas (5.2) or (7.2).
4. Repeat starting with step 2 to convergence or within given number of iterations p_{iter}.
5. Compute the final gas dynamics values at t^{n+1}.

The matter of preparing an initial quasi-uniform mesh, including the procedure of untangling the initial prepared folded mesh, is considered in [12, 24].

We now consider how to select c_a. As shown in section 5 the present method generates an unfolded mesh at any c_a and L_1- or L_2-errors depend on the shock thickness. Besides, used here the Godunov-type solver smears the shocks within 2 to

3 cells. Theoretically the thinner the discontinuity (i.e., the larger c_a) is, the higher is the accuracy we get. Therefore, there are no grounds for an automatic choice of c_a in contrast to, for instance, [37], where a smoothing of the solution and variation of the number of grid points in the shock zone are used. In practical computations, however, definition of c_a should satisfy some reasonable requirements. On the one hand c_a must not be too large, otherwise the mesh begins "to feel" weak compression and rarefaction waves that leads to undesirable and useless distortion of the cells. Besides, the larger the value of c_a, the less iterative the parameter τ needs to be in order not to leave the admissible set of convex grids (see Tables 10.1 and 10.2), and the more number of iterations p_{iter} we have to perform. On the other hand, if c_a is too small, then the mesh cannot be condensed in the vicinity of shocks of interest. Thus, we should choose c_a so that we avoid these lacks in mesh adaptation. Further, as it will be shown in section 10, on the one hand to get a substantial win in accuracy we need to have very strong grid lines condensing near the shocks. On the other hand, in gas dynamics calculations at strong grid lines clustering the admissible time step $\triangle t$ becomes rather small. It is not burdensome to steady flows. But when computing unsteady flows too small a $\triangle t$ will cause the overall time of modeling to increase significantly. To avoid a too large running time in [4, 5] the basic calculation of the 2-D unsteady flow was performed with not very large c_a when $\triangle t \approx 0.1 \triangle t_u$, where $\triangle t_u$ is the step on the quasi-uniform mesh. At some time before the control time c_a is increased so that $\triangle t \approx 0.01 \triangle t_u$. And such a technique gave very high resolution of the shocks and contact discontinuity. One more reason not to set c_a too large is that by controlling cell sizes we can eliminate only the errors caused by shock smearing; see sections 10.3 and 10.4. There also exist the errors gained throughout the domains of smooth flow. And it is useless to try to reduce the first type of error to zero.

Numerical experiments have shown that to get a suitable mesh it is sufficient to define c_a to be a constant, linear, or piecewise linear function along some distinctive direction in Ω, and its values are within the interval $0.05 \le c_a \le 0.5$.

The above discussion shows that both in theory and in practice the parameter c_a is free and the user should find an optimal solution taking into account the particularity of the concrete problem.

As it was shown in section 5 in the 2-D case, it may happen that the iteration procedure (7.2) will not converge, and we define the number of iterations p_{iter} at every time step. In real 2-D flow computations one should set $p_{\text{iter}} = 8$ to 10 for the rapidly developing unsteady flows and $p_{\text{iter}} = 1$ to 2 for the nearly steady flows [5].

8. Redistribution of nodes along the boundary curve. There are several ways to redistribute the grid nodes along the boundary $\partial\Omega$ during adaptation. The simplest one is a fixed position of every point on $\partial\Omega$, referred to as "fixed position." When moving the interior nodes towards a discontinuity, some instability in mesh generation and, consequently, in the flow near the points where the discontinuity joins $\partial\Omega$ can arise. In the next method the boundary nodes are treated as interior and the vectors of shift are projected onto $\partial\Omega$ [26]; we call it "unconstrained minimization." This way can be used only if the discontinuity is nearly orthogonal to $\partial\Omega$. If not, then, when condensing, the boundary nodes overlap, adjacent cells degenerate, and modeling breaks. The next method consists of using the 1-D functional (7.1) [12] referred to as "1-D minimization." It is more robust than the two methods discussed above and can usually be used at adaptation. However, as it has been shown in section 5, the 1-D and 2-D functionals are inconsistent. By this reason the parameters of adaptation c_a and τ should be selected separately. It requires additional work

and is particularly cumbersome when modeling unsteady flows. Sometimes we get undesirable displacement of the boundary nodes up to their overlap.

It is required that we perform redistribution of the interior and boundary nodes consistently. In the suggested method we perform constrained minimization of the discrete functional (3.1) under constraints defining $\partial\Omega$, referred to as "constrained minimization." Constrained minimization on $\partial\Omega$ has been also applied in [29]. We minimize the functional [3]

$$\tilde{I}_1^h = \sum_{i=1}^{i_{\max}} \sum_{k=1}^{4} \frac{1}{4} [F_k]_i + \sum_{l\in\mathcal{L}} \lambda_l G_l = I_1^h + \sum_{l\in\mathcal{L}} \lambda_l G_l, \tag{8.1}$$

where the constraints $G_l = G(x_l, y_l) = 0$ define $\partial\Omega$, λ_l are the Lagrange multipliers, and $\mathcal{L}$ is the set of the boundary nodes. Since the function $G(x, y)$ is assumed piecewise differentiable, the functional $\tilde{I}_1^h$ holds the infinite barrier on the boundary of the set of convex grids as I_1^h does if $f \in C^1$. If f is of the class of discontinuous functions, then the analysis from section 5 can be applied here.

If the set of convex grids is not empty, the system of algebraic equations has at least one solution that is the convex mesh

$$R_x = \frac{\partial I_1^h}{\partial x_i} + \lambda_i \frac{\partial G_i}{\partial x_i} = 0, \quad R_y = \frac{\partial I_1^h}{\partial y_i} + \lambda_i \frac{\partial G_i}{\partial y_i} = 0; \quad G_i = 0; \tag{8.2}$$

here $\lambda_i = 0$ if $i \notin \mathcal{L}$ and constraints are defined for the boundary nodes $i \in \mathcal{L}$.

Consider the method of minimizing the functional (8.1) assuming the grid to be convex at the pth step of the iterative procedure. We use the quasi-Newton procedure to find the coordinates x_i^{p+1}, y_i^{p+1} of the ith node from the system (8.2)

$$\begin{aligned}
&\tau R_x + \frac{\partial R_x}{\partial x_i}(x_i^{p+1} - x_i^p) + \frac{\partial R_x}{\partial y_i}(y_i^{p+1} - y_i^p) + \frac{\partial R_x}{\partial \lambda_i}(\lambda_i^{p+1} - \lambda_i^p) = 0,\\
&\tau R_y + \frac{\partial R_y}{\partial x_i}(x_i^{p+1} - x_i^p) + \frac{\partial R_y}{\partial y_i}(y_i^{p+1} - y_i^p) + \frac{\partial R_y}{\partial \lambda_i}(\lambda_i^{p+1} - \lambda_i^p) = 0,\\
&\tau G_i + \frac{\partial G_i}{\partial x_i}(x_i^{p+1} - x_i^p) + \frac{\partial G_i}{\partial y_i}(y_i^{p+1} - y_i^p) = 0,
\end{aligned} \tag{8.3}$$

where

$$\begin{aligned}
&\frac{\partial R_x}{\partial x_i} = \frac{\partial^2 I_1^h}{\partial x_i^2} + \lambda_i \frac{\partial^2 G_i}{\partial x_i^2}, \quad \frac{\partial R_x}{\partial y_i} = \frac{\partial^2 I_1^h}{\partial x_i \partial y_i} + \lambda_i \frac{\partial^2 G_i}{\partial x_i \partial y_i}, \quad \frac{\partial R_x}{\partial \lambda_i} = \frac{\partial G_i}{\partial x_i},\\
&\frac{\partial R_y}{\partial x_i} = \frac{\partial^2 I_1^h}{\partial x_i \partial y_i} + \lambda_i \frac{\partial^2 G_i}{\partial x_i \partial y_i}, \quad \frac{\partial R_y}{\partial y_i} = \frac{\partial^2 I_1^h}{\partial y_i^2} + \lambda_i \frac{\partial^2 G_i}{\partial y_i^2}, \quad \frac{\partial R_y}{\partial \lambda_i} = \frac{\partial G_i}{\partial y_i}.
\end{aligned}$$

Resolving the last equation of (8.3) about $y_i^{p+1} - y_i^p$ and substituting it in the two remaining equations, we get the system

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_i^{p+1} - x_i^p \\ \lambda_i^{p+1} - \lambda_i^p \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix},$$

where

$$\begin{aligned}
a_{11} &= \frac{\partial R_x}{\partial x_i} - \frac{\partial R_x}{\partial y_i}\frac{\partial G_i}{\partial x_i} \Big/ \frac{\partial G_i}{\partial y_i}, \quad a_{12} = \frac{\partial G_i}{\partial x_i}, \quad a_{13} = \tau \left[\frac{\partial R_x}{\partial y_i} G_i \Big/ \frac{\partial G_i}{\partial y_i} - R_x \right],\\
a_{21} &= \frac{\partial R_y}{\partial x_i} - \frac{\partial R_y}{\partial y_i}\frac{\partial G_i}{\partial x_i} \Big/ \frac{\partial G_i}{\partial y_i}, \quad a_{22} = \frac{\partial G_i}{\partial y_i}, \quad a_{23} = \tau \left[\frac{\partial R_y}{\partial y_i} G_i \Big/ \frac{\partial G_i}{\partial y_i} - R_y \right].
\end{aligned}$$

Denoting

$$\triangle = a_{11}a_{22} - a_{12}a_{21}, \quad \triangle_1 = a_{13}a_{22} - a_{23}a_{12}, \quad \triangle_2 = a_{11}a_{23} - a_{21}a_{13},$$

we obtain

$$x_i^{p+1} = x_i^p + \triangle_1/\triangle, \quad \lambda_i^{p+1} = \lambda_i^p + \triangle_2/\triangle, \tag{8.4}$$

and y_i^{p+1} is determined from the third equation of (8.3). If the constraints are resolved about y in the form $G(x, y) = y - g(x) = 0$, then

$$\frac{\partial G_i}{\partial x_i} = -\frac{\partial g_i}{\partial x_i}, \quad \frac{\partial G_i}{\partial y_i} = 1,$$

and above formulas are simplified. Constraints can be resolved about x in the form $G(x, y) = x - \tilde{g}(y) = 0$ and then (here it is better to resolve the third equation of (8.3) about $x_i^{p+1} - x_i^p$)

$$\frac{\partial G_i}{\partial x_i} = 1, \quad \frac{\partial G_i}{\partial y_i} = -\frac{\partial \tilde{g}_i}{\partial y_i}.$$

These two forms of $G(x, y)$ can substitute for each other. For example, on the part of $\partial\Omega$ that is nearly parallel to the axis x the boundary should be defined in the form $y = g(x)$, and where $\partial\Omega$ is nearly parallel to the axis y it should be defined as $x = \tilde{g}(y)$.

If $\partial\Omega$ is given by parametric functions $x = x(t)$, $y = y(t)$ or tabular values $(x, y)_i$, the following algorithm can be used. When calculating the coordinates of the ith node, in the interval (x_{i-1}, x_{i+1}) we construct an interpolating parabola $t = t(x)$ using the values in three nodes $i - 1, i, i + 1$. From (8.4) we compute an intermediate value $\tilde{x}_i^{p+1}$, further from the interpolation formula we determine $t_i = t(\tilde{x}_i^{p+1})$ and final values x_i^{p+1}, y_i^{p+1} from the parametric formulas.

Another way of redistributing the nodes along $\partial\Omega$, given as parametric functions or by tabular values, employs an unconstrained minimization of the functional in parametric form and is based on solving the following system of algebraic equations [3], referred to as "parametric minimization,"

$$R_t = R_x \frac{\partial x_i}{\partial t_i} + R_y \frac{\partial y_i}{\partial t_i} = 0,$$

via the quasi-Newton procedure

$$\tau R_t + \frac{\partial R_t}{\partial t_i}(t_i^{p+1} - t_i^p) = 0. \tag{8.5}$$

Here

$$\frac{\partial R_t}{\partial t_i} = \frac{\partial R_x}{\partial x_i}\left(\frac{\partial x_i}{\partial t_i}\right)^2 + \frac{\partial R_y}{\partial y_i}\left(\frac{\partial y_i}{\partial t_i}\right)^2 + \left(\frac{\partial R_x}{\partial y_i} + \frac{\partial R_y}{\partial x_i}\right)\frac{\partial x_i}{\partial t_i}\frac{\partial y_i}{\partial t_i}$$
$$+ R_x \frac{\partial^2 x_i}{\partial t_i^2} + R_y \frac{\partial^2 y_i}{\partial t_i^2}, \quad R_x = \frac{\partial I_1^h}{\partial x_i}, \quad R_y = \frac{\partial I_1^h}{\partial y_i}.$$

To the analytical control functions constrained and parametric minimization give similar results. Real-world 2-D flow computations have shown it is better to perform adaptation along the boundary using constrained minimization (8.3), (8.4) since the procedure (8.5) does not always ensure consistent redistribution of the nodes in Ω and on $\partial\Omega$.

9. Flow solver. In this section we briefly describe the Godunov linear flux correction (GLFC) scheme [2] to compute the 2-D gas flow in the Euler approach.

Equations of gas dynamics are written in the integral form which can be derived by transformation of the volume integrals in the space x-y-t to the surface integrals by virtue of Gauss's theorem as shown below for the law of conservation of mass:

$$\iiint_V \left[\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \boldsymbol{V})\right] dV = \oiint_{\partial V} \rho dxdy + \rho u dydt + \rho v dtdx = 0;$$

here V is an arbitrary control volume, homeomorphic sphere in space x-y-t, ∂V is the boundary of V.

Hence, the laws of conservation of mass, momentum, and total energy, can be written in the integral form, or generalized formulation [20], as follows:

$$\oiint_{\partial V} \boldsymbol{\sigma} dxdy + \boldsymbol{a} dydt + \boldsymbol{b} dtdx = \boldsymbol{0}, \tag{9.1}$$

where

$$\boldsymbol{\sigma} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{bmatrix}.$$

Here u and v are the velocity components, p and ρ are the pressure and density. The total energy $E = \rho[e + 0.5(u^2 + v^2)]$; e is the specific internal energy. The equation of state is $p = (\gamma - 1)\rho e$, where γ is the ratio of specific heats. Denote the vector-valued unknown functions as $\boldsymbol{f} = (u, v, p, \rho)^T$. The conservation laws (9.1) hold for any parameters $\boldsymbol{f}$, both smooth and discontinuous, governing a real gas flow.

We introduce the curvilinear moving grid in space x-y-t and consider the hexahedral computing cell; see Figure 7. The bottom face of the cell (or control volume) is taken at time level n and the top face at level $n+1$; four lateral faces generally form ruled surfaces rather than simple planes.

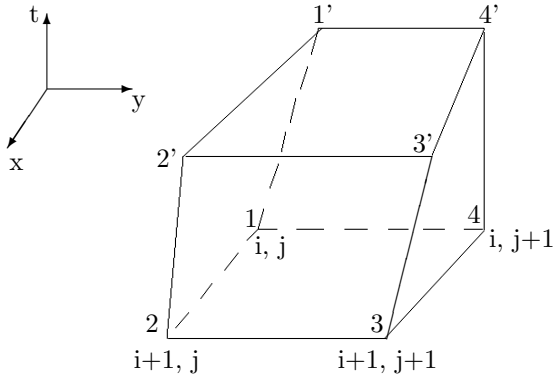


FIG. 7. *Computing cell.*

Integrating (9.1) over the oriented surface, being the boundary ∂V of the computing cell, we obtain a cell-centered finite-volume discretization of the governing equations

$$\boldsymbol{\sigma}^{n+1} A^{n+1} - \boldsymbol{\sigma}^n A^n + \boldsymbol{Q}_{411'4'} + \boldsymbol{Q}_{233'2'} + \boldsymbol{Q}_{122'1'} + \boldsymbol{Q}_{344'3'} = \boldsymbol{0}, \tag{9.2}$$

where $\boldsymbol{\sigma}^{n+1}$ and $\boldsymbol{\sigma}^n$ are the average values at time t^{n+1} and t^n in the center of the top and bottom faces, A^{n+1} and A^n are the areas of these faces. Each of four vector values $\boldsymbol{Q}_{411'4'}$, $\boldsymbol{Q}_{233'2'}$, $\boldsymbol{Q}_{122'1'}$, and $\boldsymbol{Q}_{344'3'}$ is an average flux of mass, momentum, and energy through the corresponding intercell surface in the direction of the outward normal vector. Unlike the original Godunov scheme [20] where the fluxes in (9.2) are taken at time t^n, in the present scheme those values are computed at $t^{n+1/2}$ that provides it with the second-order accuracy in time.

For example, at the face $122'1'$ the value $\boldsymbol{Q}_{122'1'}$ has the following structure:

$$\boldsymbol{Q}_{122'1'} = \boldsymbol{\sigma}^{n+1/2}A^{xy} + \boldsymbol{a}^{n+1/2}A^{yt} + \boldsymbol{b}^{n+1/2}A^{tx}, \tag{9.3}$$

where A^{xy}, A^{yt}, A^{tx} are the areas of projections of the face $122'1'$ onto the coordinate planes x-y, y-t, and t-x, respectively, given by

$$A^{xy} = \iint_{122'1'} dxdy = \frac{1}{2}[(x_{2'} - x_1)(y_{1'} - y_2) - (x_{1'} - x_2)(y_{2'} - y_1)],$$
$$A^{yt} = \iint_{122'1'} dydt = \frac{1}{2}\triangle t(y_{2'} + y_2 - y_1 - y_{1'}),$$
$$A^{tx} = \iint_{122'1'} dtdx = -\frac{1}{2}\triangle t(x_{2'} + x_2 - x_1 - x_{1'}),$$

which are obtained from the formula for the quadrangle 1234

$$A_{1234} = A(x_1, y_1; x_2, y_2; x_3, y_3; x_4, y_4) = \frac{1}{2}[(x_3 - x_1)(y_4 - y_2) - (x_4 - x_2)(y_3 - y_1)]$$

when running along its contour in an anticlockwise manner, time step $\triangle t = t^{n+1} - t^n$.

The values $\boldsymbol{f}^{n+1}$ are updated by two stages using time splitting. In the first stage, predictor, via (9.2) we compute the intermediate values $\tilde{\boldsymbol{f}}^{n+1}$. Here we apply the piecewise linear interpolation along each curvilinear coordinate line ξ, passing through the center of cells $j + 1/2 = \text{const}$, and η, passing through the center of cells $i + 1/2 = \text{const}$; i.e., determine the derivatives $\boldsymbol{f}_\xi$ and $\boldsymbol{f}_\eta$ in every cell to get the fluxes via (9.3) on the lateral faces with the second-order accuracy in space but still at t^n. To suppress spurious oscillations in the vicinity of discontinuities a monotonicity algorithm is applied [2]. At the second stage, corrector, using $\boldsymbol{f}^n$, $\tilde{\boldsymbol{f}}^{n+1}$ and derivatives $\boldsymbol{f}_\xi$, $\boldsymbol{f}_\eta$ at t^n, we get prewave values on both sides of each of four lateral faces of the cell at $t^{n+1/2}$. After solving the Riemann problem we obtain $\boldsymbol{f}^{n+1/2}$, the values at each of four lateral faces, and again calculate the fluxes from (9.3). Then, via (9.2), we compute the final values $\boldsymbol{f}^{n+1}$.

In the $(i + 1/2, j + 1/2)$th cell the admissible time step $\triangle t_{i+1/2,j+1/2}$ is defined via [20] as

$$\triangle t_{i+1/2,j+1/2} = \frac{\triangle t' \ \triangle t''}{\triangle t' + \triangle t''},$$

where

$$\triangle t' = \frac{h'}{\max(d^{II}_{41} - w_{41}; -d^{I}_{23} - w_{23})}, \qquad \triangle t'' = \frac{h''}{\max(d^{II}_{12} - w_{12}; -d^{I}_{34} - w_{34})},$$
$$h' = \frac{A_{1234}}{0.5\sqrt{(x_4 + x_3 - x_1 - x_2)^2 + (y_4 + y_3 - y_1 - y_2)^2}},$$
$$h'' = \frac{A_{1234}}{0.5\sqrt{(x_3 + x_2 - x_4 - x_1)^2 + (y_3 + y_2 - y_4 - y_1)^2}}.$$

Here $\triangle t^{'}$ and $\triangle t^{''}$ are the admissible time steps to the 1-D schemes in the ξ and η directions, respectively, $h^{'}, h^{''}$ are the "average heights" of the bottom face A_{1234}, w is the velocity of the corresponding cell edge, e.g., w_{12} is the velocity of the edge 12 which defines inclination of the face $122'1'$; see Figure 7. Next, d_{12}^{II} and d_{41}^{II} are the "extreme right wave" speeds defined from solving the Riemann problem to the faces $122'1'$ and $11'4'4$, respectively; d_{23}^{I} and d_{34}^{I} are the "extreme left wave" speeds to the faces $233'2'$ and $344'3'$, respectively.

As an admissible time step we take the minimal of all cells

$$\triangle t = \nu \min_{i,j} \triangle t_{i+1/2,j+1/2}.$$

The coefficient ν is less than 1 (usually $0.5 \le \nu \le 0.9$) and it is introduced as a correction to the nonlinearity of the problem. Note that the time step depends on both postwave values and the velocity of every intercell face. In computations the value of $\triangle t$, obtained at the preceding time step, is used for the next time step. By this reason, in the case of essentially nonstationary processes, the coefficient ν should be greatly decreased.

The GLFC scheme is of second-order accuracy in time and space in the domains of smooth flow. We get the values $\boldsymbol{f}^{n+1}$ directly on the moving mesh and need not perform interpolation at t^{n+1} from one mesh to the other.

10. Examples of modeling.

10.1. Analytical control function. First we demonstrate a simple test illustrating the inconsistency of redistributing the boundary and interior nodes when using various methods from section 8 and vice versa, i.e., their consistency when using another.

The 50×50 adaptive mesh is generated in the unit square $0 < x, y < 1$ when the control function is defined to be

$$f(x,y) = \begin{cases} 1 & \text{if} \quad y < 0.5, \\ 0 & \text{if} \quad y \ge 0.5. \end{cases}$$

Fragments of the adapted meshes in the vicinity of the discontinuity are presented in Figure 8. In the first case the coefficient $c_a = 0.1$; see Figure 8(a)–(c). When we apply fixed position and 1-D minimization methods of redistributing the boundary nodes (see Figure 8(a)–(b)), the horizontal grid lines are not parallel and in the case of the other 3 methods they are parallel. In the next case the coefficient $c_a = 0.15$; see Figure 8(d)–(f). To the fixed position method the coordinate lines become more bent; see Figure 8(d). Using 1-D minimization leads the boundary nodes to overlap (see Figure 8(e)), i.e., the mesh to fold. This happens due to the inconsistency of the nodes' redistribution in Ω and on $\partial\Omega$ despite the fact that 1-D and 2-D algorithms separately provide unfolded grid generation. In this test unconstrained minimization gives the same result as constrained and parametric minimization due to the discontinuity is orthogonal to $\partial\Omega$ and here the horizontal lines almost merge near the discontinuity and remain parallel, keeping the mesh unfolded; see Figure 8(f).

In the next example the discontinuity is not orthogonal to $\partial\Omega$. The control function is defined as

$$f(x,y) = \begin{cases} 1 & \text{if} \quad y > 5x - 2, \\ 0 & \text{if} \quad y \le 5x - 2. \end{cases}$$

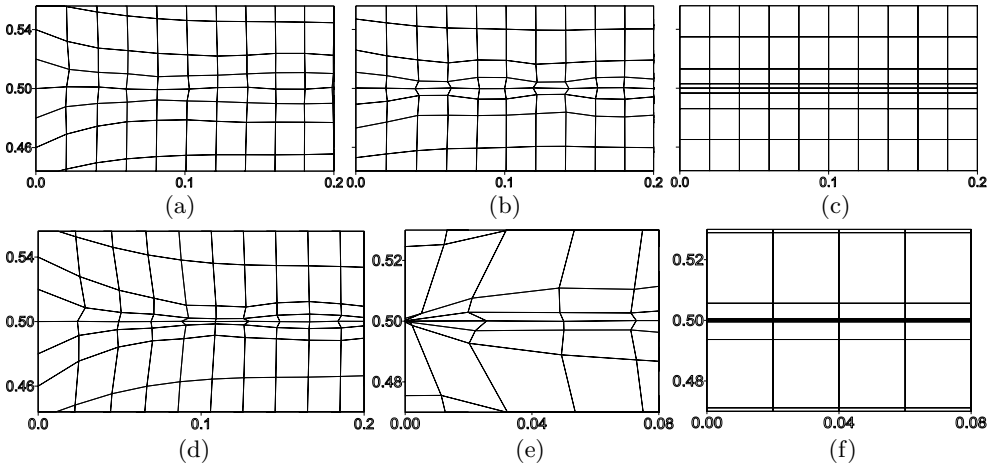


FIG. 8. *Fragment of the adapted mesh. The boundary nodes are redistributed using fixed position* (a), (d), 1-*D minimization* (b), (e), *and unconstrained or constrained or parametric minimization* (c), (f). *Coefficient* $c_a = 0.1$ *in the cases* (a)–(c) *and* $c_a = 0.15$ *in the cases* (d)–(f); *iterative parameter* $\tau = 0.15$.

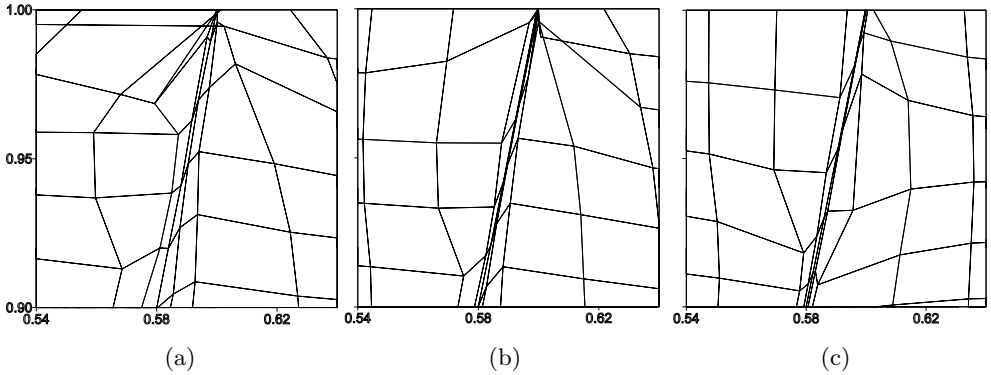


FIG. 9. *Fragment of the adapted mesh. The boundary nodes are redistributed using unconstrained* (a), 1-*D* (b), *and constrained or parametric minimization* (c) *methods. Coefficient* $c_a = 0.3$; *iterative parameter* $\tau = 0.1$.

Fragments of the adapted meshes near the top boundary are presented in Figure 9. Here using unconstrained and 1-D minimization leads the boundary nodes to overlap in several tenths of mesh iterations (see Figure 9(a)–(b)), respectively. Constrained and parametric minimization maintain an unfolded mesh; see Figure 9(c). To the analytical control functions constrained and parametric minimization give similar results.

10.2. Contact discontinuity. To observe how the adaptive procedure handles the contact discontinuities the Cauchy problem for the 1-D linear advection equation has been computed with the Godunov scheme; see Godunov and Ryaben'kii [19]. Adaptation is executed via (5.2). It appears that once the grid points have been condensed towards the contact discontinuity the thickness of discontinuity smeared increases with time proportionally to $\sqrt{t}$ as the theoretical estimation gives on the

TABLE 10.1

IVP to the inviscid Burger's equation (4.1). *Adaptation is performed by minimization of* I_1^h. *Here* $h_{\min}$ *is the smallest interval,* n *is the number of time steps,* τ *is the iterative parameter,* $||Er||_{L_1}$ *is the error.*

	Godunov's scheme				GLFC scheme			
c_a	$h_{\min}$	n	τ	$\|\|Er\|\|_{L_1}$	$h_{\min}$	n	τ	$\|\|Er\|\|_{L_1}$
0	0.1	21	-	0.078046	0.1	21	-	0.023551
0.1	0.056821	30	0.95	0.039784	0.049486	35	0.95	0.016755
0.25	0.022115	77	0.95	0.011285	0.020587	86	0.95	0.005337
0.5	0.010712	179	0.8	0.004249	0.010979	191	0.8	0.003395
1	0.004683	406	0.5	0.000540	0.004637	419	0.4	0.000204
2	0.002382	897	0.3	0.000297	0.001924	903	0.2	0.000043
4	0.001116	2065	0.15	0.000187	0.000918	1835	0.1	0.000049
8	0.000526	4438	0.07	0.000057	0.000624	3733	0.07	0.000068
16	0.000209	9355	0.05	0.000018	0.000246	7612	0.05	0.000065

uniform fixed mesh.

10.3. Inviscid Burger's equation. Consider the IVP (4.1) with initial values $u_l = 2$, $u_l = 1$. The initial uniform mesh has spacing $h = 0.1$ and zones number $i_{\max} = 50$. The boundary nodes are fixed. Calculations are performed up to $t = 1$ using the first-order Godunov scheme (4.4) and second-order GLFC scheme [2]. In this case the mesh structure differs a bit from the one in section 4. The shock smears over 3 cells with approximately similar length $h_{\min}$. To the left and right of those 3 cells there is a cell with intermediate length and the remaining 45 cells are similar (if a sufficient number of mesh iterations are provided). In computations, presented in Table 10.1, every time step includes $p_{\text{iter}} = 100$ mesh iterations.

The data of Table 10.1 show that increasing the coefficient of adaptation c_a from 0 to 16 decreases the thickness of smearing, equal approximately to $3h_{\min}$, by 478 times for the Godunov scheme and by 369 times for the GLFC scheme; that leads the accuracy to increase by factors of 4335 and 362 for those schemes, respectively. We see if on the uniform mesh the second-order scheme delivers a higher accuracy than the first-order one, on the strongly compressed mesh the situation changes. Moreover, to the GLFC scheme the accuracy is not increased at $c_a > 2$. This can be explained by the presence of spurious oscillations near the discontinuity to the second-order scheme.

10.4. Shock tube problem. Consider the 1-D flow of ideal gas with initial parameters $(u, p, \rho)_l = (2.253928, 2.296074, 1.7811378)$ in the left 3 cells and $(u, p, \rho)_r = (1.5, 1, 1)$ selected in such a way that at $t > 0$ these two domains are divided by the shock wave, moving from the left to the right. The ratio of specific heats $\gamma = 1.4$. Initial uniform mesh is as in section 10.3 and calculations are performed up to $t = 1$ with the same schemes. Every time step includes $p_{\text{iter}} = 100$ mesh iterations.

The data of Table 10.2 show that increasing the coefficient of adaptation c_a from 0 to 4 decreases thickness of the shock smeared, equal approximately $3h_{\min}$, by 80 times and accuracy increases by the factor of 194 and 138 for the Godunov and GLFC schemes, respectively. The second-order scheme gives the lesser error $||Er||_{L_1}$ estimated by the density. Further increase of c_a does not deliver a higher accuracy since the error, caused by shock smearing, becomes much less than the error gained throughout the other cells. If on the uniform mesh $c_a = 0$ the second-order scheme provides a higher accuracy than the first-order one by a factor of 3.1, at $c_a = 4$ the ratio of errors falls to 2.2. The last column shows $\triangle\rho$, the amplitude of spurious

TABLE 10.2

Shock tube problem. Adaptation is performed by minimization of I_1^h except the last row where I_2^h is used. Here $h_{\min}$ is the smallest interval, n is the number of time steps, τ is the iterative parameter, $||Er||_{L_1}$ is the error estimated by ρ, $\triangle\rho$ is the amplitude of spurious oscillations.

	Godunov's scheme				GLFC scheme				
c_a	$h_{\min}$	n	τ	$\|\|Er\|\|_{L_1}$	$h_{\min}$	n	τ	$\|\|Er\|\|_{L_1}$	$\triangle\rho, \%$
0	0.1	37	-	0.075736	0.1	38	-	0.024554	1.10
0.1	0.07469	46	0.9	0.088294	0.05827	54	0.8	0.028529	0.51
0.25	0.04080	87	0.9	0.088695	0.02784	113	0.5	0.017749	0.23
0.5	0.01583	211	0.7	0.025053	0.01307	248	0.5	0.011859	0.08
1	0.00702	486	0.7	0.005929	0.00604	584	0.5	0.001864	0.03
2	0.00259	1307	0.5	0.000581	0.00242	1396	0.4	0.000317	0.04
4	0.00126	3086	0.2	0.000391	0.00124	3245	0.2	0.000178	0.01
functional I_2^h									
0.25	0.00103	4891	0.13	0.003058	0.00103	5196	0.13	0.000897	0.06

TABLE 10.3

Shock tube problem. Parameters $\tau = 0.5$, $c_a = 1$.

$p_{\text{iter}} = 10$, $||Er||_{L_1} = 0.006681$

i	$h_{i+1/2}$	$\rho_{i+1/2}$	$\rho^{\text{exact}}_{i+1/2}$
37	0.10264	1.78201	1.78114
38	0.10547	1.78128	1.78114
39	0.10980	1.78037	1.78114
40	0.11837	1.77947	1.78114
41	0.09642	1.77808	1.78114
42	0.01553	1.70929	1.78114
43	0.01635	1.17676	1.00000
44	0.02131	1.00001	1.00000
45	0.04872	1.00000	1.00000
46	0.25688	1.00000	1.00000
47	0.27041	1.00000	1.00000
48	0.28275	1.00000	1.00000
49	0.29084	1.00000	1.00000

$p_{\text{iter}} = 100$, $||Er||_{L_1} = 0.002563$

i	$h_{i+1/2}$	$\rho_{i+1/2}$	$\rho^{\text{exact}}_{i+1/2}$
30	0.10704	1.78127	1.78114
31	0.10727	1.78120	1.78114
32	0.10751	1.78093	1.78114
33	0.10692	1.77965	1.78114
34	0.01884	1.77406	1.78114
35	0.00666	1.61892	1.78114
36	0.00604	1.00487	1.00000
37	0.00822	1.00000	1.00000
38	0.07492	1.00000	1.00000
39	0.11228	1.00000	1.00000
40	0.11305	1.00000	1.00000
41	0.11383	1.00000	1.00000
42	0.11459	1.00000	1.00000

oscillations to the density in the vicinity of the shock computed with the GLFC scheme. We see adaptation allows us to reduce significantly $\triangle\rho$ from 1.1% down to 0.01%. Using the functional I_2^h causes strong grid nodes clustering at rather small c_a and, consequently, $||Er||_{L_1}$ to reduce.

Note in the two examples above p_{iter} at every time step was equal to 100 to get similar cell spacings in the shock zone and domains of constant flow parameters to see what adaptation provides in the limit at the extreme large values of c_a. In Table 10.3 we present two computations of this problem with the GLFC scheme when $p_{\text{iter}} = 10$ and 100 with parameters $\tau = 0.5$, $c_a = 1$. The spacings $h_{i+1/2}$, values of $\rho_{i+1/2}$, and exact values of $\rho^{\text{exact}}_{i+1/2}$ in the segments center are shown in the vicinity of the shock. It can be seen that in the first case the value of p_{iter} is insufficient and nodes did not manage to come to their "final" positions judging by the values of $h_{i+1/2}$. Spacings before and behind the shock differ by a factor of 3. By this reason $h_{\min}$ is larger by a factor of ≈ 2.7 than the one in the second case, and therefore the L_1-error is too large. Thus, in the second case, by winning in accuracy we lose in the number of iterations p_{iter}.

10.5. Flow in a channel. The following test demonstrates the capability of the grid adaptation technique to reduce computer costs while obtaining the same accuracy as on the fixed quasi-uniform mesh. Into the channel, shown in Figure 10, from the

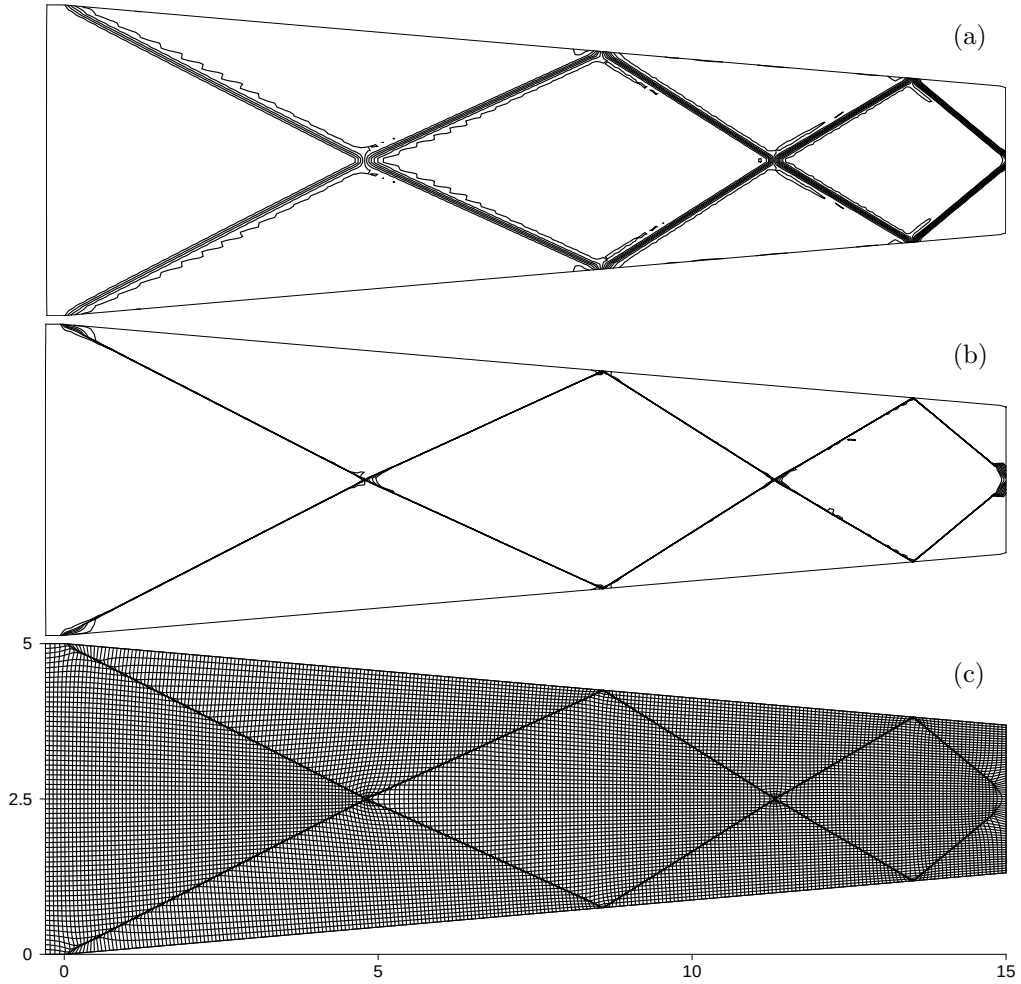


FIG. 10. *Supersonic flow in the channel.* ρ *is used as a control function. Density contours* (*from* 1 *to* 2.7 *with* $\triangle\rho = 0.05$) *computed on the quasi-uniform* (a) *and adapted* (b) *meshes* 264×64; *adapted mesh* 264×64 (c).

left is introduced an ideal gas with $M_\infty = 2.5$. There are two edges in the top and bottom with inclination of 5°. The steady flow has the following structure: two shock waves attached to the top of wedges intersect each other, reflect from the walls, again intersect, etc., dividing the domain into a set of subdomains which contain constant flow parameters.

The problem is calculated with the GLFC scheme on the successively refined grids with 64×16, 128×32, 256×64, and 512×128 zones number. At the first stage the calculation on the quasi-uniform mesh is performed until the solution achieves its steady state. Then, at the second stage, we switch to the adaptive procedure. ρ is used as a control function. At every time step we perform one mesh iteration, i.e., $p_{\text{iter}} = 1$. To provide the best accuracy of computations c_a is selected so to get similar grid clustering towards the shocks within the whole flow and this is achieved by defining a linear/(piecewise linear) dependence of c_a on the x-coordinate.

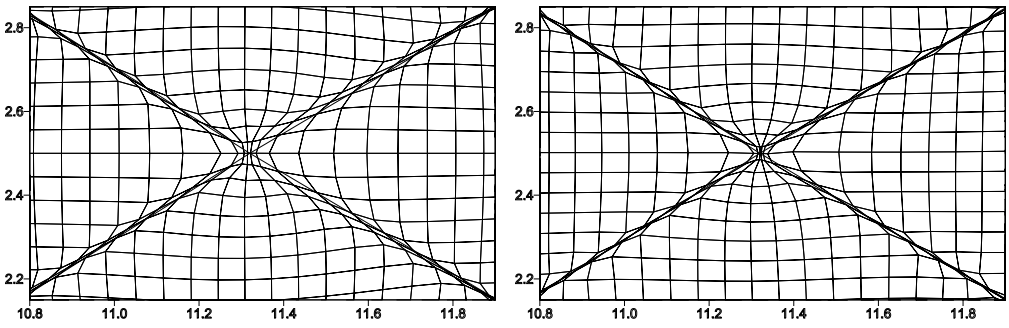


FIG. 11. *Fragments of adapted mesh in* 200 *time steps* (a) *and final mesh* (b). *Two diagonal straight-lines indicate position of shocks for the exact solution.*

TABLE 10.4

Dependence of error, estimated by ρ, *on zones number computed to quasi-uniform and adaptive meshes.*

	Error $\|\|Er\|\|_{L_1}$	
Zones number	Quasi-uniform mesh	Adaptive mesh
64×16	1.5598	0.4070
128×32	0.8274	0.1666
256×64	0.4290	0.0846
512×128	0.2207	0.0444

Figure 10 shows the density contours, when modeling on the quasi-uniform and adapted meshes 256×64 and adapted mesh. Here the first stage, to get a steady state on the quasi-uniform mesh by $t = 15$, takes 1739 time steps. Applying adaptation we get strong grid lines condensing in 200 time steps and a gain in accuracy by a factor of 2. The coefficient of adaptation is defined to be $c_a = 0.2 - 0.15(15 - x)/15$; iterative parameter $\tau = 0.3$. But condensed grid lines are still far from the shock location in the exact solution; see Figure 11(a). We have to perform 2000 time steps more to obtain maximal accuracy. After this rather long shocks "capturing" by the grid lines we get an additional gain in accuracy by a factor of 2.5; see the final mesh in Figure 11(b). To increase the grid velocity we use the functional I_2^h, and this reduces the time of shocks capturing by a factor of 1.7. Nevertheless, during the last 300 steps we use the functional I_1^h, since I_2^h does not provide very strong grid nodes condensing. Note that using I_2^h on the rough mesh 64×16 does not provide an increase of the grid velocity. In real computing the steady problems, when the exact solution is not known, the user shall watch whether the location of a shock changes with time; if yes, using the functional I_2^h can increase the grid velocity; if no, one should at once apply I_1^h. In Figure 11(a)–(b) it can be seen that the shock waves are smeared over three cells as in the 1-D case. Near the shocks the cells are very narrow and the maximal aspect ratio reaches 40. As for the 1-D case, due to adaptation, the amplitude of oscillations in the solution is reduced by many factors of ten.

Knowing the exact solution, one can estimate the error; see Table 10.4. We see adaptation allows us to increase the accuracy by a factor of 4 to 5. Keeping in mind that doubling the points in each space direction on the quasi-uniform mesh causes the accuracy to increase by a factor of 2, we can estimate that adaptation in the last three variants is equivalent to the mesh refinement by a factor of 5. Thus, adaptation allows us to gain in CPU memory by a factor of 25 and running time by a factor of

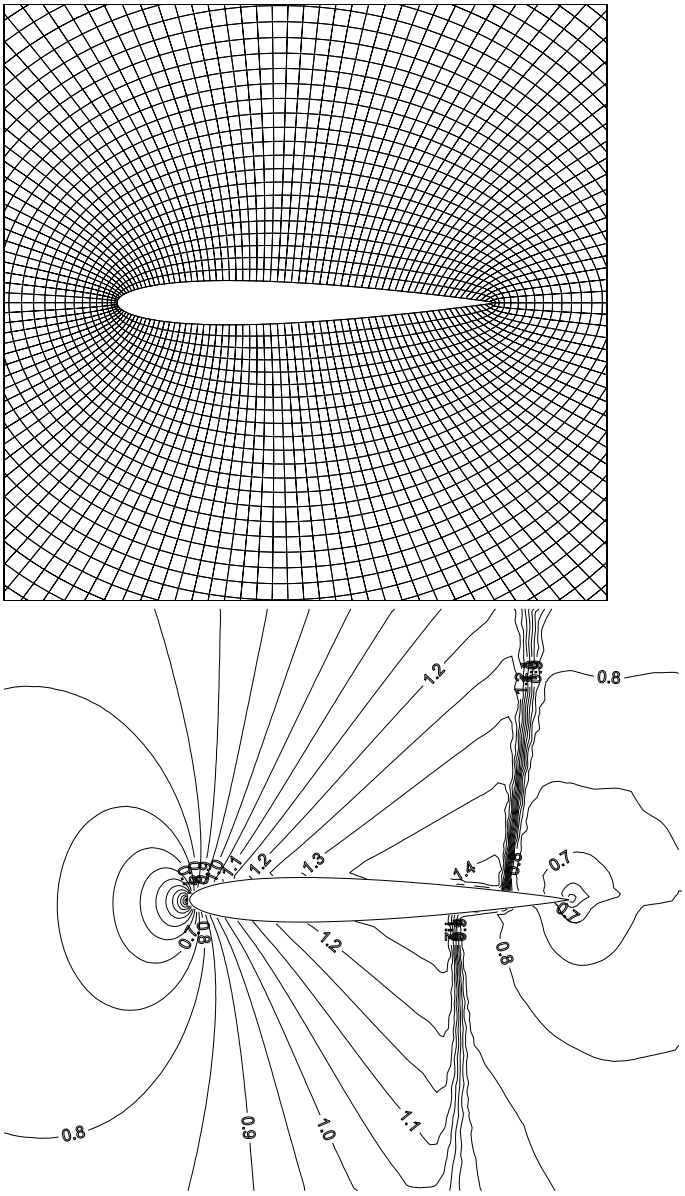


FIG. 12. *Transonic flow around NACA*0012 *airfoil. Quasi-uniform mesh* (a) *and Mach number contours* (b).

50 to 60 taking into account the additional time steps required at adaptation.

10.6. Flow over an airfoil. The GLFC scheme with an adaptive procedure is applied to calculating transonic and supersonic Euler flow over an NACA0012 airfoil. The first test is a transonic case with $M_\infty = 0.85$ and angle of attack $\alpha = 1^\circ$. Figure 12 presents the quasi-uniform O-mesh 140×80 and plots the Mach number contours. We see that shock waves, one (stronger) on the upper side of the airfoil and the other (weaker) on the low side, are rather thick. Figure 13 presents the adapted mesh and

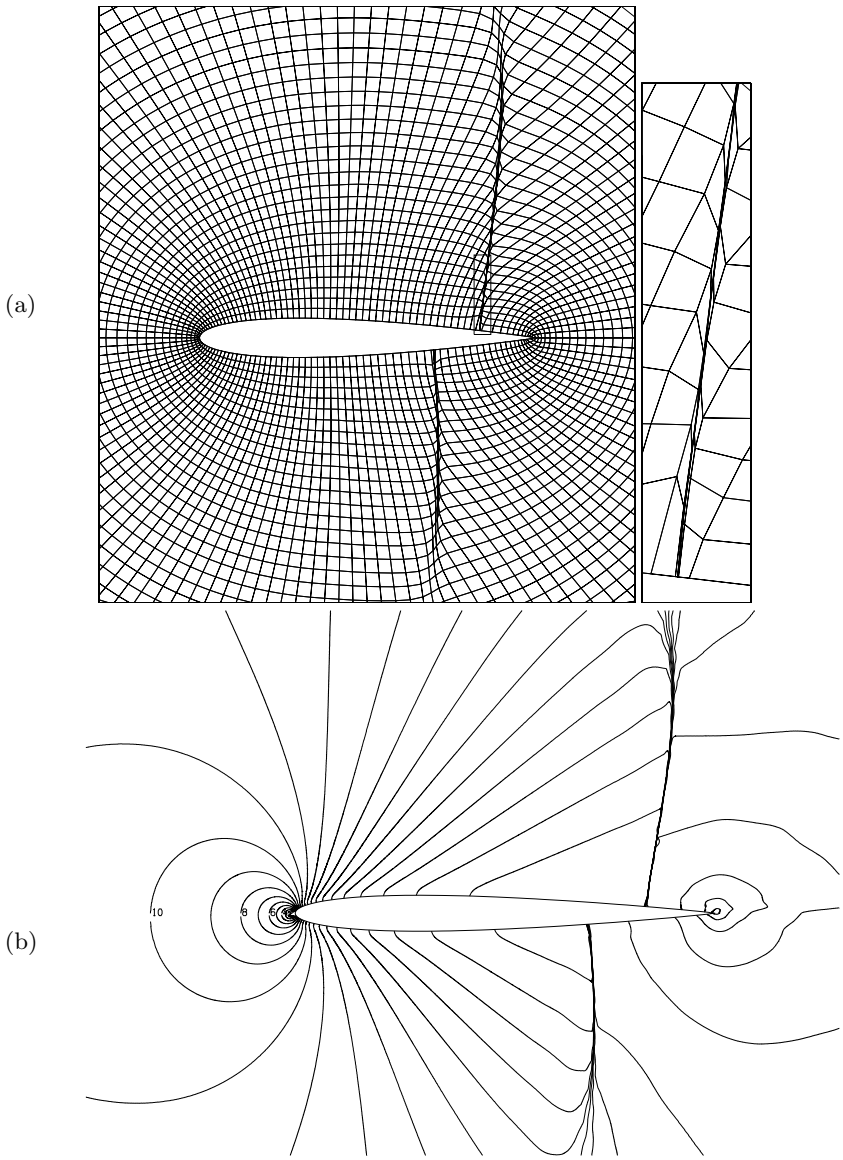


FIG. 13. *Transonic flow around NACA*0012 *airfoil. Adapted mesh* (a) *and Mach number contours* (b).

Mach number contours calculated on this grid. As a control function we use ρ. The coefficient of adaptation c_a is defined to be

$$c_a = \begin{cases} c_{\min} + (c_{\max} - c_{\min})|y|/0.8 & \text{if } |y| \le 0.8, \\ c_{\max} & \text{if } |y| > 0.8, \end{cases}$$

$$c_{\min} = 0.1, \quad c_{\max} = \begin{cases} 0.2 & \text{if } y < 0, \\ 0.15 & \text{if } y > 0. \end{cases}$$

A linear dependence of c_a on y is used to strengthen adaptation in the domains where the shocks grow weaker. Around the leading edge where there are rarefaction

waves we switch off adaptation by setting $c_a = 0$ to exclude cells distortion. Iterative parameter $\tau = 0.3$; the number of iterations at every time step $p_{\text{iter}} = 1$. Constrained minimization is applied to redistribute the boundary nodes along the airfoil contour and along the line $y = 0$ passing from the trailing edge which the boundaries of the parametric square $\xi = 0, 1$ correspond to. Using 1-D or parametric minimization with the same parameters c_a, τ on $\partial\Omega$ as in Ω leads the boundary nodes to overlap in the vicinity of the shocks. Thickness of the shocks is reduced by 50 times in comparison with the nonadapted mesh that provides us with capturing the discontinuities very accurately. The first stage takes 7345 time steps to compute by $t = 20$; adaptation does 150 steps that is about 4% of the first stage time.

Another test is a supersonic flow over the same airfoil with $M_\infty = 1.3$ and $\alpha = 0°$. Figure 14 plots the Mach number contours computed on the quasi-uniform and adapted O-meshs 120×50. As can be seen in Figure 14(a), a strong bow shock wave appears in front of the airfoil leading edge and two weak shocks emanate from the trailing edge. Using adaptation provides us with a very strong reduction in the bow shock thickness and a rather strong reduction in the trailing edge shocks thickness that is demonstrated by the both Mach number contours in Figure 14(b) and adapted grid in Figure 15. The coefficient of adaptation c_a is defined to be

$$c_a = \begin{cases} c_{\min} + (c_{\max} - c_{\min})|y|/1.5 & \text{if } |y| \le 1.5, \\ c_{\max} & \text{if } |y| > 1.5, \end{cases}$$

$$c_{\min} = \begin{cases} 0.05 & \text{if } x \le 0.6, \\ 0.1 & \text{if } x > 0.6, \end{cases} \qquad c_{\max} = 0.15,$$

and as above we set $c_a = 0$ near the airfoil leading edge. Parameters $\tau = 0.3$, $p_{\text{iter}} = 1$.

Grid points clustering allows us to hope that we nearly eliminated the errors caused by shock waves smearing and increased significantly the accuracy of computations. The first stage takes 5882 time steps to compute by $t = 5$; adaptation does 200 steps, which is about 7% of the first stage time. In both tests we use only minimization of the functional I_1^h.

11. Concluding remarks. Results of computations presented show adaptive-harmonic grid generation significantly increases the accuracy of calculations in comparison with modeling on quasi-uniform meshes due to reducing the thickness of smearing to the shock waves, while keeping the same simple grid structure and requiring less computer costs.

Theoretical analysis based on the three-point model of adaptation has shown minimization of the regularized discrete functional in the 1-D and 2-D cases delivers strong grid lines compression towards the discontinuities, while keeping the mesh unfolded. The errors of computations can be conditionally divided into two kinds. The first kind of errors are gained in the vicinity of the shock waves and the second throughout the subdomains of smooth flow. The value of the first in the integral norms L_1 or L_2 is proportional to the thickness of the shocks smeared and, consequently, grid clustering enables us to eliminate those errors or, at least, provides that their magnitude is insignificant in comparison with the second type errors, the value of which depends on the numerical scheme accuracy.

Constrained minimization leads to consistent redistribution of the boundary and interior mesh nodes that increases the reliability of the adaptive procedure and modeling.

In 2-D tests, presented in sections 10.5 and 10.6, the mesh with strong grid lines compression looks like a set of subdomains into which the flow is smooth and the

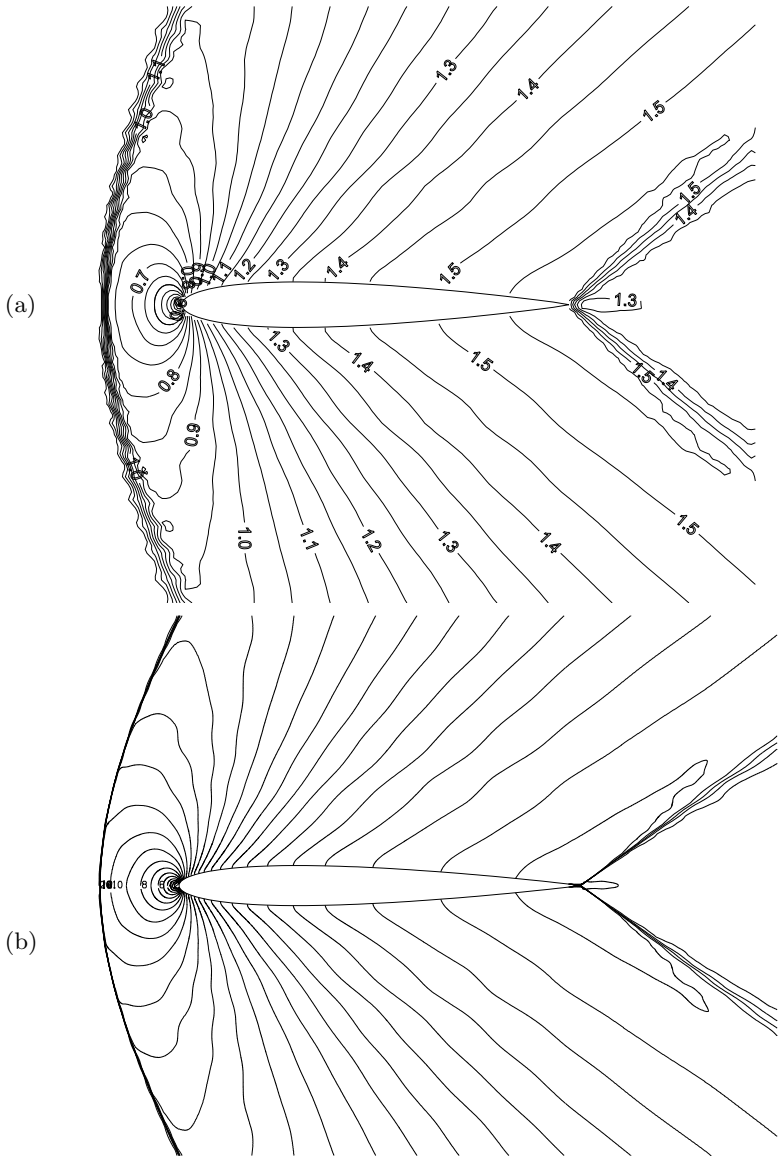


FIG. 14. *Supersonic flow around NACA*0012 *airfoil. Mach number contours computed on quasi-uniform* (a) *and adapted* (b) *grids.*

mesh is quasi-uniform. Boundaries of those subdomains coincide with the shocks and are built automatically by the condensed grid lines. If we draw a closed contour along the shock, one side to the left of the discontinuity and another to the right, and direct the contour width to zero, from the system of conservation laws, written in the integral form (9.1), we get in the limit the Rankine–Hugoniot conditions across the shock [33]. The line passing through the centers of narrow cells, stretched along the shocks, looks like that contour. To draw such a limiting contour we need two adjacent narrow cells in the shock zone as the three-point model of adaptation provides. In real computations (see sections 10.5 and 10.6) the shocks are smeared over 2 to 3

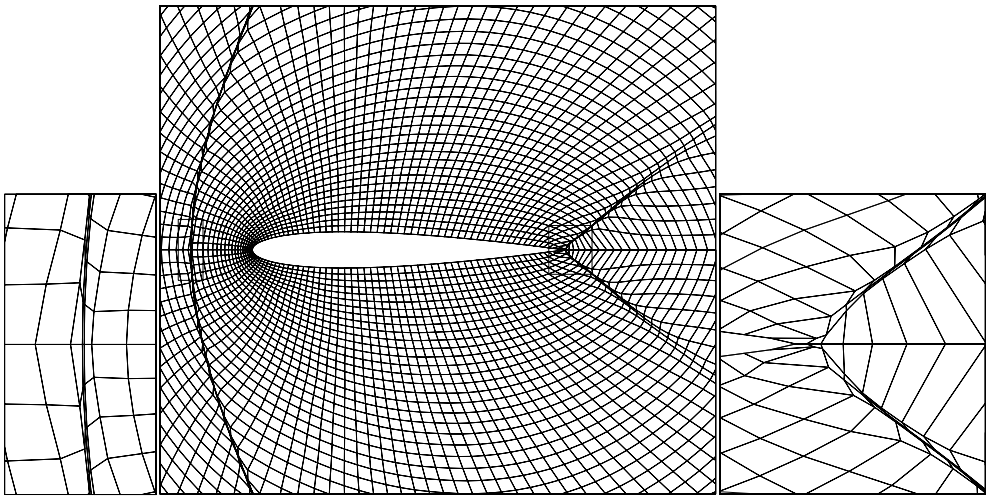


FIG. 15. *Mesh around NACA*0012 *airfoil after adaptation for supersonic flow calculations.*

cells and this is nearly the same as what we would like to have in the ideal case, when the shocks are treated using the Rankine–Hugoniot conditions with a special "shock-fitting" procedure and the subdomains of smooth flow are computed using the numerical scheme. The adaptive procedure, changing the cells width in the shock zone, adjusts automatically the flow solver to those two kinds of computations that increases significantly accuracy of modeling.

Performed in section 5, theoretical analysis shows, in contrast to the boundary layers or shocks smeared in viscous flows, that in the vicinity of the shocks in the hyperbolic problems that the adjacent cells, one located in the shock zone and the other in the domain of smooth solution, with sizes in the direction of normal towards the shock differing by orders of magnitude, do not deteriorate the accuracy of the solution. Therefore, it is not necessary to change the cells' size gradually when passing across the shock. If to apply a numerical scheme using an artificial viscosity, the shock is smeared over several cells and we get a mesh clustered gradually towards the shock. But when decreasing the artificial viscosity and tending the smooth solution to discontinuous, the mesh will come to the above described structure when cell size changes sharply while approaching the shock. From this point of view when computing on the adaptive grids the flow solvers, using piecewise polynomial approximation of the functions, especially Godunov-type schemes, have some advantages over the others based on continuous distribution of the flow parameters.

From the above we can state, probably, that the adapted meshes to the Euler and Navier–Stokes flows must have a different structure and be generated in a different way.

Acknowledgments. The author thanks Dr. Sergey Ivanenko for his attention to the work and fruitful discussions.

REFERENCES

[1] B. N. AZARENOK, *Adaptive moving grids in supersonic flow simulation*, in Proceedings of the 7th International Conference on Numerical Grid Generation in Computational Field

Simulations, B. K. Soni, J. Haeuser, J. F. Thompson, and P. Eiseman, eds., Whistler, BC, Canada, 2000, pp. 629–638.

[2] B. N. Azarenok, *Realization of a second-order Godunov's scheme*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 1031–1052.

[3] B. N. Azarenok, *Adaptive moving grids in problems of gas dynamics*, in Proceedings of the International Conferernce on Optimization of Finite-Element Approximations, Splines and Wavelets, S. A. Ivanenko and V. A. Garanzha, eds., St. Petersburg State University, St. Petersburg, Russia, 2001, pp. 30–44; also available online from http://www.ccas.ru/gridgen/.

[4] B. N. Azarenok and S. A. Ivanenko, *Application of adaptive grids in numerical analysis of time-dependent problems in gas dynamics*, Comput. Math. Math. Phys., 40 (2000), pp. 1330–1349.

[5] B. N. Azarenok and S. A. Ivanenko, *Application of moving adaptive grids for numerical solution of nonstationary problems in gas dynamics*, Internat. J. Numer. Methods Fluids, 39 (2002), pp. 1–22; also available online from http://www.math.ntnu.no/conservation/2001/043.html/.

[6] M. J. Baines, *Moving Finite Elements*, Clarendon Press, Oxford, 1994.

[7] J. U. Brackbill and J. S. Saltzman, *Adaptive zoning for singular problems in two dimensions*, J. Comput. Phys., 46 (1982), pp. 342–368.

[8] W. Cao, W. Huang, and R. D. Russell, *A study of monitor functions for two-dimensional adaptive mesh generation*, SIAM J. Sci. Comput., 20 (1999), pp. 1978–1994.

[9] W. M. Cao, W. Z. Huang, and R. D. Russell, *An r-adaptive finite element method based upon moving mesh PDEs*, J. Comput. Phys., 149 (1999), pp. 221–244.

[10] G. Carey, *Computational Grids: Generation, Adaptation, and Solution Strategies*, Taylor and Francis, Washington, DC, 1997.

[11] A. A. Charakhch'yan and S. A. Ivanenko, *Curvilinear grids of convex quadrilaterals*, Comput. Math. Math. Phys. 28 (1988), pp. 126–133.

[12] A. A. Charakhch'yan and S. A. Ivanenko, *A variational form of the Winslow grid generator*, J. Comput. Phys., 136 (1997), pp. 385–398.

[13] A. S. Dvinsky, *Adaptive grid generation from harmonic maps on Riemannian manifolds*, J. Comput. Phys., 95 (1991), pp. 450–476.

[14] H. A. Dwyer, B. R. Sanders, and F. Raiszadek, *Ignition and flame propagation studies with adaptive numerical grids*, Combustion and Flame, 52 (1984), pp. 11–23.

[15] P. S. Eisman, *Adaptive grid generation*, Comput. Methods Appl. Mech. Engrg., 64 (1987), pp. 321–376.

[16] J. E. Eells and L. Lemaire, *Another report on harmonic mapps*, Bull. London Math. Soc., 20 (1988), p. 387.

[17] F. T. Farrell and L. E. Jones, *Some non-homeomorphic harmonic homotopy equivalences*, Bull. London Math. Soc., 28 (1996), pp. 177–182.

[18] V. A. Garanzha and I. E. Kaporin, *Regularization of the barrier variational grid generation method*, Comput. Math. Math. Phys., 39 (1999), pp. 1489–1503.

[19] S. K. Godunov and V. S. Ryabenkii, *Difference Schemes: An Introduction to the University Theory*, North-Holland, Amsterdam, 1987.

[20] S. K. Godunov, A. V. Zabrodin, M. Ya. Ivanov, A. N. Kraiko, and G. P. Prokopov, *Numerical Solution of Multi-Dimensional Problems in Gas Dynamics*, Izdat, Nauka, Moscow, 1976 (in Russian); S. K. Godunov, A. V. Zabrodin, M. Ya. Ivanov, A. N. Kraiko, and G. P. Prokopov, *Résolution Numérique des Problémes Multidimensionnels de la Dynamique des Gaz*, Mir, Moscow, 1979 (in French).

[21] D. F. Hawken, J. J. Gottlieb, and J. S. Hansen, *Review of some adaptive node-movement techniques in finite-element and finite-difference solutions of partial differential equations*, J. Comput. Phys., 95 (1991), pp. 254–302.

[22] W. Z. Huang, *Practical aspects of formulation and solution of moving mesh PDEs*, J. Comput. Phys., 171 (2001), pp. 753–775.

[23] S. A. Ivanenko, *Adaptive grids and grids on surfaces*, Comput. Math. Math. Phys., 33 (1993), pp. 1179–1193.

[24] S. A. Ivanenko, *Harmonic mappings*, in Handbook of Grid Generation, J. F. Thompson et al., eds., CRC Press, Boca Raton, FL, 1998, Chapter 8.

[25] S. A. Ivanenko, *Optimality principle for nondegenerate grids*, in Proceedings of the International Conferernce on Optimization of Finite-Element Approximations, Splines and Wavelets, S. A. Ivanenko and V. A. Garanzha, eds., St. Petersburg State University, St. Petersburg, Russia, 2001, pp. 85–99; also available online from http://www.ccas.ru/gridgen/.

[26] O.-P. Jacquotte, *Grid optimization methods for quality improvement and adaptation*, in Handbook of Grid Generation, J. F. Thompson et al., eds., CRC Press, Boca Raton, FL,

1998, Chapter 33.
[27] P. Knupp and S. Steinberg, *Fundamentals of Grid Generation*, CRC Press, Boca Raton, FL, 1994.
[28] R. Li, T. Tang, and P. Zhang, *Moving mesh methods in multiple dimensions based on harmonic maps*, J. Comput. Phys., 170 (2001), pp. 562–588.
[29] R. Li, T. Tang, and P. Zhang, *A Moving Mesh Finite Element Algorithm for Singular Problems in Two and Three Dimensions*, preprint, Norwegian University of Science and Technology, 2001; also available online from http://www.math.ntnu.no/conservation/.
[30] F. Liu, S. Ji, and G. Liao, *An adaptive grid method and its application to steady Euler flow calculations*, SIAM J. Sci. Comput., 20 (1998), pp. 811–825.
[31] V. D. Liseikin, *On generation of regular grids on n-dimensional surfaces*, Comput. Math. Math. Phys., 31 (1991), pp. 47–57.
[32] V. D. Liseikin, *Grid Generation Methods*, Springer-Verlag, New York, 1999.
[33] B. L. Rozhdestvenskii and N. N. Janenko, *Systems of Quasilinear Equations and Their Applications to Gas Dynamics*, Translations of Mathematical Monographs 55, AMS, Providence, RI, 1983.
[34] J. S. Sampson, *Some properties and applications of harmonic mappings*, Ann. Sci. École Norm. Sup. (4), 11 (1978), pp. 211–228.
[35] R. Schoen and S. T. Yau, *On univalent harmonic maps between surfaces*, Invent. Math. 44 (1978), pp. 265–278.
[36] S. P. Spekreijse, R. Hagmeijer, and J. M. Boerstoel, *Adaptive grid generation by using Laplace-Beltrami operator on a monitoring surface*, in Proceedings of the 5th International Conference on Numerical Grid Generation in Computational Field Simulations, B. K. Soni, J. Haeuser, J. F. Thompson, P. R. Eiseman, eds., Missisipi State University, MS, 1996, ISGG, pp. 137–146.
[37] J. M. Stockie, J. A. Mackenzie, and R. D. Russell, *A moving mesh method for one-dimensional hyperbolic conservation law*, SIAM J. Sci. Comput., 22 (2001), pp. 1791–1813.
[38] H. Tang and T. Tang, *Moving mesh methods for one- and two-dimensional hyperbolic conservation laws*, SIAM J. Numer. Anal., submitted.
[39] J. F. Thompson, *A survey of dynamically-adaptive grids in the numerical solution of partial differential equations*, Appl. Numer. Math., 1 (1985), pp. 3–27.
[40] A. Winslow, *Numerical solution of the quasi-linear Poisson equation in a nonuniform triangle mesh*, J. Comput. Phys., 1 (1966), pp. 149–172.
[41] P. A. Zegeling, *Moving grid techniques*, in Handbook of Grid Generation, J. F. Thompson et al., eds., CRC Press, Boca Raton, FL, 1998, Chapter 37.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 683–701

ON A NUMERICAL LIAPUNOV–SCHMIDT SPECTRAL METHOD AND ITS APPLICATION TO BIOLOGICAL PATTERN FORMATION*

K. BÖHMER†, C. GEIGER‡, AND J. D. RODRIGUEZ‡

Abstract. Spectral expansions are used to provide a basis which preserves continuous symmetries. We show that spectral methods satisfy the conditions for convergence of numerical Liapunov–Schmidt methods. An explicit algorithm for the calculation of stationary bifurcation scenarios near primary instabilities in general continuous symmetric equations is given. The above convergence is extended to Γ-equivalent discrete and original bifurcation scenarios. The method is applied to a biologically motivated reaction-diffusion system with spherical symmetry forming patterns. A specific singularity of a generic steady state bifurcation is investigated in detail.

Key words. numerical Liapunov–Schmidt methods, spectral methods, symmetry, symmetry breaking, equivariant bifurcation, reaction-diffusion equations, biological pattern formation

AMS subject classifications. Primary, 65P30; Secondary, 35B32, J20, 50,K55, 37G10, 40, M20, N25, 41A10, 42A10

PII. S0036142998339526

1. Introduction.

We discuss numerical Liapunov–Schmidt spectral methods for a symmetric bifurcation problem $G(x) = G(u, \lambda) = 0$ of the form

$$G : \mathcal{D}(G) \subset \mathcal{X} = \mathcal{E} \times \mathbf{R}^q \to \hat{\mathcal{E}}, \qquad \mathcal{E} \subseteq \hat{\mathcal{E}} \text{ are Hilbert spaces,} \tag{1.1}$$

$\hat{\mathcal{E}}$ w.r.t. $\langle \cdot, \cdot \rangle$, and $\mathcal{E}$ w.r.t. a usually stronger norm $\langle \cdot, \cdot \rangle_{\mathcal{E}}$. G transforms equivariantly w.r.t. an (infinite) dimensional representation γ of a symmetry group Γ, i.e., $\gamma G(u, \lambda) = G(\gamma u, \lambda)$, and $\langle \cdot, \cdot \rangle$ is invariant w.r.t. Γ, i.e., $\langle \gamma u, \gamma v \rangle = \langle u, v \rangle \ \forall \ \ \gamma \in \Gamma$, $\forall\ u, v \in \hat{\mathcal{E}}$. We study G near a bifurcation point $x_0 = (u_0, \lambda_0)$ with rank-deficient Jacobian G':

$$G(x_0) = 0,\ G'_0 = G'(x_0) = (G^0_u, G^0_\lambda),\ \dim \mathcal{N}(G'_0) = \mu + q > q. \tag{1.2}$$

Here $\mathcal{N}(G'_0)$ denotes the kernel of G'_0. Let x_0 be a fully symmetric solution of (1.2), i.e., $\gamma x_0 = (\gamma u_0, \lambda_0) = x_0\ \forall \gamma \in \Gamma$. We assume G'_0 to be a Fredholm operator and $\langle \cdot . \cdot \rangle$ to be Γ-invariant; see [27]. For most realistic applications analytical approaches become intractable and discretization methods must be used. However, near the singularity conventional discretization methods fail for several reasons, loss of stability being the most important. Also, consistency inappropriately relates only the original operator to its discretization but not to its derivatives. Finally, discretization methods act as a perturbation and will therefore destroy all generic and nongeneric symmetric bifurcation scenarios; see, e.g., [34, 7, 16, 17, 35, 30, 36, 44, 33, 18]. Comprehensive discussions of general discretization methods in bifurcation theory, using difference methods and finite element methods for both asymmetric and discrete symmetric cases, have been recently provided in [8, 13, 1, 2, 9, 4]. In the presence of continuous

*Received by the editors December 20, 1999; accepted for publication (in revised form) December 6, 2001; published electronically June 26, 2002. Support for the authors was provided by DFG.
http://www.siam.org/journals/sinum/40-2/33952.html

†Fachbereich Mathematik und Informatik, Philipps-Universitat Marburg, Hans-Meerwein-Strasse, 35032 Marburg, Germany (boehmer@mathematik.uni-marburg.de).

‡NTSI-Europe, Einhornstr. 9, 72138 Kirchentellinsfurt, Germany (cgeiger@ntsi.com, jrodriguez@ntsi.com).

symmetries a symmetry preserving discretization can only be obtained with spectral methods. Another attractive feature of spectral methods is their high accuracy. Nice presentations of spectral methods are given in [29, 19, 5].

History of numerical bifurcation. The earlier results on numerical bifurcation theory do not include spectral methods in general. Specifically, de-aliasing techniques, as well as many important operators, are not discussed. The two most advanced approaches are given in [16, 17] and [30]; see [18] as well. In both papers very restrictive conditions have been used: Assuming an application of inverse operators they require $G : \hat{\mathcal{E}} \times \mathbf{R}^q \to \hat{\mathcal{E}}$, $G(u, \lambda) = u + T(u, \lambda)$ with compact T. This is problematic, since analytically equivalent numerical methods may behave very differently. Then, using convergent approximation operators $P^N : \hat{\mathcal{E}} \to \hat{\mathcal{E}}^N \subset \hat{\mathcal{E}}$, they study discretizations of the form

$$G^N(u, \lambda) = u + P^N T(u, \lambda) = 0 \text{ with } P^N T = T P^N. \tag{1.3}$$

These assumptions allow very elegant convergence proofs of discrete bifurcation scenarios. However, the condition (1.3) excludes many important discretization methods as well as many physically relevant operators, e.g., Navier–Stokes equations.

To cover all these problems and to allow a direct discretization of (1.1), (1.2), the *numerical Liapunov–Schmidt method* has been introduced in [8, 13, 4, 15, 9, 10]. Since the bifurcation effects are determined by an interplay of kernels, ranges, and complements of linear operators and nonlinear terms, we have to make sure that the discrete version of the original nonlinear problem indeed reproduces these effects. To achieve this goal we need the concepts of consistent differentiability and the stability of the bordered systems. Consistent differentiability requires consistency for G and its derivatives as well. In [8] stability for bordered systems was proved by referring to [40, 41]. [14] presents an independent proof, which is valid for spectral methods and for Navier–Stokes equations as well.

The purpose of this paper is to show that the numerical Liapunov–Schmidt spectral method reproduces, under physically appropriate conditions, the bifurcation, stability, and symmetry properties of the original problem. For general discretization methods only convergence of the scenarios exists [15]. In section 2 we show that the concepts of consistent differentiability and stability for the bordered systems hold for a numerical Liapunov–Schmidt spectral method based on spectral approximations. We introduce an efficient algorithm for the numerical reduction of bifurcation problems. Explicit expressions for the calculation of n-determined bifurcation equations are given. In section 3 we introduce a three dimensional reaction-diffusion equation posed on a spherical domain which arises in biological pattern formation. This example is very instructive since it models all possible complications: The symmetry group of the model, $O(3)$, is continuous and non-Abelian. It also requires the splitting of the domain into a periodic and a bounded nonperiodic part. Due to the importance of the $O(3)$-symmetry group, the bifurcation structure of a number of instabilities has been previously investigated [21, 26, 25, 24, 28]. Using these results we can completely classify the bifurcation structure in our examples once the bifurcation equations have been derived by the numerical Liapunov–Schmidt algorithm. This demonstrates another strength of combining symmetry preserving reduction, numerical Liapunov–Schmidt methods, and equivariant local bifurcation theory. If a classification for a bifurcation problem with a given symmetry has been obtained once, the results are universal and can be applied to any physical problem with this specific symmetry. We then apply the numerical Liapunov–Schmidt spectral method to the $l = 3$ instability of the model equations in section 4.

2. Numerical Liapunov–Schmidt methods.

2.1. Definitions and notations. Even a brief introduction into spectral methods would be beyond the scope of this paper. Therefore, we refer the reader to the competent presentations of spectral methods given in [29, 19, 5, 23, 42]. An extended version of this section can be found in [11]. We start with the introduction of spectral approximating spaces and projection operators. Let $\mathcal{E} \subset H^n_w(\Omega), \hat{\mathcal{E}} \subset L^2_w(\Omega), \Omega \subset \mathbf{R}^d$; $H^n_w(\Omega)$ are Sobolev spaces w.r.t. a weight function w. We assume a fixed weight function w and, hence, do not indicate it any more. Let

$$\{\varphi_k\}_{k\in\mathbf{Z}^d_0\subset\mathbf{Z}^d} \text{ be a complete orthogonal basis for } \mathcal{E},\ \hat{\mathcal{E}} \tag{2.1}$$

for $\hat{\mathcal{E}}$ w.r.t. $\langle\cdot,\cdot\rangle$, and for $\mathcal{E}$ w.r.t. $\langle\cdot,\cdot\rangle_{\mathcal{E}}$. The $\varphi_k(x)$ are real- or complex-valued functions. We define finite dimensional approximating spaces

$$\mathcal{E}^N = \operatorname{span}\{\varphi_k\}_{k\in\mathbf{K}^N} \subset \mathcal{E} \subseteq \hat{\mathcal{E}} \text{ with } u^N \in \mathcal{E}^N, \tag{2.2}$$

where $k = (k_1, \dots, k_d) \in \mathbf{K}^N$ is a d dimensional multi-index in a finite subset $\mathbf{K}^N \subset \mathbf{Z}^d_0$, and $k \in \mathbf{K}^N$ with $|k_i| \le N_i$ and $N = (N_1, \dots, N_d) \in \mathbf{N}^d_0$, $\hat{N} := |\mathbf{K}^N|$, $\tilde{N} := \min\{N_1, \dots, N_d\}$. The usage of multi-index notation is important since it appropriately reflects the structure of the basis functions $\varphi_k(x)$. This structure is certainly a consequence of the symmetry of (1.1). The structure of $\varphi_k(x)$ and the values of N can be determined using the theory of symmetric spaces. For a very instructive introduction to this theory see [6]. In section 3 we give the explicit form of $\varphi_k(x)$ for $O(3)$-symmetric spaces. Now, every $u \in \mathcal{E}$ (or $\hat{\mathcal{E}}$) is, alternatively, *approximated by truncation* $T^N u$ as

$$T^N u = T^N \left(\sum_{k\in\mathbf{Z}^d_0} \hat{a}_k \varphi_k \right) := \sum_{k\in\mathbf{K}^N} \hat{a}_k \varphi_k \in \mathcal{E}^N$$

or by (unique) *interpolation* in distinct points $y_j \in \Omega$, j are multi-indices, with

$$I^N : \mathcal{E} \to \mathcal{E}^N \text{ unique by } (I^N u - u)|_{y_j} = 0, \qquad j \in \mathbf{J}^N \subset \mathbf{Z}^d, \quad |\mathbf{J}^N| = \hat{N} = |\mathbf{K}^N|.$$

Sometimes, we use the notations $T^N = T(\mathcal{E}^N)$ and $I^N = I(\mathcal{E}^N)$. Now, $\langle\cdot,\cdot\rangle$ is approximated by the (Gaussian) quadrature rule defined as

$$\langle u, v\rangle^N := \sum_{j\in\mathbf{J}^N} u(y_j)\bar{v}(y_j)w_j, \quad (\|u\|^N_0)^2 := \langle u, u\rangle^N. \tag{2.3}$$

$\langle u, v\rangle$, $\langle u, v\rangle^N$ are Γ-invariant, and the operators $T^N, I^N : \mathcal{E}, \hat{\mathcal{E}} \to \mathcal{E}^N$ are Γ-equivariant and are orthogonal projectors w.r.t. $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle^N$; hence the corresponding error functions $T^N u - u$ and $I^N u - u$ satisfy

$$T^N u - u \perp \mathcal{E}^N \text{ and } I^N u - u \perp^N \mathcal{E}^N. \tag{2.4}$$

In general, different symmetry preserving boundary conditions $B_1(u) = 0, B_2(z) = 0$ are imposed on (1.1), (1.2). Therefore, we must replace $\mathcal{E}$ and $\hat{\mathcal{E}}$ by spaces of appropriate ansatz- and test-functions

$$\mathcal{V}^N_1 := \{u \in \mathcal{E}^N : B_1(u) = 0\} \text{ and } \mathcal{V}^N_2 := \{z \in \hat{\mathcal{E}}^N : B_2(z) = 0\}$$

with $\dim \mathcal{V}_1^N = \dim \mathcal{V}_2^N$. Now we define Γ-equivariant orthogonal projectors $P_i^N, \tilde{P}_i^N : \mathcal{E} \to \mathcal{V}_i^N$ such that

$$\langle P_i^N u - u, v^N\rangle,\ \langle \tilde{P}_i^N u - u, v^N\rangle^N = 0 \quad \forall\ v^N \in \mathcal{V}_i^N, \qquad i = 1, 2, \tag{2.5}$$

where $\tilde{P}_i^N, i = 1, 2$, are well defined whenever $u(y_j), v^N(y_j)$ are known. Obviously, for $\mathcal{E}_1^N = \mathcal{V}_1^N, \mathcal{E}_2^N = \mathcal{V}_2^N$ it follows that $P_i^N = T(\mathcal{E}_i^N),\ \tilde{P}_i^N = I(\mathcal{E}_i^N),\ i = 1, 2$. Applying the projections to our bifurcation problem (1.2) we get the following two discrete problems: With $x_0^N = (u_0^N,\ \lambda_0) \in \mathcal{V}_1^N \times \mathbf{R}^q$

$$\text{determine } x_0^N \text{ such that } \ P_2^N G(x_0^N) = 0 \text{ or } G(x_0^N) \perp \mathcal{V}_2^N, \tag{2.6}$$

$$\text{determine } x_0^N \text{ such that } \ \tilde{P}_2^N G(x_0^N) = 0 \text{ or } G(x_0^N) \perp^N \mathcal{V}_2^N. \tag{2.7}$$

Remark. Often $G(u^N)$ is replaced by an approximate operator $\tilde{G}^N(u^N)$ (in particular, if the nonlinear parts of G are evaluated in pseudospectral and collocation methods or with de-aliasing techniques; see, e.g., (2.11), (2.16)). In fact, collocation methods are perturbed Galerkin methods: They are obtained by choosing in (2.5), (2.7) the v^N for $\tilde{P}_2^N$ as the Lagrange basis $v_i^N(y_j) = \delta_{i,j}, i, j \in \mathbf{J}^N$.

The equivariant *interpolation, truncation*, and *projection operators* have the following properties. They hold for smooth $u \in H_w^m(\Omega), 0 \le n \le m$, with $\ell = 0$ and n corresponding to the norms in $\hat{\mathcal{E}}$ and $\mathcal{E}$, respectively,

$$\|T^N u - u\|_\ell \le \|I^N u - u\|_\ell, \|P_i^N u - u\|_\ell, \|\tilde{P}_i^N u - u\|_\ell = \mathcal{O}(\tilde{N}^{-m+\iota^K(\ell)}\|u\|_\ell), \tag{2.8}$$

and $K = F, C, L$ indicate Fourier (or exponential), Chebyshev-, and Legendre-approximations in (2.2), with $\iota^F(n) = n, \iota^C(n) = 2n, \iota^L(n) = 2n + d/2$. If different variables for multivariate u require different approximations, we replace $\iota^K(n)$ by the maximal value. For quadrature approximations with Gauss–Lobatto points we use the equiboundedness of $\|T^{N-\nu}u - u\|_\ell / \|T^N u - u\|_\ell$ for fixed ν, usually $\nu = 1$. Therefore, only the interpolation errors in (2.8) have to be considered.

2.2. Consistent differentiability. To avoid too many technicalities, we assume (1.1) in the form

$$\begin{aligned} G(x) = G(u, \lambda) = L_0 u + \lambda R(u) &= L_0 u + \lambda R_e\left(u, \nabla u, \int_{\Omega_0} u\right) \\ &= L_0 u + \lambda\left(u^2 + \nabla u\left(u + \int_{\Omega_0} u\right) + g\right) \end{aligned} \tag{2.9}$$

with $u_0 = 0$ and a bounded linear operator $L_0 = G_u(x_0)$ and a nonlinear operator R. For the general case see [11]. Now we evaluate (2.9) for spectral and collocation methods indicated by an indices s and c, respectively. In the spectral approach $x_0^N \in \mathcal{V}_1^N \times \mathbf{R}$ are determined as

$$\begin{aligned} \tilde{G}_s^N(x_0^N) :&= P_2^N \tilde{G}^N(x_0^N) = P_2^N T^N G(x_0^N) = 0, \quad \text{or} \\ \tilde{G}_s^N(x_0^N) &= T^N G(x_0^N) \perp \mathcal{V}_2^N. \end{aligned} \tag{2.10}$$

For collocation methods we need point evaluations in (2.9) (see (2.13)):

$$\begin{aligned} G^N(u^N, \lambda)(y_j) &:= (L_0^N u^N)(y_j) + \lambda R^N(u^N)(y_j), \\ L_0^N u^N &:= T^N\ (L_0) u^N = L_0(u^N), \end{aligned} \tag{2.11}$$

$$\begin{aligned} R^N(u^N)(y_j) := &((u^N)^2(y_j) + (\nabla u^N)(y_j)) \\ &\times (u^N(y_j) + Q^N u^N + g(y_j)), \qquad j \in \mathbf{J}^N. \end{aligned} \tag{2.12}$$

In spectral approach terms like $(L_0(u^N))(y_j), (\nabla u^N)(y_j)$ and integrals are not evaluated directly but via some linear approximation operators, e.g., the Fourier collocation derivative (see [19]) or quadrature formulas, $Q^N u^N$; see (2.13). We denote these approximate *linear operators and functionals* (evaluated in y_j) as

$$\begin{aligned} &(L_0^N u^N)(y_j) \approx (L_0 u^N)(y_j), \\ &(L_1^N u^N)(y_j) \approx (\nabla u^N)(y_j), \text{ and } Q^N u^N = \sum_{i \in \mathbf{J}^N} u^N(y_i) w_i, \end{aligned} \tag{2.13}$$

the quadrature approximation. We introduce the restriction operator

$$\rho^N : C(\Omega) \to \mathbf{R}^{\hat{N}}, (\rho^N(u))(y_j) := u(y_j), \quad j \in \mathbf{J}^N, \tag{2.14}$$

where $C(\Omega)$ denotes continuous scalar functions on Ω. Then we re-interpret the approximated $G^N(x^N)$ in (2.11) (see (2.9)) as

$$\begin{aligned} G_c^N(x^N) &:= \rho^N \tilde{G}_c^N(x^N) := \rho^N L_0^N(u^N) + \lambda \rho^N R^N(u^N) \\ &:= \rho^N L_0^N(u^N) + \lambda \rho^N R_e^N(u^N, L_1^N u^N, Q^N u^N). \end{aligned} \tag{2.15}$$

We insert the $L_i^N u^N$, $i = 0, 1$, in (2.13) into R_e. To reveal the structure of $R^N(u^N)$, note that

$$\begin{aligned} \rho^N R^N(u^N) &:= R_e(\rho^N u^N, \rho^N(L_1 u^N), Q^N u^N) \\ &= \rho^N R_e(u^N, L_1^N u^N, Q^N u^N) + \mathcal{O}(\|I^N R(u^N) - R(u^N)\|_0) \\ &= \rho^N R(u^N) + \mathcal{O}(\|I^N R(u^N) - R(u^N)\|_0), \end{aligned} \tag{2.16}$$

and there may possibly be other approximations such as Fourier collocation derivatives or de-aliasing techniques into R to obtain an R_e or even an approximation $R_a \approx R_e$. Corresponding relations hold for the partial derivatives of R, R^N. It is possible to include de-aliasing techniques into this formalism (see [11]) as well. With the equivalent definition of $\tilde{P}_z^N u$ for smooth u via functions and point evaluations (see (2.5)) we formulate the collocation equations as follows: Determine $x_0^N \in \mathcal{V}_1^N \times \mathbf{R}$ such that

$$\tilde{P}_2^N G_c^N(x_0^N) = \tilde{P}_2^N \rho^N \tilde{G}_c^N(x_0^N) = 0 \text{ or } G_c^N(x_0^N) \perp^N \mathcal{V}_2^N. \tag{2.17}$$

THEOREM 2.1. *Let, for the nonlinear operator $G \in C^r(\mathcal{D}(G))$, $G(x_0) = 0$, $\|x_0 - x\|_n$ be small. Let its spectral approximation $G^N = G_s^N$ or G_c^N, with or without de-aliasing (see* (2.11)*), satisfy* (2.16)*. Then the spectral operator G^N is* consistent *and* r-times consistently differentiable *with G, that is, for $j = 1, \dots, r, 0 \le n \le m$,*

$$\begin{aligned} &\|G^N(P_1^N x) - P_2^N G x\|_0 = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)} \|x\|_m), \\ &\|(G^N)^{(j)}(P_1^N x) P_1^N x_1 \cdot \dots \cdot P_1^N x_j - P_2^N G^{(j)}(x) x_1 \cdot \dots \cdot x_j\|_0 \\ &\quad = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)} \|x_1\|_m \cdot \dots \cdot \|x_j\|_m (1 + \|x\|_m)) \end{aligned} \tag{2.18}$$

for $x = (u, \lambda), x_j = (u_j, \mu_j) \in \mathcal{X} = \mathcal{E} \times \mathbf{R}^q$, and $u, u_1, \dots, u_j \in H_w^m(\Omega) \cap \mathcal{E}$. Analogous results hold for the $\tilde{P}_1^N, \tilde{P}_2^N$ and G, G^N combinations. All these operators, derivatives, and $\mathcal{O}$-terms are Σ-equivariant for $u_j \in$ Fix$(\mathcal{E}^\Sigma)$ (see (2.30)*) for $\Sigma = \Gamma$ or a subgroup $\Sigma \subseteq \Gamma$.*

Proof. The estimates (2.8) imply for $u \in H^m_w(\Omega) \cap \mathcal{E}$, similarly in $\hat{\mathcal{E}}$,

$$\|I_N x - x\|_n + \|P_1^N x - x\|_n + \|\tilde{P}_1^N x - x\|_n = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m). \tag{2.19}$$

With $G(u,\lambda) \in H^{m-n}_w(\Omega)$ and (2.9), (2.16), we obtain for both cases

$$\begin{aligned} &\|G_s^N(P_1^N x) - P_2^N G(x)\|_0,\ \|G_c^N(\tilde{P}_1^N x) - \tilde{P}_2^N G(x)\|_0 \\ &\qquad = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m), \end{aligned} \tag{2.20}$$

with all terms Σ-equivariant. The $\tilde{P}_1^N, \tilde{P}_2^N$, and $\tilde{G}^N$ results are obtained here and below in a full analogy.

We prove (2.18)(ii) only for the more complicated case of G_c^N. We have $v_i = x_i = (u_i, \lambda_i)$ or $v_i = u_i$:

$$(L_i v_1)^{(j)}(v_2) = \delta_{1j} L_i v_2, \quad (L_i^N v_1^N)^{(j)}(v_2^N) = \delta_{1j} L_i^N v_2^N \tag{2.21}$$

for $i = 0, 1$ and for $j \geq 1$. (2.9), (2.16) for the partials imply with the partials $\partial_i R$ that

$$\begin{aligned} \big(G_c^N\big)'\big(\tilde{P}_1^N x\big)\tilde{P}_1^N x_1 &= \rho^N(T^N L_0)\tilde{P}_1^N x_1 \\ &\quad +\lambda\rho^N(\partial_1 R^N(\tilde{P}_1^N x), \partial_2 R^N(\tilde{P}_1^N x), \partial_3 R^N(\tilde{P}_1^N x)) \\ &\quad \cdot(\tilde{P}_1^N x_1, L_1^N \tilde{P}_1^N x_1, Q^N \tilde{P}_1^N x_1)^T \\ &= \rho^N(L_0, (\partial_1 R(x), \partial_2 R(x), \partial_3 R(x))(1 + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m))) \\ &\quad \cdot((x_1, L_1 x_1, l x_1) + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x_1\|_m))^T \\ &= \rho^N(G'(x)(1 + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m))) \\ &\quad \cdot(x_1 + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x_1\|_m)) \\ &= \tilde{P}_2^N(G'(x)(1 + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m))) \\ &\quad \cdot(x_1 + \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x_1\|_m)). \end{aligned}$$

So, we have proved (2.18)(ii) for $j = 1$. This holds for G_s^N and the higher derivatives as well. ∎

Consistent differentiability and stability for bordered systems imply convergent bifurcation scenarios for numerical Liapunov–Schmidt methods [15]. In [14] we have proved this stability for a large class of discretization methods and operator equations, including spectral methods for elliptic and Navier–Stokes operators.

2.3. Bifurcation equations and truncated Liapunov–Schmidt methods. Without loss of generality we have distinguished one specific parameter λ as a bifurcation parameter. So, for the remainder of the paper we assume $\lambda \in \mathbf{R}^1$. We want to identify the bifurcation scenario for the bifurcation equation from its truncated discretizations. First we review an iterative Liapunov–Schmidt method that generates the jets (truncated Taylor series) of the bifurcation equation up to the required order. Let $E^\nu_{m,1}$ denote the module of germs [27, 28] at $(0,0)$ of C^∞ vector or matrix-valued functions $\mathbf{R}^m \times \mathbf{R} \to \mathbf{R}^\nu$, over the ring $E^1_{m,1}$. We study Γ-equivariant bifurcation problems

$$\mathcal{F}_\Gamma := \{f \in E^m_{m,1} \ :\ f(0,0) = 0,\ Df(0,0) = 0,\ \gamma f(u,\lambda) = f(\gamma u, \lambda)\ \forall\ \gamma \in \Gamma\}.$$

We identify elements in $f, g \in \mathcal{F}_\Gamma$ that are merely deformations of each other and denote them to be Γ-equivalent (we write $f \overset{\Gamma}{\sim} g$); see [28]. We write $j_k g$ for the truncated Taylor expansion of g w.r.t. all its arguments up to order k (the so called

k-jet). If $j_k g \sim g$, for minimal $k \in \mathbf{N}$, then g is said to be k-determined. We define a pseudonorm

$$\|g\|_k^0 = \sum_{|i|+j\le k} \left| \frac{\partial^{|i|+j} g}{\partial u^i \partial \lambda^j}(0,0) \right|. \tag{2.22}$$

Here we consider $g \in \mathcal{F}_\Gamma$ to be structurally stable, i.e., there exists $\varepsilon > 0$ such that all perturbations $f \in \mathcal{F}_\Gamma$ with $\|f-g\|_k^0 < \varepsilon$ satisfy $f \sim g$. As a consequence Γ must be absolutely irreducibly represented on $\mathbf{R}^m$; see [28]. Structural stability of g guarantees that qualitative results of bifurcation theory can be observed in the underlying physical model.

We consider the 1-parameter Γ-equivariant operator equation (2.9) with the singularity in $x_0 = (0,0)$:

$$G(0,0) = 0,\ G_0' := \partial G(0,0) = (\partial_u G(0,0), \partial_\lambda G(0,0)) =: L =: (\partial_u G_0, \partial_\lambda G_0),$$

G_0' is a Γ-equivariant Fredholm operator of index 1 with a kernel of dimension $m+1$, $m \ge 1$. Note that by the definitions $R(x) = O(\|x\|^2)$. $\mathcal{F}_\Gamma$ excludes turning point bifurcations, i.e., we assume the existence of a trivial solution from which the nontrivial branches bifurcate. Using the Fredholm condition we split into Γ-invariant orthogonal subspaces

$$\mathcal{X} = \mathcal{E} \times \mathbf{R} = \mathcal{N}(G_0') \oplus \operatorname{im}(G_0'^*), \quad \hat{\mathcal{E}} = \mathcal{N}(G_0'^*) \oplus \operatorname{im}(G_0'), \quad \operatorname{im}(G_0'^*) \subset \mathcal{E} \times \{0\},$$

with * always indicating the adjoint operator. We define Γ-equivariant projections (see, e.g., [43, 27])

$$\begin{aligned} &\mathcal{Q} : \mathcal{E} \times \mathbf{R} \to \operatorname{im}(G_0'^*),\ \ \mathcal{N}(\mathcal{Q}) = \mathcal{N}(G_0') = \mathcal{N}(\partial_u G_0) \times \mathbf{R}; \\ &\hat{\mathcal{Q}} : \hat{\mathcal{E}} \to \operatorname{im}(G_0'),\ \mathcal{N}(\hat{\mathcal{Q}}) = \mathcal{N}(G_0'^*),\ \text{e.g., } \hat{\mathcal{Q}}\hat{u} := \hat{u} - \textstyle\sum_{i=1}^{\mu} \langle \hat{\psi}_i, \hat{u} \rangle \hat{\psi}_i. \end{aligned} \tag{2.23}$$

Let $x = x_{\mathcal{N}} + w$ with $x_{\mathcal{N}} \in \mathcal{N}(G_0')$, $w \in \operatorname{im}(G_0'^*)$, and note that $G(x) = 0$ if and only if

$$\hat{\mathcal{Q}} G(x_{\mathcal{N}} + w) = G_0' w + \hat{\mathcal{Q}} R(x_{\mathcal{N}} + w) = 0 \ \text{ and} \tag{2.24}$$

$$(I - \hat{\mathcal{Q}}) G(x_{\mathcal{N}} + w) \ = (I - \hat{\mathcal{Q}}) R(x_{\mathcal{N}} + w) = 0. \tag{2.25}$$

The bordered system (2.24) is uniquely solvable (for small $x_{\mathcal{N}}$) yielding $w(x_{\mathcal{N}}) = w \in \operatorname{im}(G_0'^*)$. This is substituted into (2.25) to give the Γ-equivariant *bifurcation equation*:

$$B(x_{\mathcal{N}}) := (I - \hat{\mathcal{Q}}) G(x_{\mathcal{N}} + w(x_{\mathcal{N}})) = (I - \hat{\mathcal{Q}}) R(x_{\mathcal{N}} + w(x_{\mathcal{N}})) = 0. \tag{2.26}$$

For the following iteration method, theorem, and modification, see [28] and [3].

ALGORITHM 1 (truncated Liapunov–Schmidt method). Let $w_1(x_{\mathcal{N}}) = 0$.

Iteration. For $k = 2, 3, \ldots$ until determinacy do

Define the truncated bifurcation equation of order k,

$$B_k : \mathcal{N}(G_0') \to \mathcal{N}(G_0'^*), \quad B_k(x_{\mathcal{N}}) := (I - \hat{\mathcal{Q}}) j_k R(x_{\mathcal{N}} + w_{k-1}(x_{\mathcal{N}})). \tag{2.27}$$

We usually identify $x_{\mathcal{N}} \in \mathcal{N}(G_0')$ with $(\alpha, \lambda) \in \mathbf{R}^{m+1}$, $\mathcal{N}(G_0'^*)$ with $\mathbf{R}^m$, and maintain the equivariance for α. Generate the next w_k by

$$G_0' w_k(x_{\mathcal{N}}) := \hat{\mathcal{Q}} j_k R(x_{\mathcal{N}} + w_{k-1}(x_{\mathcal{N}})), \qquad w_k \in \operatorname{im}(G_0'^*). \tag{2.28}$$

THEOREM 2.2. *For the iteration defined in* (2.28), *we have*

$$w_k(x_{\mathcal{N}}) = j_k w(x_{\mathcal{N}}), \text{ and } B_k(x_{\mathcal{N}}) = j_k B(x_{\mathcal{N}}).$$

The truncated bifurcation equations (2.27) will transform equivariantly under a finite dimensional representation of the symmetry group of (2.26).

2.4. Equivariant numerical Liapunov–Schmidt methods. To guarantee that the *discretization of* (1.1) *and of all introduced projection operators inherit the* Γ*-equivariance, we need* to reflect this in the discretization method and in $\langle\cdot,\cdot\rangle^N$. In particular, $\mathcal{E}^N$, $\hat{\mathcal{E}}^N$ and $\mathcal{V}_1^N, \mathcal{V}_2^N$ need to be closed under the action of Γ, and $G^N, \tilde{G}^N$, and all the above discrete operators, e.g., T^N, have to be Γ-equivariant, and

$$\begin{aligned} &G^N(\gamma u^N, \lambda) = \gamma G^N(u^N, \lambda) \text{ and, e.g., } T^N \gamma u^N = \gamma T^N u^N, \text{ and} \\ &\langle u^N, v^N\rangle^N = \langle \gamma u^N, \gamma v^N\rangle^N \ \forall \gamma \in \Gamma \text{ and } u^N, v^N \in \mathcal{E}^N. \end{aligned} \tag{2.29}$$

In particular, Γ-invariant approximating subspaces have to be chosen such that (2.29) is satisfied. For spectral methods this is achieved by the finite dimensional bases which provide a representation of, e.g., the spherical symmetries. We assume that the symmetry group acts only in the periodic direction; hence the radial dependence of the functions has no symmetry properties. Therefore, the full problem will be discretized by a product-ansatz with a harmonic basis in the periodic directions and a basis of orthogonal polynomials in the inhomogeneous directions. For other methods, the so-called *symmetry respecting bases* are studied in [22, 2]. Fixed point spaces within $\mathcal{E}^N$ and $\hat{\mathcal{E}}^N$ are defined for subgroups $\Sigma \subseteq \Gamma$ in the usual way:

$$\begin{aligned} &\mathcal{E}^{N,\Sigma} := \mathrm{Fix}^{\Sigma}(\mathcal{E}^N) := \{u^N \in \mathcal{E}^N \ : \ u^N = \sigma u^N \ \forall\ \sigma \in \Sigma\}, \\ &\mathcal{X}^{N,\Sigma} := \mathcal{E}_0^{N,\Sigma} \times \mathbf{R}. \end{aligned} \tag{2.30}$$

For the discrete problem $G^N(x^N)$ on fixed point spaces (2.30) we have

$$G^{N,\Sigma} := G^N|_{\mathcal{X}^{N,\Sigma}} : \mathcal{X}^{N,\Sigma} \to \hat{\mathcal{E}}^{N,\Sigma},$$

G^N and its derivatives (and remainder terms) evaluated at x^N are Σ_{x^N}-equivariant; $\Sigma_{x^N} = \{\gamma \in \Gamma, \gamma x_N = x_N\}$ is the isotropy subgroup of x^N. If G^N is stable or r-times consistently differentiable, the same is true for $G^{N,\Sigma}$.

Throughout we have assumed that we study the singularity of the original problem at $(0,0) = x_0$, i.e., $G(0,0) = G(x_0) = 0$ with

$$\begin{aligned} &\partial_\lambda G(0,0) = 0, \ \mathcal{N}(G_0'^*) = \text{span}\,\{\hat{\psi}_1, \dots, \hat{\psi}_\mu\} \quad \text{and} \\ &\mathcal{N}(G_0') = \mathcal{N}(\partial_u G(0,0)) \times \mathbf{R} = \text{span}\,\{\psi_1, \dots, \psi_\mu\} \times \mathbf{R}. \end{aligned} \tag{2.31}$$

Let

$$G^N(x_0^N) = 0, \quad x_0^N \approx (0,0) \text{ and } x_0^N \in \mathrm{Fix}^{\Gamma}(\mathcal{E}^N) = \mathcal{E}^{N,\Gamma}. \tag{2.32}$$

For a simple bifurcation point of G ($\mu = 1$ in (2.31)) and for bifurcation from a trivial solution it has been shown in [17] that G^N has a bifurcation point in $x_0^N = (0,0) + \mathcal{O}(h^p)$. The general problem is studied in [8, 13, 15]. We introduce

$$\hat{G}^N(x^N) := G^N(x_0^N + x^N), \quad (\hat{G}^N)^{(j)}(x^N) = (G^N)^{(j)}(x^N), \qquad j \geq 1, \tag{2.33}$$

and drop the $\hat{}$ below without violating Theorem 2.1. Thus, without loss of generality, $(0,0)$ is also a singular point of G^N. For higher singularities of G it has been shown in [1] and [8] that for $\partial_\lambda G^N(0,0)=0$ there exist $\psi_i \approx \psi_i^N$ and $\hat{\psi}_i \approx \hat{\psi}_i^N$ such that $\psi_i^N \notin \mathcal{N}(G_0'), \hat{\psi}_i^N \notin \mathcal{N}(G_0'^*)$. These approximations for the kernels of $\partial_u G_0$ and $(\partial_u G_0)^*$ allow the application of the generalized Liapunov–Schmidt methods. Using the equivariance of the bifurcation problem and its discretization we can assume the following generic form:

$$
\begin{aligned}
&\partial_\lambda G^N(0,0)=0,\\
&\mathcal{N}(G_0^{N\prime}) = \mathcal{N}(\partial_{u^N} G^N(0,0)) \times \mathbf{R} = \operatorname{span}\{\psi_1^N,\dots,\psi_\mu^N\}\times\mathbf{R} =: \mathcal{N}^N,\\
&\mathcal{N}(G_0^{N\prime *}) = \operatorname{span}\{\hat{\psi}_1^N,\dots,\hat{\psi}_\mu^N\} =: \hat{\mathcal{N}}^N \quad \text{with again} \qquad (2.34)\\
&\|\psi_i^N - T^N\psi_i\|_n = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|\psi_i\|_m),\\
&\|\hat{\psi}_i^N - \hat{T}^N\hat{\psi}_i\|_n = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|\hat{\psi}_i\|_m), \qquad i=1,\dots,\mu=\dim\mathcal{N}(G_0'^*).
\end{aligned}
$$

We assume Γ-invariant pairings and

$$
\langle T^N u, T^N v\rangle^N = \langle u,v\rangle + \mathcal{O}(N^{-m+\iota^K(n)}\|u\|_m\|v\|_m). \tag{2.35}
$$

With the systematic replacement of the original operators, bilinear forms, projectors, spaces, and functions by their respective discrete version (labeled by N), we translate Algorithm 1 and Theorem 2.2 into their discrete counterparts.

ALGORITHM 2 (numerical truncated Liapunov–Schmidt spectral method).

Iteration. For $k=2,3,\dots$ until determinacy perform the two steps:

$$
B_k^N : \mathcal{N}(G_0^{N\prime}) \to \mathcal{N}(G_0'^*), \quad B_k^N(x_\mathcal{N}^N) := (I-\hat{\mathcal{Q}}^N) j_k R^N(x_\mathcal{N}^N + w_{k-1}^N(x_\mathcal{N}^N)), \tag{2.36}
$$

$$
G_0^{N\prime} w_k^N(x_\mathcal{N}^N) := \hat{\mathcal{Q}}^N j_k R^N(x_\mathcal{N}^N + w_{k-1}^N(x_\mathcal{N}^N)), \qquad w_k^N \in \operatorname{im}(G_0'^*). \tag{2.37}
$$

THEOREM 2.3. *Let G and its discretization G^N satisfy the following (natural) conditions:*

1. *G and G^N are r-times continuously and consistently differentiable (see (2.18)), $r \ge k$, the determinacy of the problem.*
2. *The original and discrete kernels are related via (2.31) and (2.34).*
3. *The original problem is a Γ-equivariant bifurcation problem (see (1.1) and (1.2)) and satisfies (2.31).*
4. *G^N is a spectral, pseudospectral, or collocation method with or without dealiasing; hence it is a Γ-equivariant discretization of G, that is, G^N is based on the Γ-invariant subspaces and pairings (see (2.35)) and Γ-equivariant operators, projectors, and $\mathcal{O}$-terms (see (2.2), (2.16), (2.2)) which are related, e.g., $0 \le n \le m$, by*

$$
T^N\mathcal{Q}x - \mathcal{Q}^N T^N x = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x\|_m) \quad \text{for} \quad x=(u,\lambda),\ u\in H^m(\Omega).
$$

Then, similarly to (2.8), the exact, discrete, and truncated results are related as

$$
\begin{aligned}
&w_k^N(x_\mathcal{N}) = j_k w^N(x_\mathcal{N}) \ \text{ and } \ B_k^N(x_\mathcal{N}) = j_k B^N(x_\mathcal{N}),\\
&\|w_k^N(T^N x_\mathcal{N}) - T^N w_k(x_\mathcal{N})\|_n = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x_\mathcal{N}\|_m), \qquad (2.38)\\
&\|B_k^N(T^N x_\mathcal{N}) - T^N B_k(x_\mathcal{N})\|_k^0 = \mathcal{O}(\tilde{N}^{-m+\iota^K(n)}\|x_\mathcal{N}\|_m),
\end{aligned}
$$

with Γ*-equivariant* $B_k^N(T^N x_{\mathcal{N}}), B_k(x_{\mathcal{N}})$ *determining* Γ*-equivalent bifurcation scenarios, since* B_k *and* B_k^N *are elements in* $\mathcal{F}_\Gamma$*. As norms for* w_k^N *we use* (2.37)*; for* B_k^N *we use* (2.22)*. Without the* Γ*-invariant and equivariant conditions,* $\partial_\lambda G^N(0,0) = 0$, (2.34)*, and* (2.22)*, we still obtain convergence of the discrete to the exact bifurcation scenarios.*

Proof. Since all the error-terms $\mathcal{O}$ are Γ-equivariant, the bordered system is stable, the discretization is r-times consistently differentiable, and the problem is structurally stable w.r.t. the seminorm (2.22), the results of this theorem are immediate consequences of the given conditions. ∎

The validity of the stability assumption for the number of bordering conditions can be monitored by comparing results for different values of N.

ALGORITHM 3 (symmetric Liapunov–Schmidt method).

Initiation. For harmonic spectral expansions the harmonic component of the critical eigenvector consists of a reducible representation of the symmetry group Γ. Let the dimension of the critical eigenspace be N_0; then the spectral space $\mathcal{E}^{N_0}$ is of dimension N_0 with maximum wave number l_{N_0}. The nonlinear interactions in the k-jet $j_k R^N(u)$ generate wave numbers of order kl_{N_0}.

Iteration. Remove aliasing errors by padded transforms of the nonlinear operator R (see [19, 11]), where $N_1 \geq (k+1)l_{N_0}/2-1$ for one dimensional trigonometric Fourier transforms and $N_{lon,1} \geq (2k+1)l_{N_0}+1$, $N_{lat,1} \geq \big((2k+1)l_{N_0}+1\big)/2$ for spherical harmonic transforms. Remove the truncation error by the projection $\hat{T}^{N_1}, \hat{T}^{N_{k-1}}$; then

$$j_2 R^N(u) : \mathcal{E}^{N_0} \to \mathcal{E}^{N_1}, \quad j_k R^N(u) : \mathcal{E}^{N_{k-2}} \to \mathcal{E}^{N_{k-1}}.$$

Repeat until determinacy. Thereby one generates the sequence of approximating spaces $\mathcal{E}^{N_m}$ with $\mathcal{E}^{N_0} \subset \mathcal{E}^{N_1} \subset \cdots \subset \mathcal{E}^{N_m}$.

3. Application to a reaction-diffusion system in biology.

3.1. The model equations. We consider a coupled system of nonlinear reaction-diffusion equations. The specific model we chose was originally suggested in [39]:

$$\frac{\partial c_1}{\partial t} = D_1 \nabla^2 c_1 + 1 - c_1 c_2^\alpha, \tag{3.1}$$

$$\frac{\partial c_2}{\partial t} = D_2 \nabla^2 c_2 + \beta\big(c_1 c_2^\alpha - c_2\big). \tag{3.2}$$

It is defined on a spherical domain in $\mathbf{R}^3$ of radius R. We abbreviate it in the form

$$\frac{\partial c}{\partial t} = G(c,\beta) := \tilde{L}(\beta)c + \tilde{N}(c,\beta), \quad \frac{\partial c}{\partial r} = 0 \text{ at } r = R \text{ and bounded at } r = 0, \tag{3.3}$$

with a two-component mixture $c = \{c_1, c_2\}$. These problems arise in the context of mathematical biology, e.g., in Turing's theory of pattern formation [31]. In the specific model [39] we choose, with α and λ related as in subsection 3.3,

$$\tilde{L}(\beta) = \begin{pmatrix} D_1\nabla^2 - 1 & -\alpha \\ \beta & D_2\nabla^2 + \beta(\alpha - 1) \end{pmatrix} \quad \text{with } \tilde{L} := \tilde{L}(\beta_c),$$

$$G_0' = (\tilde{L}, 0),\ \mathcal{N}(G_0') = \mathcal{N}(\tilde{L}) \times \mathbf{R}, \tag{3.4}$$

satisfying (2.31), and the nonlinear part is $\tilde{N}(c,\beta) = (-c_1 c_2^\alpha, \beta c_1 c_2^\alpha)^T$. Following [39] we assume a fixed effective Hill constant $\alpha = 3$ and the diffusion constants as

$D_1 = 0.15$ and $D_2 = 0.015$. The distinguished bifurcation parameter β represents a rate constant measuring the level of enzyme activity. The isotropy of the model induces the $O(3)$-equivariance of the model equations. We discuss stationary solutions.

3.2. Computation of the linearized eigenvalue problem. The concentration vector yields the homogeneous solution $c_1 = c_2 = 1$ for arbitrary β, $G_0 = G(c_1 = c_2 = 1, \beta) = 0$. With $\partial_\lambda G_0 = 0$, $\mathcal{N}(G_0')$ satisfies (2.31). For an extended study, see [12]. A linear stability calculation shows that $c_1 = c_2 = 1$ is stable for small β. Instabilities or bifurcation occur whenever an eigenvalue σ of $\tilde{L}v = \sigma v$ crosses the imaginary axis. We solve the eigenvalue problem with boundary conditions from (3.3) for the $f_l(r)$ and $f_l^N(r)$ below, via

$$v = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_l(r) Y_{lm}(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} (f_{l1}(r), f_{l2}(r))^T Y_{lm}(\theta, \phi) \in \mathcal{E}. \tag{3.5}$$

With $\nabla^2 Y_{lm} = -l(l+1) Y_{lm}$, $(Y_{km}, Y_{ln})_2 = \delta_{kl} \cdot \delta_{mn}$, the linear eigenvalue problem $\tilde{L}v = \sigma v$ can be split and reduced to the radial eigenvalue problem

$$\tilde{L} f_l Y_{lm} = Y_{lm} \tilde{L}_l f_l = Y_{lm} \sigma_l f_l, \text{ and reduced to } \tilde{L}_l f_l = \sigma_l f_l \text{ for fixed } l; \tag{3.6}$$

here σ_l is an eigenvalue of multiplicity $2l+1$ and 1 for $\tilde{L}$ and $\tilde{L}_l$, respectively, and

$$d_l^2 = \frac{d^2}{dr^2} + \frac{2d}{rdr} - \frac{l(l+1)}{r^2} \text{ replaces } \nabla^2 \text{ in (3.4) to define } \tilde{L}_l. \tag{3.7}$$

(3.6) may be solved exactly in terms of spherical Bessel functions [20], or numerically by a Chebyshev-tau or Chebyshev-collocation method with the Gauss–Radau points applied to the ansatz $f_l^N(r) = \sum_{n=0}^{N-1} \hat{f}_n \; T_n(r), \hat{f}_n \in \mathbf{R}^2$ for the $\hat{f}_n \in \mathbf{R}^2$. Mind that this transition from $f_l(r)$ to $f_l^N(r)$ causes the discretization errors discussed below. The critical or stability curves $\beta = \beta_l(R)$ in the (β, R) plane for each value of the wave number l satisfy $\Re\{\sigma_l\} = 0 = \sigma_l$, since for the parameter range of our interest the eigenvalues are always real. The minimal β_l selects the critical wave number l_0 and determines the critical rate constant $\beta_c(R) = \min_l \beta_l(R)$. The critical curves $\beta(R)$ are shown in Figure 3.1 for the $l = 1, 2$, and 3 spherical harmonics. Generically, there is a unique l for each R. However, when the critical value of β_c occurs at the intersection point of two stability curves, e.g., for $l_{0,1} = 1$ and $l_{0,2} = 2$ in $R \approx 0.83$, the dimension of the kernel is equal to $2(l_{0,1} + l_{0,2} + 1)$. Moreover, two radial eigenvectors corresponding to a single irreducible representation, l_0, may be simultaneously unstable with a resulting kernel of dimension $4l_0 + 2$. For the parameter values in Figure 3.1 that does not appear at criticality.

3.3. Application of the Liapunov–Schmidt reduction. For the Liapunov–Schmidt procedure and a given shell radius R the minimal critical curve occurs generically for a single $l = l_0$; see section 4 for $l = l_0 = 3$. Redefining $\lambda := \beta - \beta_c$, $u = (u_1, u_2)^T := c - (1,1)^T$ the critical value is transformed to $\lambda_c = \lambda = 0$ with $\Re\{\sigma_{l_0}\} = 0 = \sigma_{l_0}$. This allows the standard form for (3.3) as

$$G : \mathcal{X} = \mathcal{E} \times \mathbf{R} \to \hat{\mathcal{E}}, \; G(u, \lambda) = \tilde{L}u + R(u, \lambda), \; G_0' = (\tilde{L}, 0), \; \tilde{L} = \partial_u G(0, 0); \tag{3.8}$$

$\tilde{L}$ is a Fredholm operator of index 0, $R(\cdot, \cdot)$ is the nonlinear operator

$$R(u, \lambda) = \begin{pmatrix} -\alpha u_1 u_2 - (u_1 + 1)\mathcal{R}(u_2) \\ (\lambda + \beta_c)(\alpha u_1 u_2 + (u_1 + 1)\mathcal{R}(u_2)) + \lambda(u_1 + (\alpha - 1)u_2) \end{pmatrix}, \tag{3.9}$$

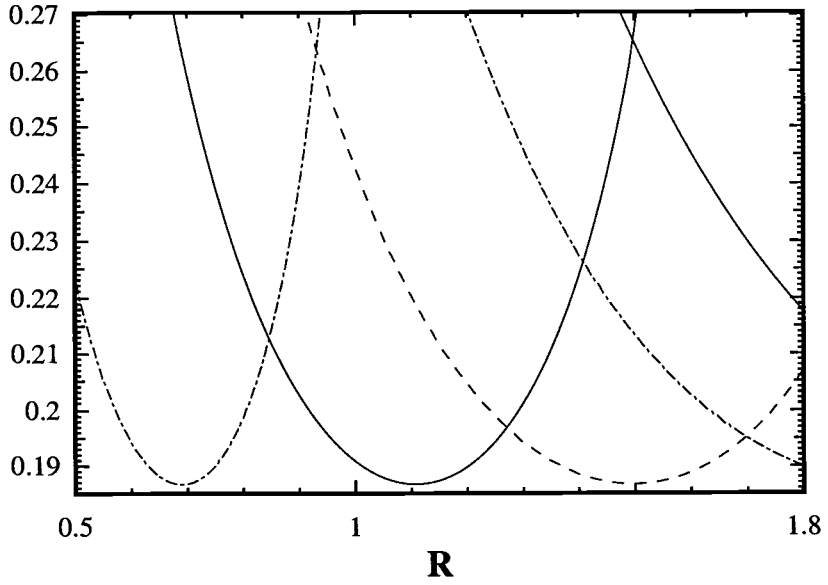


FIG. 3.1. *Linear stability diagram, β versus R, for $l = 1, 2$, and $= 3$ (dot-dashed, solid, and dashed lines, resp.). For $l = 1, 2$ two dot-dashed and solid lines are shown for different radial modes.*

and α is an integer. The Taylor expansions for $\mathcal{R}(u_2)$ and $R(u, \lambda)$ are given as

$$\mathcal{R}(u_2) \equiv \sum_{k=2}^{\alpha} \frac{\alpha!}{k!(\alpha-k)!} u_2^k, \quad R(u,\lambda) = \sum_{i=1} \sum_{j=0} R_{ij}(u)\lambda^j, \tag{3.10}$$

with $R_{ij}(u)$ an i-linear operator in u and $R_{10} = 0$; see (3.9) and (3.10). We restrict ourselves to steady state solutions of (3.8) and obtain

$$\begin{aligned} \mathcal{N} \equiv \mathcal{N}\big(\partial_u G(0,0)\big) = \mathcal{N}(\tilde{L}) = \operatorname{span}\{\psi_{l_0 -l_0}, \dots, \psi_{l_0 l_0}\} \\ = \operatorname{span}\{f^0_{l_0} Y_{l_0 -l_0}, \dots, f^0_{l_0} Y_{l_0 l_0}\}, \quad \mathcal{N}\big(G_0'\big) = \mathcal{N} \times \mathbf{R}, \quad \tilde{L}_{l_0} f^0_{l_0} = 0, \end{aligned} \tag{3.11}$$

$f^0_{l_0} = (f^0_{l_0 1}, f^0_{l_0 2})^T$, and $\dim \mathcal{N}(\tilde{L}) = \mu = 2l_0 + 1$.

Bifurcating branches may then be determined by the truncated Liapunov–Schmidt reduction. For $v \in \mathcal{N} = \mathcal{N}(\tilde{L}) = \mathcal{N}(\tilde{L}^*)$ (see (3.4)) we have

$$\begin{aligned} v = z\psi = (v_j)_{j=1}^2 = \left(\sum_{m=-l_0}^{l_0} z_m f^0_{l_0 j}(r) Y_{l_0 m}(\vartheta, \phi) \right)_{j=1}^2 \\ \in \mathcal{N} = \mathcal{N}(\tilde{L}) = \mathcal{N}(\tilde{L}^*). \end{aligned} \tag{3.12}$$

$z_m \in \mathbf{C}$, $z = (z_{-l_0}, \dots, z_{l_0})^T \in \mathbf{C}^{2l_0+1}$ satisfy the reality condition $z_{-m} = (-1)^m \bar{z}_m$.

Since $\tilde{L} = \tilde{L}^*$ is self-adjoint, the above algorithms can be simplified: We split a solution $(u, \lambda) \in \mathcal{E} \times \mathbf{R}$ of (3.8) into $u = v + w$, with $v \in \mathcal{N}(\tilde{L}) = \mathcal{N}$ and $w \in \mathcal{M} \equiv \operatorname{im}(G_0'^*) = \operatorname{im}(\tilde{L}^*) = \operatorname{im}(\tilde{L})$, i.e., $\mathcal{M}$ is the orthogonal complement of $\mathcal{N}$ in (3.11). In (2.23) we had defined the Γ-equivariant projectors $\mathcal{Q}, \hat{\mathcal{Q}}$. Using the same operator $\mathcal{P}$

in $\mathcal{E}$ and $\hat{\mathcal{E}}$, we want to slightly change the notation into

$$\mathcal{P} : \hat{\mathcal{E}} \to \mathcal{N} = \mathcal{N}(\tilde{L}) = \mathcal{N}(\tilde{L}^*) \text{ and hence } \mathcal{Q} \equiv I - \mathcal{P} = \hat{\mathcal{Q}} : \hat{\mathcal{E}} \to \mathcal{M}, \tag{3.13}$$

with complementary orthogonal projectors $\mathcal{P}, \mathcal{Q}$, and for $u \in \mathcal{E}$ or $u \in \hat{\mathcal{E}}$

$$\mathcal{P}u = \sum_{m=-l_0}^{l_0} \langle u, \psi_{l_0 m} \rangle \psi_{l_0 m}, \quad \mathcal{Q}u = (I - \mathcal{P})u = u - \mathcal{P}u, \quad \mathcal{P}(u, \lambda) = (\mathcal{P}u, \lambda). \tag{3.14}$$

The $O(3)$-symmetry of (3.8) leads to the equivariance of G w.r.t. an infinite dimensional representation T_γ, i.e., $T_\gamma G(u, \lambda) = G(T_\gamma u, \lambda)\ \forall \gamma \in O(3)$. T_γ induces the finite dimensional irreducible representation $T_{l_0,\gamma}$ on the five dimensional subspace $\mathcal{N}$ (see (3.12)) spanned by the critical spherical harmonics for l_0 via

$$T_\gamma v = T_\gamma \sum_{m=-l_0}^{l_0} z_m f_{l_0}^0 Y_{l_0 m} = T_\gamma z\psi = \sum_{m=-l_0}^{l_0} z_m f_{l_0}^0 T_\gamma Y_{l_0 m} = \sum_{m=-l_0}^{l_0} (T_{l_0,\gamma} z)_m f_{l_0}^0 Y_{l_0 m}.$$

This implies the relation $T_\gamma z\psi = (T_{l_0,\gamma} z, \psi)$ for an element $z\psi = v \in \mathcal{N}$. It also follows that $T_\gamma \mathcal{P}u = \mathcal{P}T_\gamma u$ and $T_\gamma \mathcal{Q}u = \mathcal{Q}T_\gamma u$. The bifurcation equations (2.26) are obtained by projecting (3.8) with $\mathcal{P}$ ($\mathcal{P}^*$ for the non-self-adjoint case) onto $\mathcal{N}$ to get

$$B(z, \lambda) = \mathcal{P}R(z\psi + w(z, \lambda)). \tag{3.15}$$

The successive approximations to the bifurcation equations are calculated as a Taylor series in z and in λ until determinacy. From (2.27) and (3.10) we get

$$\begin{aligned} B_k(z, \lambda) &= \mathcal{P} j_k R(z\psi + w_{k-1}(z\psi, \lambda), \lambda) \\ &= \mathcal{P} j_k \sum_{i=2} R_{i0}(z\psi + w_{k-1}(z, \lambda), \lambda) + \lambda \mathcal{P} j_{k-1} \sum_{i=1} R_{i1}(z\psi + w_{k-1}(z, \lambda), \lambda). \end{aligned} \tag{3.16}$$

Using the theory of invariants, it can be shown that the equivariance property permits the series expansion (3.16) to be decomposed at each order i in z into a finite set of m_l^i-equivariant homogeneous polynomials $Z_{lm}^{ij}(z)$ of degree i which are independent over $\mathbf{R}$ $(1 \le j \le m_l^i)$ [21, 28]. The polynomials $Z_{lm}^{ij}(z)$ have the equivariance property that $T_{l,\gamma} Z_{lm}^{ik}(z) = Z_l^{ik}(T_{l_0,\gamma} z)$ where $T_{l,\gamma}\ l \in \mathbf{N}_0$ are the irreducible representations of $O(3)$ where l is the order of the spherical harmonics Y_{lm}. This will be exemplified in section 4.1 for second order terms (cf. (4.4)). Since m_l^i is small for lower orders this results in a dramatic simplification of the Taylor series. Since in previous investigations a complete classification of all the generic singularities with $l < 5$ has been achieved [21, 24], we can directly apply these results to our examples. This is because a given singularity is independent from its physical origin, i.e., it depends only on the specific degeneracy, the symmetry group, and the representation of the symmetry group. Then a local bifurcation analysis yields three important pieces of information: the determinacy of the bifurcation problem and the corresponding normal form, the universal unfolding of the normal form, and the local solution structure of the bifurcation equations. Consequently, from the results of [21, 24] we know not only the determinacy of the problem but also which of the terms of the Taylor expansion in the bifurcation equations need to be calculated. Furthermore, once we calculate the required terms, we can apply the results of the local analysis again in order to determine the qualitative branching behavior. In cases in which a complete classification of the local branching behavior of a given singularity does not exist, the determinacy of the problem can be also determined numerically [15, 38].

4. Generic bifurcations in the $l = 3$ representation.

4.1. Calculation of the bifurcation equations. If we fix the radius of the sphere to $R = 1.55$, then in (cf. section 3.2) the minimal critical curve occurs for $l_0 = 3$ and therefore $\dim\mathcal{N} = 5$ and $z = (z_{-3}, \dots, z_3)^T$, $B_k = (B_{k,-3}, \dots, B_{k,3})^T \in \mathcal{N}$. Because generic bifurcations with $l = 3$ critical are three-determined [28] we must compute only up to third order terms.

Following section 3 we derive the bifurcation equation near critical rate constant. For odd values of l the action of the reflective component of $O(3)$ is nontrivial and the equivariants of even order vanish identically. Therefore we can immediately turn to the calculation of the third order terms. Note that this requires the computation of the function $w_2 = w_2(v, \lambda) = w_2(z\psi, \lambda)$ (cf. Algorithms 1–3 and (2.28) ff.), which we start now:

$$G_0' w_2(z\psi, \lambda) = \mathcal{Q} j_2 R(z\psi, \lambda), \quad j_2 R(z\psi, \lambda) = \lambda R_{11}(z\psi) + R_{20}(z\psi). \tag{4.1}$$

The term R_{11} will induce a term of order $\mathcal{O}(\lambda\|z\|^2)$ and $\mathcal{O}(\lambda^2\|z\|)$ in the bifurcation equations. From the classification in [21, 24] we know that these terms do not appear in the normal form and therefore do not effect the local branching behavior. Therefore we use (3.9), (3.10), (3.12) and simplify

$$j_2 R(z\psi, \lambda) = R_{20}(v = z\psi) = q \sum_{m', m''=-3}^{3} z_{m'} z_{m''} Y_{3m'} Y_{3m''}, \tag{4.2}$$

with $q = q(r) := (-1, \beta_c)^T \alpha (f_{31}^0(r) f_{32}^0(r) + \frac{\alpha-1}{2}(f_{32}^0(r))^2)$. With

$$c_{m'm''m}^{l'l''l} := \int_0^{2\pi} \int_0^{\pi} Y_{l'm'} Y_{l''m''} \bar{Y}_{lm} \, \sin\theta d\theta \, d\phi$$

we get

$$\mathcal{Q} R_{20}(z\psi) = R_{20}(z\psi) - \int_0^R q(r) f_3^0 dr \sum_{m=3}^{3} \left(\sum_{m', m''=-3}^{3} c_{m'm''m}^{333} z_{m'} z_{m''} \right) Y_{3m} = R_{20}(z\psi),$$

since the sum over the $c_{m'm''m}^{333} z_{m'} z_{m''}$ vanishes for all $m = -3, \dots, 3$. From (4.1) we get the reduced equation

$$\begin{aligned} G_0' w_2(z\psi) &= q \sum_{m'm''=-3}^{3} z_{m'} z_{m''} Y_{3m'} Y_{3m''} = q \sum_{l=0}^{\infty} \sum_{-l}^{l} Z_{lm}^{21} \\ &= q \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{m'm''=-3}^{3} c_{m'm''m}^{333} z_{m'} z_{m''} Y_{lm} = q \sum_{l=0,2,4,6} \sum_{-l}^{l} Z_{lm}^{21}. \end{aligned} \tag{4.3}$$

The linearity and $O(3)$-equivariance of (4.3) w_2 admits a product-ansatz

$$w_2(r, \theta, \phi) = \sum_{l=0,2,4,6} f_l^{21}(r) \sum_{m=-l}^{l} Z_{lm}^{21} Y_{lm}(\theta, \phi) \perp \mathcal{N}. \tag{4.4}$$

With this ansatz and (4.3) we get a differential equation (see (3.6), (3.7)),

$$G_0' w_2 = \tilde{L} w_2(r, \theta, \phi) = \sum_{l=0,2,4,6} (\tilde{L}_l f_l^{21}) \sum_{m=-l}^{l} Z_{lm}^{21} Y_{lm} = q(r) \sum_{l=0,2,4,6} \sum_{m=-l}^{l} Z_{lm}^{21} Y_{lm}.$$

Hence we solve instead the better uniquely solvable equations for $l = 0, 2, 4, 6$.

$$\tilde{L}_l f_l^{21}(r) = q(r), \quad l = 0, 2, 4, 6, \text{ where } f_l^{21} = (f_{l1}^{21} f_{l2}^{21})^T, \tag{4.5}$$

for the above boundary conditions; see (3.3). The solutions f_l^{21} depend only on l, r and not on m. The solution w_2 is a second order polynomial in z. The projection of (3.8) onto $\mathcal{N}$ finally yields the bifurcation equations. They are given to order $\mathcal{O}(|z|^3)$, $\mathcal{O}(|z|\lambda)$, and $\mathcal{O}(|z|^2\lambda)$ (see (3.9), (3.10), (3.13), (3.14)) as

$$\begin{aligned}
B_3(z,\lambda) &= \lambda \mathcal{P} R_{11}(z\psi) + \mathcal{P} R_{20}(z\psi + w_2(z\psi,\lambda)) + \mathcal{P} R_{30}(z\psi),\\
B_{3,m}(z,\lambda) &= B_{2,m}(z,\lambda) + \langle R_{30}(z\psi), \psi_m\rangle\\
&= \lambda \langle R_{11}(z\psi), \psi_m\rangle + \langle R_{20}(z\psi + w_2(z\psi,\lambda)), \psi_m\rangle + \langle R_{30}(z\psi), \psi_m\rangle\\
&= \lambda b_1 z_m + b_2 \sum_{m'm''=-3}^{3} c^{333}_{m'm''m} z_{m'} z_{m''}\\
&\quad + \sum_{l'} \tilde{b}_{3l'} \sum_{m'm''=-3}^{3} c^{l'33}_{m'm''m} Z^{21}_{l'm'} z_{m''}\\
&\quad + \tilde{b}_{30} \sum_{m'm''m'''=-3}^{3} z_{m'} z_{m''} z_{m'''} \int_0^{\pi}\int_0^{2\pi} Y_{3m'} Y_{3m''} Y_{3m'''} \bar{Y}_{3m} \sin\theta d\theta\, d\phi,
\end{aligned} \tag{4.6}$$

with

$$b_2 = \int_0^R \alpha\big(f_{31}^0 f_{32}^0 + (\alpha-1)(f_{32}^0)^2/2\big)\big(\beta_c f_{32}^0 - f_{31}^0\big)\, r^2 dr\,,$$

$$\tilde{b}_{3l} = \int_0^R \frac{\alpha}{2}\big(f_{l1}^{21} f_{32}^0 + f_{l2}^{21} f_{31}^0 + (\alpha-1) f_{l2}^{21} f_{32}^0\big)\big(\beta_c f_{32}^0 - f_{31}^0\big)\, r^2 dr\,,$$

$$\tilde{b}_{30} = \int_0^R \frac{\alpha(\alpha-1)}{2}\big(f_{31}^0 (f_{32}^0)^2 + (\alpha-2)(f_{32}^0)^3/3\big)\big(\beta_c f_{32}^0 - f_{31}^0\big)\, r^2 dr.$$

For the calculation at third order there are 7^3 terms per component of z and thus 7^3 undetermined coefficients. Using invariant theory, it can be shown that all coefficients can be expressed in terms of the two third order coefficients b_{31} and b_{32}. More precisely, the three-jet of the bifurcation equations B_3 can then be decomposed into only two equivariant polynomials. These two polynomials are given by linear combinations of the third order terms in (4.6).

4.2. Approximation errors. The approximation errors for the N-term Chebyshev spectral approximant yielding the critical rate constant β_c^N and the coefficient b_1^N are shown in Table 4.1. The converged values $\beta_c^N = \beta_c^\infty, b_1 = b_1^\infty$ (with errors $\approx 10^{-12}$) occur for $N - 1 = 18$. The numerical computation of the radial function is realized through the approximant B_3^N. Here N refers to the $(N+1)$-term Chebyshev expansion in the radial direction. The dependence of the linear operator on the angular variables (θ, ϕ) is eliminated by a discretization to a collocation grid chosen to eliminate aliasing error and is therefore exact. For cubic polynomials the elimination of aliased terms requires that $N_{lon} \geq 4l_0 + 1 = 13$ and $N_{lat} \geq (4l_0 + 1)/2 = 13/2$. We therefore choose $N_{lon} = 16$ and $N_{lat} = 8$ to allow the use of the FFT. The integrals

TABLE 4.1

Relative approximation error for the critical rate constant β_c and linear coefficient b_1, where $b_1 = 1.5096774921862$, $\beta_c = 0.1871719916317956$, and $R = 1.55$.

N	$(\beta_c^N - \beta_c)/\beta_c$	$(b_1^N - b_1)/b_1$
4	$-9.1957157200481\times10^{-4}$	$2.5571536421600\times10^{-2}$
6	$2.1262913035586\times10^{-5}$	$-3.2373852614000\times10^{-3}$
8	$-4.1598759731309\times10^{-7}$	$-1.6182126620001\times10^{-4}$
10	$5.1089623920308\times10^{-9}$	$9.4810774999843\times10^{-6}$
12	$-4.2916614706456\times10^{-11}$	$-3.9807099994071\times10^{-7}$

TABLE 4.2

Relative approximation error of the coefficients b_{31} and b_{32}. Converged values are $b_{31} = 1.0779308458571$ and $b_{32} = 1.6624854128710$.

N	$(b_{31}^N - b_{31})/b_{31}$	$(b_{32}^N - b_{32})/b_{32}$
8	$2.4447858094000\times10^{-2}$	$3.6081193658000\times10^{-2}$
10	$3.3783372601999\times10^{-3}$	$6.7765473284001\times10^{-3}$
12	$8.9734578699874\times10^{-5}$	$3.7732165149995\times10^{-4}$
14	$-2.0430003599969\times10^{-5}$	$-7.4240502500045\times10^{-5}$
16	$7.4269237000379\times10^{-6}$	$1.7987259500085\times10^{-5}$
18	$-1.4799865999393\times10^{-6}$	$-3.1605985999494\times10^{-6}$
20	$2.1271289996072\times10^{-7}$	$4.3438650010330\times10^{-7}$
22	$-2.3774499924301\times10^{-8}$	$-4.7782700018928\times10^{-8}$
24	$2.1693999929795\times10^{-9}$	$4.3290000473206\times10^{-9}$

defining the projection operator $\mathcal{P}$ are then approximated as summations over the collocation grid points. The resulting third order coefficients b_{31} and b_{32} and the approximation error for increasing numbers of terms N of the Chebyshev expansion are shown in Table 4.2. Since our computations have been designed to preserve equivariance we are thus assured by the arguments presented in section 3 that a one-to-one

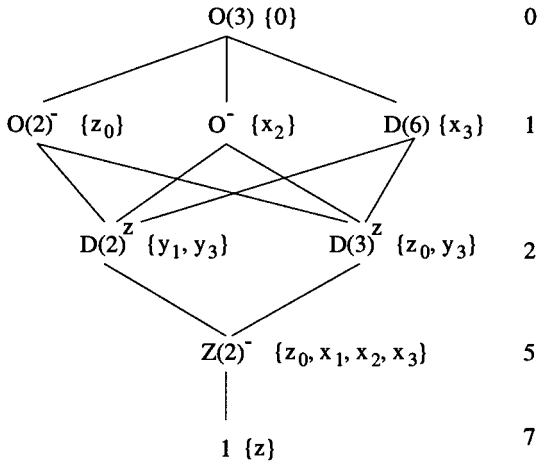


FIG. 4.1. *Lattice of isotropy subgroups for the $l = 3$ representation: For each isotropy subgroup a simple representative fix point subspace is given. The numbers at the right denote the dimension of the subspaces of a given hierarchy. The lines indicate which symmetries of a lower hierarchy are included in the symmetries of a larger hierarchy.*

correspondence exists between the discretized solutions $x^N \in \mathcal{E}^N \times \mathbf{R}$ and solutions $x \in \mathcal{E} \times \mathbf{R}$ with $\|x - x^N\| = \mathcal{O}(N^{-m+n}\|x\|_m)$.

4.3. Solution branches. Given the computation of the Taylor series coefficients the bifurcation equations are completely specified:

$$B_{3,m}(z, \lambda) = 0, \quad m = -3, \ldots, 3 \tag{4.7}$$

with $B_{3,m}$ given in (4.6). Stability and symmetry of bifurcating solution branches of (4.7) have been completely classified in [21]. Nevertheless, in order to apply the results of the Liapunov–Schmidt reduction we must present the result of [21] in an explicit form suitable to our analysis. The solutions of (4.7) can be ordered by their symmetries of the lattice of isotropy subgroup (cf. Figure 4.1). For an exact definition for the isotropy subgroups groups of Figure 4.1, c.f. [32, 28]. In [21] it is shown that in the generic case there are only solutions with symmetries of the maximal isotropy subgroups. We now restrict $B_{3,m}(z, \lambda)$ to each of the fixed point subspaces in Figure 4.1. Then a solution of the bifurcation equations can be written in the form $\lambda = -(c_{1k}b_{31} + c_{2k}b_{32})x_k^2/b_1$ with $k = 0$ for the octahedral symmetric (O^-) branch, $k = 2$ for the axisymmetric $O(2)^-$ branch, and $k = 3$ for the dihedral symmetric ($D(6)^d$) branch, respectively, and coefficients $c_{10} = 1$, $c_{20} = 9/(121\pi)$, $c_{12} = 2$, $c_{22} = 42/(121\pi)$, $c_{13} = 2$, $c_{23} = (9/242\pi)$. With the values for b_{31} and b_{32} we get three subcritical bifurcation branches. We also determine the stability of the three nontrivial branches in the full 7 dimensional $l = 3$ representation space. The result is shown in the bifurcation digram of Figure 4.2 where the sign of the eigenvalues of the Jacobian along the bifurcation branches as well as the symmetry of the branches is indicated. All the solution branches are unstable. The branches with discrete symmetry have three zero eigenvalues in the direction of the continuous group orbits. The solution with octahedral symmetry has only two distinct eigenvalues, one of which is of multiplicity three and is negative. The second eigenvalue is of multiplicity one

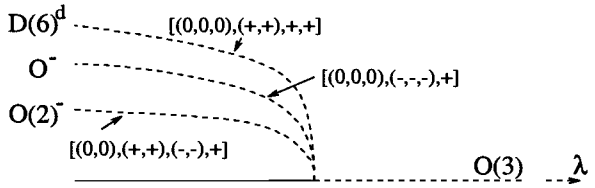


FIG. 4.2. *Bifurcation diagram for generic bifurcations in the $l = 3$ case: The sign of the eigenvalues of the Jacobian along each solution branch is given. Multiple eigenvalues are collected into parentheses.*

and is positive. This solution can be expected to stabilize in a saddle node bifurcation at finite amplitudes.

Acknowledgments. We want to thank the referees for their constructive reports, Prof. B. Schmitt for helpful discussions, and Mrs. Muth for her patient typing.

REFERENCES

[1] E. ALLGOWER AND K. BÖHMER, *Resolving singular nonlinear equations*, Rocky Mountain J. Math., 18 (1988), pp. 225–268.

[2] E.L. ALLGOWER, K. BÖHMER, K. GEORG, AND R. MIRANDA, *Exploiting symmetry in boundary element methods*, SIAM J. Numer. Anal., 29 (1992), pp. 534–552.

[3] P. ASHWIN, Ph.D. thesis, Math. Institute, University of Warwick, Warwick, UK, 1991.

[4] P. ASHWIN, K. BÖHMER, AND Z. MEI, *A numerical Liapunov Schmidt method with applications to Hopf bifurcation on a square*, Math. Comp., 64 (1995), pp. 649–670.

[5] C. BERNARDI AND Y. MADAY, *Spectral methods*, in Handb. Numer. Anal. V, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[6] A.O. BARUT AND R. RACZKA, *Theory of Group Representations and Applications*, 2nd ed., PWM-Polish Scientific, Warsaw, 1980.

[7] W.J. BEYN, *Global bifurcations and their numerical computation,* in Continuation and Bifurcation: Numerical Techniques and Applications, D. Roose, A. Spence, and B. de Dier, eds., Kluwer Academic, Dordrecht, The Netherlands, 1990, pp. 160–181.

[8] K. BÖHMER, *On a numerical Liapunov-Schmidt method for operator equations*, Computing, 53 (1993), pp. 237–269.

[9] K. BÖHMER, *On hybrid methods for bifurcation studies for general operator equations*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, Berlin, Heidelberg, New York, 2001, pp. 73–107.

[10] K. BÖHMER, *On numerical bifurcation studies for general operator equations*, in International Conference on Differential Equations, 1999, Vol. 2, J. Sprekels, B. Fiedler, and K. Gröger, eds., World Scientific, Singapore, 2000, pp. 877–883.

[11] K. BÖHMER, C. GEIGER, AND J. RODRIGUEZ, *On a Numerical Liapunov-Schmidt Spectral Method, Part* I: *Review of Spectral Methods*, Technical report, Schwerpunktprogramm der DFG, Dynamik: Analysis, Effiziente Simulation und Ergodentheorie, 1996.

[12] K. BÖHMER, C. GEIGER, AND J. RODRIGUEZ, *On a Numerical Liapunov-Schmidt Spectral Method, Part* II: *The Reduction Method and Its Applications*, Technical report, Schwerpunktprogramm der DFG, Dynamik: Analysis, Effiziente Simulation und Ergodentheorie, 1997.

[13] K. BÖHMER AND Z. MEI, *On a numerical Lyapunov-Schmidt method*, in Computational Solutions of Nonlinear Systems of Equations, E.L. Allgower and K. Georg, eds., Lectures in Appl. Math. 26, AMS, Providence, RI, 1990, pp. 79–98.

[14] K. BÖHMER AND N. SASSMANNSHAUSEN, *Stability for generalized Petrov-Galerkin methods applied to bifurcation*, ZAMM Z. Angew. Math. Mech., submitted.

[15] K. BÖHMER AND N. SASSMANNSHAUSEN, *Numerical Liapunov-Schmidt spectral method for k-determined problems*, in Computational Methods and Bifurcation Theory with Applications, Comput. Methods Appl. Mech. Engrg. 170, T. Healey, ed., North-Holland, Amsterdam, 1999, pp. 277–312.

[16] F. Brezzi, I. Rappaz, and P.A. Raviart, *Finite dimensional approximations of nonlinear problems, part* II: *Limit points*, Numer. Math., 37, (1981), pp. 1–28.
[17] F. Brezzi, I. Rappaz, and P.A. Raviart, *Finite dimensional approximations of nonlinear problems, part* III: *Simple bifurcation points*, Numer. Math., 38 (1981), pp. 1–30.
[18] G. Caloz and J. Rappaz, *Numerical analysis for nonlinear and bifurcation problems*, in Handbook of Numerical Analysis, Techniques of Scientific Computing, Part 2, Handb. Numer. Anal. V, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 487–637.
[19] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, *Spectral Methods in Fluid Dynamics,* Springer-Verlag, New York, 1988.
[20] S. Chandrasekhar, *Hygrodynamic and Hygromagnetic Stability,* Oxford University Press, Oxford, 1961.
[21] P. Chossat, R. Lauterbach, and I. Melbourne, *Steady-state bifurcation with O(3)-symmetry*, Arch. Ration. Mech. Anal., 113 (1990), pp. 313–376.
[22] A. Fässler and E. Stiefel, *Group Theoretical Methods and Their Applications*, Birkhäuser, Boston, MA, 1992.
[23] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, Cambridge, UK, 1996.
[24] C. Geiger, *Strukturbildung in nichtlinearen dissipativen Systemen mit sphärischer Symmetrie*, Dissertation, Universität Tübingen, Tübingen, Germany, 1994.
[25] C. Geiger, G. Dangelmayr, and J.D. Rodriguez, *$O(3)$–Equivariant Bifurcation in the $l = 4$ Representation*, preprint, 1996.
[26] C. Geiger, G. Dangelmayr, J.D. Rodriguez, and W. Güttinger, *Symmetry breaking bifurcations in the spherical Benard problem, Part* I: *Results from singularity theory,* in Pattern Formation: Symmetry Methods and Applications, Fields Inst. Commun. 5, J. Chadan, M. Golubitsky, W. Langford, and B. Wetton, eds., AMS, Providence, RI, 1995, pp. 225–237.
[27] M. Golubitsky and D. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. I, Springer-Verlag, New York, 1986.
[28] M. Golubitsky, I. Stewart, and D. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. II, Springer-Verlag, New York, 1988.
[29] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, PA, 1977.
[30] A. Griewank and G.W. Reddien, *Computation of cusp singularities for operator equations and their discretization*, J. Comput. Appl. Math (1989), pp. 133–153.
[31] A. Hunding, in Cell to Cell Signaling: From Experiments to Theoretical Models, A. Goldbeter, ed., Academic Press, London, 1989.
[32] E. Ihrig and M. Golubitsky, *Pattern selection with O(3) symmetry*, Phys. D, 13 (1984), pp. 1–33.
[33] A.D. Jepson and A. Spence, *On a reduction process for nonlinear equations*, SIAM J. Math. Anal., 20 (1989), pp. 39–56.
[34] J.P. Keener and H.B. Keller, *Perturbed bifurcation theory,* Arch. Ration. Mech. Anal., 50 (1974), pp. 159–175.
[35] Z. Mei, *Bifurcations of a simplified buckling problem and the effect of discretizations*, Manuscripta Math., 71 (1991), pp. 225–252.
[36] G. Raugel, *Approximation numérique de problemes nonlinéaires*, Thesis, Universitée de Rennes I, Rennes, France, 1985.
[37] J.D. Rodriguez, C. Geiger, G. Dangelmayr and W. Güttinger, *Symmetry breaking bifurcations in the spherical Benard problem, Part* II: *Numerical results,* in Pattern Formation: Symmetry Methods and Applications, Fields Inst. Commun. 5, J. Chadan, M. Golubitsky, W. Langford, and B. Wetton, eds., AMS, Providence, RI, 1995, pp. 239–253.
[38] A. Schwarzer, *Iterationsmethoden für grosse, dünn besetzte Eigenwertproblemeskalierungstechniken für k-bestimmte Verzweigungsprobleme und Iterationsmethoden*, Dissertation, Philipps-Universität Marburg, Marburg, Germany, 1997.
[39] E.E. Selkov, Eur. J. Biochem., 4 (1968), pp. 79–86.
[40] F. Stummel, *Diskrete Konvergenz linearer Operatoren*, I, Math. Ann., 190 (1970), pp. 45–92.
[41] F. Stummel, *Diskrete Konvergenz linearer Operatoren*, II, Math. Z., 120 (1971), pp. 231–264.
[42] L.N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, PA, 2000.
[43] A. Vanderbauwhede, *Local Bifurcation and Symmetry*, Res. Notes Math. 75, Pitman, Boston, 1982.
[44] B. Werner, *Eigenvalue problems with the symmetry of a group and bifurcations*, in Continuation and Bifurcations: Numerical Treatment and Applications, D. Roose, B. de Dier, A. Spence, eds., Kluwer Academic, Dordrecht, 1990, pp. 71–88.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 702–733

A NUMERICAL SCHEME FOR IMPACT PROBLEMS I: THE ONE-DIMENSIONAL CASE*

LAETITIA PAOLI† AND MICHELLE SCHATZMAN‡

Abstract. We consider a mechanical system with impact and one degree of freedom. The system is not necessarily Lagrangian. The representative point is subject to the constraint $u(t) \in \mathbb{R}^+$ for all t. We assume that, at impact, the velocity is reversed and multiplied by a given coefficient of restitution $e \in [0,1]$. We define a numerical scheme which enables us to approximate the solutions of the Cauchy problem: this is an ad hoc scheme which does not require a systematic search for the times of impact. We prove the convergence of this numerical scheme to a solution. Many of the features of this proof will be reused in the nonconvex, multidimensional case, written in generalized coordinates, given in the companion paper [L. Paoli and M. Schatzman, *SIAM J. Numer. Anal.*, 40 (2002), pp. 734–768]. We present some numerical results obtained with the scheme for a spring-dashpot system and we compare them to the results obtained by impact detection and penalization.

Key words. impact, coefficient of restitution, numerical scheme, convergence, local existence, global existence

AMS subject classifications. Primary, 65J10, 65M20, 65B05; Secondary, 17B09, 46N20, 47D03

PII. S0036142900378728

1. Introduction.

We study in this article a numerical approximation of dynamics with impact with one degree of freedom when the representative point u is subject to the constraint

$$u \in K = [0, \infty).$$

Let f be a continuous function from $[0,T] \times \mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ which is locally Lipschitz continuous with respect to its last two arguments.

The free dynamics of the system are written as

$$\ddot{u} = f(\cdot, u, \dot{u}). \tag{1.1}$$

This system is more general than the system obtained in Lagrangian mechanics, since we want to include possible dissipative terms in the dynamics of the problem under discussion. There is no need for a mass matrix: in the one degree of freedom case, the velocity is always proportional to the impulsion, and an obvious change of variable enables us to forget about any other metric other than the Euclidean one.

Let us describe now the system satisfied by the problem with impact: we replace (1.1) with

$$\ddot{u} = \mu + f(\cdot, u, \dot{u}), \tag{1.2}$$

and since we cannot expect to have global solutions in general, μ is an unknown nonnegative measure on $[t_0, t_0 + \overline{\tau}]$ which describes the reaction of the constraints; μ

*Received by the editors September 27, 2000; accepted for publication (in revised form) October 26, 2001; published electronically July 24, 2002.
http://www.siam.org/journals/sinum/40-2/37872.html

†UMR 5585 CNRS Analyse Numérique, Faculté des Sciences, Université Jean Monnet, 23 Rue du Docteur Paul Michelon, 42023 Saint-Etienne Cedex 2, France (paoli@anumsun1.univ-st-etienne.fr).
‡UMR 5585 CNRS Analyse Numérique, Université Lyon 1, 69622 Villeurbanne Cedex, France (schatz@maply.univ-lyon1.fr).

has the following properties:

$$\operatorname{supp}(\mu) \subset \{t \in [t_0, t_0 + \overline{\tau}] : u(t) = 0\}, \tag{1.3a}$$

$$\mu \geq 0. \tag{1.3b}$$

We require the following functional properties for u:

$$u \text{ is a continuous nonnegative function on } [t_0, t_0 + \overline{\tau}], \tag{1.4a}$$

$$\dot{u} \text{ is of bounded variation over } [t_0, t_0 + \overline{\tau}]. \tag{1.4b}$$

We have to make a supplementary assumption in order to have a complete description of the impact; we choose a constitutive law of the impact using a coefficient of restitution. Thus we will assume that there exists $e \in [0, 1]$ such that $\dot{u}(t + 0)$ is equal to $-e$ times $\dot{u}(t - 0)$. In other words, we have

$$\dot{u}(t + 0) = -e\dot{u}(t - 0). \tag{1.5}$$

The set of admissible initial data $\mathbb{D}$ will be

$$\mathbb{D} = \big\{(t_0, u_0, v_0) \in [0, T) \times K \times \mathbb{R} : \text{if } u_0 = 0, \text{ then } v_0 \geq 0\big\}.$$

This choice is equivalent to the convention that there is no impact at the initial time t_0.

Given initial conditions $(t_0, u_0, v_0) \in \mathbb{D}$, we require that the following Cauchy data be satisfied:

$$u(t_0) = u_0 \tag{1.6}$$

and

$$\dot{u}(t_0) = v_0. \tag{1.7}$$

For all initial data $(t_0, u_0, v_0) \in \mathbb{D}$ we will obtain the existence of a local solution to (1.2), (1.3a), (1.3b), and (1.5) belonging to the functional class defined by (1.4a) and (1.4b) and satisfying the initial conditions (1.6) and (1.7).

The existence of this local solution is obtained by defining a numerical scheme, whose convergence will be shown in appropriate functional spaces; the limit of the approximation will be a solution of our problem. The projection on K is given by

$$P_K(x) = \max(x, 0) = x^+. \tag{1.8}$$

Given two positive numbers $h^* \leq 1$ and T, assume that F is a continuous function from $[0, T] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times [0, h^*]$ to $\mathbb{R}$, which is locally Lipschitz continuous with respect to its second, third, and fourth arguments. Assume, moreover, that F is consistent with f, i.e., that for all $t \in [0, T]$, for all u and v in $\mathbb{R}$,

$$F(t, u, u, v, 0) = f(t, u, v). \tag{1.9}$$

We approximate the solution of (1.2), (1.3a), (1.3b), and (1.4a), (1.4b), (1.5), (1.6), (1.7) by the following numerical scheme: the initial values y^0 and y^1 are given by the initial position

$$y^0 = u_0, \tag{1.10}$$

and the position at the first time step

$$y^1 = u_0 + hv_0 + hz(h), \tag{1.11}$$

where $z(h)$ tends to 0 as h tends to 0.

Henceforth, we will systematically use the notation

$$t^m = t_0 + mh.$$

Given y^{m-1} and y^m, y^{m+1} is defined by the relations

$$y^{m+1} = -ey^{m-1} + \left(2y^m - (1-e)y^{m-1} + h^2F^m\right)^+ \tag{1.12}$$

and

$$F^m = F\left(t^m, y^m, y^{m-1}, \frac{y^{m+1} - y^{m-1}}{2h}, h\right). \tag{1.13}$$

The reader should be aware at this point that, given y^{m-1} and y^m, the existence of y^{m+1} is an easy consequence of the strict contraction theorem; but we might be in trouble here, since F is locally Lipschitz continuous and we are not sure about the existence of a solution on a finite time interval. This existence is not a trivial question and it depends on estimates which are at the heart of our subject.

A commentary on the construction of this scheme from the point of view of convex analysis will be useful here. We refer to the book of Rockafellar [25] for more information on the basic ideas in convex analysis to be used below.

Recall that the indicator function ψ_K of the closed convex set K is defined by

$$\psi_K(x) = \begin{cases} 0 & \text{if } x \in K, \\ +\infty & \text{otherwise,} \end{cases} \tag{1.14}$$

and its subdifferential $\partial\psi_K$ is a function from K to the set of closed convex sets given by

$$\partial\psi_K(x) = \begin{cases} \{0\} & \text{if } x \in \text{int}(K), \\ \mathbb{R}^- & \text{if } x \in \partial K. \end{cases} \tag{1.15}$$

For all $\lambda > 0$, the multivalued equation

$$x + \lambda\partial\psi_K(x) \ni f \tag{1.16}$$

has a unique solution given by

$$x = P_K f. \tag{1.17}$$

In the announcement [22], we assumed that the set of constraints K was convex and the geometry was Euclidean and d-dimensional, and we had defined a numerical scheme by the multivalued equation

$$\frac{y^{m+1} - 2y^m + y^{m-1}}{h^2} + \partial\psi_K\left(\frac{y^{m+1} + ey^{m-1}}{1+e}\right) \ni F^m. \tag{1.18}$$

We may rewrite (1.18) as

$$\begin{aligned} &\frac{y^{m+1}+ey^{m-1}}{1+e}+\frac{h^2}{1+e}\partial\psi_K\left(\frac{y^{m+1}+ey^{m-1}}{1+e}\right) \\ &\quad\ni \frac{2y^m-(1-e)y^{m-1}+h^2F^m}{1+e}, \end{aligned} \tag{1.19}$$

which reduces, thanks to (1.16), (1.17), and (1.8), to relation (1.12). In contrast with the original proof of convergence of this scheme, as written in [19], the proof presented here is written in such a way that many of its features can be reused in the nonconvex, multidimensional case, written in generalized coordinates. This more general proof will be given in the companion paper [20].

Let us outline now the structure of the article and of the proofs. The main estimates are given by Lemma 2.1 in section 2.

Then we find two constants A and τ such that for initial data given by (1.10)–(1.11), and for all small enough h and all $m \le \tau/h$, the discrete velocity is bounded:

$$\sup \left|(y^{m+1}-y^m)/h\right| \le A.$$

In sections 4, 5, 6, and 7, we prove estimates on the discrete acceleration, we establish the variational properties of the limit of the numerical scheme, and we study the transmission of energy at impact, as well as the passage to the limit for the initial conditions. All these results are obtained under the assumption that on a certain time interval starting at t_0, the discrete velocity is bounded independently of the time step.

As a preliminary to the global existence proof, we give a priori estimates on problem (1.2)–(1.7) in section 8, which is completely independent from the remainder of the article.

In section 9, we establish a very weak semicontinuity for the supremum of the local norm of the discrete velocities; this result enables us to obtain a global existence and convergence theorem.

This article contains theoretical results and reports also on some of the numerical implementations.

The existence result obtained here is a generalization of [26], [2], [24], [19], [17]. The numerical scheme has been implemented in the one-dimensional case, and the results were reported in [19], [18], [23]. In all these articles, we compared the performances of this scheme with those of a method based on the detection of impact. When the impact times are isolated, the algorithm by detection of impacts is more precise than the present scheme. As soon as the restitution coefficient is strictly less than one, we find systematically nonisolated impact times. In all cases, the present scheme is substantially faster. Since the phenomena that we want to approximate are highly nonlinear and often very sensitive to the initial data, the issue of precision is not necessarily crucial. Our numerical experiments show that the performance of the present numerical scheme is quite satisfactory from the point of view of qualitative conclusions.

The aforementioned references also concern the multidimensional case.

Let us remark that many articles have been devoted to the problem treated here under the assumption of inelasticity, i.e., a situation where the normal component of the impulsion vanishes after the impact. Moreau applied Gauss's principle of least constraint to unilateral problems in order to justify his choice of inelastic impact

[11], which eventually led him to sweeping processes [14], followed by [12], [13]. Dry friction enters in Moreau's work as [15]; frictionless inelastic impact starts as [16], and the mathematical theory is tackled by M. Monteiro Marques in a series of articles: his main contributions are [9] for the general theory of differential inclusions, [10] for one-dimensional dynamics with friction, and [5] which adds percussion to the previous framework; this work is improved as [6], where dynamics of n particles on a plane with normal friction are considered. The discretization approach has been taken up by Kunze and Monteiro Marques in [4], but most significantly by Stewart and Trinkle: they use that approach in [27], [29], and [30]. The real coronation is the beautiful and difficult article of Stewart [28], which concludes the study of dynamics with friction and inelastic impact for a finite number of degrees of freedom and one constraint; his results are not quite so precise in the case of many constraints but are still very important.

The philosophy of this long list of works is somewhat different from ours: we feel that not all impacts are inelastic, and we were originally motivated by continuous media; thus, we wanted to develop methods which work well for stiff systems of ordinary differential equations. From this point of view, any method which has to calculate with some precision the impact times is doomed to failure. On the other hand, the precision of the method presented here needs improvement, and globally it would make sense to agree on benchmarks which would enable the end-user to decide between different numerical methods.

Our approach is not the only possible one. In particular, in recent work, Mabrouk in [7] and [8] defines a numerical scheme for vibro-impact as a tool for proving existence of dynamics with impact; he allows for elastic, partially elastic, or inelastic impact. In his work, the mass matrix is assumed to be scalar, and therefore the metric is Euclidean. The idea is to discretize Moreau's formulation which describes the constraint in terms of velocity instead of describing it in terms of position, as we do here.

From a practical point of view a nice property of Mabrouk's scheme is that the velocity is reversed immediately upon impact. However, the number of steps during which the representative point of the system is outside of the set of constraints can be very large. For $e = 0$, the representative point of the system can even leave for ever the set of constraints, while remaining close to it.

Therefore, numerical simulations will probably decide which of all these numerical schemes gives the most reliable simulations of dynamics with impact. It may well be that a future scheme will conciliate the two distinct approaches and perform better than them: the road is open to researchers to try their ingenuity on these challenging problems.

2. The heart of the estimates. In the one-dimensional case, the main estimate on the numerical scheme is described in the following lemma; we recall the definition

$$r^+ = \max(r, 0).$$

LEMMA 2.1. *Let the real-valued sequence* $\left(y^m\right)_m$ *satisfy the following recurrence relation for all* $m \geq 1$*:*

$$y^{m+1} = -ey^{m-1} + \left(2y^m - (1-e)y^{m-1}\right)^+ + h^2\lambda^m. \tag{2.1}$$

Then, for all $m \geq 2$*, the discrete velocity*

$$\eta^m = \left(y^{m+1} - y^m\right)/h \tag{2.2}$$

satisfies the estimate

$$|\eta^m| \le \max\bigl(|\eta^{m-1}|, e\,|\eta^{m-2}|\bigr) + h\,|\lambda^m| + h\,|\lambda^{m-1}|. \tag{2.3}$$

Proof. Assume first that $2y^m-(1-e)y^{m-1}$ is nonnegative, and substitute $y^{m+1} = y^m + h\eta^m, y^{m-1} = y^m - h\eta^{m-1}$ into (2.1); we obtain

$$\eta^m = \eta^{m-1} + h\lambda^m$$

so that

$$|\eta^m| \le |\eta^{m-1}| + h\,|\lambda^m|. \tag{2.4}$$

Assume now that $2y^m - (1-e)y^{m-1}$ is strictly negative. On the one hand, (2.1) implies the relation

$$\eta^m = e\eta^{m-1} - \frac{1+e}{h}y^m + h\lambda^m;$$

the assumption on the sign of $2y^m - (1-e)y^{m-1}$ is equivalent to

$$\frac{(1+e)y^m}{h} < -(1-e)\eta^{m-1},$$

and therefore

$$\eta^m > \eta^{m-1} + h\lambda^m. \tag{2.5}$$

On the other hand, we subtract from the relation

$$y^{m+1} + ey^{m-1} = h^2\lambda^m$$

the inequality implied by (2.1) with m substituted by $m-1$:

$$y^m + ey^{m-2} \ge h^2\lambda^{m-1},$$

and we infer that

$$\eta^m \le -e\eta^{m-2} + h\bigl(\lambda^m - \lambda^{m-1}\bigr). \tag{2.6}$$

When we summarize (2.4), (2.5), and (2.6), we find (2.3). ∎

3. Existence of a discrete solution and estimates on the discrete velocity. We systematically use the floor and ceiling notations: when r is a real number, the floor $\lfloor r\rfloor$ of r is the largest integer at most equal to r, and the ceiling $\lceil r\rceil$ is the smallest integer at least equal to r. In this section we prove that for h and τ small enough, relations (1.12) and (1.13) uniquely define a numerical solution while $(m+1)h \le \tau$; moreover the discrete velocity of this solution is bounded independently of h.

The idea of this result is to show the existence by the Brouwer fixed point argument and the uniqueness by local considerations.

We say that a pair of numbers y^0 and y^1 satisfy the property $P(a,h)$ if the following conditions are true:

$$|y^0| \le a, \quad |y^1| \le a, \tag{3.1a}$$

$$|y^1 - y^0| \le ah, \tag{3.1b}$$

$$\begin{cases} \text{There exists } y^2 \text{ such that } |y^2 - y^1| \le ah \\ \text{and } y^2 + ey^0 - \bigl(2y^1 - (1-e)y^0 + h^2F(t^1, y^0, y^1, (y^2-y^0)/2h, h)\bigr)^+ = 0. \end{cases} \tag{3.1c}$$

THEOREM 3.1. *For all $a > 0$ and for all $A > a$, there exists $\tau > 0$ such that, for all $t_0 \in [0,T)$ and for all y^0 and y^1 satisfying property $P(a,h)$, there exists a numerical solution of the scheme* (1.12) *and* (1.13) *for $0 \le mh \le \tau$, which satisfies, moreover, the estimate*

$$\forall m \in \{0,\dots,\lfloor \tau/h \rfloor - 1\}, \quad |y^{m+1} - y^m| \le Ah. \tag{3.2}$$

Proof. Define

$$C_1 = \sup\{|F(t,u,u',0,h)| : 0 \le t \le T, \quad |u| \le a, \quad |u'| \le a, \quad 0 \le h \le h^*\}.$$

Let L be the local Lipschitz constant of F defined by

$$\begin{aligned} &\forall (t,u_i,u_i',v_i,h) \in [0,T] \times [-a-AT, a+AT]^2 \times [-A,A] \times [0,h^*], \quad i = 1,2, \\ &|F(t,u_1,u_1',v_1,h) - F(t,u_2,u_2',v_2,h)| \le L\big(|u_1-u_2| + |u_1'-u_2'| + |v_1-v_2|\big). \end{aligned}$$

Choose $\tau > 0$ such that

$$2\tau\big(C_1 + LA + 2LA\tau\big) \le A - a. \tag{3.3}$$

We will apply a Brouwer fixed point argument; we choose y^2 according to (3.1c), and we define a compact convex set B_h by

$$\begin{aligned} B_h = \big\{ \widehat{y} = \big(\widehat{y}^m\big)_{0 \le mh \le \tau} : \widehat{y}^0 = y^0, \widehat{y}^1 = y^1, \widehat{y}^2 = y^2, \\ \forall m \in \{1,\dots,\lfloor \tau/h \rfloor - 1\} : |\widehat{y}^{m+1} - \widehat{y}^m| \le Ah \big\}. \end{aligned}$$

Assuming that $\widehat{y}$ belongs to B_h, we define $\widehat{F}$ by

$$\widehat{F}^m = F(t^m, \widehat{y}^m, \widehat{y}^{m-1}, (\widehat{y}^{m+1} - \widehat{y}^{m-1})/2h, h), \quad m \in \{1,\dots,\lfloor \tau/h \rfloor - 1\}.$$

We now write the numerical scheme

$$y^{m+1} + ey^{m-1} - \big(2y^m - (1-e)y^{m-1} + h^2\widehat{F}^m\big)^+ = 0, \tag{3.4}$$

which can be put under the form (2.1), provided that we define

$$h^2\lambda^m = \big(2y^m - (1-e)y^{m-1} + h^2\widehat{F}^m\big)^+ - \big(2y^m - (1-e)y^{m-1}\big)^+. \tag{3.5}$$

It should be remarked that (3.4) possesses a unique solution, since it is explicit in y^{m+1}, and that if the mapping $\widehat{y} \mapsto y$ possesses a fixed point, this fixed point is precisely the numerical solution sought here. We estimate the discrete velocity η^m thanks to Lemma 2.1: the number λ^m is estimated by

$$\begin{aligned} &|\lambda^m| \le |\widehat{F}^m| \\ &\le |F(t^m, y^0, y^0, 0, h)| + L\big(|\widehat{y}^m - y^0| + |\widehat{y}^{m-1} - y^0| + |\widehat{\eta}^m|/2 + |\widehat{\eta}^{m-1}|/2\big), \end{aligned}$$

and the assumption $\widehat{y} \in B_h$ guarantees that

$$|\lambda^m| \le C_1 + 2LA\tau + LA. \tag{3.6}$$

Estimate (2.3) implies

$$|\eta^m| \le \max\big(|\eta^0|, |\eta^1|\big) + 2\big(C_1 + 2LA\tau + LA\big)mh$$

by discrete integration. We may conclude now that for $(m+1)h \le \tau$,

$$|\eta^m| \le a + 2\big(C_1 + 2LA\tau + LA\big)\tau,$$

and thanks to assumption (3.3), y also belongs to B_h. This mapping is clearly continuous, which implies the existence of a fixed point thanks to Brouwer's fixed point theorem. □

There remains two easy lemmas; the first one settles for h small the question of the uniqueness of the numerical solution.

LEMMA 3.2. *Under the hypotheses of Theorem* 3.1*, there exists* $h_1 > 0$ *such that for all* $h \in (0, h_1]$*, for all* y^0 *and* y^1 *satisfying condition* $P(a,h)$*, the numerical solution of* (1.12) *and* (1.13) *satisfying estimate* (3.2) *is unique.*

Proof. Given y^{m-1} and y^m, the discrete velocity η^m is a fixed point of the mapping

$$\begin{aligned}\eta \mapsto h^{-1}\Big(&-ey^{m-1} - y^m \\ &+ \big(2y^m - (1-e)y^{m-1} + h^2F(t^m, y^m, y^{m-1}, (\eta+\eta^{m-1})/2, h)\big)^+\Big).\end{aligned} \tag{3.7}$$

Let L be the Lipschitz constant of the mapping

$$z \mapsto F(t, y, y', z, h)$$

for $t \in [0,T]$, y and y' in $[-a-TA, a+TA]$, $|z| \le A$, and $0 \le h \le h^*$. Then the Lipschitz constant of the mapping (3.7) is $hL/2$, and therefore, if $h_1 < 2/M$, the uniqueness of η^m is guaranteed and the lemma is proved. □

The second lemma establishes that under conditions (1.10) and (1.11), property $P(a,h)$ holds.

LEMMA 3.3. *Assume that* $y^0 = u_0$ *and* $y^1 = u_0 + hv_0 + ho(h)$ *as in* (1.11)*; then there exists* $h_1 > 0$ *such that for all* $h \in (0, h_1]$*, there exists a unique* y^2 *such that*

$$y^2 + ey^0 - \big(2y^1 - (1-e)y^0 + h^2F(t^1, y^0, y^1, (y^2 - y^0)/2h, h)\big)^+ = 0$$

and

$$\max(|y^1 - y^0|, |y^2 - y^1|) \le (3|v_0| + 1)h.$$

Proof. Define a mapping

$$z \mapsto h^{-1}\Big(\big((1+e)y^0 + 2\eta^0 h + h^2F(t^1, y^0, y^1, z/2, h)\big)^+ - (1+e)y^0\Big), \tag{3.8}$$

which is slightly different from the mapping (3.7), since its fixed point will be $(y^2 - y^0)/h$. Standard arguments show that it is possible to find $h_1 > 0$ such that for all $h \in (0, h_1]$ and all y^1 in $[y^0 - 1, y^0 + 1]$, the mapping (3.8) is a strict contraction from the ball of radius $2|v_0| + 1/2$ to itself. We set $y^2 = y^0 + hz$, where z is the fixed point of the above mapping. Then it is clear that for h small enough,

$$|\eta^2| = |(y^2 - y^1)/h| \le |z| + |\eta^0|,$$

and the lemma is proved. □

As a consequence of Theorem 3.1 and Lemma 3.2, we have the following result.

PROPOSITION 3.4. *For all* $(t_0, u_0, v_0) \in \mathbb{D}$*, there exists* $A > 0$*,* $\tau \in (0, T - t_0]$*, and* $h_1 \in (0, h^*]$ *such that for all* $m \in \{0, \dots, \lfloor \tau/h \rfloor\}$*,* y^m *is uniquely defined by* (1.10)*,* (1.11)*, and the recursive formulas* (1.12)*,* (1.13) *and satisfies the estimate*

$$\forall m \in \big\{0, \dots, \lfloor \tau/h \rfloor - 1\big\}, \quad |y^{m+1} - y^m| \le Ah.$$

Proof. The main observation is that we have to choose A as a function of the initial data. Thanks to Lemma 3.3, it suffices to take

$$A \ge \max\big(3|v_0| + 1, |u_0| + 1\big);$$

the remainder of the argument is clear. ∎

4. Estimates on the discrete acceleration. In this section and the three following ones, we assume that there exist strictly positive numbers τ, A, and h_1 and a subsequence of time steps to which correspond solutions of the numerical scheme defined by (1.10), (1.11), (1.12), and (1.13), which satisfy the estimate, for all $h \le h_1$,

$$\forall l \in \{0, \dots, P - 1\}, \quad \big|y^{l+1} - y^l\big| \le Ah, \tag{4.1}$$

where

$$P = \lfloor \tau/h \rfloor.$$

Here we estimate the discrete total variation of the sequence $\big(\eta^m\big)_m$.

THEOREM 4.1. *Under assumption* (4.1)*, there exists a constant* C_2 *such that for all* $h \le h_1$

$$\sum_{m=1}^{P-1} \big|\eta^m - \eta^{m-1}\big| \le C_2. \tag{4.2}$$

Proof. The constant C_3 is taken as a majorant of $|F^m|$; we can take it as equal to

$$\begin{aligned} C_3 = \max\big\{|F(t, y, y', z, h)| : t \in [0, T], |y - u_0| \le AT, \\ |y' - u_0| \le AT, |z| \le A, 0 \le h \le h^*\big\}. \end{aligned} \tag{4.3}$$

We put the numerical scheme under the form (2.1) by defining λ^m through

$$h^2\lambda^m = \big(2y^m - (1 - e)y^{m-1} + h^2F^m\big)^+ - \big(2y^m - (1 - e)y^{m-1}\big)^+, \tag{4.4}$$

which differs slightly from (3.5), since it involves F^m instead of $\widehat{F}^m$. The number λ^m is estimated by

$$|\lambda^m| \le |F^m| \le C_3.$$

We observe that

$$\eta^m - \eta^{m-1} = h\lambda^m + (2y^m - (1 - e)y^{m-1})^-/h. \tag{4.5}$$

Therefore, by the triangle inequality,

$$|\eta^m - \eta^{m-1}| \le hC_3 + (2y^m - (1 - e)y^{m-1})^-/h.$$

Using (4.5) again, we obtain

$$|\eta^m - \eta^{m-1}| \le 2hC_3 + \eta^m - \eta^{m-1}. \tag{4.6}$$

We observe that we have the elements of a telescoping sum: we sum (4.6) for m varying from 1 to $P-1$ and we get

$$\sum_{m=1}^{P-1} |\eta^m - \eta^{m-1}| \le 2hC_3P + \eta^{P-1} - \eta^0 \le 2C_3\tau + 2A. \qquad \square$$

5. Variational properties of the limit of the numerical scheme. In this section, we work under the assumption (4.1). We define a function u_h by affine interpolation, as follows:

$$\begin{cases} u_h(t) = y^m + (t - t_0 - mh)\dfrac{y^{m+1} - y^m}{h} \\ \qquad\qquad \text{for } t - t_0 \in [mh, (m+1)h), \ 0 \le m \le P-1, \\ u_h(t) = y^P \quad \text{for } t - t_0 \in [Ph, \tau]. \end{cases} \tag{5.1}$$

We also define a measure F_h as the following sum of Dirac masses:

$$F_h(t) = \sum_{m=1}^{P-1} hF^m \delta(t - t_0 - mh). \tag{5.2}$$

In this section we prove that the sequence $(u_h)_h$ converges in an appropriate sense to a function u which satisfies (1.2) to (1.4b) with τ instead of $\overline{\tau}$. We delay the proof of (1.5), the transmission condition at impacts, to a later section.

There are three steps in the convergence proof: the first is to prove that the limit u exists in an appropriate sense and takes its values in K; in the second step, we show that $\dot{u}_h$ is of bounded variation uniformly in h and that F_h converges to $f(\cdot, u, \dot{u})$ weakly in the space of $\mathbb{R}$-valued measures. The last step is the characterization of the measure $\mu = \ddot{u} - f(\cdot, u, \dot{u})$: there we show that μ satisfies conditions (1.3a) and (1.3b).

LEMMA 5.1. *From all sequence of functions $(u_h)_h$ indexed by a sequence h tending to 0, it is possible to extract a subsequence, still denoted by $(u_h)_h$, such that*

$$u_h \to u \quad \textit{in } C^0([t_0, t_0+\tau]) \textit{ strong}, \tag{5.3}$$

$$\dot{u}_h \to \dot{u} \quad \textit{in } L^\infty([t_0, t_0+\tau]) \textit{ weak } *. \tag{5.4}$$

The function u takes its values in K.

Proof. Thanks to assumption (4.1), we know that $(u_h)_{0<h\le h_1}$ is uniformly Lipschitz continuous over $[t_0, t_0+\tau]$. Therefore, we may extract a subsequence, still denoted by u_h, such that (5.3) and (5.4) hold. Thus u belongs to $W^{1,\infty}([t_0, t_0+\tau]) \cap C^0([t_0, t_0+\tau])$, which means that u is a Lipschitz continuous function [1]. For all m belonging to $\{1, \dots, P-1\}$, we have that

$$\frac{y^{m+1} + ey^{m-1}}{1+e} = y^m + h\frac{\eta^m - e\eta^{m-1}}{1+e} \ge 0. \tag{5.5}$$

It follows that, for all $m \in \{1, \dots, P-1\}$, the Euclidean distance between y^m and K can be estimated as follows:

$$(y^m)^- \le h\left|\eta^m - e\eta^{m-1}\right| / (1+e) \le hA. \tag{5.6}$$

Thanks to the definition (5.1), we can see that for all $t \in [t_0, t_0+\tau]$ the Euclidean distance between $u_h(t)$ and K is estimated by $2hA$. This allows us to pass to the limit when h tends to 0 and to conclude. □

The next lemma describes the convergence of the measures involved in our problem; we denote by $M^1\big((t_0,t_0+\tau)\big)$ the space of bounded measures over $(t_0,t_0+\tau)$.

LEMMA 5.2. *The measures $\ddot{u}_h$ and F_h converge weakly $*$ in $M^1\big((t_0,t_0+\tau)\big)$ to $\ddot{u}$ and $f(\cdot,u,\dot{u})$, respectively.*

Proof. The measure $\ddot{u}_h$ is a sum of Dirac measures on $(t_0,t_0+\tau)$; more precisely, we have

$$\ddot{u}_h(t) = \sum_{m=1}^{P-1} (\eta^m - \eta^{m-1})\delta(t-t_0-mh) - \eta^{P-1}\delta(t-t_0-Ph),$$

and the total variation of $\dot{u}_h$ on $(t_0,t_0+\tau)$ is estimated by

$$TV(\dot{u}_h) \le \sum_{m=1}^{P-1} \left|\eta^m - \eta^{m-1}\right| + \left|\eta^{P-1}\right|.$$

Theorem 4.1 implies that $(\dot{u}_h)_{0<h\le h_1}$ is a bounded family in $BV\big((t_0,t_0+\tau)\big)$, the space of functions of bounded variation over $(t_0,t_0+\tau)$. Using Helly's theorem, we can extract another subsequence $\big(\dot{u}_h\big)_h$ which converges, except perhaps on a countable set of points, to a function of bounded variation. Hence

$$\dot{u} \in BV\big((t_0,t_0+\tau)\big).$$

Moreover,

$$\ddot{u}_h \to \ddot{u} \quad \text{ weakly } * \text{ in } M^1\big((t_0,t_0+\tau)\big).$$

Lebesgue's theorem implies that $\dot{u}_h$ converges to $\dot{u}$ in $L^1\big(t_0,t_0+\tau\big)$. We extend $\dot{u}_h$ and $\dot{u}$ to $\mathbb{R}$ by 0 outside of $(t_0,t_0+\tau)$ and still denote the respective extensions by $\dot{u}_h$ and $\dot{u}$. The set $\{\dot{u}_h : h \in (0,h_1]\} \cup \{\dot{u}\}$ is a compact subset of $L^1(\mathbb{R})$. The classical characterization of compact subsets of $L^1(\mathbb{R})$ [3] implies that

$$\lim_{\theta\to 0} \sup_{0<h\le h_1} \int_{\mathbb{R}} |\dot{u}_h(t-\theta) - \dot{u}_h(t)|\, dt = 0.$$

Letting $\theta = h$, we can see that $\dot{u}_h(\cdot - h)$ converges to $\dot{u}$ in $L^1\big(\mathbb{R}\big)$. Let us define an approximate velocity v_h on $\mathbb{R}$ by

$$v_h(t) = \frac{\dot{u}_h(t-h+0) + \dot{u}_h(t+0)}{2}.$$

The sequence v_h converges to $\dot{u}$ in $L^1\big(\mathbb{R}\big)$. Moreover, for all $t \in [t^m, t^{m+1})$ and for all $m \in \{1,\dots,P-1\}$, we have the identity

$$v_h(t) = \frac{\eta^m + \eta^{m-1}}{2}.$$

We immediately have the following estimates for all $t \in (t_0,t_0+\tau)$ and all $h \in (0,h_1]$:

$$|v_h(t)| \le A, \quad |u_h(t) - u_0| \le A(t-t_0) \le A\tau. \tag{5.7}$$

Let ψ be a continuous function over $[0,T]$ with compact support included in $(t_0, t_0+\tau)$. For all small enough h, the support of ψ is included in $[t_0+h, t_0+(P-1)h]$. The duality product $\langle F_h, \psi\rangle$ has the expression

$$\langle F_h, \psi\rangle = \sum_{m=1}^{P-1} h\psi(t_0+mh)F^m. \tag{5.8}$$

We wish to compare the expression (5.8) to

$$\int_{t_0}^{t_0+\tau} \psi f(\cdot, u, \dot{u})\, dt. \tag{5.9}$$

We compare the right-hand side of (5.8) which is basically a numerical quadrature by the formula of rectangles to an appropriate integral. Let us rewrite the individual terms of the right-hand side of (5.8) as

$$h\psi(t^m)F^m = \int_{t^m}^{t^{m+1}} \psi(t)F^m\, dt + \int_{t^m}^{t^{m+1}} \big(\psi(t^m)-\psi(t)\big)F^m\, dt. \tag{5.10}$$

Consider now the second term on the right-hand side of (5.10). Recalling estimate (4.3),

$$\max_{0\le m\le n} |F^m| \le C_3, \tag{5.11}$$

and denoting by ω_ψ the modulus of continuity of ψ we can see that

$$\left| \int_{t^m}^{t^{m+1}} \big(\psi(t^m)-\psi(t)\big)F^m\, dt \right| \le C_3\omega_\psi(h)h. \tag{5.12}$$

We consider now the first term on the right-hand side of (5.10), which we would like to compare to expression (5.9). Thanks to the consistency assumption (1.9) we have the following inequalities, for all $t\in[t^m, t^{m+1})$ and all $m\in\{1,\dots,P-1\}$:

$$\begin{aligned}
&\big|F^m - f\big(t, u_h(t), v_h(t)\big)\big| \\
&\quad\le \big|F\big(t^m, y^m, y^{m-1}, v_h(t^m), h\big) - F\big(t^m, u_h(t), u_h(t), v_h(t^m), h\big)\big| \\
&\quad+ \big|F\big(t^m, u_h(t), u_h(t), v_h(t^m), h\big) - F\big(t^m, u_h(t), u_h(t), v_h(t), 0\big)\big| \\
&\quad+ \big|f\big(t^m, u_h(t), v_h(t)\big) - f\big(t, u_h(t), v_h(t)\big)\big|.
\end{aligned}$$

Denote by $\mathcal{D}$ the set

$$\begin{aligned}
\mathcal{D} = \big\{(t,u_1,u_2,v,h) : 0\le t\le T,\quad &|u_1-u_0|\le AT, \\
&|u_2-u_0|\le AT,\quad |v|\le A,\quad 0\le h\le h^*\big\}.
\end{aligned}$$

Let L be the Lipschitz constant of $(u_1,u_2)\mapsto F(t,u_1,u_2,v,h)$ restricted to $\mathcal{D}$ and let ω_F be the modulus of continuity of F on $\mathcal{D}$. With these notations, we can see that

$$\begin{aligned}
&\big|F^m - f\big(t, u_h(t), v_h(t)\big)\big| \\
&\quad\le L\big(|y^m-u_h(t)| + \big|y^{m-1}-u_h(t)\big|\big) + 2\omega_F(h).
\end{aligned} \tag{5.13}$$

Since $(u_h)_h$ converges strongly in $C^0([t_0, t_0+\tau])$ and $(v_h)_h$ converges strongly to $\dot{u}$ in $L^1(\mathbb{R})$ and almost everywhere on $(t_0, t_0+\tau)$, we see that $f(\cdot, u_h, v_h)$ tends to $f(\cdot, u, \dot{u})$ strongly in $L^1(t_0, t_0+\tau)$ and almost everywhere on $(t_0, t_0+\tau)$. We summarize relations (5.12) and (5.13) together with the above convergence result, and we find that

$$\begin{aligned}
&\left| \langle F_h, \psi \rangle - \int_{t_0}^{t_0+\tau} \psi f(\cdot, u, \dot{u})\, dt \right| \\
&\leq \int_{t_0}^{t_0+\tau} \left| f\big(\cdot, u_h, v_h\big) - f\big(\cdot, u, \dot{u}\big) \right| |\psi|\ dt \\
&\quad + C_3 \omega_\psi(h)\tau + (3LAh + 2\omega_F(h)) \int_{t_0}^{t_0+\tau} |\psi|\ dt,
\end{aligned}$$

which concludes the proof. □

Let us prove now that the measure μ has the required variational properties.

LEMMA 5.3. *The measure μ satisfies properties* (1.3a) *and* (1.3b).

Proof. Define

$$\mu_h = \ddot{u}_h - F_h;$$

μ_h is a sum of Dirac measures on $(t_0, t_0+\tau)$. More precisely

$$\begin{aligned}
\mu_h = &\sum_{m=1}^{P-1} \big(\eta^m - \eta^{m-1} - hF^m\big)\delta(t - t_0 - mh) \\
&- \eta^{P-1}\delta(t - t_0 - Ph).
\end{aligned}$$

With all the previous results, we know that μ_h converges to $\mu = \ddot{u} - f(\cdot, u, p)$ weakly * in $M^1\big((t_0, t_0+\tau)\big)$. Let us prove property (1.3a). Assume that τ_0 is a point of $(t_0, t_0+\tau)$ such that $u(\tau_0) > 0$. Then, by continuity of u, there exist $\varepsilon > 0$ and $\rho > 0$ such that

$$\forall t \in (\tau_0 - \varepsilon, \tau_0 + \varepsilon), \quad u(t) \geq 3\rho.$$

Since the sequence $(u_h)_h$ converges uniformly to u as h tends to 0, we can decrease h_1 so that

$$\forall h \in (0, h_1], \quad \forall t \in (\tau_0 - \varepsilon, \tau_0 + \varepsilon), \quad u_h(t) \geq 2\rho.$$

Replacing y^{m-1} by $y^m - h\eta^{m-1}$, we have

$$2y^m - (1-e)y^{m-1} + h^2F^m = (1+e)y^m + (1-e)h\eta^{m-1} + h^2F^m,$$

and relations (4.3) and (4.1) imply that

$$2y^m - (1-e)y^{m-1} + h^2F^m \geq (1+e)y^m - (1-e)hA - h^2C_3.$$

Possibly decreasing h_1, we have

$$\forall h \in (0, h_1], \quad \forall t^m \in (\tau_0 - \varepsilon, \tau_0 + \varepsilon), \quad 2y^m - (1-e)y^{m-1} + h^2F^m \geq \rho,$$

and thus

$$\forall h \in (0, h_1], \quad \forall t^m \in (\tau_0 - \varepsilon, \tau_0 + \varepsilon), \quad \eta^m - \eta^{m-1} - hF^m = 0.$$

This proves that, for h small enough, the support of μ_h does not intersect the open set $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$ and therefore relation (1.3a). In order to conclude the proof, we observe that

$$\eta^m - \eta^{m-1} - hF^m = \frac{1}{h}(2y^m - (1-e)y^{m-1} + h^2F^m)^- \geq 0.$$

Thus, for all $\tau' \in (0, \tau)$, the measure μ_h is nonnegative on $(t_0, t_0 + \tau')$ for h small enough, which implies by a straightforward passage to the limit that μ is nonnegative. This concludes the proof of the lemma. □

6. Transmission of energy during impact. The basic assumption is still the one made at the beginning of section 4.

Let $\overline{\tau} \in (0, \tau)$ be such that $u(t_0 + \overline{\tau})$ vanishes. Write $\overline{t} = t_0 + \overline{\tau}$.

We will prove the relation

$$\dot{u}(\overline{t} + 0) = -e\dot{u}(\overline{t} - 0)$$

by performing a precise analysis of the transmission of the energy by the scheme.

Possibly decreasing h_1, there exists a nonempty interval $[\tau_{-5}, \tau_2]$ containing $\overline{\tau}$ and included in $[h, (P-1)h]$. The apparently strange notations τ_{-5} and τ_2 have been chosen in view of the upcoming construction of Lemmas 6.1 and 6.2, where we will consider relative times

$$\tau_{-5} < \cdots < \tau_{-1} < \overline{\tau} < \tau_1 < \tau_2.$$

Define

$$P = \lceil \tau_{-5}/h \rceil + 1 \text{ and } Q = \lfloor \tau_2/h \rfloor - 1.$$

The measure $\ddot{u}_h$ is a sum of Dirac measures on $(t_0 + \tau_{-5}, t_0 + \tau_2)$. We define two measures ω_h and λ_h on $(t_0 + \tau_{-5}, t_0 + \tau_2)$ by

$$\omega_h(t) = \sum_{m=P}^{Q} \frac{\big(-2y^m + (1-e)y^{m-1}\big)^+}{h} \delta(t - t_0 - mh)$$

and

$$\lambda_h(t) = \sum_{m=P}^{Q} h\lambda^m \delta(t - t_0 - mh).$$

We have

$$\ddot{u}_h = \omega_h + \lambda_h,$$

and it is obvious that ω_h is a nonnegative measure.

Since the real numbers λ^m are bounded independently of h and m, the measure by $|\lambda_h|$ of any subinterval $[a, b]$ of $(t_0 + \tau_{-5}, t_0 + \tau_2)$ is bounded by $C_3(b - a + h)$, and it is clear therefore that there exists a function $\lambda \in L^\infty(t_0 + \tau_{-5}, t_0 + \tau_2)$ and a subsequence λ_h converging to λ in the weak * topology of $M^1\big((t_0 + \tau_{-5}, t_0 + \tau_2)\big)$.

The measure ω_h converges in the weak * topology of $M^1\big((t_0 + \tau_{-5}, t_0 + \tau_2)\big)$ to a nonnegative measure ω and in the limit

$$\ddot{u} = \omega + \lambda, \tag{6.1}$$

while

$$|\lambda|_{L^\infty} \leq C_3. \tag{6.2}$$

Since u is nonnegative on $(t_0+\tau_{-5}, t_0+\tau_2)$ and $u(\bar t)$ vanishes, we must have

$$\dot u(\bar t+0) \geq 0, \quad \dot u(\bar t-0) \leq 0.$$

On the other hand, $\dot u(\bar t+0) - \dot u(\bar t-0)$ is equal to $\omega(\{\bar t\})$; if $\omega(\{\bar t\})$ vanishes, we have

$$\dot u(\bar t+0) = \dot u(\bar t-0) = 0,$$

and the identity

$$\dot u(\bar t+0) = -e\dot u(\bar t-0)$$

holds. Therefore, the only interesting case is when

$$\omega(\{\bar t\}) > 0. \tag{6.3}$$

The following two lemmas enable us to prove in two steps that the velocity is reversed according to the law described by (1.5). Lemma 6.1 shows that if ω has a Dirac mass at $\bar t$, then the left velocity at $\bar t$ is outgoing; Lemma 6.2 indeed shows that (1.5) holds.

LEMMA 6.1. *If $\omega(\{\bar t\})$ is strictly positive, then $\dot u(\bar t-0)$ is strictly negative.*

Proof. The idea of the proof is to find two successive times $t^{m-1} \leq t^m < \bar t$ for which we can write down an estimate on the discrete velocities and then to use Lemma 2.1 to perform a discrete integration and to obtain a contradiction. We must deal with the fact that $\dot u_h$ does not converge uniformly to $\dot u$.

Without loss of generality, we may assume that $\dot u$ is continuous on the right and that for all $h \leq h_1$, $\dot u_h$ is also continuous from the right. According to Helly's theorem, there exists a countable set D such that

$$\dot u_h(t) \to \dot u(t) \quad \forall t \text{ such that } t-\bar t \in (\tau_{-5}, \tau_2) \setminus D.$$

Assume that $\dot u(\bar t-0)$ vanishes; therefore, $\dot u(\bar t+0)$ is strictly positive. Choose $\alpha = \dot u(\bar t+0)/4$, and let τ_{-4} and τ_1 be such that

$$\begin{aligned} &\tau_{-5} \leq \tau_{-4} < \overline{\tau} < \tau_1 \leq \tau_2, \\ &6C_3\big(\tau_1 - \tau_{-4}\big) \leq \alpha, \end{aligned} \tag{6.4}$$

and

$$\omega\big([t_0+\tau_{-4}, \bar t)\big) \leq \alpha, \quad \omega\big((\bar t, t_0+\tau_1]\big) \leq \alpha. \tag{6.5}$$

An integration of (6.1) on appropriate intervals yields

$$\forall t \in (t_0+\tau_{-4}, \bar t), \quad |\dot u(t \pm 0)| \leq \alpha + C_3(\bar t - t), \tag{6.6}$$

$$\forall t \in (\bar t, t_0+\tau_1), \quad \dot u(t \pm 0) \geq 2\omega\big(\{\bar t\}\big) - \alpha - C_3(t-\bar t). \tag{6.7}$$

Choose $\tau_{-3} \in (\tau_{-4}, \overline{\tau}) \setminus D$ and $\tau_{-1} \in (\tau_{-3}, \overline{\tau}) \setminus D$; since ω_h is a nonnegative measure, we have the following inequality for all $\tau' \in (\tau_{-3}, \tau_{-1})$ and all $\tau'' \in (\tau', \tau_{-1})$:

$$\begin{aligned} |\dot u_h(t_0+\tau') - \dot u_h(t_0+\tau'')| &\leq \omega_h((t_0+\tau', t_0+\tau'']) + C_3(\tau''-\tau'+h) \\ &\leq \omega_h([t_0+\tau_{-3}, t_0+\tau_{-1}]) + C_3(\tau''-\tau'+h). \end{aligned}$$

We integrate $\omega_h - \omega$ on the interval $\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]$; since the measures ω and ω_h do not charge $t_0+\tau_{-3}$ and $t_0+\tau_{-1}$, we find that

$$\begin{aligned}&\omega_h\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big) - \omega\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big)\\ &\quad = \dot u_h\big(t_0+\tau_{-1}\big) - \dot u_h\big(t_0+\tau_{-3}\big) - \dot u\big(t_0+\tau_{-1}\big) + \dot u\big(t_0+\tau_{-3}\big)\\ &\quad + \lambda\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big) - \lambda_h\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big),\end{aligned}$$

and therefore

$$\begin{aligned}&\omega_h\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big)\\ &\quad \le \omega\big(\big[t_0+\tau_{-3}, t_0+\tau_{-1}\big]\big) + |\dot u_h(t_0+\tau_{-1}) - \dot u(t_0+\tau_{-1})|\\ &\quad + |\dot u_h(t_0+\tau_{-3}) - \dot u(t_0+\tau_{-3})| + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big)+h\big).\end{aligned}$$

Choose now $\tau_{-2} \in \big(\tau_{-3}, \tau_{-1}\big) \setminus D$; then, for h small enough, $t^m = h\lfloor \tau_2/h\rfloor$ and $t^{m-1} = t^m - h$ belong to the interval (τ_{-3}, τ_{-1}), and therefore

$$\big|\dot u_h(t^m) - \dot u_h(t^{m-1})\big| \le \alpha + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big)+3h\big) + \varepsilon_h, \tag{6.8}$$

where ε_h tends to 0 as h tends to 0. On the other hand, $\dot u_h(t_0+\tau_{-2})$ tends to $\dot u(t_0+\tau_{-2})$ and therefore, thanks to relation (6.6), there exists a family ε'_h such that

$$|\dot u_h(t_0+\tau_{-2})| = |\dot u_h(t^m)| \le \alpha + C_3\big(\overline\tau - \tau_{-2}\big) + \varepsilon'_h,$$

which is equivalent to

$$|\eta^m| \le \alpha + C_3\big(\overline\tau - \tau_{-2}\big) + \varepsilon'_h; \tag{6.9}$$

we infer from (6.8) and (6.9) that

$$\big|\eta^{m-1}\big| \le 2\alpha + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big) + \overline\tau - \tau_{-2} + 3h\big) + \varepsilon_h + \varepsilon'_h.$$

Thus, for all $n \ge m$ we infer from Lemma 2.1 that

$$\begin{aligned}|\eta^n| \le{}& 2\alpha + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big) + 3h\\ &+ \overline\tau - \tau_{-2} + 2\big(t^n - t^m\big)\big) + \varepsilon_h + \varepsilon'_h.\end{aligned}$$

Therefore, in the limit, for all $t \ge t_0 + \tau_{-2}$

$$|\dot u(t)| \le 2\alpha + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big) + \overline\tau - \tau_{-2} + 2\big(t - t_0 - \tau_{-2}\big)\big),$$

and for all $t \in [t_0+\tau_{-2}, t_0+\tau_1]$

$$|\dot u(t)| \le 2\alpha + C_3\big(2\big(\tau_{-1}-\tau_{-3}\big) + \overline\tau - \tau_{-2} + 2\big(\tau_1 - \tau_{-2}\big)\big). \tag{6.10}$$

On the other hand, relation (6.7) implies that for all $t \in (\overline t, t_0+\tau_1)$,

$$|\dot u(t)| \ge 3\alpha - C_3\big(\tau_1 - \overline\tau\big). \tag{6.11}$$

Under assumption (6.4), relation (6.11) contradicts relation (6.10). □

We can conclude now the local study of the reflection of the velocity by the following lemma.

LEMMA 6.2. *If* $\omega\big(\{\overline{t}\}\big)$ *is strictly positive, then*

$$\dot{u}(\overline{t}) = -e\dot{u}(\overline{t}-0). \tag{6.12}$$

Proof. Since $\dot{u}(\overline{t}-0)$ is strictly negative, there exists a real number τ_{-3} such that $u(t)$ is strictly positive on $[t_0+\tau_{-3},\overline{t}) \subset [t_0+\tau_{-5},\overline{t})$. For all $\tau_{-2} \in (\tau_{-3},\overline{\tau})$, there exists $\tau_{-1} \in (\tau_{-2},\overline{\tau})$ and $h_1 > 0$ such that

$$\forall h \in (0,h_1], \quad \forall t \in [t_0+\tau_{-2}, t_0+\tau_{-1}), \quad u_h(t) \geq \frac{u(t_0+\tau_{-2})}{2}. \tag{6.13}$$

We prove now that there exists a maximal integer

$$m \in \{\lfloor \tau_{-3}/h \rfloor, \dots, \lfloor (\tau_2)/h \rfloor\}$$

such that

$$\forall l \in \big\{\lfloor \tau_{-3}/h \rfloor, \dots, m-1\big\}, \quad 2y^l - (1-e)y^{l-1} \geq 0, \tag{6.14}$$

and denoting

$$\rho_h = t^{m-1} - t_0, \tag{6.15}$$

the time ρ_h satisfies

$$\lim_{h\to 0} \rho_h = \overline{\tau}. \tag{6.16}$$

Let us first observe that for all small enough h and all t_l belonging to $[t_0+\tau_{-2}, t_0+\tau_{-1})$ we have

$$2y^l - (1-e)y^{l-1} \geq 0. \tag{6.17}$$

Indeed,

$$\begin{aligned} 2y^l - (1-e)y^{l-1} &= (1+e)y^l + (1-e)h\eta^{l-1} \\ &\geq \frac{1+e}{2}u(t_0+\tau_{-2}) - h(1-e)A, \end{aligned}$$

and if $2A(1-e)h \leq (1+e)u(t_0+\tau_{-2})$, we can see that (6.17) holds. Therefore m exists and

$$\liminf \rho_h \geq \overline{\tau}.$$

On the other hand, if there existed $\tau_1 > \overline{\tau}$ such that for all $t^m \in [t_0+\tau_{-3}, t_0+\tau_1]$ we had (6.17), then ω_h would vanish on $(t_0+\tau_{-3}, t_0+\tau_1)$, which contradicts assumption (6.3). Therefore, we have shown that

$$\limsup \rho_h \leq \overline{\tau},$$

i.e., (6.16). We integrate (4.4) discretely, and we find that for $t \in \big[t_0+\tau_{-3}, t_0+\rho_h\big]$

$$\begin{aligned} u_h(t) =& u_h(t_0+\rho_h) - (t_0+\rho_h-t)\dot{u}_h(t_0+\rho_h) \\ &+ \int_t^{t_0+\rho_h} \lambda_h((s, t_0+\rho_h])\, ds. \end{aligned} \tag{6.18}$$

In the limit we have

$$u(t) = u(t_0+\overline{\tau}) - (\overline{\tau}+t_0-t)\lim_{h\downarrow 0}\dot{u}_h(t_0+\rho_h+0) + \int_t^{t_0+\overline{\tau}}\int_s^{t_0+\overline{\tau}}\lambda(r)\,dr\,ds. \tag{6.19}$$

The comparison of (6.18) and (6.19) shows that

$$\lim_{h\downarrow 0}\dot{u}_h(t_0+\rho_h+0) = \lim_{h\downarrow 0}\eta^{m-1} = \dot{u}(\overline{t}-0). \tag{6.20}$$

Our purpose now is to obtain very precise estimates on the behavior of y_h beyond $t_0+\rho_h$. Thanks to the maximality of m, we have the relation

$$y^{m+1} = -ey^{m-1} + h^2\lambda^m; \tag{6.21}$$

let us estimate $2y^{m+1}-(1-e)y^m$. We substitute the value of y^{m+1} given by (6.21) into this expression, and we also use (4.4) with m replaced by $m-1$; we find that

$$\begin{aligned} &2y^{m+1}-(1-e)y^m \\ &\quad = -\bigl[2y^{m-1}-(1-e)y^{m-2}\bigr] - (1-e)h^2\lambda^{m-1} + 2h^2\lambda^m. \end{aligned}$$

We apply relation (2.1) for $n=m+1$ and we find that

$$\begin{aligned} \eta^{m+1}+e\eta^{m-1} &= h\bigl(\lambda^{m+1}-\lambda^m\bigr) \\ &\quad + \bigl\{-\bigl[2y^{m-1}-(1-e)y^{m-2}\bigr]h^{-1} - (1-e)h\lambda^{m-1} + 2h\lambda^m\bigr\}^+. \end{aligned}$$

Therefore, we have

$$\eta^{m+1}+e\eta^{m-1} \geq -2hC_3.$$

On the other hand, if $\xi = -\bigl[2y^{m-1}-(1-e)y^{m-2}\bigr]h^{-1} - (1-e)h\lambda^{m-1} + 2h\lambda^m$ is less than or equal to 0,

$$\bigl|\eta^{m+1}+e\eta^{m-1}\bigr| \leq 2hC_3;$$

if ξ is positive, then the sign condition on $2y^{m-1}-(1-e)y^{m-2}$ implies that

$$\eta^{m+1}+e\eta^{m-1} \leq h\bigl(\lambda^{m+1}+\lambda^m\bigr) - (1-e)h\lambda^{m-1}.$$

Thus, we have shown that

$$\bigl|\eta^{m+1}+e\eta^{m-1}\bigr| \leq 3C_3h. \tag{6.22}$$

If e is strictly positive, then for all small enough h,

$$\eta^{m+1} \geq e\,\bigl|\dot{u}(\overline{t}-0)\bigr|\,/2.$$

Let us estimate now the expression $2y^{m+2}-(1-e)y^{m+1}$: we have

$$2y^{m+2}-(1-e)y^{m+1} = -e\bigl[2y^m-(1-e)y^{m-1}\bigr] + O(h^2).$$

If $2y^{m+2}-(1-e)y^{m+1}$ is nonnegative, then

$$y^{m+3} = 2y^{m+2} - y^{m+1} + h^2\lambda^{m+2}.$$

We must estimate $2y^{m+3} - (1-e)y^{m+2}$:

$$\begin{aligned} &2y^{m+3} - (1-e)y^{m+2} - 2y^{m+2} + (1-e)y^{m+1} \\ &= h\big(2\eta^{m+2} - (1-e)\eta^{m+1}\big) \\ &= h(1+e)\eta^{m+1} + 2h^2\lambda^{m+2}, \end{aligned}$$

and therefore $2y^{m+3} - (1-e)y^{m+2}$ is nonnegative for all small enough h; the repetition of the argument shows that there exists $\theta > 0$ such that for all small enough h and all $n \in \{m+2, \dots, m + \lfloor \theta/h \rfloor\}$, the expression $2y^{n+1} - (1-e)y^n$ is nonnegative, and thus we have the relations

$$y^n = y^{m+1} + h(n-m-1)\eta^{m+1} + \sum_{j=m+2}^{n-1} \big(n-j\big)h^2\lambda^j.$$

On the other hand, if $2y^{m+2} - (1-e)y^{m+1}$ is negative, we must have

$$y^m = -\frac{(1-e)h\eta^{m-1}}{1+e} + O(h^2),$$

and therefore

$$y^{m-1} = -\frac{2h\eta^{m-1}}{1+e} + O(h^2).$$

These relations and the assumption on the sign of $2y^{m+2} - (1-e)y^{m+1}$ imply that

$$2y^{m+3} - (1-e)y^{m+2} = -\frac{\big(4e^2 + e(1-e)^2\big)h\eta^{m-1}}{1+e} + O(h^2), \tag{6.23}$$

which is strictly positive for h small enough. But now, we can see that

$$y^{m+3} - y^{m+2} = -eh\eta^{m-1} + O(h^2),$$

which is strictly positive for small enough h, and therefore $2y^{m+4} - (1-e)y^{m+3}$ is strictly positive for h small enough, since

$$2y^{m+4} - (1-e)y^{m+3} \geq -he(1+e)\eta^{m-1} + O(h^2);$$

the same argument as above shows now that there exists $\theta > 0$ such that for all $n \in \{m+3, \dots, m + \lfloor \theta/h \rfloor\}$,

$$y^n = y^{m+2} + h(n-m-2)\eta^{m+2} + \sum_{j=m+3}^{n-1} \big(n-j\big)h^2\lambda^j.$$

If we let $\rho'_h = t^{m+1} - t_0$ in the first case and $\rho'_h = t^{m+2} - t_0$ in the second case, we have now for $\rho'_h \leq t - t_0 \leq \rho'_h + \theta - h$

$$u_h(t) = u_h(t_0 + \rho'_h) + (t - \rho'_h - t_0)\dot{u}_h(t_0 + \rho'_h) + \int_{t_0+\rho'_h}^{t} \lambda_h((s,t])\, ds \tag{6.24}$$

and

$$u_h(t_0 + \rho'_h) = O(h), \quad \dot{u}_h(t_0 + \rho'_h) = -e\eta^{m-1} + O(h). \tag{6.25}$$

Passing to the limit in (6.24), we can see that

$$\dot{u}(\bar{t}+0) = -e\dot{u}(\bar{t}-0).$$

If we assume now that e vanishes, relation (6.22) implies

$$\eta^{m+1} = O(h).$$

We observe that Lemma 2.1 implies that for all n

$$|\eta^n| \leq \left|\eta^{n-1}\right| + 2C_3 h,$$

which implies immediately that for $n \geq m+1$

$$|\eta^n| \leq \left|\eta^{m+1}\right| + 2hC_3(n-m-1),$$

which proves by a straightforward passage to the limit that

$$\dot{u}(\bar{t}+0) = 0.$$

This completes the proof of the lemma. □

7. Initial conditions. In this section we prove that the solution that we have constructed satisfies the initial conditions; we work under the hypotheses stated at the beginning of section 4.

LEMMA 7.1. *The function u satisfies the initial conditions*

$$u(t_0) = u_0, \quad \dot{u}(t_0+0) = v_0.$$

Proof. By uniform convergence of u_h to u, it is clear that $u(t_0)$ is equal to u_0. There remains to show that the initial condition on the velocity is satisfied.

Assume first $u_0 > 0$; then there exist $h_1 > 0$ and $\tau_1 > 0$ such that for all $h \in (0, h_1]$ and for all $t - t_0 \in [0, \tau_1]$

$$u_h(t) \geq u_0/2.$$

Then, for all $t^m - t_0 \in (0, \tau_1]$, $2y^m - (1-e)y^{m-1} + h^2F^m$ belongs to K for h small enough; we indeed have

$$\begin{aligned} 2y^m - (1-e)y^{m-1} + h^2F^m &\geq (1+e)y^m - (1-e)hA - h^2C_3 \\ &\geq (1+e)u_0/2 - (1-e)hA - h^2C_3, \end{aligned}$$

which is strictly positive for h small enough. Thus the constraints are not active for $t_0 \leq t^m \leq t_0 + \tau_1$ and the convergence is clear.

In the second case, u_0 vanishes; we have taken admissible initial conditions, so that

$$v_0 \geq 0.$$

Let us show that

$$\dot{u}(t_0+0) = v_0,$$

considering two cases: $v_0 > 0$ and $v_0 = 0$. When v_0 vanishes, we have

$$y^1 = y^0 + hv_0 + O(h^2) = O(h^2)$$

and

$$\begin{aligned} y^2 &= -ey^0 + \big(2y^1 - (1-e)y^0\big)^+ + h^2\lambda^1 \\ &= 2\big(y^1\big)^+ + h^2\lambda^1 = O(h^2). \end{aligned}$$

Thus,

$$\eta^0 = O(h), \quad \eta^1 = O(h),$$

and relation (2.3) implies

$$|\eta^m| \le O(h) + 2C_3h(m-1);$$

therefore, a passage to the limit immediately gives

$$\dot{u}(t_0 + 0) = 0.$$

If, on the other hand, v_0 is strictly positive, then

$$2y^1 - (1-e)y^0 = 2y^1 = 2hv_0 + O(h^2),$$

which is strictly positive if h is small enough. Let $\{1, \dots, m\}$ be the maximal interval such that

$$2y^n - (1-e)y^{n-1} > 0 \quad \text{if } n \le m.$$

Then, for all $n \in \{1, \dots, m\}$,

$$\eta^n - \eta^{n-1} = h\lambda^n,$$

which implies by discrete integration that

$$\eta^n \ge \eta^0 - hnC_3,$$

as long as n belongs to $\{1, \dots, m\}$. Moreover, if we choose any $\tau_1 < v_0/(2C_3)$ and if n is at most equal to $\min\big(m, \lfloor \tau_1/h \rfloor\big)$, we can see that

$$y^n = y_0 + h\big(\eta^0 + \cdots + \eta^{n-1}\big) \ge \frac{hnv_0}{2}$$

for all small enough values of h.

In particular, for all $n \le \min\big(m, \lfloor \tau_1/h \rfloor\big)$,

$$2y^n - (1-e)y^{n-1} \ge \frac{(1+e)hnv_0}{2} - (1-e)hA,$$

which proves that m is at least equal to $\lfloor \tau_1/h \rfloor$. Therefore, ω_h vanishes on the interval $(t_0, t_0 + \tau_1 - h)$; in the limit, ω vanishes on $(t_0, t_0 + \tau_1)$ and therefore

$$\dot{u}(t_0 + 0) = v_0,$$

which completes the proof of the lemma. □

8. A priori estimates. In this section we prove that solutions of the problem (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), and (1.7) satisfy an a priori estimate on an interval with nonempty interior.

LEMMA 8.1. *Let R be strictly larger than $|v_0|$. Then there exists $\tau(R) > 0$ such that for all solution u of* (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), *and* (1.7) *defined on $[t_0, t_0 + \tau]$, the following estimates hold:*

$$\forall t \in [t_0, t_0 + \min(\tau, \tau(R))], \quad |u(t) - u_0| \le R, \quad |\dot u(t)| \le R. \tag{8.1}$$

Proof. The measure μ appearing in (1.3b) can be decomposed in the sum of an atomic part μ_a and a diffuse part μ_d. There might be a continuous singular part in the measure μ, and it is convenient to see μ as the sum of the derivative of a jump function and of the derivative of a continuous function. At each point of the support of μ_a we have

$$|\dot u(t+0)| \le |\dot u(t-0)| \tag{8.2}$$

thanks to relation (1.5). On any interval (t_1, t_2) which does not intersect the support of μ_a, we multiply relation (1.2) by $\dot u$, and we find that

$$\frac{d}{dt}\frac{1}{2}|\dot u|^2 = \dot u f(\cdot, u, \dot u). \tag{8.3}$$

Define

$$z = |\dot u|.$$

Relations (8.2) and (8.3) imply that in the sense of measures

$$z\dot z \le \dot u f(\cdot, u, \dot u). \tag{8.4}$$

Our purpose now is to transform (8.4) into a differential inequality. We write

$$\dot u f(t, u, \dot u) = \dot u\big[f(t, u, \dot u) - f(t, u_0, 0) + f(t, u_0, 0)\big].$$

Define

$$g(t) = |f(t, u_0, 0)|,$$

fix $R > |v_0|$, and let $\omega(R)$ be the Lipschitz constant of $(u, v) \mapsto f(t, u, v)$ for $t \in [0, T]$ and $\max(|u - u_0|, |v|) \le R$. By construction, ω is continuous and it is an increasing function of R.

If $t_0 \le t \le t_0 + \tau$ and if $\max(|u(t) - u_0|, |\dot u(t)|) \le R$ on $[t_0, t_0 + \tau]$, we have the inequality

$$z\dot z \le |\dot u f(\cdot, u, \dot u)| \le z\big(g + \omega(R)(|u - u_0| + |\dot u|)\big).$$

But we can estimate $u(t) - u_0$:

$$|u(t) - u_0| \le \int_{t_0}^{t} |\dot u(s)|\, ds \le \int_{t_0}^{t} z\, ds.$$

Therefore we have the estimate

$$|\dot u f(\cdot, u, \dot u)| \le zg + z\omega(R)\left(\int_{t_0}^{t} z\, ds + z\right),$$

and we conclude that z satisfies the differential inequality

$$\dot{z} \leq g + \omega(R) \left[\int_{t_0}^{t} z \, ds + z \right].$$

Let $\widehat{z}$ be the solution of the integro-differential equation

$$\frac{d\widehat{z}}{dt} = g + \omega(R) \left[\int_{t_0}^{t} \widehat{z} \, ds + \widehat{z} \right], \quad \widehat{z}(t_0) = |v_0|.$$

Such a $\widehat{z}$ exists and is unique, by very classical arguments, and it is also a majorant of z. Let $\tau(R)$ be the largest number in $(0, T - t_0]$ such that

$$\forall t \in [t_0, t_0 + \tau(R)], \quad \widehat{z}(t) \leq R, \quad \int_{t_0}^{t} \widehat{z}(s) \, ds \leq R.$$

Such a number exists since $\widehat{z}(t_0)$ is strictly inferior to R. On the interval $[t_0, t_0 + \min(\tau, \tau(R))]$ we have the desired estimate. □

9. Global results. We summarize the results obtained so far in the following proposition.

PROPOSITION 9.1. *Assume that there exist strictly positive numbers τ, A, and $h_1 > 0$ and a sequence of solutions of the numerical scheme defined by* (1.10), (1.11), (1.12), *and* (1.13), *which satisfies the estimate* (4.1). *Then it is possible to extract from the sequence u_h defined by* (5.1) *a subsequence which converges to a solution of* (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), *and* (1.7). *The convergence holds in the following sense: u_h converges uniformly to u_h on $[t_0, t_0 + \tau]$; $\dot{u}_h$ converges to $\dot{u}$ in $L^\infty(t_0, t_0 + \tau)$ weakly $*$ and almost everywhere on $[t_0, t_0 + \tau]$; and $\ddot{u}_h$ converges to $\ddot{u}$ in the weak $*$ topology of measures. Moreover, for all $\overline{\tau} \in (0, \tau]$, we have the following convergence:*

$$\begin{aligned} &\limsup_{h \downarrow 0} \, \sup \big\{ |\eta^m| : t_0 \leq t^m \leq t_0 + \overline{\tau} \big\} \\ &\qquad \leq \operatorname{ess\,sup} \big\{ |\dot{u}(t)| : t_0 \leq t \leq t_0 + \overline{\tau} \big\}. \end{aligned} \tag{9.1}$$

Proof. The only statement which deserves a proof is the last one; if it is not true, there exists $\gamma > 0$, a sequence of time steps still denoted by h, and a sequence of integers $m(h)$ such that

$$\left| \eta^{m(h)} \right| \geq \operatorname{ess\,sup} \big\{ |\dot{u}(t)| : t_0 \leq t \leq t_0 + \overline{\tau} \big\} + \gamma. \tag{9.2}$$

Without loss of generality, we may assume that $hm(h)$ tends to $\tau_2 \in [0, \overline{\tau}]$.

First, τ_2 cannot be equal to 0: we have learnt in section 7 that there exists a constant C_4 and a time τ_1 such that for all $h \leq h_1$ and all $m \leq \tau_1/h$,

$$\left| \eta^m - \eta^0 \right| \leq C_4 mh.$$

In particular, this estimate implies that

$$\left| \eta^{m(h)} \right| = |v_0| + O(mh);$$

but $|v_0|$ is at most equal to $\operatorname{ess\,sup} \big\{ |\dot{u}(t)| : t_0 \leq t \leq t_0 + \tau \big\}$, which contradicts (9.2). In the same fashion, we cannot have $u(t_0 + \tau_2) > 0$; if it were the case, we could find an

interval $[\tau_1, \tau_3]$ containing τ_2 and $h_1 > 0$ such that for all $h \in (0, h_1]$, $u_h([\tau_1, \tau_3])$ belongs to the interior of K. But, in this case, $\dot{u}_h$ converges uniformly to $\dot{u}$ in $C^0([\tau_1, \tau_3])$ and this contradicts again (9.2).

Thus, we assume that τ_2 is strictly positive and that $u(t_0 + \tau_2)$ vanishes.

We infer from (2.3) that

$$|\eta^{m+1}| \leq \max\bigl(|\eta^m|, |\eta^{m-1}|\bigr) + 2C_3 h.$$

We now use (9.2). We can see that for all $m \leq m(h)$,

$$|\eta^{m(h)}| \leq \max\bigl(|\eta^m|, |\eta^{m-1}|\bigr) + 2C_3\bigl(m(h) - m\bigr)h,$$

so that

$$\max\bigl(|\eta^m|, |\eta^{m-1}|\bigr) \geq |\eta^{m(h)}| - 2C_3\bigl(m(h) - m\bigr)h.$$

If $\tau_4 < \tau_2$ is such that

$$\tau_2 - \tau_4 \leq \gamma/4C_3,$$

we can see that for all $m \in \{\lceil \tau_4/h \rceil, \ldots, m(h)\}$, the following estimate holds:

$$\max\bigl(|\eta^m|, |\eta^{m-1}|\bigr) \geq \operatorname{ess\,sup}\bigl\{|\dot{u}(t)| : t_0 \leq t \leq t_0 + \overline{\tau}\bigr\} + \gamma/2. \tag{9.3}$$

But the function $|\dot{u}_h|$ converges almost everywhere on $[t_0, t_0 + \tau]$ to $|\dot{u}|$, and so does $\max(|\dot{u}_h(\cdot - h)|, |\dot{u}_h|)$. Therefore, in the limit, relation (9.3) leads to

$$\begin{aligned}\liminf_{h \downarrow 0} \operatorname{ess\,sup}\bigl\{|\dot{u}_h(t)| : t \in [t_0 + \tau_4, t_0 + \tau_2]\bigr\}\\ \geq \operatorname{ess\,sup}\bigl\{|u(t)| : t_0 \leq t \leq t_0 + \overline{\tau}\bigr\} + \gamma/2,\end{aligned}$$

which is a contradiction. □

A corollary can be inferred imediately from this proposition and Proposition 3.4.

COROLLARY 9.2. *For all admissible initial conditions (t_0, u_0, p_0), there exists $\tau > 0$ and a solution of* (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), *and* (1.7) *defined on $[t_0, t_0 + \tau]$.*

We have proved above the existence of a nonempty interval on which the numerical scheme converges to a solution of (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), and (1.7). On the other hand, Lemma 8.1 gives a priori estimates on the solution of such a problem.

We couple now the a priori estimates with the local convergence result to obtain a global result.

THEOREM 9.3. *Let R be strictly larger than $|v_0|$, and let $\tau(R)$ be given as in Lemma* 8.1. *Then, for all small enough h, the solution y^m of the numerical scheme* (1.10), (1.11), (1.12), (1.13) *is defined on a discrete interval $\{0, \ldots, m(h)\}$ such that*

$$\liminf_{h \to 0} hm(h) \geq \tau(R);$$

moreover, the approximation u_h converges to a solution u of the continuous time equation, i.e., (1.2), (1.3a), (1.3b), (1.4a), (1.4b), (1.5), (1.6), *and* (1.7), *which is defined on $[t_0, t_0 + \tau(R)]$.*

Proof. Let A be given by

$$A = \max\bigl(3R+1, u_0 + T(3R+1)\bigr).$$

Let $\{0, \dots, m(h)\}$ be the maximal discrete time interval for which the numerical scheme (1.10), (1.11), (1.12), (1.13) has a solution satisfying the estimate

$$\forall m \in \bigl\{0, \dots, m(h)-1\bigr\}, \quad |y^{m+1} - y^m| \leq Ah.$$

Let

$$\tau_1 = \liminf_{h \to 0} hm(h).$$

We know from Proposition 3.4 that τ_1 is at least equal to some number $\tau > 0$. Assume that τ_1 is strictly inferior to $\tau(R)$. Proposition 9.1 implies in particular that for all $\varepsilon > 0$

$$\begin{aligned}\limsup_{h \to 0} \bigl\{\sup|\eta^m| : t_0 \leq t^m \leq t_0 + \tau_1 - \varepsilon\bigr\} \\ \leq \operatorname{ess\,sup}\bigl\{|\dot{u}(t)| : t_0 \leq t \leq t_0 + \tau_1 - \varepsilon\bigr\} \leq R,\end{aligned}$$

thanks to the a priori estimates proved in Lemma 8.1. Since the above inequalities hold for all $\varepsilon > 0$, we see that

$$\limsup_{h \to 0} \bigl\{\sup|\eta^m| : t_0 \leq t^m \leq t_0 + \tau_1\bigr\} \leq R.$$

Let

$$a = \max\bigl(R + 1/2, u_0 + T(R+1/2)\bigr);$$

Theorem 3.1 implies the existence of $\tau_2 > 0$ such that if $\widehat{y}^0$ and $\widehat{y}^1$ satisfy property $P(a,h)$, and $\widehat{t}_0$ is any time in $[0,T)$, then there exists a numerical solution of (1.12), (1.13) which satisfies

$$\forall m \in \{0, \dots, \lfloor \tau_2/h \rfloor\}, \quad \bigl|\widehat{y}^{m+1} - \widehat{y}^m\bigr| \leq Ah.$$

We denote

$$l(h) = \lfloor (\tau_1 - \tau_2/2)/h \rfloor,$$

and we initialize with the following choices:

$$\widehat{t}_0 = t_0 + hl(h), \quad \widehat{y}^0 = y^{l(h)}, \quad \widehat{y}^1 = y^{l(h)+1}.$$

With these data, we know that $\widehat{y}^m$ exists for $0 \leq mh \leq \tau_2$, so that the numerical solution y^m is extended up to $\lfloor (\tau_1 + \tau_2/2)/h \rfloor - 1$, and therefore,

$$\liminf_{h \to 0} hm(h) \geq \tau_1 + \tau_2/2,$$

which is a contradiction. ∎

10. Numerical experiments. In this section, we report about the numerical implementation of our scheme and of the impact detection method; we compare these results to results obtained for a penalized version of our model, using a freely available scientific computation package.

In view of its practical importance and of its ease of programming, we have limited ourselves to vibro-impact, i.e.,

$$f(t,u,\dot{u}) = a\cos\omega t - u - 2\alpha\dot{u}$$

with constraint set

$$K = [u_{\min}, +\infty).$$

The following numerical values are kept constant throughout the numerical experiments:

$$\alpha = 0.5, \quad a = 1, \quad \omega = 50, \quad e = 0.5.$$

Observe that $u_{\min}/a$ is the relevant parameter. We have observed in previous work [17], [19], [18] that if we systematically choose

$$u(0) = u_{\min}, \quad \dot{u}(0) = 0.1,$$

the variation of the parameter $u_{\min}$ triggers a variety of dynamical behaviors: periodic solutions, period doubling, and chaotic attractors.

All the penalty computations presented here have been implemented as `SCILAB` programs, a free high level scientific computation software developed and distributed by INRIA (http://www-rocq.inria.fr/scilab). Some of the other computations have been performed in FORTRAN.

10.1. Implementation particulars. The impact detection scheme goes as follows: starting from initial data t_j, $u(t_j) = u_{\min}$, $\dot{u}(t_j)$, the solution of the linear problem

$$\ddot{u} + 2\alpha\dot{u} + u = a\cos\omega t$$

is found explicitly; a nonlinear solver finds the first zero $t_{j+1} > t_j$ of $t \mapsto u(t) - u(t_j)$ and this instant is called t_{j+1}. We let

$$u(t_{j+1} + 0) = -eu(t_{j+1} - 0),$$

and we restart the process.

The foregoing description is slightly too rough: if there is an accumulation of impact instants, we have to define a threshold of velocity at t_j, under which we set the solution equal to $u_{\min}$, as long as $\cos\omega t$ remains negative. Observe that the detection method is potentially accurate to machine precision, since the nonlinear solvers for a scalar function are extremely precise, and the threshold velocity can be taken very small.

The numerical scheme is implemented as follows:

$$\begin{aligned} y^{n+1} &= -ey^{n-1} + \max\bigl((1+e)u_{\min}, x^n\bigr), \\ x^n &= \frac{h^2 a}{1+\alpha h}\cos(\omega t_n) + \frac{2-h^2}{1+\alpha h}y^n - \frac{(1-e)-\alpha h(1+e)}{1+\alpha h}y^{n-1}. \end{aligned}$$

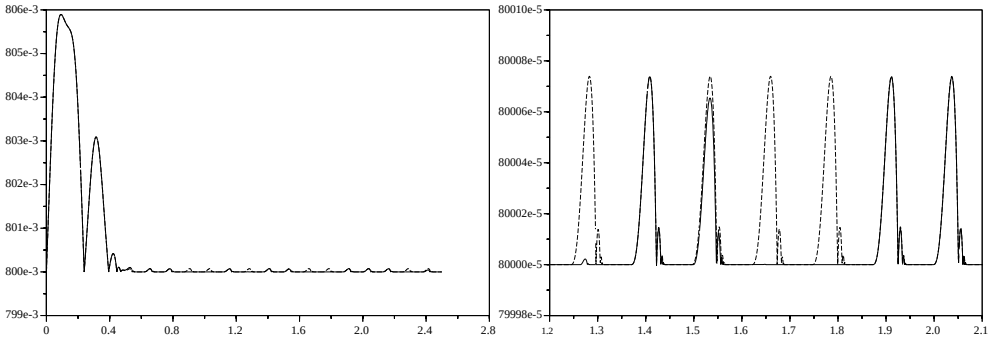


FIG. 10.1. *The penalized solution on the left and its zoom on the right for time step* $h = 1.25\,10^{-5}$ *and* $\varepsilon = 10^{-6}/256$.

The reader will check that this form is equivalent to (1.12) and (1.13); one of the advantages of this numerical scheme is that it is absolutely trivial to program.

For the penalty method, we applied the `SCILAB` function `ode` with the option `stiff` to the ordinary differential system

$$\begin{aligned} &\dot u = v,\\ &\dot v = a\cos\omega t - u - 2\alpha v + \frac{(u-u_{\min})^-}{\varepsilon} + 2\frac{\beta\,\mathrm{sign}^-\,(u-u_{\min})}{\sqrt{\varepsilon}}, \end{aligned} \tag{10.1}$$

where

$$r^- = -\min(r,0), \quad \mathrm{sign}^-\,(r) = \begin{cases} -1 & \text{if } r<0,\\ 0 & \text{otherwise.}\end{cases}$$

The parameter β is defined in terms of the restitution coefficient e by

$$\beta = -\frac{\ln e}{\sqrt{\pi^2 + (\ln e)^2}}.$$

It has been proved in [21] that the solution of (10.1) converges to a solution of (1.2), (1.3a), (1.3b), (1.4a), (1.4b), and (1.5). The choice of β is also justified in that reference.

10.2. Periodic solution: $u_{\min} = 0.8$. The solution obtained by the scheme and impact detection agree satisfactorily: they both converge as time increases to a periodic solution, with an infinite number of impacts per period—of which we calculate only a finite number, of course!

The penalized approximation is satisfactory for not too small values of the penalty parameter, but for very small values of the parameter, it completely misses some of the periods (see Figure 10.1). The choice of parameters corresponds to $\sqrt{\varepsilon}/h = 5$, which is quite sufficient for a good numerical approximation of the rebound.

As a matter of comparison we show in Figure 10.2 that the scheme and the detection method coincide very precisely.

We cannot offer much in the way of the explanations, since we did not go into the details of the `SCILAB` code to understand its inner workings; though it is an open package, with freely accessible code, we treated it as a black box.

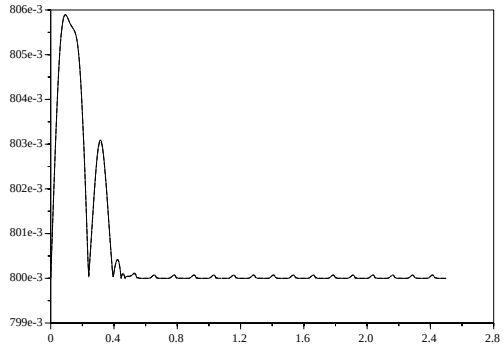


FIG. 10.2. *The scheme for* $h = 0.0003125$ *superposed to the solution by detection.*

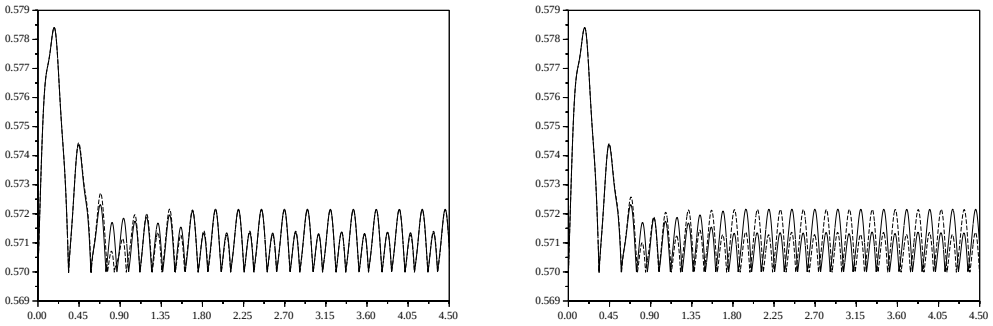


FIG. 10.3. *On the left: detection (solid) and scheme (dotted); on the right: detection (solid) and penalty (dotted).*

10.3. Period doubling: $u_{\min} = 0.57$. We observe period doubling slightly above this value, and the scheme continues to agree satisfactorily with the detection method.

The penalty method gives also period doubling, but, depending on ε and the time step, we may obtain either a good approximation of the period doubled solution obtained by the previous two approaches or its translate by a half (doubled) period, i.e., $2\pi/\omega$.

For instance, Figure 10.3 shows the results of the calculation with the detection scheme, our ad hoc scheme with a time step $h = 5\,10^{-4}$, and the penalty method for $\varepsilon = 10^{-7}$ with a time step $h = 2\,10^{-5}$.

The results of the numerical scheme do not seem to depend on the time step, once convergence is experimentally achieved; see Figure 10.4, left, where the solution is significantly improved by decreasing the time step from $6.25\,10^{-4}$ to $5.\,10^{-4}$. In contrast, on the right, decreasing the time step with the same penalty parameter $\varepsilon = 10^{-6}$ does not give a significant improvement: for the larger time parameter, the beginning of the numerical solution is bad, and for the smaller one, we hit a phase shifted solution.

The penalized solution is also very sensitive to the choice of the penalty parameter; see Figure 10.5, with the same time step of $2\,10^{-5}$ and penalty parameters of 10^{-7} and 10^{-8}.

The results of the penalty method keep depending on the choice of the time step and ε (see Figures 10.4, right, and 10.5), and we have not been able to establish the

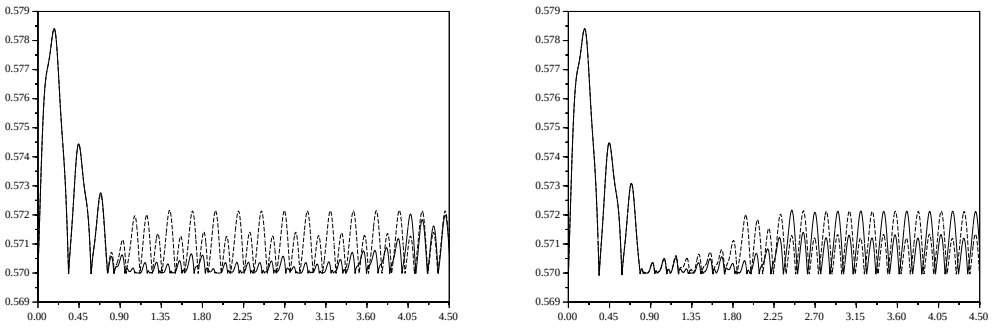


FIG. 10.4. *Left: the scheme with* $h = 6.25\,10^{-4}$ *(solid line) and with* $h = 5\,10^{-4}$ *(dashed line); right: the penalty method with* $h = 2\,10^{-5}, \varepsilon = 10^{-6}$ *(solid line) and with* $h = 10^{-4}, \varepsilon = 10^{-6}$ *(dashed line).*

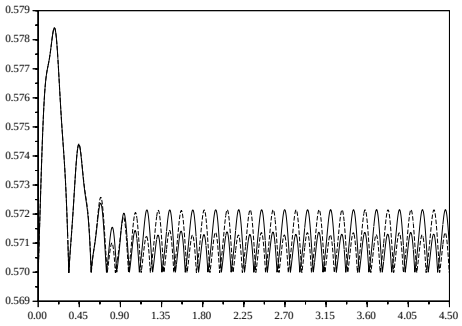


FIG. 10.5. *The penalty method with* $h = 2\,10^{-5}, \varepsilon = 10^{-8}$ *(solid line) and with* $h = 2\,10^{-5}, \varepsilon = 10^{-7}$ *(dashed line).*

pattern of dependency.

10.4. Strange attractor: $u_{\min} = 0.54$. We picture a stroboscopic view of the attractors by displaying the sequences

$$\big(u((2k+1)\pi/\omega), \dot{u}((2k+1)\pi/\omega)\big)$$

for 1989 values of k.

Figure 10.6 displays a superposition of the computation by the detection method and the scheme, in FORTRAN double precision. Figure 10.7 displays a superposition of the `SCILAB` computation by the penalty method and by the scheme; it is somewhat surprising to see so few points of the scheme in this last computation, while the sizes of vectors are identical: k varies from 1 to 1989, corresponding to a final time of 250 (and not to the bicentennial of the French Revolution). We believe that this phenomenon may be due to a bad control in `SCILAB` of the format of the numbers.

Nevertheless, these figures show a very satisfactory agreement between the three methods.

10.5. Numerical conclusion. We would like to stress the qualitative properties of numerical schemes in a dynamical systems framework: this scientific program has been started by several authors, and we refer to [31] for an overview of the subject. However, the methods of analysis rely heavily on a smoothness assumption which is not satisfied here, and therefore, they do not apply. Thus, at the present moment, we

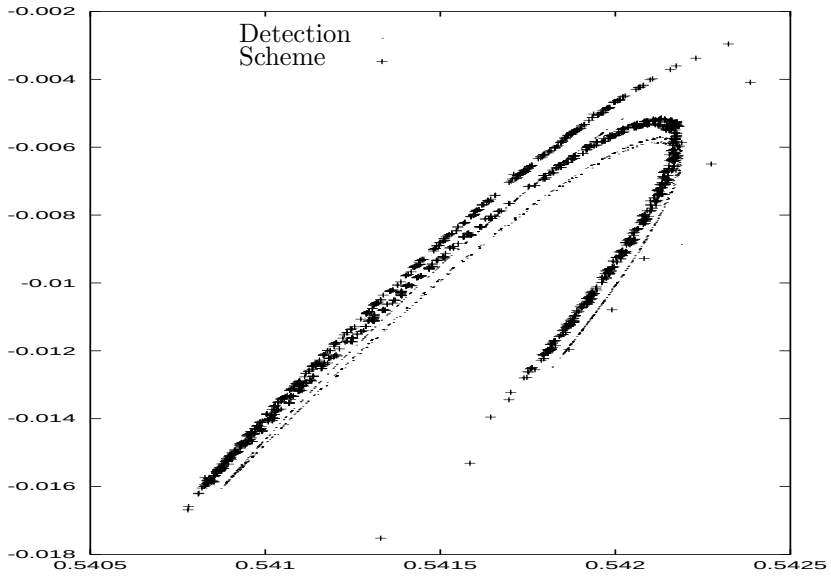


FIG. 10.6. *A superposition of the stroboscopic pictures of the attractor obtained by the detection method (dots) and the scheme (points):* `FORTRAN` *calculation with time step* 510^{-4}.

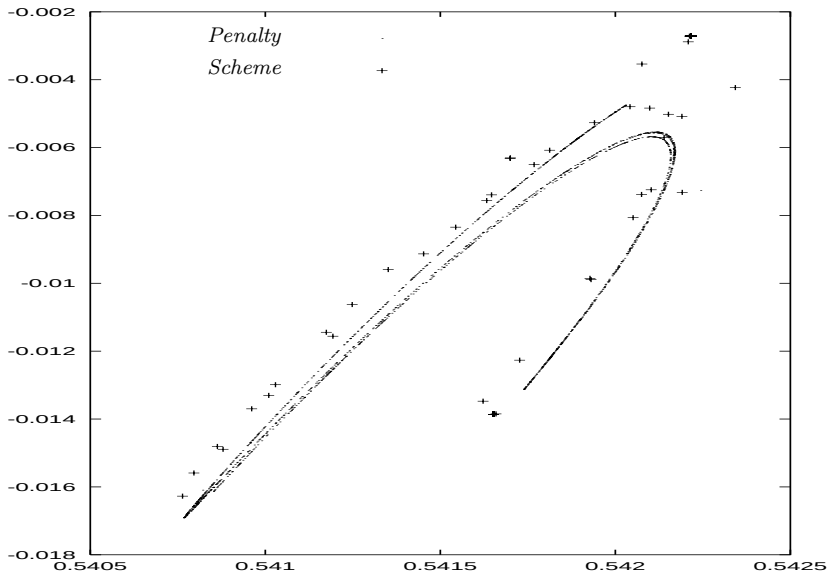


FIG. 10.7. *A superposition of the stroboscopic pictures of the attractor obtained by the penalty method (dots) and the scheme (points):* `SCILAB` *calculation with time step* $\pi/2500$.

are reduced to experimental numerics, but we should keep in mind a rational approach to the numerical analysis of dynamical systems, focusing not only on the accuracy of a given method for finite time intervals but also on the qualitative properties of the method for long time ranges. The provisional conclusion for our one-dimensional case is that the penalty method performs reasonably well, but it misses details, while the

detection method enables us to benchmark our ad hoc scheme and to certify that it does not miss details. Our scheme is quite easy to implement, and it contains no black box; however, it is still missing some bells and whistles, such as time step control. In any case, our scheme is of very low order: order one with respect to the position and order zero with respect to the velocity. The reason is that the velocity is discontinuous and the numerical impact times are usually distinct from their limit. Schemes with a better order of convergence should approximate very precisely the impact times, but one cannot but wonder whether it is really necessary to do so.

REFERENCES

[1] H. Brézis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1983.

[2] G. Buttazzo and D. Percivale, *On the approximation of the elastic bounce problem on Riemannian manifolds*, J. Differential Equations, 47 (1983), pp. 227–245.

[3] N. Dunford and J. T. Schwartz, *Linear operators. Spectral theory. Selfadjoint operators in Hilbert space. Part* II, John Wiley and Sons, New York, 1988.

[4] M. Kunze and M. D. P. Monteiro Marques, *On the discretization of degenerate sweeping processes*, Portugal. Math., 55 (1998), pp. 219–232.

[5] M. Laghdir and M. D. P. Monteiro Marques, *Dynamics of a particle with damping, friction, and percussional effects*, J. Math. Anal. Appl., 196 (1995), pp. 902–920.

[6] M. Laghdir and M. D. P. Monteiro Marques, *Measure-differential inclusions in percussional dynamics*, J. Convex Anal., 4 (1997), pp. 381–393.

[7] M. Mabrouk, *Liaisons unilatérales et chocs élastiques quelconques: un résultat d'existence*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1353–1357.

[8] M. Mabrouk, *A unified variational model for the dynamics of perfect unilateral constraints*, Eur. J. Mech. A Solids, 17 (1998), pp. 819–842.

[9] M. D. P. Monteiro Marques, *Differential Inclusions in Nonsmooth Mechanical Problems. Shocks and Dry friction*, Birkhäuser Verlag, Basel, 1993.

[10] M. D. P. Monteiro Marques, *An existence, uniqueness and regularity study of the dynamics of systems with one-dimensional friction*, Eur. J. Mech. A Solids, 13 (1994), pp. 277–306.

[11] J.-J. Moreau, *Les liaisons unilatérales et le principe de Gauss*, C. R. Acad. Sci. Paris, 256 (1963), pp. 871–874.

[12] J.-J. Moreau, *Rafle par un convexe variable.* I, Secrétariat des Math. 118, U.É.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, France, 1971.

[13] J.-J. Moreau, *Rafle par un convexe variable.* II, Secrétariat des Math. 122, Univ. Sci. Tech. Languedoc, Montpellier, France, 1972.

[14] J.-J. Moreau, *Evolution problem associated with a moving convex set in a Hilbert space*, J. Differential Equations, 26 (1977), pp. 347–374.

[15] J.-J. Moreau, *Application of convex analysis to some problems of dry friction*, in Trends in Applications of Pure Mathematics to Mechanics, Vol. II Pitman, Boston, 1979, pp. 263–280.

[16] J.-J. Moreau, *Liaisons unilatérales sans frottement et chocs inélastiques*, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 296 (1983), pp. 1473–1476.

[17] M. Panet, L. Paoli, and M. Schatzman, *Vibrations with an obstacle and a finite number of degrees of freedom*, in Proceedings of the International Symposium on Identification of Nonlinear Mechanical Systems from Dynamic Tests—Euromech 280, L. Jezequel and C.-H. Lamarque, eds., Balkema, Rotterdam, 1992.

[18] M. Panet, L. Paoli, and M. Schatzman, *Theoretical and numerical study for a model of vibrations with unilateral constraints*, in Contact Mechanics, M. Raous, M. Jean, and J.-J. Moreau, eds., Plenum Press, New York, 1995, pp. 457–464.

[19] L. Paoli, *Analyse numérique de vibrations avec contraintes unilatérales*, Ph.D. thesis, Université Claude Bernard—Lyon 1, Lyon, France, 1993.

[20] L. Paoli and M. Schatzman, *A numerical scheme for impact problems:* II. *The multidimensional case*, SIAM J. Numer. Anal., 40 (2002), pp. 734–768.

[21] L. Paoli and M. Schatzman, *Mouvement à un nombre fini de degrés de liberté avec contraintes unilatérales: cas avec perte d'énergie*, M2AN Math. Model. Numer. Anal., 27 (1993), pp. 673–717.

[22] L. Paoli and M. Schatzman, *Schéma numérique pour un modèle de vibrations avec contraintes unilatérales et perte d'énergie aux impacts, en dimension finie*, C. R. Acad. Sc. Paris Sér

I Math., 317 (1993), pp. 211–215.
[23] L. Paoli and M. Schatzman, *Resonance in impact problems. Recent advances in contact mechanics*, Math. Comput. Modelling, 28 (1998), pp. 385–406.
[24] D. Percivale, *Bounce problem with weak hypotheses of regularity*, Ann. Mat. Pura Appl. (4), 143 (1986), pp. 259–274.
[25] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1997.
[26] M. Schatzman, *A class of nonlinear differential equations of second order in time*, Nonlinear Anal., 2 (1978), pp. 355–373.
[27] D. E. Stewart, *A numerical method for friction problems with multiple contacts*, J. Austral. Math. Soc. Ser. B, 37 (1996), pp. 288–308.
[28] D. E. Stewart, *Convergence of a time-stepping scheme for rigid-body dynamics and resolution of Painlevé's problem*, Arch. Ration. Mech. Anal., 145 (1998), pp. 215–260.
[29] D. E. Stewart and J. C. Trinkle, *An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2673–2691.
[30] D. E. Stewart and J. C. Trinkle, *Dynamics, friction, and complementarity problems*, in Complementarity and Variational Problems (Baltimore, MD, 1995), SIAM, Philadelphia, PA, 1997, pp. 425–439.
[31] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 734–768

A NUMERICAL SCHEME FOR IMPACT PROBLEMS II: THE MULTIDIMENSIONAL CASE*

LAETITIA PAOLI† AND MICHELLE SCHATZMAN‡

Abstract. We consider a mechanical system with impact and n degrees of freedom, written in generalized coordinates. The system is not necessarily Lagrangian. The representative point is subject to a constraint: it must stay inside a closed set K with boundary of class C^3. We assume that, at impact, the tangential component of the impulsion is conserved, while its normal coordinate is reflected and multiplied by a given coefficient of restitution $e \in [0,1]$: the mechanically relevant notion of orthogonality is defined in terms of the local metric for the impulsions (local cotangent metric). We define a numerical scheme which enables us to approximate the solutions of the Cauchy problem: this is a generalization of the scheme presented in the companion paper [L. Paoli and M. Schatzman, *SIAM J. Numer. Anal.*, 40 (2002), pp. 702–733]. We prove the convergence of this numerical scheme to a solution, which also yields an existence result. Without any a priori estimates, the convergence and the existence are local; with some a priori estimates, the convergence and the existence are proved on intervals depending exclusively on these estimates. The technique of proof uses a localization of the scheme close to the boundary of K; this idea is classical for a differential system studied in the framework of flows of a vector field. It is much more difficult to implement here because finite differences schemes are only approximately local: straightening the boundary creates quadratic terms which cause all the difficulties of the proof.

Key words. impact, coefficient of restitution, numerical scheme, convergence, local existence, global existence

AMS subject classifications. Primary, 65J10, 65M20, 65B05; Secondary, 17B09, 46N20, 47D03

PII. S003614290037873X

1. Introduction.

In this article we study a numerical approximation of dynamics with impact with a finite number of degrees of freedom and a smooth constraint.

The set of constraints is denoted K and satisfies the following assumptions:

(1.1a) K is a closed subset of $\mathbb{R}^d$ with a nonempty interior;

(1.1b) $\begin{cases} \text{the boundary } \partial K \text{ of } K \text{ is an embedded submanifold} \\ \text{of class } C^3 \text{ of } \mathbb{R}^d; \end{cases}$

(1.1c) K lies on only one side of ∂K.

It is possible to find a function ϕ of class C^3 such that

$$K = \{u \in \mathbb{R}^d : \phi(u) \geq 0\}$$

and the differential $d\phi$ does not vanish on $\partial K = \{u \in \mathbb{R}^d : \phi(u) = 0\}$.

Let f be a continuous function from $[0,T]\times\mathbb{R}^d\times\mathbb{R}^d$ to $\mathbb{R}^d$ which is locally Lipschitz continuous with respect to its last two arguments, and let $M(u)$ be the mass matrix:

*Received by the editors September 27, 2000; accepted for publication (in revised form) October 26, 2001; published electronically July 24, 2002.
http://www.siam.org/journals/sinum/40-2/37873.html
†UMR 5585 CNRS et Equipe d'Analyse Numérique, Faculté des Sciences, Université Jean Monnet, 23 Rue du Docteur Paul Michelon, 42023 Saint-Etienne Cedex 2, France (paoli@anumsun1.univ-st-etienne.fr).
‡UMR 5585 CNRS-MAPLY, Université Lyon 1, 69622 Villeurbanne Cedex, France (schatz@maply.univ-lyon1.fr).

$u \mapsto M(u)$ is a mapping of class C^3 from $\mathbb{R}^d$ to the set of symmetric positive definite matrices.

The free dynamics of the system are written in generalized coordinates as

$$M(u)\ddot{u} = f(\cdot, u, p), \quad p = M(u)\dot{u}. \tag{1.2}$$

This system is more general than the system obtained in Lagrangian mechanics, since we want to include possible dissipative terms in the dynamics of the problem under discussion.

Let us give the few geometric notations which are absolutely necessary here, since we use a Riemannian metric: the cotangent bundle $T^*\mathbb{R}^d$ is identified with $\mathbb{R}^d \times \mathbb{R}^d$, and its elements are denoted as pairs (u, ξ); at each point u of $\mathbb{R}^d$ the metric tensor for tangent vectors is defined by the matrix $M(u)$, and the metric tensor for cotangent vectors is defined by the matrix $M(u)^{-1}$. The scalar product of two vectors x and y in the tangent space at u is denoted by $\langle x, y\rangle_u$; coordinatewise it can be expressed as $x^T M(u) y$, where x and y are column vectors. The scalar product of two vectors ξ and η in the cotangent space at u is denoted by $\langle \xi, \eta\rangle_u^*$ and coordinatewise it is equal to $\xi^T M(u)^{-1}\eta$. The corresponding norms of vectors and covectors are denoted, respectively, by $|x|_u$ and $|\xi|_u^*$.

Therefore, a cotangent vector (u, ξ) belonging to $T^*\mathbb{R}^d$ is orthogonal to the cotangent vector (u, η) iff $\langle \xi, \eta\rangle_u^*$ vanishes.

With these notations, if the velocity of the system is $\dot{u}$, the generalized impulsion is $M(u)\dot{u} = p$ and (u, p) belongs to the cotangent space $T^*\mathbb{R}^d$. Whenever we take the orthogonal of a vector or a vector subspace of the tangent or the cotangent space at u, we always use the relevant metric tensor; therefore it is important to know which of the vectors under consideration are cotangent and which are tangent. Of course, all the differential forms are cotangent vectors.

Let us describe now the system satisfied by the problem with impact: we replace (1.2) by

$$M(u)\ddot{u} = \mu + f(\cdot, u, p), \tag{1.3}$$

and since we cannot expect to have global solutions in general, μ is an unknown measure on $[t_0, t_0 + \overline{\tau}]$ with values in $\mathbb{R}^d$ which describes the reaction of the constraints and has the following properties: if $d\phi$ denotes the differential of ϕ, then

$$\operatorname{supp}(\mu) \subset \{t \in [t_0, t_0 + \overline{\tau}] : \phi(u(t)) = 0\}, \tag{1.4a}$$

$$\mu = \lambda d\phi(u), \tag{1.4b}$$

$$\lambda \geq 0 \quad |d\phi(u)| \text{ almost everywhere on } [t_0, t_0 + \overline{\tau}]. \tag{1.4c}$$

We require the following functional properties for u:

$$\begin{cases} u \text{ is a continuous function taking its values in } K \\ \forall\ t \in [t_0, t_0 + \overline{\tau}], \end{cases} \tag{1.5a}$$

$$\dot{u} \text{ is of bounded variation over } [t_0, t_0 + \overline{\tau}]. \tag{1.5b}$$

If $\dot{u}$ is of bounded variation, p is also of bounded variation. Assume that $u(t)$ belongs to ∂K; we decompose $p(t-0)$ and $p(t+0)$ on $\mathbb{R}d\phi(u(t)) \oplus d\phi(u(t))^\perp$; here the $\perp$ sign means the orthogonality with respect to the local cotangent metric. We integrate (1.3) on a small neighborhood of t, and relation (1.4b) implies that the component of $p(t-0)$ on $d\phi(u(t))^\perp$ is conserved.

Therefore, we have to make a supplementary assumption in order to have a complete description of the impact; we choose a constitutive law of the impact using a coefficient of restitution. Thus we will assume that there exists $e \in [0,1]$ such that the component of $p(t+0)$ along $\mathbb{R}d\phi(u)$ is equal to $-e$ times the component of $p(t-0)$ on $\mathbb{R}d\phi(u)$. In other words, we have

$$p(t+0) = p(t-0) - (1+e)\frac{\langle d\phi(u(t)), p(t-0)\rangle^*_{u(t)}}{\langle d\phi(u(t)), d\phi(u(t))\rangle^*_{u(t)}} d\phi(u(t)). \tag{1.6}$$

The set of admissible initial data $\mathbb{D}$ will be

$$\begin{aligned}\mathbb{D} = \Big\{ &(t_0, u_0, p_0) \in [0,T) \times K \times \mathbb{R}^d : \\ &\text{if } u_0 \in \partial K, \text{ then } \langle p_0, d\phi(u_0)\rangle^*_{u_0} \geq 0 \Big\}.\end{aligned} \tag{1.7}$$

This choice is equivalent to the convention that there is no impact at the initial time t_0.

Given initial conditions $(t_0, u_0, p_0) \in \mathbb{D}$, we require that the following Cauchy data be satisfied:

$$u(t_0) = u_0 \tag{1.8}$$

and

$$p(t_0) = p_0. \tag{1.9}$$

For all initial data $(t_0, u_0, p_0) \in \mathbb{D}$ we will obtain the existence of a local solution to (1.3), (1.4a), (1.4b), (1.4c), and (1.6) belonging to the functional class defined by (1.5a) and (1.5b) and satisfying the initial conditions (1.8) and (1.9).

The existence of this local solution is obtained by defining a numerical scheme, whose convergence will be shown in appropriate functional spaces; the limit of the approximation will be a solution of our problem.

The distance on $\mathbb{R}^d$ is defined with the help of the Riemannian metric: if $s \mapsto u(s)$ is a C^1 mapping from $[a,b]$ to $\mathbb{R}^d$, the Riemannian length of the image of u is

$$\ell(u) = \int_a^b |\dot{u}(s)|_{u(s)}\, ds.$$

This curve length is invariant by a diffeomorphic change of parameter. Therefore, we may assume that $a = 0$ and $b = 1$. The distance from x to y is the lower bound of the length of the curves from x to y, or in other words

$$\operatorname{dist}(x,y) = \inf\{\ell(u) : \quad u \in C^1([0,1]), \quad u(0) = x, \quad u(1) = y\}.$$

It is classical that the lower bound is attained on the geodesics for the given Riemannian metric; it is also known that for each point x there exists $r > 0$ such that, if $\operatorname{dist}(x,y) \leq r$, there is only one geodesic from x to y.

We denote by $\operatorname{dist}(x,E)$ the Riemannian distance of a point x to a set E.

Under assumptions (1.1), a projection on ∂K can be defined uniquely on an appropriate neighborhood of ∂K; more precisely, for all compact $\mathcal{C} \subset \partial K$, there exists a neighborhood of $\mathcal{C}$ on which the projection $P_{\partial K}$ is uniquely defined, and

there exists a unique geodesic joining a point of this neighborhood to its projection. This projection $P_{\partial K}$ is characterized by the relation

$$\forall y \in \partial K, \quad \operatorname{dist}(P_{\partial K} x, x) \leq \operatorname{dist}(y, x). \tag{1.10}$$

This projection is of class C^2.

For all x in ∂K, denote by $N(x)$ the interior unit normal vector: this means that $|N(x)|_x$ is equal to 1 and that it is orthogonal to the tangent space at $P_{\partial K} x$ with respect to the scalar product in the tangent space; i.e., for all y such that $d\phi(x)y$ vanishes, $\langle y, N(x)\rangle_x = 0$. The smoothness of ∂K implies that the mapping $z \mapsto N(z)$ is of class C^2.

When the geodesic from x to $P_{\partial K} x$ is unique it is tangent at $P_{\partial K} x$ to $N(P_{\partial K} x)$.

Starting from this projection on ∂K, we can define a projection on K as follows: for each compact $\mathcal{C}$ included in K, there exists a relatively compact neighborhood $\mathcal{U}$ of $\mathcal{C}$ on which P_K is defined by

$$P_K(x) = \begin{cases} P_{\partial K}(x) & \text{if } x \notin K, \\ x & \text{otherwise.} \end{cases} \tag{1.11}$$

The reader will check that P_K is Lipschitz continuous over $\mathcal{U}$ and that $P_K x$ realizes the minimum of the distance from x to K.

Given two positive numbers $h^* \leq 1$ and T, assume that F is a continuous function from $[0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times [0, h^*]$ to $\mathbb{R}^d$, which is locally Lipschitz continuous with respect to its second, third, and fourth arguments; assume, moreover, that F is consistent with f, i.e., that for all $t \in [0,T]$, for all u and v in $\mathbb{R}^d$

$$F(t, u, u, v, 0) = M(u)^{-1} f(t, u, M(u)v). \tag{1.12}$$

We approximate the solution of (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.8), (1.9) by the following numerical scheme: the initial values U^0 and U^1 are given by the initial position

$$U^0 = u_0, \tag{1.13}$$

and the position at the first time step

$$U^1 = u_0 + hM(u_0)^{-1} p_0 + hz(h), \tag{1.14}$$

where $z(h)$ tends to 0 as h tends to 0.

We will systematically use, henceforth, the notation

$$t_m = t_0 + mh. \tag{1.15}$$

Given U^{m-1} and U^m, U^{m+1} is defined by the relations

$$U^{m+1} = -eU^{m-1} + (1+e) P_K \left(\frac{2U^m - (1-e)U^{m-1} + h^2 F^m}{1+e} \right) \tag{1.16}$$

and

$$F^m = F\left(t_m, U^m, U^{m-1}, \frac{U^{m+1} - U^{m-1}}{2h}, h \right) \tag{1.17}$$

provided that U^{m+1} is unique in a neighborhood of U^m. Definition (1.16) and (1.17) is the obvious generalization of (1.12) and (1.13) of [9].

In the announcement [11], where we assumed that the set of constraints K was convex and the geometry was Euclidean, we had defined the numerical scheme by relation (1.16), where P_K was the Euclidean projection on the convex set K.

The boundary ∂K is smooth, and as we expect that for small h, the U^m's will stay close to K, we still have a projection of $(2U^m - (1-e)U^{m-1} + h^2F^m)/(1+e)$ on K, and thus we start from (1.16) to define the numerical scheme.

Let us define

$$W^m = \frac{2U^m - (1-e)U^{m-1} + h^2F^m}{1+e}, \tag{1.18}$$

which will be used in many places in the upcoming proofs. With this definition, (1.16) is rewritten as

$$U^{m+1} = -eU^{m-1} + (1+e)P_K(W^m).$$

Hence, if we define

$$Z^m = \frac{U^{m+1} + eU^{m-1}}{1+e}, \tag{1.19}$$

we find that

$$Z^m = P_K(W^m). \tag{1.20}$$

Another way of writing (1.16) is to define the discrete velocity V^m by

$$V^m = \frac{U^{m+1} - U^m}{h}. \tag{1.21}$$

Then, (1.16) can be rewritten as

$$V^m - V^{m-1} - hF^m = \frac{(1+e)(Z^m - W^m)}{h}. \tag{1.22}$$

A strict contraction argument in $\mathbb{R}^d$ gives the existence of a unique U^m for small values of m and h. As the projections on K and F are only Lipschitz continuous, the iteration of a fixed point argument might request smaller and smaller bounds on the time step h, and there is no guarantee that we could integrate numerically on a time interval bounded from below, for any initial time step size.

Let us outline now the structure of the article and of the proofs. In section 2, we will straighten the boundary, a natural geometrical idea.

While the system (1.3)–(1.6) is nicely transformed under a diffeomorphism, the numerical scheme (1.13), (1.14), (1.16), and (1.17) does not behave well under diffeomorphisms. The reason is that a numerical scheme is not a local object: when we define a discrete velocity by subtracting U^m from U^{m+1}, we use locally a vector structure which is not intrinsic from the point of view of differential geometry. In particular, if we apply a diffeomorphism to the numerical scheme, we will find another numerical scheme which will look much more complicated than the previous one, since it will contain a number of small term which show the lack of an intrinsic description of the scheme. After a very technical proof, we find two constants C_3 and

τ such that for initial data in a neighborhood of a compact subset of the admissible set, and for all small enough h and all $m \leq \tau/h$, the discrete velocity is bounded:

$$\sup |V^m| \leq C_3.$$

Since uniqueness is not true in general [2], [16], and hypotheses of analyticity are often but not always used for the proof of uniqueness [13], [15], [17], [4], [1], the proof of convergence of the numerical approximation is delicate also for this reason.

However, there is a bonus: all the effort made to prove the local convergence of the numerical scheme provides us with a local existence proof for our problem. In section 3 we prove estimates on the discrete acceleration; then in section 4 we establish the variational properties of the limit of the numerical scheme. Sections 5 and 6 are devoted to the study of the transmission of energy at impact and the passage to the limit for the initial conditions. All these results are obtained under the assumption that on a certain time interval starting at t_0, the discrete velocity is bounded independently of the time step.

As a preliminary to the global existence proof, we give a priori estimates on problem (1.3)–(1.9) in section 7, which is completely independent from the remainder of the article.

In section 8, we establish a very weak semicontinuity for the supremum of the local norm of the discrete velocities; this result enables us to obtain a global existence and convergence theorem.

A note on the strategy of proof seems necessary here: an essential device is the straightening of the boundary of K. However, it is not always convenient to work in the straightened coordinates; in particular, whenever we need very precise estimates involving P_K, the physical coordinates are better. This seeming inconsistency in the redaction of our proofs is a compromise whereby we tried to save space without losing clarity.

We present some numerical simulations in section 9: we simulated the dropped bar experiment of Stoianovici and Hurmuzlu [18], using generalized coordinates and therefore a nontrivial mass matrix. We compared the results of our ad hoc scheme to an impact detection method; the reader is referred to the original article [18] for a comparison with the experiments. Much more information on our simulations has been given in [12].

The numerical scheme analyzed here has been implemented in the case of a trivial mass matrix in [8], [7]. Our ad hoc scheme is substantially faster than an impact detection method.

The existence result obtained here is a generalization of [16], [3], [14], [8], [10], and of [6] and [5], who also constructs a solution with the help of a numerical method; however, he does not consider a nonconstant mass matrix.

We refer the reader to the introduction of [9] for more bibliographical references.

2. Existence of $(U^m)_{0 \leq m \leq \lfloor \tau/h \rfloor}$ for some $\tau > 0$. We systematically use the floor and ceiling notations: when r is a real number, the floor $\lfloor r \rfloor$ of r is the largest integer at most equal to r, and the ceiling $\lceil r \rceil$ is the smallest integer at least equal to r.

The main result of this section is the existence of a number $\tau > 0$ such that for all small enough h and all $m \leq \lfloor \tau/h \rfloor$ there exists indeed a discrete solution of (1.16) and (1.17), whose discrete velocity is bounded independently of h. In fact, we prove a stronger result: provided that the first two discrete velocities are bounded, we find a uniform lower bound on τ when the initial position belongs to a neighborhood of a

compact subset of K. As in the one dimensional case [9], the idea is to show the existence by Brouwer's fixed point argument and the uniqueness by local considerations. However, the proof is longer and much more technical: the cost of straightening the boundary is the appearance of a host of supplementary terms of geometrical origin that we have to control. At a crucial point of the proof, we use Lemma 2.1 of [9] to estimate the normal component of the velocity, thereby effectively reducing the N dimensional problem to the one dimensional case.

We say that U^0 and U^1 satisfy condition $E\big(\overline{u}, r_0, C_2, h\big)$ if

$$\big|U^0 - \overline{u}\big| \le r_0, \quad \big|U^1 - \overline{u}\big| \le r_0, \quad \big|U^1 - U^0\big| \le C_2 h, \tag{2.1}$$

and that, moreover, U^2 is uniquely defined in $B(\overline{u}, r_0)$ by

$$\frac{U^2 + eU^0}{1+e} = P_K\left(\frac{2U^1 - (1-e)U^0 + h^2 F^1}{1+e}\right), \tag{2.2}$$

and the following inequalities are satisfied:

$$\big|U^2 - \overline{u}\big| \le r_0, \quad \big|U^2 - U^1\big| \le C_2 h. \tag{2.3}$$

When $\overline{u}$ belongs to ∂K, we need local coordinates in which the projection P_K is particularly simple. They are defined in the following fashion: we choose a coordinate frame in $\mathbb{R}^d$ such that

- $\overline{u} = 0$;
- the tangent hyperplane to ∂K at 0 is the hyperplane of the first $d-1$ coordinates;
- the positive direction of the dth coordinate axis points inside K.

For a d dimensional vector x, we will use the notation

$$x' = (x_1, \dots, x_{d-1}).$$

Locally, ∂K is a graph over the hyperplane of the first $d-1$ coordinates, and it can be parameterized as

$$\chi(x') = \begin{pmatrix} x' \\ H(x') \end{pmatrix},$$

where x' belongs to $\mathbb{R}^{d-1}$, H is of class C^3, and $DH(0)$ vanishes. Let $s \mapsto \psi(s,z)$ be the parameterization of the geodesic starting at $z \in \partial K$ with an initial velocity equal to $-N(z)$ which satisfies

$$\left|\frac{\partial \psi}{\partial s}(s,z)\right|_{\psi(s,z)} = 1. \tag{2.4}$$

Let Ψ be defined by

$$\Psi(x', y) = \psi(-y, \chi(x')); \tag{2.5}$$

the function Ψ is of class C^2 in a neighborhood of 0; its derivative at 0 has the block representation

$$D\Psi(0,0) = \left(\begin{array}{c|c} \mathbf{1}_{d-1} & N(0) \\ 0 & \end{array}\right); \tag{2.6}$$

it is invertible, since $N(0)$ does not belong to the tangent plane at 0 to ∂K. Thus Ψ is a local diffeomorphism from a neighborhood $\mathcal{U}$ of 0 to a neighborhood $\Psi(\mathcal{U})$ of 0. In particular, we may assume that $\mathcal{U}$ contains a compact neighborhood of 0 of the form $\overline{\mathcal{O}} \times [-r_1, r_1]$, where $\mathcal{O}$ is an open neighborhood of 0 in $\mathbb{R}^{d-1}$.

The inverse diffeomorphism of Ψ is denoted by Φ, and we decompose it as

$$\Phi(x) = \begin{pmatrix} S(x) \\ Y(x) \end{pmatrix}, \tag{2.7}$$

where S takes its values in $\mathbb{R}^{d-1}$ and Y takes its values in $\mathbb{R}$. If x belongs to

$$\mathcal{F} = \Psi(\overline{\mathcal{O}} \times [-r_1, r_1]),$$

the projection $P_K(x)$ is given by

$$P_K(x) = \Psi \begin{pmatrix} S(x) \\ Y(x)^+ \end{pmatrix}. \tag{2.8}$$

With these preparations, we are able to prove the main local estimates.

THEOREM 2.1. *For all $\overline{u} \in K$, for all $C_2 > 0$, there exist two positive numbers, $r_1 < r_2$ and three numbers $\tau > 0$, $h_1 > 0$, and $C_3 < \infty$ such that for all $h \in (0, h_1]$ and all $t_0 \in [0, T)$, for all U^0 and U^1, satisfying the condition $E\big(\overline{u}, r_1, C_2, h\big)$, U^m is uniquely defined in $B(\overline{u}, r_2)$, for all $m \le \lfloor \min(\tau, T - t_0)/h \rfloor$, and $|V^m|$ is bounded by C_3 independently of h for $0 \le m \le \lfloor \min(\tau, T - t_0)/h \rfloor - 1$.*

Proof. The theorem decomposes into an easy part and a difficult part. The easy part is when $\overline{u}$ belongs to the interior of K.

First case: $\overline{u} \in \mathbf{int}(K)$. We choose $r_0 > 0$ such that the ball of center $\overline{u}$ and radius $2r_0$ is included in K. Let C_3 be greater than C_2. We define the numbers C_1 and L by

$$\begin{aligned} C_1 = \sup\big\{ |F(t, u, u', 0, h)| : t \in [0, T],\ |u - \overline{u}| \le 2r_0, \\ |u' - \overline{u}| \le 2r_0,\ h \in [0, h^*]\big\}, \end{aligned} \tag{2.9}$$

$$\begin{aligned} L = \sup\Bigg\{ \frac{|F(t, u, u', v, h) - F(t, u, u', v', h)|}{|v - v'|} : t \in [0, T],\ |u - \overline{u}| \le 2r_0, \\ |u' - \overline{u}| \le 2r_0, |v| \le C_3,\ |v'| \le C_3,\ h \in [0, h^*],\ v \ne v' \Bigg\}. \end{aligned} \tag{2.10}$$

We choose $r_1 = r_0/2$ and $r_2 = r_0$. Assume that τ satisfies the following inequalities:

$$\begin{aligned} &\tau\big(C_1 + LC_3 + 2LC_3\tau\big) < C_3 - C_2, \\ &\tau C_3 \le \frac{r_0}{2}, \quad 0 < \tau < T - t_0. \end{aligned} \tag{2.11}$$

Let h_1 be a nonnegative number such that

$$\frac{2h_1 C_3}{1 + e} + \frac{h_1^2}{1 + e}\big(C_1 + LC_3 + 2LC_3\tau\big) < r_0. \tag{2.12}$$

Let U^0, U^1 satisfy the condition $E\big(\overline{u}, r_1, C_2, h\big)$. We claim that for all $h \in (0, h_1]$ we can find a solution of

$$U^{m+1} - 2U^m + U^{m-1} = h^2 F\left(t_m, U^{m-1}, U^m, \big(V^m + V^{m-1}\big)/2, h\right) \tag{2.13}$$

for all $m \in \{1, \dots, \lfloor \tau/h \rfloor - 1\}$ which satisfies the estimates

$$\forall m \in \{0, \dots, \lfloor \tau/h \rfloor\}, \quad |U^m - \overline{u}| \leq r_2, \tag{2.14}$$

$$\forall m \in \{0, \dots, \lfloor \tau/h \rfloor - 1\}, \quad |V^m| \leq C_3. \tag{2.15}$$

In this construction, we seek a solution without considering the constraints, and we prove eventually that they are satisfied.

We will apply Brouwer's fixed point argument in order to prove the existence of a solution of (2.13). Let $h \in (0, h_1]$ and define a compact convex set B_h by

$$B_h = \Big\{ \widehat{U} = (\widehat{U}^m)_{0 \leq m \leq \lfloor \tau/h \rfloor} : \widehat{U}^0 = U^0, \widehat{U}^1 = U^1, \widehat{U}^2 = U^2, \\ \forall m \in \{1, \dots, \lfloor \tau/h \rfloor - 1\}, \quad \left| \widehat{U}^{m+1} - \widehat{U}^m \right| \leq C_3 h \Big\}.$$

Assuming that $\widehat{U}$ belongs to B_h, we define $\widehat{F}$ by

$$\forall m \in \{1, \dots, \lfloor \tau/h \rfloor - 1\}, \qquad \widehat{F}^m = F\left(t_m, \widehat{U}^m, \widehat{U}^{m-1}, \frac{\widehat{U}^{m+1} - \widehat{U}^{m-1}}{2h}, h \right).$$

We write now the numerical scheme

$$\forall m \in \{2, \dots, \lfloor \tau/h \rfloor - 1\}, \qquad U^{m+1} - 2U^m + U^{m-1} = h^2 \widehat{F}^m.$$

Since $\widehat{U}$ belongs to B_h, we estimate $\left| \widehat{F}^m \right|$ by

$$\left| \widehat{F}^m \right| \leq \left| F(t_m, \widehat{U}^0, \widehat{U}^0, 0, h) \right| + L \left(\left| \widehat{U}^m - \widehat{U}^0 \right| + \left| \widehat{U}^{m-1} - \widehat{U}^0 \right| + \left| \frac{\widehat{U}^{m+1} - \widehat{U}^{m-1}}{2h} \right| \right) \\ \leq C_1 + 2LC_3\tau + LC_3.$$

By definition of V^m we have

$$|V^m| = \left| V^{m-1} + h\widehat{F}^m \right| \leq \left| V^{m-1} \right| + h \left| \widehat{F}^m \right|,$$

and by discrete integration

$$|V^m| \leq \left| V^0 \right| + \tau\big(C_1 + 2LC_3\tau + LC_3\big) \leq C_3,$$

thanks to (2.11) and (2.1). Hence U belongs to B_h and the mapping $\widehat{U} \mapsto U$ is clearly continuous, which implies the existence of a fixed point in B_h thanks to Brouwer's fixed point theorem. This fixed point is a solution of (2.13) satisfying (2.15). Moreover, for all $m \in \{0, \dots, \lfloor \tau/h \rfloor\}$ we have

$$|U^m - \overline{u}| \leq \sum_{p=0}^{m-1} h\,|V^p| + \left| U^0 - \overline{u} \right| \leq C_3\tau + r_0/2 \leq r_0$$

thanks to (2.11), and thus (2.14) holds.

Let us prove that the vector W^m defined by (1.18) belongs to K: since

$$W^m - \overline{u} = U^{m-1} + \frac{2h}{1+e} V^{m-1} + \frac{h^2 F^m}{1+e} - \overline{u},$$

we have the estimate

$$|W^m - \overline{u}| \le r_0 + \frac{2hC_3}{1+e} + \frac{h^2}{1+e}\big(C_1 + LC_3 + 2L\tau C_3\big).$$

With (2.12) we infer that W^m belongs to K, and the sequence U^m satisfies (1.16)–(1.17).

Second case: $\overline{u} \in \partial K$. We define on $\mathbb{R}^d$ a norm denoted by $\|\ \|$ as follows:

$$x = \begin{pmatrix} x' \\ x_d \end{pmatrix}, \quad \|x\| = \max\big(|x'|\,, |x_d|\big).$$

Pick $R_1 > 0$ such that Ψ is a diffeomorphism from a closed neighborhood

$$\mathcal{B} = \big\{x : \|x - \Phi(\overline{u})\| \le R_1\big\}$$

to its image and such that $\Psi(\mathcal{B})$ is included in an Euclidean ball $B(\overline{u}, r_0)$ such that P_K is Lipschitz continuous on $B(\overline{u}, 2r_0)$; denote by γ the Lipschitz constant of P_K on $B(\overline{u}, 2r_0)$.

Define Λ by

$$\Lambda = \max\left\{\sup\left\{\frac{|D\Psi(x)x_1|}{\|x_1\|} : x \in \mathcal{B}, x_1 \ne 0\right\}, \sup\left\{\frac{|D\Phi(u)u_1|}{\|u_1\|} : u \in \Psi(\mathcal{B}), u_1 \ne 0\right\}\right\}$$

and

$$C_4 = \max\left\{\sup\left\{\frac{\big|D^2\Psi(x)x_1 \otimes x_2\big|}{2\,\|x_1\|\,\|x_2\|} : x \in \mathcal{B}, x_1 \ne 0, x_2 \ne 0\right\}, \sup\left\{\frac{\big|D^2\Phi(u)u_1 \otimes u_2\big|}{2\,\|u_1\|\,\|u_2\|} : u \in \Psi(\mathcal{B}), u_1 \ne 0, u_2 \ne 0\right\}\right\}.$$

A continuity argument shows that the compact set $\Psi(\mathcal{B})$ contains the ball of radius R_1/Λ about $\overline{u}$.

We will give now a description of the scheme (1.16), (1.17) in the new coordinates. We define

$$X^m = \Phi(U^m).$$

Assume, therefore, that

$$U^{m+1},\ U^m,\ U^{m-1},\ W^m, \text{ and } \frac{U^m + eU^{m-1}}{1+e} \text{ belong to } \Psi(\mathcal{B}). \tag{2.16}$$

We know that (1.16) is equivalent to

$$\frac{U^{m+1} + eU^{m-1}}{1+e} = P_K(W^m). \tag{2.17}$$

We map (2.17) by Φ, and we calculate the Taylor expansion of either side of (2.17) around U^m. The left-hand side of (2.17) can be rewritten as

$$U^m + h\frac{V^m - eV^{m-1}}{1+e},$$

and therefore

$$\Phi\left(U^m + h\frac{V^m - eV^{m-1}}{1+e}\right) = \Phi(U^m) + D\Phi(U^m)h\frac{V^m - eV^{m-1}}{1+e} + I^m, \tag{2.18}$$

where

$$\|I^m\| \le C_4 \left|\frac{h\big(V^m - eV^{m-1}\big)}{1+e}\right|^2.$$

But $U^{m+1} = U^m + hV^m$, so that another Taylor expansion gives

$$\Phi\big(U^{m+1}\big) = \Phi(U^m) + D\Phi(U^m)hV^m + \widehat{I}^m$$

with

$$\left\|\widehat{I}^m\right\| \le C_4 \left|hV^m\right|^2.$$

Thus

$$D\Phi(U^m)hV^m = \Phi(U^{m+1}) - \Phi(U^m) - \widehat{I}^m. \tag{2.19}$$

A similar calculation gives

$$-D\Phi(U^m)hV^{m-1} = \Phi(U^{m-1}) - \Phi(U^m) - \widetilde{I}^m, \tag{2.20}$$

with

$$\left\|\widetilde{I}^m\right\| \le C_4 \left|hV^{m-1}\right|^2.$$

If we substitute (2.19) and (2.20) into (2.18), we find that

$$\Phi\big((U^{m+1} + eU^{m-1})/(1+e)\big) = \frac{X^{m+1} + eX^{m-1}}{1+e} - \frac{\overline{I}^m}{1+e},$$

where

$$\overline{I}^m = \widehat{I}^m + e\widetilde{I}^m - (1+e)I^m,$$

and we have the estimate

$$\left\|\overline{I}^m\right\| \le C_4 h^2\big(|V^m|^2 + e\left|V^{m-1}\right|^2 + (1+e)^{-1}\left|V^m - eV^{m-1}\right|^2\big). \tag{2.21}$$

Consider now the right-hand side of (2.17). By definition of V^{m-1}, we have the identity

$$W^m = U^m + \frac{(1-e)hV^{m-1} + h^2F^m}{1+e}, \tag{2.22}$$

and a Taylor expansion gives

$$\Phi(W^m) = \Phi(U^m) + D\Phi(U^m)\frac{(1-e)hV^{m-1} + h^2F^m}{1+e} + J^m, \tag{2.23}$$

with

$$\|J^m\| \le C_4 \left| \frac{(1-e)hV^{m-1} + h^2 F^m}{1+e} \right|^2.$$

We substitute (2.20) into (2.23), and we obtain

$$\Phi(W^m) = \frac{2X^m - (1-e)X^{m-1} + h^2 D\Phi(U^m)F^m}{1+e} + \frac{\overline{J}^m}{1+e},$$

where

$$\overline{J}^m = (1+e)J^m + (1-e)\widetilde{I}^m,$$

so that

$$\left\|\overline{J}^m\right\| \le C_4 \left[\frac{\left|(1-e)hV^{m-1} + h^2 F^m\right|^2}{1+e} + (1-e)h^2 \left|V^{m-1}\right|^2 \right].$$

We have to estimate $\|\overline{I}^m\| + \|\overline{J}^m\|$; by elementary inequalities,

$$\left\|\overline{I}^m\right\| + \left\|\overline{J}^m\right\| \le C_4 h^2 \left[\frac{2(1-e)^2 \left|V^{m-1}\right|^2 + 2h^2 \left|F^m\right|^2}{1+e} \right.$$
$$\left. + (1-e)\left|V^{m-1}\right|^2 + \left|V^m\right|^2 + e\left|V^{m-1}\right|^2 + \frac{2\left|V^m\right|^2 + 2e^2\left|V^{m-1}\right|^2}{1+e} \right].$$

The coefficient of $\left|V^{m-1}\right|^2$ in the above bracket is

$$\frac{2(1-e)^2}{1+e} + 1 + \frac{2e^2}{1+e},$$

and since for $e \in [0,1]$, $(1-e)^2 \le 1-e^2$, this coefficient is at most equal to 3. The coefficient of $|V^m|^2$ in the same bracket is equal to $1+2/(1+e)$, which is also at most equal to 3. Therefore

$$\left\|\overline{I}^m\right\| + \left\|\overline{J}^m\right\| \le C_4 h^2 \left[3\left|V^m\right|^2 + 3\left|V^{m-1}\right|^2 + 2h^2\left|F^m\right|^2\right]. \tag{2.24}$$

Thanks to the properties of P_K,

$$\Phi\left(\frac{U^{m+1} + eU^{m-1}}{1+e}\right) = \Phi(P_K W^m) = \begin{pmatrix} S(W^m) \\ Y(W^m)^+ \end{pmatrix}. \tag{2.25}$$

Define

$$s^m = S(U^m) = \left[X^m\right]', \quad y^m = Y(U^m) = X_d^m.$$

In these new coordinates, we have

$$s^{m+1} - 2s^m + s^{m-1} = h^2 \kappa^m, \tag{2.26}$$

$$y^{m+1} + ey^{m-1} - \left(2y^m - (1-e)y^{m-1}\right)^+ = h^2 \lambda^m, \tag{2.27}$$

and κ^m and λ^m are given by

$$\begin{aligned} h^2\kappa^m &= \left[h^2 D\Phi(U^m)F^m + \overline{I}^m + \overline{J}^m\right]', \\ h^2\lambda^m &= \left[2y^m - (1-e)y^{m-1} + \left(h^2 D\Phi(U^m)F^m + \overline{J}^m\right)_d\right]^+ \\ &\qquad - \left(2y^m - (1-e)y^{m-1}\right)^+ + \overline{I}^m_d. \end{aligned}$$

Therefore, we have the estimates:

$$\max\left(|\kappa^m|, |\lambda^m|\right) \le \Lambda |F^m| + C_4\left(3|V^m|^2 + 3\left|V^{m-1}\right|^2 + 2h^2|F^m|^2\right). \tag{2.28}$$

We define ξ^m and ζ^m by

$$\xi^m = \begin{pmatrix} \sigma^m \\ \eta^m \end{pmatrix} = \frac{X^{m+1} - X^m}{h}, \quad \zeta^m = \begin{pmatrix} \kappa^m \\ \lambda^m \end{pmatrix}.$$

Now let q be a number which satisfies

$$q > \Lambda C_2.$$

Let $C_3 = \Lambda q$, and let C_1 and L be, respectively, as in (2.9) and (2.10). If we assume beyond (2.16) that

$$\max\left(\left|V^{m-1}\right|, |V^m|\right) \le C_3, \tag{2.29}$$

we have the estimate

$$|F^m| \le C_1 + L\left(|V^m| + \left|V^{m-1}\right|\right)/2;$$

by elementary inequalities,

$$|F^m|^2 \le 2C_1^2 + L^2\left(|V^m|^2 + \left|V^{m-1}\right|^2\right),$$

and therefore, if we define

$$C_5 = \frac{\Lambda}{2} + C_1^2\left(\Lambda + 4h_1^2 C_4\right), \quad C_6 = \left(\left(\frac{\Lambda}{2} + 2h_1^2 C_4\right)L^2 + 3C_4\right)\Lambda^2,$$

we have shown that under assumptions (2.16) and (2.29), the following inequality holds:

$$\|\zeta^m\| \le C_5 + C_6\left(\|\xi^m\|^2 + \left\|\xi^{m-1}\right\|^2\right). \tag{2.30}$$

Let τ be a number which satisfies the following inequalities:

$$\begin{aligned} &\Lambda C_2 + 2\tau(C_5 + 2C_6 q^2) < q, \\ &\tau q < R_1/3\Lambda^2, \quad 0 < \tau < T - t_0. \end{aligned} \tag{2.31}$$

Let h_1 be a nonnegative number such that

$$\max\left(\frac{1-e}{1+e}\Lambda q h_1 + \frac{h_1^2}{1+e}(C_1 + L\Lambda q), h_1 q\right) < \frac{R_1}{3\Lambda}. \tag{2.32}$$

We will prove that, for all $h \in (0, h_1]$ and all $t_0 \in [0, T)$, for all U^0, U^1 satisfying the condition $E\big(\overline{u}, r_1, C_2, h\big)$ with $r_1 = R_1/3\Lambda^3$, U^m is uniquely defined in $B(\overline{u}, r_2)$, for all $m \le \lfloor \tau/h \rfloor$ with $r_2 = 2R_1/3\Lambda$, and $|V^m|$ is bounded by C_3 for $0 \le m \le \lfloor \tau/h \rfloor - 1$.

We will apply once again Brouwer's fixed point argument. Let $h \in (0, h_1]$ and define a compact convex set B_h by

$$B_h = \Big\{ \widehat{X} = (\widehat{X}^m)_{0 \le m \le \lfloor \tau/h \rfloor} : \widehat{X}^0 = \Phi(U^0), \widehat{X}^1 = \Phi(U^1), \widehat{X}^2 = \Phi(U^2), \\ \forall m \in \{1, \dots, \lfloor \tau/h \rfloor - 1\}, \quad \Big\| \widehat{X}^{m+1} - \widehat{X}^m \Big\| \le qh \Big\}.$$

Let $\widehat{X}$ be in B_h. Since $\Lambda \ge 1$ and U^0, U^1 satisfy the condition $E\big(\overline{u}, r_1, C_2, h\big)$, we have

$$\{U^0, U^1, U^2\} \subset B(\overline{u}, r_1) \subset B(\overline{u}, R_1/\Lambda) \subset \Psi(\mathcal{B}),$$

and for all $m \in \{3, \dots, \lfloor \tau/h \rfloor\}$ we have

$$\begin{aligned} \Big\| \widehat{X}^m - \Phi(\overline{u}) \Big\| &\le \Big\| \widehat{X}^m - \widehat{X}^0 \Big\| + \Big\| \widehat{X}^0 - \Phi(\overline{u}) \Big\| \\ &\le \tau q + \Lambda \big| U^0 - \overline{u} \big| \le \frac{2R_1}{3\Lambda^2} < R_1. \end{aligned} \tag{2.33}$$

It follows that $\widehat{X}^m$ belongs to $\mathcal{B}$ for all $m \in \{0, \dots, \lfloor \tau/h \rfloor\}$. For all $m \in \{3, \dots, \lfloor \tau/h \rfloor\}$ we define

$$U^m = \Psi(\widehat{X}^m).$$

We clearly have $U^m \in \Psi(\mathcal{B})$ for all $m \in \{0, \dots, \lfloor \tau/h \rfloor\}$ and

$$\big| U^{m+1} - U^m \big| = \Lambda \Big\| \widehat{X}^{m+1} - \widehat{X}^m \Big\| \le \Lambda q h \le C_3 h.$$

Hence we have the following estimate:

$$|F^m| \le C_1 + \frac{L}{2} \big(|V^m| + \big| V^{m-1} \big| \big) \le C_1 + L\Lambda q.$$

Using (2.22) we infer that

$$\begin{aligned} |W^m - \overline{u}| &\le |U^m - \overline{u}| + \left| \frac{(1-e)hV^{m-1} + h^2 F^m}{1+e} \right| \\ &\le \Lambda \Big\| \widehat{X}^m - \Phi(\overline{u}) \Big\| + \frac{(1-e)h}{1+e} \Lambda q + \frac{h^2}{1+e} (C_1 + L\Lambda q) \le \frac{R_1}{\Lambda}, \end{aligned}$$

thanks to (2.32) and (2.33). Moreover,

$$\left| \frac{U^{m+1} + eU^{m-1}}{1+e} - \overline{u} \right| \le |U^m - \overline{u}| + h \left| \frac{V^m - eV^{m-1}}{1+e} \right|$$

and (2.32) and (2.33) also imply that

$$\left| \frac{U^{m+1} + eU^{m-1}}{1+e} - \overline{u} \right| \le \Lambda \Big\| \widehat{X}^m - \Phi(\overline{u}) \Big\| + hq \le \frac{R_1}{\Lambda}.$$

Consequently assumption (2.16) is satisfied and we define I^m, $\widehat{I}^m$, $\widetilde{I}^m$, and J^m as in (2.18), (2.19), (2.20), and (2.23). We now write the numerical scheme (2.26)–(2.27) and define $(X^m)_{0\leq m\leq \lfloor \tau/h\rfloor}$ by

$$X^0=\widehat{X}^0,\ X^1=\widehat{X}^1,\ X^2=\widehat{X}^2,\quad \text{and,}\quad \text{for } 3\leq m\leq \lfloor \tau/h\rfloor,\quad X^m=\begin{pmatrix} s^m\\ y^m\end{pmatrix}.$$

It should be remarked that if the mapping $\widehat{X}\mapsto X$ possesses a fixed point X in B_h, then $U=\Psi(X)$ is precisely the numerical solution sought here. Using (2.30) we have the following estimate:

$$\|\zeta^m\|\leq C_5+C_6\left(\left\|\widehat{\xi}^m\right\|^2+\left\|\widehat{\xi}^{m-1}\right\|^2\right)\leq C_5+2C_6q^2.$$

Let us estimate now the discrete velocity $\|\xi^m\|$. By definition of σ^m we have

$$|\sigma^m|\leq \left|\sigma^{m-1}\right|+h\left\|\zeta^m\right\|.$$

For the normal component η^m, we apply Lemma 2.1 of the companion paper [9] and get

$$|\eta^m|\leq \max\left(\left|\eta^{m-1}\right|,e\left|\eta^{m-2}\right|\right)+h\left\|\zeta^m\right\|+h\left\|\zeta^{m-1}\right\|.$$

It follows that

$$\|\xi^m\|\leq \max\left(\left\|\xi^{m-1}\right\|,e\left\|\xi^{m-2}\right\|\right)+h\left\|\zeta^m\right\|+h\left\|\zeta^{m-1}\right\| \tag{2.34}$$

and thus

$$\|\xi^m\|\leq \max\left(\left\|\xi^{m-1}\right\|,\left\|\xi^{m-2}\right\|\right)+2h(C_5+2C_6q^2).$$

By discrete integration we obtain

$$\|\xi^m\|\leq \max\left(\left\|\xi^{1}\right\|,\left\|\xi^{0}\right\|\right)+2\tau(C_5+2C_6q^2),$$

and we have

$$\max\left(\left\|\xi^{1}\right\|,\left\|\xi^{0}\right\|\right)\leq \Lambda\max\left(\left|\frac{U^1-U^0}{h}\right|,\left|\frac{U^2-U^1}{h}\right|\right)\leq \Lambda C_2.$$

Hence (2.31) implies that X belongs to B_h. The mapping $\widehat{X}\mapsto X$ is clearly continuous and Brouwer's fixed point theorem implies the existence of a fixed point in B_h.

We prove uniqueness in both cases by local considerations. Given U^{m-1} and U^m, the discrete velocity V^m is a fixed point of the mapping

$$\begin{aligned} v\mapsto h^{-1}\Bigg[&-eU^{m-1}-U^m\\ &+(1+e)P_K\left(\frac{2U^m-(1-e)U^{m-1}+h^2F(t_m,U^m,U^{m-1},(v+V^{m-1})/2h,h)}{1+e}\right)\Bigg].\end{aligned}$$

The Lipschitz constant of this mapping is $h\gamma L/2$, and therefore if $h_1<2/(\gamma L)$, the uniqueness of V^m is guaranteed. ∎

The next lemma establishes the existence of U^2 under appropriate assumptions on U^0 and U^1.

LEMMA 2.2. *For all $(t_0, u_0, p_0) \in \mathbb{D}$, define $v_0 = M(u_0)^{-1}p_0$; then for all U^1 satisfying* (1.14) *and for all small enough h, there exists a unique solution U^2 of* (1.16) *for $m = 1$ satisfying*

$$\left|U^2 - U^1\right|_{u_0} \leq 3|v_0|_{u_0}h + h.$$

Proof. Let $r > 0$ be such that P_K is Lipschitz continuous on

$$B_{u_0}(u_0, r) = \left\{u \in \mathbb{R}^d : |u - u_0|_{u_0} \leq r\right\}.$$

Define $\widetilde{C}_1$ by

$$\widetilde{C}_1 = \max\Big\{ |F(t, u, u', 0, h)|_{u_0} : t \in [0, T],\ |u - u_0|_{u_0} \leq r,\\ |u' - u_0|_{u_0} \leq r,\ h \in [0, h^*]\Big\},$$

and let $\widetilde{L}$ be the Lipschitz constant defined by

$$\widetilde{L} = \sup\left\{ \frac{|F(t, u, u', v, h) - F(t, u, u', v', h)|_{u_0}}{|v - v'|_{u_0}} : |u' - u_0|_{u_0} \leq r,\ |u' - u_0|_{u_0} \leq r,\right.\\ \left. |v|_{u_0} \leq 2\,|v_0|_{u_0} + 1,\ |v'|_{u_0} \leq 2\,|v_0|_{u_0} + 1,\ v \neq v',\ h \in [0, h^*]\right\}.$$

Finally, let $\widetilde{\gamma}$ be the Lipschitz constant of P_K defined by

$$\widetilde{\gamma} = \sup\left\{ \frac{|P_K u - P_K u'|_{u_0}}{|u - u'|_{u_0}} : |u - u_0|_{u_0} \leq r,\ |u - u_0|_{u_0} \leq r,\ u \neq u' \right\}.$$

There exists a function $z(t)$ which is bounded in a neighborhood of 0 such that for small positive values of t

$$P_K(u_0 + tv_0) = u_0 + tv_0 + t^2 z(t); \tag{2.35}$$

indeed, if v_0 vanishes, or if u_0 belongs to $\mathrm{int}(K)$, or if u_0 belongs to ∂K and the scalar product $\langle v_0, N(u_0)\rangle_{u_0}$ is strictly positive, z vanishes; if u_0 belongs to ∂K and $\langle v_0, N(u_0)\rangle_{u_0}$ vanishes, while v_0 does not vanish, (2.35) is a consequence of the smoothness of $P_{\partial K}$ in a neighborhood of u_0. For the values of t for which $u_0 + tv_0$ belongs to K, z vanishes; for the values of t for which $u_0 + tv_0$ does not belong to K, a Taylor expansion shows that

$$P_{\partial K}(u_0 + tv_0) = u_0 + tv_0 + O(t^2)$$

and hence (2.35). With the change of variable $v = (U^2 - U^0)/h$, (1.16) is equivalent to $v = G(v)$, where the function G is defined by

$$G(v) = \frac{1 + e}{h}\left[P_K\left(U^0 + \frac{2h}{1 + e}V^0 + \frac{h^2}{1 + e}F\left(t_1, U^1, U^0, v, h\right)\right) - U^0\right].$$

Let us check that G is a strict contraction on $B_{u_0}\big(0, 2\,|v_0|_{u_0} + 1/2\big)$. If $|v|_{u_0} \leq 2\,|v_0|_{u_0} + 1/2$, for h small enough, we can use the definitions of $\widetilde{L}$ and $\widetilde{C}_1$:

$$\left|F\left(t_1, U^1, U^0, v, h\right)\right|_{u_0} \leq \widetilde{C}_1 + \widetilde{L}\big(2\,|v_0|_{u_0} + 1/2\big). \tag{2.36}$$

We estimate $G(v)$ as follows: by the triangle inequality and the Lipschitz condition on P_K,

$$|G(v)|_{u_0} \leq \frac{1+e}{h}\widetilde{\gamma}\left|\frac{2h\big(V^0-v_0\big)}{1+e}+\frac{h^2}{1+e}F^1\right|_{u_0}+\frac{1+e}{h}\left|P_K\left(u_0+\frac{2hv_0}{1+e}\right)-U^0\right|_{u_0}.$$

We apply (2.35), (1.14), and (2.36), and we find that

$$|G(v)|_{u_0} \leq \widetilde{\gamma}\Big[2\,|z(h)|_{u_0}+h\big(\widetilde{C}_1+2\widetilde{L}\,|v_0|_{u_0}+\widetilde{L}\big)\Big]+2\,|v_0|_{u_0}+\frac{4h}{1+e}\left|z\left(\frac{2hv_0}{1+e}\right)\right|_{u_0}.$$

Therefore, for h small enough, G maps $B_{u_0}\big(0,2\,|v_0|_{u_0}+1/2\big)$ to itself. Moreover, the Lipschitz constant of G on this ball is at most equal to $\widetilde{\gamma}\widetilde{L}h$. This proves that G has a fixed point in $B_{u_0}\big(0,2\,|v_0|_{u_0}+1/2\big)$ for small enough values of h. We set $U^2=U^0+hv$, where v is the fixed point of G in $B_{u_0}\big(0,2\,|v_0|_{u_0}+1/2\big)$. Then it is clear that, for h small enough,

$$\big|(U^2-U^1)/h\big|_{u_0} \leq 2|v_0|_{u_0}+1/2+\big|(U^1-U^0)/h\big|_{u_0},$$

and the lemma is proved. □

If we put together Theorem 2.1 and Lemma 2.2, we obtain a local existence result for the scheme.

THEOREM 2.3. *For all $(t_0,u_0,M(u_0)v_0)\in\mathbb{D}$, for all U^1 satisfying* (1.14)*, there exist $\tau>0$, $C_3<\infty$, and $h_1>0$ such that for all $h\in(0,h_1]$, there exists a unique solution of* (1.16) *and* (1.17) *for all $m\leq\lfloor\tau/h\rfloor-1$, which satisfies the estimate*

$$\forall l\leq\lfloor\tau/h\rfloor,\quad \big|V^l\big|\leq C_3. \tag{2.37}$$

Proof. Let us check that U^0 and U^1 satisfy condition $E(u_0,r_1,C_2,h)$. Lemma 2.2 and assumption (1.14) on U^1 imply that

$$\big|U^1-u_0\big|_{u_0}\leq h\big(|z(h)|_{u_0}+|v_0|_{u_0}\big)$$

and

$$\big|U^2-u_0\big|_{u_0}\leq h\big(2\,|v_0|_{u_0}+1/2\big).$$

Choose $C_2\geq\big(3\,|v_0|_{u_0}+2\big)\,\big\|M(u_0)^{-1/2}\big\|$; U^0 and U^1 satisfy condition $E(u_0,r_1,C_2,h)$ for small enough values of h. Then it is clear that Theorem 2.1 applies. □

It is convenient to give a uniformized version of Theorem 2.1.

THEOREM 2.4. *For all compact subsets $\mathcal{C}$ of K, for all $C_2>0$, there exist positive numbers r_1, $r_2>r_1$, τ, C_3, and h_1 such that for all $t_0\in[0,T)$, for all $\overline{u}\in\mathcal{C}$, for all $h\leq h_1$, and for all U^0 and U^1 satisfying condition $E(\overline{u},r_1,C_2,h)$, relations* (1.16) *and* (1.17) *define uniquely in the ball $B(\overline{u},r_2)$ under condition* (2.37) *the vectors U^m for $2\leq m\leq\lfloor\min(\tau,T-t_0)/h\rfloor$.*

Proof. Let C_2 be a strictly positive number. Any element u of $\mathcal{C}$ is included in an open ball $\mathrm{int}(B(u,r_1(u)))$ such that Theorem 2.1 holds. We cover $\mathcal{C}$ by a finite number of balls $\mathrm{int}(B(u_j,r_1(u_j)/2))$ with associated numbers $r_2(u_j)$, $\tau(u_j)$, $h_1(u_j)$, and $C_3(u_j)$. If we let

$$r_1=\frac{1}{2}\min\{r_1(u_j):1\leq j\leq J\},$$

then any $\overline{u} \in \mathcal{C}$ belongs to a ball $B(u_j, r_1(u_j)/2)$, and in particular $B(\overline{u}, r_1)$ is included in $B(u_j, r_1(u_j))$. If we take

$$\tau = \min_j \tau(u_j),\ r_2 = \max_j r_2(u_j), \quad h_1 = \min_j h_1(u_j),\ C_3 = \max_j C_3(u_j),$$

it is immediate that the theorem holds, thanks to Theorem 2.1. □

3. Estimates on the acceleration. In this section and the three following ones, we assume that there exist strictly positive numbers τ, C_3, and h_1, and a subsequence of time steps to which correspond solutions of the numerical scheme defined by (1.13), (1.14), (1.16), and (1.17), which satisfy the estimate, for all $h \le h_1$,

$$\forall l \in \{0, \dots, P-1\}, \quad \left|U^{l+1} - U^l\right| \le C_3 h, \tag{3.1}$$

where $P = \lfloor \min(\tau, T - t_0)/h \rfloor$. Here we estimate the discrete total variation of the sequence $\left(V^m\right)_m$. It is also convenient to define the function $w_h(t)$ on $[t_0, t_0 + \tau]$ by

$$\begin{aligned} &w_h(t_m) = W^m, \quad w_h \text{ is continuous and it is affine} \\ &\text{on each interval } [t_m, t_{m+1}) \text{ and constant on } [t_P, t_0 + \tau]. \end{aligned}$$

THEOREM 3.1. *Under assumption* (3.1)*, there exists a constant* C_7 *such that for all* $h \le h_1$

$$\sum_{m=1}^{P-1} \left|V^m - V^{m-1}\right| \le C_7. \tag{3.2}$$

Proof. The idea of the proof is exactly the same as in section 4 of [9], up to geometric complications.

Let $\mathcal{C}$ be the compact set $K \cap B(u_0, C_3\tau)$ and let r_1 be as in Theorem 2.4; cover $\mathcal{C}$ with a finite number of balls $B(u_j, r_1/4)$; observe that, thanks to the Ascoli–Arzelá theorem, the set $\mathcal{W}$ of functions $(w_h)_{0<h_1\le h}$ is relatively compact in $C^0([t_0, t_0 + \tau])$. The set of limit points of $(w_h)_{0<h\le h_1}$ as h tends to 0 is also a compact set, which we shall denote by $\mathcal{W}_\infty$. There exists a finite subset $w^1, \dots, w^I$ of $\mathcal{W}_\infty$ such that

$$\forall w \in \mathcal{W}_\infty \quad \inf\{\|w - w^i\|_{C^0[t_0, t_0+\tau]} : 1 \le i \le I\} \le r_1/4.$$

For each $i \in \{1, \dots, I\}$, it is possible to find a finite increasing sequence of times

$$0 = \tau(i,0) < \cdots < \tau(i,k) < \cdots < \tau(i, \kappa(i)) = \tau$$

such that

$$w^i([t_0 + \tau(i,k), t_0 + \tau(i,k+1)]) \subset B\big(u_{j(i,k)}, r_1/4\big).$$

Thus, for all $w \in \mathcal{W}_\infty$, there exists $i \in \{1, \dots, I\}$ such that for all $k \in \{0, \dots, \kappa(i) - 1\}$,

$$w([t_0 + \tau(i,k), t_0 + \tau(i,k+1)]) \subset B\big(u_{j(i,k)}, r_1/2\big).$$

Therefore, we can decrease h_1 so that

$$\begin{aligned} &\forall h \in (0, h_1], \quad \exists i \in \{1, \dots, I\}, \quad \forall k \in \{1, \dots, \kappa(i) - 1\}, \\ &\forall t \in [t_0 + \tau(i,k), t_0 + \tau(i,k+1)] \quad w_h(t) \in B\big(u_{j(i,k)}, 3r_1/4\big), \end{aligned}$$

and thanks to (2.22) and to (3.1), we can decrease h_1 such that

$$\forall h \in (0, h_1], \quad \exists i \in \{1, \dots, I\}, \quad \forall k \in \{1, \dots, \kappa(i) - 1\},$$
$$\forall l \in \{\lfloor \tau(i,k)/h \rfloor, \dots, \lfloor \tau(i,k+1)/h \rfloor\}, \quad U^l \in B\big(u_{j(i,k)}, r_1\big).$$

We simplify the notations by letting

$$P = \lfloor \tau(i,k)/h \rfloor, \quad Q = \lfloor \tau(i,k+1)/h \rfloor,$$

and we take C_1 as in (2.9), where $\overline{u}$ is set equal to $u_{j(i,k)}$, r_0 is set equal to r_1, and C_3 is set equal to C_3.

Now, we have to consider two cases.

First case: $B(\overline{u}, r_1) \cap \partial K = \emptyset$. We have the inequality

$$|F^m| \le C_1 + LC_3; \tag{3.3}$$

hence, thanks to (1.16), we have the inequality

$$\big|V^m - V^{m-1}\big| \le h\big(C_1 + LC_3\big),$$

and therefore

$$\sum_{m=P+1}^{Q} \big|V^m - V^{m-1}\big| \le \big(\tau(i,k+1) + 2h - \tau(i,k)\big)\big(C_1 + LC_3\big). \tag{3.4}$$

Second case: $B(\overline{u}, r_1) \cap \partial K \neq \emptyset$. We observe that, thanks to (2.30), we have the estimate

$$\forall m \in \{P+1, \dots, Q-1\}, \quad \max\big(|\kappa^m|, |\lambda^m|\big) \le C_9, \tag{3.5}$$

where

$$C_9 = C_5 + 2C_6\Lambda^2 C_3^2.$$

The estimates on the first $d-1$ components of the velocity in the straightened coordinates are immediate:

$$\sum_{m=P+1}^{Q-1} \left| \frac{s^{m+1} - s^m}{h} - \frac{s^m - s^{m-1}}{h} \right| \le \big(\tau(i,k+1) + 2h - \tau(i,k)\big)C_9. \tag{3.6}$$

In order to estimate the last coordinate we use the same idea as in the one dimensional case (see Theorem 3.1 [9]). We observe that

$$\eta^m - \eta^{m-1} = h\lambda^m + \big(2y^m - (1-e)y^{m-1}\big)^-/h \tag{3.7}$$

and therefore, by the triangle inequality,

$$\big|\eta^m - \eta^{m-1}\big| \le hC_9 + \big(2y^m - (1-e)y^{m-1}\big)^-/h,$$

and using (3.7) again,

$$\big|\eta^m - \eta^{m-1}\big| \le 2hC_9 + \eta^m - \eta^{m-1}.$$

We have the elements of a telescoping sum and we obtain

$$\sum_{m=P+1}^{Q-1} \left|\eta^m - \eta^{m-1}\right| \leq 2C_9 h(Q-P) + 2C_3\Lambda.$$

Summarizing this relation with (3.6), we can see that

$$\sum_{m=P+1}^{Q} \left|V^m - V^{m-1}\right| \leq \Lambda C_9\big(3\tau(i,k+1) - 3\tau(i,k) + 6h\big) + 2C_3\Lambda^2. \tag{3.8}$$

Relations (3.4) and (3.8) do not depend on $h \leq h_1$; since we have only a finite number of these estimates, the theorem is proved. □

4. Variational properties of the limit of the numerical scheme. In this section, we work under the assumption (3.1). Recall that $P = \lfloor \tau/h \rfloor$. We define a function u_h by affine interpolation as follows:

$$\begin{cases} u_h(t) = U^m + (t - t_0 - mh)\dfrac{U^{m+1} - U^m}{h} \\ \qquad\qquad \text{for } t - t_0 \in [mh, (m+1)h\big),\ 0 \leq m \leq P-1, \\ u_h(t) = U^P \quad \text{for } t - t_0 \in [Ph, \tau]. \end{cases} \tag{4.1}$$

We also define a measure F_h as the following sum of Dirac masses:

$$F_h(t) = \sum_{m=1}^{P-1} hF^m \delta(t - t_0 - mh). \tag{4.2}$$

In this section we prove that the sequence $(u_h)_h$ converges in an appropriate sense to a function u which satisfies (1.3) to (1.5b) with τ instead of $\overline{\tau}$. We delay the proof of (1.6), the transmission condition at impacts, to a later section.

As in the one dimensional case there are three steps in the convergence proof: the first is to prove that the limit u exists in an appropriate sense and takes its values in K; in the second step, we show that $\dot{u}_h$ is of bounded variation uniformly in h and that F_h converges to $M(u)^{-1} f(\cdot, u, M(u)\dot{u})$ weakly in the space of $\mathbb{R}^d$-valued measures. The last step is the characterization of the measure $\mu = M(u)\ddot{u} - f(\cdot, u, M(u)\dot{u})$: there we show that μ satisfies conditions (1.4a), (1.4b), and (1.4c). The main ideas of the proofs are the same as in the one dimensional case; we will essentially give the differences in the proofs.

LEMMA 4.1. *From all sequence of functions $(u_h)_h$ indexed by a sequence h tending to 0, it is possible to extract a subsequence, still denoted by $(u_h)_h$ such that*

$$u_h \to u \quad \textit{in } C^0([t_0, t_0+\tau]) \textit{ strong}, \tag{4.3}$$

$$\dot{u}_h \to \dot{u} \quad \textit{in } L^\infty([t_0, t_0+\tau]) \textit{ weak } *. \tag{4.4}$$

The function u takes its values in K.

Proof. We generalize the proof of Lemma 5.1 of [9], as follows: instead of (5.5), we have for all m belonging to $\{1, \ldots, P-1\}$

$$Z^m = \frac{U^{m+1} + eU^{m-1}}{1+e} = U^m + h\frac{V^m - eV^{m-1}}{1+e}; \tag{4.5}$$

hence $U^m = Z^m - h(V^m - eV^{m-1})/(1+e)$. By definition of the scheme, we have $Z^m = P_K(W^m)$ (see (1.20)), and thus Z^m belongs to K. It follows that, for all $m \in \{1, \dots, P-1\}$, the Euclidean distance between U^m and K can be estimated as follows:

$$\min\big\{|U^m - u| : u \in K\big\} \le h\left|V^m - eV^{m-1}\right|/(1+e) \le hC_3. \qquad \square \tag{4.6}$$

The next lemma describes the convergence of the measures involved in our problem; we denote by $M^1\big((t_0, t_0+\tau)\big)$ the space of bounded measures over $(t_0, t_0+\tau)$ with values in $\mathbb{R}^d$.

LEMMA 4.2. *The measures $\ddot{u}_h$ and F_h converge weakly $*$ in $M^1\big((t_0, t_0+\tau)\big)$, respectively, to $\ddot{u}$ and $M(u)^{-1}f(\cdot, u, M(u)\dot{u})$.*

Proof. The proof is a direct generalization of the proof of Lemma 5.2 of [9]; the only modifications are the following: η^m is replaced by V^m and $f(t, u, \dot{u})$ is replaced by $M(u)^{-1}f(t, u, M(u)\dot{u})$. $\square$

Let us prove now that the measure μ has the required variational properties.

LEMMA 4.3. *The measure μ satisfies properties* (1.4a), (1.4b), *and* (1.4c).

Proof. The measure

$$\mu_h = M(u_h)\big(\ddot{u}_h - F_h\big)$$

is a sum of Dirac measures on $(t_0, t_0+\tau)$ given by

$$\mu_h = \sum_{m=1}^{P-1} M(U^m)\big(V^m - V^{m-1} - hF^m\big)\delta(t - t_0 - mh) - M\big(U^P\big)V^{P-1}\delta(t - t_0 - Ph)$$

and it converges to $\mu = M(u)\ddot{u} - f(\cdot, u, p)$ weakly $*$ in $M^1\big((t_0, t_0+\tau)\big)$. The proof of property (1.4a) is analogous to the proof of property (1.3a) of the companion paper (Lemma 5.3. of [9]).

The foregoing proof is precisely the one where one would be tempted to use straightened coordinates; however, this is an inefficient choice, because of the simplicity of (1.22) in the original coordinates.

Assume now that $u_1 = u(t_1)$ belongs to ∂K, and let $B(u_1, r_1)$ be a ball having the properties of Theorem 2.1; assume that the image of (τ_1, τ_2) by u_h and w_h is included in this ball for all small enough h. We rewrite conditions (1.4b) and (1.4c) as follows: for all continuous function ψ with compact support included in $(t_0, t_0+\tau)$ and taking its values in $\mathbb{R}^d$ the following implication holds:

$$\forall t \in (t_0, t_0+\tau), \quad d\phi(u(t))\psi(t) \ge 0 \Longrightarrow \langle \mu, \psi \rangle \ge 0. \tag{4.7}$$

The reader will check the equivalence of (1.4b) and (1.4c) with (4.7). We infer from relation (4.6) that

$$|Y(U^m)| \le \Lambda h C_3;$$

therefore, there exists a constant C_{10} such that

$$|Y(W^m)| \le hC_{10}.$$

Since (4.7) is local, it is enough to check it in the neighborhood of any $t_1 \in (t_0, t_0+\tau)$. Let

$$P = \lceil \tau_1/h \rceil, \quad Q = \lfloor \tau_2/h \rfloor,$$

and

$$\mathcal{P} = \big\{ m \in \{P, \dots, Q\} : W^m \notin K \big\}, \quad \mathcal{P}' = \{P, \dots, Q\} \setminus \mathcal{P}.$$

We observe that if m belongs to $\mathcal{P}'$, then

$$V^m - V^{m-1} - hF^m = 0.$$

Therefore, we have the identity

$$\begin{aligned} &\sum_{m=P}^{Q} \langle V^m - V^{m-1} - hF^m, \psi(t_m) \rangle_{U^m} \\ &\quad = \sum_{m \in \mathcal{P}} \langle V^m - V^{m-1} - hF^m, \psi(t_m) \rangle_{U^m}. \end{aligned}$$

We recall relation (1.22). Relation (2.25) implies that

$$\Phi(Z^m) - \Phi(W^m) = \begin{pmatrix} 0 \\ Y(W^m)^- \end{pmatrix},$$

and therefore

$$\left| Z^m - W^m - D\Phi(W^m) \begin{pmatrix} 0 \\ Y(W^m)^- \end{pmatrix} \right| \le C_4 \, |Z^m - W^m|^2 \, \Lambda^2. \tag{4.8}$$

On the other hand, the definition of Ψ is such that the dth column of $D\Psi(Z^m)$ is equal to $N(Z^m)$; therefore

$$\begin{aligned} &\left| D\Psi(W^m) \begin{pmatrix} 0 \\ Y(W^m)^- \end{pmatrix} - N(Z^m) Y(W^m)^- \right| \\ &\quad \le 2C_4 \, |Z^m - W^m| \, Y(W^m)^-. \end{aligned} \tag{4.9}$$

We combine (4.8) and (4.9) to get

$$\begin{aligned} &\big| Z^m - W^m - Y(W^m)^- N(Z^m) \big| \\ &\quad \le C_4 \big(2Y(W^m)^- + \Lambda^2 \, |Z^m - W^m| \big) \, |Z^m - W^m| \\ &\quad \le \frac{C_4 C_{10} \big(2 + \Lambda^3\big) h^2}{1+e} \big| V^m - V^{m-1} - hF^m \big|, \end{aligned}$$

and thus there exists C_{11} such that for all $m \in \mathcal{P}$

$$\big| Z^m - W^m - Y(W^m)^- N(Z^m) \big| \le h^2 C_{11} \big| V^m - V^{m-1} - hF^m \big|.$$

We now can see that

$$\begin{aligned}
&\sum_{m\in\mathcal{P}} \langle V^m - V^{m-1} - hF^m, \psi(t_m)\rangle_{U^m} \\
&= \frac{1+e}{h} \sum_{m\in\mathcal{P}} \langle Z^m - W^m, \psi(t_m)\rangle_{U^m} \\
&\geq \frac{1+e}{h} \sum_{m\in\mathcal{P}} Y(W^m)^- \langle N(Z^m), \psi(t_m)\rangle_{U^m} \\
&\quad - C_{11}h(1+e) \max_{P\leq m\leq Q} \big(\|M(U^m)\|\,|\psi(t_m)|\big) \sum_{m\in\mathcal{P}} \left|V^m - V^{m-1} - hF^m\right|,
\end{aligned}$$

which implies by a straightforward passage to the limit that $\langle \mu, \psi\rangle$ is nonnegative. This concludes the proof of the lemma. ∎

5. Transmission of energy during impact. The basic assumption is still the one made at the beginning of section 3.

Let $\overline{\tau} \in (0,\tau)$ be such that $u(t_0+\overline{\tau})$ belongs to ∂K. Write $\overline{t} = t_0 + \overline{\tau}$. We decompose $p(\overline{t}\pm 0)$ into a normal component $p_N(\overline{t}\pm 0)$ belonging to $\mathbb{R}d\phi(u(\overline{t}))$ and a tangential part $p_T(\overline{t}\pm 0)$ belonging to the orthogonal of $d\phi(u(\overline{t}))$ in the cotangent metric at $u(\overline{t})$.

In this section, we shall prove that

$$p_T(\overline{t}+0) = p_T(\overline{t}-0) \text{ and } p_N(\overline{t}+0) = -ep_N(\overline{t}-0), \tag{5.1}$$

where e is the restitution coefficient of the problem.

The conservation of the tangential component of the impulsion is proved in next lemma.

LEMMA 5.1. *Assume that $\overline{\tau} \in (0,\tau)$ is such that $u(\overline{\tau})$ belongs to ∂K. Then*

$$p_T(\overline{t}+0) = p_T(\overline{t}-0).$$

Proof. Thanks to Lemma 4.3, we know that

$$M(u)\ddot{u} = \mu + f(\cdot, u, p) \tag{5.2}$$

and that there exists a nonnegative measure λ such that

$$\mu = \lambda d\phi(u). \tag{5.3}$$

We take the measure of the set $\{\overline{t}\}$ by the two sides of (5.2), and we find that

$$M(u(\overline{t}))\big(\dot{u}(\overline{t}+0) - \dot{u}(\overline{t}-0)\big) = \mu(\{\overline{t}\}),$$

which immediately implies that $p(\overline{t}+0) - p(\overline{t}-0)$ is parallel to $d\phi(u(\overline{t}))$ and proves the lemma. ∎

Let $\overline{u} = u(\overline{t})$ and let $B(\overline{u}, r_1)$ and $B(\overline{u}, r_2)$ have the properties of Theorem 2.1. There exists an interval $[\tau_-, \tau_+]$ containing $\overline{\tau}$ in its interior such that for all small enough h, $u_h\big([t_0+\tau_-, t_0+\tau_+]\big)$ is included in $B(\overline{u}, r_1)$.

Define

$$P = \lceil \tau_-/h \rceil + 1, \quad Q = \lfloor \tau_+/h \rfloor - 1,$$

and let x_h be obtained from the X^m by affine interpolation, for $P \le m \le Q$. We infer from estimates (3.1) and (3.2) the estimates

$$\max_{P\le m\le Q}\left|\frac{X^{m+1}-X^m}{h}\right| \le \Lambda C_3,$$
$$\sum_{m=P}^{Q}\left|\frac{X^{m+1}-X^m}{h}-\frac{X^m-X^{m-1}}{h}\right| \le \Lambda C_7.$$

Therefore, we have the following convergences:

$$x_h \to x \text{ strongly in } C^0\big([t_0+\tau_-,t_0+\tau_+]\big);$$
$$\dot{x}_h \to \dot{x} \text{ except on a countable set and weakly } * \text{ in } L^\infty\big([t_0+\tau_-,t_0+\tau_+]\big);$$
$$\ddot{x}_h \to \ddot{x} \text{ weakly in } M^1\big([t_0+\tau_-,t_0+\tau_+]\big).$$

Write for all $h \le h_1$

$$x_h = \begin{pmatrix} s_h \\ y_h \end{pmatrix}, \quad x = \begin{pmatrix} s \\ y \end{pmatrix},$$

where the s_h's and s take their values in $\mathbb{R}^{d-1}$ and the y_h's and y are real valued functions. We do not have $x_h = \Phi(u_h)$, because x_h is a linear interpolation of the sequence $X^m = \Phi(U^m)$, and $\Phi(u_h)$ is the image of the linear interpolation of the sequence U^m. However, we can estimate the difference $x_h - \Phi(u_h)$.

LEMMA 5.2. *For all $t \in [t_0+\tau_-, t_0+\tau_+]$ belonging to $[t_m, t_{m+1}]$, we have*

$$x_h(t) - \Phi\big(u_h(t)\big) \le 2C_4C_3^2 h \min(t-t_m, t_{m+1}-t).$$

Proof. We observe that $x_h(t_m) = X^m$ and that

$$\left|\frac{d}{dt}\big[x_h(t)-\Phi\big(u_h(t)\big)\big]\ \big|_{t=t_m+0}\right| = \left|\frac{\Phi(U^{m+1})-\Phi(U^m)-hD\Phi(U^m)V^m}{h}\right| \le hC_3^2C_4.$$

Moreover, for all $t \in [t_m, t_{m+1})$

$$\left|\frac{d^2}{dt^2}\Big[x_h(t)-\Phi\big(u_h(t)\big)\Big]\right| = \big|D^2\Phi\big(U^m+(t-t_m)V^m\big)V^m\otimes V^m\big| \le 2C_3^2C_4.$$

Therefore, a straightforward integration yields

$$\big|x_h(t)-\Phi\big(u_h(t)\big)\big| \le C_3^2C_4\big(h(t-t_m)+(t-t_m)^2\big),$$

which implies

$$\big|x_h(t)-\Phi\big(u_h(t)\big)\big| \le 2C_3^2C_4h(t-t_m).$$

We can write the analogous estimate on the interval $[t, t_{m+1})$, which concludes the proof. ∎

As a consequence of Lemma 5.2 we obtain

$$\forall t \in [t_0+\tau_-, t_0+\tau_+], \quad x(t) = \Phi\big(u(t)\big)$$

and

$$\forall t \in (t_0+\tau_-, t_0+\tau_+), \quad \dot{x}(t\pm 0) = D\Phi\big(u(t)\big)\dot{u}(t\pm 0).$$

By virtue of relation (2.6),

$$\dot{u}(\bar{t}\pm 0) = \begin{pmatrix} \dot{s}(\bar{t}\pm 0) \\ 0 \end{pmatrix} + \dot{y}(\bar{t}\pm 0)N(\overline{u}).$$

We can rewrite this relation in terms of p_N and p_T:

$$p_T(\bar{t}\pm 0) = M(\overline{u})\begin{pmatrix} \dot{s}(\bar{t}\pm 0) \\ 0 \end{pmatrix}, \quad p_N(\bar{t}\pm 0) = \dot{y}(\bar{t}\pm 0)M(\overline{u})N(\overline{u}).$$

Lemma 5.1 implies $\dot{s}(\bar{t}+0) = \dot{s}(\bar{t}-0)$. In order to achieve the proof of relation (5.1), it remains to prove the scalar relation

$$\dot{y}(\bar{t}+0) = -e\dot{y}(\bar{t}-0). \tag{5.4}$$

But this relation is an immediate consequence of the precise analysis of the transmission of energy performed in the one dimensional case in [9] (see section 6).

6. Initial conditions. In this section we prove that the solution that we have constructed satisfies the initial conditions; we work under the hypotheses stated at the beginning of section 3.

LEMMA 6.1. *The function u satisfies the initial conditions*

$$u(t_0) = u_0, \quad p(t_0+0) = p_0.$$

Proof. By uniform convergence of u_h to u, it is clear that $u(t_0)$ is equal to u_0. There remains to show that the initial condition on the impulsion is satisfied.

First assume that u_0 belongs to the interior of K; then there exist $h_1 > 0$ and $\tau_1 > 0$ such that for all $h \in (0, h_1]$ and for all $t - t_0 \in [0, \tau_1]$

$$|u_h(t) - u_0| \le \frac{1}{2}\inf\{|u_0 - y| : y \notin K\}.$$

Then for all $t_m - t_0$ belonging to $(0, \tau_1]$, $\big(2U^m - (1-e)U^{m-1} + h^2F^m\big)/(1+e)$ belongs to K for h small enough; we have indeed

$$\begin{aligned}\left|\frac{2U^m - (1-e)U^{m-1} + h^2F^m}{1+e} - u_0\right| \\ \le \frac{1-e}{1+e}hC_3 + \frac{1}{2}\inf\{|u_0 - y| : y \notin K\} + \frac{h^2}{1+e}C_8,\end{aligned}$$

which is strictly inferior to $\inf\{|u_0 - y| : y \notin K\}$ for h small enough. Thus the constraints are not active for $0 \le t_m \le \tau_1$ and the convergence is a classical result.

In the second case, assume that u_0 belongs to ∂K; we have taken admissible initial conditions so that

$$\langle p_0, d\phi(u_0)\rangle^*_{u_0} \ge 0.$$

We use the construction and notations of section 2: Φ, Ψ, X^m, s^m, y^m, and ζ^m have the same signification as there.

Taylor's formula yields

$$\xi^0 = \frac{X^1 - X^0}{h} = D\Phi(u_0)\frac{U^1 - u_0}{h} + O(h),$$

and the definition (1.14) of U^1 gives

$$\xi^0 = D\Phi(u_0)M(u_0)^{-1}p_0 + O(h). \tag{6.1}$$

Write

$$\begin{pmatrix}\sigma_0\\ \eta_0\end{pmatrix} = D\Phi(u_0)M(u_0)^{-1}p_0.$$

Then the normal and tangential components of the impulsion are given by

$$p_{0T} = M(u_0)\begin{pmatrix}\sigma_0\\ 0\end{pmatrix} \text{ and } p_{0N} = \eta_0 M(u_0)N(u_0).$$

We wish to prove $p(t_0+0) = p_0$, which is equivalent to

$$\dot{x}(t_0+0) = \begin{pmatrix}\dot{s}(t_0+0)\\ \dot{y}(t_0+0)\end{pmatrix} = \begin{pmatrix}\sigma_0\\ \eta_0\end{pmatrix}.$$

We recall relation (2.26). Relation (6.1) implies that

$$\sigma^0 = \big(D\Phi(u_0)M(u_0)^{-1}p_0\big)' + O(h),$$

and together with (2.26), we obtain in the limit

$$\dot{s}(t) = \big(D\Phi(u_0)M(u_0)^{-1}p_0\big)' + O(t-t_0),$$

i.e., $\dot{s}(t_0+0) = \sigma_0$. Finally, using the same arguments as in the one dimensional case (see Lemma 7.1 in [9]), we obtain

$$\dot{y}(t_0+0) = \eta_0,$$

which concludes the proof. □

7. A priori estimates. In this section we prove that solutions of the problem (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), and (1.9) satisfy an a priori estimate on an interval with a nonempty interior.

LEMMA 7.1. *Let R be strictly larger than $|p_0|^*_{u_0}$. Then there exists $\tau(R) > 0$ such that for all solution u of* (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), *and* (1.9) *defined on $[t_0, t_0+\tau]$, the following estimates hold:*

$$\forall t \in [t_0, t_0 + \min(\tau, \tau(R))], \quad |u(t) - u_0| \le R, \quad |p(t)|^*_{u(t)} \le R. \tag{7.1}$$

Proof. Once again, the main ideas of the proof are the same as in the one dimensional case, but the estimates are more complex. The measure λ appearing in (1.4b) can be decomposed in the sum of an atomic part λ_a and a diffuse part λ_d. At each point of the support of λ_a we have

$$|p(t+0)|^*_{u(t)} \le |p(t-0)|^*_{u(t)} \tag{7.2}$$

thanks to relation (1.6). On any interval (t_1, t_2) which does not intersect the support of λ_a, we multiply relation (1.3) by $\dot{u}^T$ on the left, and we find that

$$\frac{d}{dt}\frac{1}{2}\dot{u}^T M(u)\dot{u} = \dot{u}^T f(\cdot, u, p) + \frac{1}{2}\dot{u}^T (DM(u)\dot{u})\dot{u}. \tag{7.3}$$

Define

$$E(u,p) = \frac{1}{2}\langle p, p\rangle_u^*, \quad z = |p|_u^*.$$

It is convenient to recall that

$$|p|_u^* = |M(u)^{-1/2} p| = |M(u)^{1/2}\dot{u}|.$$

Relations (7.2) and (7.3) imply that in the sense of measures

$$z\dot{z} = \dot{E} \le \dot{u}^T f(\cdot, u, p) + \frac{1}{2}\dot{u}^T \big(DM(u)\dot{u}\big)\dot{u}. \tag{7.4}$$

Our purpose now is to transform (7.4) into a differential inequality. Let $\chi(u)$ be the norm of the bilinear mapping

$$(v_1, v_2) \mapsto M(u)^{-1/2}(DM(u)M(u)^{-1/2}v_1)M(u)^{-1/2}v_2.$$

With this definition,

$$\big|\dot{u}^T \big(DM(u)\dot{u}\big)\dot{u}\big| \le \chi(u) z^3.$$

We write now

$$\begin{aligned}\dot{u}^T f(t,u,p) &= \dot{u}^T M(u)^{1/2} M(u)^{-1/2} f(t,u,p)\\ &= \dot{u}^T M(u)^{1/2}\big[M(u)^{-1/2} f(t,u,p) - M(u_0)^{-1/2} f(t,u_0,0) + M(u_0)^{-1/2} f(t,u_0,0)\big].\end{aligned}$$

Define

$$g(t) = |M(u_0)^{-1/2} f(t, u_0, 0)|.$$

Fix $R > |p_0|_{u_0}^*$ and let $\omega(R)$ be the Lipschitz constant of $(u,p) \mapsto M(u)^{-1/2} f(t,u,p)$ for $t \in [0,T]$ and $\max(|u-u_0|, |p|_u^*) \le R$; more precisely,

$$\begin{aligned}\omega(R) = \sup\Bigg\{ &\frac{|M(u_1)^{-1/2} f(t,u_1,p_1) - M(u_2)^{-1/2} f(t,u_2,p_2)|}{|u_1-u_2| + |p_1-p_2|} : 0 \le t \le T,\\ &\max(|u_1-u_0|, |u_2-u_0|, |p_1|_{u_1}^*, |p_2|_{u_2}^*) \le R, u_1 \ne u_2 \text{ or } p_1 \ne p_2\Bigg\}.\end{aligned}$$

By construction, ω is continuous and is an increasing function of τ and R.

If $t_0 \le t \le t_0 + \tau$ and if $\max(|u(t)-u_0|, |p(t)|_{u(t)}^*) \le R$ on $[t_0, t_0+\tau]$, we have the inequality

$$\big|\dot{u}^T f(\cdot, u, p)\big| \le z\big(g + \omega(R)(|u-u_0| + |p|)\big).$$

But we can estimate $u(t) - u_0$:

$$|u(t) - u_0| \le \int_{t_0}^{t} |\dot{u}(s)|\, ds \le \int_{t_0}^{t} \|M(u)^{-1/2}\| z\, ds.$$

Therefore we have the estimate

$$|\dot{u}^T f(\cdot, u, p)| \le zg + z\omega(R)\left(\int_{t_0}^{t} \|M(u)^{-1/2}\| z\, ds + \|M(u)^{1/2}\| z\right),$$

and we conclude that z satisfies the differential inequality

$$\dot{z} \le g + \omega(R)\left[\int_{t_0}^{t} \|M(u)^{-1/2}\| z\, ds + \|M(u)^{1/2}\| z\right] + \frac{1}{2}\chi(u) z^2.$$

Set

$$\alpha = \sup\big\{\|M(u)^{1/2}\| : |u - u_0| \le R\big\}, \tag{7.5}$$

$$\beta = \sup\big\{\|M(u)^{-1/2}\| : |u - u_0| \le R\big\}, \tag{7.6}$$

$$\gamma = 2\sup\big\{\chi(u) : |u - u_0| \le R\big\}.$$

While $t \le t_0 + \tau$ and $\max(|u(t) - u_0|, |p(t)|^*_{u(t)}) \le R$, z satisfies the following differential inequality:

$$\dot{z} \le g + \omega(R)\left[\beta \int_{t_0}^{t} z\, ds + \alpha z\right] + \gamma z^2. \tag{7.7}$$

Consider the integrodifferential equation

$$\dot{y} = g + \omega(R)\left(\beta \int_{t_0}^{t} y\, ds + \alpha y\right) + \gamma |y|^2, \tag{7.8}$$

with the initial condition $y(t_0) = z(t_0)$. It has a unique maximal solution which blows up in finite time, as soon as γ is strictly positive and $\sup|g|$ is strictly positive. Let $\tau(R) \in (0, T - t_0]$ be the largest time for which

$$\forall t \in [t_0, t_0 + \tau(R)], \quad y(t) \le R, \quad \beta \int_{t_0}^{t} y\, ds \le R.$$

Such a number exists since $y(t_0) < R$. Then we can compare the solution z of (7.7) and the solution y of (7.8), and we find immediately that

$$\forall t \in \big[t_0, t_0 + \min\big(\tau, \tau(R)\big)\big], \quad z(t) \le y(t). \tag{7.9}$$

This concludes the proof of the lemma. □

8. Global results. We summarize the results obtained so far in the following proposition.

PROPOSITION 8.1. *Assume that there exist strictly positive numbers τ, C_3, and $h_1 > 0$, and a sequence of solutions of the numerical scheme defined by* (1.13), (1.14), (1.16), *and* (1.17), *which satisfies the estimate* (3.1). *Then it is possible to extract from the sequence u_h defined by* (4.1) *a subsequence which converges to a solution of* (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), *and* (1.9). *The convergence*

holds in the following sense: u_h converges uniformly to u on $[t_0, t_0+\tau]$; $\dot{u}_h$ converges to $\dot{u}$ in $L^\infty(t_0, t_0+\tau)$ weakly $*$ and almost everywhere on $[t_0, t_0+\tau]$; and $\ddot{u}_h$ converges to $\ddot{u}$ in the weak $*$ topology of measures. Moreover, for all $\bar{\tau} \in (0, \tau]$, we have the following convergence:

$$\begin{aligned} &\limsup_{h\downarrow 0} \sup\big\{|V^m|_{U^m} : t_0 \le t_m \le t_0 + \bar{\tau}\big\} \\ &\quad \le \operatorname{ess\,sup}\big\{|\dot{u}(t)|_{u(t)} : t_0 \le t \le t_0 + \bar{\tau}\big\}. \end{aligned} \tag{8.1}$$

Proof. The proof of this theorem is for the most part identical to the proof of Proposition 9.1 of [9]; assume the last statement to be false; then there exist $\tau_2 > 0$, $\gamma > 0$, and a sequence of time steps h and a sequence of integers $m(h)$ such that $hm(h)$ converges to τ_2 and

$$\Big|V^{m(h)}\Big|^2_{U^{m(h)}} \ge \operatorname{ess\,sup}\big\{|\dot{u}(t)|^2_{u(t)} : t_0 \le t \le t_0 + \bar{\tau}\big\} + \gamma. \tag{8.2}$$

Moreover, $u(t_0 + \tau_2)$ belongs to ∂K.

Choose a coordinate system such that the origin is at $u(t_0 + \tau_2)$; let Ψ be the diffeomorphism defined at (2.5). In this case, $D\Psi(0)$ is given by (2.6). Define

$$\beta^m = \big(\xi^m\big)^T D\Psi(0)^T M(0) D\Psi(0) \xi^m.$$

Let us compare β^m to $|V^m|^2_{U^m}$; it is convenient to define

$$\widetilde{V}^m = D\Psi(0)\xi^m;$$

then

$$\begin{aligned} |V^m|^2_{U^m} - \beta^m &= \big(V^m\big)^T M(U^m) V^m - \big(\widetilde{V}^m\big)^T M(0) \widetilde{V}^m \\ &= \big(V^m\big)^T \big(M(U^m) - M(0)\big) V^m - \big(V^m - \widetilde{V}^m\big) M(0) \big(V^m - \widetilde{V}^m\big) \\ &\quad + 2\big(V^m - \widetilde{V}^m\big)^T M(0) V^m. \end{aligned}$$

We observe that

$$|U^m| \le \|u_h - u\| + C_3 \left|mh - \tau_2\right|, \quad \|X^m\| \le \Lambda \|U^m\|$$

and that

$$\left|V^m - \widetilde{V}^m\right| \le C_4 \|\xi^m\| \left[2\Lambda \|X^m\| + \left\|X^m - X^{m-1}\right\|\right].$$

These observations enable us to estimate the difference: there exists a constant C_{12} such that

$$\left||V^m|^2_{U^m} - \beta^m\right| \le C_{12}\big(h + \|u - u_h\|_{C^0([t_0, t_0+\tau])} + |mh - \tau_2|\big).$$

We infer from (2.34) that there exists a constant C_{13} such that

$$\beta^{m+1} \le \min\big(\beta^m, \beta^{m-1}\big) + C_{13} h.$$

We now use (8.2): we can see that for all $m \le m(h)$,

$$\beta^{m(h)} \le \max\big(\beta^m, \beta^{m-1}\big) + C_{13}\big(m(h) - m\big)h,$$

so that

$$\max\bigl(|V^m|^2_{U^m}, |V^{m-1}|^2_{U^{m-1}}\bigr) \geq \beta^{m(h)} - C_{13}\bigl(m(h)-m\bigr)h - C_{12}\bigl(h + \|u-u_h\|_{C^0([t_0,t_0+\tau])} + |mh-\tau_2|\bigr).$$

If $\tau_4 < \tau_2$ is such that

$$\tau_2 - \tau_4 \leq \gamma/(4C_{13}),$$

and if

$$C_{12}\bigl(h + \|u-u_h\|_{C^0([t_0,t_0+\tau])} + |mh-\tau_2|\bigr) \leq \gamma/4,$$

we can see that for all small enough h and all $m \in \{\lceil \tau_4/h\rceil, \dots, m(h)\}$ the following estimate holds:

$$\max\bigl(|V^m|_{U^m}, |V^{m-1}|_{U^{m-1}}\bigr) \geq \operatorname{ess\,sup}\bigl\{|\dot u(t)|^2_{u(t)} : t_0 \leq t \leq t_0+\overline{\tau}\bigr\} + \gamma/4. \tag{8.3}$$

But the function v_h defined by

$$v_h(t) = |V^m|^2_{U^m} \text{ if } t \in [mh, (m+1)h)$$

converges almost everywhere on $[t_0, t_0+\tau]$ to $|\dot u(t)|^2_{u(t)}$; so does $\max(v_h(t-h), v_h(t))$. Therefore, in the limit, relation (8.3) leads to

$$\liminf_{h\downarrow 0} \operatorname*{ess\,sup}_{t\in[t_0+\tau_4, t_0+\tau_2]} v_h(t) \geq \operatorname{ess\,sup}\bigl\{|\dot u(t)|^2_{u(t)} : t_0 \leq t \leq t_0+\overline{\tau}\bigr\} + \gamma/4,$$

which is a contradiction.

Once again, we have been tempted to work in straightened coordinates; but it does not shorten the proof, and it introduces more notations. Therefore, it was hardly worth the effort. □

A corollary can be inferred immediately from this proposition and Theorem 2.3.

COROLLARY 8.2. *For all admissible initial conditions u_0 and p_0, there exists $\tau > 0$ and a solution of* (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), *and* (1.9) *defined on $[t_0, t_0+\tau]$.*

Above we have proved the existence of a nonempty interval on which the numerical scheme converges to a solution of (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), and (1.9). On the other hand, Lemma 7.1 gives a priori estimates on the solution of such a problem.

We couple now the a priori estimates with the local convergence result to obtain a global result.

THEOREM 8.3. *Let R be strictly larger than $|p_0|^*_{u_0}$, and let $\tau(R)$ be given as in Lemma* 7.1. *Then, for all small enough h, the solution U^m of the numerical scheme* (1.13), (1.14), (1.16), (1.17) *is defined on a discrete interval $\{0, \dots, m(h)\}$, such that*

$$\liminf_{h\to 0} hm(h) \to \tau(R);$$

moreover, the approximation u_h converges to a solution u of the continuous time equation, i.e., (1.3), (1.4a), (1.4b), (1.4c), (1.5a), (1.5b), (1.6), (1.8), *and* (1.9), *which is defined on $[t_0, t_0+\tau(R)]$.*

Proof. Let C_2 be given by

$$C_2 = \max\left\{ (3\,|v_0|_{u_0} + 2) \left\| M(u_0)^{-1/2} \right\|, 1 + R \max\left\{ \left\| M(u)^{-1/2} \right\| ; |u - u_0| \le R \right\} \right\}.$$

We know from Theorem 2.4 that there exist nonnegative numbers r_1, τ_2, C_3, and h_1 such that for all $\widehat{t}_0 \in [0,T)$, for all $\overline{u} \in K \cap B(u_0, R+1)$, and for all $h \in (0, h_1]$, if $\widehat{U}^0$, $\widehat{U}^1$ satisfy the condition $E(\overline{u}, r_1, C_2, h)$, then the numerical scheme (1.16)–(1.17) has a unique solution which satisfies the estimate

$$\forall m \in \{0, \dots, \lfloor \tau_2/h \rfloor\} \quad \left| \widehat{U}^{m+1} - \widehat{U}^m \right| \le C_3 h.$$

Let $\{0, \dots, m(h)\}$ be the maximal discrete time interval for which the numerical scheme (1.13), (1.14), (1.16), (1.17) has a solution satisfying

$$|V^m| \le C_3.$$

Let

$$\tau_1 = \liminf_{h \to 0} h m(h).$$

From Theorem 2.1 we know that τ_1 is at least equal to some number $\tau > 0$. Assume that τ_1 is strictly inferior to $\tau(R)$. Proposition 8.1 and the a priori estimates proved in Lemma 7.1 imply that, for h small enough and for all $\varepsilon > 0$,

$$\begin{aligned} &\forall t_m \in [t_0, t_0 + \tau_1 - \varepsilon], \qquad U^m \in B(u_0, R+1), \\ &\text{and } |V^m|^2_{U^m} \le \operatorname{ess\,sup}\left\{ |\dot{u}(t)|^2_{u(t)} : t_0 \le t_m \le t_0 + \tau_1 - \varepsilon \right\} \le R^2 + 1. \end{aligned}$$

Since the above estimates hold for all $\varepsilon > 0$, we see that

$$\begin{aligned} &\forall t_m \in [t_0, t_0 + \tau_1], \qquad U^m \in B(u_0, R+1), \\ &\text{and } |V^m|^2 \le 1 + R \max\left\{ \left\| M(u)^{-1/2} \right\| : |u - u_0| \le R \right\}. \end{aligned}$$

We denote

$$\ell(h) = \lfloor (\tau_1 - \tau_2/2)/h \rfloor$$

and we re-initialize with the following choices:

$$\widehat{t}_0 = t_0 + \ell(h) h, \quad \widehat{U}^0 = U^{\ell(h)}, \quad \widehat{U}^1 = U^{\ell(h)+1}.$$

With these data, we know that $\widehat{U}^m$ exists for $0 \le mh \le \tau_2$, so that the numerical solution U^m is extended up to $\lfloor (\tau_1 + \tau_2/2)/h \rfloor - 1$, and therefore

$$\liminf_{h \to 0} h m(h) \ge \tau_1 + \tau_2/2,$$

which is a contradiction. ∎

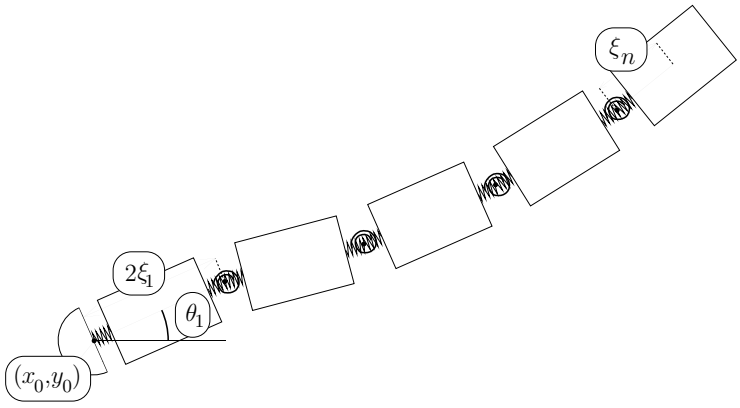


FIG. 9.1. *The discretization of the bar and the generalized coordinates.*

9. A numerical experiment: The dropped bar of Stoianovici and Hurmuzlu. In a recent paper [18], D. Stoianovici and Y. Hurmuzlu report experiments on the dynamics of a slender impacting bar, where they performed measurements of the apparent restitution coefficient of the bar.

The experiments consisted of dropping a bar on a rigid horizontal massive support: the velocity vanished initially and the bar initially made an angle θ with respect to the horizontal. This angle was varied in the experiments. The main conclusion of [18] was that the apparent coefficient of restitution, calculated by taking the ratio of the normal velocity of the impacting point after impact to the normal velocity of the impacting point before impact, varied as a function of θ.

We use a mechanical model of a bar with a finite number of degrees of freedom: the bar is discretized by a finite number n of identical cylindrical segments (of circular section) and a half-sphere at the impacting end; the nearest segment to the impacting half-sphere is joined by a linear spring; every other segment is joined to its neighbors by three springs and an articulation; see Figure 9.1.

Let θ_i be the angle between the ith segment and the horizontal; let $2\xi_i$ be the length of the ith segment plus the springs, except for the nth segment, where the convention is displayed on Figure 9.1. Denote by (x_0, y_0) the coordinates of the center of the impacting hemisphere. With these notations, the center of gravity of the ith segment has coordinates

$$x_i = x_0 + 2\sum_{j=1}^{i-1} \xi_j \cos\theta_j + \xi_i \cos\theta_i, \quad y_i = y_0 + 2\sum_{j=1}^{i-1} \xi_j \sin\theta_j + \xi_i \sin\theta_i.$$

We have made the following choice of generalized coordinates:

$$u = \big(x_0, y_0, \theta_1, \ldots, \theta_n, \xi_1, \ldots, \xi_n\big).$$

Whenever convenient, we will denote by u_i the ith coordinate of u.

The kinetic energy and the potential energy are defined, respectively, by

$$T = \frac{\mu}{2}\left(\dot{x}_0^2 + \dot{y}_0^2\right) + \frac{m}{2}\sum_{i=1}^{n}\left(\dot{x}_i^2 + \dot{y}_i^2\right) + \frac{J\dot{\theta}_1^2}{2} + \frac{I}{2}\sum_{i=1}^{n}\dot{\theta}_i^2,$$

$$U = \mu g y_0 + mg\sum_{i=1}^{n} y_i + k\sum_{i=1}^{n-1}\left(\xi_i - L\right)^2 + \frac{k}{2}\left(\xi_n - L\right)^2 + \frac{\Gamma}{2}\sum_{i=1}^{n-1}\left(\theta_i - \theta_{i+1}\right)^2.$$

Here m is the mass of each segment, μ is the mass of the end hemisphere, and I and J are appropriate moments of inertia.

We use all the physical constants chosen by the authors of [18], and the length of the bar is 200 mm. Our discretization used $n = 9$ segments, i.e., 20 degrees of freedom.

The set of constraints is

$$K = \{u \in \mathbb{R}^{2n+2} : \phi(u) = y_0 - R \geq 0\}.$$

Let e_2 be the second vector of the canonical basis of $\mathbb{R}^{2n+2}$; let

$$v(u) = M(u)^{-1}e_2,$$

and denote by $v_i(u)$ the ith component of $v(u)$. With these notations, the transmission law at impact is given for all $i = 1, \ldots, 2n+2$ by

$$\dot{u}_i(t+0) = \dot{u}_i(t-0) - (1+e)\frac{v_i(u)}{v_2(u)}\dot{u}_2(t-0).$$

For $i = 2$, we find

$$\dot{u}_2(t+0) = -e\dot{u}_2(t-0),$$

which is precisely Newton's law for the impacting degree of freedom.

In the practical implementation of the scheme, we freeze the metric at U^n, and therefore we utilize the projection P_K^n onto K with respect to the metric defined by $M(U^n)$; at each step of the scheme we solve

$$\frac{U^{n+1} + eU^{n-1}}{1+e} = P_K^n\left(\frac{2U^n - (1-e)U^{n-1} + h^2F^n}{1+e}\right)$$

by Newton's method. Observe that in our experiments we took $e = 1$.

In order to compare our ad hoc scheme to methods whose reliability is known, we set up a detection method: we integrate the free flight by a Newmark scheme, and we seek the impact time by looking for zeros of the parabolic interpolation of $y_0 - R$.

In Figure 9.2, we show the apparent restitution coefficient obtained by our two methods. There are many microimpacts; the apparent coefficient of restitution is the ratio of the vertical velocity of the impacting point after all the microimpacts to the vertical velocity before impact; we plot this ratio as a function of θ. In this simulation, the ad hoc scheme is approximately 40% faster than the detection scheme.

We checked that numerically the total energy is conserved; the result plotted here implies that a significant part of the global kinetic energy is transferred to the continuous medium modes.

The reader is referred to [18] for a comparison with the experimental results.

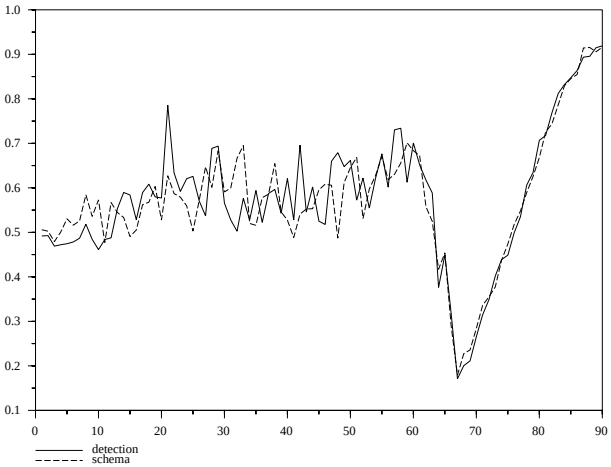


FIG. 9.2. *The apparent coefficient of restitution as a function of the angle θ of the bar with the horizontal.*

REFERENCES

[1] P. BALLARD, *Dynamique des systèmes mécaniques avec liaisons unilatérales parfaites*, C. R. Acad. Sci. Paris Sér. IIb, 327 (1999), pp. 953–958.

[2] A. BRESSAN, *Questioni di regolarità e di unicità del moto in presenza di vincoli olonomi unilaterali*, Rend. Sem. Mat. Univ. Padova, 29 (1959), pp. 271–315.

[3] G. BUTTAZZO AND D. PERCIVALE, *On the approximation of the elastic bounce problem on Riemannian manifolds*, J. Differential Equations, 47 (1983), pp. 227–245.

[4] M. CARRIERO AND E. PASCALI, *Uniqueness of the one-dimensional bounce problem as a generic property in* $L^1([0, T]; \mathbf{r})$, Boll. Unione Mat. Ital. Sez. A (6), 1 (1982), pp. 87–91.

[5] M. MABROUK, *Liaisons unilatérales et chocs élastiques quelconques: un résultat d'existence*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1353–1357.

[6] M. MABROUK, *A unified variational model for the dynamics of perfect unilateral constraints*, Eur. J. Mech. A Solids, 17 (1998), pp. 819–842.

[7] M. PANET, L. PAOLI, AND M. SCHATZMAN, *Theoretical and numerical study for a model of vibrations with unilateral constraints*, in Contact Mechanics, M. Raous, M. Jean, and J.-J. Moreau, eds., Plenum Press, New York, 1995, pp. 457–464.

[8] L. PAOLI, *Analyse numérique de vibrations avec contraintes unilatérales*, Ph.D. thesis, Université Claude Bernard—Lyon 1, Lyon, France, 1993.

[9] L. PAOLI AND M. SCHATZMAN, *A numerical scheme for impact problems:* I. *The one-dimensional case*, SIAM J. Numer. Anal., 40 (2002), pp. 702–733.

[10] L. PAOLI AND M. SCHATZMAN, *Mouvement à un nombre fini de degrés de liberté avec contraintes unilatérales : cas avec perte d'énergie*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 673–717.

[11] L. PAOLI AND M. SCHATZMAN, *Schéma numérique pour un modèle de vibrations avec contraintes unilatérales et perte d'énergie aux impacts, en dimension finie*, C.R. Acad. Sc. Paris Ser. I Math., 317 (1993), pp. 211–215.

[12] L. PAOLI AND M. SCHATZMAN, *Dynamics of an impacting bar: a model and numerics*, in Proceedings of the European Congress on Computational Mechanics, W. Wunderlich, ed., CD-ROM, Munich, 1999.

[13] D. PERCIVALE, *Uniqueness in the elastic bounce problem*, J. Differential Equations, 56 (1985), pp. 206–215.

[14] D. PERCIVALE, *Bounce problem with weak hypotheses of regularity*, Ann. Mat. Pura Appl. (4), 143 (1986), pp. 259–274.

[15] D. PERCIVALE, *Uniqueness in the elastic bounce problem.* II, J. Differential Equations, 90 (1991), pp. 304–315.

[16] M. Schatzman, *A class of nonlinear differential equations of second order in time*, Nonlinear Anal., 2 (1978), pp. 355–373.
[17] M. Schatzman, *Uniqueness and continuous dependence on data for one-dimensional impact problems*, *Recent advances in contact mechanics*, Math. Comput. Modelling, 28 (1998), pp. 1–18.
[18] D. Stoianovici and Y. Hurmuzlu, *A critical study of the applicability of rigid body collision theory*, Trans. ASME J. Appl. Mech., 63 (1996), pp. 307–316.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 769–791

A LOCAL DISCONTINUOUS GALERKIN METHOD FOR KdV TYPE EQUATIONS*

JUE YAN† AND CHI-WANG SHU†

Abstract. In this paper we develop a local discontinuous Galerkin method for solving KdV type equations containing third derivative terms in one and two space dimensions. The method is based on the framework of the discontinuous Galerkin method for conservation laws and the local discontinuous Galerkin method for viscous equations containing second derivatives; however, the guiding principle for intercell fluxes and nonlinear stability is new. We prove L^2 stability and a cell entropy inequality for the square entropy for a class of nonlinear PDEs of this type in both one and multiple space dimensions, and we give an error estimate for the linear cases in the one-dimensional case. The stability result holds in the limit case when the coefficients to the third derivative terms vanish; hence the method is especially suitable for problems which are "convection dominated," i.e., those with small second and third derivative terms. Numerical examples are shown to illustrate the capability of this method. The method has the usual advantage of local discontinuous Galerkin methods, namely, it is extremely local and hence efficient for parallel implementations and easy for *h-p* adaptivity.

Key words. discontinuous Galerkin method, KdV equation, stability, error estimate

AMS subject classifications. 65M60, 35Q53

PII. S0036142901390378

1. Introduction.

In this paper we develop a local discontinuous Galerkin method for solving KdV type equations containing third derivative terms in one and multiple space dimensions. An example of such a PDE is the original KdV equation [20],

$$U_t + (\alpha U + \beta U^2)_x + \sigma U_{xxx} = 0, \tag{1.1}$$

where α, β, and σ are constants. In this paper we use capital letters such as U, V, Q, etc. to denote the solutions to the PDEs and use lowercase letters to denote the numerical solutions. Our scheme can be designed and proven stable for more general nonlinearities, namely,

$$U_t + f(U)_x + (r'(U)g(r(U)_x)_x)_x = 0 \tag{1.2}$$

in one space dimension for arbitrary (smooth) functions f, g, and r, and

$$U_t + \sum_{i=1}^{d} f_i(U)_{x_i} + \sum_{i=1}^{d} \left(r_i'(U) \sum_{j=1}^{d} g_{ij}(r_i(U)_{x_i})_{x_j} \right)_{x_i} = 0 \tag{1.3}$$

in multiple space dimensions for arbitrary (smooth) functions f_i, g_{ij}, and r_i, $i, j = 1, \ldots, d$.

*Received by the editors June 2, 2001; accepted for publication (in revised form) February 11, 2002; published electronically July 24, 2002. This research was supported by ARO grant DAAD19-00-1-0405; NSF grants DMS-9804985 and ECS-9906606; NASA Langley grant NCC1-01035 and contract NAS1-97046 while the second author was in residence at ICASE, NASA Langley Research Center, Hampton, VA; and by AFOSR grant F49620-99-1-0077.

http://www.siam.org/journals/sinum/40-2/39037.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (yjue@cfm.brown.edu, shu@cfm.brown.edu).

KdV type equations describe the propagation of waves in a variety of nonlinear, dispersive media and appear often in applications. See, e.g., [1]. Various numerical methods have been proposed and used in practice to solve this type of equation; see, e.g., [4, 5, 19]. However, in many situations, such as in quantum hydrodynamic models of semiconductor device simulations [16] and in the dispersive limit of conservation laws [21], the third derivative terms might have small or even zero coefficients in some parts of the domain. We will call such cases "convection dominated." The design of stable, efficient, and high order methods, especially those for the "convection dominated" cases, i.e., when the third derivative terms are small ($|\sigma| \ll 1$ in (1.1)), remains a challenge, as the success of such methods depends crucially on correct treatment of this singular limit. The situation is similar to convection diffusion problems. While a large class of methods suitable for parabolic problems also works for convection diffusion problems when diffusion dominates, these methods may not work well when convection dominates.

The discontinuous Galerkin method is a class of finite element methods using completely discontinuous piecewise polynomial space for the numerical solution and the test functions. One certainly needs to use more degrees of freedom because of the discontinuities at the element boundaries; however, this also gives one room to design suitable interelement boundary treatments (the so-called fluxes) to obtain highly accurate and stable methods in many difficult situations.

The first discontinuous Galerkin method was introduced in 1973 by Reed and Hill [24] in the framework of neutron transport (steady state linear hyperbolic equations). Convergence for such methods was proven in, e.g., [22]. A major development of the discontinuous Galerkin method was carried out by Cockburn et al. in a series of papers [11, 10, 8, 12] in which a framework was established to easily solve *nonlinear* time dependent hyperbolic conservation laws (i.e., (1.2) and (1.3) without the third derivative terms) using explicit, nonlinearly stable high order Runge–Kutta time discretizations [26] and discontinuous Galerkin discretization in space with exact or approximate Riemann solvers as interface fluxes and TVB (total variation bounded) nonlinear limiters [25] to achieve nonoscillatory properties for strong shocks. See also [17] for a discontinuous Galerkin method with additional "shock capturing" terms.

The discontinuous Galerkin method has found rapid applications in such diverse areas as aeroacoustics, electromagnetism, gas dynamics, granular flows, magnetohydrodynamics, meteorology, modeling of shallow water, oceanography, oil recovery simulation, semiconductor device simulation, transport of contaminant in porous media, turbomachinery, turbulent flows, viscoelastic flows, and weather forecasting, among many others. Good references for the discontinuous Galerkin method and its recent development include the survey paper [9], other papers therein, and the review paper [14].

The original discontinuous Galerkin method was designed to solve first order hyperbolic problems. A simple example to illustrate its essential ideas is the linear transport equation

$$U_t + U_x = 0. \tag{1.4}$$

Let's denote the mesh by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ for $j = 1, \ldots, N$, with the center of the cell denoted by $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$ and the size of each cell by $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. We will denote $\Delta x = \max_j \Delta x_j$. If we multiply (1.4) by an arbitrary test function

$V(x)$, integrate over the interval I_j, and integrate by parts, we get

$$\int_{I_j} U_t V dx - \int_{I_j} U V_x dx + U(x_{j+\frac{1}{2}}, t)V(x_{j+\frac{1}{2}}) - U(x_{j-\frac{1}{2}}, t)V(x_{j-\frac{1}{2}}) = 0. \tag{1.5}$$

This is the starting point for designing the discontinuous Galerkin method. We replace both the solution U and the test function V by piecewise polynomials of degree at most k and denote them by u and v. That is, $u, v \in \mathcal{V}_{\Delta x}$, where

$$\mathcal{V}_{\Delta x} = \{v : \ v \text{ is a polynomial of degree at most } k \text{ for } x \in I_j, \ \ j = 1, \ldots, N\}. \tag{1.6}$$

When U and V in (1.5) are replaced by u and v, there is ambiguity in the last two terms of (1.5) involving the boundary values at $x_{j\pm\frac{1}{2}}$, as both the solution u and the test function v are *discontinuous* exactly at these boundary points. The idea is to treat these terms with an upwinding mechanism (information from characteristics) borrowed from successful high resolution finite volume schemes. Thus u at the interfaces $x_{j\pm\frac{1}{2}}$ is given by a single valued numerical flux $\hat{u}_{j\pm\frac{1}{2}} = u^-_{j\pm\frac{1}{2}}$, determined from upwinding, and the v's at the interfaces $x_{j\pm\frac{1}{2}}$ are given by the values taken from inside the cell I_j, namely, $v^-_{j+\frac{1}{2}}$ and $v^+_{j-\frac{1}{2}}$. Notice that we use v^- and v^+ to denote the left and right limits of v, respectively, at the interface where v is discontinuous. For more general nonlinear fluxes $f(u)$, the only difference is that the single valued flux $\hat{f}_{j+\frac{1}{2}}$ would be taken as a monotone flux depending on both $u^-_{j+\frac{1}{2}}$ and $u^+_{j+\frac{1}{2}}$ (exact or approximate Riemann solvers in the system case). The resulting method of the method-of-lines ODE is then discretized by the nonlinearly stable high order Runge–Kutta time discretizations [26]. Nonlinear TVB limiters [25] may be used if the solution contains strong discontinuities. The scheme thus obtained, for solving hyperbolic conservation laws ((1.2) and (1.3) without the third derivative terms), has the following attractive properties:

1. It can be easily designed for any order of accuracy. In fact, the order of accuracy can be locally determined in each cell, thus allowing for efficient p adaptivity.
2. It can be used on arbitrary triangulations, even those with hanging nodes, thus allowing for efficient h adaptivity.
3. It is extremely local in data communications. The evolution of the solution in each cell needs to communicate only with the immediate neighbors, regardless of the order of accuracy, thus allowing for efficient parallel implementations. See, e.g., [3].
4. It has excellent provable nonlinear stability. One can prove a strong L^2 stability and a cell entropy inequality for the square entropy, for the general nonlinear cases, for any orders of accuracy on arbitrary triangulations in any space dimension, without the need for nonlinear limiters [18]. See also [17] for a nonlinear stability result of a discontinuous Galerkin method with extra "shock capturing" terms.

In [13] these discontinuous Galerkin methods were generalized to solve convection diffusion problems containing second derivative terms. This generalization was motivated by the successful numerical experiments of Bassi and Rebay [2] for the compressible Navier–Stokes equations. The idea can be illustrated with the simple heat equation

$$U_t - U_{xx} = 0 \tag{1.7}$$

which we rewrite into a first order system

$$U_t - Q_x = 0, \qquad Q - U_x = 0. \tag{1.8}$$

We can then *formally* use the same discontinuous Galerkin method for the convection equation to solve (1.8), resulting in the following scheme: Find $u, q \in \mathcal{V}_{\Delta x}$ such that, for all test functions $v, w \in \mathcal{V}_{\Delta x}$,

$$\begin{aligned} &\int_{I_j} u_t v dx + \int_{I_j} q v_x dx - \hat{q}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0, \\ &\int_{I_j} q w dx + \int_{I_j} u w_x dx - \hat{u}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} = 0. \end{aligned} \tag{1.9}$$

However, there is no longer an upwinding mechanism or characteristics to guide the design of the fluxes $\hat{u}_{j+\frac{1}{2}}$ and $\hat{q}_{j+\frac{1}{2}}$. The crucial part in designing a stable and accurate algorithm (1.9) is a correct design of these fluxes. In [13], criteria are given for these fluxes to guarantee stability, convergence, and a suboptimal error estimate of order k in the L^2 norm for piecewise polynomials of degree k. The (most natural) central fluxes

$$\hat{u}_{j+\frac{1}{2}} = \frac{1}{2}\left(u^-_{j+\frac{1}{2}} + u^+_{j+\frac{1}{2}}\right), \qquad \hat{q}_{j+\frac{1}{2}} = \frac{1}{2}\left(q^-_{j+\frac{1}{2}} + q^+_{j+\frac{1}{2}}\right) \tag{1.10}$$

would satisfy these criteria and give a scheme which is indeed suboptimal in the order of accuracy for odd k (i.e., the accuracy is order k rather than the expected order $k+1$ for odd k). This deficiency, however, is easily removed by a clever choice of fluxes, proposed in [13]:

$$\hat{u}_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = q^+_{j+\frac{1}{2}}; \tag{1.11}$$

i.e., we alternatively take the left and right limits for the fluxes in u and q (we could of course also take the pair $u^+_{j+\frac{1}{2}}$ and $q^-_{j+\frac{1}{2}}$ as the fluxes). Notice that the evaluation of (1.11) is simpler than that of the central fluxes in (1.10), and this easily generalizes to multispace dimensions on arbitrary triangulations. The accuracy now becomes the optimal order $k+1$ for both even and odd k; see [6] for a proof of this in the general h-p context. Notice that this alternating way of choosing fluxes may render the resulting system nonsymmetric except for the P^0 case, although it still will be positive definite. This is not a problem if the method is discretized explicitly in time or if an implicit time discretization is solved by an iterative method suitable for positive definite matrices.

The schemes thus designed for the heat equation (1.7), or in fact for the most general multidimensional nonlinear convection diffusion equations (nonlinear in both the first derivative convection part and the second derivation diffusion part), retain *all* four nice properties listed above for the method used on convection equations. Moreover, the appearance of the auxiliary variable q is superficial: when a local basis is chosen in cell I_j, q is eliminated and the actual scheme for u takes a form similar to that for convection alone. This is a major advantage of the scheme over the traditional "mixed methods," and it is the reason that the scheme is termed *local* discontinuous Galerkin method in [13]. Even though the auxiliary variable q can be locally eliminated, it does approximate the derivative of the solution u to the same order of accuracy, thus matching the advantage of traditional "mixed methods" on this.

The purpose of this paper is to develop a similar local discontinuous Galerkin method for the KdV type equations (1.1), (1.2), and (1.3) containing third derivative terms. Our objective is to design a method retaining again all four nice properties listed above for the method used on convection and convection diffusion equations, plus having the feature of being local, namely, the auxiliary variables introduced to approximate the first and second derivatives of the solution could be locally eliminated.

The organization of the paper is as follows. In subsection 1.1 we give a short "preview" of the proposed method on a simple linear equation to motivate the ideas. In section 2 we describe the method for the one-dimensional case and prove its nonlinear L^2 stability and a cell entropy inequality, as well as an error estimate for the linear case. In section 3 the multiple space dimension case is considered, where the nonlinear stability is given for the general triangulations. In section 4 we provide several numerical examples to illustrate the capability of the method. Concluding remarks and remarks about future work are given in section 5.

1.1. A preview of the method. We will give a "preview" of the method on the simple linear equation

$$U_t + U_{xxx} = 0 \tag{1.12}$$

which we again rewrite into a first order system

$$U_t + P_x = 0, \quad P - Q_x = 0, \quad Q - U_x = 0. \tag{1.13}$$

We can then *formally* use the same discontinuous Galerkin method for the convection equation to solve (1.13), resulting in the following scheme: Find $u, p, q \in \mathcal{V}_{\Delta x}$ such that, for all test functions $v, w, z \in \mathcal{V}_{\Delta x}$,

$$\begin{aligned}
&\int_{I_j} u_t v dx - \int_{I_j} p v_x dx + \hat{p}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - \hat{p}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0,\\
&\int_{I_j} p w dx + \int_{I_j} q w_x dx - \hat{q}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} = 0,\\
&\int_{I_j} q z dx + \int_{I_j} u z_x dx - \hat{u}_{j+\frac{1}{2}} z^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} z^+_{j-\frac{1}{2}} = 0.
\end{aligned} \tag{1.14}$$

However, the fluxes $\hat{p}_{j+\frac{1}{2}}$, $\hat{q}_{j+\frac{1}{2}}$, and $\hat{u}_{j+\frac{1}{2}}$ must be designed based on guiding principles different than the first order convection or second order diffusion cases. The crucial part in designing a stable and accurate algorithm (1.14) is again a correct design of these fluxes. It turns out that the simple choices

$$\hat{p}_{j+\frac{1}{2}} = p^+_{j+\frac{1}{2}}, \quad \hat{q}_{j+\frac{1}{2}} = q^+_{j+\frac{1}{2}}, \quad \hat{u}_{j+\frac{1}{2}} = u^-_{j+\frac{1}{2}}, \tag{1.15}$$

which are partially motivated by upwinding (a simple wave solution to (1.12) moves from right to left), would guarantee stability and convergence, as will be proven later in this paper and also clearly seen in Example 4.1 in section 4.

We remark that the choice for the fluxes (1.15) is not unique. In fact, the crucial part is to take $\hat{p}$ and $\hat{u}$ from opposite sides and to take $\hat{q}$ from the right. Thus

$$\hat{p}_{j+\frac{1}{2}} = p^-_{j+\frac{1}{2}}, \quad \hat{q}_{j+\frac{1}{2}} = q^+_{j+\frac{1}{2}}, \quad \hat{u}_{j+\frac{1}{2}} = u^+_{j+\frac{1}{2}}$$

would also work.

2. The local discontinuous Galerkin method for the one-dimensional case. In this section, we present and analyze the local discontinuous Galerkin method for the following one-dimensional nonlinear problem:

$$U_t + f(U)_x + (r'(U)g(r(U)_x)_x)_x = 0, \qquad 0 \le x \le 1, \tag{2.1}$$

with an initial condition

$$U(x,0) = U^0(x), \qquad 0 \le x \le 1, \tag{2.2}$$

and periodic boundary conditions. Here $f(U)$, $r(U)$, and $g(Q)$ (with Q defined by (2.3) below) are arbitrary (smooth) nonlinear functions. Notice that the assumption of periodic boundary conditions is for simplicity only and is not essential: the method can be easily designed for nonperiodic boundary conditions. Also notice that both the linear equation (1.12) and the KdV equation (1.1) are special cases of (2.1).

To define the local discontinuous Galerkin method, we first introduce the new variables

$$Q = r(U)_x, \qquad P = g(Q)_x \tag{2.3}$$

and rewrite (2.1) as a first order system:

$$U_t + (f(U) + r'(U)P)_x = 0, \quad P - g(Q)_x = 0, \quad Q - r(U)_x = 0. \tag{2.4}$$

The local discontinuous Galerkin method is obtained by discretizing the above system with the discontinuous Galerkin method. This is achieved by multiplying the three equations in (2.4) by three test functions v, w, z, respectively, integrating over the interval I_j, and integrating by parts. We also need to pay special attention to the boundary terms resulting from the procedure of integration by parts, as mentioned in the previous section. Thus we seek piecewise polynomial solutions $u, p, q \in \mathcal{V}_{\Delta x}$, where $\mathcal{V}_{\Delta x}$ is defined in (1.6) and consists of piecewise polynomials of degree up to k in each cell I_j such that for all test functions $v, w, z \in \mathcal{V}_{\Delta x}$ we have, for $1 \le j \le N$,

$$\begin{aligned}
&\int_{I_j} u_t v dx - \int_{I_j} (f(u) + r'(u)p) v_x dx + \left(\hat{f} + \widehat{r'}\hat{p}\right)_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - \left(\hat{f} + \widehat{r'}\hat{p}\right)_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0,\\
&\int_{I_j} p w dx + \int_{I_j} g(q) w_x dx - \hat{g}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} + \hat{g}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} = 0,\\
&\int_{I_j} q z dx + \int_{I_j} r(u) z_x dx - \hat{r}_{j+\frac{1}{2}} z^-_{j+\frac{1}{2}} + \hat{r}_{j-\frac{1}{2}} z^+_{j-\frac{1}{2}} = 0.
\end{aligned} \tag{2.5}$$

The only ambiguity in the algorithm (2.5) now is the definition of the numerical fluxes (the "hats"), which should be designed based on guiding principles different than the first order convection or second order diffusion cases to ensure stability. It turns out that we can take the following simple choices (we omit the subscripts $j \pm \frac{1}{2}$ in the definition of the fluxes, as all quantities are evaluated at the interfaces $x_{j\pm\frac{1}{2}}$):

$$\hat{f} = \hat{f}(u^-, u^+), \quad \widehat{r'} = \frac{r(u^+) - r(u^-)}{u^+ - u^-}, \quad \hat{p} = p^+, \quad \hat{g} = \hat{g}(q^-, q^+), \quad \hat{r} = r(u^-), \tag{2.6}$$

where $\hat{f}(u^-, u^+)$ is a monotone flux for $f(u)$, namely, $\hat{f}(u^-, u^+)$ is a Lipschitz continuous function in both arguments u^- and u^+, is consistent with $f(u)$ in the sense that

$\hat{f}(u,u) = f(u)$, and is a nondecreasing function in u^- and a nonincreasing function in u^+. Likewise, $-\hat{g}(q^-,q^+)$ is a monotone flux for $-g(q)$, namely, $\hat{g}(q^-,q^+)$ is a Lipschitz continuous function in both arguments q^- and q^+, is consistent with $g(q)$ in the sense that $\hat{g}(q,q) = g(q)$, and is a nonincreasing function in q^- and a nondecreasing function in q^+. Examples of monotone fluxes which are suitable for discontinuous Galerkin methods can be found in, e.g., [11]. We could, for example, use the simple Lax–Friedrichs flux

$$\hat{f}(u^-,u^+) = \frac{1}{2}\left(f(u^-) + f(u^+) - \alpha(u^+ - u^-)\right), \qquad \alpha = \max_u |f'(u)|, \tag{2.7}$$

where the maximum is taken over a relevant range of u. The algorithm is now well defined.

We remark that the choice for the fluxes (2.6) is not unique. In fact, the crucial part is to take $\hat{p}$ and $\hat{r}$ from opposite sides. Thus

$$\hat{f} = \hat{f}(u^-,u^+), \quad \widehat{r'} = \frac{r(u^+) - r(u^-)}{u^+ - u^-}, \quad \hat{p} = p^-, \quad \hat{g} = \hat{g}(q^-,q^+), \quad \hat{r} = r(u^+)$$

would also work. Since information for the third order equation (2.1) flows in preferred directions, the choice of monotone fluxes above is heuristically justified based on upwind considerations. Of course, the rigorous justification for the choice of fluxes comes from the stability results to be proven in Proposition 2.1 below.

We also remark that the algorithm (2.5)–(2.6) is very easy for numerical implementation. Given u, one first uses the third equation in (2.5) to obtain q. This is achieved locally: q in I_j can be obtained with the information of u in the cells I_j and I_{j-1}. The second equation in (2.5) is then used to obtain p locally: p in I_j can be obtained with the information of q in (at most) the cells I_j, I_{j-1}, and I_{j+1}. Finally, the update of the solution u is obtained using the first equation in (2.5), again locally, namely, the update of u in I_j can be obtained with the information of u in (at most) the cells I_j, I_{j-1}, and I_{j+1} and that of p in the cells I_j and I_{j+1}.

We have the following "cell entropy inequality" for the scheme (2.5)–(2.6). This is a generalization of the cell entropy inequality obtained in [18] for the discontinuous Galerkin method applied to hyperbolic conservation laws ((2.1) with $g(q) = 0$).

PROPOSITION 2.1 (cell entropy inequality). *There exist numerical entropy fluxes $\hat{H}_{j+\frac{1}{2}}$ such that the solution to the scheme* (2.5)–(2.6) *satisfies*

$$\frac{d}{dt}\int_{I_j}\left(\frac{u^2(x,t)}{2}\right)dx + \left(\hat{H}_{j+\frac{1}{2}} - \hat{H}_{j-\frac{1}{2}}\right) \le 0. \tag{2.8}$$

Proof. We sum up the three equalities in (2.5) and introduce the notation

$$\begin{aligned} B_j(u,p,q;v,w,z) = &\int_{I_j} u_t v dx - \int_{I_j}(f(u) + r'(u)p)v_x dx + \left(\hat{f} + \widehat{r'}\hat{p}\right)_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} \\ &- \left(\hat{f} + \widehat{r'}\hat{p}\right)_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} + \int_{I_j} pw dx + \int_{I_j} g(q)w_x dx - \hat{g}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} \\ &+ \hat{g}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} + \int_{I_j} qz dx + \int_{I_j} r(u)z_x dx - \hat{r}_{j+\frac{1}{2}} z^-_{j+\frac{1}{2}} + \hat{r}_{j-\frac{1}{2}} z^+_{j-\frac{1}{2}}. \end{aligned} \tag{2.9}$$

Clearly, the solutions u, p, q of the scheme (2.5)–(2.6) satisfy

$$B_j(u,p,q;v,w,z) = 0 \tag{2.10}$$

for all $v, w, z \in \mathcal{V}_{\Delta x}$. We then take

$$v = u, \quad w = q, \quad z = -p$$

to obtain, after some algebraic manipulations,

$$0 = B_j(u, p, q; u, q, -p) = \frac{d}{dt} \int_{I_j} \left(\frac{u^2(x,t)}{2} \right) dx + \left(\hat{H}_{j+\frac{1}{2}} - \hat{H}_{j-\frac{1}{2}} \right) + \Theta_{j-\frac{1}{2}}$$

with the numerical entropy flux $\hat{H}$ defined by

$$\hat{H} = -F(u^-) + G(q^-) - r(u^-)p^- + \left(\hat{f} + \widehat{r'}\hat{p} \right) u^- - \hat{g}q^- + \hat{r}p^- \tag{2.11}$$

and the extra term Θ given by

$$\Theta = [F(u) - G(q) + r(u)p] - \left(\hat{f} + \widehat{r'}\hat{p} \right) [u] + \hat{g}[q] - \hat{r}[p].$$

Here

$$F(u) = \int^u f(u)du, \qquad G(q) = \int^q g(q)dq,$$

and

$$[v] = v^+ - v^-$$

denotes the jump of v. Notice that we have dropped the subscripts about the location $j - \frac{1}{2}$ or $j + \frac{1}{2}$, as all these quantities are defined at a single interface and depend only on the left and right values at that interface. Now all we need to do is verify $\Theta \geq 0$. To this end, we notice that, with the definition (2.6) of the numerical fluxes and with simple algebraic manipulations, we easily obtain

$$[r(u)p] - \widehat{r'}\hat{p}[u] - \hat{r}[p] = 0,$$

and hence

$$\begin{aligned} \Theta &= [F(u)] - \hat{f}[u] - [G(q)] + \hat{g}[q] \\ &= \int_{u^-}^{u^+} \left(f(s) - \hat{f}(u^-, u^+) \right) ds - \int_{q^-}^{q^+} \left(g(s) - \hat{g}(q^-, q^+) \right) ds \\ &\geq 0, \end{aligned} \tag{2.12}$$

where the last inequality follows from the monotonicity of the fluxes $\hat{f}$ and $-\hat{g}$. This finishes the proof. ∎

Now the L^2 stability of the method is a trivial corollary as follows.

COROLLARY 2.2 (L^2 stability). *The solution to the scheme* (2.5)–(2.6) *satisfies the L^2 stability*

$$\frac{d}{dt} \int_0^1 \left(\frac{u^2(x,t)}{2} \right) dx \leq 0. \tag{2.13}$$

Proof. For the proof, we simply add up (2.8) over j. ∎

Regarding time discretizations, if we denote the semidiscrete local discontinuous Galerkin method (2.5)–(2.6) by

$$u_t = R(u),$$

then the following implicit θ scheme:

$$\frac{u^{n+1}-u^n}{\Delta t} = R(u^{n+\theta}), \qquad u^{n+\theta} = (1-\theta)u^n + \theta u^{n+1} \tag{2.14}$$

also will satisfy the same cell entropy inequality and L^2 stability as long as $\frac{1}{2} \le \theta \le 1$. Notice that this includes the first order backward Euler and second order Crank–Nicholson implicit time discretizations as special cases. See [18] for the purely hyperbolic case.

PROPOSITION 2.3 (implicit time discretization). *The cell entropy inequality and the L^2 stability also hold for the time discretization* (2.14) *with $\frac{1}{2} \le \theta \le 1$ for the scheme* (2.5)–(2.6). *That is,*

$$\int_{I_j} \left(\frac{(u^{n+1}(x))^2 - (u^n(x))^2}{2\Delta t} \right) dx + \hat{H}^{n+\theta}_{j+\frac{1}{2}} - \hat{H}^{n+\theta}_{j-\frac{1}{2}} \le 0 \tag{2.15}$$

and

$$\int_0^1 (u^{n+1}(x))^2 dx \le \int_0^1 (u^n(x))^2 dx. \tag{2.16}$$

Proof. If we take the test functions at $n+\theta$, e.g., $v = u^{n+\theta}$ given by (2.14), we obtain, just as before,

$$\int_{I_j} \frac{u^{n+1}(x) - u^n(x)}{\Delta t} u^{n+\theta} dx + \hat{H}^{n+\theta}_{j+\frac{1}{2}} - \hat{H}^{n+\theta}_{j-\frac{1}{2}} \le 0,$$

which can be rewritten as

$$\int_{I_j} \left(\frac{(u^{n+1}(x))^2 - (u^n(x))^2}{2\Delta t} \right) dx + \hat{H}^{n+\theta}_{j+\frac{1}{2}} - \hat{H}^{n+\theta}_{j-\frac{1}{2}} + \left(\theta - \frac{1}{2}\right) \int_{I_j} \left(\frac{(u^{n+1}(x) - u^n(x))^2}{\Delta t} \right) dx \le 0.$$

Thus, a sufficient condition for obtaining the cell entropy inequality (2.15) is just $\theta \ge \frac{1}{2}$. Again, (2.16) is obtained simply by adding up (2.15) over j. ∎

The stability result obtained here can be used to get an error estimate in L^2 for the numerical solution u when (2.1) is linear. Without loss of generality, we may take $f(U) = U$, $g(Q) = Q$, and $r(U) = U$, resulting in the equation

$$U_t + U_x + U_{xxx} = 0. \tag{2.17}$$

We have the following result, where C here and below denotes a generic constant which may be of different values at different locations.

PROPOSITION 2.4 (error estimate). *The error for the scheme* (2.5)–(2.6) *applied to the linear PDE* (2.17) *satisfies*

$$\sqrt{\int_0^1 \left(U(x,t) - u(x,t)\right)^2 dx} \le C\Delta x^{k+\frac{1}{2}}, \tag{2.18}$$

where the constant C depends on the first $k+3$ derivatives of U and time t.

Proof. First, we notice that, in this linear case, most monotone fluxes simply become upwinding,

$$\hat{f}(u^-, u^+) = u^-, \qquad \hat{g}(q^-, q^+) = q^+,$$

and this is what we will assume. It is then easy to work out the exact form of Θ in (2.12) for the cell entropy inequality to be

$$\Theta = \frac{1}{2}\left([u]^2 + [q]^2\right). \tag{2.19}$$

We now notice that the exact solution of the PDE (2.17), U, $Q = U_x$, and $P = U_{xx}$ clearly satisfies

$$B_j(U, P, Q; v, w, z) = 0$$

for all $v, w, z \in \mathcal{V}_{\Delta x}$, where B_j is defined by (2.9). Taking the difference between the above equality and (2.10), we obtain the *error equation*

$$B_j(U - u, P - p, Q - q; v, w, z) = 0 \tag{2.20}$$

for all $v, w, z \in \mathcal{V}_{\Delta x}$. As usual, this error equation is the basic starting point of error estimates.

We now take

$$v = \mathcal{S}U - u, \quad w = \mathcal{P}Q - q, \quad z = p - \mathcal{P}P \tag{2.21}$$

in the error equation (2.20). Here $\mathcal{P}$ is the standard L^2 projection into $\mathcal{V}_{\Delta x}$; that is, for each j,

$$\int_{I_j} (\mathcal{P}r(x) - r(x))s(x)dx = 0 \quad \forall s \in P^k,$$

where P^k denotes the space of all polynomials of degree at most k. In other words, the difference between the projection $\mathcal{P}r$ and the original function r is orthogonal to all polynomials of degree up to k in each interval. $\mathcal{S}$ is a special projection into $\mathcal{V}_{\Delta x}$ which satisfies, for each j,

$$\int_{I_j} (\mathcal{S}r(x) - r(x))s(x)dx = 0 \quad \forall s \in P^{k-1} \quad \text{and} \quad \mathcal{S}r(x^-_{j+1/2}) = r(x^-_{j+1/2});$$

in other words, the difference between the projection $\mathcal{S}r$ and the original function r is orthogonal to all polynomials of degree up to $k-1$ in each interval, and the projection agrees with the function at the right boundary in each interval. This special projection is needed for U because we have no control over the jumps of p in the cell entropy inequality; see (2.19). Substituting (2.21) into the error equation (2.20) and moving terms, we obtain

$$B_j(v, -z, w; v, w, z) = B_j(v^e, -z^e, w^e; v, w, z), \tag{2.22}$$

where v, w, z are given by (2.21), and v^e, w^e, z^e are given by

$$v^e = \mathcal{S}U - U, \quad w^e = \mathcal{P}Q - Q, \quad z^e = P - \mathcal{P}P. \tag{2.23}$$

By the same argument as that used for the cell entropy inequality, the left-hand side of (2.22) becomes

$$B_j(v,-z,w;v,w,z)=\frac{d}{dt}\int_{I_j}\left(\frac{v^2}{2}\right)dx+\left(\hat{H}_{j+\frac{1}{2}}-\hat{H}_{j-\frac{1}{2}}\right)+\Theta_{j-\frac{1}{2}}, \tag{2.24}$$

where, by (2.19),

$$\Theta_{j-\frac{1}{2}}=\frac{1}{2}\left([v]^2_{j-\frac{1}{2}}+[w]^2_{j-\frac{1}{2}}\right). \tag{2.25}$$

The right-hand side of (2.22) can be written out as

$$B_j(v^e,-z^e,w^e;v,w,z)=\mathcal{I}+\mathcal{II}+\mathcal{III}+\mathcal{IV}, \tag{2.26}$$

where

$$\mathcal{I}=\int_{I_j}v_t^e v dx, \tag{2.27}$$

$$\mathcal{II}=-\int_{I_j}z^e w dx+\int_{I_j}w^e z dx-\int_{I_j}(v^e-z^e)v_x dx+\int_{I_j}w^e w_x dx+\int_{I_j}v^e z_x dx, \tag{2.28}$$

$$\mathcal{III}=-\left(\left(v^e_{j-\frac{1}{2}}\right)^-\!-\left(z^e_{j-\frac{1}{2}}\right)^+\right)[v]_{j-\frac{1}{2}}+\left(w^e_{j-\frac{1}{2}}\right)^+[w]_{j-\frac{1}{2}}+\left(v^e_{j-\frac{1}{2}}\right)^-[z]_{j-\frac{1}{2}}, \tag{2.29}$$

and

$$\mathcal{IV}=\hat{h}_{j+\frac{1}{2}}-\hat{h}_{j-\frac{1}{2}} \tag{2.30}$$

for some flux function $\hat{h}$. Notice that v,w,z are given by (2.21) and v^e,w^e,x^e are given by (2.23), respectively.

Now, by using the simple inequality $ab\le\frac{1}{2}(a^2+b^2)$, and the standard approximation theory on $v_t^e=(\mathcal{S}U-U)_t$ (see, e.g., [7]), we have

$$\mathcal{I}\le C\Delta x_j^{2k+3}+\int_{I_j}\left(\frac{v^2}{2}\right)dx.$$

Because $\mathcal{P}$ is a local L^2 projection, and $\mathcal{S}$, even though not a local L^2 projection, does have the property that $w-\mathcal{S}w$ is locally orthogonal to all polynomials of degree up to $k-1$, all the terms in $\mathcal{II}$ are actually zero. The last term in $\mathcal{III}$ is zero because of the special interpolating property of the projection $\mathcal{S}$. An application of the simple inequality $ab\le\frac{1}{2}(a^2+b^2)$ for the first two terms in $\mathcal{III}$ and standard approximation theory on the point values of $v^e-z^e=(\mathcal{S}U-U)+(\mathcal{P}P-P)$ and of $w^e=\mathcal{P}Q-Q$ (see, e.g., [7]) then gives

$$\mathcal{III}\le C(\Delta x_{j-1}^{2k+2}+\Delta x_j^{2k+2})+\frac{1}{4}\left([v]^2+[w]^2\right).$$

Finally, $\mathcal{IV}$ only contains flux difference terms which will vanish upon a summation in j.

Combining all these and summing over j we obtain the inequality

$$\frac{d}{dt}\int_0^1 \left(\frac{v^2}{2}\right) dx + \frac{1}{4}\left([v]^2 + [w]^2\right) \le C\Delta x^{2k+1} + \int_0^1 \left(\frac{v^2}{2}\right) dx.$$

An integration in t plus the standard approximation theory on $v^e = \mathcal{S}U - U$ then gives the desired error estimate (2.18). □

We remark that the actual numerical computations in section 4 demonstrate that the order of accuracy is $k+1$ in both L^2 and L^∞ norms. It is not clear if the error estimate (2.18) is sharp. The trick to getting the extra half-order in the error estimate is through the choice of projections in (2.23), similar to $\mathcal{S}$ there but satisfying the point condition on the left boundary. This would eliminate the boundary terms in $\mathcal{III}$ but would cause problems in the control of volume integrals in $\mathcal{II}$. Notice that for first order linear hyperbolic equations, the optimal order of accuracy is $k+1$ in one dimension, and also in tensor product multidimensional cases [22], while the optimal order is $k + 1/2$ for general multidimensional cases [23].

3. The local discontinuous Galerkin method for the multiple dimensional case. In this section, we generalize the scheme discussed in the previous section to multiple space dimensions $x = (x_1, \ldots, x_d)$. We solve the following nonlinear problem:

(3.1)
$$U_t + \sum_{i=1}^{d} f_i(U)_{x_i} + \sum_{i=1}^{d} \left(r_i'(U) \sum_{j=1}^{d} g_{ij}(r_i(U)_{x_i})_{x_j} \right)_{x_i} = 0, \quad 0 \le x_i \le 1, \quad i = 1, \ldots, d,$$

with an initial condition

$$U(x,0) = U^0(x), \quad 0 \le x_i \le 1, \quad i = 1, \ldots, d, \tag{3.2}$$

and periodic boundary conditions. Here $f_i(U)$, $r_i(U)$, and $g_{ij}(Q)$ are arbitrary (smooth) nonlinear functions. Notice that the assumption of a box geometry and periodic boundary conditions is for simplicity only and is not essential: the method can be easily designed for arbitrary domain and for nonperiodic boundary conditions.

Let's denote a triangulation of the unit box by $\mathcal{T}_{\Delta x}$, consisting of nonoverlapping polyhedra completely covering the unit box. Hanging nodes are allowed. Here Δx measures the longest edge of all polyhedra in $\mathcal{T}_{\Delta x}$. We again denote the finite element space by

$$\mathcal{V}_{\Delta x}^d = \{v : \ v \text{ is a polynomial of degree at most } k \text{ for } x \in K \ \forall K \in \mathcal{T}_{\Delta x}\}. \tag{3.3}$$

Similar to the one-dimensional case, to define the local discontinuous Galerkin method we first introduce the new variables

$$Q_i = r_i(U)_{x_i}, \quad P_i = \sum_{j=1}^{d} g_{ij}(Q_i)_{x_j}, \quad i = 1, \ldots, d, \tag{3.4}$$

and rewrite (3.1) as a first order system:

$$\begin{aligned} &U_t + \sum_{i=1}^{d} (f_i(U) + r_i'(U)P_i)_{x_i} = 0, \\ &P_i - \sum_{j=1}^{d} g_{ij}(Q_i)_{x_j} = 0, \quad Q_i - r_i(U)_{x_i} = 0, \quad i = 1, \ldots, d. \end{aligned} \tag{3.5}$$

The local discontinuous Galerkin method is obtained by discretizing the above system with the discontinuous Galerkin method. This is achieved by multiplying the equations in (3.5) by test functions v, w_i, z_i for $i = 1, \ldots, d$, respectively, integrating over an element $K \in \mathcal{T}_{\Delta x}$, and integrating by parts. We again need to pay special attention to the boundary terms resulting from the procedure of integration by parts, as in the one-dimensional case. Thus we seek piecewise polynomial solutions $u, p_i, q_i \in V_{\Delta x}^d$, where $V_{\Delta x}^d$ is defined in (3.3), such that for all test functions $v, w_i, z_i \in V_{\Delta x}^d$ we have

$$\int_K u_t v dx - \sum_{i=1}^d \int_K (f_i(u) + r_i'(u)p_i) v_{x_i} dx + \int_{\partial K} \widehat{h_{n_K}} v^{int_K} ds = 0,$$

$$\int_K p_i w_i dx + \sum_{j=1}^d \int_K g_{ij}(q_i)(w_i)_{x_j} dx - \int_{\partial K} \widehat{g_{i,n_K}} w^{int_K} ds = 0, \qquad i = 1, \ldots, d, \tag{3.6}$$

$$\int_K q_i z_i dx + \int_K r_i(u)(z_i)_{x_i} dx - \int_{\partial K} \widehat{r_{i,n_K}} z^{int_K} ds = 0, \qquad i = 1, \ldots, d,$$

where ∂K is the boundary of element K, and the numerical fluxes (the "hats") are defined similar to the one-dimensional cases, namely,

$$\widehat{h_{n_K}} = \widehat{f_{n_K,K}}(u^{int_K}, u^{ext_K}) + \frac{\sum_{i=1}^d \left(r_i(u^{ext_K}) - r_i(u^{int_K})\right) p_i^+ n_{i,K}}{u^{ext_K} - u^{int_K}},$$

$$\widehat{g_{i,n_K}} = \widehat{g_{i,n_K,K}}(q^{int_K}, q^{ext_K}), \quad \widehat{r_{i,n_K}} = r_i(u^-)\, n_{i,K}. \tag{3.7}$$

Here $n_K = (n_{1,K}, \ldots, n_{d,K})$ is the outward unit normal for element K along the element boundary ∂K, u^{int_K} denotes the value of u evaluated from inside the element K, and u^{ext_K} denotes the value of u evaluated from outside the element K (inside the neighboring element). On the other hand, p^+ denotes the value of p evaluated from a predesignated "plus" side along an edge e, which is always the boundaries of two neighboring elements. For example, we could choose a fixed vector β, which is not parallel with any normals of element boundaries, and then designate the "plus" side to be the side at the end of the arrow of the normal n with $n \cdot \beta > 0$; see Figure 3.1. $\widehat{f_{n_K,K}}(u^{int_K}, u^{ext_K})$ is a monotone flux for $f_{n_K}(u) = \sum_{i=1}^d f_i(u) n_{i,K}$, namely, $\widehat{f_{n_K,K}}(u^{int_K}, u^{ext_K})$ is a Lipschitz continuous function in both arguments u^{int_K} and u^{ext_K}, is consistent with $f_{n_K}(u)$ in the sense that $\widehat{f_{n_K}}(u, u) = f_{n_K}(u)$, and is a nondecreasing function in u^{int_K} and a nonincreasing function in u^{ext_K}. Moreover, it is conservative (that is, there is only one flux at each edge shared by two elements, added to the residue for one and subtracted from the residue for another), namely,

$$\widehat{f_{n_K,K}}(a, b) = -\widehat{f_{n_{K'},K'}}(b, a),$$

where K and K' share the same edge where the flux is computed and hence $n_{K'} = -n_K$. Likewise, $-\widehat{g_{i,n_K,K}}(q_i^{int_K}, q_i^{ext_K})$ is a monotone flux for $-g_{i,n_K}(q_i) = -\sum_{j=1}^d g_{ij}(q) n_{j,K}$. Notice that we can again use the one-dimensional monotone fluxes as in the previous section. For example, we can use the simple Lax–Friedrichs flux

$$\widehat{f_{n_K,K}}(u^{int_K}, u^{ext_K}) = \frac{1}{2}\left(\sum_{i=1}^d \left(f_i(u^{int_K}) + f_i(u^{ext_K})\right) n_{i,K} - \alpha(u^{ext_K} - u^{int_K})\right), \tag{3.8}$$

$$\alpha = \max_u |f_{n_K}'(u)|,$$

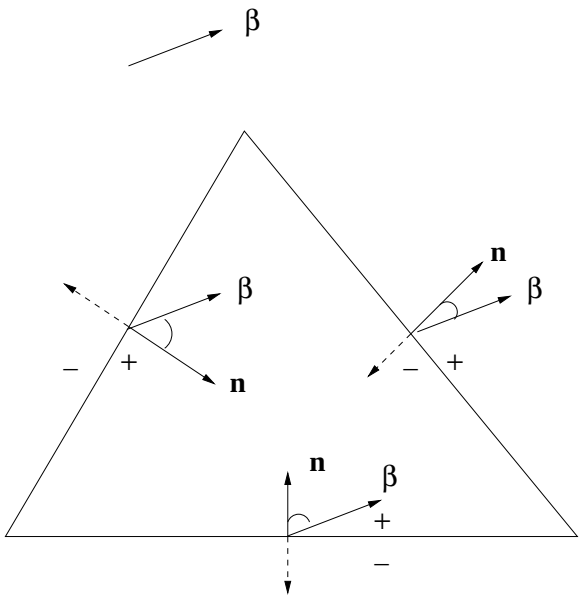


FIG. 3.1. *Illustration of the definition of "plus" and "minus" sides determined by a predetermined vector* β.

where the maximum is taken over a relevant range of u. The algorithm is now well defined.

Again, the algorithm (3.6)–(3.7) is very easy for numerical implementation. Given u, one first locally solves for the q_i, then locally solves for the p_i, and finally locally solves for the update of u. All the advantages listed for the method for the one-dimensional case are still valid in this multiple dimensional case.

We still have the following "cell entropy inequality" for the scheme (3.6)–(3.7). The proof follows along the same lines as that for the one-dimensional case, so we will omit it.

PROPOSITION 3.1 (cell entropy inequality). *There exist conservative numerical entropy fluxes* $\widehat{H_{n_K,K}}$ *such that the solution to the scheme* (3.6)–(3.7) *satisfies*

$$\frac{d}{dt}\int_K \left(\frac{u^2(x,t)}{2}\right) dx + \int_{\partial K} \widehat{H_{n_K,K}} ds \le 0. \tag{3.9}$$

The definition of the numerical entropy flux $\widehat{H_{n_K,K}}$ is similar to (2.11) for the one-dimensional case and is a bit longer in notation; thus we do not write it out. It is not important what the exact form of this numerical entropy flux is, as long as it is a "flux", i.e., shared by both sides of the edge and coming into the edge from one cell and out to the other, thus summing to zero when both cells are considered.

The L^2 stability of the method is then again a trivial corollary by summing up the cell entropy inequalities over K as follows.

COROLLARY 3.2 (L^2 stability). *The solution to the scheme* (3.6)–(3.7) *satisfies the* L^2 *stability*

$$\frac{d}{dt}\int_\Omega \left(\frac{u^2(x,t)}{2}\right) dx \le 0. \tag{3.10}$$

The same cell entropy inequality also holds for the implicit time discretizations as follows.

PROPOSITION 3.3 (implicit time discretization). *The cell entropy inequality and the L^2 stability also hold for the time discretization* (2.14) *with* $\frac{1}{2} \le \theta \le 1$ *for the scheme* (3.6)–(3.7). *That is,*

$$\int_K \left(\frac{(u^{n+1}(x))^2 - (u^n(x))^2}{2\Delta t} \right) dx + \int_{\partial K} \widehat{H^{n+\theta}_{n_K,K}} ds \le 0 \tag{3.11}$$

and

$$\int_\Omega (u^{n+1}(x))^2 dx \le \int_\Omega (u^n(x))^2 dx. \tag{3.12}$$

Unfortunately, we could not obtain an error estimate similar to the one-dimensional case because of the lack of a similar suitable projection $\mathcal{S}$. However, numerical examples in the next section verify that the accuracy holds as in the one-dimensional case. We remark that we could change the scheme to add an extra penalty term for the jumps of p (i.e., modifying the definition of the flux $\hat{r}$ in (2.6) by adding an additional term $-c[p]$, where c is a positive constant); then the error estimate proof would proceed in a straightforward way since there will be an additional $[p]^2$ term in the definition of Θ in (2.19). However, the resulting scheme would be much less attractive because it will lose the local solvability property for q and p. In order to get q and p one would then need to solve a global system, thus losing the main advantage of a local discontinuous Galerkin method. We do not give further details of this modified scheme here because it has little practical value, but interested readers can find similar issues addressed for the convection diffusion problems in [13].

4. Numerical examples. In this section we provide a few preliminary numerical examples to illustrate the accuracy and capability of the method. Attention has not been paid to efficiency in time discretizations, so an explicit third order Runge–Kutta method [26] is used, with small time steps so that spatial errors always dominate. We have also computed them with implicit time discretizations obtaining essentially the same results. Study of suitable implicit time discretizations which have efficient iterative solvers maintaining the local structure of the method is the subject of future work.

We would like to illustrate through these numerical examples the high order accuracy of the methods for both one-dimensional and two-dimensional linear and nonlinear problems. We would also like to illustrate the good behavior of the method for the so-called convection dominated cases, namely, the case where the coefficients of the third derivative terms are small.

Example 4.1. We compute the solution of the linear one-dimensional equation

$$U_t + U_{xxx} = 0 \tag{4.1}$$

with an initial condition $U(x,0) = \sin(x)$ and periodic boundary conditions (with 2π periodicity). The exact solution is given by $U(x,t) = \sin(x+t)$. Both uniform meshes and nonuniform meshes are used. The nonuniform meshes in this and later examples are a repeated pattern of $0.9\Delta x$ and $1.1\Delta x$ with an even number of elements. The L^2 and L^∞ errors, $||U-u||_{L^2}$ and $||U-u||_{L^\infty}$, and the numerical order of accuracy are contained in Table 4.1 for the uniform mesh case and in Table 4.2 for the nonuniform mesh case. We can clearly see that the method with P^k elements gives a uniform

TABLE 4.1

$U_t + U_{xxx} = 0$. $U(x,0) = \sin(x)$. Periodic boundary conditions. L^2 and L^∞ errors. Uniform meshes with N cells. Local discontinuous Galerkin methods with $k = 0, 1, 2, 3$. $t = 1$.

k		**N=10**	**N=20**		**N=40**		**N=80**	
		error	error	order	error	order	error	order
0	L^2	2.2534E-01	1.2042E-01	0.91	6.2185E-02	0.95	3.1582E-02	0.98
	L^∞	4.3137E-01	2.1977E-01	0.97	1.1082E-01	0.98	5.5376E-02	1.00
1	L^2	1.7150E-02	4.2865E-03	2.00	1.0716E-03	2.00	2.6792E-04	1.99
	L^∞	5.8467E-02	1.5757E-02	1.89	4.0487E-03	1.96	1.0210E-03	1.99
2	L^2	8.5803E-04	1.0823E-04	2.98	1.3559E-05	2.99	1.6958E-06	3.00
	L^∞	4.0673E-03	5.1029E-04	2.99	6.4490E-05	2.98	8.0722E-06	3.00
3	L^2	3.3463E-05	2.1035E-06	3.99	1.3166E-07	3.99	8.2365E-09	3.99
	L^∞	1.8185E-04	1.1157E-05	3.97	7.2362E-07	3.99	4.5593E-08	3.99

TABLE 4.2

$U_t + U_{xxx} = 0$. $U(x,0) = \sin(x)$. Periodic boundary conditions. L^2 and L^∞ errors. Non-uniform meshes with N cells. Local discontinuous Galerkin methods with $k = 0, 1, 2, 3$. $t = 1$.

k		**N=10**	**N=20**		**N=40**		**N=80**	
		error	error	order	error	order	error	order
0	L^2	2.2222E-01	1.2014E-01	0.88	6.2532E-02	0.94	3.1900E-02	0.97
	L^∞	4.3282E-01	2.2006E-01	0.97	1.1210E-01	0.97	5.8810E-02	0.93
1	L^2	2.0144E-02	5.2347E-03	1.94	1.3322E-03	1.97	3.3592E-04	1.98
	L^∞	8.8110E-02	2.3302E-02	1.93	5.9387E-03	1.97	1.4969E-03	1.98
2	L^2	9.8394E-04	1.1974E-04	3.03	1.4953E-05	3.00	1.8687E-06	3.00
	L^∞	5.2984E-03	6.8421E-04	2.95	8.5138E-05	3.00	1.0728E-05	2.99
3	L^2	7.3589E-05	4.6509E-06	3.98	2.9191E-06	3.99	2.0141E-08	3.86
	L^∞	3.4438E-04	2.2260E-05	3.95	1.3992E-06	3.99	9.1039E-08	3.94

$(k+1)$th order of accuracy in both norms for both the uniform and the nonuniform meshes.

Example 4.2. We compute the solution of the linear two-dimensional equation

$$U_t + U_{xxx} + U_{yyy} = 0 \tag{4.2}$$

with an initial condition $U(x,y,0) = \sin(x+y)$ and periodic boundary conditions (with 2π periodicity) in both directions. The exact solution is given by $U(x,y,t) = \sin(x+y+2t)$. Both uniform and nonuniform rectangular meshes are used. The nonuniform meshes are a repeated pattern of $0.9\Delta x$ and $1.1\Delta x$, in both directions, with an even number of edges in both directions. The L^2 and L^∞ errors and the numerical order of accuracy are contained in Table 4.3 for the uniform mesh case and in Table 4.4 for the nonuniform mesh case. We can clearly see again that the method with P^k elements gives a uniform $(k+1)$th order of accuracy for both the uniform and the nonuniform meshes.

Example 4.3. In order to see the accuracy of the method for nonlinear problems, we compute the classical soliton solution of the KdV equation

$$U_t - 3\left(U^2\right)_x + U_{xxx} = 0 \tag{4.3}$$

in $-10 \le x \le 12$. The initial condition is given by

$$U(x,0) = -2\,\mathrm{sech}^2(x).$$

TABLE 4.3

$U_t + U_{xxx} + U_{yyy} = 0$. $U(x, y, 0) = \sin(x + y)$. Periodic boundary conditions. L^2 and L^∞ errors. Uniform meshes with $N \times N$ cells. Local discontinuous Galerkin methods with $k = 0, 1, 2, 3$. $t = 1$.

k		**10×10**	**20×20**		**40×40**	
		error	error	order	error	order
0	L^2	3.5528E-01	2.0535E-01	0.79	1.1090E-01	0.89
	L^∞	7.1359E-01	4.0190E-01	0.82	2.1165E-01	0.92
1	L^2	3.3603E-02	9.0904E-03	1.89	2.4084E-03	1.92
	L^∞	2.2074E-01	6.1899E-02	1.83	1.5962E-02	1.95
2	L^2	3.8750E-03	4.8463E-04	2.99	6.0501E-05	3.00
	L^∞	3.9084E-02	4.8902E-03	2.99	6.1274E-04	2.99
3	L^2	4.1491E-04	2.6426E-05	3.97	1.6550E-06	3.99
	L^∞	4.2847E-03	2.8478E-04	3.91	1.7846E-05	3.99

TABLE 4.4

$U_t + U_{xxx} + U_{yyy} = 0$. $U(x, y, 0) = \sin(x + y)$. Periodic boundary conditions. L^2 and L^∞ errors. Nonuniform meshes with $N \times N$ cells. Local discontinuous Galerkin methods with $k = 0, 1, 2, 3$. $t = 1$.

k		**10×10**	**20×20**		**40×40**	
		error	error	order	error	order
0	L^2	3.5963E-01	2.0788E-01	0.79	1.1228E-01	0.88
	L^∞	7.3869E-01	4.0713E-01	0.85	2.1681E-01	0.91
1	L^2	3.4590E-02	9.1681E-03	1.92	2.3412E-03	1.97
	L^∞	2.5815E-01	7.2978E-02	1.82	1.8533E-02	1.97
2	L^2	4.0949E-03	5.1285E-04	2.99	6.4054E-05	3.00
	L^∞	5.0429E-02	6.3078E-03	2.99	8.0584E-04	2.97
3	L^2	4.5434E-04	2.8854E-05	3.97	1.8080E-06	3.99
	L^∞	6.0982E-03	4.0321E-04	3.92	2.5340E-05	3.99

The exact solution is

$$U(x, t) = -2\,\mathrm{sech}^2(x - 4t).$$

Table 4.5 (uniform mesh) and Table 4.6 (nonuniform mesh) give the errors of numerical solution at $t = 0.5$ using the boundary condition

$$U(-10, t) = g_1(t), \quad U_x(12, t) = g_2(t), \quad U_{xx}(12, t) = g_3(t), \tag{4.4}$$

where $g_i(t)$ corresponds to the data from the exact solution. Notice that the local discontinuous Galerkin method allows for an easy implementation of such boundary conditions. We can see from these tables that the orders of accuracy are comparable to that for the linear case.

Example 4.4. In order to see the accuracy of the method for nonlinear problems with small coefficient for the third derivative term, we compute the soliton solution of the generalized KdV equation [5]

$$U_t + U_x + \left(\frac{U^4}{4}\right)_x + \epsilon U_{xxx} = 0 \tag{4.5}$$

in $-2 \le x \le 3$, where we take $\epsilon = 0.2058 \times 10^{-4}$. The initial condition is given by

$$U(x, 0) = A\,\mathrm{sech}^{\frac{2}{3}}\left(K(x - x_0)\right) \tag{4.6}$$

TABLE 4.5

The KdV equation $U_t - 3(U^2)_x + U_{xxx} = 0$. $U(x,0) = -2\ \mathrm{sech}^2(x)$. *Boundary condition* (4.4). L^2 *and* L^∞ *errors. Uniform meshes with* N *cells. Local discontinuous Galerkin methods with* $k = 0, 1, 2, 3$. $t = 0.5$.

k		N=40	N=80		N=160		N=320	
		error	error	order	error	order	error	order
0	L^2	2.5292E-01	1.9098E-01	0.40	1.3019E-01	0.55	7.9780E-02	0.71
	L^∞	9.0170E-01	6.8651E-01	0.39	4.6405E-01	0.56	2.8531E-01	0.70
1	L^2	2.6512E-02	4.6652E-03	2.50	1.0108E-03	2.20	2.5906E-04	1.96
	L^∞	1.4748E-01	3.4625E-02	2.09	1.1840E-02	1.55	3.3239E-03	1.83
2	L^2	1.5317E-03	1.8083E-04	3.08	2.2642E-05	2.99	2.8335E-06	2.99
	L^∞	1.7486E-02	2.7505E-03	2.66	3.5575E-04	2.95	4.4397E-05	3.00
3	L^2	2.0631E-04	1.3981E-05	3.88	8.9054E-07	3.97	5.6029E-08	3.99
	L^∞	2.0155E-03	2.1462E-04	3.23	1.4461E-05	3.89	9.1140E-07	3.98

TABLE 4.6

The KdV equation $U_t - 3(U^2)_x + U_{xxx} = 0$. $U(x,0) = -2\ \mathrm{sech}^2(x)$. *Boundary condition* (4.4). L^2 *and* L^∞ *errors. Nonuniform meshes with* N *cells. Local discontinuous Galerkin methods with* $k = 0, 1, 2, 3$. $t = 0.5$.

k		N=40	N=80		N=160		N=320	
		error	error	order	error	order	error	order
0	L^2	2.4530E-01	1.9004E-01	0.37	1.3390E-01	0.50	8.4635E-02	0.66
	L^∞	1.0172E+00	7.6826E-01	0.40	5.3383E-01	0.52	3.3655E-01	0.66
1	L^2	2.7042E-02	4.9065E-03	2.46	1.0555E-03	2.21	2.6978E-04	1.97
	L^∞	1.4490E-01	4.1570E-02	1.80	1.3925E-02	1.57	3.9129E-03	1.83
2	L^2	1.9493E-03	2.0134E-04	3.27	2.4926E-05	3.01	3.1208E-06	2.99
	L^∞	2.2876E-02	3.5163E-03	2.70	4.7161E-04	2.89	5.9033E-05	2.99
3	L^2	3.0402E-04	1.5462E-05	4.29	1.0064E-06	3.94	6.3370E-08	3.99
	L^∞	2.7735E-03	2.1464E-04	3.69	1.8358E-05	3.55	1.3119E-06	3.80

TABLE 4.7

The GKdV equation (4.5) *with initial condition* (4.6) *and boundary condition* (4.7). L^2 *and* L^∞ *errors. Nonuniform meshes with* N *cells. Local discontinuous Galerkin methods with* $k = 0, 1, 2, 3$. $t = 1$.

k		N=160	N=320		N=640		N=1280	
		error	error	order	error	order	error	order
0	L^2	1.6566E-02	1.1259E-02	0.56	7.0817E-03	0.67	4.1526E-03	0.77
	L^∞	9.3056E-02	6.6829E-02	0.48	4.4502E-02	0.58	2.7539E-02	0.69
1	L^2	3.8554E-04	6.0675E-05	2.66	1.1784E-05	2.36	2.8635E-06	2.04
	L^∞	3.2635E-03	6.2508E-04	2.38	2.2689E-04	1.47	6.4595E-05	1.81
2	L^2	8.2907E-06	9.5348E-07	3.12	1.1895E-07	3.00	1.5290E-08	2.96
	L^∞	1.6684E-04	2.2545E-05	2.88	3.0858E-06	2.87	3.9503E-07	2.97
3	L^2	1.7005E-06	1.3664E-07	3.63	3.0527E-09	5.48	1.9206E-10	3.99
	L^∞	1.7607E-05	1.3291E-06	3.72	8.3962E-08	3.98	5.2861E-09	3.99

with $A = 0.2275$, $x_0 = 0.5$, and $K = 3\left(\frac{A^3}{40\epsilon}\right)^{\frac{1}{2}}$. The exact solution is

$$U(x,t) = A\,\mathrm{sech}^{\frac{2}{3}}\left(K(x - x_0) - \omega t\right),$$

where $\omega = K\left(1 + \frac{A^3}{10}\right)$. We compute the solution using the boundary condition

$$U(-2,t) = g_1(t), \quad U_x(3,t) = g_2(t), \quad U_{xx}(3,t) = g_3(t) \tag{4.7}$$

with a nonuniform mesh. The result is contained in Table 4.7. The scheme clearly

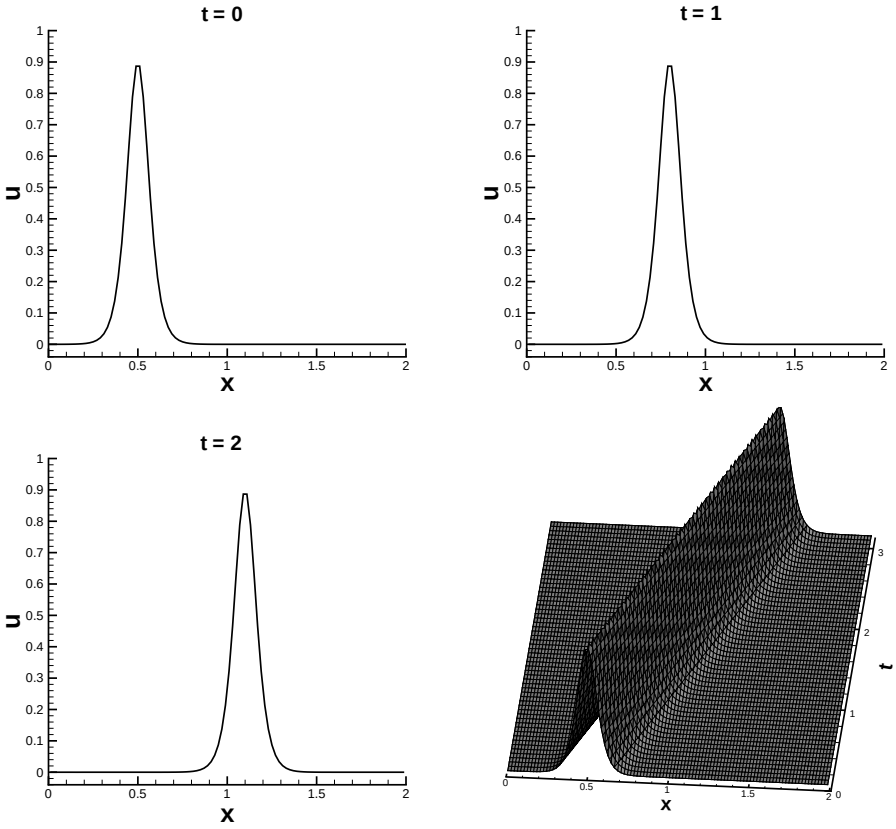


FIG. 4.1. *Single soliton profiles. Solutions of* (4.8) *with initial condition* (4.9) *and periodic boundary conditions in* $[0, 2]$ *using* P^2 *elements with* 100 *cells. Top left: solution at* $t = 0$; *top right:* $t = 1$; *bottom left:* $t = 2$; *bottom right: space time graph of the solution up to* $t = 3$.

demonstrates an order of accuracy of $k + 1$ for P^k elements in both the L^2 and L^∞ norms for this problem with strong nonlinearity in the first derivative term, small dispersive term, and nonperiodic boundary conditions.

Example 4.5. In this example we compute the classical soliton solutions of the KdV equation

$$U_t + \left(\frac{U^2}{2}\right)_x + \epsilon U_{xxx} = 0. \tag{4.8}$$

The examples are those used in [15].

The single soliton case has the initial condition

$$U_0(x) = 3c\,\mathrm{sech}^2\left(k(x - x_0)\right) \tag{4.9}$$

with $c = 0.3$, $x_0 = 0.5$, $k = (1/2)\sqrt{c/\epsilon}$, and $\epsilon = 5 \times 10^{-4}$. The solution is computed in $x \in [0, 2]$ with periodic boundary conditions, using P^2 elements with 100 cells, and is shown in Figure 4.1.

The double soliton collision case has the initial condition

$$U_0(x) = 3c_1\,\mathrm{sech}^2\left(k_1(x - x_1)\right) + 3c_2\,\mathrm{sech}^2\left(k_2(x - x_2)\right) \tag{4.10}$$

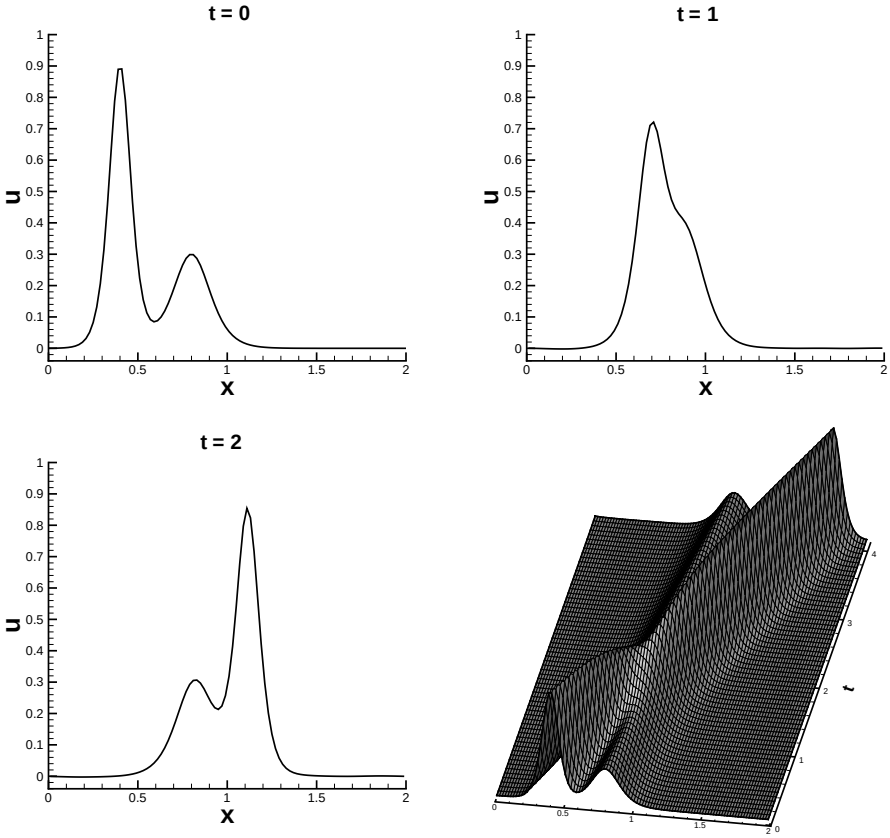


FIG. 4.2. *Double soliton collision profiles. Solutions of* (4.8) *with initial condition* (4.10) *and periodic boundary conditions in* $[0,2]$ *using* P^2 *elements with* 100 *cells. Top left: solution at* $t=0$; *top right:* $t=1$; *bottom left:* $t=2$; *bottom right: space time graph of the solution up to* $t=4$.

with $c_1 = 0.3$, $c_2 = 0.1$, $x_1 = 0.4$, $x_2 = 0.8$, $k_i = (1/2)\sqrt{c_i/\epsilon}$ for $i = 1, 2$, and $\epsilon = 4.84 \times 10^{-4}$. The solution is computed in $x \in [0, 2]$ with periodic boundary conditions, using P^2 elements with 100 cells, and is shown in Figure 4.2.

The triple soliton splitting case has the initial condition

$$U_0(x) = \frac{2}{3}\operatorname{sech}^2\left(\frac{x-1}{\sqrt{108\epsilon}}\right) \tag{4.11}$$

with $\epsilon = 10^{-4}$. The solution is computed in $x \in [0, 3]$ with periodic boundary conditions and is shown in Figure 4.3.

Example 4.6. We compute in this example the KdV zero dispersion limit of conservation laws. The equation is (4.8) with an initial condition

$$U(x, 0) = 2 + 0.5\sin(2\pi x) \tag{4.12}$$

for $x \in [0, 1]$ with periodic boundary conditions, and we are interested in the limit when $\epsilon \to 0^+$. Theoretical and numerical discussions about this limit can be found in [21] and [27]. Here we are concerned mainly with the capability of our numerical method in resolving the small scale solution structures in this limit when ϵ is small.

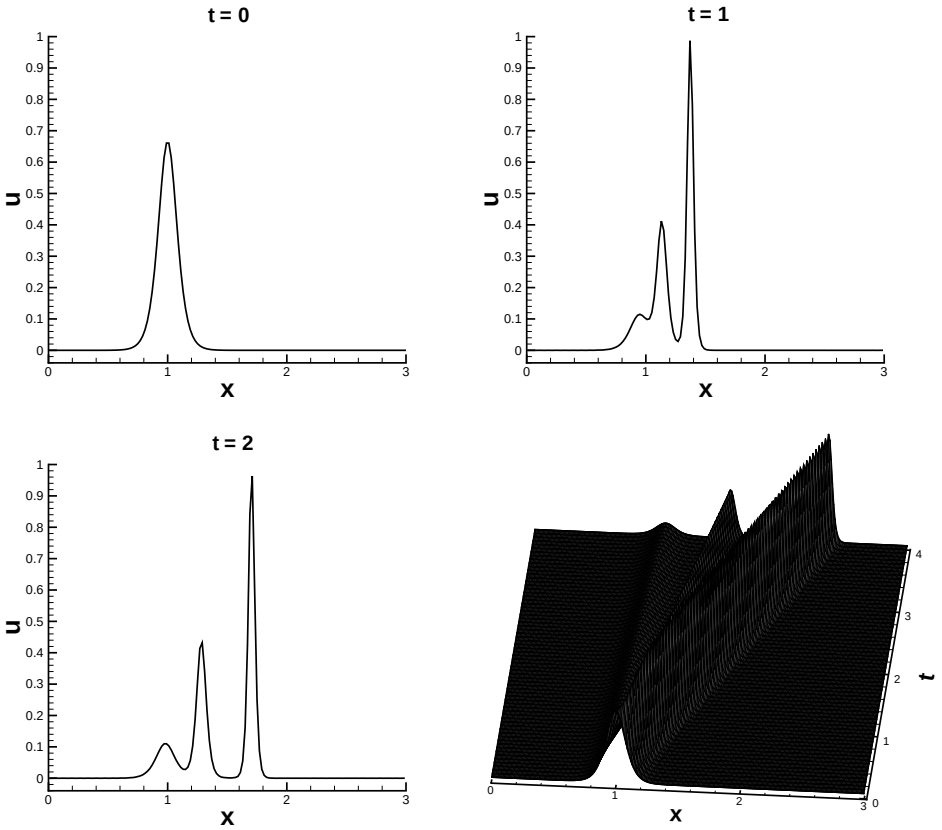


FIG. 4.3. *Triple soliton splitting profiles. Solutions of* (4.8) *with initial condition* (4.11) *and periodic boundary conditions in* $[0, 3]$ *using* P^2 *elements with* 150 *cells. Top left: solution at* $t = 0$; *top right:* $t = 1$; *bottom left:* $t = 2$; *bottom right: space time graph of the solution up to* $t = 4$.

For this purpose we compute the solution to $t = 0.5$ with $\epsilon = 10^{-4}, 10^{-5}, 10^{-6}$, and 10^{-7} using P^2 elements with 300 cells for the first two cases, 800 cells for the third case, and 1700 cells for the last case. We have verified that these are "converged" solutions in the sense that further increasing the number of cells does not change the solutions graphically. These solutions are shown in Figure 4.4. Notice the physical "oscillations" which are typical in such dispersive limits; see, e.g., [21]. Clearly our method is very suitable for computing such solutions.

5. Concluding remarks. We have designed a class of local discontinuous Galerkin methods for solving KdV type equations containing third derivatives and have proven their stability for any space dimensions for a general class of nonlinear equations. Numerical examples are shown to illustrate the accuracy and capability of the methods, especially for the convection dominated cases where the coefficients of the third derivative terms are small. Efficient implicit time discretizations which have efficient iterative solvers maintaining the local structure of the method, an accuracy enhancement study, and more numerical experiments with physically interesting problems constitute an ongoing project.

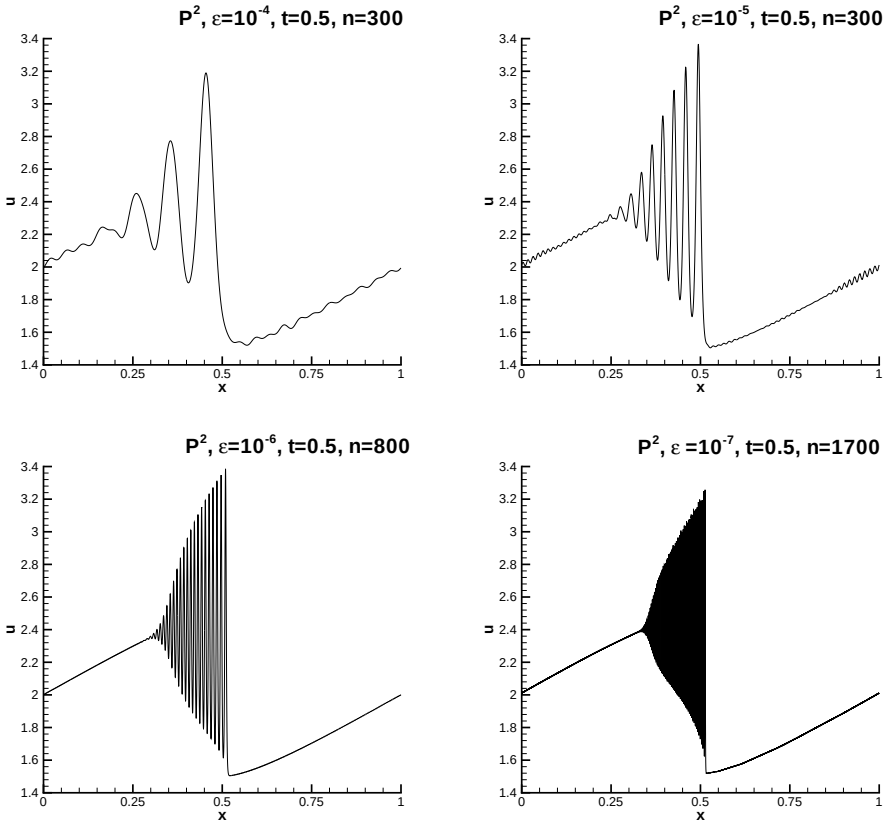


FIG. 4.4. *Zero dispersion limit of conservation laws. Solutions of* (4.8) *with initial condition* (4.12) *and periodic boundary conditions in* $[0,1]$ *using* P^2 *elements at* $t = 0.5$. *Top left:* $\epsilon = 10^{-4}$ *with* 300 *cells; top right:* $\epsilon = 10^{-5}$ *with* 300 *cells; bottom left:* $\epsilon = 10^{-6}$ *with* 800 *cells; bottom right:* $\epsilon = 10^{-7}$ *with* 1700 *cells.*

Acknowledgments. We would like to thank Bernardo Cockburn for his valuable help in the discussion about the projection $\mathcal{S}$ and thank Andrew Majda for pointing out reference [21] and test cases of zero dispersive limits of conservation laws in Example 4.6.

REFERENCES

[1] T. B. BENJAMIN, J. L. BONA, AND J. J. MAHONY, *Model equations for long waves in nonlinear, dispersive systems*, Phil. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[2] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[3] R. BISWAS, K. D. DEVINE, AND J. FLAHERTY, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math., 14 (1994), pp. 255–283.

[4] J. L. BONA, V. A. DOUGALIS, AND O. A. KARAKASHIAN, *Fully discrete Galerkin methods for the Korteweg-de Vries equation*, Comput. Math. Appl. Ser. A, 12 (1986), pp. 859–884.

[5] J. L. BONA, V. A. DOUGALIS, O. A. KARAKASHIAN, AND W. R. MCKINNEY, *Conservative, high-order numerical schemes for the generalized Korteweg-de Vries equation*, Phil. Trans. Roy. Soc. London Ser. A, 351 (1995), pp. 107–164.

[6] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab, *Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems*, Math. Comp., 71 (2002), pp. 455–478.

[7] P. Ciarlet, *The Finite Element Method For Elliptic Problems*, North-Holland, Amsterdam, 1975.

[8] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[9] B. Cockburn, G. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lecture Notes in Comput. Sci. Engrg. 11, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 3–50.

[10] B. Cockburn, S.-Y. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[11] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[12] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[13] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[14] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[15] A. Debussche and J. Printems, *Numerical simulation of the stochastic Korteweg-de Vries equation*, Phys. D, 134 (1999), pp. 200–226.

[16] C. L. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.

[17] J. Jaffré, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.

[18] G.-S. Jiang and C.-W. Shu, *On cell entropy inequality for discontinuous Galerkin methods*, Math. Comp., 62 (1994), pp. 531–538.

[19] O. A. Karakashian and W. McKinney, *On the approximation of solutions of the generalized Korteweg-de Vries-Burgers equation. Solitons, nonlinear wave equations and computation*, Math. Comput. Simulation, 37 (1994), pp. 405–416.

[20] D. J. Korteweg and G. de Vries, *On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves*, Philosophical Magazine, 39 (1895), pp. 422–443.

[21] P. D. Lax, C. D. Levermore, and S. Venakides, *The generation and propagation of oscillations in dispersive initial value problems and their limiting behavior*, in Important Developments in Soliton Theory, Springer Ser. Nonlinear Dynam., A. S. Fokas and V. E. Zakharov, eds., Springer-Verlag, Berlin, 1993, pp. 205–241.

[22] P. LeSaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–145.

[23] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[24] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[25] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comp., 49 (1987), pp.105–121.

[26] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[27] S. Venakides, *The zero dispersion limit of the Korteweg-de Vries equation with periodic initial data*, Trans. Amer. Math. Soc., 301 (1987), pp. 189–226.

SIAM J. NUMER. ANAL.
Vol. 40, No. 2, pp. 792–804

A CONVERGENCE ANALYSIS OF DYKSTRA'S ALGORITHM FOR POLYHEDRAL SETS*

CHRIS PERKINS†

Abstract. Let H be a nonempty closed convex set in a Hilbert space X determined by the intersection of a finite number of closed half spaces. It is well known that given an $x_0 \in X$, Dykstra's algorithm applied to this collection of closed half spaces generates a sequence of iterates that converge to $P_H(x_0)$, the orthogonal projection of x_0 onto H. The iterates, however, do not necessarily lie in H. We propose a combined Dykstra–conjugate-gradient method such that, given an $\varepsilon > 0$, the algorithm computes an $x \in H$ with $\|x - P_H(x_0)\| < \varepsilon$. Moreover, for each iterate x_m of Dykstra's algorithm we calculate a bound for $\|x_m - P_H(x_0)\|$ that approaches zero as m tends to infinity. Applications are made to computing bounds for $\|x_m - P_H(x_0)\|$ where H is a polyhedral cone. Numerical results are presented from a sample isotone regression problem.

Key words. Dykstra's algorithm, best approximation from polyhedra, Hildreth's algorithm, linear inequalities

AMS subject classifications. Primary, 41A65; Secondary, 65F10, 15A39

PII. S0036142900367557

1. Introduction.

Let X be a Hilbert space and $K_0, K_1, \ldots, K_{n-1}$ be n closed convex sets in X whose intersection, K, is nonempty. Given an $x_0 \in X$, the Boyle–Dykstra algorithm [3] generates a sequence of iterates $\{x_m\}$ whose limit is the orthogonal projection of x_0 onto K. Defining P_{K_i} to be the orthogonal projection onto K_i, letting $[m]$ denote m mod n, and setting

$$e_{-n} = e_{-(n-1)} = \cdots = e_{-1} = \vec{0},$$

Dykstra's algorithm can be written as

$$\begin{aligned} x_{m+1} &= P_{K_{[m]}}(x_m + e_{m-n}), \\ e_m &= (x_m + e_{m-n}) - x_{m+1} \end{aligned}$$

for $m = 0, 1, \ldots$. If the n closed convex sets are half spaces of the form $H_i = \{x \mid \langle x, z_i \rangle \leq f_i\}$, where $z_i \in X$ and $f_i \in \Re$, then the formula for the $(m+1)$st iterate of Dykstra's algorithm can be written as $x_{m+1} = x_m + \xi_m z_{[m]}$, where $\xi_m \in \Re$.

For this finite collection of intersecting closed half spaces, Iusem and De Pierro [11] showed that the Dykstra algorithm converged linearly. Subsequently Deutsch and Hundal [6] have sharpened the rate of convergence bound established in [11] and have shown that

$$\|x_m - P_H(x_0)\| \leq \rho c^m, \tag{1.1}$$

where $H = \cap_{i=0}^{n-1} H_i$, ρ is a constant, $0 \leq c < 1$, and $m = 0, 1, \ldots$. In general, the use of the Deutsch–Hundal result to explicitly bound $\|x_m - P_H(x_0)\|$ is problematic in that although c can be calculated a priori, ρ cannot be. This means that in practice when computing x_m, inequality (1.1) cannot be used to calculate a bound for $\|x_m - P_H(x_0)\|$

*Received by the editors January 10, 2000; accepted for publication (in revised form) October 22, 2001; published electronically July 24, 2002.
http://www.siam.org/journals/sinum/40-2/36755.html
†13656 Spring Mill Boulevard, Carmel, IN 46032 (christopher.perkins@roche.com).

for $m = 0, 1, \ldots$. Xu comments in [15] that the applicability of Dykstra's algorithm for polyhedron approximation is restrictive unless it is possible to find an active bound for ρ. This lack of a computable error bound for the Dykstra iterates is the motivation for this manuscript. Here we develop a method to explicitly bound $\|x_m - P_H(x_0)\|$ by a ρ_m, where $\lim_{m\to\infty} \rho_m = 0$. This convergence analysis allows the algorithm to be terminated when x_m is within a prescribed distance of $P_H(x_0)$.

In section 2, we first review the equivalence between Dykstra's algorithm for polyhedral sets in the general Hilbert space setting and the finite dimensional Hildreth algorithm [10]. This equivalence is critical to our computation of the error bounds for the iterates. The finite dimensional version allows us to use a linear algebraic analysis to determine a ρ_m such that $\|x_m - P_H(x_0)\| \le \rho_m$ with $\lim_{m\to\infty} \rho_m = 0$. Given that it is possible to determine when an x_m is arbitrarily close to $P_H(x_0)$, we are then able to calculate an $x \in H$ such that $\|x - P_H(x_0)\| < \varepsilon$.

Section 3 shows how the previously established results can be modified if $\{z_i\}$ is linearly independent. Other implementation details are discussed and the N-convex regression problems are reviewed. Numerical results are presented from a sample isotone regression problem.

There are other methods available for computing $P_H(x_0)$. These include solving a related least distance quadratic programming problem [7]. More recently, Xu has published an algorithm with an error analysis that also computes the nearest point mapping for polyhedrons in a finite number of iterations [15]. Unfortunately, the use of Xu's algorithm requires a priori knowledge of a point in H. Such a point can first be calculated using the algorithm proposed by T. S. Motzkin [14] and discussed in Agmon [1].

2. Definitions, lemmas, and theory. In 1988 Han [9] rediscovered Hildreth's algorithm [10] and commented that the arguments used to show convergence in Euclidean space can be used in the general Hilbert space setting where the intersecting closed convex sets are half spaces. Deutsch and Hundal [6] later reported that in Euclidean space, where the intersection of closed convex sets in polyhedral, Dykstra's algorithm reduces to Hildreth's algorithm.

In fact, the Deutsch–Hundal statement can be strengthened, and Han's result is immediately applicable to the general Hilbert space setting. This is due to the fact that in Hilbert space, Dykstra's algorithm applied to a finite set of intersecting closed half spaces is equivalent to Hildreth's algorithm applied to a related set of linear inequalities. We next review this equivalence.

Let X be a Hilbert space and $H_0, H_1, \ldots, H_{n-1}$ be closed half spaces of the form $H_i = \{x \mid \langle x, z_i\rangle \le f_i\}$ for $0 \le i \le n-1$, where $z_i \in X$, $\|z_i\| = 1$, $f_i \in \Re$, and $H = \cap_{i=0}^{n-1} H_i$ is nonempty. The boundary of each closed half space H_i is the hyperplane B_i, where $B_i = \{x \mid \langle x, z_i\rangle = f_i\}$.

LEMMA 2.1. *Given $x \in X$,*

$$P_{H_i}(x) = \begin{cases} x & \text{if } x \in H_i, \\ x - (\langle x, z_i\rangle - f_i) z_i & \text{if } x \notin H_i, \end{cases}$$

$$P_{B_i}(x) = x - (\langle x, z_i\rangle - f_i) z_i,$$

$$\mathrm{d}(x, B_i) = |\langle x, z_i\rangle - f_i|$$

for $i = 0, 1, \ldots, n-1$. □

LEMMA 2.2. *Given an $x_0 \in X$, the iterates of Dykstra's algorithm lie in* $\operatorname{span}\{x_0, z_0, \ldots, z_{n-1}\}$.

Proof. This follows from Dykstra's algorithm, the fact that the closed convex sets are half spaces, Lemma 2.1, and a simple inductive argument. □

By Lemma 2.2, the Dykstra iterates $\{x_m\}$ are contained in a finite dimensional subspace of X. Thus we can use Gram–Schmidt to generate an orthonormal basis $\{w_j\}_{j=1}^{k}$ for $\text{span}\{x_0, z_0, \ldots, z_{n-1}\}$, and we can write

$$x_0 = \sum_{j=1}^{k} \gamma_j w_j \tag{2.1}$$

as well as

$$z_i = \sum_{j=1}^{k} \alpha_{ij} w_j \tag{2.2}$$

for $0 \leq i \leq n-1$.

Since any finite dimensional Hilbert space is isometric to $l_2(k)$, we reduce the convergence analysis of Dykstra's algorithm in X to an equivalent analysis in $\Re^k$.

Defining

$$y_0 = (\gamma_1, \ldots, \gamma_k)^{\mathrm{T}} \tag{2.3}$$

and

$$a_i = (\alpha_{i1}, \ldots, \alpha_{ik})^{\mathrm{T}} \tag{2.4}$$

for $0 \leq i \leq n-1$, it is possible to define the n closed half spaces in $\Re^k$ by

$$h_i = \{y \mid \langle y, a_i \rangle \leq f_i\} \tag{2.5}$$

and the corresponding boundaries by

$$b_i = \{y \mid \langle y, a_i \rangle = f_i\}. \tag{2.6}$$

We remark that Dykstra's algorithm applied to $h_0, h_1, \ldots, h_{n-1}$ is the same as Hildreth's algorithm applied to $Ay \leq f$, where $A^{\mathrm{T}} = [a_0, \ldots, a_{n-1}]$ and $f = (f_0, \ldots, f_{n-1})^{\mathrm{T}}$.

Throughout what follows, x_0 will be defined as in (2.1), $\{x_m\}$ will denote the sequence of Dykstra iterates generated by the closed half spaces $H_0, \ldots, H_{n-1}$ in X. Similarly, y_0 will be defined as in (2.3) and $\{y_m\}$ will denote the sequence of Dykstra iterates generated by the closed half spaces $h_0, \ldots, h_{n-1}$ in $\Re^k$ as specified in (2.5).

Let $S \subset \{0, 1, \ldots, n-1\}$. We define $\tilde{S} = \{i \mid 0 \leq i \leq n-1 \text{ and } i \notin S\}$ and $|S|$ to be the cardinality of set S. We introduce the notation H_S (resp., h_S) to denote $\cap_{i \in S} H_i$ (resp., $\cap_{i \in S} h_i$) and B_S (resp., b_S) to denote $\cap_{i \in S} B_i$ (resp., $\cap_{i \in S} b_i$). For notational simplicity, if $S = \{0, 1, \ldots, n-1\}$, the subscript S is omitted. A_S^{T} and f_S are defined to be

$$A_S^{\mathrm{T}} = [a_{s_1}, \ldots, a_{s_{|S|}}\}, \tag{2.7}$$

$$f_S = (f_{s_1}, \ldots, f_{s_{|S|}})^{\mathrm{T}}, \tag{2.8}$$

where $s_j \in S$, $1 \leq j \leq |S|$ with $s_j < s_{j+1}$. Let $C \subset \{0, 1, \ldots, n-1\}$ contain the indices of the critical boundaries. That is, $C = \{i \mid P_h(y_0) \in b_i\}$. If $C = \emptyset$, then $P_h(y_0) \in \text{int}(h)$ and $P_h(y_0) = y_0$. Henceforth we will assume that $C \neq \emptyset$.

Let $S \subset \{0, 1, \ldots, n-1\}$, $S \neq \emptyset$, and let A_S^{T} and f_S be as defined in (2.7) and (2.8). We denote the residual $r_{m,S} = f_S - A_S y_m$. If $f_S \in \mathrm{Range}(A_S)$ (equivalently $b_S \neq \emptyset$), then by (2.6),

$$\begin{aligned} r_{m,S} &= A_S P_{b_S}(y_m) - A_S y_m, \\ &= A_S(P_{b_S}(y_m) - y_m). \end{aligned}$$

A_S^+ is the Moore–Penrose generalized inverse of A_S, and $A_S^+ A_S$ is the projection onto the orthogonal complement of the null space of A_S (denoted $\mathrm{N}(A_S)^\perp$) [8]. If $b_S \neq \emptyset$, then $P_{b_S}(y_m) - y_m$ is an element of $\mathrm{N}(A_S)^\perp$,

$$A_S^+ A_S(P_{b_S}(y_m) - y_m) = P_{b_S}(y_m) - y_m,$$

and $A_S^\perp r_{m,S} = P_{b_S}(y_m) - y_m$. Therefore,

$$\|y_m - P_{b_S}(y_m)\| = \|A_S^+ r_{m,S}\| \leq \|A_S^+\| \, \|r_{m,S}\| \tag{2.9}$$

with $\|A_S^+\|$ being the inverse of the smallest nonzero singular value of A_S [8].

For each iterate y_m of Dykstra's algorithm we determine the existence of a set C_m with the properties

$$f_{C_m} \in \mathrm{Range}(A_{C_m}), \tag{2.10a}$$

$$\text{for all } j \in \tilde{C}_m, \quad 2\|A_{C_m}^+\| \, \|r_{m,C_m}\| < f_j - \langle y_m, a_j \rangle, \tag{2.10b}$$

$$\text{for all } j \in \tilde{C}_m, \quad e_{\pi_{m,j}} = \vec{0},$$

$$\text{where } \pi_{m,j} \in \{i \mid m - n \leq i \leq m-1 \text{ and } [i] = j\}. \tag{2.10c}$$

If no such C_m exists, we define the bound, ρ_m, on $\|y_m - P_h(y_0)\|$ to be ∞. In Theorem 2.16 we prove that for sufficiently large m, the set C satisfies properties (2.10a)–(2.10c).

For notational simplicity throughout what follows, we will use the notation r_m in place of r_{m,C_m} or $r_{m,C}$ where appropriate. We next assume that $C_m \neq \emptyset$ and develop the theory to bound $\|y_m - P_h(y_0)\|$ by ρ_m, where

$$\rho_m = 2\|A_{C_m}^+\| \, \|r_m\|. \tag{2.11}$$

Thereafter, in Theorem 2.17 we will show sufficient conditions on the selection of $\{C_m\}$ to guarantee the bound $\rho_m = 2\|A_{C_m}^+\| \, \|r_m\|$ tends to zero.

By the construction of C_m and inequality (2.9),

$$2\|y_m - P_{b_{C_m}}(y_m)\| \leq 2\|A_{C_m}^+\| \, \|r_m\| \tag{2.12}$$

$$< f_j - \langle y_m, a_j \rangle \tag{2.13}$$

$$= \mathrm{d}(y_m, b_j)$$

for all $j \in \tilde{C}_m$.

The following lemma shows that if $C_m \neq \{0, 1, \ldots, n-1\}$, then the closed ball centered at $P_{b_{C_m}}(y_m)$ with radius $\mathrm{d}(y_m, P_{b_{C_m}}(y_m))$ is contained in the $\mathrm{int}(h_{\tilde{C}_m})$. To facilitate subsequent discussion, we define

$$\mathcal{B}_m = \mathcal{B}[P_{b_{C_m}}(y_m), \mathrm{d}(y_m, P_{B_{C_m}}(y_m))]. \tag{2.14}$$

LEMMA 2.3. *If $C_m \neq \{0, 1, \ldots, n-1\}$, then the closed ball $\mathcal{B}_m \subset \text{int}(h_{\tilde{C}_m})$.*

Proof. We will show that, given any element y contained in the closed ball $\mathcal{B}_m$, $\langle y, a_j \rangle < f_j$ for every $j \in \tilde{C}_m$. This will imply that $y \in \text{int}(h_{\tilde{C}_m})$. Let $y \in \mathcal{B}_m$ and define $\varepsilon_1 = \text{d}(y_m, P_{b_{C_m}}(y_m))$. Then $P_{b_{C_m}}(y_m) = y_m + \varepsilon_1 v_1$ for some unit vector v_1. Since $y \in \mathcal{B}_m$, $y = P_{b_{C_m}}(y_m) + \varepsilon_2 v_2$ for $0 \leq \varepsilon_2 \leq \text{d}(y_m, P_{b_{C_m}}(y_m)) = \varepsilon_1$ and some unit vector v_2. Let $j \in \tilde{C}_m$. Then

$$\begin{aligned}
\langle y, a_j \rangle &= \langle P_{b_{C_m}}(y_m) + \varepsilon_2 v_2, a_j \rangle \\
&= \langle y_m + \varepsilon_1 v_1 + \varepsilon_2 v_2, a_j \rangle \\
&= \langle y_m, a_j \rangle + \varepsilon_1 \langle v_1, a_j \rangle + \varepsilon_2 \langle v_2, a_j \rangle \\
&\leq \langle y_m, a_j \rangle + \varepsilon_1 \|v_1\| \, \|a_j\| + \varepsilon_2 \|v_2\| \, \|a_j\| \\
&= \langle y_m, a_j \rangle + \varepsilon_1 + \varepsilon_2 \\
&\leq \langle y_m, a_j \rangle + 2\varepsilon_1 \\
&= \langle y_m, a_j \rangle + 2\text{d}(y_m, P_{b_{C_m}}(y_m)) \\
&= \langle y_m, a_j \rangle + 2\|y_m - P_{b_{C_m}}(y_m)\| \\
&< \langle y_m, a_j \rangle + f_j - \langle y_m, a_j \rangle \quad \text{by (2.13)} \\
&= f_j.
\end{aligned}$$

Thus $y \in \text{int}(h_j)$. Since j was arbitrary in $\tilde{C}_m$, $y \in \text{int}(h_{\tilde{C}_m})$, and $\mathcal{B}_m \subset \text{int}(h_{\tilde{C}_m})$. ∎

In order to show the bound stated in (2.11), we will prove that all subsequent Dykstra iterates are elements of the closed ball $\mathcal{B}_m$. We will then have $\{y_{m+k}\}_{k=0}^{\infty} \subset \mathcal{B}_m$. $\mathcal{B}_m$, by construction, is closed and must contain its limit points. Since $\{y_{m+k}\}_{k=0}^{\infty}$ converges to $P_h(y_0)$, $P_h(y_0)$ would then be contained in $\mathcal{B}_m$. This will allow us to compute a bound on $\|y_m - P_h(y_0)\|$.

The following result is proved in [6]. As in section 1, $[m]$ denotes m mod n.

RESULT 2.4. *If y_m is the mth iterate of Dykstra's algorithm, then $y_{m+1} = (1 - \lambda)y_m + \lambda P_{b_{[m]}}(y_m)$ for some λ, $0 \leq \lambda \leq 1$.*

The following proposition can be proved in part by using the characterization of best approximation from subspaces in Hilbert space [5].

PROPOSITION 2.5. *Let $y \in \Re^k$, $S \subset \{0, 1, \ldots, n-1\}$, $S \neq \emptyset$, $b_S \neq \emptyset$, $m \geq 0$ with $[m] \in S$; then for $0 \leq \lambda \leq 1$*

$$\|(1 - \lambda)y + \lambda P_{b_{[m]}}(y) - P_{b_S}(y)\| \leq \|y - P_{b_S}(y)\|.$$

Let $S \neq \emptyset$ and define the subspace

$$\hat{b}_S = \bigcap_{i \in S} \{y \mid \langle y, a_i \rangle = 0\}.$$

If $b_S \neq \emptyset$, then $P_{b_S}(y) = P_{\hat{b}_S}(y) + P_{b_S}(\vec{0})$. Since $P_{\hat{b}_S}$ is continuous, Proposition 2.6 follows.

PROPOSITION 2.6. *Let $S \neq \emptyset$ and $b_S \neq \emptyset$; then P_{b_S} is continuous.*

PROPOSITION 2.7. *Let $y \in \Re^k$, $S \neq \emptyset$, and $b_S \neq \emptyset$; then*

$$P_{b_S}\left(y + \sum_{i \in S} \xi_i a_i\right) = P_{b_S}(y).$$

Proposition 2.7 can be shown using the fact that $P_{b_S}(y) = P_{\hat{b}_S}(y) + P_{b_S}(\vec{0})$, $P_{\hat{b}_S}$ is a linear operator, and $\hat{b}_S$ is the orthogonal complement of $\text{span}_{i \in S}\{a_i\}$.

Lemma 2.8 next shows that y_m and all subsequent Dykstra iterates are elements of the closed ball $\mathcal{B}_m$. As previously defined, $\mathcal{B}_m = \mathcal{B}[P_{b_{C_m}}(y_m),\ \mathrm{d}(y_m, P_{b_{C_m}}(y_m))]$. The lemma is actually proven by showing a stronger result as stated in the induction hypothesis.

LEMMA 2.8. *The iterates of Dykstra's algorithm y_{m+k}, $k = 0, 1, \ldots$, are contained in the closed ball $\mathcal{B}_m$.*

Proof. The proof is by induction.

It is evident that $y_m \in \mathcal{B}_m$. Moreover, by the definition of C_m (2.10c), for all $y \in \tilde{C}_m$, $e_{\pi_{m,j}} = \vec{0}$, where $\pi_{m,j} \in \{i \mid m - n \leq i \leq m - 1 \text{ and } [i] = j\}$.

We will assume that for each l, $1 \leq l \leq k$, $y_{m+l} \in \mathcal{B}_m$ and for all $j \in \tilde{C}_m$, $e_{\pi_{m+l,j}} = \vec{0}$, where $\pi_{m+l,j} \in \{i \mid m + l - n \leq i \leq m + l - 1 \text{ and } [i] = j\}$.

Next we will show that $y_{m+k+1} \in \mathcal{B}_m$.

If $[m+k] \in \tilde{C}_m$, then, by Lemma 2.3 and the induction hypothesis, $y_{m+k} \in h_{[m+k]}$. Therefore

$$y_{m+k+1} = P_{h_{[m+k]}}(y_{m+k} + \vec{0}) = P_{h_{[m+k]}}(y_{m+k}) = y_{m+k}.$$

Since $y_{m+k} \in \mathcal{B}_m$, $y_{m+k+1} \in \mathcal{B}_m$.

If $[m + k] \in C_m$, then by applying Result 2.4 to the $m + k + 1$ iterate,

$$y_{m+k+1} = (1 - \lambda)y_{m+k} + \lambda P_{b_{[m+k]}}(y_{m+k})$$

for some λ, $0 \leq \lambda \leq 1$. By Proposition 2.5,

$$\|(1 - \lambda)y_{m+k} + \lambda P_{b_{[m+k]}}(y_{m+k}) - P_{b_{C_m}}(y_{m+k})\| \leq \|y_{m+k} - P_{b_{C_m}}(y_{m+k})\|$$

for $0 \leq \lambda \leq 1$. Thus,

$$\|y_{m+k+1} - P_{b_{C_m}}(y_{m+k})\| \leq \|y_{m+k} - P_{b_{C_m}}(y_{m+k})\|.$$

Using the induction hypothesis, we have $y_{m+l} = y_{m+l-1}$ whenever $[m + l - 1] \in \tilde{C}_m$. Therefore, it can be shown that $y_{m+k} = y_m + \sum_{j \in C_m} \xi_j a_j$ and, by Proposition 2.7, $P_{b_{C_m}}(y_{m+k}) = P_{b_{C_m}}(y_m)$.

Therefore

$$\|y_{m+k+1} - P_{b_{C_m}}(y_m)\| \leq \|y_{m+k} - P_{b_{C_m}}(y_m)\|.$$

Since $y_{m+k} \in \mathcal{B}_m$, it follows that $y_{m+k+1} \in \mathcal{B}_m$.

Moreover, regardless of whether $[m + k] \in C_m$ or $[m + k] \in \tilde{C}_m$ for all $j \in \tilde{C}_m$, $e_{\pi_{m+k+1,j}} = \vec{0}$, where $\pi_{m+k+1,j} \in \{i \mid m + k + 1 - n \leq i \leq m + k \text{ and } [i] = j\}$.

Therefore, for $k = 0, 1, \ldots, y_{m+k} \in \mathcal{B}_m$. ∎

We next bound $\|y_m - P_h(y_0)\|$.

LEMMA 2.9. $\|y_m - P_h(y_0)\| \leq 2\|y_m - P_{b_{C_m}}(y_m)\|$.

Proof. By Lemma 2.8, $\{y_{m+k}\}_{k=0}^{\infty} \subset \mathcal{B}_m$. Since $\mathcal{B}_m$ is closed and $\lim_{k\to\infty} y_{m+k} = P_h(y_0)$, we have that $P_h(y_0) \in \mathcal{B}_m$. By the triangle inequality,

$$\begin{aligned}\|y_m - P_h(y_0)\| &\le \|y_m - P_{b_{C_m}}(y_m)\| + \|P_{b_{C_m}}(y_m) - P_h(y_0)\| \\ &\le 2\|y_m - P_{b_{C_m}}(y_m)\|. \qquad \square\end{aligned}$$

The first major result of this section follows.

THEOREM 2.10. $\|y_m - P_h(y_0)\| \le 2\|A_{C_m}^+\|\,\|r_m\|$.

Proof. By inequality (2.12), $2\|y_m - P_{b_{C_m}}(y_m)\| \le 2\|A_{C_m}^+\|\,\|r_m\|$. By Lemma 2.9, $\|y_m - P_h(y_0)\| \le 2\|y_m - P_{b_{C_m}}(y_m)\|$ and thus $\|y_m - P_h(y_0)\| \le 2\|A_{C_m}^+\|\,\|r_m\|$. $\square$

We now show the prerequisites for proving Theorems 2.16 and 2.17. Lemma 2.11 shows that $C = \{i \mid P_h(y_0) \in b_i\}$ satisfies property (2.10a).

LEMMA 2.11. $f_C \in \text{Range}(A_C)$.

Proof. By definition of C, $P_h(y_0) \in b_C$. Therefore $A_C P_h(y_0) = f_C$ and $f_C \in \text{Range}(A_C)$. $\square$

If $C = \{0, 1, \ldots, n-1\}$, then since y_m converges to $P_h(y_0)$, there exists a smallest integer M such that, for all $m > M - n$, $\|y_m - P_h(y_0)\| < \frac{1}{3\kappa}$, where $\kappa = \|A^+\|\,\|A\|$. Otherwise $C \neq \{0, 1, \ldots, n-1\}$, and again by the convergence of y_m to $P_h(y_0)$, there exists a smallest integer M such that, for all $m > M - n$, $\|y_m - P_h(y_0)\| < \frac{\varepsilon_C}{3\kappa_C}$, where $\kappa_C = \|A_C^+\|\,\|A_C\|$ and $\varepsilon_C > 0$ such that $\mathcal{B}[P_h(y_0), \varepsilon_C] \subset \text{int}(h_{\tilde{C}})$.

LEMMA 2.12. *For all $m > M$ and $j \in \tilde{C}$, $e_{\pi_{m,j}} = \vec{0}$, where $\pi_{m,j} \in \{i \mid m - n \le i \le m - 1$ and $[i] = j\}$.*

Proof. If $C = \{0, 1, \ldots, n-1\}$, then the lemma is vacuously true. Next assume that $C \neq \{0, 1, \ldots, n-1\}$ and let $m > M$. Suppose there exists an $j \in \tilde{C}$ such that $e_{\pi_{m,j}} \neq \vec{0}$, where $\pi_{m,j} \in \{i \mid m - n \le i \le m - 1$ and $[i] = j\}$. Since $e_{\pi_{m,j}} \neq \vec{0}$, this implies that $y_{\pi_{m,j}+1} = P_{b_j}(y_{\pi_{m,j}})$ and $y_{\pi_{m,j}+1} \in b_j$. This is impossible since $\pi_{m,j} + 1 > M - n$ and $y_{\pi_{m,j}+1} \in \mathcal{B}[P_h(y_0), \varepsilon_C] \subset \text{int}(h_j)$. Clearly $\langle y_{\pi_{m,j}+1}, a_j\rangle < f_j$ and $\langle y_{\pi_{m,j}+1}, a_j\rangle = f_j$ cannot both be true. Therefore $e_{\pi_{m,j}} = \vec{0}$.

Thus for all $m > M$ and $j \in \tilde{C}$, $e_{\pi_{m,j}} = \vec{0}$, where $\pi_{m,j} \in \{i \mid m - n \le i \le m - 1$ and $[i] = j\}$. $\square$

Lemma 2.12 implies the following result.

RESULT 2.13. *For all $m > M$, $y_{m+1} = y_m$ whenever $[m] \in \tilde{C}$.*

We now prove that for sufficiently large m, $2\|A_C^+\|\,\|r_m\| < f_j - \langle y_m, a_j\rangle$ for all $j \in \tilde{C}$.

LEMMA 2.14. *If $\{y_m\}$ is the sequence of Dykstra iterates, then for all $m > M$, $P_{b_C}(y_m) = P_h(y_0)$.*

Proof. As a consequence of Result 2.13 and the definition of Dykstra's algorithm, for all $m > M$,

$$y_m = y_{M+1} + \sum_{i\in C} \xi_i a_i.$$

Using Proposition 2.7,

$$\begin{aligned}P_{b_C}(y_m) &= P_{b_C}\left(y_{M+1} + \sum_{i\in C} \xi_i a_i\right) \\ &= P_{b_C}(y_{M+1}).\end{aligned}$$

By Proposition 2.6, P_{b_C} is continuous, and thus for all $m > M$

$$\begin{aligned} P_{b_C}(y_m) &= P_{b_C}(y_{M+1}) \\ &= \lim_{m\to\infty} P_{b_C}(y_m) \\ &= P_{b_C}\left(\lim_{m\to\infty} y_m\right) \\ &= P_{b_C}(P_h(y_0)) \\ &= P_h(y_0). \end{aligned}$$

The last equality holds since $P_h(y_0) \in b_C$. ∎

As specified above, $\kappa_C = \|A_C^+\|\,\|A_C\|$, $\varepsilon_C > 0$ such that $\mathcal{B}[P_h(y_0), \varepsilon_C] \subset \operatorname{int}(h_{\tilde{C}})$, and M is selected such that for all $m > M - n$ we have $\|y_m - P_h(y_0)\| < \frac{\varepsilon_C}{3\kappa_C}$. Moreover, it can be shown that $\kappa_C \geq 1$ [12].

LEMMA 2.15. *If $m > M$, then $2\|A_C^+\|\,\|r_m\| < f_j - \langle y_m, a_j\rangle$ for all $j \in \tilde{C}$.*

Proof. If $C = \{0, 1, \ldots, n-1\}$, then the lemma is vacuously true. Next assume that $C \neq \{0, 1, \ldots, n-1\}$ and let $m > M$.

$$\begin{aligned} 2\|A_C^+\|\,\|r_m\| &= 2\|A_C^+\|\,\|f_C - A_C y_m\| \\ &= 2\|A_C^+\|\,\|A_C(P_{b_C}(y_m) - y_m)\| \\ &\leq 2\|A_C^+\|\,\|A_C\|\,\|y_m - P_{b_C}(y_m)\| \\ &\leq 2\kappa_C\|y_m - P_{b_C}(y_m)\| \\ &= 2\kappa_C\|y_m - P_h(y_0)\| \\ &< 2\kappa_C\frac{\varepsilon_C}{3\kappa_C} \\ &= \frac{2\varepsilon_C}{3}. \end{aligned}$$

It remains to be shown that $\frac{2\varepsilon_C}{3} < f_j - \langle y_m, a_j\rangle$ for all $j \in \tilde{C}$.

Suppose there exists a $j \in \tilde{C}$ such that $f_j - \langle y_m, a_j\rangle \leq \frac{2\varepsilon_C}{3}$. Then

$$\begin{aligned} \mathrm{d}(P_h(y_0), b_j) &\leq \mathrm{d}(P_h(y_0), P_{b_j}(y_m)) \\ &\leq \mathrm{d}(P_h(y_0), y_m) + \mathrm{d}(y_m, P_{b_j}(y_m)) \\ &= \|y_m - P_h(y_0)\| + \mathrm{d}(y_m, b_j) \\ &< \frac{\varepsilon_C}{3\kappa_C} + \mathrm{d}(y_m, b_j) \\ &\leq \frac{\varepsilon_C}{3} + \mathrm{d}(y_m, b_j) \\ &= \frac{\varepsilon_C}{3} + |\langle y_m, a_j\rangle - f_j| \\ &\leq \frac{\varepsilon_C}{3} + \frac{2\varepsilon_C}{3} \\ &= \varepsilon_C. \end{aligned}$$

Thus there exists an element of b_j that is within ε_C of $P_h(y_0)$. This is a contradiction since $\mathcal{B}[P_h(y_0), \varepsilon_C] \subset \text{int}(h_j)$. Therefore, for all $m > M$ and $j \in \tilde{C}$ we have that $2\|A_C^+\| \, \|r_m\| < f_j - \langle y_m, a_j \rangle$. ∎

THEOREM 2.16. *For all $m > M$, C satisfies properties* (2.10a)–(2.10c).

Proof. This is a consequence of Lemmas 2.11, 2.12, and 2.15. ∎

For the remainder of the section, we will assume that $C_m = C$ whenever $m > M$.

THEOREM 2.17. $\lim_{m\to\infty} \rho_m = 0$.

Proof. Let $m > M$. By assumption we have that $C_m = C$ and $\rho_m \leq 2\kappa_C \|y_m - P_{b_C}(y_m)\|$. By Lemma 2.14, for all $m > M$, $P_{b_C}(y_m) = P_h(y_0)$, and thus $\rho_m \leq 2\kappa_C \|y_m - P_h(y_0)\|$. Therefore

$$\lim_{m\to\infty} \rho_m \leq \lim_{m\to\infty} 2\kappa_C \|y_m - P_h(y_0)\| = 0,$$

and $\lim_{m\to\infty} \rho_m = 0$. ∎

We have now established that it is possible to construct a sequence $\{\rho_m\}$ such that $\|y_m - P_h(y_0)\| \leq \rho_m$ with $\lim_{m\to\infty} \rho_m = 0$. Next we exhibit a $y \in h$ such that $\|y - P_h(y_0)\| < \varepsilon$ for any $\varepsilon > 0$.

Given $\varepsilon > 0$, it is possible to determine an m such that C_m is nonempty and $\|y_m - P_h(y_0)\| < 2\varepsilon$. Bramley and Sameh have shown that using the conjugate-gradient method to solve $A_{C_m}^{\mathrm{T}} A_{C_m} y = A_{C_m}^{\mathrm{T}} b_{C_m}$ with an initial approximation of y_m results in a solution of $P_{b_{C_m}}(y_m)$ [2]. Using arguments in Lemma 2.9 and Theorem 2.10,

$$\|P_{b_{C_m}}(y_m) - P_h(y_0)\| < \varepsilon.$$

The following lemma shows that $P_{b_{C_m}}(y_m) \in h$.

LEMMA 2.18. *If $C_m \neq \emptyset$, then $P_{b_{C_m}}(y_m) \in h$.*

Proof. If $C_m = \{0, 1, \ldots, n-1\}$ (equivalently if $\tilde{C}_m = \emptyset$), then $P_b(y_m) \in b \subset h$. Next suppose that $\tilde{C}_m \neq \emptyset$. By Lemma 2.3, $P_{b_{C_m}}(y_m) \in \text{int}(h_{\tilde{C}_m}) \subset h_{\tilde{C}_m}$. In addition $P_{b_{C_m}}(y_m) \in b_{C_m} \subset h_{C_m}$. Therefore $P_{b_{C_m}}(y_m) \in (h_{C_m} \cap h_{\tilde{C}_m}) = h$. ∎

Thus setting $y = P_{b_{C_m}}(y_m)$ specifies a point in h with the property that $\|y - P_h(y_0)\| < \varepsilon$. Moreover, if $m > M$, then by Lemma 2.14, $y = P_{b_{C_m}}(y_m)$ is in fact $P_h(y_0)$.

3. Implementation, applications, and numerical results. The two main results of section 2 were Theorem 2.10 and Theorem 2.17. In Theorem 2.10 we showed that if $C_m \neq \emptyset$, then $\|y_m - P_h(y_0)\| \leq 2\|A_{C_m}^+\| \, \|r_m\|$, and we subsequently defined $\rho_m = 2\|A_{C_m}^+\| \, \|r_m\|$. Theorem 2.17 implied that for sufficiently large m, if $C_m = C$, then $\lim_{m\to\infty} \rho_m = 0$. We remark that for m sufficiently large and $C_m = C$, we have, using Lemma 2.14, $\|y_m - P_h(y_0)\| \leq \frac{1}{2}\rho_m$.

Next we show how to construct a sequence $\{C_m\}$ so that for large enough m, $C_m = C$. Lemma 3.1 shows that if $C_m \neq \emptyset$, then $C \subset C_m$. As previously specified, $\mathcal{B}_m = \mathcal{B}[P_{b_{C_m}}(y_m), \, \mathrm{d}(y_m, P_{b_{C_m}}(y_m))]$, $C = \{i \mid P_h(y_0) \in b_i\}$, and C_m is selected to satisfy properties (2.10a)–(2.10c).

LEMMA 3.1. *If $C_m \neq \emptyset$, then $C \subset C_m$.*

Proof. The proof is by contradiction.

Suppose there exists an m such that $C - C_m \neq \emptyset$. Let $j \in C - C_m$ and hence $j \in \tilde{C}_m$. By Lemma 2.3 the closed ball $\mathcal{B}_m \subset \text{int}(h_j)$. By Lemma 2.8, $\{y_{m+k}\}_{k=0}^{\infty} \subset \mathcal{B}_m$. Since $\mathcal{B}_m$ is closed and $\lim_{k\to\infty} y_{m+k} = P_h(y_0)$, $P_h(y_0) \in \mathcal{B}_m$. By definition of C, $P_h(y_0) \in b_j$. In addition we have shown above that $P_h(y_0) \in \mathcal{B}_m \subset \text{int}(h_j)$. It

is impossible for $P_h(y_0)$ to be both in the interior of h_j and on the boundary of h_j. Therefore, if $C_m \neq \emptyset$, we have that $C \subset C_m$. □

By Theorem 2.16, $C_m = C$ satisfies properties (2.10a)–(2.10c) for sufficiently large m. By Lemma 3.1, if $C_m \neq \emptyset$, then $C \subset C_m$. Therefore, if C_m is chosen to have the smallest cardinality of any set satisfying properties (2.10a)–(2.10c), then for sufficiently large m, C_m must be the set C.

As previously stated, we compute

$$\rho_m = 2\|A^+_{C_m}\| \, \|r_m\|,$$

where

$$\|A^+_{C_m}\| = (\sigma_{\min}(A_{C_m}))^{-1},$$

$\sigma_{\min}(A_{C_m}) \neq 0$.

If the columns of A^{T} are linearly independent, we will subsequently show how to reduce the computational effort required to bound $\|y_m - P_h(y_0)\|$.

Proposition 3.2 may be directly inferred by the repeated application of the interlacing property for singular values [12].

PROPOSITION 3.2. *Let A^{T} have linearly independent columns and $S \subset \{0, 1, \ldots, n-1\}$, $S \neq \emptyset$; then $\|A^+_S\| \leq \|A^+\|$.*

We next determine a set C_m with the properties

$$f_{C_m} \in \mathrm{Range}(A_{C_m}), \tag{3.1a}$$

$$\text{for all } j \in \tilde{C}_m, \quad 2\|A^+\| \, \|r_m\| < f_j - \langle y_m, a_j \rangle, \tag{3.1b}$$

$$\text{for all } j \in \tilde{C}_m, \quad e_{\pi_{m,j}} = \vec{0},$$

$$\text{where } \pi_{m,j} \in \{i \mid m - n \leq i \leq m-1 \ \text{ and } \ [i] = j\}. \tag{3.1c}$$

C_m is well defined as $C_m = \{0, 1, \ldots, n-1\}$ vacuously satisfies properties (3.1a)–(3.1c). Using arguments similar to those presented in section 2, it can be shown that $\|y_m - P_h(y_0)\| \leq 2\|A^+\| \, \|r_m\|$. Thus, in order to bound the error of each Dykstra iterate we need only determine $\sigma_{\min}(A)$ and $\|r_m\|$.

PROPOSITION 3.3. *Let $S \subset \{0, 1, \ldots, n-1\}$, $S \neq \emptyset$, then $\|A_S\| \leq \|A\|$.*

Assuming that the columns of A^{T} are linearly independent, using Propositions 3.2 and 3.3 and selecting C_m to satisfy properties (3.1a)–(3.1c),

$$\begin{aligned}
\|y_m - P_h(y_0)\| &\leq 2\|A^+\| \, \|r_m\| \\
&= 2\|A^+\| \, \|A_{C_m}(y_m - P_{h_{C_m}}(y_m))\| \\
&\leq 2\|A^+\| \, \|A_{C_m}\| \, \|y_m - P_{h_{C_m}}(y_m)\| \\
&\leq 2\|A^+\| \, \|A\| \, \|y_m - P_{h_{C_m}}(y_m)\| \\
&= 2\kappa\|y_m - P_{h_{C_m}}(y_m)\|,
\end{aligned}$$

where $\kappa = \|A^+\| \, \|A\|$.

In order for the bound $2\|A^+\| \, \|r_m\|$ to be useful, $\{C_m\}$ must be selected such that $\lim_{m\to\infty} \|A^+\| \, \|r_m\| = 0$. To guarantee that

$$\lim_{m\to\infty} 2\kappa\|y_m - P_{h_{C_m}}(y_m)\| = 0,$$

it suffices that for sufficiently large m, $C_m = C$. As in section 2, with such a selection of $\{C_m\}$, $\lim_{m\to\infty} 2\|A^+\|\,\|r_m\| = 0$.

An important class of problems in statistical inference is to find the N-convex regression for a real valued function g_0 defined on a finite subset of $\Re$. Let $t_1 < t_2 < \cdots < t_q$, $q \geq N+1$. For a given $y_0 = (g_0(t_1), \ldots, g_0(t_q))^{\mathrm{T}}$ in $\Re^q$ we want to determine the best approximation to y_0 from the set of N-convex functions in $\Re^q$.

The real valued function g defined on $t_1, \ldots, t_q$ is N-convex if for any set of $N+1$ points $t_{i_1} < t_{i_2} < \cdots < t_{i_{N+1}}$, the Nth order divided difference, $g[t_{i_1}, \ldots, t_{i_{N+1}}] \geq 0$. This type of approximation problem generates a rectangular system of linear inequalities $Ay \leq f$ such that $\{y \mid Ay \leq f\}$ is a closed convex cone.

For N-convex regression problems it is possible to show that if $h = \{y \in \Re^q \mid y \text{ is } N\text{-convex}\}$, then $h = \cap_{i=0}^{q-N-1} h_i$, where $h_i = \{y \in \Re^q \mid \langle y, z_i\rangle \leq 0\}$, $b_i = \{y \in \Re^q \mid \langle y, z_i\rangle = 0\}$, and

$$\hat{z}_i(j) = \begin{cases} (-1)^{j-i+N}\binom{N}{j-i-1} & \text{whenever } \ 0 \leq j-i-1 \leq N, \\ 0 & \text{otherwise} \end{cases}$$

with

$$z_i = \frac{\hat{z}_i}{\|\hat{z}_i\|}. \tag{3.2}$$

The following proposition is easily proved.

PROPOSITION 3.4. *The* $\{z_i\}_{i=0}^{q-N-1}$ *as defined in* (3.2) *is linearly independent.*

Thus it is possible, given $y_0 = (g_0(t_1), \ldots, g_0(t_q))^{\mathrm{T}}$, to use Dykstra's algorithm to approximate $P_h(y_0)$ and to bound the norm of the error for each Dykstra iterate, $\|y_k - P_h(y_0)\|$, using the previously developed theory.

The set h of monotonically increasing functions can be expressed using a system of linear inequalities $h = \{y \in \Re^q \mid Ay \leq \vec{0}\}$, where $A^{\mathrm{T}} \in \Re^{q\times q-1}$. Let $\hat{e} \in \Re^q$ and for $i = 1, \ldots, q$, $j = 1, \ldots, q$, define

$$\hat{e}_i(j) = \begin{cases} 0 & \text{whenever } \ i = j, \\ 1 & \text{otherwise.} \end{cases}$$

Then $A^{\mathrm{T}} = [a_0, \ldots, a_{q-2}]$, where $a_i = \frac{1}{\sqrt{2}}(\hat{e}_{i+2} - \hat{e}_{i+1})$ for $i = 0, \ldots, q-2$.

By Proposition 3.4, the columns of A^{T} are linearly independent. In order to bound the norm of the error of the Dykstra iterates when approximating $P_h(y_0)$, we need only to calculate $\|A^+\|$ and $\|r_m\|$. As previously stated, $\|A^+\| = (\sigma_{\min}(A))^{-1}$ and since $(\sigma_{\min}(A))^{-1} = \lambda_{\min}(AA^{\mathrm{T}})^{-\frac{1}{2}}$, we need to estimate the smallest eigenvalue of AA^{T}. The matrix AA^{T} is tridiag$(-0.5, 1, -0.5)$, and thus $\lambda_{\min}(AA^{\mathrm{T}})$ can be approximated by using the QR algorithm with Givens rotations.

We present an example of monotonic regression and exhibit error bounds on $\|y_m - P_h(y_0)\|$. Here $y_0 = (g_0(t_1), \ldots, g_0(t_q))^{\mathrm{T}}$, where $g_0(t_i) \geq g_0(t_j)$ whenever $1 \leq i < j \leq q$. It can be shown that $P_h(y_0) = (g(t_1), \ldots, (g(t_1))^{\mathrm{T}}$, where $g(t_i) = \frac{1}{q}\sum_{j=1}^{q} g_0(t_j)$ for $1 \leq i \leq q$. Thus it is possible to compare $\|y_m - P_h(y_0)\|$ to the error bounds calculated using the previously derived theory. In the example $q = 31$, $g_0(t_i) = 16 - i$ for $1 \leq i \leq 31$ and $C = \{0, 1, \ldots, 30\}$. Table 3.1 shows $\|y_m - P_h(y_0)\|$ and $\rho_m = 2\|A^+\|\,\|r_m\|$ for selected values of m.

In order to compute the error bounds for this illustration, we need to determine for each iterate a set C_m that satisfies properties (3.1a)–(3.1c). For $m = 1, \ldots,$ it

TABLE 3.1.

m	$\|y_m - P_h(y_0)\|$	$\rho_m = 2\|A^+\| \, \|r_m\|$
6.0 E +3	6.3504 E +0	1.2778 E +1
1.2 E +4	8.1138 E −1	1.6326 E +0
1.8 E +4	1.0367 E −1	2.0854 E −1
2.4 E +4	1.3245 E −2	2.6651 E −2
3.0 E +4	1.6923 E −3	3.4051 E −3
3.6 E +4	2.1622 E −4	4.3506 E −4
4.2 E +4	2.7625 E −5	5.5586 E −5
4.8 E +4	3.5296 E −6	7.1020 E −6
5.4 E +4	4.5096 E −7	9.0740 E −7
6.0 E +e	5.7617 E −8	1.1593 E −7

can be shown that $C_m = \{0, 1, \ldots, 30\}$ satisfies properties (3.1a)–(3.1c). In addition, it is impossible to construct a proper subset of C_m which also satisfies these same properties. Thus, to compute a bound on $\|y_m - P_h(y_0)\|$ which tends to zero, we need only to calculate $\|r_m\| = \|Ay_m\|$ and multiply by the constant $2\|A^+\|$. In general, in order to guarantee that $\lim_{m\to\infty} \rho_m = 0$, the sequence of sets C_m must be selected judiciously. As previously discussed in this section, selecting C_m to have the smallest cardinality of any set satisfying properties (2.10a)–(2.10c) guarantees that $\lim_{m\to\infty} \rho_m = 0$.

The previous discussion shows that it is possible to compute a bound in the general Hilbert space setting on $\|x_m - P_H(x_0)\|$ that tends to zero where x_n is the nth iterate of Dykstra's algorithm and H is the intersection of a finite number of closed half spaces. The convergence is studied via an equivalent problem in $\Re^k$. This approach is well suited for those occasions where H is a closed convex cone uniquely determined by intersecting half spaces. The analysis is of particular interest to the author when using Dykstra's algorithm to estimate attribute utility values in nonmetric trade-off experiments.

REFERENCES

[1] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[2] R. BRAMLEY AND A. SAMEH, *Row projection methods for large nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 168–193.

[3] J. P. BOYLE AND R. L. DYKSTRA, *A method for finding projections onto the intersection of convex sets in Hilbert spaces*, in Advances in Order Restricted Statistical Inference, Lecture Notes in Statist. 37, Springer-Verlag, Berlin, 1985, pp. 28–47.

[4] R. L. DYKSTRA, *An algorithm for restricted least squares regression*, J. Amer. Statist. Assoc., 78 (1993), pp. 837–842.

[5] F. DEUTSCH, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[6] F. DEUTSCH AND H. HUNDAL, *The rate of convergence of Dykstra's cyclic projections algorithm: The polyhedral case*, Numer. Funct. Anal. Optim., 15 (1994), pp. 537–565.

[7] R. FLETCHER, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, Chichester, UK, 1991.

[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[9] S. P. HAN, *A successive projection method*, Math. Program., 40 (1988), pp. 1–14.

[10] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 36 (erratum), 79–85.

[11] A. N. IUSEM AND A. R. DE PIERRO, *On the convergence properties of Hildreth's quadratic programming algorithm*, Math. Program., 47 (1990), pp. 37–51.

[12] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge,

UK, 1985.
[13] B. F. Mitchell, V. F. Dem'yanov, and V. N. Malozemov, *Finding the point of a polyhedron closest to the origin*, SIAM J. Control, 12 (1974), pp. 19–26.
[14] T. S. Motzkin and J. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.
[15] S. Xu, *Successive approximate algorithm for best approximation from a polyhedron*, J. Approx. Theory, 93 (1998), pp. 415–433.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 805–819

POSTPROCESSING THE LINEAR FINITE ELEMENT METHOD*

JAVIER DE FRUTOS† AND JULIA NOVO†

Abstract. We extend the idea of the postprocessing Galerkin methods for dissipative evolution equations to the case of the linear finite element method. The postprocessing technique has been developed earlier for spectral methods and for higher order finite element methods.

The analysis shows that this procedure improves the order of convergence of the piecewise linear Galerkin finite element approximation in the H^1 norm. We show by means of numerical experiments that there is no improvement in the order of convergence in the L^2 norm.

Key words. dissipative equations, postprocessing linear finite element methods

AMS subject classifications. 65M60, 65M15, 65M20

PII. S0036142900375438

1. Introduction.

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded domain with a smooth boundary. We consider dissipative partial differential equations which can be written in the abstract form

$$u_t + \nu Au + F(u) = f, \tag{1.1}$$

in a suitable Hilbert space H. In (1.1) $\nu > 0$ is a scalar, $-A$ is usually the Laplace operator with appropriate boundary conditions, although other elliptic operators can also be considered, F is a nonlinear term (typically a reaction or a convection term), and f is a function that can be time-dependent. Throughout the paper we suppose that for each initial condition $u(\cdot, 0) = u_0$, smooth enough, (1.1) has a unique solution defined in some interval $[0, T]$.

Recently, a postprocessing technique has been introduced to increase the efficiency of Galerkin methods of spectral type [4], including the p-version of the finite element method (spectral element method); see [10], [7], [8], [9]. Postprocessed methods yield greater accuracy than standard Galerkin schemes at nearly the same computational cost so that the postprocessed Galerkin method really improves the efficiency of the method to which it is applied.

In [11] the postprocessing technique has been extended to the h-version of the finite element method for dissipative partial differential equations. There, the authors prove that the postprocessed method has a higher rate of convergence than the standard finite element method when other than piecewise linear (say, quadratic, cubic, ...) finite elements are used. The main reason for this limitation is that in [11] the emphasis is on error analysis in the $L^2(\Omega)$ norm, whereas, in the case of piecewise linear finite elements, the improvement in the rate of convergence is observed only in the $H^1(\Omega)$ norm (the energy norm).

In the present paper we prove that the postprocessing technique applied to the linear finite element method improves the rate of convergence of this method when the error is measured in the $H^1(\Omega)$ norm. Furthermore, we present some numerical

*Received by the editors July 27, 2001; accepted for publication (in revised form) March 11, 2002; published electronically August 1, 2002.

http://www.siam.org/journals/sinum/40-3/37543.html

†Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, Valladolid 47011, Spain (frutos@mac.cie.uva.es, jnovo@mac.cie.uva.es). The research of the first author was partly supported by project MCYT BFM2001-2138. The research of the second author was partly supported by projects DGICYT PB98-074 and MCYT BFM2001-2138.

experiments showing that the lack of gain in the $L^2(\Omega)$ norm is really observed in practice and thus is not merely a consequence of the technique employed in the proofs. We wish to note that this fact has not been previously reported. We point out that our technique of proof is related to but different from the one employed in [11].

For the sake of simplicity, in this paper we shall concentrate on equations of convection-diffusion type as, for example,

$$u_t - \nu\Delta u + (u \cdot \nabla)u = f, \tag{1.2}$$

or reaction-diffusion type, such as

$$u_t - \nu\Delta u + g(u) = f, \tag{1.3}$$

on a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, subject to homogeneous Dirichlet boundary conditions. We confine ourselves to reaction terms of polynomial type

$$g(u) = \sum_{j=0}^{2p-1} b_j u^j, \qquad b_{2p-1} > 0,$$

although, in both cases, more general types of nonlinearities can also be treated along very similar lines. The Hilbert space in the abstract formulation (1.1) is $L^2(\Omega)^n$, with $n = 1$ for (1.3) or $n = d$ for (1.2). In what follows we will denote by $F(u)$ either $F(u) = (u \cdot \nabla)u$ or $F(u) = g(u)$.

In this paper we show that the gain in the convergence rate that is obtained through postprocessing the linear finite element method is the same for reaction-diffusion and convection-diffusion equations. This fact is confirmed by the numerical experiments we present. In [11] the improvement obtained by postprocessing reaction-diffusion type equations was superior to the one obtained by postprocessing convection-diffusion equations, but only when the postprocessing was applied after using the finite element method with polynomials of degree greater than or equal to 3, and boundary data approximations were performed with a sufficiently high order of accuracy by means of superparametric finite elements.

The rest of the paper is as follows. Section 2 contains some standard preliminary material; in section 3 the postprocessed method is introduced. Section 4 is devoted to the proof of the main results of the paper. Finally, some numerical experiments are presented in section 5.

2. Preliminaries and notation. Let $W^{s,p}(\Omega)$, $H^s(\Omega)$, $p \geq 1$, $s \geq 0$ be the standard Sobolev spaces. For simplicity of notation, when we consider (1.2), we shall write $W^{s,p}(\Omega)$, $H^s(\Omega)$, or $L^p(\Omega)$ instead of $W^{s,p}(\Omega)^d$, $H^s(\Omega)^d$, or $L^p(\Omega)^d$. For reasons of convenience we denote by $\|\cdot\|_0$, $\|\cdot\|_\infty$, and $\|\cdot\|_1$ the norms of the spaces $L^2(\Omega)$, $L^\infty(\Omega)$, and $H^1(\Omega)$, respectively. Let $H_0^1(\Omega)$ be the space of functions in $H^1(\Omega)$ with null trace at the boundary of Ω. The norm of its dual space will be denoted by $\|\cdot\|_{-1}$.

We will frequently make use of the following Sobolev inequalities:

$$\|u\|_\infty \leq C\|u\|_{H^{d/2+\epsilon}(\Omega)}, \qquad u \in H^{d/2+\epsilon}(\Omega) \quad \epsilon > 0, \tag{2.1}$$

$$\|u\|_{W^{j,q}(\Omega)} \leq C\|u\|_{W^{j+s,p}(\Omega)}, \quad \frac{1}{p} \geq \frac{1}{q} \geq \frac{1}{p} - \frac{s}{d}, \quad u \in W^{j+s,p}(\Omega). \tag{2.2}$$

We refer to [1, Lemma 5.15] for (2.1) and to [1, Theorem 5.4], [2, Theorem 6.5.1] for (2.2).

In $H_0^1(\Omega)$ we consider the bilinear form induced by $A = -\triangle$, that is,

$$a(u,v) = (A^{1/2}u, A^{1/2}v) = (\nabla u, \nabla v), \qquad u, v \in H_0^1(\Omega).$$

It is well known that $a(\cdot,\cdot)$ is continuous and coercive in $H_0^1(\Omega)$.

Let us fix a positive time T. The error estimates in section 4 and the constant in (2.8) below depend on the constant

$$K(u) = \max(K_1(u), K_2(u)), \tag{2.3}$$

where

$$K_1(u) = \max_{0\le t\le T} \|u(\cdot,t)\|_{H^2(\Omega)}, \qquad K_2(u) = \max_{0\le t\le T} \|u_t(\cdot,t)\|_{H^2(\Omega)}.$$

Notice that we implicitly assume that the solution is sufficiently regular. Equation (1.1) has sufficiently smooth solutions; in fact, the solution is analytic in time if we assume that the nonlinear terms and the data of the problem are smooth and satisfy certain compatibility conditions. We refer to [12] for results concerning the regularity properties of solutions of dissipative partial differential equations.

In order to simplify the description of the postprocessed method, in the following we shall suppose that Ω is a convex polygonal or polyhedral domain, although the results we present are applicable to the general case in which Ω is a bounded domain with a smooth boundary.

Let $\mathcal{T}_h = (\tau_i^h, \phi_i^h)$, $h > 0$, be a family of partitions of Ω; the parameter h is the maximum diameter of the elements τ_i^h in $\mathcal{T}_h$ and ϕ_i^h are affine mappings of the reference simplex τ_0 onto τ_i^h. For $r \ge 2$ we consider the finite element spaces

$$S_{h,r} = \left\{ \chi_h \in C(\overline{\Omega}) \ \mid \chi_{h|\tau_i^h} \circ \phi_i^h \in \mathbb{P}^{r-1}(\tau_0),\ \chi_h(x) = 0\ \forall x \in \partial\Omega \right\},$$

where $\mathbb{P}^{r-1}(\tau_0)$ denotes the space of polynomials of degree $r-1$ or less on τ_0.

We shall restrict ourselves to quasi-uniform meshes $\mathcal{T}_h$ so that the following inverse estimate holds for any element $\tau \in \mathcal{T}_h$ and $v_h \in S_{h,2}$ (see, e.g., [13]):

$$\|v_h\|_{W^{m,p}(\tau)} \le C h^{l-m-d(\frac{1}{q}-\frac{1}{p})} \|v_h\|_{W^{l,q}(\tau)}, \quad 0 \le l \le m \le 1, \quad 1 \le q \le p \le \infty. \tag{2.4}$$

There exists a piecewise linear (quasi) interpolant such that, for $1 \le p \le \infty$, $0 \le m \le s \le 2$ and any $u \in L^1(\Omega) \cap W^{s,p}(\Omega)$ [3],

$$\|u - I_h(u)\|_{W^{m,p}(\Omega)} \le C h^{s-m} \|u\|_{W^{s,p}(\Omega)}. \tag{2.5}$$

Let $a_h(\cdot,\cdot)$ be the bilinear form on $S_{h,r}$ defined by

$$a_h(\chi_h, \psi_h) = (\nabla\chi_h, \nabla\psi_h) \quad \forall \chi_h, \psi_h \in S_{h,r}.$$

We shall denote by A_h the associated positive, self-adjoint operator, that is,

$$a_h(\chi_h, \psi_h) = (A_h\chi_h, \psi_h) \quad \forall \chi_h, \psi_h \in S_{h,r}.$$

The standard $L^2(\Omega)$ orthogonal projection and the elliptic projection onto $S_{h,r}$ are denoted by $P_{h,r}$ and $R_{h,r}$, respectively. Recall that, for $u \in H_0^1(\Omega)$,

$$a_h(R_{h,r}(u), \chi_h) = a(u, \chi_h) \quad \forall \chi_h \in S_{h,r}.$$

We shall use the bound [14], [5]

$$\|u - R_{h,2}(u)\|_0 + h\|u - R_{h,2}(u)\|_1 \leq Ch^2\|u\|_{H^2(\Omega)}, \qquad u \in D(A). \tag{2.6}$$

Let u be the solution of (1.2) or (1.3) with initial condition $u(\cdot,0) = u_0$. Then the Galerkin linear finite element approximation to u is $u_h : [0,T] \to S_{h,2}$, $u_h(0) = R_{h,2}(u_0)$ such that, for all $t \in [0,T]$ and $\varphi_h \in S_{h,2}$,

$$((u_h)_t, \varphi_h) + \nu(\nabla u_h, \nabla\varphi_h) + (F(u_h), \varphi_h) = (f, \varphi_h). \tag{2.7}$$

The next error estimate for the approximation u_h is well known:

$$\|u(t) - u_h(t)\|_0 + h\|u(t) - u_h(t)\|_1 \leq C(K(u))h^2, \qquad 0 \leq t \leq T. \tag{2.8}$$

Finally, set $B(u,v) = (u \cdot \nabla)v$. We shall make use of the following inequality [6, Remark 6.2]. For each $u \in L^\infty(\Omega)$, $v \in H^1(\Omega)$, and $w \in L^2(\Omega)$,

$$|(B(u,v),w)| \leq C\|u\|_\infty\|v\|_1\|w\|_0. \tag{2.9}$$

3. The postprocessed method. Fix $T > 0$ and let us suppose that the Galerkin approximation $u_h(T) \in S_{h,2}$ has been computed solving (2.7). We consider an improved finite element space $\widetilde{S}_h$. Two choices are possible for the new space:

(1) $\widetilde{S}_h = S_{h',2}$, $h' < h$, the linear finite element space that is obtained from $S_{h,2}$ by refining the partition. That is, every element τ_i^h is divided into a finite number of elements $\tau_j^{h'}$.

(2) If the solution u of (1.2) or (1.3) with initial condition u_0 belongs to $H^3(\Omega) \cap H_0^1(\Omega)$ we can take $\widetilde{S}_h = S_{h,3}$, the space of piecewise quadratic polynomials over the same grid.

The postprocessed approximation $\tilde{u}_h(T) \in \widetilde{S}_h$ is the solution of the discrete elliptic problem

$$\nu(\nabla\tilde{u}_h(T), \nabla\varphi_h) = -(F(u_h(T)), \varphi_h) - ((u_h(T))_t, \varphi_h) + (f, \varphi_h) \tag{3.1}$$

for all $\varphi_h \in \widetilde{S}_h$.

Remark 3.1. Let W_h be the orthogonal complement of $S_{h,2}$ in $\widetilde{S}_h$ with respect to the inner product in $H_0^1(\Omega)$ (i.e., $(\nabla\cdot, \nabla\cdot)$). Then the postprocessed approximation $\tilde{u}_h(T)$ at time $T > 0$ can also be obtained adding to the Galerkin approximation $u_h(T)$, the *approximate inertial manifold* $\Phi_h(u_h(T)) \in W_h$ defined by

$$\nu(\nabla\Phi_h(u_h(T)), \nabla\psi_h) = -(F(u_h(T)), \psi_h) - ((u_h(T))_t, \psi_h) + (f, \psi_h) \tag{3.2}$$

for all $\psi_h \in W_h$.

4. Analysis of the postprocessed method. In this section we establish the rate of convergence of the postprocessed linear finite element method in the $H^1(\Omega)$ norm. We shall prove that the postprocessed approximation improves, up to one unit in terms of h (up to a logarithmic term), the rate of convergence of the Galerkin linear finite element method for both choices of the refined finite element space $\widetilde{S}_h$.

The proof of the two main theorems will require several preparatory lemmas which we now state.

LEMMA 4.1. *Let $g \in L^\infty([0,T], S_{h,2})$. For all $t \in [0,T]$ we have*

$$\left\| \int_0^t e^{-(t-s)\nu A_h} A_h g(s)\, \mathrm{d}s \right\|_0 \leq C\nu^{-1}\log(1/h) \max_{0\leq t\leq T} \|g(t)\|_0,$$

where the constant C does not depend on h.

Proof. We refer to [8, Lemma 4.7] since the proof can be obtained by the same steps. The only requirement needed in that proof is the bound

$$(A_h v_h, v_h) \le Ch^{-2}\|v_h\|_0^2, \qquad v_h \in S_{h,2},$$

which is a direct consequence of the inverse inequality (2.4). ∎

LEMMA 4.2. *Let u be the solution of* (1.2) *with initial condition u_0 and let u_h be the Galerkin linear finite element approximation to u. Then, if $d = 2$, the following bound holds with $l = -1, 0$:*

$$\|B(u_h, u_h) - B(u, u)\|_l \le C\|u - u_h\|_{l+1}\|u\|_{H^{3/2}(\Omega)}. \tag{4.1}$$

If $d = 3$, for $l = -1, 0$, and $\epsilon \in (0, 1/2]$,

$$\|B(u_h, u_h) - B(u, u)\|_l \le C\|u - u_h\|_{l+1}\|u\|_{H^{3/2+\epsilon}(\Omega)}. \tag{4.2}$$

Proof. We first notice that

$$\|B(u_h, u_h) - B(u, u)\|_l \le \|B(u_h - u, u_h)\|_l + \|B(u, u_h - u)\|_l. \tag{4.3}$$

Let us prove (4.1). Taking into account that in this case $d = 2$, we first observe that

$$\|\nabla u_h\|_{L^4(\Omega)} \le C\|u\|_{H^{3/2}(\Omega)}. \tag{4.4}$$

To obtain this bound we use (2.8), (2.4), (2.5), and (2.2) to get

$$\begin{aligned}
\|\nabla u_h\|_{L^4(\Omega)} &\le \|\nabla(u_h - I_h u)\|_{L^4(\Omega)} + \|\nabla(I_h u - u)\|_{L^4(\Omega)} + \|\nabla u\|_{L^4(\Omega)} \\
&\le Ch^{-1/2}\|\nabla(u_h - I_h u)\|_{L^2(\Omega)} + C\|u\|_{W^{1,4}(\Omega)} + \|\nabla u\|_{L^4(\Omega)} \\
&\le Ch^{-1/2}\left(\|\nabla(u_h - u)\|_{L^2(\Omega)} + \|\nabla(u - I_h u)\|_{L^2(\Omega)}\right) + C\|u\|_{W^{1,4}(\Omega)} \\
&\le Ch^{-1/2}h^{1/2}\|u\|_{H^{3/2}(\Omega)} + C\|u\|_{W^{1,4}(\Omega)} \\
&\le C\|u\|_{H^{3/2}(\Omega)}.
\end{aligned}$$

Then we have

$$\begin{aligned}
|(B(u_h - u, u_h), \varphi)| &\le \|u - u_h\|_0\|\nabla u_h\|_{L^4(\Omega)}\|\varphi\|_{L^4(\Omega)} \\
&\le C\|u - u_h\|_0\|u\|_{H^{3/2}(\Omega)}\|\varphi\|_{H^{1/2}(\Omega)} \\
&\le C\|u - u_h\|_0\|u\|_{H^{3/2}(\Omega)}\|\varphi\|_1,
\end{aligned}$$

where we have used (4.4) and (2.2). Consequently, the first term in (4.3) is bounded by

$$\|B(u_h - u, u_h)\|_{-1} \le C\|u - u_h\|_0\|u\|_{H^{3/2}(\Omega)}.$$

The proof for the case $l = -1$ is completed by using (2.9), (2.1), and (2.2) to estimate the second term in (4.3). Thus we get

$$\begin{aligned}
|(B(u, u_h - u), \varphi)| &\le \|(B(u, \varphi), u_h - u)| + |(\mathrm{div}(u)(u_h - u), \varphi)| \\
&\le \|u - u_h\|_0\left(\|\varphi\|_1\|u\|_\infty + \|\nabla u\|_{L^4(\Omega)}\|\varphi\|_{L^4(\Omega)}\right) \\
&\le C\|u - u_h\|_0\|u\|_{H^{3/2}(\Omega)}\|\varphi\|_1.
\end{aligned}$$

The case $l = 0$ is treated in a similar way. We have

$$|(B(u_h - u, u_h), \varphi)| \leq \|u_h - u\|_{L^4(\Omega)} \|\nabla u_h\|_{L^4(\Omega)} \|\varphi\|_0.$$

This gives

$$\begin{aligned} \|B(u_h - u, u_h)\|_0 &\leq C\|u_h - u\|_{H^{1/2}(\Omega)} \|u\|_{H^{3/2}(\Omega)} \\ &\leq C\|u_h - u\|_1 \|u\|_{H^{3/2}(\Omega)} \end{aligned}$$

after using (4.4) and (2.2). Finally, using (2.9),

$$|(B(u, u_h - u), \varphi)| \leq C\|u\|_\infty \|u - u_h\|_1 \|\varphi\|_0, \tag{4.5}$$

and then, using (2.1), we get

$$\|B(u, u_h - u)\|_0 \leq C\|u\|_{H^{1+\epsilon}(\Omega)} \|u - u_h\|_1, \qquad \epsilon > 0.$$

This completes the proof of (4.1).

The proof of (4.2) follows along the same lines as before. We begin with the case $l = -1$. Using (2.2) we get

$$\begin{aligned} |(B(u_h - u, u_h), \varphi)| &\leq \|u - u_h\|_0 \|\nabla u_h\|_{L^3(\Omega)} \|\varphi\|_{L^6(\Omega)} \\ &\leq C\|u - u_h\|_0 \|u\|_{H^{3/2}(\Omega)} \|\varphi\|_1, \end{aligned}$$

where we have used the estimate

$$\|\nabla u_h\|_{L^3(\Omega)} \leq C\|u\|_{H^{3/2}(\Omega)}. \tag{4.6}$$

This bound is obtained using again (2.8), (2.4), (2.5), and (2.2) in the following way:

$$\begin{aligned} \|\nabla u_h\|_{L^3(\Omega)} &\leq \|\nabla(u_h - I_h u)\|_{L^3(\Omega)} + \|\nabla(I_h u - u)\|_{L^3(\Omega)} + \|\nabla u\|_{L^3(\Omega)} \\ &\leq Ch^{-1/2} \|\nabla(u_h - I_h u)\|_{L^2(\Omega)} + C\|u\|_{W^{1,3}(\Omega)} + \|\nabla u\|_{L^3(\Omega)} \\ &\leq Ch^{-1/2} \left(\|\nabla(u_h - u)\|_{L^2(\Omega)} + \|\nabla(u - I_h u)\|_{L^2(\Omega)} \right) + C\|u\|_{W^{1,3}(\Omega)} \\ &\leq Ch^{-1/2} h^{1/2} \|u\|_{H^{3/2}(\Omega)} + C\|u\|_{W^{1,3}(\Omega)} \\ &\leq C\|u\|_{H^{3/2}(\Omega)}. \end{aligned}$$

For the second term in (4.3) we use first (2.9) and then (2.1) and (2.2):

$$\begin{aligned} |(B(u, u_h - u), \varphi)| &\leq |(B(u, \varphi), u_h - u)| + |(\mathrm{div}(u)(u_h - u), \varphi)| \\ &\leq \|u - u_h\|_0 \left(\|\varphi\|_1 \|u\|_\infty + \|\nabla u\|_{L^3(\Omega)} \|\varphi\|_{L^6(\Omega)} \right) \\ &\leq C\|u - u_h\|_0 \|u\|_{H^{3/2+\epsilon}(\Omega)} \|\varphi\|_1, \ \epsilon > 0. \end{aligned}$$

The case $l = 0$ can be treated in a similar way as in the proof of (4.1). The only difference is that now, after using (4.6) and (2.2), we have

$$\begin{aligned} |(B(u_h - u, u_h), \varphi)| &\leq \|u_h - u\|_{L^6(\Omega)} \|\nabla u_h\|_{L^3(\Omega)} \|\varphi\|_0 \\ &\leq C\|u_h - u\|_1 \|u\|_{H^{3/2}(\Omega)} \|\varphi\|_0 \end{aligned}$$

so that in this case also

$$\|B(u_h - u, u_h)\|_0 \leq C\|u_h - u\|_1 \|u\|_{H^{3/2}(\Omega)}.$$

Finally, using (4.5) and (2.1) we obtain

$$\|B(u,u_h-u)\|_0 \le C\|u_h-u\|_1\|u\|_{H^{3/2+\epsilon}(\Omega)}, \qquad \epsilon>0,$$

which completes the proof. ∎

LEMMA 4.3. *Let u be the solution of* (1.3) *with initial condition u_0 and let u_h be the Galerkin linear finite element approximation to u. Then, if $d=2$, the following bound holds with $l=-1,0$:*

$$\|g(u_h)-g(u)\|_l \le C(\|u\|_1)\|u-u_h\|_{l+1}. \tag{4.7}$$

If $d=3$, for $l=-1,0$,

$$\|g(u_h)-g(u)\|_l \le C(\|u\|_{H^{3/2}(\Omega)})\|u-u_h\|_{l+1}. \tag{4.8}$$

Proof. Let us first observe that

$$g(u_h)-g(u)=(u_h-u)\left(\sum_{j=1}^{2p-1} b_j\left(\sum_{k=1}^{j} u_h^{j-k}u^{k-1}\right)\right)=(u_h-u)G(u_h,u).$$

Since, for $d=2,3$, using (2.2),

$$\begin{aligned}|(g(u_h)-g(u),\varphi)| &= |((u_h-u)G(u_h,u),\varphi)|\\ &\le \|u_h-u\|_0\|G(u_h,u)\|_{L^4(\Omega)}\|\varphi\|_{L^4(\Omega)}\\ &\le \|u_h-u\|_0\|G(u_h,u)\|_{L^4(\Omega)}\|\varphi\|_1\end{aligned}$$

and

$$\begin{aligned}\|G(u_h,u)\|_{L^4(\Omega)} &\le \sum_{j=1}^{2p-1}|b_j|\sum_{k=1}^{j}\|u_h^{j-k}u^{k-1}\|_{L^4(\Omega)}\\ &\le \sum_{j=1}^{2p-1}|b_j|\sum_{k=1}^{j}\|u_h\|^{j-k}_{L^{8(j-k)}(\Omega)}\|u\|^{k-1}_{L^{8(k-1)}(\Omega)},\end{aligned}$$

the estimate (4.7) with $l=-1$ is readily obtained by taking into account that, after using (2.2), $\|u_h\|_{L^{8(j-k)}(\Omega)}$ and $\|u\|_{L^{8(k-1)}(\Omega)}$ are bounded in terms of $\|u_h\|_1$ and $\|u\|_1$, respectively, and then $\|G(u_h,u)\|_{L^4(\Omega)}\le C(\|u\|_1)$.

To obtain (4.8), we again use (2.2) to get

$$\|u\|_{L^{8(k-1)}(\Omega)} \le C\|u\|_{H^{3/2}(\Omega)}.$$

Now using (2.4), (2.5), (2.8), and (2.2), we have, for $q=8(j-k)$,

$$\begin{aligned}\|u_h\|_{L^q(\Omega)} &\le \|u_h-I_hu\|_{L^q(\Omega)}+\|I_hu-u\|_{L^q(\Omega)}+\|u\|_{L^q(\Omega)}\\ &\le Ch^{-3/2}\|u_h-I_hu\|_{L^2(\Omega)}+C\|u\|_{L^q(\Omega)}+\|u\|_{L^q(\Omega)}\\ &\le Ch^{-3/2}h^{3/2}\|u\|_{H^{3/2}(\Omega)}+C\|u\|_{H^{3/2}(\Omega)},\end{aligned}$$

which gives $\|G(u_h,u)\|_{L^4(\Omega)}\le C(\|u\|_{H^{3/2}(\Omega)})$. This completes the proof in the case $l=-1$. The corresponding results for $l=0$ are obtained in a similar way, taking into account that

$$|(g(u_h)-g(u),\varphi)| \le \|u_h-u\|_{L^4(\Omega)}\|G(u_h,u)\|_{L^4(\Omega)}\|\varphi\|_0$$

and again using (2.2). ∎

We now state the first main result of the paper which yields a superconvergence result. More precisely, we prove that the $H^1(\Omega)$ norm of the difference between the elliptic projection of the exact solution and the Galerkin approximation to this solution is $O(h^2)$, up to a logarithmic term.

THEOREM 4.4. *Fix $T > 0$, let u be the solution of* (1.2) *or* (1.3) *with initial condition u_0, let u_h be its Galerkin linear finite element approximation* (2.7)*, and let $r_h = R_{h,2}(u)$ be the elliptic projection of u onto $S_{h,2}$. Then*

$$\max_{0\le t\le T} \|u_h - r_h\|_1 \le C(K(u))\nu^{-1}\log(1/h)h^2,$$

where $K(u)$ is the constant in (2.3).

Proof. It is easy to see that the error $e_h = u_h - r_h$ satisfies, for all $\varphi_h \in S_{h,2}$,

$$((e_h)_t, \varphi_h) = -\nu(\nabla e_h, \nabla\varphi_h) + (F(u_h) - F(u), \varphi_h) + (\tau, \varphi_h), \tag{4.9}$$

where $\tau = (r_h)_t - u_t$. Hence,

$$\begin{aligned} e_h(t) = e^{-\nu A_h t}e_h(0) &+ \int_0^t e^{-\nu A_h(t-s)}P_{h,2}(F(u_h) - F(u))\,\mathrm{d}s \\ &+ \int_0^t e^{-\nu A_h(t-s)}P_{h,2}\tau\,\mathrm{d}s. \end{aligned} \tag{4.10}$$

Taking into account that $e_h(0) = 0$ and applying $A_h^{1/2}$ to both sides of (4.10), we obtain

$$\begin{aligned} A_h^{1/2}e_h(t) = &\int_0^t e^{-\nu A_h(t-s)}A_h^{1/2}P_{h,2}(F(u_h) - F(u))\,\mathrm{d}s \\ &+ \int_0^t e^{-\nu A_h(t-s)}A_h^{1/2}P_{h,2}((r_h)_t - u_t)\,\mathrm{d}s. \end{aligned} \tag{4.11}$$

On the other hand, it is clear that for every function $f \in L^2(\Omega)$ one has

$$\|A_h^{-1/2}P_{h,2}f\|_0 \le C\|f\|_{-1}.$$

Then, using Lemma 4.1, we have

$$\begin{aligned} \left\|\int_0^t e^{-\nu A_h(t-s)}A_h^{1/2}P_{h,2}f\,\mathrm{d}s\right\|_0 &= \left\|\int_0^t e^{-\nu A_h(t-s)}A_hA_h^{-1/2}P_{h,2}f\,\mathrm{d}s\right\|_0 \\ &\le C\nu^{-1}\log(1/h)\|A_h^{-1/2}P_{h,2}f\|_0 \\ &\le C\nu^{-1}\log(1/h)\|f\|_{-1}. \end{aligned}$$

Applying the above inequality to the terms on the right-hand side of (4.11) we obtain

$$\begin{aligned} \|A_h^{1/2}e_h\|_0 &\le C\nu^{-1}\log(1/h)\left(\|F(u_h) - F(u)\|_{-1} + \|(r_h)_t - u_t\|_{-1}\right) \\ &\le C\nu^{-1}\log(1/h)\left(C(K(u))\|u_h - u\|_0 + \|(r_h)_t - u_t\|_0\right), \end{aligned}$$

where in the last inequality we have used Lemma 4.2 or Lemma 4.3. Finally, by (2.8) and (2.6) we get

$$\|A_h^{1/2}e_h\|_0 \le C\nu^{-1}\log(1/h)C(K(u))h^2,$$

which proves the theorem, on taking into account that since $e_h \in H_0^1(\Omega)$ the norm $\|A_h^{1/2} e_h\|$ is equivalent to $\|e_h\|_1$. ∎

We next state and prove another technical lemma and then the main theorem of the paper that gives the rate of convergence of the postprocessed method.

LEMMA 4.5. *Let u be the solution of* (1.2) *or* (1.3) *with initial condition u_0 and u_h its Galerkin linear finite element approximation defined by* (2.7). *Then*

$$\max_{0\le t\le T} \|(u_h)_t - u_t\|_0 \le C(K(u)) \log(1/h) h.$$

Proof. Let us first observe that $\|(u_h)_t - u_t\|_0 \le \|(e_h)_t\|_0 + \|(r_h)_t - u_t\|_0$. The second term, using (2.6), is bounded by $CK(u)h^2$. We now consider the first term. From (4.9) and taking $\varphi_h = (e_h)_t$ we obtain

$$\begin{aligned}\|(e_h)_t\|_0^2 &\le C\nu \|e_h\|_1 \|(e_h)_t\|_1 + \|F(u_h) - F(u)\|_0 \|(e_h)_t\|_0 + \|(r_h)_t - u_t\|_0 \|(e_h)_t\|_0 \\ &\le C\|(e_h)_t\|_0 \left(\nu \|e_h\|_1 h^{-1} + \|F(u_h) - F(u)\|_0 + \|(r_h)_t - u_t\|_0\right),\end{aligned}$$

where we have used the inverse inequality (2.4). Then

$$\|(e_h)_t\|_0 \le Ch^{-1}\nu \|e_h\|_1 + \|F(u_h) - F(u)\|_0 + \|(r_h)_t - u_t\|_0.$$

Now, taking into account Theorem 4.4, Lemma 4.2 or Lemma 4.3, (2.8), and (2.6) we get

$$\|(e_h)_t\|_0 \le C(K(u)) \log(1/h) h + C(K(u))h + CK(u)h^2,$$

which is the desired conclusion. ∎

THEOREM 4.6. *Fix $T > 0$, let u be the solution of* (1.2) *or* (1.3) *with initial condition u_0, and let $\tilde{u}_h(T) = u_h(T) + \Phi_h(u_h(T)) \in \widetilde{S}_h$ be the postprocessed approximation to $u(T)$ defined in* (3.1). *If $\widetilde{S}_h = S_{h',2}$, then*

$$\|u(T) - \tilde{u}_h(T)\|_1 \le C(K(u))\nu^{-1} \log(1/h) h^2 + C\|u(\cdot, T)\|_{H^2(\Omega)} h'. \tag{4.12}$$

If $u(\cdot, T) \in H^3(\Omega) \cap H_0^1(\Omega)$ and $\widetilde{S}_h = S_{h,3}$, then

$$\|u(T) - \tilde{u}_h(T)\|_1 \le C(K(u))\nu^{-1} \log(1/h) h^2 + C\|u(\cdot, T)\|_{H^3(\Omega)} h^2. \tag{4.13}$$

Proof. From the decomposition $\widetilde{S}_h = S_{h,2} \oplus W_h$ (see Remark 3.1) we define Ψ_h such that $\tilde{r}_h = r_h + \Psi_h$, where $\tilde{r}_h$ stands for the elliptic projection of u over the space $\widetilde{S}_h$. Let us denote $\Phi_h(u_h(T))$ by Φ_h. Then the error of the postprocessed approximation can be decomposed in the following way:

$$\|u - (u_h + \Phi_h)\|_1 \le \|u - \tilde{r}_h\|_1 + \|r_h - u_h\|_1 + \|\Psi_h - \Phi_h\|_1. \tag{4.14}$$

The second term on the right-hand side above was estimated in Theorem 4.4.

To bound the third term we first observe that

$$\nu(\nabla \Psi_h, \nabla \psi_h) = -(F(u), \psi_h) - (u_t, \psi_h) + (f, \psi_h) \quad \forall \psi_h \in W_h. \tag{4.15}$$

Subtracting (4.15) from (3.2) we obtain

$$\nu(\nabla(\Phi_h - \Psi_h), \nabla \psi_h) = (F(u) - F(u_h), \psi_h) + ((u - u_h)_t, \psi_h)$$

for all $\psi_h \in W_h$. Using the coercivity of $a(\cdot,\cdot)$,

$$\nu\alpha\|\Phi_h - \Psi_h\|_1 \leq \|F(u) - F(u_h)\|_{-1} + Ch\|(u_h - u)_t\|_0. \tag{4.16}$$

In this last inequality we used that

$$\|w\|_0 \leq Ch\|w\|_1 \quad \forall w \in W_h. \tag{4.17}$$

To prove (4.17) we observe that if $w \in W_h$, then $w = w - R_{h,2}(w)$ so that (4.17) follows from (2.6). Now, using Lemma 4.2 or Lemma 4.3 and Lemma 4.5 in (4.16) we get

$$\begin{aligned}\nu\alpha\|\Phi_h - \Psi_h\|_1 &\leq C(K(u))\|u - u_h\|_0 + C(K(u))\log(1/h)h^2 \\ &\leq C(K(u))h^2 + C(K(u))\log(1/h)h^2\end{aligned}$$

after using (2.8) in the last inequality.

To complete the proof it remains only to bound the first term of (4.14). We distinguish two cases. If $\widetilde{S}_h = S_{h',2}$, then using (2.6) this term is bounded by $C\|u(\cdot,T)\|_{H^2(\Omega)}h'$ and (4.12) is obtained. In the case $\widetilde{S}_h = S_{h,3}$ we use the well-known bound [5]

$$\|u - R_{h,3}(u)\|_1 \leq Ch^2\|u\|_{H^3(\Omega)}, \tag{4.18}$$

instead of (2.6), to deduce that this term is bounded by $C\|u(\cdot,T)\|_{H^3(\Omega)}h^2$ so that the desired estimate (4.13) is obtained. □

Remark 4.1. We observe from Theorem 4.6 that the highest possible rate of convergence in the $H^1(\Omega)$ norm for the postprocessed method is $O(h^2)$ up to a logarithmic term. Then using the postprocessing technique we can obtain an improvement over the standard linear finite element error bound which is $O(h)$ in the $H^1(\Omega)$ norm (see (2.8)). However, to obtain this rate of convergence by postprocessing in the space $S_{h',2}$, that is, by means of some refinement of the mesh, we need $h' = O(h^2)$. This is a very demanding requirement especially in two or three dimensions. Fortunately, for practical computations, only a slightly refined mesh is usually sufficient in order to observe a considerable reduction of the error in the $H^1(\Omega)$ norm (see the next section). Since the refinement of the mesh is performed only at the final time T, the extra cost of the postprocessing is almost negligible when compared with the cost of integrating in time from $t = 0$ to $t = T$ on the coarser mesh.

When the solution u is smooth enough we can use quadratic polynomials over the same grid once the time integration is completed. In view of (4.13) this procedure always improves the rate of convergence of the standard Galerkin method by a single power of h (up to a logarithmic term).

For a general domain with smooth boundary, one can still benefit from using the postprocessed method. In this case, each $\mathcal{T}_h$ is a partition of a domain Ω_h that approximates Ω. For an affine family of finite elements

$$\delta(h) = \max_{x\in\Omega_h} \operatorname{dist}(x, \partial\Omega) \leq ch^2$$

so that (2.6) remains valid; see [14, p. 61]. Therefore, it is possible to obtain exactly the same bounds as in Theorem 4.6, postprocessing with $\widetilde{S}_h = S_{h',2}$. Postprocessing with piecewise quadratic polynomials requires that boundary data approximations are

performed to a sufficiently high order of accuracy since the bound (4.18) is no longer valid in a straight mesh.

Finally, we remark that the one-dimensional case can be treated in a much easier way. In this case one can either refine the partition or increase the degree of the polynomials used in the postprocessing step at the final time level.

5. Numerical experiments. In this section we present some numerical experiments to assess some features of the postprocessed linear finite element method. We begin with a simple one-dimensional example which will help us to illuminate the results we have presented in previous sections.

Let us consider the one-dimensional Burgers equation

$$u_t - \nu u_{xx} + u u_x = 0, \qquad 0 < x < 1,$$

subject to homogeneous Dirichlet boundary conditions $u(0,t) = u(1,t) = 0$, $t \geq 0$. We take as the initial condition $u(x,0) = \sin(2\pi x)$. The value of ν in these experiments is $\nu = 0.01$, but similar results have been obtained for other values of the diffusion parameter.

We use linear finite elements in space over a uniform partition of $[0,1]$ and integrate in time using the *backward differentiation formulae* implemented in the variable-step variable-order mode. The final time chosen is $T = 1$. The Galerkin approximation at the final time is postprocessed using linear finite elements but over a refined partition. If the Galerkin approximation has been calculated over a partition of size h, then $h' \approx h^2$ has been taken as the new diameter of the refined partition for the postprocessing step.

The "exact" solution u has been computed using the Galerkin method with a sufficiently small mesh size and a sufficiently small value of the tolerance in the time integrator in order to ensure that the errors in time and space are negligible. Then the computed solution may be taken as exact, for all intents and purposes, as it is sufficiently more accurate than those shown in the experiments.

In Figure 5.1 we show, in a logarithmic scale, the $L^2(0,1)$ and $H^1(0,1)$ norms of the relative errors for the linear finite element method (continuous line) and the postprocessed method (discontinuous line) against the number $N = 1/h$ of subintervals in the partition used in the calculations. We have used circles to represent the errors measured in the $L^2(0,1)$ norm and diamonds for the errors in the $H^1(0,1)$ norm. For each value of h the experiments were carried out with decreasing values of the time integration tolerance until further reduction of the tolerance did not reduce the error any further. This means that, at this point, the time discretization error is negligible when compared with the spatial error. In the figure it is observed that the postprocessed method *does not reduce the error in the* $L^2(0,1)$ *norm*, while *a considerable reduction of the error is obtained in the* $H^1(0,1)$ *norm*. In fact, the slopes of the lines in the convergence diagram in the $L^2(0,1)$ norm are -1.99 for the standard Galerkin scheme and -1.95 for the postprocessed Galerkin method. That is, the order of convergence in $L^2(0,1)$ is the same for both methods. The slopes in the $H^1(0,1)$ norm are -1.01 and -1.94 for the standard and postprocessed Galerkin method, respectively. This result is in agreement with the bound on the error of the postprocessed method obtained in Theorem 4.6 that predicts quadratic convergence in $H^1(0,1)$ in contrast to the linear convergence of the Galerkin method in the same norm. Notice that, as we mentioned before, the value of h' in this experiment is taken to be $O(h^2)$.

In Figure 5.2 we show the same errors as before (for the $H^1(0,1)$ norm only)

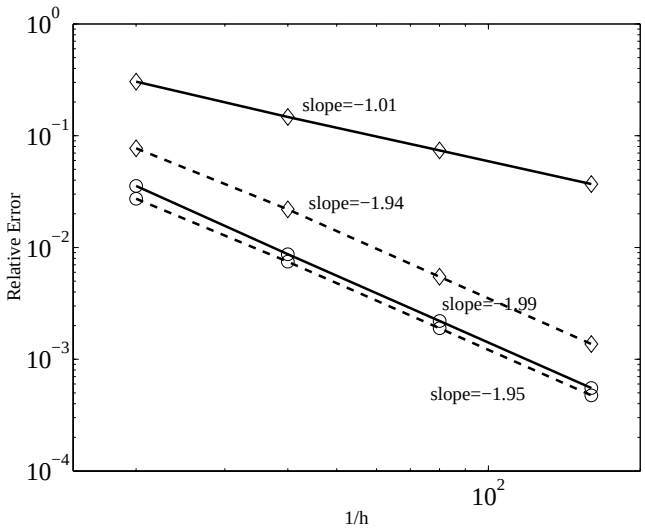


FIG. 5.1. *Burgers equation, $\nu = 0.01$; relative errors against $N = 1/h$; continuous line: Galerkin method, discontinuous line: postprocessed method; circle: $L^2(\Omega)$ norm; diamond: $H^1(\Omega)$ norm.*

but plotted against the CPU time that was required to compute the corresponding solutions. The CPU time shown was the smallest among those yielding a given error but using different values of the time integration tolerance. It can be seen that, for a given accuracy, the postprocessed method achieves a dramatic reduction in the CPU time required to compute the corresponding solution. Thus in this experiment the postprocessing procedure really improves the efficiency of the standard linear finite element method in the $H^1(0,1)$ norm. We remark that the semidiscrete Galerkin equations (2.7) are a stiff system of ordinary differential equations that must be integrated by means of some implicit time stepping procedure. Thus at each time step one has to solve a nonlinear system of equations by means of some Newton or quasi-Newton iteration. The cost of the postprocessing step is the same as an extra single (quasi-) Newton iteration on the refined mesh. Although in this experiment the mesh size we have used to postprocess is much smaller than the one used to obtain the Galerkin solution, the most expensive part of the algorithm is the time integration procedure which is carried out on the coarse mesh. Furthermore, as we show in the next experiment, usually only a slightly refined mesh is sufficient for ensuring a substantial reduction in the error. This fact greatly reduces in practice the cost of the postprocessing step.

Next we present a two-dimensional example. We now consider the following reaction-diffusion equation with a cubic nonlinearity:

$$u_t - \nu\Delta u - u + u^3 = f(t,\mathbf{x}), \qquad \mathbf{x} \in \Omega,$$

in the domain $\Omega = (-1,1) \times (-1,1)$. We impose a homogeneous Dirichlet boundary condition on u. The value of ν in the experiment is $\nu = 0.01$ and the final time is $T = 1$. We chose the forcing term so that the exact solution is $u(x,y,t) = \cos(2\pi t)\sin(\pi x)\sin(\pi y)$.

In the calculations we took a regular triangulation of Ω induced by the set of nodes $(-1+2j/N, -1+2k/N)$, $0 \le j,k \le N$, where $N = 2/h$ is an integer. The meshes were

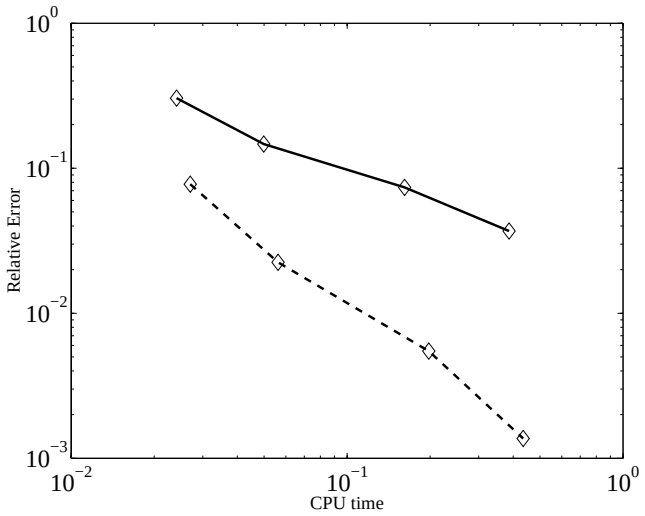


FIG. 5.2. *Burgers equation, $\nu = 0.01$; relative errors in the $H^1(\Omega)$ norm against CPU time; continuous line: Galerkin method, discontinuous line: postprocessed method.*

generated using the finite element package of MATLAB. We use the MATLAB time integrator ODE15s for the time integration of the semidiscrete Galerkin equations.

In Figure 5.3 we show the $L^2(\Omega)$ and $H^1(\Omega)$ norms of the relative errors for the Galerkin and postprocessed Galerkin methods using the same symbols as before: a continuous line for the Galerkin method and a discontinuous (dashed or dash-dotted) line for the postprocessed method, circles for the $L^2(\Omega)$ norm, and diamonds for the $H^1(\Omega)$ norm, respectively. In the postprocessing step we first construct a uniform refinement of the partition taking $h' = h/8$, that is, three regular refinements of the mesh used in the Galerkin equations. We observe in the figure that the postprocessing procedure does not improve the error in the $L^2(\Omega)$ norm. Indeed, for each value of h, the $L^2(\Omega)$ norm of the error of the postprocessed method (dashed line, circles) is significantly larger than the corresponding error of the Galerkin method (continuous line, circles), although both exhibit quadratic convergence. The slopes in the convergence diagram are -1.98 for the $L^2(\Omega)$ norm of the Galerkin error and -1.89 for the $L^2(\Omega)$ norm of the postprocessed error.

If we consider the errors achieved in the $H^1(\Omega)$ norm the situation is completely different. Now the postprocessed errors (dashed line, diamonds) are below the Galerkin errors (continuous line, diamonds) except for the first point that corresponds to $h = 2/8$. Furthermore, the difference between Galerkin and postprocessed Galerkin errors increases as the value of h decreases. The computed slope for the line representing the postprocessed errors in the $H^1(\Omega)$ norm is -1.83 showing that the convergence is almost quadratic. This result is again in agreement with the error bounds of Theorem 4.6. Indeed, we could have obtained exactly quadratic convergence with the postprocessed method had we postprocessed the last two points in the picture using a smaller value of h'. However, from a practical point of view, although the postprocessed step is carried out only at the final time T, the use of a mesh of size $h' \approx h^2$ for ensuring quadratic convergence is not very useful. Fortunately, one still benefits from using the postprocessed method in two dimensions even if $h' \gg h^2$. To show this fact we have depicted in Figure 5.3, using diamonds joined by a dashed-

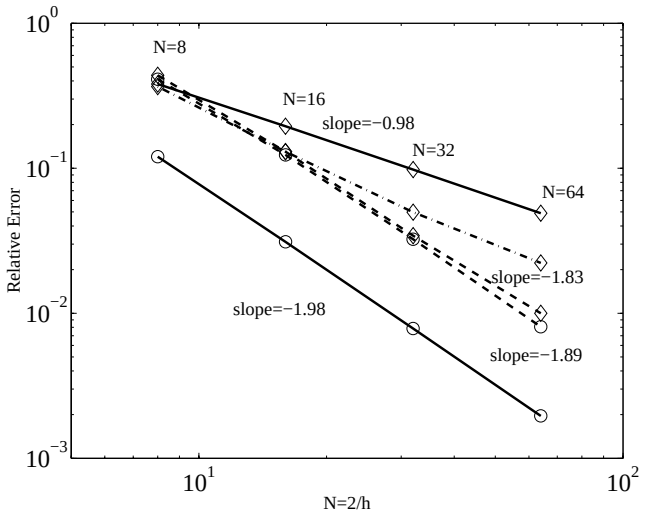


FIG. 5.3. *Reaction-diffusion equation, $\nu = 0.01$; relative errors against $N = 2/h$; continuous line: Galerkin method, discontinuous line: postprocessed method; circle: $L^2(\Omega)$ norm, diamond: $H^1(\Omega)$ norm.*

dotted line, the $H^1(\Omega)$ norm of the errors in the postprocessed solution based on linear elements over a refined mesh of size $h' = h/2$ (only one regular refinement). Notice then that by postprocessing at the final time $T = 1$ the Galerkin solution that has been computed using a mesh of size $h = 2/16$ we obtain the same error as if we had used a mesh of size $h = 2/32$ over the full time interval $[0, T]$ to calculate the Galerkin solution. The same result is true for other values of h; compare, for example, the results with $h = 2/32$ (postprocessed with $h' = 2/64$) and $h = 2/64$. *Postprocessing the Galerkin approximation with $h' = h/2$ one can achieve a smaller error than by computing the Galerkin solution on a mesh of half the spacing.* This version of applying the postprocessed method is clearly an efficient way to improve the accuracy of the Galerkin method in the $H^1(\Omega)$ norm. Notice that now the cost of the postprocessing step is only slightly larger than that of one (quasi-) Newton iteration in one step of the time integration procedure, and, furthermore, the postprocessed method is applied only once at time T where the numerical solution is required.

Acknowledgment. The authors thank Professor Endre Süli for his valuable suggestions, especially concerning the proof of Lemma 4.2, which is a significant improvement over a previous version.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, 1976.
[3] C. BERNARDI, *Optimal finite-element interpolation on curved domains*, SIAM J. Numer. Anal., 26 (1989), pp. 1212–1240.
[4] C. CANUTO, M. Y. HUSSAINI, A. QUARTERONI, AND T. A. ZANG, *Spectral Methods in Fluid Dynamics*, Springer Ser. Comput. Phys., Springer-Verlag, Berlin, 1988.
[5] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 17–351.

[6] P. Constantin and C. Foias, *Navier-Stokes Equations*, Chicago Lectures in Math., University of Chicago Press, Chicago, IL, 1988.

[7] J. de Frutos, B. García-Archilla, and J. Novo, *A postprocessed Galerkin method with Chebyshev and Legendre polynomials,* Numer. Math., 86 (2000), pp. 419–442.

[8] J. de Frutos and J. Novo, *A spectral element method for the Navier–Stokes equations with improved accuracy,* SIAM J. Numer. Anal., 38 (2000), pp. 799–819.

[9] J. de Frutos and J. Novo, *A postprocess based improvement of the spectral element method*, Appl. Numer. Math., 33 (2000), pp. 217–223.

[10] B. García-Archilla, J. Novo, and E. S. Titi, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds,* SIAM J. Numer. Anal., 35 (1998), pp. 941–972.

[11] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case,* SIAM J. Numer. Anal., 37 (2000), pp 470–499.

[12] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

[13] A. H. Schatz and L. B. Wahlbin, *On the quasi-optimality in L^∞ of the H^1-projection into finite element spaces,* Math. Comput., 38 (1982), pp. 1–21.

[14] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, Berlin, 1995.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 820–871

STABILITY OF DISCRETE SHOCKS FOR DIFFERENCE APPROXIMATIONS TO SYSTEMS OF CONSERVATION LAWS*

DANIEL MICHELSON†

Abstract. The asymptotic stability of weak discrete stationary shocks for systems of conservation laws in one space dimension is proved. The difference approximation should be conservative, dissipative, and kth order accurate in space with odd k. The problem is considered in a finite interval $|x| \le \ell$ with appropriate boundary conditions, where ℓ is large compared with the width of the shock layer $\varepsilon^{-1} = |u_R - u_L|^{-1/k}$. The proof is based on the assumption that the corresponding continuous shocks for the scalar problem $u_t + uu_x = -(i\partial_x)^{k+1}u$ are stable. The latter is known to be true for $k = 1$ and $k = 3$.

Key words. discrete shocks, systems of conservation laws, asymptotic stability, viscous shock profiles, high order dissipation

AMS subject classifications. 65M12, 65M06, 35L65, 35L67

PII. S0036142900377577

1. Introduction.

Consider a system of conservation laws

$$u_t + f(u)_x = 0, \tag{1.1}$$

where $f\colon R^n \to R^n$ is a smooth vector function and the unknown function $u = u(x,t) \in R^n$ depends on $-\infty < x < \infty$ and $t > 0$. By planar stationary shock, one means the solutions

$$u = u_L, \quad x < x_0; \qquad u = u_R, \quad x > x_0, \tag{1.2}$$

where u_L, u_R satisfy the Rankine–Hugoniot condition

$$f(u_L) = f(u_R). \tag{1.3}$$

Suppose that (1.1) is approximated by a dissipative system

$$u_t + f(u)_x = -(i\partial_x)^{\frac{k+1}{2}}\Big[A(u)(i\partial_x)^{\frac{k+1}{2}}\Big]u, \qquad k \text{ odd}. \tag{1.4}$$

Stationary solutions u_{st} of (1.4) which attain the above limits u_L and u_R as $x \to -\infty$ and $x \to \infty$ correspondingly are called stationary viscous shocks. (The traveling shocks by a change of variables $x - c \cdot t \to x$ are reduced to the stationary ones.) If the difference $|u_R - u_L|$ is small, the shock is called weak. We assume that u_L, u_R lie in a small neighborhood of a point u_0 such that the differential $df[u_0]$ of f at u_0 has *distinct real* eigenvalues with a zero eigenvalue $\lambda_1(u_0) = 0$. Without loss of generality, we may assume that

$$df[u_0] = \operatorname{diag}\Big(\lambda_1(u_0), \lambda_2(u_0), \dots, \lambda_n(u_0)\Big). \tag{1.5}$$

*Received by the editors September 5, 2000; accepted for publication (in revised form) February 20, 2002; published electronically August 1, 2002.
http://www.siam.org/journals/sinum/40-3/37757.html
†Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (daniel_michel@wisdom.weizmann.ac.il).

The eigenvalue $\lambda_1(u)$ also should be *genuinely nonlinear*, i.e., the directional derivative

$$d\lambda_1[u_0] \cdot e_1 \neq 0, \qquad e_1 = (1, 0, \ldots, 0)^T. \tag{1.6}$$

We assume that the pair u_L, u_R is entropy-satisfying, i.e.,

$$\lambda_1(u_L) > 0 > \lambda_1(u_R), \tag{1.7}$$

and the right-hand side (r.h.s.) of (1.4) is dissipative in a proper sense (see the dissipativity assumption in section 2). Under the above assumptions, the existence of viscous shocks for $k = 1$ and $k = 3$ follows from [1]. Using the idea of Conley's index [2] and the reduction to the central manifold as in [1], one can actually prove the existence of viscous shocks for any odd k.

A natural question to ask is whether these shocks are asymptotically stable. Namely, given an initial condition $u(x, 0)$ which is a small perturbation of the stationary viscous shock $u_{st}(x)$, will $u(x, t)$ tend to $u_{st}(x)$ as $t \to \infty$? For $k = 1$ this problem was studied by Goodman in [3]. Under the assumption of zero mean perturbation, i.e., $\int_{-\infty}^{\infty}(u(x, 0) - u_{st}(x))dx = 0$, he proved that the shock is asymptotically stable. Liu in [4] removed the above restriction but assumed that u_L and u_R are connected by a nondegenerate sum of n shocks. Then he showed that $u(x, t)$ tends asymptotically to a sum of traveling viscous shocks. As far as we know, for $k = 3$ the stability problem was never considered.

Lately, while studying the stability of Bunsen flames [5] we became involved in the stability problem for the shock solution of the scalar equation

$$u_t + \left(\frac{u^2}{2}\right)_x = -u_{xxxx}. \tag{1.8}$$

The stationary shock u_{st} satisfies the equation

$$u_{xxx} = \frac{1}{2}(1 - u^2), \qquad u(\mp\infty) = \pm 1, \tag{1.9}$$

and is not monotone. As a result, the energy method in the L_2 space is not applicable to (1.8). Instead, we used a rigorous computer-assisted proof to show that the linear problem

$$su + (u_{st} \cdot u)_x + u_{xxxx} = 0, \qquad -\infty < x < \infty, \tag{1.10}$$

for complex s with $\operatorname{Re} s \geq 0$ does not have solutions which decrease exponentially as $|x| \to \infty$. (See Theorem 5.1 in [5].) Recently, Engelberg [6] proved analytically that the solution u_{st} of (1.8) is stable in a norm with a weight function $w(x)$ which grows exponentially as $|x| \to \infty$. It seems that our result could be deduced directly from [6] or by strengthening the result in [6]. Unfortunately, the weighted norm estimates are not applicable to systems as in (1.4). Instead, we study the existence and asymptotic stability of a stationary solution u_{st} on a finite interval $|x| \leq \ell$.

In the case $1 < n_1 < n$ (n_1 is defined at (1.14) below), in order to prove the existence of stationary solutions u_{st}, one needs the restriction

$$\delta_0^{-1}\varepsilon^{-1} \leq \ell \leq \delta_0\varepsilon^{-k} \tag{1.11}$$

and $k \geq 3$. (For stability of u_{st} one needs even more, i.e., $\ell \leq \delta_0\varepsilon^{-2}$.) In cases $n_1 = 1$ and $n_1 = n$, ℓ should satisfy

$$\delta_0^{-1}\varepsilon^{-1} \leq \ell \leq \delta_0\varepsilon^{-k-1} \tag{1.12}$$

and $k \geq 1$. Here ε^{-1} is proportional to the width of the boundary layer

$$\varepsilon^{-1} \sim \Big(\lambda_1(u_{_L})\Big)^{-1/k} \sim |u_{_R} - u_{_L}|^{-1/k}, \tag{1.13}$$

and δ_0 is a small constant. The boundary conditions should include

$$P\Big(f(u) + (i\partial_x)^{\frac{k-1}{2}}\Big[A(u)(i\partial_x)^{\frac{k+1}{2}}u\Big]\Big) = 0, \qquad x = \pm\ell, \tag{1.14}$$

where $P\colon R^n \to R^n$ is a projector on the first n_1 components of u. This implies that the integral $\int_{-\ell}^{\ell} Pudx$ is conserved in time. Note that u_{st} tends exponentially fast to $u_{_L}$, $u_{_R}$ with respect to the variable εx. Thus the boundaries $x = \pm\ell$ are practically at infinity.

In this paper we will prove asymptotic stability only in the cases $n_1 = 1$ and $n_1 = n$ since, in the case $1 < n_1 < n$, the result is much weaker and requires special treatment. However, the existence of a stationary solution u_{st} will be proved for a general n_1. The perturbation u of u_{st} should have the same mean value $\int_{-\ell}^{\ell} Pudx$ as u_{st}. But this is not a restriction on the perturbation since there is an n_1 parameter family of u_{st} depending on this mean value (see (2.89) in what follows).

The real motivation for the shock stability problem comes from computations. Hence, instead of the continuous model (1.4), we will consider a discrete approximation to (1.1). The continuous problem could be solved in a similar way. The discrete setting will follow the one in [7], where we proved the existence of weak discrete shocks for dissipative approximations with $k = 1$ and 3. Our numerical experiments for a 2×2 system of polytropic gas indicate that u_{st} also is stable for strong shocks if $k = 1$ or $k = 3$. We conjecture that the same holds for all odd k. Thus, we consider in this paper the case of a general odd k, under the assumption that the eigenvalue problem in (1.10) with $(i\partial_x)^{k+1}u$ derivative has no eigensolutions with $\operatorname{Re} s \geq 0$. The precise statement of our results will be given in the next section.

One should mention the most recent papers on the subject of stability of viscous shocks: Kreiss and Kreiss [10], Zumbrun and Howard [11], and Liu and Yu [12]. All three papers deal with the case of second order viscosity only. In [10] and [11] the continuous problem is studied. In [10] the Jacobian of the flux is assumed to be the same at $\pm\infty$, which does not hold in real physical applications. The existence and linear stability of the shock is also postulated. In [11, p. 760] the Evans function criterion (condition D) is postulated. It holds indeed in the case of weak shocks; however, it is not clear whether the estimates in [11] give an efficient bound on the size of the perturbation of the stationary shock as the strength of the shock tends to 0. The paper of Liu and Yu [12] is closest to our results. Their estimates, unlike ours, hold on the whole line. They also prove existence of moving shocks while we consider only stationary ones. However, their uniform dissipativity assumption is much stronger than our "characteristic-component" dissipativity assumption in (2.16) below. It would be a great challenge to extend the results of [12] to the case of fourth order dissipation and to the boundary value problems.

Let us conclude this section with a practical remark. The most popular numerical schemes for shock wave computations produce monotone shock profiles. As such, they exclude high order approximations unless the scheme is modified outside the shock layer. Our result indicates that the high order schemes, although producing one overshoot of about 20% in the shock layer, converge exponentially fast to the stationary solution. Since the special oscillations have a canonical form, they could be filtered out at the end of computation without sacrificing the overall high accuracy.

2. The difference approximation. The system of conservation laws (1.1) is approximated by a difference scheme

$$G\Big(\big\{E^j u(x,t)\big\}\Big) = 0, \qquad j = (j_1, j_2) \in J \subset \mathbb{Z}^2, \qquad j_2 \le 0. \tag{2.1}$$

Here J is a finite set; $G = G(\{u_j\})$ is a smooth (at least C^3) vector function of vector variables $u_j \in R^n$, $j \in J$; and $E^j = E_x^{j_1} E_t^{j_2}$ is a shift operator

$$E^j u(x,t) = u(x + j_1 h, t + j_2 h). \tag{2.2}$$

We consider $u(x,t)$ in (2.1) as a grid-function defined on a uniform grid with a mesh size h,

$$D_h = I_h \times (R_+)_h, \tag{2.3}$$

in the half-strip $[-\ell, \ell] \times [0, \infty)$. Since the step size h does not enter the problem explicitly, we will assume that $h = 1$ so that 2ℓ is equal to the number of grid points in I_h. The scheme G should be *conservative*, i.e.,

$$G\Big(\big\{E^j u(x,t)\big\}\Big) = (E_x - I) G_1\Big(\big\{E^j u(x,t)\big\}\Big) + (E_t - I) G_2\Big(\big\{E^j u(x,t)\big\}\Big) \tag{2.4}$$

with multi-index j in G_1, G_2 varying over corresponding subsets of J. We assume that G is *consistent* with (1.1), i.e.,

$$G_1\Big(\{u\}\Big) = f(u), \qquad G_2\Big(\{u\}\Big) = u, \tag{2.5}$$

where $\{u\}$ stands for the set $\{u_j\}$ of vectors $u_j = u$. Since G_2 approximates u, it is natural to assume that

$$G_2\Big(\{u_j\}\Big) \text{ is a linear function.} \tag{2.6}$$

(This assumption is not essential but somewhat simplifies the proof.)

Since the domain is bounded, there are boundary conditions

$$S_{_L}\Big(\big\{E^j u(-\ell, t)\big\}, u_{_L}\Big) = 0, \qquad S_{_R}\Big(\big\{E^j u(\ell, t)\big\}, u_{_R}\Big) = 0, \tag{2.7}$$

where $S_{_L}$ and $S_{_R}$ depend smoothly on its arguments and j belongs to a finite set. The range of $j = (j_1, j_2)$ in $S_{_L}$ is such that $j_1 \ge 0$, $j_2 \le 0$, while in $S_{_R}$, $j_1 \le 0$, $j_2 \le 0$. The boundary conditions should be *consistent* with a constant solution, i.e.,

$$S_{_L}\Big(\{u\}, u\Big) \equiv S_{_R}\Big(\{u\}, u\Big) \equiv 0. \tag{2.8}$$

There are also initial conditions

$$u(x,t) = u_{in}(x,t), \qquad x = j_1 \in [-\ell, \ell], \qquad t = j_2 \in [0, \Delta j_2 - 1], \tag{2.9}$$

where $-\Delta j_2$ is the minimal value of the index j_2 in the functions G, $S_{_L}$, and $S_{_R}$. Note that we can always add to G, $S_{_L}$, $S_{_R}$ dummy variables u_j so that they have a common minimal j_2.

We consider a family of initial boundary value problems (IBVPs) that depend on the parameter $\varepsilon \sim (\lambda_1(u_{_L}))^{1/k}$ with $u_{_L}$, $u_{_R}$ satisfying (1.3) and (1.7) so that

$u_{_L}(\varepsilon = 0) = u_{_R}(\varepsilon = 0) = u_0$ and ℓ lies in the bounds of (1.11) or (1.12). More precisely,

$$u_{_L} = u_0 + \mu e_1, \quad u_{_R} = u_0 - \mu e_1 + O(\mu^2), \quad \mu = b\varepsilon^k / \Big(d\lambda_1[u_0] \cdot e_1 \Big), \tag{2.10}$$

where b is the (positive) dissipation coefficient defined in (2.16) below and e_1 is the unit vector as in (1.6).

Equation (2.1) is defined for (x,t) in a subdomain $\mathring{D}_h \subset D_h$ such that $(x+j_1, t+j_2)$ lie in D_h for all $j \in J$. Denote by $x_{-\ell}$, $x_\ell - 1$ the left and right end points of $\mathring{D}_h$, and sum the equation in (2.1) for $x \in [x_{-\ell}, x_\ell - 1]$. We obtain the global conservation law

$$(E_t - I) \sum G_2\Big(\big\{ E^j u(x,t) \big\} \Big) = G_1\Big(\big\{ E^j u(x_{-\ell}, t) \big\} \Big) - G_1\Big(\big\{ E^j u(x_\ell, t) \big\} \Big). \tag{2.11}$$

It would be natural to assume that

$$G_1\Big(\big\{ E^j u(x_{\pm\ell}) \big\} \Big) - f(u_{_{R,L}}) = 0; \tag{2.12}$$

i.e., (2.12) is a part of the boundary conditions in (2.7). More generally, we will assume that

$$P\Big(G_1\Big(\big\{ E^j u(x_{\pm\ell}, t) \big\} \Big) - f(u_{_{R,L}}) \Big) = 0, \tag{2.13}$$

where $P = P(\mu)\colon R^n \to R^n$ is a projector which depends smoothly on μ such that $P(0)$ is the standard projection on the first n_1 components of u. As a result, the solution of IBVPs (2.1), (2.7), (2.9) satisfies

$$P \sum G_2\Big(\big\{ E^j u(x,t) \big\} \Big) = P \sum G_2\Big(\{ E^j u_{in} \} \Big). \tag{2.14}$$

Let us linearize the above IBVP at the constant state $u = u_0$ and $\mu = 0$ with arbitrary ℓ. The resulting constant coefficient difference operators will be denoted by $dG[u_0](E_x, E_t)$, $dS[u_0](E_x, E_t)$, or simply $dG[u_0]$, $dS[u_0]$. The Laplace–Fourier symbol of $dG[u_0]$ is defined as $dG[u_0](e^{i\xi}, e^s)$. As in [7], we impose the following assumptions.

Dissipativity assumption. The symbol $dG[u](e^{i\xi}, e^s)$ is nonsingular for all pairs (ξ, s) with $\operatorname{Re} s \geq 0$, but $s = 0$ and $\xi = 0 \bmod 2\pi$, where u is any vector in a neighborhood of u_0.

Remark 1. Actually, we need the above dissipativity only at $u = u_0$, and the usual stability at u close to u_0.

Accuracy assumption. The difference operator $dG[u_0]$, when restricted to the x variable, is exactly a kth order accurate approximation of $df[u_0]\frac{\partial}{\partial x}$ in the direction $e_1 = (1, 0, \dots, 0)^{^T}$. More precisely,

$$dG[u_0](e^{i\xi}, 1)e_1 = i\xi df[u_0]e_1 + O(\xi^{k+1}) = O(\xi^{k+1}), \tag{2.15}$$

where the first component $O_1(\xi^{k+1})$ of $O(\xi^{k+1})$ satisfies

$$O_1(\xi^{k+1}) = b\xi^{k+1} + O(\xi^{k+2}), \qquad b \neq 0. \tag{2.16}$$

It is easy to show (see Proposition 2.1 in [7]) that the dissipativity assumption then implies $b > 0$. Note that the scheme need not be kth order accurate in t.

Obviously, one should assume that the linear problem

$$dG[u_0]u = F, \qquad dS_{_L}[u_0]u = g_{_L}, \qquad dS_{_R}[u_0]u = g_{_R} \tag{2.17}$$

is solvable in time. Namely, let $dG^{(0)}$, $dS_{_L}^{(0)}$, $dS_{_R}^{(0)}$ be the upper parts of dG, $dS_{_L}$, $dS_{_R}$ in t. We should then consider only the half-line problems

$$dG^{(0)}[u_0]u = F, \quad x \geq -\ell; \qquad dS_{_L}^{(0)}[u_0]u(-\ell) = g \tag{2.18}$$

and

$$dG^{(0)}[u_0]u = F, \quad x \leq \ell; \qquad dS_{_R}^{(0)}[u_0]u(\ell) = g. \tag{2.19}$$

Solvability assumption. For any F and g, problems (2.18) and (2.19) have a unique solution u satisfying the estimate

$$\|u\|_2^2 \leq K\Big(\|F\|_2^2 + |g|^2\Big), \tag{2.20}$$

where $\|\cdot\|_2$ is the ℓ_2 norm and K is independent of F and g. Theorem 1.1 in [8] states an equivalent algebraic (Lopatinsky) condition on $dG^{(0)}$, $dS^{(0)}$. If ℓ is large enough, this implies the Lopatinsky condition for the combined (2.18), (2.19) problem in $[-\ell, \ell]$. The last, in turn, implies estimate (2.20) in norm $\|\cdot\|_p$, $p \geq 1$, for the combined problem. Then, by the inverse function theorem, the nonlinear IBVP (2.1), (2.7), (2.9) is solvable for u in a small neighborhood of u_0.

Finally, let us state the stability condition. Apply to (2.17) the Laplace–Fourier transform in t; i.e., consider

$$dG[u_0](E_x, e^s)u = F, \quad dS_{_L}[u_0](E_x, e^s) = g_{_L}, \quad dS_{_R}[u_0](E_x, e^s)u = g_{_R}. \tag{2.21}$$

The weak Lopatinsky condition. For $\operatorname{Re} s \geq 0$, $s \neq 0$, the corresponding half-line problems have no nontrivial solutions in ℓ_2.

As in the solvability case, for $\operatorname{Re} s \geq 0$ and $|s| > \delta$ we obtain for problem (2.21) a uniform estimate (2.20). (Actually, the norm $\|\cdot\|_2$ could be replaced by any norm $\|\cdot\|_p$, $p \geq 1$.) The crucial point, of course, is $s = 0$. Since $\lambda_1(u_0) = 0$, the boundaries $x = \pm\ell$ are characteristic. We will see in the next section that, for $s = s'\varepsilon^{k+1}$, the homogeneous solution of the linearized shock problem in proper coordinates consists of three parts: y_1, $\{y_i\}$, $2 \leq i \leq n$, and $y_{_{I,II}}$. The first part y_1 is a discrete approximation to a solution of the linearized scalar shock equation

$$s'y_1 + \partial_\tau(y_{sh}y_1) + (-1)^{\frac{k+1}{2}}\partial_\tau^{k+1}y_1 = 0, \tag{2.22}$$

where y_{sh} is the solution of the problem

$$(-1)^{\frac{k+1}{2}}\partial_\tau^k y_{sh} = \tfrac{1}{2}(1 - y_{sh}^2), \qquad y_{sh}(-\infty) = 1, \quad y_{sh}(\infty) = -1, \tag{2.23}$$

and $\tau = \varepsilon x$. The second, noncharacteristic part approximates the solution of the equation

$$sy_i + \lambda_i(u_0)\partial_x y_i = 0, \qquad 2 \leq i \leq n, \tag{2.24}$$

where $\lambda_i(u_0)$ are the nonzero eigenvalues of $df[u_0]$. The third, boundary layer part corresponds to the exponentially decreasing ($y_{_I}$) and increasing ($y_{_{II}}$) solutions of the equation

$$dG_1[u_0](E_x, 1)u = 0. \tag{2.25}$$

The variables y_i, $1 \le i \le n$, modulo $O(s)$ terms affect only u_i components of u, but $y_{I,II}$, due to a possible coupling in the dissipative term, may affect all components of u. The boundary condition is designed so that the problem for $y_- = (\{y_i\}, i \ge 2; y_{I,II})$ almost decouples from the one for y_1. Toward that end we assume that the boundary operators dS_L and dS_R split into three parts,

$$\begin{aligned} &\text{(a)} \quad P dG_1 u = 0, \\ &\text{(b)} \quad dS_1 u = 0, \\ &\text{(c)} \quad dS_2 u = 0. \end{aligned} \tag{2.26}$$

Represent the projector $P(\mu)$ as an orthogonal sum

$$P(\mu) = P_1(\mu) \oplus P_-(\mu) \tag{2.27}$$

such that $P_1(0)$ is the standard projection on the unit vector e_1. The equation $P_1 dG_1 u = 0$ in (2.26)(a) is used to integrate (2.22) with respect to τ. The remaining components of (2.26)(a) are used to specify the hyperbolic variables $y_2, \dots, y_n$ at the inflow boundaries, which are

$$x = -\ell \quad \text{if} \quad \lambda_i(u_0) > 0 \qquad \text{and} \qquad x = \ell \quad \text{if} \quad \lambda_i(u_0) < 0. \tag{2.28}$$

There are $n_1 - 1$ unused conditions in (2.26)(a) which are replaced by the conservation laws

$$P_- \sum_x dG_2[u_0](E_x, E_t) u = 0. \tag{2.29}$$

The operator $dS_1[u_0]$ has exactly $\frac{k-1}{2}$ components at each end $x = \pm\ell$ and supplies the boundary conditions for (2.22). Denote by $dS_{1j}^{(1)}[u_0]$ the entry in the first column and jth row of $dS_1[u_0]$. We assume that for $1 \le j \le \frac{k-1}{2}$,

$$dS_{1j}^{(1)}[u_0] = c_j (E_x - I)^{d_j} + \text{ higher powers of } (E_x - I), \tag{2.30}$$

where

$$\begin{aligned} &0 \le d_j \le k-1 \quad \text{for} \quad n_1 = 1 \quad \text{and} \\ &0 \le d_j \le k-2 \quad \text{for} \quad 1 < n_1 \le n \end{aligned} \tag{2.31}$$

are all distinct and $c_j \ne 0$. The boundary condition (2.26)(b) thus approximates for small ε the conditions $\partial_\tau^{d_j} y_1 = 0$ at $\tau = -\varepsilon\ell$ or $\tau = \varepsilon\ell$. The resulting boundary value problem for (2.22) should satisfy the one-sided Lopatinsky conditions, namely, the following.

The scalar shock stability condition. For complex s with $\operatorname{Re} s \ge 0$ and real $\delta \ge 0$ such that $|s| + \delta \ne 0$, let $\lambda_1, \lambda_2, \dots, \lambda_{\frac{k+1}{2}}$ be the roots of the equation

$$s \pm \delta\lambda + (-1)^{\frac{k+1}{2}} \lambda^{k+1} = 0 \tag{2.32}$$

with corresponding $\pm \operatorname{Re} \lambda \le 0$. Then for all such s, the determinant

$$\det(\lambda_i^{d_j+1}), \qquad 1 \le i \le \frac{k+1}{2}, \quad 0 \le j \le \frac{k-1}{2}, \quad d_0 \stackrel{df}{=} -1 \tag{2.33}$$

is nonzero. (In the case $s = 0$, $\lambda_1 = 0$ and $\lambda_1^{d_0+1} \stackrel{df}{=} 1$.)

This condition is obviously satisfied if $d_j = j - 1$. If λ_i is a multiple root, then besides $\lambda_i^{d_j+1}$ we have also the rows $(d/d\lambda)^m(\lambda^{d_j+1})|_{\lambda=\lambda_i}$, where $1 \le m < r_j$, and r_j is the multiplicity of the root λ_i. Note that the exponents d_j at $x = -\ell$ and $x = \ell$ need not be the same. Due to a scaling, the parameter δ, if nonzero, could be chosen to be $\delta = 1$. Note also that for $k = 1$ the set of boundary conditions (2.26)(b) is empty.

The boundary conditions (2.26)(c) and the conservation laws (2.29) are used to define the boundary layer $y_{I,II}$ and the remaining hyperbolic variables $y_{n_1+1}, \dots, y_n$. We wish the contribution of y_1 to y_- for $s = s'\varepsilon^{k+1}$ to be of order $O(\varepsilon^k)$. Therefore, we assume that the first column $dS_2^{(1)}$ of dS_2 satisfies

$$dS_2^{(1)}[u_0](E_x, 1) \text{ is divisible by } (E_x - 1)^k \tag{2.34}$$

and, similarly,

$$P_-(\mu = 0)dG_2^{(1)}[u_0](E_x, 1) \text{ is divisible by } (E_x - 1)^{k+1}\ . \tag{2.35}$$

The above two conditions (2.34), (2.35) will be called the *decoupling conditions* since they imply that the Lopatinsky condition for the whole solution y decouples into one for y_1 as in (2.22) and one for the y_- variables.

Now consider the homogeneous problem $dG[u_0](E_x, e^s)u = 0$ for small $|s|$ with $\operatorname{Re} s \ge 0$. Denote the solutions that correspond to the noncharacteristic hyperbolic components y_i, $2 \le i \le n$, by φ_i and the ones that correspond to $y_{I,II}$ by $\varphi_I(x)$, $\varphi_{II}(x)$. Thus, if we disregard the contribution of y_1, the general solution u is a linear combination

$$u(x) = \sum_{i=2}^{n} \varphi_i(x)c_i + \varphi_I(x)c_I + \varphi_{II}(x)c_{II}. \tag{2.36}$$

If $\lambda_i(u_0) > 0$, then

$$\begin{aligned} \varphi_i(x)c_i &= X_i(s)\exp\Big(\mu_i(s)(x+\ell)\Big)c_i = X_i(x)y_i(x), \\ X_i(s) &= e_i + O(s), \quad e_i - \text{unit vector}. \end{aligned} \tag{2.37}$$

Here $\mu_i(s) = \frac{-s}{\lambda_i(u_0)+O(s^2)}$ and, by the dissipativity assumption, $\operatorname{Re}\mu_i(s) < 0$. Since s could be arbitrarily small and ℓ arbitrarily large, the factor $\rho_i = \exp(2\ell\mu_i(s))$ could be any complex number with the only restriction $|\rho_i| < 1$. For $\varphi_I c_I$ we have

$$\varphi_I(x)c_I = X_I(s)\Big(M_I(s)\Big)^{x+\ell}c_I = X_I(s)y_I(x), \tag{2.38}$$

where $\|M_I\| < 1$. Similar relations hold for $\varphi_i c_i$ in the case $\lambda_i(u_0) < 0$ and for $\varphi_{II}c_{II}$. One can assume that $X_i(s)$, $X_{I,II}(s)$ depend analytically on s and that the columns φ_i, $\varphi_{I,II}$ are independent as functions of x. Note that $\varphi_I(\ell)$ and $\varphi_{II}(-\ell)$ are exponentially small so that their contribution to the boundary condition will be (formally) neglected. On the other hand, the solutions $\varphi_i(x)$ couple together the boundary conditions at $x = -\ell$ and $x = \ell$. Let us split the hyperbolic variables y_i, $2 \le i \le n$, into four groups $y^{(j_1,j_2)}$, $j_1 = 1, 2$, $j_2 = 1, 2$, where $j_1 = 1$ if $i \le n_1$, and 2 if $i > n_1$, while $j_2 = 1$ if $\lambda_i(u_0) > 0$, and 2 if $\lambda_i(u_0) < 0$. Substitute the solution u in

(2.36) into the boundary conditions $P_- dG_1[u_0]u = 0$ in (2.26)(a). In terms of the y variables, we obtain

$$\begin{aligned}
&\text{(a)} \quad \Big(\lambda^{(1,1)} + O(s)\Big) y^{(1,1)}(-\ell) = sB^{(1)}\Big(y(-\ell); s\Big),\\
&\text{(b)} \quad \Big(\lambda^{(1,2)} + O(s)\Big) y^{(1,2)}(\ell) = sB^{(2)}\Big(y(\ell); s\Big),\\
&\text{(c)} \quad B^{(1)}\Big(y(\ell); s\Big) - \rho^{(1,1)} B^{(1)}\Big(y(-\ell); s\Big) = 0,\\
&\text{(d)} \quad B^{(2)}\Big(y(-\ell); s\Big) - \rho^{(1,2)} B^{(2)}\Big(y(\ell); s\Big) = 0.
\end{aligned} \tag{2.39}$$

Here $\rho^{(1,1)} = \operatorname{diag}(\rho_i)$, $\lambda^{(1,1)} = \operatorname{diag}(\lambda_i(u_0))$, where ρ_i, λ_i correspond to the $(1,1)$ group and similarly for the group $(1,2)$. The functions $B^{(1)}$, $B^{(2)}$ depend linearly on $y_- = (y_j, 2 \le j \le n;\ y_I, y_{II})$ and analytically on s. As $s \to 0$, (2.39) becomes

$$\begin{aligned}
&\text{(a)} \quad y^{(1,1)}(-\ell) = 0,\\
&\text{(b)} \quad y^{(1,2)}(\ell) = 0,\\
&\text{(c)} \quad B^{(1)}\Big(y(\ell); 0\Big) - \rho^{(1,1)} B^{(1)}\Big(y(-\ell); 0\Big) = 0,\\
&\text{(d)} \quad B^{(2)}\Big(y(-\ell); 0\Big) - \rho^{(1,2)} B^{(2)}\Big(y(\ell); 0\Big) = 0.
\end{aligned} \tag{2.40}$$

In the vectors $y(-\ell)$, $y(\ell)$, we assume

$$\begin{aligned}
&y_{II}(-\ell) = y_I(\ell) = 0, \qquad && y^{(2,1)}(\ell) = \rho^{(2,1)} y^{(2,1)}(-\ell),\\
&&& y^{(2,2)}(-\ell) = \rho^{(2,2)} y^{(2,2)}(\ell).
\end{aligned} \tag{2.41}$$

To find the function B, one should observe that

$$\begin{aligned}
&s^{-1} dG_1[u_0](E_x, e^s)\varphi_I(x) c_I\\
&\quad = -s^{-1}(e^s - 1) dG_2[u_0](M_I, e^s)(M_I - 1)^{-1} X_I(s) y_I(x)\\
&\quad = -dG_2[u_0](M_I, 1)(M_I - 1)^{-1} X_I(s) y_I(x) + O(s) y_I(x),
\end{aligned} \tag{2.42}$$

where the matrix M_I has eigenvalues inside the unit circle.

A similar formula holds for $\varphi_{II}(x)$. For φ_i we have

$$\begin{aligned}
&s^{-1} dG_1[u_0](E_x, e^s)\varphi_i(x) c_i\\
&\quad = -\Big(1 + O(s)\Big) dG_2[u_0](e^{\mu_i(s)}, e^s)\mu_i^{-1}(s) X_i(s) y_i(x).
\end{aligned} \tag{2.43}$$

The matrices $B^{(1)}$ and $B^{(2)}$ are obtained by projecting the r.h.s.'s of (2.42), (2.43) onto the components of u corresponding to the groups $y^{(1,1)}$ and $y^{(1,2)}$. Notice that the r.h.s. of (2.42) is the summation formula for $\sum dG_2[u_0]\varphi_I(x)c_I$ and similarly for (2.43). Thus (2.39)(a)–(d) are equivalent to the equations obtained by substitution of u in (2.36) into $(n_1 - 1)$ equations $P_- dG_1(u) = 0$ and $(n_1 - 1)$ equations in (2.29). By consistency, $dG_2[u_0](E_x, e^s)X_i(s) = e_i + O(s)$. If $i > n_1$, then $Pe_i = 0$ so that indeed B depends analytically on s. In a particular case when the symbols $dG_j[u_0](\kappa, e^s)$, $j = 1, 2$, *are diagonal* modulo quadratic terms in $(\kappa - 1)$ and s, the contribution of φ_i to B is $O(s)$ so that y_i, $2 \le i \le n$, *do not enter* (2.40)(c),(d). For example, if the scheme G is at least second order accurate, one can rearrange G_1 and G_2 so that the

above holds. Another trivial case when y_i do not enter (2.40)(c),(d) is $n_1 = n$. The remaining boundary conditions (2.26)(c) at $s = 0$,

$$dS_2[u_0](E_x, 1)u = 0, \tag{2.44}$$

involve, generally speaking, all components of $y_-(\pm\ell)$.

The generalized stability condition. The system of equations (2.40), (2.41), (2.44) is nonsingular for all complex ρ_i, $n_1 < i \le n$, with $|\rho_i| \le 1$.

Usually the numerical boundary conditions consist of the physical ones and artificial ones. The artificial boundary conditions are extrapolatory and vanish on constant solutions. The physical ones represent some rules of reflection, i.e., express the ingoing characteristic components in terms of the outgoing ones. Thus, let us assume that the boundary conditions (2.44) split into

$$\begin{aligned} &\text{(a)} \quad dS^{(1)}_{2,\pm\ell}(E_x, 1)u = 0, \\ &\text{(b)} \quad dS^{(2)}_{2,\pm\ell}(E_x, 1)u = 0, \end{aligned} \tag{2.45}$$

such that

$$dS^{(2)}_{2,\pm\ell}(1, 1)(I - P) = 0, \tag{2.46}$$

while $dS^{(1)}_{2,\pm\ell}(1, 1)u = 0$ provide a full set of $n - n_1$ equations for $y^{(2,1)}(-\ell)$, $y^{(2,2)}(\ell)$ correspondingly. As above, we assume also that $y^{(2,1)}$, $y^{(2,2)}$ do not enter (2.40)(c),(d). Thus (2.40)(c),(d) and (2.45)(b) provide a full set of boundary conditions for $y_I(-\ell)$, $y_{II}(\ell)$. The generalized stability condition then implies that $y_I(-\ell), y_{II}(\ell) = 0$. In other words, we obtain the following condition.

The stationary boundary layer condition. The equation $dG_1[u_0](E_x, 1)u = 0$ has no boundary layer solutions u which satisfy (2.45)(b) and the conservation laws

$$P_- \sum_x dG_2[u_0](E_x, 1)u = 0.$$

By (2.45)(a), the vectors $y^{(2,1)}(-\ell)$, $y^{(2,2)}(\ell)$ could be expressed as

$$\begin{aligned} &\text{(a)} \quad y^{(2,1)}(-\ell) = R^{(1)} y^{(2,2)}(-\ell), \\ &\text{(b)} \quad y^{(2,2)}(\ell) = R^{(2)} y^{(2,1)}(\ell). \end{aligned} \tag{2.47}$$

The generalized stability condition now implies that

$$\det(R^{(1)}\rho^{(2,2)}R^{(2)}\rho^{(2,1)} - I) \ne 0. \tag{2.48}$$

The last clearly holds if

$$\|R^{(1)}\| \cdot \|R^{(2)}\| < 1; \tag{2.49}$$

i.e., the reflection boundary conditions in (2.47) are dissipative. It is well known that if the boundary conditions are nondissipative, a system of hyperbolic differential equations may have exponentially growing solutions. Clearly, (2.49) and the stationary boundary layer condition imply the generalized stability condition, provided

$$P_- dG_2[u_0](1, 1)(I - P) = 0 \tag{2.50}$$

and (2.46) hold. Thus, for a wide class of boundary conditions we obtain $\{\rho_i\}$ independent criteria of stability. In particular, if the scheme $dG[u_0]$ is diagonal and if the boundary conditions do not mix different components of u, then the stability criteria become very simple. Indeed, the problem becomes scalar. Condition (2.48) is absent. Since the conservation law $\sum dG_2[u_0]u = 0$ is scalar, either $dS^{(2)}_{2,-\ell}(E_x, 1)u = 0$ uniquely define $y_{I}(-\ell)$ or $dS^{(2)}_{2,\ell}(E_x, 1)u = 0$ uniquely define $y_{II}(\ell)$. In any case the boundary conditions for $y_{I}(-\ell)$, $y_{II}(\ell)$ are uncoupled. Then one can use for $S^{(2)}_2$ boundary conditions which satisfy scheme-independent stability criteria. For s away from zero the eigenvalues of $dG[u_0](E_x, e^s)$ are away from the unit circle. The Lopatinsky condition for the half-line problems in (2.21) again holds for these boundary conditions.

Remark 2. Notice that the global conservation laws (2.14), though formally following from (2.13), actually fail for large time due to the round-off errors. Therefore, at the outflow boundary, the ith component of (2.12) for $2 \le i \le n_1$ should be replaced by the ith component of (2.14). (The outflow boundary is the opposite of the inflow one in (2.28).) For $i = 1$ one can do the replacement at either boundary.

Examples. Let us approximate the derivative $f(u)_x$ in (1.1) by

$$F\Big(\{E^\alpha u\}\Big) = \frac{1}{\Delta x}\Big(D_x^{(k+1)} f(u) + (-1)^{k_1} K (E_x^{1/2} - E_x^{-1/2})^{k+1} u\Big), \qquad k_1 = \frac{k+1}{2}, \tag{2.51}$$

where $(\Delta x)^{-1} D_x^{(k+1)}$ is a $k+1$ order central difference approximation of the operator ∂_x and $K = O^*(1)$ is a positive constant. For example,

$$D_x^{(k+1)} = \sum_{j=1}^{k_1} c_j (E_x^j - E_x^{-j}), \qquad c_j = \frac{(-1)^{j-1}(k_1!)^2}{j(k_1 - j)!(k_1 + j)!}, \tag{2.52}$$

has a maximal accuracy for the $k+2$ point lattice. Since $k+1$ is even, the dissipative term $(-1)^{k_1}(E_x^{1/2} - E_x^{-1/2})^{k+1}$ contains only integer powers of E_x. Now we replace (1.1) by the equation

$$u_t + F(u) = 0 \qquad \Big(F(u) \text{ is shorthand for } F\big(\{E^\alpha u\}\big)\Big) \tag{2.53}$$

in the space of grid-function $u(x)$, $x \in I_h$, and approximate it by an mth order ODE solver with $m \ge k$. For example, one can use the mth order Adams–Bashforth multistep method

$$(I - E_t^{-1}) + \Delta t \sum_{i=1}^{m} d_i E_t^{-i} F(u) = 0. \tag{2.54}$$

Clearly, $F(u) = (E_x - I)F_1(u)$, and hence the scheme is in conservation form (2.4). Let $d\widehat{F}[u_0]$ be the Fourier symbol of the differential of F at $[u_0]$. The eigenvalues λ of $\Delta t d\widehat{F}[u_0]$ are

$$\lambda = \frac{\Delta t}{\Delta x}\left(\lambda_i(u_0)\sqrt{-1}\sum_{j=1}^{k_1} 2c_j \sin j\xi + K(2\sin\xi/2)^{k+1}\right). \tag{2.55}$$

For the stability of the Cauchy problem, one needs that the roots z of the characteristic equation

$$1 - z^{-1} + \sum d_i z^{-1} \lambda = 0 \tag{2.56}$$

satisfy $|z| \le 1$. In other words, $-\lambda$ should belong to the domain Ω of absolute stability of the multistep method. Since the method is of order m, a point $-\lambda$ near 0 with $\operatorname{Re}\lambda \ge K_1 |\operatorname{Im}\lambda|^{m+1}$ belongs to Ω. Thus, if $m > k+1$, then $-\lambda$ in (2.55) for small ξ always belongs to Ω. If $m = k+1$, then it might be necessary to increase K or to decrease $\frac{\Delta t}{\Delta x}$. For nonsmall values of ξ the condition $-\lambda \in \Omega$ restricts from above the values of $|\lambda_i(u_0)\frac{\Delta t}{\Delta x}|$ and $K\frac{\Delta t}{\Delta x}$. The scheme clearly satisfies the accuracy and dissipativity assumptions. The boundary conditions could be set as follows: The zero flux conditions (2.12),

$$P\Big(F_1\Big(u(x_{\pm\ell}, t)\Big) - f(u_{L,R})\Big) = 0, \qquad x_{-\ell} = -\ell + k_1, \qquad x_\ell = \ell - k_1, \tag{2.57}$$

are used to compute $Pu(\pm\ell, t)$. Hence K should be such that

$$K \ne |c_{k_1} \lambda_i(u_0)|, \qquad 1 \le i \le n_1. \tag{2.58}$$

For $1 \le i \le n_1$ we need $k_1 - 1$ more conditions to compute $u_i(x,t)$, $-\ell + 1 \le x < -\ell + k_1$. These conditions are

$$u_i(-\ell+1, t) = u_{L,i}, \quad (E_x - I)^j u_i(-\ell+1, t) = 0, \quad 1 \le j \le k_1 - 2. \tag{2.59}$$

If $i > n_1$ and $\lambda_i(u_0) > 0$, i.e., $x = -\ell$ is the inflow boundary, then the conditions are

$$u_i(-\ell, t) = u_{L,i}, \qquad (E_x - I)^j u_i(-\ell, t) = 0, \qquad 1 \le j \le k_1 - 1, \tag{2.60}$$

and in the outflow case,

$$(E_x - I)^j u_i(-\ell, t) = 0, \qquad 1 \le j \le k_1, \tag{2.61}$$

or, instead,

$$(E_x - 1)^d E_x^j u_i(-\ell, t) = 0, \qquad 0 \le j \le k_1 - 1, \quad d \ge 1 \text{ an integer}. \tag{2.62}$$

Similar conditions are imposed at the right boundary. Note that in (2.62) we used a fixed type of extrapolation translated along the boundary points. Such boundary conditions were suggested by [9]. Since the scheme is explicit, it is easy to see that the boundary conditions define all the boundary values of $u(x)$; i.e., the scheme is solvable. Let us now check the weak Lopatinsky condition. Since the scheme is diagonal it suffices to consider the case of scalar $f(u) = \lambda_i u$. Due to the stability of the Cauchy problem, the characteristic equation $dG[u_0](\kappa, e^s) = 0$ for $\operatorname{Re} s \ge 0$ has only solutions κ inside or outside the unit circle. Let κ_ν, $1 \le \nu \le N$, be the roots $|\kappa| < 1$ counted with their multiplicities. By solvability, $N = k_1$ is the number of boundary conditions. If all κ_ν are distinct, then the substitution of the general decreasing homogeneous solution $\varphi(x) = \sum c_\nu \kappa_\nu^{(x+\ell)}$ into (2.62) results in a matrix $D = \{(\kappa_\nu - 1)^d \kappa_\nu^j\}$, which is proportional to a Vandermonde matrix and hence is nonsingular. If some κ_ν is a multiple root, then the matrix D will contain the corresponding column and its

derivatives; hence again D is nonsingular. In view of (2.42), the boundary condition (2.57) implies

$$\sum_{\nu} \kappa_\nu^{k_1} (\kappa_\nu - 1)^{-1} c_\nu = 0. \tag{2.63}$$

The factor $\kappa_\nu^{k_1}$ is due to the negative power $E_x^{-k_1}$ in (2.51). The boundary conditions in (2.59) result in a system

$$\sum_{\nu} \kappa_\nu (\kappa_\nu - 1)^j c_\nu = 0, \qquad 0 \le j \le k_1 - 2. \tag{2.64}$$

Since $\kappa^{k_1}(\kappa-1)^{-1} = \kappa(\kappa-1)^{-1} + \sum_{j=0}^{k_1-2} \operatorname{coef} \cdot \kappa(\kappa-1)^j$, the matrix which corresponds to (2.63), (2.64) is equivalent to $D = \{\kappa_\nu(\kappa_\nu - 1)^j\}$, $-1 \le j \le k_1 - 2$, and is nonsingular, provided $\kappa_\nu \ne 0$. The last follows from (2.58). Notice that for our proof it is essential that the boundary conditions (2.59) are applied to $x = -\ell + 1$ and not to $x = -\ell$. Clearly, conditions (2.59) for $i = 1$ are of type (2.30). The scalar shock stability condition in (2.32) is satisfied since $d_j = j - 1$. The decoupling condition in (2.34), (2.35) is satisfied trivially since the corresponding columns are empty. The stationary boundary layer condition was already verified above. Since the scheme is diagonal and (2.46) holds trivially, the generalized stability condition follows. Thus our difference approximation satisfies all the requirements of asymptotic stability of the weak discrete shocks.

Another important method of solution of (2.53) is by an explicit r-stage Runge–Kutta method

$$k_q = \Delta t F\left(u + \sum_{j=1}^{q-1} \alpha_{qj} k_j\right), \quad E_t u = u + \sum_{q=1}^{r} \alpha_{r+1,q} k_q, \quad \alpha_{q,q-1} \ne 0. \tag{2.65}$$

Again, for stability of the Cauchy problem the eigenvalues λ in (2.55) should lie in the domain Ω of absolute stability of the Runge–Kutta method. For example, for a four stage fourth order Runge–Kutta method the domain Ω is

$$\left|\sum_{i=1}^{4} \frac{\lambda^i}{i!}\right| \le 1. \tag{2.66}$$

It is easy to check that Ω includes an imaginary interval around zero (precisely $(-\sqrt{8}i, \sqrt{8}i)$). Hence the Cauchy stability holds for any odd k and any K, provided Δt small enough. In case of a two step second order Runge–Kutta method, Ω is

$$\left|1 + \lambda + \frac{\lambda^2}{2}\right| \le 1. \tag{2.67}$$

The boundary of Ω for small λ satisfies

$$\operatorname{Re}\lambda = -\frac{(\operatorname{Im}\lambda)^4}{4} + O(\operatorname{Im}\lambda)^5; \tag{2.68}$$

hence for $k = 3$ and sufficiently small $\frac{\Delta t}{\Delta x}$ the Cauchy stability holds. For the sake of the global conservation law we impose the boundary conditions

$$PF_1\left(u + \sum_{j=1}^{q-1} \alpha_{q_j} k_j\right)\Bigg|_{x = x_{\pm\ell}} - Pf(u_{L,R}) = 0, \qquad 2 \le q \le r + 1. \tag{2.69}$$

These conditions are used to determine the values of $Pk_q(\pm\ell)$, $1 \le q \le r$. Notice that the equation in (2.69) for $q = r + 1$ implies

$$PF_1(E_t u)|_{x=x_{\pm\ell}} - Pf(u_{L,R}) = 0. \tag{2.70}$$

The boundary conditions (2.59)–(2.62) are replaced correspondingly by

$$P_i(E_x - I)^j k_q(-\ell + 1, t) = 0, \qquad 0 \le j \le k_1 - 2, \tag{2.71}$$

$$P_i(E_x - I)^j k_q(-\ell, t) = 0, \qquad 0 \le j \le k_1 - 1, \tag{2.72}$$

$$P_i(E_x - I)^j k_q(-\ell, t) = 0, \qquad 1 \le j \le k_1, \tag{2.73}$$

and

$$P_i(E_x - I)^d E_x^j k_q(-\ell, t) = 0, \qquad 1 \le j \le k_1 - 1, \tag{2.74}$$

where P_i are the standard projections and $1 \le q \le r$. We assume that $u(t = 0)$ satisfies (2.59)–(2.62). Hence $u(t)$ satisfies (2.57) and (2.59)–(2.62). Notice that for linear F_1 the equations in (2.69) become

$$PF_1(k_q)|_{x=x_{\pm\ell}} = 0, \qquad 1 \le q \le r. \tag{2.75}$$

Thus all k_q satisfy the same homogeneous boundary conditions. We can consider the operator F as acting in the space of grid-functions satisfying (2.71)–(2.75). Denote the extended operator by $\mathcal{F}$. Then

$$E_t u = p(\Delta t \mathcal{F})u, \tag{2.76}$$

where p is an rth degree polynomial and $p(0) = 1$. We can consider (2.76) also on half-lines $x \ge -\ell$ or $x \le \ell$. The weak Lopatinsky condition for (2.76) is satisfied if the operator $\Delta t\mathcal{F}$ restricted to the above half-lines does not have eigenvalues $\lambda \ne 0$, $\lambda \notin \mathrm{Int}(\Omega)$. The analysis now proceeds as in the multistep case. Due to the stability of the Cauchy problem, for λ as above the roots κ of the characteristic equation

$$\det \left| \lambda I - \Delta t dF[u_0](\kappa) \right| = 0 \tag{2.77}$$

do not lie on the unit circle. The substitution of the roots $|\kappa| < 1$ into the boundary conditions (2.71)–(2.75) leads to the Vandermonde-type matrices. In view of (2.58), the boundary conditions (2.69), (2.71)–(2.74) uniquely define the boundary values of k_q and hence the scheme is solvable. According to Remark 2, the boundary conditions (2.57) at the outflow boundaries should be replaced by the global conservation law

$$\sum_{x_{-\ell} \le x \le x_\ell - 1} Pu(x, t) = \text{ const.}, \tag{2.78}$$

and correspondingly, (2.69) should be replaced by

$$\sum_{-x_\ell \le x \le x_\ell - 1} Pk_q(x) = 0. \tag{2.79}$$

We again denote the operator $\mathcal{F}$ with modified boundary conditions by $\mathcal{F}$. Since the scheme is diagonal, the generalized stability condition decouples into corresponding conditions for the half-line problems. Namely, for $\lambda_i(u_0) > 0$ (i.e., the inflow boundary at $x = -\ell$) the problem for the ith component $P_i dG[u_0](E_x, 1)u = 0$, $x \geq -\ell$, with the boundary conditions (2.71)–(2.74) and (2.75) at $x = x_{-\ell}$ should have no bounded solutions u. Since the resulting problem could be written in a polynomial form

$$P_i\Big(p(\Delta t\mathcal{F}) - 1\Big)u = P_i \prod_{j=1}^{r} (\Delta t\mathcal{F} - \mu_j)u = 0, \tag{2.80}$$

it is enough to show that $P_i(\Delta t\mathcal{F} - \mu_j)u = 0$ does not have solutions $u \in \ell_\infty$ on the half-line. Here $\mu_1 = 0$ and the remaining $\mu_j \neq 0$ belong to the boundary of Ω. The case $\mu_j \neq 0$ was already considered above, and the case $\mu_j = 0$ was also considered when we studied the multistep scheme. For the problem on the half-line $x \leq \ell$ the boundary conditions $P_i F_1(k_q) = 0$ are replaced by

$$\sum_{-\infty \leq x \leq x_\ell - 1} P_i k_q(x) = 0. \tag{2.81}$$

Now the resulting problem should have no solutions in ℓ_1 on the half-line. Again we arrive at the equations $P_i(\Delta t\mathcal{F} - \mu_j)u = 0$ which were studied above in the multistep case.

Remark 3. The Runge–Kutta scheme employed here is not exactly of type (2.1) with boundary conditions (2.13). Indeed, if we eliminate the intermediate steps k_q, the resulting scheme could be rewritten as

$$(E_t - I)u + (E_x - I)G_1\Big(\big\{E^j u(x,t)\big\}, x\Big) = 0, \tag{2.82}$$

where the flux G_1 changes its form as a discrete function of x. For inner points $-\ell + k_1 r < x < \ell - k_1 r$ the flux G_1 depends on $E^j u$, $-k_1 r \leq j \leq k_1 r - 1$, while at $x \equiv x_{\pm\ell}$, $G_1 = \Delta t F_1$. One can consider the equations in (2.82) for boundary points (i.e., noninner points) as artificial boundary conditions which are included in (2.45)(b). The remaining boundary conditions (2.57), (2.59)–(2.62) are used to compute $E_t u$ for $-\ell \leq x \leq -\ell + k_1$ and $\ell - k_1 \leq x \leq \ell$. The first condition in (2.60) belongs to (2.45)(a), the conditions in (2.59) for $i = 1$ belong to (2.26)(b), (2.57) stands for (2.26)(a), and the remaining conditions in (2.59)–(2.62) belong to (2.45)(b). Now formula (2.42) does not hold. However, if $u = \varphi_{_I}(x)c_{_I}$ in (2.38) is a solution of the equation

$$\Big(sI + (E_x - I)dG_1[u_0, x]\Big)u = 0 \tag{2.83}$$

for all $-\ell \leq x$, then for $s = 0$, $dG_1[u_0, x]u \equiv 0$, and for small $|s|$, $dG_1[u_0, x]u = O(s)$. Hence the functions $B^{(1)}$, $B^{(2)}$ in (2.39) are well defined. With these remarks in mind, our general theory applies also to the Runge–Kutta scheme.

Finally, let us check the generalized stability condition for the first component $P_1 u$. Since the problem is scalar, it is enough to check the stationary boundary layer condition. The projector P_- vanishes on the first component; hence there is no global conservation law condition. Thus the boundary layers at $x = \pm\ell$ are uncoupled. Hence we can assume that the boundary layer u is defined for $x \geq -\ell$, decays exponentially as $x \to \infty$, and satisfies the equation $(E_x - I)dG_1[u_0, x]u = 0$ for all $x \geq -\ell$ (see (2.83)) but not conditions (2.59) and (2.57) for the first component. If we represent

the polynomial p in (2.76) as $p(z) = 1 + zp_1(z)$, then the solution u satisfies the functional equation

$$\mathcal{F}_0 p_1(\Delta t \mathcal{F}) u = 0, \tag{2.84}$$

where $\mathcal{F}_0$ is given by the difference operator $dF[u_0]$ on the half-line $x \geq -\ell$ without boundary conditions. Since $\lambda_1(u_0) = 0$, $\mathcal{F}_0 \approx (E_x - 1)^{k+1}$. Hence u satisfies the equation $p_1(\Delta t \mathcal{F})u = \prod_{j=2}^{r}(\Delta t \mathcal{F} - \mu_j)u = 0$. It was shown above that this equation has no bounded solutions. The decoupling condition in (2.35) is absent for diagonal schemes. The boundary operator in (2.34) is given by $P_1(E_x - I)dG_1[u_0, x]$, which is divisible by $(E_x - I)^{k+1}$. Thus the asymptotic stability theory applies also to the above Runge–Kutta approximation.

Our final assumption pertains to the continuous problem (2.22).

The continuous stability hypothesis. The problem (2.23) with odd k has a solution y_{sh} such that for $\operatorname{Re} s' \geq 0$, problem (2.22) has no solutions $y_1 \in H^{k+1}(R^1)$ with $\int_{-\infty}^{\infty} y_1 dx = 0$.

The above hypothesis holds trivially for $k = 1$ and has been proved in [5] for $k = 3$. In both cases the solution y_{sh} is unique (up to translation). Now we can state the main results.

THEOREM 2.1 (the existence of stationary shocks). *Let all the above assumptions hold for $s = 0$. Then there exist ε_0 and δ_0 such that for $\varepsilon < \varepsilon_0$ and $\ell > \delta_0^{-1}\varepsilon^{-1}$ problem* (2.1), (2.2) *has an $n_1 \geq 1$ parameter family of stationary solutions u_{st}, where $n_1 = \dim \operatorname{Im} P$. This solution has the form*

$$u_{st} = u_{sh} + \Delta u_{st}, \tag{2.85}$$

where

$$u_{sh} = \tfrac{1}{2}(u_{L} + u_{R}) + \tfrac{1}{2}(u_{L} - u_{R}) y_{sh}\Big(\varepsilon(x - x_0)\Big), \quad \varepsilon = \Big(\lambda_1(u_{L})\Big)^{1/k}, \tag{2.86}$$

$y_{sh}(\tau)$ is a solution of (2.23), *$u_{L,R} = u_0 + O(\varepsilon^k)$ are as in* (1.3), *and*

$$\varepsilon \|\Delta u_{st}\|_1 + \|\Delta u_{st}\|_\infty \leq \delta_0 \varepsilon^k \ . \tag{2.87}$$

For $n_1 = 1$, $k \geq 1$, and for $1 < n_1 < n$, $k \geq 3$, in addition we have

$$\|(E_x - I)\Delta u_{st}\|_1 \leq K\varepsilon^{2k-1} + \delta_0 \varepsilon^k. \tag{2.88}$$

Here $\|\cdot\|_p$ is the usual ℓ_p norm over the interval $[-\ell, \ell]$.

THEOREM 2.2. *Let all the above assumptions hold for all $\operatorname{Re} s \geq 0$. Then there exist ε_0 and δ_0 such that for all $\varepsilon < \varepsilon_0$ and all ℓ as in* (1.12) *for $n_1 = 1$ or $n_1 = n$, the solution of problems* (2.1), (2.7), (2.9) *with initial condition u_{in} in a small neighborhood of u_{st} tends to u_{st} as $t \to \infty$, provided*

$$P \sum G_2\Big(\{E^j u_{st}\}\Big) = P \sum G_2\Big(\{E^j u_{in}\}\Big). \tag{2.89}$$

The last condition uniquely selects the appropriate u_{st} out of the n_1 parameter family of stationary solutions. The precise size of the neighborhood of u_{st} and the rate of the exponential decay of $|u - u_{st}|$ is given in section 6, estimates (6.26), (6.27), (6.39), (6.40).

3. The linearized problem.

3.1. Preliminary transformations and the normal block form. Let u_{st} be a stationary solution as in (2.85)–(2.87), and consider the linearized problem

$$\begin{aligned} &\text{(a)} \quad dG[u_{st}](E_x, E_t)u = F, \\ &\text{(b)} \quad dS[u_{st}](E_x, E_t)u = g = (g_{_L}, g_{_R}, g_{\rm sum}). \end{aligned} \tag{3.1}$$

Here $dG[u_{st}](E_x, E_t)$ is the differential of $G(\{E^j u(x,t)\})$ at $u = u_{st}$ and similarly for dS. The boundary operator dS consists of the differential of $S_{_L}$ and $S_{_R}$ in (2.7) and of the left-hand side (l.h.s.) of (2.14) with respect to $u(x,t)$. The partition of g above corresponds to the above partition of dS. We will study the resolvent problem

$$\begin{aligned} &\text{(a)} \quad dG[u_{st}](E_x, e^s)u = F, \\ &\text{(b)} \quad dS[u_{st}](E_x, e^s)u = g, \end{aligned} \tag{3.2}$$

where s is a complex parameter with $\operatorname{Re} s \ge 0$. Recall that $dS_{_L}$ and $dS_{_R}$ consist of three parts as in (2.26) and, correspondingly, $g_{_L}$, $g_{_R}$ split as

$$g_{_L} = (g_{_{L,0}}, g_{_{L,1}}, g_{_{L,2}}), \qquad g_{_R} = (g_{_{R,0}}, g_{_{R,1}}, g_{_{R,2}}). \tag{3.3}$$

Clearly the system in (3.2) is overdetermined since

$$P \sum_{x=x_{-\ell}}^{x_\ell - 1} F(x) = (1 - e^{-s}) g_{\rm sum} + g_{_{R,0}} - g_{_{L,0}}. \tag{3.4}$$

To make it uniquely solvable one should proceed as in Remark 2.

For $\operatorname{Re} s \ge 0$ and $s \ne 0$ the roots κ of the characteristic equation $\det dG[u](\kappa, e^s) = 0$ (here u is a constant grid-function) are away from the unit circle $|\kappa| = 1$, and problem (3.2) is trivial. Hence we will consider the more difficult case of small $|s|$ (actually, the most difficult is the subcase of $|s| < K\varepsilon^{k+1}$). In order to eliminate the zero eigenvalue $s' = 0$ in (2.22) one should integrate at least the first component of (3.2)(a). We found it convenient to integrate all components of (3.2)(a). Namely, represent

$$dG_2(E_x, e^s) = dG_2(1, e^s) + (E_x - I)\Delta dG_2(E_x, e^s) \tag{3.5}$$

and define $U(x)$ by

$$(E_x - I)U = u, \qquad U(x_{-\ell}) = -\Big(dG_2(1, e^s)\Big)^{-1} \Delta dG_2(E_x, e^s) u(x_{-\ell}) \tag{3.6}$$

so that U satisfies the boundary condition

$$dG_2(E_x, e^s) U(x_{-\ell}) = 0. \tag{3.7}$$

The summation of (3.2)(a) from $x_{-\ell}$ until $x - 1 \le x_\ell - 1$ yields

$$\begin{aligned} &dG[u_{st}(x)](E_x, e^s)U(x) \\ &\quad \stackrel{df}{=} (e^s - 1) dG_2(E_x, e^s)U(x) + dG_1[u_{st}(x)](E_x, e^s)(E_x - I)U(x) \\ &\quad = \sum_{x_{-\ell}}^{x-1} F(\xi) + dG_1[u_{st}(x_{-\ell})](E_x, e^s) u(x_{-\ell}) = F^{(1)}(x). \end{aligned} \tag{3.8}$$

Recall that by (2.13),

$$PdG_1[u_{st}(x_{-\ell})](E_x, e^s)u(x_{-\ell}) = g_{L,0}. \tag{3.9}$$

It will turn out that Δu_{st} in (2.85) is a negligible term, and hence u_{st} in (3.8) and (3.2)(b) will be replaced by u_{sh}. The resulting difference operators dG_1 and dS still depend on u_{sh} at several grid points $x + j_1$. We replace $u_{sh}(x + j_1)$ in dG_1 and dS by $u_{sh}(x)$ so that the coefficients of dG_1 and dS depend only on a point value of $u_{sh}(x)$, and we consider the difference problem

$$\begin{aligned} &\text{(a)} \quad dG(\eta)(E_x, e^s)U(x) \\ &\qquad \stackrel{df}{=} \Big[(e^s - 1)dG_2(E_x, e^s) + dG_1(\eta)(E_x, e^s)(E_x - 1)\Big]U \\ &\qquad = F^{(1)}(x), \qquad x_{-\ell} + 1 \le x \le x_\ell, \\ &\text{(b)} \quad dS(\eta)u = g. \end{aligned} \tag{3.10}$$

Here

$$\eta = u_{sh}(x) - u_0 = \mu e_1 y_{sh}\Big(\varepsilon(x - x_0)\Big) + O(\mu^2) \tag{3.11}$$

(see (2.10)) is a small parameter. The new grid-functions $F^{(1)}$ and g relate to the old ones in (3.8) and (3.2)(b) by

$$F^{(1)}_{\text{new}} = F^{(1)}_{\text{old}} + \Big(O(\Delta u_{st}) + O(\eta')\Big)u, \tag{3.12}$$

$$g_{\text{new}} = g_{\text{old}} + \Big(O(\Delta u_{st}) + O(\eta')\Big)u. \tag{3.13}$$

Remark 4. We do not modify $dG_1\big[u_{st}(x_{-\ell})\big](E_x, e^s)$ in (3.8) and (3.9). Recall that, by (2.87),

$$\|\Delta u_{st}\|_1 \le \delta_0 \varepsilon^{k-1}, \tag{3.14}$$

where δ_0 is a small constant, while

$$\|\eta'\|_1 = |\mu| \, \|\varepsilon \partial_\tau y_{sh}(\tau)\|_1 \le K|\mu| = K\varepsilon^k. \tag{3.15}$$

Without loss of generality, we may assume that the powers of E_x in dG are nonnegative,

$$dG(\eta)(E_x, e^s) = L(\eta, s, E_x) = \sum_{i=0}^{\nu_0} L_i(\eta, s)E_x^i. \tag{3.16}$$

Introduce the grid vector functions

$$\widetilde{U}(x) = \Big(U(x), E_x U(x), \dots, E_x^{\nu_0 - 1}U(x)\Big)^T, \qquad \widetilde{F} = \Big(0, 0, \dots, F^{(1)}\Big)^T \tag{3.17}$$

and difference operators

$$B = \begin{bmatrix} 0 & -I & 0 & & 0 \\ 0 & 0 & -I & & \\ & & & \ddots & \\ & & & & -I \\ L_0 & L_1 & & \cdots & L_{\nu_0 - 1} \end{bmatrix}, \qquad A = \operatorname{diag}(I, I, \dots, L_{\nu_0}). \tag{3.18}$$

Then (3.10) could be rewritten as

$$\widetilde{L}(\eta, s, E_x)\widetilde{U} = \Big(A(\eta, s)E_x + B(\eta, s)\Big)\widetilde{U} = \widetilde{F}. \tag{3.19}$$

The operators $\widetilde{L}$ and dG are related by the equivalence

$$D_2(\eta, s, E_x)\widetilde{L}(\eta, s, E_x)D_1(E_x) = L(\eta, s, E_x) \oplus I_{n(\nu_0-1)}, \tag{3.20}$$

where D_1, D_2 and their inverses are polynomial matrices in E_x, and the first block column of D_1 is $(I, E_x, \dots, E_x^{\nu_0-1})^T$. Note that we do not assume that A or B are nonsingular. The eigenvalues of the pencil of matrices $\widetilde{L}(\eta, s, \kappa) = A\kappa + B$ satisfy

$$\det \widetilde{L}(\eta, s, \kappa) = \det L(\eta, s, \kappa) = 0. \tag{3.21}$$

The infinite eigenvalues or those which are close to infinity are defined by

$$\det \kappa\widetilde{L}(\eta, s, \kappa^{-1}) = \det \kappa^{\nu_0} L(\eta, s, \kappa^{-1}) = 0. \tag{3.22}$$

At $s = \eta = 0$ and κ close to 1, $L(\eta, s, \kappa)$ becomes

$$L(0, 0, \kappa) = (\kappa - 1)\Big(\operatorname{diag}(\lambda_1(u_0), \dots, \lambda_n(u_0)) + O(\kappa - 1)\Big), \qquad \lambda_1(u_0) = 0, \tag{3.23}$$

where by (2.15) the first column of $O(\kappa-1)$ is actually $O(\kappa-1)^k$. Hence the characteristic equation (3.21) at $s = \eta = 0$ has the eigenvalue $\kappa = 1$ of multiplicity $n+k$. By the dissipativity assumption, all other eigenvalues are not on the unit circle. We will split them into two groups: I, for $|\kappa| < 1$, and II, for $|\kappa| > 1$. There are, correspondingly, two projectors

$$Q_{I}(\eta, s) = (2\pi i)^{-1} \oint_{|\kappa|=1-\delta} \Big(\kappa A(\eta, s) + B(\eta, s)\Big)^{-1} A(\eta, s) d\kappa, \tag{3.24}$$

$$Q_{II}(\eta, s) = (2\pi i)^{-1} \oint_{|\kappa|=1-\delta} \Big(A(\eta, s) + \kappa B(\eta, s)\Big)^{-1} B(\eta, s) d\kappa \tag{3.25}$$

on the invariant spaces of the pencil $A\kappa+B$ corresponding to the eigenvalues of groups I and II. Here δ is such that the eigenvalues κ of the first group and the inverses κ^{-1} of the second group lie in the disc $|\kappa| < 1 - \delta$. One can choose basis $X_{I,II}$ in $\operatorname{Im} Q_{I,II}$ which depends analytically on s and η in a neighborhood of zero such that

$$AX_{I}M_{I} + BX_{I} = 0, \qquad AX_{II} + BX_{II}M_{II} = 0, \tag{3.26}$$

where M_{I}, M_{II} also depend analytically on s and

$$\|M_{I,II}\| < 1 - \delta \tag{3.27}$$

(e.g., see pp. 14–17 in [8]). The projector Q_0 associated with the eigenvalue $\kappa = 1$ is

$$Q_0(\eta, s) = (2\pi i)^{-1} \oint_{|\kappa-1|=\delta} \Big(\kappa A(\eta, s) + B(\eta, s)\Big)^{-1} A(\eta, s) d\kappa \tag{3.28}$$

and has an $n+k$ dimensional image. In order to split $\operatorname{Im} Q_0(\eta, s)$ into simple components, we rewrite

$$\begin{aligned} L(\eta, s, \kappa) &= dG_2(\kappa, e^s)\Big((e^s - I)I + (\kappa - 1)dG_2^{-1}(\kappa, e^s)dG_1(\eta)(\kappa, e^s)\Big) \\ &= dG_2(\kappa, e^s)\Big((e^s - 1)I + (\kappa - 1)L'(\eta, s, \kappa)\Big). \end{aligned} \tag{3.29}$$

Recall that, by consistency in (2.5), $dG_2(\kappa, e^s) = I + O(\kappa - 1) + O(s)$, and hence

$$L'(\eta, s, \kappa) = \operatorname{diag}\Big(\lambda_1(u_0), \dots, \lambda_n(u_0)\Big) + O(\kappa - 1) + O(s) + O(\eta). \tag{3.30}$$

Since the eigenvalues $\lambda_i(u_0)$ are distinct, L' could be brought to a diagonal form

$$C^{-1}(\eta, s, \kappa)L'(\eta, s, \kappa)C'(\eta, s, \kappa) = \operatorname{diag}\Big(p_1'(\eta, s, \kappa), \dots, p_n'(\eta, s, \kappa)\Big), \tag{3.31}$$

where C and p_i' depend analytically on the parameters $C(0, 0, 1) = I$ and $p_i'(0, 0, 1) = \lambda_i(u_0)$. Hence, for $i = 2, \dots, n$ the matrix $L(\eta, s, \kappa)$ has eigenvalues

$$\kappa_i = 1 - s\Big(\lambda_i^{-1}(u_0) + O(s) + O(\eta)\Big) \tag{3.32}$$

with the eigenvectors

$$C_i(\eta, s, \kappa_i) = e_i + O(s) + O(\eta), \tag{3.33}$$

where C_i is the ith column of C. The matrix $\widetilde{L}(\eta, s, \kappa)$ has the corresponding eigenvectors

$$X_i(\eta, s) = (C_i, \kappa_i C_i, \dots, \kappa_i^{\nu_0 - 1} C_i)^T = (e_i, \dots, e_i)^T + O(s) + O(\eta). \tag{3.34}$$

By the dissipativity assumption, for $\operatorname{Re} s \geq 0$ and small real η,

$$\begin{aligned} &|\kappa_i(\eta, s)| \leq 1 \quad \text{if} \quad \lambda_i(u_0) > 0, \\ &|\kappa_i(\eta, s)| \geq 1 \quad \text{if} \quad \lambda_i(u_0) < 0. \end{aligned} \tag{3.35}$$

The determinant of $L'(\eta, s, \kappa)$ as follows from (2.15), (2.16) is

$$\begin{aligned} &\det L'(\eta, s, \kappa) = \Big(1 + O(\kappa - 1) + O(s)\Big) \\ &\cdot \left(\prod_{i=2}^{n} \lambda_i(u_0) b_1 (\kappa - 1)^k + O(s) + O_1(\eta) + O\Big((\kappa - 1)^{k+1}\Big) + O(\eta)O(\kappa - 1)\right), \end{aligned} \tag{3.36}$$

where

$$\begin{aligned} O_1 &= \partial_{e_1} f_1(u_0 + \eta) \prod_{i=2}^{n} \lambda_i(u_0) + O(\eta^2) \\ &= d\partial_{e_1} f_1(u_0) \cdot \eta \prod_{i=2}^{n} \lambda_i(u_0) + O(\eta^2) \end{aligned} \tag{3.37}$$

and

$$b_1 = (-1)^{\frac{k+1}{2}} b \qquad \text{(see (2.16))}. \tag{3.38}$$

Here f_1 is the first component of the flux f in (1.1). On the other hand,

$$\begin{aligned} \det L'(\eta, s, \kappa) &= \prod_{i=1}^{n} p_i'(\eta, s, \kappa) \\ &= \left(\prod_{i=2}^{n} \lambda_i(u_0) + O(\kappa - 1) + O(s) + O(\eta) \right) p_1'(\eta, s, \kappa). \end{aligned} \tag{3.39}$$

Hence

$$\begin{aligned} p_1'(\eta, s, \kappa) &= b_1(\kappa - 1)^k + d\partial_{e_1} f_1(u_0) \cdot \eta \\ &\quad + O\Big((\kappa - 1)^{k+1}\Big) + O(s) + O(\eta^2) + O(\eta) \cdot O(\kappa - 1). \end{aligned} \tag{3.40}$$

Thus the remaining $k+1$ roots κ of (3.21) in a neighborhood of $\kappa = 1$, $s, \eta = 0$ satisfy

$$\begin{aligned} p_1(\eta, s, \kappa) &= (e^s - 1) + (\kappa - 1) p_1'(\kappa, s) \\ &= b_1(\kappa - 1)^{k+1} + \Big(d\partial_{e_1} f_1(u_0)\eta \Big)(\kappa - 1) + s \\ &\quad + (\kappa - 1)\Big(O(\kappa - 1)^{k+1} \\ &\qquad + O(\eta^2) + O(\eta) O(\kappa - 1) + O(s) \Big) + O(s^2)\ . \end{aligned} \tag{3.41}$$

By the Weierstrass preparation theorem, $p_1(\eta, s, \kappa)$ is equivalent to a $(k+1)$ degree polynomial in $(\kappa - 1)$,

$$p_1(\eta, s, \kappa) = \widetilde{p}_1(\eta, s, \kappa) q_1(\eta, s, \kappa) = \left(\sum_{i=0}^{k+1} \alpha_i(\eta, s)(\kappa - 1)^i \right) q_1(\eta, s, \kappa), \tag{3.42}$$

where

$$\begin{aligned} \alpha_0(\eta, s) &= s\Big(b_1^{-1} + O(s) + O(\eta) \Big), \\ \alpha_1(\eta, s) &= \partial_{e_1} df_1(u_0) \eta b_1^{-1} + O(s) + O(\eta^2), \\ \alpha_i(\eta, s) &= O(s) + O(\eta), \qquad 2 \le i \le k, \qquad \alpha_{k+1}(\eta, s) = 1, \end{aligned} \tag{3.43}$$

and q_1 is an analytic function of η, s, $(\kappa - 1)$ which does not vanish at $\eta, s = 0$, $\kappa = 1$. We remark that in view of the accuracy assumption (2.15),

$$C_1(\eta, s, \kappa) = e_1 + C_{10}(s)(\kappa - 1)^k + O(\kappa - 1)^{k+1} + O(\eta) + O(s). \tag{3.44}$$

Without loss of generality, we may assume that the first component $C_1^{(1)}(\eta, s, \kappa) \equiv 1$.

Now we can construct the basis $X_1 = (X_{1,1}, \dots, X_{1,k+1})$ for the invariant space corresponding to the $k+1$ roots of $\widetilde{p}_1(\eta, s, \kappa - 1)$. Note that the image of the projector Q_0 is not changed if we replace the integrand $A d\kappa$ in (3.28) by $\varphi(\kappa) d\kappa$, where $\varphi(\kappa)$ is any holomorphic vector function (e.g., see [8, p. 14]). By (3.20),

$$\begin{aligned} &\oint \Big(\kappa A(\eta, s) + B(\eta, s) \Big)^{-1} \varphi(\kappa) d\kappa \\ &\qquad = \oint D_1(\kappa) \Big(L^{-1}(\eta, s, \kappa) \oplus I \Big) D_2(\eta, s, \kappa) \varphi(\kappa) d\kappa \in \operatorname{Im} Q_0. \end{aligned} \tag{3.45}$$

If we replace $D_2\varphi$ by $\binom{\varphi}{0}$, where $\dim\varphi = n$, and again replace $C^{-1}\varphi$ by $\binom{q_1\varphi}{0}$, where φ is a scalar function and q_1 is as in (3.42), we obtain the vector

$$(3.46)\qquad (2\pi i)^{-1}\oint_{|\kappa-1|=\delta} D_1(\kappa)C_1(\eta,s,\kappa)\Big(\widetilde{p}_1(\eta,s,\kappa)\Big)^{-1}\varphi(\kappa)d\kappa \in \operatorname{Im} Q_0.$$

The vectors $X_{1,i}$, $1\le i\le k+1$, are obtained by substituting, correspondingly,

$$(3.47)\qquad \varphi = \varphi_i = \sum_{j=i}^{k+1}\alpha_j(\eta,s)(\kappa-1)^{j-i}.$$

At $s=\eta=0$, by direct integration we obtain

$$(3.48)\qquad \begin{aligned} X_{1,i}(0) = \left(\frac{d}{d\kappa}\right)^{i-1}\Big(e_1,\kappa e_1,\dots,\kappa^{(\nu_0-1)}e_1\Big)^{T}\Big|_{\kappa=1} \\ +\delta_{i,k+1}\Big(C_{10}(0),C_{10}(0),\dots,C_{10}(0)\Big)^{T}. \end{aligned}$$

Application of B to $X_{1,i}(\eta,s)$ yields

$$(3.49)\qquad \begin{aligned} BX_{1,i} &= (2\pi i)^{-1}\oint_{|\kappa-1|=\delta}(\widetilde{L}-A\kappa)D_1\widetilde{p}_1^{-1}\varphi_i d\kappa \\ &= -A(2\pi i)^{-1}\oint_{|\kappa-1|=\delta} D_1C_1\widetilde{p}_1^{-1}(\varphi_i+\varphi_{i-1}-\alpha_{i-1}\varphi_{k+1})d\kappa \\ &= -A(X_{1,i}+X_{1,i-1}-\alpha_{i-1}X_{1,k+1}). \end{aligned}$$

For $i=1$ the term φ_{i-1} becomes $\widetilde{p}_1$, and hence $X_{1,i-1}$ disappears. Then the basis $X_1(\eta,s)$ satisfies

$$(3.50)\qquad AX_1M_1+BX_1=0,\qquad M_1 = I+\begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ \vdots & & & \ddots & & \\ 0 & & & & 0 & 1 \\ -\alpha_0 & & & & & -\alpha_k \end{bmatrix},$$

where both X_1 and M_1 depend analytically on η and s. If we combine all vectors X_j, $1\le j\le n$, into one basis of $\operatorname{Im} Q_0$,

$$(3.51)\quad X_0=(X_1,X_2,\dots,X_n),\qquad M_0=\oplus_{j=1}^{n}M_j,\qquad M_j=\kappa_j,\qquad j>1,$$

then

$$(3.52)\qquad AX_0M_0+BX_0=0.$$

In view of (3.34), (3.48), the columns of X_0 are linearly independent for small $|\eta|$ and $|s|$. Thus the matrix

$$(3.53)\qquad X=(X_{_0},X_{_I},X_{_{II}})$$

is nonsingular for small $|\eta|$ and $|s|$ and

$$(3.54)\qquad (A\kappa+B)X = T\begin{pmatrix} \kappa I-M_{_0} & 0 & 0 \\ 0 & \kappa I-M_{_I} & 0 \\ 0 & 0 & -\kappa M_{_{II}}+I \end{pmatrix},$$

where

$$T = (AX_0, AX_I, BX_{II}). \tag{3.55}$$

Since $A\kappa + B$ is nonsingular for κ which are not the roots of (3.21), $T = T(\eta, s)$ is nonsingular for small $|\eta|$ and $|s|$.

3.2. The a priori estimate. We now consider (3.19) with boundary conditions (3.10)(b) and (3.7), where u and U are related by (3.6). With X as in (3.53), we define the new variables $Y(x)$ and $y(x)$ by

$$\widetilde{U}(x) = X\Big(\eta(x), s\Big)Y(x), \qquad (E_x - I)Y(x) = y(x) \tag{3.56}$$

so that

$$\begin{aligned}\widetilde{u}(x) &= \Big(u(x), \dots, E_x^{\nu_0 - 1}u(x)\Big)^{T} = (E_x - I)\widetilde{U}(x)\\ &= \Big((E_x - I)X\Big(\eta(x), s\Big)\Big)E_xY(x) + X\Big(\eta(x), s\Big)(E_x - I)Y(x)\\ &= X\Big(\eta(x), s\Big)y(x) + O(\eta')E_xY(x).\end{aligned} \tag{3.57}$$

It is also possible to express Y in terms of y. Indeed, by (3.6) one can express the vector $\widetilde{U}(x_{-\ell})$ in terms of $\widetilde{u}(x_{-\ell})$. Then by (3.56), (3.57),

$$X \cdot Y(x_{-\ell}) = O^*(1)y(x_{-\ell}) + O(\eta')\Big(Y(x_{-\ell}) + y(x_{-\ell})\Big). \tag{3.58}$$

Since X is nonsingular and $O(\eta')$ is small, we obtain

$$Y(x_{-\ell}) = O^*(1)y(x_{-\ell}), \qquad Y(x) = O^*(1)y(x_{-\ell}) + \sum_{\xi = x_{-\ell}}^{x-1} y(\xi). \tag{3.59}$$

In view of (3.54), (3.19) in the Y variables becomes

$$\begin{aligned}&\begin{pmatrix} E_x - M_0 & 0 & 0\\ 0 & E_x - M_I & 0\\ 0 & 0 & -M_{II}E_x + I\end{pmatrix} Y\\ &\qquad = T^{-1}\Big(\widetilde{F} - A\Big((E_x - I)X\Big)E_xY\Big) = H.\end{aligned} \tag{3.60}$$

The vector Y is partitioned into groups Y_0, Y_I, Y_{II} and $Y_0 = (Y_1, \dots, Y_n)$, $Y_1 = (Y_{1,1}, \dots, Y_{1,k+1})$ according to X, and the same applies for vectors y and H. In the case $n_1 = n$ we will treat only system (3.60). In the case $n_1 = 1$ we will consider the equations

$$\begin{aligned}&\text{(a)} \quad (E_x - M_1)Y_1 = H_1,\\ &\text{(b)} \quad (E_x - M_j)y_j = (E_x - I)H_j + O(\eta' s)Y_j = h_j, \qquad 2 \le j \le n,\\ &\text{(c)} \quad (E_x - M_I)y_I = (E_x - I)H_I + O(\eta')Y_I = h_I,\\ &\text{(d)} \quad (-M_{II}E_x + I)y_{II} = (E_x - I)H_{II} + O(\eta')Y_{II} = h_{II}.\end{aligned} \tag{3.61}$$

(The term $O(\eta' s)$ in (3.61)(b) comes from κ_i in (3.32).) Here and elsewhere, $\eta' = \partial_x \eta(x)$, and by (3.15) and (2.86),

$$\varepsilon \|O(\eta')\|_1 + \|O(\eta')\|_\infty \le K\varepsilon^{k+1}. \tag{3.62}$$

The general solution of equation $(E_x - M_j)Y_j = H_j$, $\lambda_j > 0$, could be written as

$$\begin{aligned} &Y_j = C_j \varphi_{\eta,j} + \mathring{Y}_j, \qquad \varphi_{\eta,j} = \prod_{-\ell \le x' \le x-1} M_j\Big(\eta(x'), s\Big), \quad \varphi_{\eta,j}(-\ell) = 1, \\ &\mathring{Y}_j(x+1) = \sum_{x'=-\ell}^{x-1} \prod_{\xi=x'+1}^{x} M_j\Big(\eta(\xi), s\Big) H_j(x') + H_j(x), \end{aligned} \tag{3.63}$$

and we have similar formulas for (3.61)(b),(c):

$$y_j = c_j \varphi_{\eta,j} + \mathring{y}_j, \qquad y_{I} = c_{I} \varphi_{\eta,I} + \mathring{y}_{I}. \tag{3.64}$$

In the case $\lambda_j < 0$, we instead have $\varphi_{\eta,j}(\ell) = 1$ and similar formulas for Y_j, y_j, and y_{II}. In view of (3.27),

$$\|\mathring{y}_{I}\|_1 + \|\mathring{y}_{II}\|_1 \le K\Big(\|h_{I}\|_1 + \|h_{II}\|_1\Big) \tag{3.65}$$

and

$$\|\mathring{y}_{I}\|_\infty + \|\mathring{y}_{II}\|_\infty \le K\Big(\|h_{I}\|_\infty + \|h_{II}\|_\infty\Big). \tag{3.66}$$

Since, by (3.35) for $j > 1$ and $\lambda_j > 0$, $|M_j(\eta, s)| \le 1$, we have

$$\|\mathring{Y}_j\|_\infty \le K\|H_j\|_1 \qquad \text{and} \qquad \|\mathring{y}_j\|_\infty \le K\|h_j\|_1. \tag{3.67}$$

The coefficients C and c in (3.63), (3.64) are determined by the following boundary conditions from (3.10)(b) and (3.7):

$$\begin{aligned} &\text{(a)} \quad P_- dG_2(E_x, e^s)U(x_{-\ell}) = 0, \\ &\text{(b)} \quad P_- \sum_x dG_2(E_x, e^s)u(x) = P_- g_{\text{sum}}, \\ &\text{(c)} \quad dS_2(u_{sh})u = g_2. \end{aligned} \tag{3.68}$$

Note that in view of (3.7) we have

$$\begin{aligned} \sum_x dG_2(E_x, e^s)(E_x - I)U &= dG_2(E_x, e^s)U(x_\ell) - dG_2(E_x, e^s)U(x_{-\ell}) \\ &= dG_2(E_x, e^s)U(x_\ell), \end{aligned} \tag{3.69}$$

and hence (3.68)(b) could be rewritten as

$$P_- dG_2(E_x, e^s)U(x_\ell) = P_- g_{\text{sum}}. \tag{3.70}$$

We start with the case $n_1 = n$. The components $Y_{I,II}$ are written in the form

$$Y_{I} = C_{I} \varphi_{\eta,I} + \mathring{Y}_{I}, \qquad Y_{II} = C_{II} \varphi_{\eta,II} + \mathring{Y}_{II} \tag{3.71}$$

as Y_j in (3.63) and satisfy

$$\|\mathring{Y}_{I}\|_p + \|\mathring{Y}_{II}\|_p \le K\Big(\|H_{I}\|_p + \|H_{II}\|_p\Big), \qquad p = 1, \infty. \tag{3.72}$$

In view of (2.34), (2.35), and (3.48), the contribution of Y_1 to the l.h.s. of (3.68)(a),(c) and (3.70) is bounded by

$$\Big|(E_x - I)^{k+1} Y_1\Big| + \Big|(E_x - I) Y_{1,k+1}\Big| + |\eta'|\,|Y_1| + \Big(|s| + |\eta|\Big)\Big|(E_x - I) Y_1\Big|. \tag{3.73}$$

Note that by the generalized stability condition, the system of equations for $C_{I,II}$ and C_j, $j \ge 1$, is solvable and

$$\begin{aligned} |C_{I,II}| + \sum_{j>1} |C_j| \le K \left(|P_- g_{\rm sum}| + |g_2| + \|H_{I,II}\|_\infty + \sum_{j>1} \|H_j\|_1 \right) \\ + \text{the terms in (3.73).} \end{aligned} \tag{3.74}$$

Hence

$$\|Y_{I,II}\|_1 + \sum_{j>1} \|Y_j\|_\infty \le K\Big(|P_- g_{\rm sum}| + |g_2| + \|H\|_1\Big) + \text{the terms in (3.73).} \tag{3.75}$$

In order to estimate the Y_1 components, it is convenient to introduce the scale

$$\sigma = \varepsilon + |s|^{\frac{1}{k+1}}, \qquad \varepsilon' = \frac{\varepsilon}{\sigma}, \qquad s' = \frac{b_1^{-1} s}{\sigma^{k+1}}, \qquad \tau = x\sigma, \qquad \ell' = \sigma \ell \tag{3.76}$$

and rescale the vectors Y_1, H_1, α,

$$Y'_{1,i} = \sigma^{-i+1} Y_{1,i}, \qquad H'_{1,i} = \sigma^{-i} H_{1,i}, \qquad \alpha'_i = \sigma^{i-k-1} \alpha_i. \tag{3.77}$$

The difference equation for Y'_1 could be written as

$$D_\tau Y'_1 - J Y'_1 + \left(0, \dots, 0, \sum_{i=1}^{k+1} \alpha'_{i-1} Y'_{1,i}\right)^T = D_\tau Y'_1 - M'_1 Y'_1 = H'_1, \tag{3.78}$$

where J is a Jordan cell with zero diagonal and

$$D_\tau = \frac{(E_x - I)}{\sigma} \tag{3.79}$$

approximates the derivative $\frac{d}{d\tau}$. The equations

$$\begin{aligned} &\text{(a)} \quad P_1 dG_2(E_x, e^s) U(x_{-\ell}) = 0, \\ &\text{(b)} \quad P_1 dG_2(E_x, e^s) U(x_\ell) = P_1 g_{\rm sum}, \\ &\text{(c)} \quad dS_1 u(x) = dS_1 (E_x - I) U(x) = g_1, \qquad x = x_{-\ell} \quad \text{or} \quad x = x_\ell, \end{aligned} \tag{3.80}$$

provide $k+1$ boundary conditions for Y'_1. In view of the consistency assumption in (2.5), (3.80)(a),(b) could be written in the form

$$\begin{aligned} Y'_{1,1}(x) + O(\sigma) Y'_1(x) = O(1) Y_{I,II}(x) + \textstyle\sum_{j=2}^n \Big(O(\eta) + O(s)\Big) Y_j(x) \\ + (0, P_1 g_{\rm sum}) = g'_{1,0}, \qquad x = x_{\pm\ell}. \end{aligned} \tag{3.81}$$

In view of (2.30), (3.34), and (3.48), boundary condition (3.80)(c) could be normalized as

$$\begin{aligned} Y'_{1,d_i+2} + O(\sigma)Y'_1 &= \sigma^{-d_i-1}\left(O(1)Y_{I,II} + \sum_{j=2}^{n} \Big(O(\eta) + O(s)\Big) Y_j + O(1) g_{1,i} \right) \\ &= g'_{1,i}, \quad 1 \le i \le \frac{k-1}{2}, \quad x = x_{\pm\ell}. \end{aligned} \tag{3.82}$$

Consider the boundary value problem (3.78), (3.81), (3.82). By (3.43), (3.11), (3.38), and (2.10),

$$\begin{aligned} \sum_{i=1}^{k+1} \alpha'_{i-1} Y'_{1,i} &= s' Y'_{1,1} + \partial_{e_1} df_1(u_0) e_1 \mu b_1^{-1} \sigma^{-k} y_{sh}\Big(\varepsilon(x - x_0)\Big) Y'_{1,2} + O(\sigma) Y'_1 \\ &= s' Y'_{1,1} + (\varepsilon')^k (-1)^{\frac{k+1}{2}} y_{sh}\Big(\varepsilon(x - x_0)\Big) Y'_{1,2} + O(\sigma) Y'_1. \end{aligned} \tag{3.83}$$

Now it is clear that (3.78) approximates the differential equation

$$\partial_\tau^{k+1} Y'_{1,1} + (\varepsilon')^k (-1)^{\frac{k+1}{2}} y_{sh}\Big(\varepsilon'(\tau - \tau_0)\Big) \partial_\tau Y'_{1,1} + s' Y'_{1,1} = \sum_{i=0}^{k} \partial_\tau^i H'_{1,k+1-i}. \tag{3.84}$$

The l.h.s. of (3.84) is the integral of (2.22) with $\partial_\tau Y_{1,1} = y_1$ modulo a change of scale $\varepsilon'\tau \to \tau$ and $s'/[(\varepsilon')^{k+1}(-1)^{\frac{k+1}{2}}] \to s'$. We will show in section 4 that, under the scalar shock stability condition in (2.32) and the continuous stability hypothesis for (2.22), the problem (3.78), (3.81), (3.82) satisfies the estimate

$$\sum_{i+j \le k+2} \|D_\tau^i Y'_{1,j}\|_\infty \le K \left(\sum_{i+j \le k+1} \|D_\tau^i H'_{1,j}\|_{1,\sigma} + |g'_1| \right), \tag{3.85}$$

where $\|H\|_{1,\sigma} = \sum_\tau |H(\tau)| \cdot \sigma$ is the usual ℓ_1 norm on the grid-interval $[-\ell', \ell']$ with step size σ. It is assumed that $\ell\varepsilon \ge \delta_0^{-1}$ is sufficiently large. The constant K is independent of ℓ and σ. For $|s| \gg \varepsilon^{k+1}$ one can replace the norm $\|\cdot\|_{1,\sigma}$ in the r.h.s. of (3.85) by the norm $\|\cdot\|_\infty$ so that (3.85) will become the usual estimate in $\|\cdot\|_\infty$ for elliptic boundary value problems.

By means of (3.85) we can estimate the terms in (3.73). Indeed,

$$\begin{aligned} &\Big|(E_x - I)^{k+1} Y_1\Big| + \Big|(E_x - I) Y_{1,k+1}\Big| + |\eta'|\,|Y_1| + \Big(|s| + |\eta|\Big)\Big|(E_x - I) Y_1\Big| \\ &\le \sum_{j=1}^{k+1} \Big(\|(E_x - I)^{k+2-j} Y_{1,j}\|_\infty + \varepsilon^{k+1} \|Y_{1,j}\|_\infty + \sigma^k \|(E_x - I) Y_{1,j}\|_\infty \Big) \\ &\le \sigma^{k+1} \sum_{i+j \le k+2} \|D_\tau^i Y'_{1,j}\|_\infty \le K\sigma^{k+1} \left(\sum_{i+j \le k+1} \|D_\tau^i H'_{1,j}\|_{1,\sigma} + |g'_1| \right) \\ &\le K\Big(\sigma \|H_1\|_1 + \sigma^{k+1} |g'_1| \Big). \end{aligned} \tag{3.86}$$

Thus, from (3.75) and (3.85) we obtain a combined estimate

$$\|Y_{I,II}\|_1 + \sum_{j>1} \|Y_j\|_\infty \le K\Big(\|H\|_1 + \sigma^{k+1} |g'_1| + |P_- g_{\text{sum}}| + |g_2| \Big). \tag{3.87}$$

The grid-function H in (3.60) is $O(1)\widetilde{F} + O(\eta')Y$; hence

$$\|H\|_1 \le K\Big(\|\widetilde{F}\|_1 + \|\eta'\|_1 \cdot \|Y\|_\infty\Big) \le K\Big(\|\widetilde{F}\|_1 + \varepsilon^k\|Y\|_\infty\Big). \tag{3.88}$$

The term $\varepsilon^k\|Y\|_\infty$ is absorbed in the l.h.s. of (3.87) with the exception of the component $\varepsilon^k\|Y_1\|_\infty$. Now we should estimate more carefully the contribution of $O(\eta')Y$ to the r.h.s. of (3.85). We have

$$\|D_\tau^i\sigma^{-j}O(\eta')Y\|_{1,\sigma} \le K\sigma^{-i-j}\varepsilon^k \cdot \sigma\|Y\|_\infty \le K\|Y\|_\infty. \tag{3.89}$$

This is a sufficiently nice estimate for the components $Y_{I,II}$, Y_j, $j \ge 2$, and $Y_{1,j} = \sigma^{j-1}Y'_{1,j}$, $j > 1$. The latter are dominated by the l.h.s. of (3.85). For $Y_{1,1}$,

$$\|D_\tau^i\sigma^{-j}O(\eta')Y_{1,1}\|_{1,\sigma} \le K\sigma^{-j}\varepsilon^k \cdot \sigma\sum_{i_1\le i}\|D_\tau^{i_1}Y_{1,1}\|_\infty. \tag{3.90}$$

The worst case is $i = 0$, $j = k+1$. This comes from the contribution of $T^{-1}A(E_x - I)X_{1,1}E_xY_{1,1}$ in (3.60) to the component $H_{1,k+1}$. Note, however, that this contribution is not $O(\eta')$ but $O(\eta's)$. Indeed, for $s = 0$ the vectors $X_j(\eta)$, $j \ge 2$, and $X_{1,1}(\eta)$ are eigenvectors of the pencil $A\kappa + B$ with eigenvalues $\kappa = 1$ and form a basis in the space V_0 of repeated ν_0+1 tuples $(u, u, \dots, u)$, $u \in \mathbb{C}^n$. Since $(E_x - I)X_{1,1}(\eta) \in V_0$, it is a linear combination of the above $X_j(\eta)$, $X_{1,1}(\eta)$. Hence the above $O(\eta')Y_{1,1}$ term for $s = 0$ contributes only to H_j, $j \ge 2$, and $H_{1,1}$. Thus estimate (3.85) implies

$$\sum_{ij\le k+2}\|D_\tau^iY'_{1,j}\|_\infty \le K\left(\sigma^{-k}\|\widetilde{F}\|_1 + |g'_1| + \|Y_{I,II}\|_\infty + \sum_{j\ge2}\|Y_j\|_\infty\right) \tag{3.91}$$

and

$$\begin{aligned}\|Y_{I,II}\|_1 + \sum_{j>1}\|Y_j\|_\infty + \sigma^k\sum_{i+j\le k+2}\|D_\tau^iY'_{1,j}\|_\infty \\ \le K\Big(\|\widetilde{F}\|_1 + \sigma^k|g'_1| + |P_-g_{\mathrm{sum}}| + |g_2|\Big).\end{aligned} \tag{3.92}$$

Recall that $\widetilde{F}$ is related to the original F by formulas (3.17), (3.12), and (3.8). Hence

$$\begin{aligned}\|\widetilde{F}\|_1 \le \left\|\sum_{x_{-\ell}}^{x-1}F(\xi)\right\|_1 + \Big|dG_1[u_{st}(x_{-\ell})](E_x, e^s)u(x_{-\ell})\Big| \cdot \ell \\ +\|O(\Delta u_{st})u\|_1 + \|O(\eta')u\|_1.\end{aligned} \tag{3.93}$$

Since $n_1 = n$, PdG_1u in (2.26)(a) is dG_1u, and in the notation of (3.2), (3.3) we have

$$dG_1\Big[u_{st}(x_{-\ell})\Big](E_x, e^s)u(x_{-\ell}) = g_{L,0}. \tag{3.94}$$

Next, by (3.14), (3.15),

$$\|O(\Delta u_{st})u\|_1 + \|O(\eta')u\|_1 \le \delta_0\varepsilon^{k-1}\|u\|_\infty + K\varepsilon^k\|u\|_\infty. \tag{3.95}$$

The grid-function u could be expressed in terms of Y (see (3.57)). These terms are negligible with respect to the l.h.s. of (3.92). The only problematic term comes from the contribution of $(E_x - I)Y_1$ to u (see (3.57)). However,

$$\delta_0\varepsilon^{k-1}\|(E_x - I)Y_{1,j}\|_\infty \le \delta_0\varepsilon^{k-1}\sigma^j\|D_\tau Y'_{1,j}\|_\infty \ll \sigma^k\|D_\tau Y'_{1,j}\|_\infty. \tag{3.96}$$

From (3.82) we obtain

$$|\sigma^k g'_{1,i}| \le \sigma^{k-d_i-1}\left(\|Y_{I,II}\|_\infty + \sigma^k \sum_{j=2}^{n} \|Y_j\|_\infty + |g_{1,i}|\right). \tag{3.97}$$

Because of the $Y_{I,II}$ term, we have to assume that

$$0 \le d_i \le k-2, \qquad 1 \le i \le \frac{k-1}{2}. \tag{3.98}$$

Note that for $k = 1$ the set of $\{d_i\}$ is empty. For odd $k \ge 3$ the restriction in (3.98) does not exclude the set of boundary conditions suggested in (2.59). Hence, the terms with $Y_{I,II}$, Y_j in $\sigma^k g'_{1,i}$ are negligible. Altogether, from (3.98) we arrive at the estimate

$$\begin{aligned}&\|Y_{I,II}\|_1 + \sum_{j>1}\|Y_j\|_\infty + \sigma^k \sum_{i+j\le k+2} \|D_\tau^i Y'_{1,j}\|_\infty \\ &\le K\left(\left\|\sum_{x-\ell}^{x-1} F(\xi)\right\|_1 + \sigma^k \sum |\sigma^{-d_i-1} g_{1,i}| + |P_- g_{\mathrm{sum}}| + \sigma^k |P_1 g_{\mathrm{sum}}| + |g_2| + |g_{L,0}|\cdot \ell\right).\end{aligned} \tag{3.99}$$

From here we can obtain estimates for the variable U. By (3.48), the contribution of Y_1 to the components U_j, $j > 1$, is bounded by $(O(s) + O(\eta))Y_1 + O(1)Y_{1,k+1}$. Hence

$$\sum_{j>1}\|U_j\|_\infty + \sum_{i\le k}\|\sigma^{k-i}(E_x - I)^i U_1\|_\infty \le \text{r.h.s. of (3.99)}. \tag{3.100}$$

Now we consider the most important case, $n_1 = 1$. The single global conservation law prevents the shock wave from "shifting." Let us return to (3.61)(b)–(d) with the solutions (3.64). Since the projector P_- is zero, the coefficients c_j, $j > 1$, and $c_{I,II}$ are determined by the boundary conditions (3.68)(c). In order to express these boundary conditions in terms of y_j, $y_{I,II}$, we use the relations (3.57) and (3.59). Hence

$$\begin{aligned}&dS_2(u_{sh})\left(\sum_{j>1} X_j y_j + X_I y_I + X_{II} y_{II}\right) \\ &\qquad = g_2 + O(\eta')Y - dS_2(u_{sh}) X_1 y_1.\end{aligned} \tag{3.101}$$

Recall that the contribution of $y_1 = (E_x - I)Y_1$ to the r.h.s. of (3.101) is bounded by the terms in (3.73). Hence

$$\begin{aligned}&\sum_{j>1}|c_j| + |c_I| + |c_{II}| \\ &\le K\cdot\left(|g_2| + \sum_{j>1}\|\mathring{y}_j\|_\infty + \|\mathring{y}_{I,II}\|_\infty + \varepsilon^{k+1}\|Y\|_\infty + \text{the terms in (3.73)}\right),\end{aligned} \tag{3.102}$$

and by (3.65), (3.67),

$$\begin{aligned}&\sum_{j>1}\|y_j\|_\infty + \|y_{I,II}\|_1 \\ &\le K\cdot\left(\|h_{I,II}\|_1 + \sum_{j>1}\|h_j\|_1 + |g_2| + \varepsilon^{k+1}\|Y\|_\infty + \text{terms in (3.73)}\right).\end{aligned} \tag{3.103}$$

The variable Y_1 was already estimated in (3.85). Unfortunately, the contribution of the constant term $dG_1 \cdot u(x_{-\ell})$ in (3.8) to the norm $\|H'_{1,j}\|_{1,\sigma}$ in (3.85) for $j = k$ results in terms like $\|y_i\|_\infty \ell \sigma^{-k+1}$, $i > 1$. This term contributes to $\varepsilon^{k+1}\|Y_1\|_\infty$ a value $\ell\varepsilon^2|y_j|_\infty$, which is not negligible if $\ell = o(1)\varepsilon^{-k-1}$, $k > 1$. Therefore we will improve estimate (3.85) in the following manner. Define the variables

$$\widetilde{Y}'_{1,i} = Y'_{1,i} + H'_{1,i-1}, \qquad i \geq 2; \qquad \widetilde{Y}'_{1,1} = Y'_{1,1}. \tag{3.104}$$

Then $\widetilde{Y}'_1$ satisfies system (3.78) with H'_1 replaced by

$$\begin{aligned} &\widetilde{H}'_{1,1} = 0; \qquad \widetilde{H}'_{1,i} = D_\tau H'_{1,i-1}, \qquad 2 \leq i \leq k; \\ &\widetilde{H}'_{1,k+1} = H'_{1,k+1} + D_\tau H'_{1,k} + \textstyle\sum_{i=1}^k \alpha'_i H'_{1,i}. \end{aligned} \tag{3.105}$$

By (3.43) and (3.77),

$$\sum_{i=1}^k \alpha'_i H'_{1,i} = \sum_{i=1}^k \sigma^{i-k-1}\Big(O(\eta) + O(s)\Big)\sigma^{-i} H_{1,i} = O(\sigma^{-1})H_1. \tag{3.106}$$

In boundary conditions (3.81), (3.82), one should replace $g'_{1,i}$ by

$$\widetilde{g}'_{1,i} = g'_{1,i} + H'_{1,d_i+1} + O(\sigma)H'_1. \tag{3.107}$$

Now, from the basic estimate (3.85) for $\widetilde{Y}'_1$, we obtain

$$\begin{aligned} \sum_{i+j\leq k+2} \|D_\tau^i Y'_{1,j}\|_\infty \leq K\Bigg(&\sum_{i+j\leq k} \|D_\tau^i D_\tau H'_{1,j}\|_{1,\sigma} + \|H'_{1,k+1}\|_{1,\sigma} \\ &+ \|\sigma^{-1}H_1\|_{1,\sigma} + \sum_{\substack{i+j\leq k+1 \\ j\leq k}} \|D_\tau^i H'_{1,j}\|_\infty + |g'_1| \Bigg). \end{aligned} \tag{3.108}$$

In order to estimate the component $H_{1,k+1}$, we will need the following elementary lemma.

LEMMA 3.1. *The $H_{1,k+1}$ component of the vector H in* (3.60) *depends on $F^{(1)}$ in* (3.10)(a) *as*

$$H_{1,k+1} = O(1)P_1 F^{(1)} + \Big(O(\eta) + O(s)\Big)(I - P_1)F^{(1)}. \tag{3.109}$$

(Here P_1 is the projection on the first component.)

Proof. Let $\eta, s = 0$. Take $F^{(1)}$ to be a constant grid-function with $P_1 F^{(1)} = 0$. Then (3.10)(a) has a solution $P_j U = x \cdot P_j F^{(1)}/\lambda_j$, $j \geq 2$, $P_1 U = 0$. Here P_j is the projection on the jth component. The corresponding vector functions Y and H defined by (3.17), (3.56), and (3.60) satisfy system (3.60). In particular, Y_1 satisfies $(E_x - I)Y_1 - JY_1 = H_1$. Notice, however, that $\widetilde{U}$ in (3.17) is a first order polynomial in x and so is Y. On the other hand, $F^{(1)}$, $\widetilde{F}$, and H are constants. Hence, $JY_1 = H_1 - (E_x - I)Y_1$ is constant. In particular, $Y_{1,k+1}$ is constant and $H_{1,k+1} = (E_x - I)Y_{1,k+1} = 0$. Since $H(x)$ depends only on $F^{(1)}(x)$, we obtain that $H_{1,k+1}$ for $s, \eta = 0$ is independent of $(I - P_1)F^{(1)}$. Hence for small s, η, $H_{1,k+1}$ satisfies (3.109).

Now we can express the r.h.s. of (3.108) in terms of $F^{(1)}$ in (3.10)(a) as

$$\begin{aligned}\sum_{i+j\le k+2}\|D_\tau^i Y'_{1,j}\|_\infty \le K\Big(&\|\sigma^{-k}(E_x-I)F^{(1)}\|_1+\|\sigma^{-k}P_1F^{(1)}\|_1\\ &+\|F^{(1)}\|_1+\|\sigma^{-k-1}(E_x-I)F^{(1)}\|_\infty+\|\sigma^{-k}F^{(1)}\|_\infty+|g'_1|+\|Y\|_\infty\Big).\end{aligned}\tag{3.110}$$

We already have shown in (3.90) that the actual contribution of $Y_{1,1}$ to $\|Y\|_\infty$ is negligible. Thus $\|Y\|_\infty$ above could be replaced by $\|Y_{I,II}\|_\infty+\sum_{j\ge2}\|Y_j\|_\infty$. The contribution of $dG_1u(x_{-\ell})$ in (3.8) to the r.h.s. of (3.110) is bounded by

$$\Big|\sigma^{-k}P_1dG_1u(x_{-\ell})\Big|\ell+\Big|(I-P_1)dG_1u(x_{-\ell})\Big|\cdot(\ell+\sigma^{-k}).\tag{3.111}$$

By the accuracy assumption in (2.15) and formulas (3.48), (3.57),

$$\begin{aligned}\Big|(I-P_1)dG_1[u_0]u(x_{-\ell})\Big|&\le K\Big(|(I-P_1)u|_\infty+|(E_x-I)^{k+1}P_1u|_\infty\Big)\\ &\le K\left(\|y_{I,II}\|_\infty+\sum_{j\ge2}\|y_j\|_\infty+\|y_{1,k+1}\|_\infty+\sigma^k\|y\|_\infty+\varepsilon^{k+1}\|Y\|_\infty\right).\end{aligned}\tag{3.112}$$

Since $dG_1u(x_{-\ell})$ is actually $dG_1[u_{st}(x_{-\ell})](E_x,e^s)u(x_{-e})$, by (2.86)–(2.87) we should add to the above the term $K\varepsilon^k\|y\|_\infty$, which is absorbed in $\sigma^k\|y\|_\infty$. The component $P_1dG_1u(x_{-\ell})$ is nothing but $P_1g_{L,0}$ (see (3.3)). Next, the contribution of the term $(O(\Delta u_{st})+O(\eta'))u$ in (3.12) to the r.h.s. of (3.110) is bounded by

$$\begin{aligned}K\Big(&\Big\|\sigma^{-k}\Big(O(\Delta u_{st})+O(\eta')\Big)u\Big\|_1\\ &+\Big\|\sigma^{-k-1}\Big(O(\Delta u_{st})+O(\eta')\Big)u\Big\|_\infty\Big)\le K\delta_0\sigma^{-1}\|u\|_\infty.\end{aligned}\tag{3.113}$$

What is left from $F^{(1)}$ in (3.8) is the sum $\sum_{x_{-\ell}}^{x-1}F(\xi)$. Its contribution to the r.h.s. of (3.110) is bounded by

$$\begin{aligned}K\Big(&\|\sigma^{-k}F\|_1+\|\sigma^{-k-1}F\|_\infty\\ &+\Big\|\sigma^{-k}P_1\sum F(\xi)\Big\|_1+\Big\|\sum F(\xi)\Big\|_1+\Big\|\sigma^{-k}\sum F(\xi)\Big\|_\infty\Big).\end{aligned}\tag{3.114}$$

Finally, consider the term g'_1. Instead of (3.82), we rewrite boundary conditions (3.80)(c) as

$$\begin{aligned}Y'_{1,d_i+2}+O(\sigma)Y'_1&=\sigma^{-d_i-1}\left(O(1)y_{I,II}+\sum_{j\ge2}O(1)y_j+O(\eta')Y+O(1)g_{1,i}\right)\\ &=g'_{1,i},\qquad i\ge1.\end{aligned}\tag{3.115}$$

Hence

$$|g'_{1,i}|\le K\sigma^{-d_i-1}\left(\|y_{I,II}\|_\infty+\sum_{j\ge2}\|y_j\|_\infty+\varepsilon^{i+1}\|Y\|_\infty+|g_{1,i}|\right),\quad i\ge1.\tag{3.116}$$

By (3.81),

$$|g'_{1,0}| \le K\left(\|Y_{I,II}\|_\infty + \sigma^k \sum_{j\ge 2}\|Y_j\|_\infty + |P_1 g_{\rm sum}|\right). \tag{3.117}$$

Altogether,

$$\begin{aligned}\sum_{i+j\le k+2}\|D_\tau^i Y'_{1,j}\|_\infty \le K\Big(&\|\sigma^{-k}F\|_1 + \|\sigma^{-k-1}F\|_\infty + \Big\|\sigma^{-k}P_1\sum F(\xi)\Big\|_1\\ &+\Big\|\sum F(\xi)\Big\|_1 + \Big\|\sigma^{-k}\sum F(\xi)\Big\|_\infty + \sigma^{-k}|g_{L,0}|\cdot\ell + |P_1 g_{\rm sum}|\\ &+\sum_{i\ge 1}\sigma^{-d_i-1}|g_{1,i}| + (\ell+\sigma^{-k})\Big(\|y_{I,II}\|_\infty + \sum_{j\ge 2}\|y_j\|_\infty + \|y_{1,k+1}\|_\infty\\ &+\sigma^k\|y_1\|_\infty + \varepsilon^{k+1}\|Y\|_\infty\Big) + \delta_0\sigma^{-1}\|u\|_\infty + \|Y_{I,II}\|_\infty + \sum_{j\ge 2}\|Y_j\|_\infty\Big).\end{aligned} \tag{3.118}$$

Unlike in (3.98), we can relax the restriction on d_i to be

$$0 \le d_i \le k-1, \qquad 1 \le i \le \frac{k-1}{2}. \tag{3.119}$$

The contribution of the y_1 and Y_1 components to the r.h.s. of (3.118) is bounded by

$$\begin{aligned}K\Big((\ell+\sigma^{-k})\Big(&\|\sigma^{k+1}D_\tau Y'_{1,k+1}\|_\infty + \sigma^k\|\sigma D_\tau Y_1\|_\infty + \varepsilon^{k+1}\|Y_1\|_\infty\Big)\\ &+\delta_0\sigma^{-1}\|\sigma D_\tau Y_1\|_\infty\Big).\end{aligned} \tag{3.120}$$

Recall that we consider $|s| \le K\varepsilon^{k+1}$. Since $\ell\varepsilon^{k+1} \le \delta_0$, for sufficiently small δ_0, $K\ell\sigma^{k+1}$ is small. Hence the terms in (3.120) are dominated by the l.h.s. of (3.118). By (3.59),

$$\|Y_{I,II}\|_\infty + \sum_{j\ge 2}\|Y_j\|_\infty \le K\|y\|_\infty + \ell\left(\|y_{I,II}\|_\infty + \sum_{j\ge 2}\|y_j\|_\infty\right). \tag{3.121}$$

Hence all y, u, and Y terms in the r.h.s. of (3.118) could be replaced by

$$K(\ell+\sigma^{-k})\left(\|y_{I,II}\|_\infty + \sum_{j\ge 2}\|y_j\|_\infty\right). \tag{3.122}$$

We now return to estimate (3.103). In view of (3.121), the term $\varepsilon^{k+1}(\|Y_{I,II}\|_\infty + \sum_{j\ge 2}\|Y_j\|_\infty)$ is dominated by the l.h.s. of (3.103) (modulo the negligible term $K\varepsilon^{k+1}\|y_1\|_\infty$). By (3.86), the terms in (3.73) are bounded by σ^{k+1} times the l.h.s. of

(3.118). The same holds for $\varepsilon^{k+1}\|Y_1\|_\infty$. The term $\|h\|_1$ is estimated as

$$\begin{aligned}
&\|h_{I,II}\|_1+\sum_{j\ge 2}\|h_j\|_1\le\|(E_x-I)H_{I,II}\|_1+\sum_{j\ge 2}\|(E_x-I)H_j\|_1\\
&\quad+\|O(\eta')\|_1\Big(\|Y_{I,II}\|_\infty+|s|\sum_{j\ge 2}\|Y_j\|_\infty\Big)\\
&\quad\le K\Big(\|(E_x-I)\widetilde F\|_1+\|O(\eta')\|_1\|\widetilde F\|_\infty+\|O(\eta')\|_1\\
&\qquad\cdot\Big(\|y\|_\infty+\varepsilon\|Y\|_\infty\Big)+\varepsilon^k\Big(\|y_{I,II}\|_1+\sigma^{k+1}\ell\sum_{j\ge 2}\|y_j\|_\infty+\sigma^{k+1}\|y\|_\infty\Big)\Big).
\end{aligned}\tag{3.123}$$

Clearly, the $y_{I,II}$ and y_j, $j\ge 2$, terms are negligible, and the y_1, Y_1 terms are bounded by σ^{k+1} times the l.h.s. of (3.118). Now consider the $\widetilde F\approx F^{(1)}$ terms in the r.h.s. of (3.123). The contribution of $dG_1u(x_{-\ell})$ in (3.8) to $\|O(\eta')\|_1\|\widetilde F\|_\infty$ is

$$\begin{aligned}
\varepsilon^k\Big|dG_1[u_{st}](x_{-\ell})(E_x,e^s)u(x_{-\ell})\Big|&\le\varepsilon^k\Big(\|y\|_\infty+\varepsilon^{k+1}\|Y\|_\infty\Big)\\
&\le\varepsilon^k\Big(\|y\|_\infty+\varepsilon^{k+1}\|Y\|_\infty\Big)\le\varepsilon^k(1+\varepsilon^{k+1}\ell)\|y\|_\infty
\end{aligned}\tag{3.124}$$

and is already included in the $\varepsilon^k\|y\|_\infty$ term of the r.h.s. of (3.123). The contribution of $(O(\Delta u_{st})+O(\eta'))u$ in (3.12) is bounded by

$$\begin{aligned}
&\Big\|\Big((E_x-I)\Big(O(\Delta u_{st})+O(\eta')\Big)\Big)\Big\|_1\cdot\|u\|_\infty\\
&\qquad+\Big(\|\Delta u_{st}\|_1+\|\eta'\|_1\Big)\|(E_x-I)u\|_\infty\\
&\qquad\le K\varepsilon^k\|u\|_\infty+\delta_0\varepsilon^{k-1}\|(E_x-I)u\|_\infty
\end{aligned}\tag{3.125}$$

(see (2.87), (2.88)). The norms of u and $(E_x-I)u$ could be expressed in terms of y and Y using (3.57). Namely,

$$\|u\|_\infty\le\|y\|_\infty+\varepsilon^{k+1}\|Y\|_\infty\tag{3.126}$$

and

$$\|(E_x-I)u\|_\infty\le K\Big(\varepsilon^{k+1}\|y\|_\infty+\|(E_x-I)y\|_\infty+\varepsilon^{k+2}\|Y\|_\infty\Big).\tag{3.127}$$

The contribution of y_j, $j\ge 2$, and $y_{I,II}$ components to the r.h.s. of (3.125) is negligible compared with the l.h.s. of (3.103). For the Y_1 component, the worst term is

$$K\varepsilon^k\|y_1\|_\infty+\delta_0\varepsilon^{k-1}\|(E_x-I)y_1\|_\infty\le K\sigma^{k+1}\Big(\|D_\tau Y_1\|_\infty+\|D_\tau^2Y_1\|_\infty\Big)\tag{3.128}$$

and is bounded by $K\sigma^{k+1}$ times the l.h.s. of (3.118). Finally, in the $|g_2|$ term at the r.h.s. of (3.103) we have the contribution of $(O(\Delta u_{st})+O(\eta'))u$ in (3.13). This is bounded by $\delta_0\varepsilon^k\|u\|_\infty$.

We are ready to derive the combined estimate. Multiply (3.118) by $K_1\sigma^{k+1}$, where K_1 is a large constant, and add the result to (3.103). We obtain

$$\begin{aligned}
&K_1\sigma^{k+1}\sum_{i+j\le k+2}\|D_\tau^i Y'_{1,j}\|_\infty+\sum_{j>1}\|y_j\|_\infty+\|y_{I,II}\|_1\\
&\quad\le K_1K\sigma^{k+1}\bigg(\|\sigma^{-k}F\|_1+\|\sigma^{-k-1}F\|_\infty+\Big\|\sigma^{-k}P_1\sum F(\xi)\Big\|_1\\
&\qquad+\Big\|\sum F(\xi)\Big\|_1+\Big\|\sigma^{-k}\sum F(\xi)\Big\|_\infty\\
&\qquad+\sigma^{-k}|g_{L,0}|\cdot\ell+|P_1g_{\rm sum}|+\sum_{i\ge1}\sigma^{-d_i-1}|g_{1,i}|\bigg)\\
&\quad+K\Big(\|F\|_1+\varepsilon^k\Big\|\sum F(\xi)\|_\infty+|g_2|\Big).
\end{aligned}\tag{3.129}$$

Indeed, $K_1\sigma^{k+1}$ times the terms in (3.122) results in $K_1K(\ell\sigma^{k+1}+\sigma)(\|y_{I,II}\|_\infty+\sum_{j>1}\|y_j\|_\infty)$. Since $\sigma\approx\varepsilon$, these terms are bounded by the l.h.s. of (3.129), provided $K_1K\delta_0\ll1$. In order to neutralize the contribution of Y_1 to the r.h.s. of (3.103), it is enough that $K_1>K$. Since $\sigma^{k+1}\ell\approx\varepsilon^{k+1}\ell\ll1$, estimate (3.129) could be rewritten in an equivalent form

$$\begin{aligned}
&\sigma^{k+1}\sum_{i+j\le k+2}\|D_\tau^i Y'_{1,j}\|_\infty+\sum_{j>1}\|y_j\|_\infty+\|y_{I,II}\|_1\\
&\quad\le K\bigg(\|F\|_1+\sigma\Big\|P_1\sum F(\xi)\Big\|_1+\sigma\ell|g_{L,0}|\\
&\qquad+\sigma^{k+1}|P_1g_{\rm sum}|+\sum_{i\ge1}\sigma^{k-d_i}|g_{1,i}|+|g_2|\bigg).
\end{aligned}\tag{3.130}$$

Since $\sigma D_\tau Y_1=(E_x-I)Y_1=y_1$,

$$\sigma^k\sum_{i+j\le k+1}\|D_\tau^i\sigma^{-j+1}y_{1,j}\|_\infty+\|y_{I,II}\|_1+\sum_{j>1}\|y_j\|_\infty\le\text{r.h.s. of (3.130)}.\tag{3.131}$$

In the original u variables, we obtain

$$\sum_{j>1}\|u_j\|_\infty+\sum_{i=-1}^{k}\|\sigma^{k-i}(E_x-I)^iu_1\|_\infty\le\text{r.h.s. of (3.130)},\tag{3.132}$$

where $(E_x-I)^{-1}u_1\stackrel{df}{=}\sum_\xi u_1(\xi)$.

In case $K\varepsilon^{k+1}<|s|<\delta$, the norm $\|\cdot\|_1$ in (3.110) is replaced by the $\sigma^{-1}\|\cdot\|_\infty$ norm. As a result, the terms $\|\sigma^{-k}P_1\sum F(\xi)\|_1+\|\sum F(\xi)\|_1$ in (3.129) are replaced by $\sigma^{-k-1}\|P_1\sum F(\xi)\|_\infty+\sigma^{-1}\|\sum F(\xi)\|_\infty\le\sigma^{-k-1}\|F\|_1$. Therefore we obtain estimate (3.130) without the term $\sigma\|P_1\sum F(\xi)\|_1$. Finally, instead of (3.132) we obtain

$$\begin{aligned}
&\sum_{j>1}\|u_j\|_\infty+\sum_{i=-1}^{k}\|\sigma^{k-i}(E_x-I)^iu_1\|_\infty\\
&\quad\le K\left(\|F\|_1+|g_{L,0}|+\sigma^{k+1}|P_1g_{\rm sum}|+\sum_{i\ge1}\sigma^{k-d_i}|g_{1,i}|+|g_2|\right).
\end{aligned}\tag{3.133}$$

In order to prove exponential decay of the solution u, we need to extend estimates (3.100), (3.132), and (3.133) into the domain

$$0 \geq \operatorname{Re} s \geq -\frac{\delta\varepsilon^k}{\ell}. \tag{3.134}$$

Indeed, the bound $\delta\varepsilon^k/\ell$ cannot be increased, since the differential equation in (3.84) (with r.h.s. $= 0$) has for small $|s'|$ and large $|\tau|$ the exponential solutions with exponents $\approx s'\tau/(\varepsilon')^k$. If $\operatorname{Re} s' < 0$, then these exponents reach on the interval $[-\ell, \ell]$ the values of the order $\exp((\operatorname{Re} s)/\varepsilon^k \cdot \ell)$. Since these values have to be bounded, we must impose the bound in (3.134). In order to extend the estimates, we write $s = \operatorname{Im} s + \operatorname{Re} s$ and move the $O(\operatorname{Re} s)u$ terms in (3.2)(a),(b) to the r.h.s. of the equations. Note that the $O(s)u$ term in (3.2)(a) is

$$\begin{aligned} O(s)u &= (e^s - 1)dG_2(u) + O(s)(E_x - I)O(u) \\ &= (e^s - 1)u + (E_x - I)O(s)u + O(s^2)u, \end{aligned} \tag{3.135}$$

where $O(s^2)$ has constant coefficients. Hence, the contribution of the $O(\operatorname{Re} s)u$ term to the r.h.s. of (3.99) is bounded by

$$|\operatorname{Re} s| \left(\left\| \sum_\xi u(\xi) \right\|_1 + \|u\|_\infty \right) \leq \delta\varepsilon^k \|U\|_\infty \tag{3.136}$$

and is bounded by the l.h.s. of (3.100). In case $n_1 = 1$, $|s| \leq K\varepsilon^{k+1}$, the contribution of $O(\operatorname{Re} s)u$ to the r.h.s. of (3.130) is estimated as

$$\begin{aligned} &|\operatorname{Re} s| \left(\sigma \left\| \sum_\xi P_1 u(\xi) \right\|_1 + |s|\sigma \left\| \sum_\xi u(\xi) \right\|_1 + \|u\|_1 \right) \\ &\leq \left(\delta\sigma\varepsilon^k \left\| \sum_\xi P_1 u(\xi) \right\|_\infty + \delta\varepsilon^k \|u\|_\infty \right) \end{aligned} \tag{3.137}$$

and is bounded by the l.h.s. of (3.132). The contribution of $O(\operatorname{Re} s)u$ to the term $\|F\|_1$ in (3.133) is negligible. Clearly, for $\operatorname{Re} s$ as in (3.134) and $\ell \gg \varepsilon^{-1}$,

$$\varepsilon + |\operatorname{Im} s|^{1/(k+1)} \sim \varepsilon + |s|^{1/(k+1)}. \tag{3.138}$$

Thus, estimates (3.100), (3.132), (3.133) are valid also for s as in (3.134).

4. The basic stability estimate for the scalar problem. In this section we will prove the basic stability estimate (3.85) for problem (3.78), (3.81), (3.82). Let us first consider the analogous differential equation

$$\partial_\tau Y - MY = H, \tag{4.1}$$

where

$$M = \begin{pmatrix} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ -\alpha_0' - \alpha_1' & & & -\alpha_k' \end{pmatrix}, \qquad M \approx \sigma^{-1} \log(I + \sigma M_1'), \quad M_1' \text{ in (3.78)}.$$

The coefficients $\alpha_i' \bmod O(\sigma)$ coincide with α_i' in (3.78). They depend analytically on $y_{sh}(\varepsilon'(\tau-\tau_0))$, s', ε', and σ and $\alpha_0' = (-1)^{\frac{k+1}{2}} s' + O(\sigma)$, $\alpha_1' = (-1)^{\frac{k+1}{2}} (\varepsilon')^k y_{sh} + O(\sigma)$, $\alpha_i' = O(\sigma)$, $i \geq 2$. Equation (4.1) is solved on interval $-\ell' \leq \tau \leq \ell'$ with boundary conditions at $\tau = \pm\ell'$,

$$Y_1 + O(\sigma)Y = g_0, \qquad Y_{d_i+2} + O(\sigma)Y = g_i, \qquad 1 \leq i \leq \frac{k-1}{2}. \tag{4.2}$$

The characteristic equation for (4.1) at $\tau = \pm\infty$ is correspondingly

$$(-1)^{\frac{k+1}{2}} \lambda^{k+1} \mp (\varepsilon')^k \lambda + s' + O(\sigma) = 0 \qquad \text{(see (3.84))} \tag{4.3}$$

(where $O(\sigma)$ depends also on λ, ε', s'). We consider the parameter vector $p = (\varepsilon', s', \sigma)$ in a neighborhood of a point $p_0 = (\varepsilon_0', s_0', 0)$, where

$$\varepsilon_0' + |b_1 s_0'|^{\frac{1}{k+1}} = 1, \qquad \varepsilon_0' > 0, \qquad \operatorname{Re} s_0' \geq 0 \qquad \text{(see (3.76))}. \tag{4.4}$$

If $s_0' \neq 0$, then the roots of (4.3) split into groups I and II with $\frac{k+1}{2}$ eigenvalues λ' correspondingly in the half-planes $\operatorname{Re}\lambda < 0$ or $\operatorname{Re}\lambda > 0$. If $s_0' = 0$, then there is a simple root λ_0 near 0. In order for $\operatorname{Re}\lambda_0 \neq 0$, we should assume that $\operatorname{Re} s \geq 0$ and $\sigma \geq 0$. In order to get an estimate for (4.1), (4.2), we split the interval $(-\ell', \ell')$ into three subintervals $(-\ell', -\ell_0')$, $(-\ell_0', \ell_0')$, (ℓ_0', ℓ'). We will assume that $\ell' \geq \delta_0^{-1}$ (as in (1.11), (1.12)) and could be arbitrarily large. The number ℓ_0' will be specified in what follows. On the interval $[\ell_0', \ell']$ we consider (4.1) as a perturbation of a constant coefficient problem

$$\partial_\tau Y - M(\infty)Y = \Delta M Y + H = F, \qquad \Delta M = M - M(\infty). \tag{4.5}$$

At $\tau = \ell'$ we impose the boundary conditions (4.2). At $\tau = \ell_0'$ the boundary condition is

$$P_{I}(\tau)Y = g^{(0)}, \qquad \tau = \ell_0', \tag{4.6}$$

where $P_{I}(\tau)$ is an orthogonal projection on the $\frac{k+1}{2}$ dimensional subspace $V_{I}(\tau)$ of exponentially decreasing solutions of (4.1). Note that the "near zero" root λ_0 belongs to group II. The interval $[-\ell', -\ell_0']$ is treated in a similar way. The solution of problem (4.5), (4.6), (4.2) on $[\ell_0', \ell']$ and $[-\ell', \ell_0']$ will be expressed as

$$Y = G_1\Big(H, g(\ell'), g^{(0)}(\ell_0')\Big), \qquad Y = G_{-1}\Big(H, g(-\ell'), g^{(0)}(-\ell_0')\Big). \tag{4.7}$$

On the interval $[-\ell_0', \ell_0']$, (4.1) is solved as an initial value problem. Thus we obtain

$$Y(\ell_0') = G_0\Big(H, Y(-\ell_0')\Big). \tag{4.8}$$

Finally, we substitute the values of $Y(-\ell_0')$, $Y(\ell_0')$ from (4.7) into (4.8) and obtain a system of $k+1$ equations for $k+1$ unknown components of $g^{(0)}(\ell_0') \in V_{I}(\ell_0')$ and $g^{(0)}(-\ell_0') \in V_{II}(-\ell_0')$. This is the idea. To carry it through we should get bounds on G_1, G_{-1} and prove solvability of (4.8).

LEMMA 4.1. *The solution $Y = G_1(F, g(\ell'), g^{(0)}(\ell_0'))$ satisfies the estimate*

$$\|Y\|_\infty \leq K\Big(\|H\|_1 + |g| + |g^{(0)}|\Big), \tag{4.9}$$

where K is independent of ℓ'. For $H = 0$, $g(\ell') = 0$ the boundary value $Y(\ell'_0)$ is

$$Y(\ell'_0) = g^{(0)}(\ell'_0)\Big(1 + O(e^{-\delta(\ell'-\ell'_0)})\Big), \qquad \delta > 0 \text{ a constant.} \tag{4.10}$$

Proof. Without loss of generality, one can assume that $M(\infty)$ is in the block form $M(\infty) = M_{I} \oplus M_{II}$, where $\operatorname{Re} M_{I} < -\delta I$ and $\operatorname{Re} M_{II} \geq 0$. Similarly, we partition $Y = (Y_{I}, Y_{II})$, H, and F. First find a particular solution of (4.5) with boundary condition $Y_{I}(\ell'_0) = 0$, $Y_{II}(\ell') = 0$. Then

$$\frac{1}{2}|Y_{II}(\tau)|^2 + \int_{\tau}^{\ell'} (\operatorname{Re} M_{II}\, Y_{II}, Y_{II})d\tau = -\int_{\tau}^{\ell'} (F_{II}, Y_{II})d\tau \tag{4.11}$$

and similarly for Y_{I}. Thus

$$\tfrac{1}{2}\|Y\|_{\infty} \leq \|F\|_1 \leq \|H\|_1 + \|\Delta M\|_1 \|Y\|_{\infty}. \tag{4.12}$$

Recall that $\Delta M = O(y_{sh}(\varepsilon'(\tau - \tau_0)) - 1) = O(e^{-\delta_1(\tau-\tau_0)})$. We should choose ℓ'_0 sufficiently large so that for $\tau > \ell'_0$, $\|\Delta M\|_1 \leq \frac{1}{4}$. Thus we find a particular solution $Y^{(1)}$ such that

$$\|Y^{(1)}\|_{\infty} \leq 4\|H\|_1. \tag{4.13}$$

Now we solve the homogeneous equation $\partial_\tau Y - MY = 0$ with original boundary conditions. For Y_{II}, instead of (4.11), we obtain

$$\tfrac{1}{2}\|Y_{II}\|_{\infty}^2 \leq \|\Delta M\|_1 \|Y\|_{\infty}^2 + \tfrac{1}{2}|Y_{II}(\ell')|^2. \tag{4.14}$$

The equation for Y_{I} is multiplied by $\exp(\delta_2(\tau - \ell'_0))$, $\delta_2 < \min(\delta, \delta_1)$, and integrated. Then, instead of (4.14) we obtain

$$\begin{aligned} &\tfrac{1}{2}\Big\|Y_{I} \exp\Big(\delta_2(\tau - \ell'_0)\Big)\Big\|_{\infty}^2 \\ &\quad \leq \Big\|\Delta M \exp\Big(\delta_2(\tau - \ell'_0)\Big)\Big\|_1 \|Y\|_{\infty} \Big\|Y_{I} \exp\Big(\delta_2(\tau - \ell'_0)\Big)\Big\|_{\infty} \\ &\qquad + \tfrac{1}{2}|Y_{I}(\ell'_0)|^2 \ . \end{aligned} \tag{4.15}$$

Since ΔM decays faster than $\exp(\delta_2(\tau - \ell'_0))$ grows, we may assume also that $\|\Delta M \exp \delta_2(\tau - \ell'_0)\|_1$ is small. Since the boundary condition in (4.2) is Lopatinsky well posed at $\tau = \infty$, we can express

$$Y_{II}(\ell') = O(g) + O\Big(Y_{I}(\ell') \exp \delta_2(\ell' - \ell'_0)\Big) \exp\big(-\delta_2(\ell' - \ell'_0)\big). \tag{4.16}$$

From (4.6) we obtain

$$Y_{I}(\ell'_0) = O(g^{(0)}) + O\Big(Y_{II}(\ell'_0)\Big). \tag{4.17}$$

Indeed, $P_{I}(\ell'_0) = P_{I}(\infty) + O(\exp(-\delta_2 \ell'_0))$, while $P_{I}(\infty)$ acts as an identity on the Y_{I} component. Substitute (4.16), (4.17) into (4.14), (4.15), multiply (4.14) by $c \gg 1$, and add the result to (4.15). For large $\delta_2(\ell' - \ell'_0)$ the contribution of Y_{I} in (4.16) is negligible relative to the l.h.s. of (4.15). Hence

$$\Big\|Y_{I} \exp\Big(\delta_2(\tau - \ell'_0)\Big)\Big\|_{\infty} + \|Y_{II}\|_{\infty} \leq K\Big(|g| + |g^{(0)}|\Big). \tag{4.18}$$

The above g, $g^{(0)}$ are also affected by the solution $Y^{(1)}$ estimated in (4.13). Altogether the sum of the two solutions solves the original problem and satisfies estimate (4.9). In order to prove (4.10), we take an approximate solution $Y^{(1)} \in V_{I}$ with $Y^{(1)}(\ell_0') = g^{(0)}$. Since it decays exponentially, its contribution to the boundary condition at ℓ' is $O(\exp(-\delta(\ell' - \ell_0')))$. By estimate (4.9) we bound the correction to $Y^{(1)}$.

Now, return to (4.8):

$$\begin{aligned} &G_1\Big(H, g(\ell'), 0\Big) + g^{(0)}(\ell_0')\Big(1 + O\Big(\exp(-\delta(\ell' - \ell_0'))\Big)\Big) \\ &\quad = G_0\Big(H, G_{-1}(H, g(-\ell'), 0)\Big) \\ &\qquad + G_0\Big(0, g^{(0)}(-\ell_0')\Big(1 + O\Big(\exp(-\delta(\ell' - \ell_0'))\Big)\Big)\Big). \end{aligned} \tag{4.19}$$

Since ℓ_0' is fixed, G_0 is a bounded map. The vectors $g^{(0)}(\ell_0') \in V_{I}(\ell_0')$, $g^{(0)}(-\ell_0') \in V_{II}(-\ell_0')$. By the continuous stability hypothesis, $G_0(0, V_{II}(-\ell_0'))$ is transversal to $V_{I}(-\ell_0')$ for the value of parameter $p = p_0$. Otherwise equation $\partial_\tau Y - MY = 0$ or, equivalently, (2.22), would have a nontrivial solution which decays asymptotically at infinity. Since $\exp(-\delta(\ell' - \ell_0'))$ is small, the problem in (4.19) is boundedly solvable with respect to $g^{(0)}(\ell_0')$, $g^{(0)}(-\ell_0')$. The corresponding function Y solves the boundary value problem (4.1), (4.2) and satisfies the estimate

$$\|Y\|_\infty \le K\Big(\|H\|_1 + |g|\Big). \tag{4.20}$$

Remark 5. If $s_0' \ne 0$, then we may assume also that $\operatorname{Re} M_{II} > \delta I$. Then instead of (4.12) we can obtain a stronger estimate $\|Y\|_p + \|Y\|_\infty \le K(\|H\|_p + \|H\|_\infty)$, $p = \infty, 1$, and therefore a final estimate will be

$$\|Y\|_p + \|Y\|_\infty \le K\Big(\|H\|_p + \|H\|_\infty + |g|\Big), \qquad p = \infty, 1. \tag{4.21}$$

Now return to the discrete problem (3.78), (3.81), (3.82). We solve it exactly as the continuous one. The space $V_{I}(\tau)$ is replaced by the corresponding space $V_{I}'(\tau)$ of decreasing solutions of (3.78), and the projection P_{I} by a projection P_{I}' on $V_{I}'(\tau)$. The difference equation (3.78) is rewritten as

$$\begin{aligned} &\text{(a)} \quad E_\tau Y_{I}' - (I + \sigma M_{I}')Y_{I}' = \sigma H_{I}' + \sigma O(\Delta M')Y' = \sigma F_{I}', \\ &\text{(b)} \quad E_\tau Y_{II}' - (I + \sigma M_{II}')Y_{I}' = \sigma H_{II}' + \sigma O(\Delta M')Y' = \sigma F_{II}'. \end{aligned} \tag{4.22}$$

(For brevity we dropped the subscript 1 in (3.78).) Here $(I + \sigma M_{I}')^*(I + \sigma M_{I}') \le (1 - 2\delta)I$, $(I + \sigma M_{II}')^*(I + \sigma M_{II}') \ge I$. Multiply both sides of (4.22)(a) by $E_\tau Y_{I}' + (I + \sigma M_{I}')Y_{I}'$ and sum from ℓ_0' until τ and similarly for (4.22)(b). Instead of (4.12), (4.13), we obtain the estimate $\|Y'\|_\infty \le K\|F'\|_{1,\sigma}$ and $\|Y'^{(1)}\|_\infty \le K\|H'\|_{1,\sigma}$. The rest of the proof of Lemma 4.1 is repeated word for word. Now we arrive at the functional equation (4.19). The operator G_0 is replaced by the solution operator G_0' of the difference equation. By the standard convergence theorem for the Euler method, it follows that G_0' approximates G_0 as $\sigma \to 0$. Similarly, the stable manifold $V_{I}'(\ell_0')$ approximates $V_{I}(\ell_0')$ as $\sigma \to 0$. This result is less standard; e.g., see Theorem 5.3 in [7]. Then by continuity, G_0' maps $V_{II}'(-\ell_0')$ transversally to the subspace $V_{I}'(\ell_0')$. Finally, we obtain the estimate

$$\|Y_1'\|_\infty \le K\Big(\|H_1'\|_{1,\sigma} + |g'|\Big), \tag{4.23}$$

and in case $|s_0'| \neq 0$,

$$\|Y_1'\|_{p,\sigma} + \|Y_1'\|_\infty \leq K\Big(\|H_1'\|_{p,\sigma} + \|H_1'\|_\infty + |g'|\Big), \qquad p = 1, \infty. \tag{4.24}$$

5. Existence of stationary shocks. In [7] we proved existence of stationary discrete shocks on the whole line. In this section we will prove Theorem 2.1 about existence of an n_1 parameter family of stationary shocks on the interval $[-\ell, \ell]$. Unlike the nonstationary case, we also will treat the case $1 < n_1 < n$. The n_1 parameters are the components of the vector

$$P\sum_x G_2\Big(\{E_x^j u_{st}(x)\}\Big) = g_{\text{sum}}. \tag{5.1}$$

The function u_{st} satisfies the equation

$$G_1\Big(\{E_x^j u_{st}(x)\}\Big) = \text{const}. \tag{5.2}$$

We will represent this constant as a value $f(u_{_L} + \Delta u_{_L})$ and look for the solution in the form

$$u_{st} = u_{sh}\Big(\varepsilon(x - x_0), u_{_L} + \Delta u_{_L}\Big) + \Delta u_{st}. \tag{5.3}$$

Here $u_{sh}(\varepsilon(x - x_0), u_{_L} + \Delta u_{_L})$ is defined as in (2.86), namely,

$$\begin{aligned} u_{sh}\Big(\varepsilon(x - x_0), u_{_L} + \Delta u_{_L}\Big) &= \tfrac{1}{2}(u_{_L} + \Delta u_{_L} + u_{_R} + \Delta u_{_R}) \\ &+ \tfrac{1}{2}(u_{_L} + \Delta u_{_L} - u_{_R} - \Delta u_{_R}) y_{sh}\Big(\varepsilon(x - x_0)\Big), \\ \varepsilon &= \Big(\lambda_1(u_{_L} + \Delta u_{_L})\Big)^{1/k}, \end{aligned} \tag{5.4}$$

$\Delta u_{_R}$ is defined uniquely by the equation

$$f(u_{_L} + \Delta u_{_L}) = f(u_{_R} + \Delta u_{_R}), \tag{5.5}$$

and y_{sh} is a solution of (2.23). Then, by boundary condition (2.13),

$$P\Big(f(u_{_L} + \Delta u_{_L}) - f(u_{_L})\Big) = 0. \tag{5.6}$$

The remaining boundary conditions are

$$S_{_{L,1}}\Big(\{E_x^j u_{st}(x_{-\ell})\}, u_{_L}\Big) = 0, \qquad S_{_{R,1}}\Big(\{E_x^j u_{st}(x_\ell), u_{_R}\Big) = 0 \tag{5.7}$$

and

$$S_{_{L,2}}\Big(\{E_x^j u_{st}(x_{-\ell})\}, u_{_L}\Big) = 0, \qquad S_{_{R,2}}\Big(\{E_x^j u_{st}(x_\ell), u_{_R}\Big) = 0 \tag{5.8}$$

(see (2.7), (2.26)(b),(c)). Equation (5.2) will be rewritten as

$$dG_1[u_{sh}]\Delta u_{st} = f(u_{_L} + \Delta u_{_L}) - G_1\Big(\{E_x^j u_{sh}(\cdot, u_{_L} + \Delta u_{_L})\}\Big) + O(\Delta u_{st})^2 = F. \tag{5.9}$$

It is easy to see that for u_L, u_R as in (2.10),

$$\begin{aligned} &P_1\Big(f(u_L) - G_1\Big(\{E_x^j u_{sh}(\varepsilon x, u_L)\}\Big)\Big) = O(\varepsilon\mu^2),\\ &(I-P_1)\Big(f(u_L) - G_1\Big(\{E_x^j u_{sh}(\varepsilon x, u_L)\}\Big)\Big) = O(\mu^2). \end{aligned} \tag{5.10}$$

Note that by consistency, the r.h.s. of (5.10) vanishes at $\varepsilon x = \pm\infty$ so that actually $O(\varepsilon\mu^2)$, $O(\mu^2)$ are multiplied by an exponent of the type $e^{-\varepsilon\delta|x|}$. The above result is obviously valid if we replace u_L by $u_L + \Delta u_L$ and μ by $\lambda_1(u_L + \Delta u_L)$. However, the projectors P_1 and $(I - P_1)$ depend on the point $u_0 = u_0(u_L)$. If we change u_L to $u_L + \Delta u_L$ and do not change P_1 and $I - P_1$, then the correct formula will be

$$\begin{aligned} &\text{(a) } P_1\Big(f(u_L+\Delta u_L) - G_1\Big(\{E_x^j u_{sh}(\varepsilon x, u_L+\Delta u_L)\}\Big)\Big),\\ &\qquad = O\Big(\varepsilon\lambda_1^2(u_L+\Delta u_L)\Big) + O\Big(\lambda_1^2(u_L+\Delta u_L)\Big)O(\Delta u_L),\\ &\text{(b) } (I-P_1)\Big(f(u_L+\Delta u_L) - G_1\Big(\{E_x^j u_{sh}(\varepsilon x, u_L+\Delta u_L)\}\Big)\Big)\\ &\qquad = O\Big(\lambda_1^2(u_L+\Delta u_L)\Big). \end{aligned} \tag{5.11}$$

The contribution of $\Delta u_{sh} = u_{sh}(\varepsilon(x-x_0), u_L + \Delta u_L) - u_{sh}(\varepsilon x, u_L)$ to (5.1) is

$$\begin{aligned} \sum_x G_2\Big(\{E_x^j \Delta u_{sh}\}\Big) &= (\Delta u_L + \Delta u_R)\Big(\ell + O(1)\Big)\\ &\quad + (u_L + \Delta u_L - u_R - \Delta u_R)x_0\left(1 + O\left(\frac{e^{-\varepsilon\delta\ell}(1 - e^{-x_0\delta\varepsilon})}{x_0\varepsilon}\right)\right)\\ &= (\Delta u_L + \Delta u_R)\Big(\ell + O(1)\Big)\\ &\quad + (u_L + \Delta u_L - u_R - \Delta u_R)x_0\Big(1 + O(e^{-\varepsilon\delta\ell})\Big). \end{aligned} \tag{5.12}$$

The correspondence $u_R = u_R(u_L)$ is given by an invariant formula

$$u_R = u_L - 2\lambda_1(u_L)e_1(u_L) + O\Big(\lambda_1(u_L)\Big)^2, \tag{5.13}$$

where $e_1(u_L)$ is the eigenvector of $df[u_L]$ corresponding to $\lambda_1(u_L)$ and normalized so that

$$\nabla\lambda_1(u_L)\cdot e_1(u_L) = 1. \tag{5.14}$$

Hence

$$\begin{aligned} (P-P_1)(\Delta u_L + \Delta u_R) &= (P-P_1)2\Delta u_L + O\Big(\lambda_1(u_L)\Big)\Delta u_L,\\ (P-P_1)(u_L + \Delta u_L - u_R - \Delta u_R) &= O\Big(\lambda_1(u_L)\Big)^2 + O\Big(\lambda_1(u_L)\Big)\Delta u_L, \end{aligned} \tag{5.15}$$

$$\begin{aligned} P_1(\Delta u_L + \Delta u_R) &= 2P_1\Delta u_L - 2\Big(\nabla\lambda_1(u_L), \Delta u_L\Big) + O\Big(\lambda_1(u_L)\Big)\Delta u_L,\\ P_1(u_L + \Delta u_L - u_R - \Delta u_R) &= 2\lambda_1(u_L) + 2\Big(\nabla\lambda_1(u_L), \Delta u_L\Big)\\ &\quad + O\Big(\lambda_1(u_L)\Big)^2 + O\Big(\lambda_1(u_L)\Big)\Delta u_L. \end{aligned} \tag{5.16}$$

By (5.6) one can express $P\Delta u_{_L}$ as a function of $(I-P)\Delta u_{_L}$. Namely,

$$P\Big(df[u_{_L}]\Delta u_{_L}+O(\Delta u_{_L})^2\Big)=0, \tag{5.17}$$

where $df[u_{_L}]=T\,\mathrm{diag}(\lambda_1(u_{_L}),\lambda_2(u_{_L}),\dots,\lambda_n(u_{_L}))T^{-1}$ and $T=I+O(\mu)$. Note that the entry $(df[u_{_L}])_{11}$ is $\lambda_1(u_{_L})+O(\mu^2)=O^*(\mu)$. If $\Delta u_{_L}/\mu$ is small, then one can write the solution of (5.17) as

$$\begin{aligned}(P-P_1)\Delta u_{_L}&=O(\mu)\Big(O(P_1\Delta u_{_L})+O\Big((I-P)\Delta u_{_L}\Big)\Big),\\ P_1\Delta u_{_L}&=O\Big((I-P)\Delta u_{_L}\Big).\end{aligned} \tag{5.18}$$

From (5.1), (5.12), (5.15), (5.16), and (5.18) we obtain for Δu_{st} the integral relations

$$\begin{aligned}&(P-P_1)\sum_x G_2\Big(\{E_x^j\Delta u_{st}(x)\}\Big)\\ &\quad=(P-P_1)\Big(g_{\rm sum}-\sum_x G_2\Big(\{E_x^j u_{sh}(\varepsilon x,u_{_L})\}\Big)\Big)\\ &\qquad+O(\mu)\Big(\ell+O(1)\Big)(I-P)\Delta u_{_L}\\ &\qquad+\Big(O(\mu^2)+O(\mu)(I-P)\Delta u_{_L}\Big)x_0\Big(1+O(1)\Big)=\tau_-;\end{aligned} \tag{5.19}$$

$$\begin{aligned}&P_1\sum_x G_2\Big(\{E_x^j\Delta u_{st}(x)\}\Big)\\ &\quad=P_1\Big(g_{\rm sum}-\sum_x G_2\Big(\{E_x^j u_{sh}(x)\}\Big)\Big)\\ &\qquad+\ell\Big(1+o(1)\Big)O\Big((I-P)\Delta u_{_L}\Big)+x_0\Big(1+o(1)\Big)\\ &\qquad\cdot\Big(2\lambda_1(u_{_L})+2\Big(\nabla\lambda_1(u_{_L}),\Delta u_{_L}\Big)+O(\mu^2)+O(\mu\Delta u_{_L})\Big).\end{aligned} \tag{5.20}$$

The remaining boundary conditions for Δu_{st} follow from (5.7), (5.8),

$$dS_{_{L,1}}\Big(\{E_x^j\Delta u_{st}(x_{-\ell})\}\Big)=-S_{_{L,1}}\Big(\{E_x^j u_{sh}(x_{-\ell})\},u_{_L}\Big)+O(\Delta u_{st})^2 \tag{5.21}$$

and similarly for $dS_{_{R,1}}$, $dS_{_{L,2}}$, $dS_{_{R,2}}$. By the consistency assumption in (2.8),

$$S_{_L}\Big(\{E_x^j u_{sh}(x_{-\ell})\},u_{_L}\Big)=dS_{_L}\Delta u_{_L}+O(\Delta u_{_L})^2+O(e^{-\delta\varepsilon\ell})\Big(O(\mu)+O(\Delta u_{_L})\Big) \tag{5.22}$$

and similarly for the right boundary. Note that by (5.4) and (5.13), $(I-P_1)(u_{sh}-u_{_L}-\Delta u_{_L})=(O(\mu^2)+O(\mu\Delta u_{_L}))(y_{sh}-1)$. Hence by (2.30),

$$\begin{aligned}&S_{1j}\Big(\{E_x^j u_{sh}(x_{-\ell})\}\Big)\\ &\quad=dS_{1j}(I-P_1)\Delta u_{_L}+O(\varepsilon^{d_j}e^{-\delta\varepsilon\ell})\Big(O(\mu)+O(\Delta u_{_L})\Big)+O(\Delta u_{_L})^2\ .\end{aligned} \tag{5.23}$$

Similarly, by (2.34),

$$
\begin{aligned}
&S_2\Big(\{E_x^j u_{sh}(x_{-\ell})\}\Big)\\
&\qquad = dS_2(I-P_1)\Delta u_{\scriptscriptstyle L} + O(\varepsilon^k e^{-\delta\varepsilon\ell})\Big(O(\mu)+O(\Delta u_{\scriptscriptstyle L})\Big) + O(\Delta u_{\scriptscriptstyle L})^2\ .
\end{aligned}
\tag{5.24}
$$

We will solve (5.9) with constraints (5.18), (5.19), (5.21). Instead of fixing $g_{\rm sum}$ we will assume that τ_- in (5.19) and x_0 are given. The solution Δu_{st}, $\Delta u_{\scriptscriptstyle L}$ will thus depend on τ_- and (implicitly) on x_0. Then $(P-P_1)g_{\rm sum}$ will be determined by (5.19) and $P_1 g_{\rm sum}$ by (5.20). This way we can overcome the restriction $\ell\varepsilon^k \ll 1$. However, we need this restriction in order to obtain one-to-one correspondence between $g_{\rm sum}$ and τ_-, x_0.

In order to solve (5.9) we proceed as in section 3. Namely, transform the variables

$$
\Delta\widetilde{u}_{st}(x) = \Big(\Delta u_{st}(x), \dots, E_x^{\nu_0-1}\Delta u_{st}(x)\Big)^{\scriptscriptstyle T} = X\Big(\eta(x)\Big)y(x),
\tag{5.25}
$$

where $\eta(x) = u_{sh}(\varepsilon(x-x_0), u_{\scriptscriptstyle L}+\Delta u_{\scriptscriptstyle L})$. Then (5.9) becomes

$$
\begin{aligned}
&\begin{pmatrix} E_x - M_1 & 0 & 0\\ 0 & E_x - M_{\scriptscriptstyle I} & 0\\ 0 & 0 & E_x - M_{\scriptscriptstyle II}\end{pmatrix} y\\
&\qquad = T^{-1}\Big(\widetilde{F} - A\Big((E_x - I)X\Big)E_x y\Big) = H.
\end{aligned}
\tag{5.26}
$$

Here we have $\widetilde{F} = (0,\dots,0,F)^{\scriptscriptstyle T}$ as in (3.17), A as in (3.18), and $M_{\scriptscriptstyle I}$, $M_{\scriptscriptstyle II}$ as in (3.26). Since the operator dG_1 is $dG(s=0)/(E_x - I)$, the matrix M_1 in (5.26) is nothing but the lower $k\times k$ block of the matrix M_1 in (3.50). Namely,

$$
M_1 = I + \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ \vdots & & & \ddots & \\ 0 & & & & 1\\ -\alpha_1 & & & & -\alpha_k\end{bmatrix},
\tag{5.27}
$$

where $\alpha_i = \alpha_i(\eta, s=0)$ as in (3.43). The characteristic equation

$$
\det(M_1 - \kappa I) = (\kappa-1)^k + \sum_{i=1}^{k}\alpha_i(\eta)(\kappa-1)^{i-1} = 0
\tag{5.28}
$$

at $\varepsilon x = \pm\infty$ has roots

$$
\kappa = 1 + \varepsilon(\pm 1)^{1/k} + O(\varepsilon^2).
\tag{5.29}
$$

Unlike (4.3), now the values of $\lambda = \varepsilon^{-1}\log\kappa$ are away from the imaginary line $\operatorname{Re}\lambda = 0$. In such a situation, instead of (4.23) we have a stronger estimate (4.24). Let us first find $y_{\scriptscriptstyle I}$, $y_{\scriptscriptstyle II}$. As in (3.64), we represent $y_{\scriptscriptstyle I} = c_{\scriptscriptstyle I}\varphi_{\eta,{\scriptscriptstyle I}} + \mathring{y}_{\scriptscriptstyle I}$ and similarly for $y_{\scriptscriptstyle II}$. For $\mathring{y}_{{\scriptscriptstyle I},{\scriptscriptstyle II}}$ we have the estimate

$$
\|\mathring{y}_{\scriptscriptstyle I}\|_1 + \|\mathring{y}_{\scriptscriptstyle II}\|_1 \le K\Big(\|H_{\scriptscriptstyle I}\|_1 + \|H_{\scriptscriptstyle II}\|_1\Big).
\tag{5.30}
$$

As in (3.68), the coefficients $c_{I,II}$ are determined by the boundary conditions (5.19) (with given τ_-) and (5.21) (with $(dS_{L,2}, dS_{R,2}) = dS_2$ instead of $dS_{L,1}$). The vector $(I-P)\Delta u_L$ plays the role of c_j, $n_1 < j \le n$. By the generalized stability condition for the system (2.40), (2.41), (2.44) (with $\rho \equiv 1$), these boundary conditions are uniquely solvable in terms of $c_{I,II}$, $(I-P)\Delta u_L$. We should merely collect all terms which contribute to the r.h.s. of these equations. In view of (2.35), the contribution of y_1 to (5.19) is estimated

$$\left|(P-P_1)\Big(\sum_x e_1 y_{1,1}(x) + \sum O(\eta) y_1(x)\Big)\right| \le K\varepsilon^k \|y_1\|_1. \tag{5.31}$$

Notice that because of the reduction of order, the vectors $X_{1,i}(0)$ in (3.48), $1 \le i \le k$, are not affected by C_{10}. The contribution of y_1 to $dS_2 \cdot \Delta u_{st}$ is estimated as in (3.73) and is bounded by

$$K\Big(\|(E_x - I)^k y_1\|_\infty + \varepsilon^k \|y_1\|_\infty\Big). \tag{5.32}$$

The contribution of the particular solution $\mathring{y}_{I,II}$ to (5.19) is bounded by $K(\|H_I\|_1 + \|H_{II}\|_1)$. The nonhomogeneous terms in (5.19) and (5.24) are the ones which do not include Δu_L (since Δu_L depends on $(I-P)\Delta u_L$). Thus we obtain the estimate

$$\begin{aligned} &\|y_I\|_1 + \|y_{II}\|_1 + |\Delta u_L| + \mu^{-1}|(P-P_1)\Delta u_L| \\ &\quad \le K\Big(\|H_I\|_1 + \|H_{II}\|_1 + \varepsilon^k \|y_1\|_1 + \|(E_x - I)^k y_1\|_\infty \\ &\qquad + |\tau_-| + \varepsilon^{2k} e^{-\delta\varepsilon\ell} + |\Delta u_L|^2 + \|\Delta u_{st}\|_\infty^2\Big). \end{aligned} \tag{5.33}$$

For y_1 we have the estimate

$$\begin{aligned} &\sum_{i+j\le k+1} \Big(\|D_\tau^i y'_{1,j}\|_{1,\varepsilon} + \|D_\tau^i y'_{1,j}\|_\infty\Big) \\ &\quad \le K\Bigg(\sum_{i+j\le k} \|D_\tau^i H'_{1,j}\|_{1,\varepsilon} + \sum \varepsilon^{-d_j} |dS_{1j}(I-P_1)\Delta u_L| \\ &\qquad + O(e^{-\delta\varepsilon\ell}\mu) + O(e^{-\delta\varepsilon\ell}\Delta u_L) + \varepsilon^{-(k-1)}\Big(\|\Delta u_{st}\|_\infty^2 + |\Delta u_L|^2\Big)\Bigg), \end{aligned} \tag{5.34}$$

where $y'_{1,j} = \varepsilon^{-j+1} y_{1,j}$, $H'_{1,j} = \varepsilon^{-j} H_{1,j}$. Multiply (5.34) by $K_1 \varepsilon^{k-1}$, where $K_1 \gg K$, and add to (5.33). We obtain

$$\begin{aligned} &\|y_{I,II}\|_1 + \varepsilon^{k-1} \sum_{i+j\le k+1} \Big(\|D_\tau^i y'_{1,j}\|_{1,\varepsilon} + \|D_\tau^i y'_{1,j}\|_\infty\Big) \\ &\qquad + |\Delta u_L| + \mu^{-1}|(P-P_1)\Delta u_L| \\ &\quad \le K\Bigg(\|H_{I,II}\|_1 + \varepsilon^{k-1} \sum_{i+j\le k} \|D_\tau^i H'_{1,j}\|_{1,\varepsilon} + |\tau_-| \\ &\qquad + \varepsilon^{2k-1} e^{-\delta\varepsilon\ell} + \|\Delta u_{st}\|_\infty^2 + |\Delta u_L|^2\Bigg). \end{aligned} \tag{5.35}$$

The terms $K\varepsilon^k\|y_1\|_1 = K\varepsilon^{k-1}\|y_1\|_{1,\varepsilon}$ and $K\|(E_x - I)^k y_1\|_\infty = K\varepsilon^k\|D_\tau^k y_1\|_\infty$ are majorated by $K_1\varepsilon^{k-1}$ times the l.h.s. of (5.34). In the case $1 < n_1 < n$ we have assumed that $d_j \le k-2$; hence $\varepsilon^{k-1}\varepsilon^{-d_j}O(\Delta u_{_L})$ is negligible. The terms with H_1' are bounded by

$$\varepsilon^{k-1}\|D_\tau^i H'\|_{1,\varepsilon} = \varepsilon^k\|D_\tau^i H'_{1,j}\|_1 \le \|H_{1,j}\|_1. \tag{5.36}$$

The contribution of y to $\|H\|_1$ is bounded by

$$\|O(\eta')y\|_1 \le K\varepsilon^k\|y\|_\infty \tag{5.37}$$

and is negligible compared with the l.h.s. of (5.35). Thus in $\|H\|_1$ we are left with $\|F\|_1$, where F is defined in (5.9). Recall that the r.h.s. of (5.11) is actually multiplied by $e^{-\varepsilon\delta|x|}$. Hence

$$\begin{aligned}\|F\|_1 &\le K\Big(\varepsilon^{-1}|\lambda_1^2(u_{_L} + \Delta u_{_L})| + \|(\Delta u_{st})^2\|_1\Big)\\ &\le K\Big(\varepsilon^{2k-1} + \varepsilon^{-1}|\Delta u_{_L}|^2 + \|\Delta u_{st}\|_\infty\|\Delta u_{st}\|_1\Big).\end{aligned} \tag{5.38}$$

The biggest nonhomogeneous term on the r.h.s. of (5.35) is $K\varepsilon^{2k-1}$ in (5.38). Thus we will assume that

$$|\tau_-| \le K\varepsilon^{2k-1}. \tag{5.39}$$

From (5.35) we obtain the estimate for Δu_{st}:

$$\begin{aligned}&\|(I-P_1)\Delta u_{st}\|_1 + \varepsilon^{k-1}\|P_1\Delta u_{st}\|_\infty\\ &\quad+\varepsilon^k\|P_1\Delta u_{st}\|_1 + \varepsilon^{k-1}\|P_1(E_x - I)\Delta u_{st}\|_1\\ &\quad+|\Delta u_{_L}| + \mu^{-1}|(P-P_1)\Delta u_{_L}|\\ &\le K\Big(\varepsilon^{2k-1} + \varepsilon^{-1}|\Delta u_{_L}|^2 + \|\Delta u_{st}\|_\infty\|\Delta u_{st}\|_1\Big).\end{aligned} \tag{5.40}$$

Unfortunately, the last estimate is too weak since it gives the bounds $\|P_1\Delta u_{st}\|_\infty = O(\varepsilon^k)$, $\|P_1\Delta u_{st}\|_1 = O(\varepsilon^{k-1})$, and hence the $\|\Delta u_{st}\|_\infty \cdot \|\Delta u_{st}\|_1$ term is not negligible. Therefore, return to estimate (5.34). We remark that by Lemma 3.1 the components $(I-P_1)F$ in (5.9) affect $H_{1,k}$ in (5.34) as $O(\mu)(I-P_1)F$. By (5.11),

$$\|\varepsilon(I-P_1)F\|_1 + \|P_1F\|_1 \le K\Big(\varepsilon^{2k} + \varepsilon^{2k-1}|\Delta u_{_L}| + \|(\Delta u_{st})^2\|_1\Big). \tag{5.41}$$

Multiply estimate (5.34) by $\delta_1\varepsilon^{k-2}$, where δ_1 is a small constant, and add the result to (5.33). We obtain

$$\begin{aligned}&\|y_{_{I,II}}\|_1 + \delta_1\varepsilon^{k-2}\sum_{i+j\le k+1}\Big(\|D_\tau^i y'_{1,j}\|_{1,\varepsilon} + \|D_\tau^i y'_{1,j}\|_\infty\Big)\\ &\quad+|\Delta u_{_L}| + \mu^{-1}|(P-P_1)\Delta u_{_L}|\\ &\le K\Bigg(\|H_{_{I,II}}\|_1 + \delta_1\varepsilon^{k-2}\sum_{i+j\le k}\|D_\tau^i H'_{1,j}\|_{1,\varepsilon} + |\tau_-| + \varepsilon^{2k}e^{-\delta\varepsilon\ell}\\ &\quad+\delta_1|\Delta u_{_L}| + \delta_1 e^{-\delta\varepsilon\ell}\varepsilon^{2k-2} + \varepsilon^{-1}\delta_1\Big(\|\Delta u_{st}\|_\infty^2 + |\Delta u_{_L}|^2\Big)\Bigg)\\ &\le K\Big(\varepsilon^{2k-1} + \delta_1 e^{-\delta\varepsilon\ell}\varepsilon^{2k-2} + \delta_1|\Delta u_{_L}|\\ &\qquad+\varepsilon^{-1}|\Delta u_{_L}|^2 + \delta_1\varepsilon^{-1}\|\Delta u_{st}\|_1\|\Delta u_{st}\|_\infty + \delta_1\varepsilon^{k-1}\|y\|_\infty\Big).\end{aligned} \tag{5.42}$$

Now instead of (5.40) we obtain the estimate

$$\begin{aligned}&\|(I-P_1)\Delta u_{st}\|_1+\varepsilon^{k-1}\|P_1\Delta u_{st}\|_1+\varepsilon^{k-2}\|P_1\Delta u_{st}\|_\infty\\&\quad+\varepsilon^{k-2}\|P_1(E_x-I)\Delta u_{st}\|_1+|\Delta u_{_L}|+\mu^{-1}|(P-P_1)\Delta u_{_L}|\\&\quad\le K\Big(\varepsilon^{2k-1}+e^{-\delta\varepsilon\ell}\varepsilon^{2k-2}+\varepsilon^{-1}\|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty\\&\qquad+\varepsilon^{-1}|\Delta u_{_L}|^2+|\tau_-|\Big).\end{aligned}\tag{5.43}$$

If $\varepsilon\ell\gg 1$ then, modulo the quadratic terms,

$$\|\Delta u_{st}\|_1\ll\varepsilon^{k-1},\qquad\|\Delta u_{st}\|_\infty\ll\varepsilon^{k},\qquad|\Delta u_{_L}|\ll\varepsilon^{2k-2}.\tag{5.44}$$

Hence

$$\varepsilon^{-1}\|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty+\varepsilon^{-1}|\Delta u_{_L}|^2\ll\varepsilon^{2k-2}+\varepsilon^{4k-5}\ll\varepsilon^{2k-2},\tag{5.45}$$

provided $2k>3$. The last is true since, in the case $1<n_1<n$, we have assumed that $k>1$. Thus, in (5.43) we can disregard the quadratic terms. Now, at the r.h.s. of (5.40) we have the terms

$$\|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty+\varepsilon^{-1}|\Delta u_{_L}|^2\ll\varepsilon^{2k-1}+\varepsilon^{4k-5}\le 2\varepsilon^{2k-1}\tag{5.46}$$

since $2k-4\ge 0$. We obtain the final estimates

$$\begin{aligned}&\text{(a)}\quad\|(I-P_1)\Delta u_{st}\|_1+|\Delta u_{_L}|+\varepsilon^{-k}|(P-P_1)\Delta u_{_L}|\le K\varepsilon^{2k-1},\\&\text{(b)}\quad\varepsilon\|P_1\Delta u_{st}\|_1+\|\Delta u_{st}\|_\infty\\&\qquad+\|P_1(E_x-I)\Delta u_{st}\|_1\le K(\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k).\end{aligned}\tag{5.47}$$

Recall that the difference Δu_{st} in (2.85), in our present notation, is the sum $\Delta u_{st}+u_{sh}(\varepsilon(x-x_0),u_{_L}+\Delta u_{_L})-u_{sh}(\varepsilon x,u_{_L})$. Thus the norm $\varepsilon\|\Delta u_{st}\|_1+\|\Delta u_{st}\|_\infty+\|(E_x-I)\Delta u_{st}\|_1$ in (2.87) and (2.88) is bounded by

$$\begin{aligned}&\varepsilon\|\Delta u_{st}\|_1+\|\Delta u_{st}\|_\infty+\|(I-P_1)\Delta u_{st}\|_1\\&\quad+\|P_1(E_x-I)\Delta u_{st}\|_1+\varepsilon|\Delta u_{_L}|\cdot\ell+|\Delta u_{_L}|+K\varepsilon|x_0|\varepsilon^k\\&\qquad\le K\Big(\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k+(\varepsilon^k\ell)\varepsilon^k+\varepsilon|x_0|\varepsilon^k\Big).\end{aligned}\tag{5.48}$$

We will assume

$$\varepsilon|x_0|\ll 1.\tag{5.49}$$

Then, by (1.11), the r.h.s. of (5.48) is bounded by $\delta_0\varepsilon^k$ as in (2.87). Thus Theorem 2.1 has been proved in the case $1<n_1<n$. Equations (5.19), (5.20) define the correspondence between τ_-, x_0 and Pg_{sum}. Rescale the variables

$$\begin{aligned}&\tau_-'=\tau_-/\varepsilon^{2k-1},\qquad x_0'=\varepsilon x_0,\qquad\Delta u_{_L}'=\Delta u_{_L}/\varepsilon^{2k-1},\\&(P-P_1)g_{\text{sum}}'=(P-P_1)\Big(g_{\text{sum}}-\sum_x G_2\Big(\{E_x^j u_{sh}(\varepsilon x,u_{_L})\}\Big)\Big)/\varepsilon^{2k-1},\\&P_1g_{\text{sum}}'=P_1\Big(g_{\text{sum}}-\sum_x G_2\Big(\{E_x^j u_{sh}(\varepsilon x,u_{_L})\}\Big)\Big)/\varepsilon^{k-1}.\end{aligned}\tag{5.50}$$

Then

$$(P-P_1)g'_{\rm sum} = (\ell\varepsilon^k)O(\Delta u'_{_L}) + x'_0\Big(O(1)+O(\varepsilon^k\Delta u'_{_L})\Big) + \tau'_- \tag{5.51}$$

$$\begin{aligned} -P_1 g'_{\rm sum} = {} & (\ell\varepsilon^k)O(\Delta u'_{_L}) + x'_0\Big(2\lambda_1(u_{_L})/\varepsilon^k + o(1)\Big) \\ & -P_1\varepsilon^{-k+1}\sum_x G_2\Big(\{E_x^j\Delta u_{st}(x)\}\Big). \end{aligned} \tag{5.52}$$

By (5.47), $\Delta u'_{_L} = O(1)$, $P_1\varepsilon^{-k+1}\sum_x G_2(\{E_x^j\Delta u_{st}(x)\}) = O(\varepsilon) + O(e^{-\delta\varepsilon\ell})$. Since $\ell\varepsilon^k \ll 1$ and $\lambda_1(u_{_L})/\varepsilon^k = O^*(1)$, (5.51) and (5.52) are uniquely solvable with respect to x'_0 and τ'_-. In order to satisfy (5.39), (5.49) we should assume that

$$|P_1 g'_{\rm sum}| \ll 1, \qquad |(P-P_1)g'_{\rm sum}| \le K. \tag{5.53}$$

Thus we obtain an n_1 parameter family of stationary solutions which depend on the vector $Pg'_{\rm sum}$.

Now consider the special cases $n_1 = 1$ and $n_1 = n$ while $k \ge 1$ and $\ell\varepsilon^{k+1} \ll 1$. In the case $n_1 = 1$ the terms (5.31) which contribute to (5.19) are absent. Thus, instead of (5.33), we obtain

$$\begin{aligned} \|y_{_{I,II}}\|_1 + |\Delta u_{_L}| \le K\Big(& \|H_{_{I,II}}\|_1 + \varepsilon^k\|y_1\|_\infty + \|(E_x-I)^k y_1\|_\infty \\ & + \varepsilon^{2k}e^{-\delta\varepsilon\ell} + |\Delta u_{_L}|^2 + \|\Delta u_{st}\|_\infty^2\Big). \end{aligned} \tag{5.54}$$

Besides the norm $\|\cdot\|_1$, we can estimate the norm $\|y_{_{I,II}}\|_\infty$,

$$\begin{aligned} \|y_{_{I,II}}\|_\infty + |\Delta u_{_L}| \le K\Big(& \|H_{_{I,II}}\|_\infty + \varepsilon^k\|y_1\|_\infty + \|(E_x-I)^k y_1\|_\infty \\ & + \varepsilon^{2k}e^{-\delta\varepsilon\ell} + |\Delta u_{_L}|^2 + \|\Delta u_{st}\|_\infty^2\Big), \end{aligned} \tag{5.55}$$

$$\begin{aligned} & \|y_{_{I,II}}\|_\infty + K_1\varepsilon^k \sum_{i+j\le k+1}\Big(\|D_\tau^i y'_{1,j}\|_{1,\varepsilon} + \|D_\tau^i y'_{1,j}\|_\infty\Big) + |\Delta u_{_L}| \\ & \le K\Big(\|H_{_{I,II}}\|_\infty + \varepsilon^{2k}e^{-\delta\varepsilon\ell}\Big) + K_1K\Big(\varepsilon\|H_1\|_1 + \sum\varepsilon^{k-d_j}|\Delta u_{_L}| \\ & \qquad + \varepsilon^{2k}e^{-\delta\varepsilon\ell} + \varepsilon\Big(\|\Delta u_{st}\|_\infty^2 + |\Delta u_{_L}|^2\Big)\Big). \end{aligned} \tag{5.56}$$

The contribution of y to the norm of H at the r.h.s. of (5.55) is bounded by $K\varepsilon^{k+1}\|y\|_\infty$ $+K_1\varepsilon^{k+1}\|y\|_\infty$, which is negligible. Now, we have a relaxed condition $d_j \le k-1$. Thus $\Delta u_{_L}$ term at the r.h.s. of (5.56) could be dropped. Altogether,

$$\begin{aligned} & \|y_{_{I,II}}\|_\infty + \varepsilon^k\Big(\sum\|D_\tau^i y'_{1,j}\|_{1,\varepsilon} + \|D_\tau^i y'_{1,j}\|_\infty\Big) + |\Delta u_{_L}| \\ & \qquad \le K\Big(\|F\|_\infty + \varepsilon^{2k}e^{-\delta\varepsilon\ell} + \varepsilon\|\Delta u_{st}\|_\infty^2\Big) \\ & \qquad \le K\Big(\varepsilon^{2k} + \|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty\Big). \end{aligned} \tag{5.57}$$

Now return to (5.34). The term with H_1' in the r.h.s. of (5.34) is estimated as in (5.36), (5.37) with a correction as in (5.41). Thus we obtain

$$\begin{aligned}
&\sum_{i+j\le k+1}\Big(\|D_\tau^i y'_{1,j}\|_{1,\varepsilon}+\|D_\tau^i y'_{1,j}\|_\infty\Big)\\
&\qquad\le K\Big(\varepsilon^{-k+1}\Big(\varepsilon^{2k}+|\Delta u_{_L}|+\|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty+\varepsilon^k\|y\|_\infty\Big)\\
&\qquad\qquad+e^{-\delta\varepsilon\ell}\varepsilon^k+e^{-\delta\varepsilon\ell}|\Delta u_{_L}|+\varepsilon^{-k+1}\Big(\|\Delta u_{st}\|_\infty^2+|\Delta u_{_L}|^2\Big)\Big)\\
&\qquad\le K\Big(\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k+\varepsilon^{-k+1}\|\Delta u_{st}\|_1\cdot\|\Delta u_{st}\|_\infty\Big).
\end{aligned}\tag{5.58}$$

Then estimate (5.54) implies

$$\|y_{_{I,II}}\|_1\le K\Big(\varepsilon^{2k-1}+\|\Delta u_{st}\|_1\|\Delta u_{st}\|_\infty\Big).\tag{5.59}$$

By (5.57)–(5.59),

$$\|\Delta u_{st}\|_\infty\le K\Big(\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k+\varepsilon^{-k+1}\|\Delta u_{st}\|_1\|\Delta u_{st}\|_\infty\Big)\tag{5.60}$$

and

$$\|\Delta u_{st}\|_1\le K\Big(\varepsilon^k+e^{-\delta\varepsilon\ell}\varepsilon^{k-1}+\varepsilon^{-k}\|\Delta u_{st}\|_1\|\Delta u_{st}\|_\infty\Big).\tag{5.61}$$

Hence

$$\varepsilon^{-2k+1}\|\Delta u_{st}\|_\infty\|\Delta u_{st}\|_1=A\le K^2(\varepsilon+e^{-\delta\varepsilon\ell}+A)^2\tag{5.62}$$

or $A\le K(\varepsilon+e^{-\delta\varepsilon\ell})$. Now, by (5.57)–(5.59) we arrive at estimates (5.47). For $(I-P_1)\Delta u_{st}$, $\Delta u_{_L}$ we have a better estimate

$$\|(I-P_1)\Delta u_{st}\|_\infty+|\Delta u_{_L}|\le K(\varepsilon^{2k}+e^{-\delta\varepsilon\ell}\varepsilon^{2k-1}).\tag{5.63}$$

Then, as in (5.48),

$$\begin{aligned}
&\varepsilon\|\Delta u_{st}\|_1+\|\Delta u_{st}\|_\infty+\varepsilon|\Delta u_{_L}|\ell+K\varepsilon|x_0|\varepsilon^k\\
&\qquad\le K\Big(\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k+\varepsilon^{k+1}\ell\cdot\varepsilon^k+(\varepsilon\ell)e^{-\delta\varepsilon\ell}\varepsilon^{2k-1}+\varepsilon|x_0|\varepsilon^k\Big)\\
&\qquad\le\delta_0\varepsilon^k,
\end{aligned}\tag{5.64}$$

provided $\varepsilon^{k+1}\ell\ll 1$, $\varepsilon|x_0|\ll 1$, $\varepsilon\ell\gg 1$. The difference $(E_x-I)\Delta u_{st}$ is estimated with the aid of (5.58), (5.59), (5.62),

$$\|(E_x-I)\Delta u_{st}\|_1\le K\Big(\|y_{_{I,II}}\|_1+\varepsilon\|D_\tau y_1\|_1\Big)\le K(\varepsilon^{2k-1}+\varepsilon^{k+1}+e^{-\delta\varepsilon\ell}\varepsilon^k).\tag{5.65}$$

Thus, for $n_1=1$ we obtained the requested estimates (2.87), (2.88). The only free parameter in the definition of Δu_{st} is $x_0'=\varepsilon x_0$. It is related to $P_1 g'_{\rm sum}$ by (5.52). Note that $\Delta u'_{_L}=\Delta u_{_L}/\varepsilon^{2k-1}=O(\varepsilon)+O(e^{-\delta\varepsilon\ell})$. Hence

$$\ell\varepsilon^k O(\Delta u'_{_L})=O(\ell\varepsilon^{k+1})+O(\ell\varepsilon^k e^{-\delta\varepsilon\ell})\ll 1.\tag{5.66}$$

Thus x_0' in (5.52) is in one-to-one correspondence with $P_1 g'_{\rm sum}$, provided $|P_1 g'_{\rm sum}| \ll 1$.

It remains to consider the case $n_1 = n$, when $\Delta u_{_L} = 0$. Estimate (5.35) is valid with $\Delta u_{_L} = 0$. We can afford $d_j \le k-1$ since there are no terms $\varepsilon^{k-1}\varepsilon^{-d_j} O(\Delta u_{_L})$ at the r.h.s. of (5.35). We proceed as in (5.36)–(5.39) and arrive at (5.40) with $(\Delta u_{_L} = 0)$. All estimates (5.41)–(5.46) are valid; only the term ε^{4k-5} in (5.45), (5.46) is absent. Thus there is no restriction $k > 1$. Hence estimate (5.47) holds also for $k = 1$. Instead of (5.48) we have

$$\varepsilon\|\Delta u_{st}\|_1 + \|\Delta u_{st}\|_\infty + K\varepsilon|x_0|\varepsilon^k \le K\Big(\varepsilon^{k+1} + e^{-\delta\varepsilon\ell}\varepsilon^k + \varepsilon|x_0|\varepsilon^k\Big) \le \delta_0\varepsilon^k, \tag{5.67}$$

provided (5.49) holds. Equations (5.51), (5.52) (with $\Delta u_{_L} = 0$) define one-to-one correspondence between $g'_{\rm sum}$ and x_0' and τ'_- under the restriction (5.53).

6. Asymptotic stability of the nonlinear problem. In this section we will prove the main theorem, Theorem 2.2, about asymptotic stability of the IBVP (2.1), (2.7), (2.9). Let u be a solution of the IBVP and u_{st} a stationary solution of (2.1), (2.7) as described in section 5. Denote by Δu the difference $u - u_{st}$. Since G has the form (2.4) with linear operator G_2, Δu satisfies the equation

$$dG[u_{st}](E_x, E_t)\Delta u = (E_x - I)O(\Delta u)^2 = F. \tag{6.1}$$

In case $n_1 = n$, one should integrate the above equation. Thus we consider first the case $n_1 = 1$. Since both u and u_{st} satisfy the boundary conditions (2.7), the difference Δu satisfies

$$dS_{_{L,R}}[\Delta u_{st}]\Delta u = O(\Delta u)^2 = g_{_{L,R}}. \tag{6.2}$$

However, we remove from (6.2) the boundary condition (2.13) at $x = x_\ell$ and replace it by the integral condition (6.5) below. The initial values of Δu are

$$\Delta u_{in}(x,t) = u_{in}(x,t) - u_{st}(x), \qquad (x,t) \text{ as in (2.9)}. \tag{6.3}$$

We will assume that $g_{\rm sum}$ defined by the initial values u_{in},

$$P_1 \sum_x G_2\big(\{E^j u_{in}\}\big) = g_{\rm sum}, \tag{6.4}$$

satisfies conditions (5.53), where $g'_{\rm sum}$ is defined by (5.50). Then, as shown in section 5, one can choose u_{st} such that

$$P_1 \sum_x dG_2 \Delta u \equiv \Delta g_{\rm sum} = 0. \tag{6.5}$$

The estimates for the IBVP problem (6.1)–(6.3), (6.5) follow from estimates (3.100), (3.132), (3.133) for the resolvent problem. Although the procedure is standard, for completeness we present it here. First we rewrite the problem (6.1), (6.2), (6.5) in global operator form,

$$L(E_t)\Delta u = \sum_{i=0}^{\nu_0} L_i E_t^i \Delta u = F^{(1)} = (F, g_{_{L,R}}, \Delta g_{\rm sum}), \tag{6.6}$$

where $\nu_0 = \Delta j_2$ is defined in (2.9). Then problem (6.6) is transformed as in (3.19) into a two-level scheme

$$\widetilde{L}(E_t)\widetilde{u} = (A \cdot E_t + B)\Delta\widetilde{u} = \widetilde{F}, \tag{6.7}$$

where

$$\Delta\widetilde{u}(t) = \Big(\Delta u(t), E_t\Delta u(t), \dots, E_t^{\nu_0-1}\Delta u(t)\Big)^T, \qquad \widetilde{F} = (0, \dots, 0, F^{(1)})^T. \tag{6.8}$$

The operators $\widetilde{L}(E_t)$ and $L(E_t)$ are related by the identity (3.20). The explicit form of D_1 and D_2 is

$$D_1 = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ E_t & I & 0 & & 0 \\ E_t^2 & E_t & I & & 0 \\ \vdots & & & \ddots & \\ E_t^{\nu_0-1} & \cdots & E_t^2 & E_t & I \end{bmatrix}, \qquad D_2 = \begin{bmatrix} C_{\nu_0-1} & C_{\nu_0-2} & \cdots & C_1 & I \\ -I & 0 & & 0 & 0 \\ & -I & & & \\ \vdots & & & & \vdots \\ 0 & \cdots & & -I & 0 \end{bmatrix}, \tag{6.9}$$

where

$$C_1 = L_{\nu_0}E_t + L_{\nu_0-1}, \quad C_{\nu+1} = C_\nu E_t + L_{\nu_0-1-\nu} \quad \text{for } \nu = 1, 2, \dots, \nu_0 - 2.$$

By the solvability assumption, the operator A is invertible. Thus the problem (6.7), (6.3) becomes

$$E_t\widetilde{u} = -A^{-1}B\widetilde{u} + A^{-1}\widetilde{F}, \qquad \Delta\widetilde{u}(0) = \Delta\widetilde{u}_{in}. \tag{6.10}$$

Recall that the estimates for the resolvent problem could be extended into the domain $\operatorname{Re} s \ge -\delta\varepsilon^k/\ell$. As a result, the difference $\Delta\widetilde{u}$ will decay in time as $\exp(-s_0 t)$, $s_0 \approx \delta\varepsilon^k/\ell$. Let us fix $s_0 = \delta\varepsilon^k/2\ell$ and change the variables

$$\Delta\widetilde{u} = e^{-s_0 t}v, \qquad e^{s_0(t+1)}A^{-1}\widetilde{F}(t) = H(t). \tag{6.11}$$

Then v satisfies

$$E_t v = \widetilde{A}v + H, \qquad \widetilde{A} = -e^{s_0}A^{-1}B. \tag{6.12}$$

The solution v is a convolution

$$v(t) = \sum_{\tau=0}^{t} \widetilde{A}^\tau H(t-\tau) + \widetilde{A}^t\Delta\widetilde{u}_{in}. \tag{6.13}$$

The power $\widetilde{A}$ is computed by the integral

$$\begin{aligned} \widetilde{A}^\tau \cdot H &= (2\pi i)^{-1}\oint_{|z|=e^{-s_0}} (z - \widetilde{A})^{-1} z^\tau H dz \\ &= (2\pi i)^{-1}\oint_{|z|=e^{-s_0}} (Ae^{-s_0}z + B)^{-1} z^\tau e^{-s_0} AH dz \\ &= (2\pi i)^{-1}\oint_{|z|=e^{-2s_0}} D_1(z)\Big(L(z) \oplus I\Big)^{-1} D_2(z)(ze^{s_0})^\tau AH dz \\ &= (2\pi i)^{-1}\oint_{|z|=e^{-2s_0}} D_1(z)\Big(L^{-1}(z) \oplus 0\Big) D_2(z)(ze^{s_0})^\tau AH dz. \end{aligned} \tag{6.14}$$

Note that

$$D_2(z)AH(t) = D_2(z)e^{s_0(t+1)}\widetilde{F}(t) = e^{s_0(t+1)}\Big(F^{(1)}(t),0,\dots,0\Big). \tag{6.15}$$

Since $D_1(z)$ is bounded,

$$\|\widetilde{A}^\tau H(t-\tau)\|_\infty \le Ke^{-s_0\tau}\oint_{|z|=e^{-2s_0}} \|L^{-1}(z)F^{(1)}(t-\tau)e^{s_0(t-\tau)}\|_\infty |dz|. \tag{6.16}$$

For $z=e^s$, $\operatorname{Re} s = -2s_0$, $|\operatorname{Im} s| \le K\varepsilon^{k+1}$, $u = L^{-1}(z)F^{(1)}$ satisfies estimate (3.132) and for $K\varepsilon^{k+1} \le |\operatorname{Im} s| \le \delta$ satisfies estimate (3.133). For s away from 0, by dissipativity, the operator dG is zero order elliptic. However, because of the global condition (2.14) we still have the norm $\|F\|_1$ at the r.h.s. of (3.133). Estimate (3.133) could be strengthened, but it would not improve the estimate for the power of $\widetilde{A}$. Now return to the integral in (6.16). In the domain $|\operatorname{Im} s| \le K\varepsilon^{k+1}$, this integral is bounded by

$$\begin{aligned} &K\int \sigma^{-k}\bigg(\|F\|_1 + \sigma\|P_1\sum F(\xi)\|_1 + \sigma\ell|g_{L,0}| \\ &\qquad + \sum_{i\ge 1}\sigma^{k-d_i}|g_{1,i}| + |g_2|\bigg)|dz| \\ &\le K\varepsilon^{k+1}\varepsilon^{-k}\bigg(\|F\|_1 + \varepsilon\|P_1\sum F(\xi)\|_1 + \varepsilon\ell|g_{L,0}| \\ &\qquad + \sum_{i\ge 1}\varepsilon^{k-d_i}|g_{1,i}| + |g_2|\bigg). \end{aligned} \tag{6.17}$$

In the domain $|\operatorname{Im} s| \ge K\varepsilon^{k+1}$ the integral is bounded by

$$K\left(\|F\|_1 + |g_{L,0}| + \sum_{i\ge 1}|g_{1,i}| + |g_2|\right). \tag{6.18}$$

Altogether,

$$\begin{aligned} &\|\widetilde{A}^\tau H(t-\tau)\|_\infty \\ &\quad \le Ke^{-s_0\tau}e^{s_0(t-\tau)}\Big(\|(E_x-I)O(\Delta u)^2\|_1 + \|O(\Delta u)^2\|_\infty \\ &\qquad\qquad + \varepsilon^2\Big(\|O(\Delta u)^2\|_1 + \ell\|O(\Delta u)^2\|_\infty\Big)\Big) \\ &\quad \le Ke^{-s_0 t}\ell\|v(t-\tau)\|_\infty^2 \end{aligned} \tag{6.19}$$

and

$$\left\|\sum_{\tau=0}^{t}\widetilde{A}^\tau H(t-\tau)\right\| \le Kte^{-s_0 t}\ell \sup_{0\le\tau\le t}\|v(\tau)\|_\infty^2. \tag{6.20}$$

In the initial term $\widetilde{A}^t\Delta\widetilde{u}_{in}$ we should apply the last integral in (6.14) to the function $A\Delta\widetilde{u}_{in}$. By (3.18) and (6.9) the first block row of $D_2A\Delta\widetilde{u}_{in}$ is

$$C_{\nu_0-1}\Delta u_{in}(\nu_0-1) + \cdots + C_1\Delta u_{in}(1) + L_{\nu_0}\Delta u_{in}(0). \tag{6.21}$$

The resulting vector could be partitioned as $(F, g_{L,R}, g_{\text{sum}})$. The F component is affected by the operator $dG[u_{st}]$ in (6.1), $g_{L,R}$ by $dS_{L,R}[u_{st}]$ in (6.2), and g_{sum} by $P_1 \sum dG_2$. Recall that the variable coefficient part of dG is $(E_x - I)dG_1$. Hence the term $\|P_1 \sum F(\xi)\|_1$ is bounded by

$$\ell \|\Delta \widetilde{u}_{in}\|_\infty + \left\| \sum_\xi \Delta \widetilde{u}_{in}(\xi) \right\|_1 . \tag{6.22}$$

One should add to (6.17), (6.18) the $|g_{\text{sum}}|$ term which is bounded by $|\sum_\xi \Delta \widetilde{u}_{in}(\xi)|$. Altogether,

$$\|\widetilde{A}^t \Delta \widetilde{u}_{in}\|_\infty \le K e^{-s_0 t} \left(\varepsilon^2 \left\| \sum_\xi \Delta \widetilde{u}_{in}(\xi) \right\|_1 + \|\Delta \widetilde{u}_{in}\|_1 + \varepsilon^2 \ell \|\Delta \widetilde{u}_{in}\|_\infty \right). \tag{6.23}$$

In order to estimate $\sup_{0 \le \tau \le t} \|v(\tau)\|_\infty$, we should assume that

$$K t e^{-s_0 t} \ell \sup_{0 \le \tau \le t} \|v(\tau)\|_\infty < \tfrac{1}{2}. \tag{6.24}$$

Then in turn,

$$\|v(t)\|_\infty \le 2 \|\widetilde{A}^t \Delta \widetilde{u}_{in}\|_\infty. \tag{6.25}$$

Thus a sufficient condition for estimate (6.25) to hold for all t is

$$\begin{aligned} &\varepsilon^2 \left\| \textstyle\sum_\xi \Delta \widetilde{u}_{in}(\xi) \right\|_1 + \|\Delta \widetilde{u}_{in}\|_1 + \varepsilon^2 \ell \|\Delta \widetilde{u}_{in}\|_\infty \\ &\qquad \le \frac{\delta \ell^{-1}}{\max(t e^{-s_0 t})} \le \delta s_0 \ell^{-1} \approx \delta \varepsilon^k \ell^{-2} . \end{aligned} \tag{6.26}$$

The difference Δu decays as

$$\|\Delta u(t)\|_\infty \le K \varepsilon^k \ell^{-2} e^{-\delta \varepsilon^k \ell^{-1} t}. \tag{6.27}$$

For minimal possible $\ell \approx K \varepsilon^{-1}$, the rate of decay is

$$\|\Delta u(t)\|_\infty \le K \varepsilon^{k+2} e^{-\delta \varepsilon^{k+1} t}. \tag{6.28}$$

Now consider the case $n_1 = n$. The original nonlinear problem (2.1) is integrated,

$$G_1\Big(\{E^j (E_x - I) U(x,t)\}\Big) + (E_t - I) G_2\Big(\{E^j U(x,t)\}\Big) = f(u_{_L}), \tag{6.29}$$

where $(E_x - I)U(x,t) = u(x,t)$ and $U(x,t)$ satisfies the boundary condition

$$\begin{aligned} &\text{(a)} \quad G_2\Big(\{E^j U(x_{-\ell}, t)\}\Big) = 0, \\ &\text{(b)} \quad G_2\Big(\{E^j U(x_\ell, t)\}\Big) = g_{\text{sum}} \stackrel{df}{=} \textstyle\sum_x G_2\Big(\{E^j u_{in}(x)\}\Big). \end{aligned} \tag{6.30}$$

The boundary conditions at $x = x_{-\ell}$ are used to define the value $U(x_{-\ell}, t)$, $t \ge \nu_0$, while for $t < \nu_0$ the values $U(x_{-\ell}, t)$ are arbitrary. The remaining boundary conditions are the one in (2.26)(b),(c) or, more precisely, in the nonlinear form

$$S_1\Big((E_x - I)U\Big) = 0, \qquad S_2\Big((E_x - I)U\Big) = 0. \tag{6.31}$$

Given stationary solution u_{st}, which satisfies the conservation law $\sum_x G_2(\{E^j u_{st}(x)\}) = g_{\rm sum}$, one can define U_{st}, $(E_x - I)U_{st} = u_{st}$, which satisfies (6.30)(a) and therefore also (6.30)(b). The difference $\Delta U = U - U_{st}$ satisfies the equations

$$dG_1[u_{st}](E_x, E_t)(E_x - I)\Delta U + (E_t - I)dG_2\Delta U = F = O\Big((E_x - I)\Delta U\Big)^2, \tag{6.32}$$

$$dG_2 \cdot \Delta U = 0, \qquad x = x_{\pm\ell}, \tag{6.33}$$

and

$$dS_{1,2}[u_{st}]\Delta U = g_{1,2} = O\Big((E_x - I)\Delta U\Big)^2, \tag{6.34}$$

with initial conditions

$$\Delta U(x,t) = \Delta U_{in}, \qquad 0 \le t \le \nu_0 - 1. \tag{6.35}$$

We proceed as in the case $n_1 = 1$. Now we use estimate (3.100). We remark that the function F in (6.32) stands for the sum $\sum F(\xi)$ in (3.99). The terms $g_{\rm sum}$ and $g_{L,0}$ are 0. Therefore instead of (6.19) we obtain

$$\|\widetilde{A}^\tau H(t-\tau)\|_\infty \le Ke^{-s_0\tau}e^{s_0(t-\tau)}\Big(\|F\|_1 + |g_{1,2}|\Big) \le Ke^{-s_0 t}\ell\|v(t-\tau)\|_\infty^2. \tag{6.36}$$

The initial term is estimated as

$$\|\widetilde{A}^t\Delta\widetilde{U}_{in}\|_\infty \le Ke^{-s_0 t}\|\Delta\widetilde{U}_{in}\|_1. \tag{6.37}$$

Estimate (6.25) is replaced by

$$\|V(t)\|_\infty \le 2\|\widetilde{A}^t\Delta\widetilde{U}_{in}\|_\infty, \tag{6.38}$$

and condition (6.26) by

$$\|\Delta\widetilde{U}_{in}\|_1 \le \delta\varepsilon^k\ell^{-2}. \tag{6.39}$$

Finally, instead of (6.27) we have

$$\|\Delta U\|_1 \le K\varepsilon^k\ell^{-2}e^{-\delta\varepsilon^k\ell^{-1}t}. \tag{6.40}$$

REFERENCES

[1] N. Kopell and L.N. Howard, *Bifurcations and trajectories joining critical points*, Adv. Math., 18 (1975), pp. 306–358.

[2] C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Reg. Conf. Ser. Math. 38, AMS, Providence, RI, 1978.

[3] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Ration. Mech. Anal., 95 (1985), pp. 325–344.

[4] T.P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985), pp. 1–108.

[5] D. Michelson, *Stability of the Bunsen flame profiles in the Kuramoto-Sivashinsky equation*, SIAM J. Math. Anal., 27 (1996), pp. 765–781.

[6] S. Engelberg, *An analytical proof of the linear stability of the viscous shock profile of the Burger equation with fourth-order viscosity*, SIAM J. Math. Anal., 30 (1999), pp. 927–936.

[7] D. Michelson, *Discrete shocks for difference approximations to systems of conservation laws*, Adv. in Appl. Math., 5 (1984), pp. 433–469.
[8] D. Michelson, *Stability theory of difference approximations for multidimensional initial-boundary value problems*, Math. Comp., 40 (1983), pp. 1–45.
[9] M. Goldberg and E. Tadmor, *Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems*, Math. Comp., 32 (1978), pp. 1097–1107.
[10] G. Kreiss and H.-O. Kreiss, *Stability of systems of viscous conservation laws*, Comm. Pure Appl. Math., 51 (1998), pp. 1397–1424.
[11] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.
[12] T.P. Liu and S.-H. Yu, *Continuum shock profiles for discrete conservation laws,* I. *Construction;* II. *Stability*, Comm. Pure. Appl. Math., 52 (1999), pp. 85–128; 1047–1073.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 872–890

GEOMETRIC PROPERTIES OF RUNGE–KUTTA DISCRETIZATIONS FOR INDEX 2 DIFFERENTIAL ALGEBRAIC EQUATIONS*

JOHANNES SCHROPP†

Abstract. We analyze Runge–Kutta discretizations applied to index 2 differential algebraic equations (DAEs). The asymptotic features of the numerical and the exact solutions are compared. It is shown that Runge–Kutta methods satisfying the first order constraint condition of the DAE correctly reproduce the geometric properties of the continuous system. The proof combines embedding techniques of index 2 DAEs and ordinary differential equations (ODEs) with some invariant manifolds results of Nipp and Stoffer [*Attractive Invariant Manifolds for Maps*, SAM Research Report 92-11, ETH, Zurich, Switzerland, 1992]. The results support the favorable behavior of these Runge–Kutta methods applied to index 2 DAEs for $t \geq 0$.

Key words. differential algebraic systems, projected and half-explicit Runge–Kutta methods, invariant manifolds

AMS subject classifications. 34C05, 34C40, 65L05

PII. S0036142900376626

1. Introduction.

Differential algebraic problems of index 2 frequently arise when modeling phenomena from scientific computations. (An important class of such problems is, e.g., multibody systems with constraints on the velocity level or in the Gear–Gupta–Leimkuhler formulation [8].) They also occur as auxiliary systems for minimization problems when searching for an evolution that approaches a local minimum of an objective function restricted by algebraic constraints (see, e.g., Schropp [16]).

Quite often, analytic treatment of the system is impossible and, hence, numerical simulations become important to gain a deeper understanding of the global behavior. In particular, the question arises of which qualitative properties of the continuous system are preserved by a numerical method.

In the present paper we analyze the behavior of some widely used Runge–Kutta type discretizations applied to index 2 differential algebraic equations (DAEs) in Hessenberg form. To be more precise, we focus our interest on two different aspects. It is well known that the solution flow of the DAE takes place in a submanifold of the state times control space. We characterize how that submanifold persists under discretization with projected and half-explicit Runge–Kutta methods. We show that the discretized dynamics possesses an invariant submanifold close to the original one. Moreover, we deal with the following subject. The index 0 formulation of the DAE is an ordinary differential equation (ODE) on the manifold defined by the first order constrained condition. Runge–Kutta schemes satisfying that condition can be regarded as discrete flows on the constrained manifold. Hence, questions about the behavior of numerical methods near invariant sets like stationary points or periodic orbits on manifolds are of interest.

*Received by the editors August 4, 2000; accepted for publication (in revised form) February 4, 2002; published electronically August 1, 2002.
http://www.siam.org/journals/sinum/40-3/37662.html
†Department of Mathematics and Statistics, University of Konstanz, P.O. Box 5560, D-78434 Konstanz, Germany (johannes.schropp@uni-konstanz.de).

Our main tools are embedding and invariant manifold techniques. We embed the original DAE into another DAE of the same index such that the corresponding index 0 ODE admits a representation as a dynamical system on an open subset of the Euclidian space. Then, using the discrete invariant manifold techniques of Nipp and Stoffer [14], we mimic that approach for the projected and half-explicit Runge–Kutta dynamics. An important feature of this method is that it makes general techniques of regular perturbation methods available for DAEs. Applying the results of Beyn [2] for one-step methods in $\mathbb{R}^N$, we can establish that hyperbolic periodic orbits persist under discretization in invariant closed curves. Moreover, in Schropp [17] it is shown that the phase portrait near hyperbolic equilibria is correctly reproduced. These results underpin the use of projected or half-explicit Runge–Kutta DAE methods when dealing with the behavior of index 2 DAEs in the long time run.

Our work was largely motivated by the discretization results of Beyn [2], [3], Garay, [7], and Kloeden and Lorenz [12] for one-step methods near compact, invariant sets. Later we learned the convergence theory for DAEs from the excellent book of Hairer, Lubich, and Roche [10]. After finishing our paper we were informed about a forthcoming paper of Nipp [13], in which a persistence result of the invariant manifold of an index 2 DAE under discretization is shown for the special class of stiffly accurate Runge–Kutta methods and for linear multistep methods of backward differentiation formulae (BDF) type. Nipp's result is obtained with a different method of proof which does not allow analyzing the discrete behavior near compact invariant sets.

2. The main results. We consider the DAE

$$\begin{aligned} \dot{u} &= f(u,\lambda), \quad u(0) = u_0, \\ 0 &= g(u), \qquad \lambda(0) = \lambda_0, \end{aligned} \tag{2.1}$$

with $u \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}^l$ in Hessenberg form. Let C_b^ν denote the space of functions of class C^ν with bounded derivatives up to order ν. We make the following assumptions:

(A1) $f \in C_b^\nu(\mathbb{R}^{N+l}, \mathbb{R}^N)$, $g \in C_b^{\nu+1}(\mathbb{R}^N, \mathbb{R}^l)$ for ν sufficiently large.

(A2) There is a C_b^ν-function ψ_0 satisfying $Dg(u)f(u,\psi_0(u)) = 0$ for $u \in D_\tau := \{u \in \mathbb{R}^N |\ \|g(u)\|_2 < \tau\}$, $\tau > 0$.

(A3) $Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi_0(u))$ is invertible for $u \in D_\tau$ and the inverse has bounded norm.

In particular, problem (2.1) is of index 2, and consistent initial values for (2.1) must satisfy $g(u_0) = 0$ and $Dg(u_0)f(u_0,\psi_0(u_0)) = 0$. Additionally, condition (A3) says that $Dg(u)$ is of full rank so that the second equation of (2.1) defines the submanifold $S := \{u \in \mathbb{R}^N \mid g(u) = 0\}$ of $\mathbb{R}^N$, and the underlying index 0 ODE reads

$$\dot{u} = f(u,\psi_0(u)), \qquad u(0) = u_0 \in S \tag{2.2}$$

(for an illustration of S and the dynamics on it, see Hairer and Wanner [11, p. 458]). We denote the solution flow of (2.2) with $\bar{u}(t,u_0)$, $u_0 \in S$. Then, (A2) implies the solution flow $(\bar{u}(t,u_0), \bar{\lambda}(t,u_0))$, $\bar{\lambda}(t,u_0) = \psi_0(\bar{u}(t,u_0))$ for (2.1). This means that the manifold

$$M_0 = \{(u,\lambda) \in D_\tau \times \mathbb{R}^l \mid g(u) = 0,\ \lambda = \psi_0(u)\}$$

is the phase space of the solution flow of (2.1).

We are interested in the qualitative, geometric features of s-stage Runge–Kutta type methods with Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}, \quad A = (a_{ij})_{1\le i,j\le s} \in \mathbb{R}^{s,s}, \quad b, c \in \mathbb{R}^s, \tag{2.3}$$

and constant step size h when applied to (2.1). The Runge–Kutta method possesses stage order q if

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \quad k = 1, \dots, q, \quad i = 1, \dots, s.$$

To avoid drift problems in the discrete long time run, we must focus our interest on Runge–Kutta type methods which retain the first order constraint $g(u) = 0$. This leads us to the widely used projected Runge–Kutta methods introduced by Ascher and Petzold [1] or to the half-explicit Runge–Kutta methods due to Hairer, Lubich, and Roche [9]. For the Butcher tableau of the projected Runge–Kutta method, we impose the following conditions:

(B1) The Runge–Kutta matrix A is invertible.
(B2) $R(\infty) = 1 - b^T A^{-1} \mathbb{I}$, $\mathbb{I} = (1, \dots, 1)$, satisfies $|R(\infty)| < 1$.
(B3) The method is of classical order p and possesses stage order q with $p \geq q \geq 1$.

Applied to (2.1), the projected Runge–Kutta method has the form

$$\begin{aligned} \tilde{u}_{n+1} &= u_n + h(b^T \otimes I) \bar{f}(U^n, \Lambda^n), \\ \lambda_{n+1} &= (1 - b^T A^{-1} \mathbb{I}) \lambda_n + (b^T A^{-1} \otimes I) \Lambda^n, \end{aligned} \tag{2.4}$$

where $U^n = (U_1^n, \dots, U_s^n) \in \mathbb{R}^{Ns}$, $\Lambda^n = (\Lambda_1^n, \dots, \Lambda_s^n) \in \mathbb{R}^{ls}$ denote the solution of the algebraic system

$$\begin{aligned} U - (\mathbb{I} \otimes u_n) &= h(A \otimes I) \bar{f}(U, \Lambda), \\ 0 &= \bar{g}(U), \end{aligned} \tag{2.5}$$

and $\bar{f}, \bar{g}$ stand for $\bar{f}(U^n, \Lambda^n) = (f(U_1^n, \Lambda_1^n), \dots, f(U_s^n, \Lambda_s^n))$, $\bar{g}(U^n) = (g(U_1^n), \dots, g(U_s^n))$. Finally, the projection step

$$\begin{aligned} u_{n+1} &= \tilde{u}_{n+1} + \frac{\partial}{\partial \lambda} f(u_{n+1}, \lambda_{n+1}) \gamma, \\ 0 &= g(u_{n+1}) \end{aligned} \tag{2.6}$$

determines u_{n+1}. In (2.6) the variable γ is needed for the projection only.

A Runge–Kutta method satisfying $a_{sj} = b_j$, $j = 1, \dots, s$, is called stiffly accurate. Stiffly accurate Runge–Kutta solutions satisfy the first order constraint $g(u) = 0$ and, hence, the projection step (2.6) is superfluous.

For the half-explicit Runge–Kutta methods, that is, $a_{i,j} = 0$ for $i \leq j$ in the Butcher tableau, we assume the following:

(B1′) $a_{i+1,i} \neq 0$ for $i = 1, \dots, s-1$ and $b_s \neq 0$.
(B2′) The method is of DAE order p (see Hairer and Wanner [11, Chap. VII.6] for conditions on A, b, c).

The application of a half-explicit Runge–Kutta method to (2.1) reads as follows: Solve (2.5) in the case $a_{i,j} = 0$ for $j \geq i$ and obtain U^n and Λ_i^n, $i = 1, \dots, s-1$. Then Λ_s^n and u_{n+1} are computed by

$$\begin{aligned} u_{n+1} &= u_n + h(b^T \otimes I) \bar{f}(U^n, \Lambda^n), \\ 0 &= g(u_{n+1}). \end{aligned} \tag{2.7}$$

With the matrix

$$\tilde{A} = \begin{pmatrix} a_{21} & & & & \\ a_{31} & a_{32} & & & \\ \vdots & & \ddots & & \\ a_{s1} & \dots & \dots & a_{ss-1} & \\ b_1 & \dots & \dots & b_{s-1} & b_s \end{pmatrix} \in \mathbb{R}^{s,s} \tag{2.8}$$

and $\tilde{U}^n := (U_2^n, \dots, U_s^n, u_{n+1})$, the half-explicit Runge–Kutta scheme can be written in the compact form

$$\begin{aligned} \tilde{U}^n - \mathbb{I} \otimes u_n &= h(\tilde{A} \otimes I)\bar{f}((u_n, \tilde{U}_1^n, \dots, \tilde{U}_{s-1}^n), \Lambda^n), \\ 0 &= \bar{g}(\tilde{U}^n) \end{aligned} \tag{2.9}$$

(see, e.g., formula (4.59) in Hairer, Lubich, and Roche [10]).

To compute the λ-component one has several possibilities. The most accurate is the computation of λ from the index 2 condition, that is, $\lambda_{n+1} = \psi_0(u_{n+1})$. Here we follow the more efficient approach of Hairer, Lubich, and Roche [10]. They propose to require $c_s = 1$ and take

$$\lambda_{n+1} = \Lambda_s^n. \tag{2.10}$$

Moreover, we assume

(B3′) $\Lambda_s^n - \bar{\lambda}(h, u_n) = O(h^r)$, $r \le p$ (see, e.g., Brasey and Hairer [5] for sufficient conditions on A, b, c).

The qualitative properties of the discrete schemes are characterized in the following.

THEOREM 2.1. *Consider the DAE* (2.1) *and assume* (A1)–(A3). *Let* (u_n, λ_n) *denote the sequences generated with a projected* (*half-explicit*) *Runge–Kutta method satisfying* (B1)–(B3) (*respectively,* (B1′)–(B3′)), *when applied to* (2.1) *with consistent initial values* (u_0, λ_0).

Then there exists a positive constant h_0 *such that for* $h < h_0$ *the iterates* (u_n, λ_n) *exist for* $n \in \mathbb{N}$. *Moreover, for* $h \in]0, h_0]$ *there is a* C_b^ν*-function* $\psi_{0,h} : S \to \mathbb{R}^l$, $S = \{u \in \mathbb{R}^N \mid g(u) = 0\}$ *satisfying the following assertions:*

(i) *The set* $M_{0,h} = \{(u, \lambda) \in D_\tau \times \mathbb{R}^l \mid g(u) = 0,\ \lambda = \psi_{0,h}(u)\}$ *is invariant for the projected* (*half-explicit*) *Runge–Kutta map* (2.4)–(2.6) (*respectively,* (2.9)–(2.10)).

(ii) *The manifold* $M_{0,h}$ *is uniformly attractive with the constant* $\chi_h = |R(\infty)| + O(h^{q+1})$ $(\chi_h = 0)$, *that is,* $\|\lambda_{n+1} - \psi_{0,h}(u_{n+1})\| \le \chi_h \|\lambda_n - \psi_{0,h}(u_n)\|$ *for every discrete evolution* (u_n, λ_n) *starting sufficiently close to* M_0.

(iii) *For every initial value* (u_0, λ_0) *with* $\|\lambda_0 - \psi_0(u_0)\|$ *sufficiently small, there is* $(\tilde{u}_0, \tilde{\lambda}_0) \in M_{0,h}$ *and* $c, \hat{c} > 0$ *such that the corresponding evolutions* (u_n, λ_n) *and* $(\tilde{u}_n, \tilde{\lambda}_n)$ *satisfy*

$$\begin{aligned} \|u_i - \tilde{u}_i\| &\le c\chi_h^i \|\lambda_0 - \psi_0(u_0)\|, \quad i \in \mathbb{N}, \\ \|\lambda_i - \tilde{\lambda}_i\| &\le \hat{c}\chi_h^i \|\lambda_0 - \psi_0(u_0)\|, \quad i \in \mathbb{N}. \end{aligned}$$

(iv) $\|\psi_0(u) - \psi_{0,h}(u)\| \le Ch^q$ $[Ch^r]$ *for* $u \in S$.

Remark. The invariant manifold $M_{0,h}$ in the projected Runge–Kutta case is highly attractive if $R(\infty) = 0$. The manifold is infinitely attractive, that is, $(u_1, \lambda_1) \in M_{0,h}$ for every (u_0, λ_0) with $\|\lambda_0 - \psi_0(u_0)\|$ sufficiently small, if $\chi_h = 0$. This is valid for half-explicit and stiffly accurate Runge–Kutta methods.

Next we characterize the behavior of the projected and half-explicit Runge–Kutta methods for index 2 DAEs (2.1) in a neighborhood of hyperbolic periodic orbits. Here, we call $(\bar{u}(t,u_0),\bar{\lambda}(t,u_0))$, $\bar{\lambda}(t,u_0)=\psi_0(\bar{u}(t,u_0))$, $\bar{u}(t,u_0)=\bar{u}(t+T,u_0)$ a hyperbolic T-periodic orbit of (2.1) if $\bar{u}(t,u_0)$ is a hyperbolic T-periodic solution of (2.2). This means that the linearized T-flow mapping $T_{u_0}(S)=N(Dg(u_0))$ into itself has the simple eigenvalue 1, and all other eigenvalues are off the unit circle. Here $N(Dg(u_0))$ stands for the null space of $Dg(u_0)$, and $T_{u_0}(S)$ denotes the tangential space of the manifold S at u_0.

THEOREM 2.2. *Let the assumptions of Theorem* 2.1 *hold and let* $(\bar{u}(t,u_0),\bar{\lambda}(t,u_0))$, $\bar{\lambda}(t,u_0)=\psi_0(\bar{u}(t,u_0))$ *be a hyperbolic* T*-periodic orbit of the DAE* (2.1). *Additionally, let* (u_n,λ_n) *be generated by applying the projected Runge–Kutta scheme* (2.4)–(2.6) (*half-explicit Runge–Kutta method* (2.9)–(2.10)) *to the DAE* (2.1).

Then for sufficiently small step size h *the* u*-component of the discrete dynamics possesses an invariant curve* $\gamma^h=u^h(\mathbb{R})$, $u^h(t)=u^h(t+T)$, *satisfying*

$$\max\{\|\bar{u}(t,u_0)-u^h(t)\| \mid t\in\mathbb{R}\}\le Ch^q\ [Ch^r].$$

As a direct consequence of Theorems 2.1 and 2.2, we obtain the following.

COROLLARY 2.3. *Under the hypotheses of Theorems* 2.1 *and* 2.2, *there is an invariant curve* $S^h(\mathbb{R})=(u^h(\mathbb{R}),\psi_{0,h}(u^h(\mathbb{R})))$, $S^h(t)=S^h(t+T)$ *for the projected* (*half-explicit*) *Runge–Kutta map such that*

$$\max\{\|(\bar{u}(t,u_0),\bar{\lambda}(t,u_0))-S^h(t)\| \mid t\in\mathbb{R}\}\le Ch^q\ [Ch^r]$$

is valid.

Theorem 2.2 and Corollary 2.3 show that half-explicit or projected Runge–Kutta methods correctly reproduce the phase portrait in a neighborhood of a periodic orbit. Moreover, this result can be regarded as the analogue of Theorem 2.1 of Beyn [2] for ODEs on manifolds of the form $g(u)=0$. Using the results of Garay [7] it is shown in Schropp [17] that the continuous time-h flow of (2.1) and its corresponding projected or half-explicit Runge–Kutta time-h map are locally topologically conjugate near a hyperbolic equilibrium.

The rest of the paper is organized as follows. In section 3 we present existence results for the projected and half-explicit Runge–Kutta schemes, in section 4 Theorem 2.1 is proved, and in section 5 we give a proof of Theorem 2.2.

3. Embedding techniques for index 2 DAEs. We have seen in the previous section that the corresponding index 0 version (2.2) to an index 2 DAE (2.1) is a dynamical system on a manifold. For technical reasons it is useful to embed (2.1) into another DAE of the same index such that their corresponding index 0 ODE provides an embedding of (2.2) in an open neighborhood of S in $\mathbb{R}^N$. This allows us to attack DAE problems with well-developed ODE methods on $\mathbb{R}^N$. Results for the original DAE (2.1) are then obtained by pulling back the results for the state variable u derived on an open subset of $\mathbb{R}^N$ to the manifold S.

Assuming (A1)–(A3), an embedding of (2.2) into D_{τ_0}, $\tau_0\in\,]0,\tau]$, sufficiently small can be established as follows. Consider the DAE

$$\begin{aligned}\dot{u}&=f(u,\lambda), & u(0)&=u_0,\\ \dot{v}&=-B(u)v, & v(0)&=v_0,\\ 0&=g(u)-v, & \lambda(0)&=\lambda_0\end{aligned}\tag{3.1}$$

and suppose $\mu_2(-B(u)) \le -\eta$, $\eta > 0$ for $u \in D_\tau$ with a C_b^ν-function $B(\cdot)$ on $\mathbb{R}^N$ (e.g., choose $B \equiv I$). Here $\mu_2(C)$ stands for the logarithmic norm of a matrix $C \in \mathbb{R}^{l,l}$ (see, e.g., Dekker and Verwer [6, p. 27] for a definition). Our first aim here is to show that (A1)–(A3) imply the following assertions for the DAE (3.1):

(A1′) $f \in C_b^\nu(\mathbb{R}^{N+l}, \mathbb{R}^N)$, $g \in C_b^{\nu+1}(\mathbb{R}^N, \mathbb{R}^l)$, and $B \in C_b^\nu(\mathbb{R}^N, \mathbb{R}^{l,l})$ for ν sufficiently large.

(A2′) There is $\tau_0 \in \,]0, \tau]$ and a C_b^ν-function ψ satisfying $Dg(u)f(u, \psi(u, v)) + B(u)v = 0$ for $u \in D_{\tau_0}$ and $\|v\|_2 < \tau_0$.

(A3′) $Dg(u)\frac{\partial f}{\partial \lambda}(u, \psi(u, v))$ is invertible for $u \in D_{\tau_0}$, $\|v\|_2 < \tau_0$ and the inverse has bounded norm.

(A1′) holds trivially for (3.1). To prove (A2′) and (A3′) we need the following version of the Banach fixed point theorem in a ball which is, for later purposes, formulated in the more general concept of vector norms.

A functional $|\cdot| : W \to \mathbb{R}^k$ on a vector space W is called a generalized norm if

$$\begin{aligned} |v| \ge 0, \quad |v| = 0 \Longleftrightarrow v = 0, \\ |v_1 + v_2| \le |v_1| + |v_2|, \\ |\alpha v| = |\alpha|_{\mathbb{R}} |v| \end{aligned} \tag{3.2}$$

holds with the natural ordering "$\le$" on $\mathbb{R}^k$. Here $|\cdot|_{\mathbb{R}}$ denotes the absolute value in $\mathbb{R}$. Every norm $\|\cdot\|_*$ in $\mathbb{R}^k$ defines a norm $\|\cdot\|$ in W via $\|v\| = \| \, |v| \, \|_*$.

LEMMA 3.1. *Let $(W, |\cdot|)$ be a Banach space with generalized norm $|\cdot|$ and let $B_r(v_0) := \{v \in W \mid |v - v_0| \le r\}$ for $r > 0$. Let the map $F : B_r(v_0) \mapsto W$ be continuously differentiable with invertible $DF(v_0)$. Moreover, for some nonnegative matrices $P, K \in \mathbb{R}^{k,k}$, we assume*

$$\begin{aligned} |DF(v_0)^{-1} z| \le P|z|, \quad z \in W, \\ |(DF(v_0) - DF(v))z| \le K|z|, \quad z \in W, \quad v \in B_r(v_0), \\ P|F(v_0)| < (I - PK)r. \end{aligned}$$

Then $F(v) = 0$ has a unique solution in $B_r(v_0)$. In addition, the matrix $I - PK$ is nonsingular and we have the stability inequality

$$|v - w| \le (I - PK)^{-1} P |F(v) - F(w)| \quad \forall v, w \in B_r(v_0).$$

A proof of Lemma 3.1 can be found in Beyn and Schropp [4].

We construct ψ by applying Lemma 3.1 to the equation

$$F_{u,v}(\zeta) := Dg(u)f(u, \psi_0(u) + \zeta) + B(u)v = 0. \tag{3.3}$$

For $\zeta_0 = 0$ we can compute with $\rho := \sup\{\|B(u)\|_2 \mid u \in D_\tau\} < \infty$ the inequalities

$$\begin{aligned} \|F_{u,v}(0)\|_2 \le \rho \|v\|_2, \\ \|DF_{u,v}(0)^{-1}\|_2 \le C \end{aligned}$$

as well as $DF_{u,v}(0) - DF_{u,v}(\zeta) = Dg(u)(\frac{\partial f}{\partial \lambda}(u, \psi_0(u)) - \frac{\partial f}{\partial \lambda}(u, \psi_0(u) + \zeta))$. Hence, we obtain

$$\|DF_{u,v}(0) - DF_{u,v}(\zeta)\|_2 \le \hat{C} r_0 \quad \text{for } \|\zeta\|_2 \le r_0.$$

Obviously, the inequality

$$C\|F_{u,v}(0)\|_2 \le C\rho\|v\|_2 < (1 - C\hat{C}r_0)r_0$$

holds for $\|v\|_2, r_0 > 0$ sufficiently small. Then Lemma 3.1 guarantees that (3.3) possesses exactly one solution $\hat{\zeta}_{u,v}$ in $B_{r_0}(0)$; that is, $\psi(u,v) := \psi_0(u) + \hat{\zeta}_{u,v}$ satisfies $Dg(u)f(u,\psi(u,v)) + B(u)v = 0$ for $u \in D_{\tau_0}$, $\|v\|_2 \le \tau_0$, and an implicit function argument ensures the smoothness of ψ. The reader may notice that we have $\psi_0(u) = \psi(u,0)$ by uniqueness.

Moreover, an application of the Banach lemma with $Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi_0(u))$ and the perturbation $Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi(u,v))$ shows that $Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi(u,v))$, $\|v\|_2 < \tau_0$ is invertible, the inverse possesses a bounded norm and (A1′)–(A3′) is verified.

(A1′)–(A3′) imply that (3.1) is of index 2. Consistent initial values must satisfy $g(u_0) - v_0 = 0$ and $Dg(u_0)f(u_0,\lambda_0) + B(u_0)v_0 = 0$. The solution flow of (3.1) has the form $(\tilde{u}(t,u_0), \tilde{v}(t,u_0), \tilde{\lambda}(t,u_0))$, $u_0 \in D_{\tau_0}$, with $\tilde{v}(t,u_0) = g(\tilde{u}(t,u_0))$ and $\tilde{\lambda}(t,u_0) = \psi(\tilde{u}(t,u_0), \tilde{v}(t,u_0))$. Moreover,

$$M_e := \{(u,v,\lambda) \in D_{\tau_0} \times \mathbb{R}^{2l} \mid g(u) - v = 0,\ \lambda = \psi(u,v)\}$$

is the phase space of (3.1). Using the theory of logarithmic norms (see, e.g., Dekker and Verwer [6, Thm. 1.5.2]), we obtain that

$$\|\tilde{v}(t,v_0)\|_2 \le \|v_0\|_2 \exp(-\eta t) \tag{3.4}$$

is valid for the v-component of every solution of (3.1). In particular, with $v(0) = v_0 = 0$ problem (3.1) reduces to (2.1). After eliminating the v-variables by $g(u) = v$, the u-component of the underlying index 0 ODE of (3.1) reads

$$\dot{u} = f(u,\psi(u,g(u))), \qquad u(0) = u_0 \in D_{\tau_0} \subset \mathbb{R}^N \text{ open.} \tag{3.5}$$

Next we summarize the qualitative properties of the solutions of (3.1).

LEMMA 3.2. *Consider* (3.1) *on the phase space* M_e*, and let* (A1)–(A3) *hold.*

Then every solution of (3.1) *with initial values* $u_0 \in D_{\tau_0}$, $v_0 = g(u_0)$, *and* $\lambda_0 = \psi(u_0,v_0)$ *exists for all* $t \ge 0$. *Moreover,* $M_{0,e} = \{(u,v,\lambda) \in D_{\tau_0} \times \mathbb{R}^{2l} \mid g(u) = v = 0, \lambda = \psi(u,0)\}$ *is an invariant and globally attractive subset of the phase space* M_e.

Proof. The proof of Lemma 3.2 is a direct consequence of (3.4) and the fact that f, g, B are C_b^ν-functions. ∎

We are interested in the behavior of s-stage projected and half-explicit Runge–Kutta type methods of order p with Butcher tableau (2.3) and constant step size h when applied to (3.1). The projected Runge–Kutta method has the form

$$\begin{aligned} \tilde{u}_{n+1} &= u_n + h(b^T \otimes I)\bar{f}(U^n,\Lambda^n), \\ v_{n+1} &= v_n - h(b^T \otimes I)\bar{B}(U^n)V^n, \\ \lambda_{n+1} &= (1 - b^T A^{-1}\mathbb{I})\lambda_n + (b^T A^{-1} \otimes I)\Lambda^n, \end{aligned} \tag{3.6}$$

where $U^n = (U_1^n, \dots, U_s^n) \in \mathbb{R}^{Ns}$, $V^n = (V_1^n, \dots, V_s^n) \in \mathbb{R}^{ls}$, $\Lambda^n = (\Lambda_1^n, \dots, \Lambda_s^n) \in \mathbb{R}^{ls}$ denote the solution of the algebraic system

$$\begin{aligned} U - (\mathbb{I} \otimes u_n) &= h(A \otimes I)\bar{f}(U,\Lambda), \\ V - (\mathbb{I} \otimes v_n) &= -h(A \otimes I)\bar{B}(U)V, \\ 0 &= \bar{g}(U) - V, \end{aligned} \tag{3.7}$$

and $\bar{B}$ stands for $\bar{B}(U^n) = \operatorname{diag}(B(U_1^n), \dots, B(U_s^n))$. Finally, the projection step

$$\begin{aligned} u_{n+1} &= \tilde{u}_{n+1} + \frac{\partial}{\partial \lambda} f(u_{n+1},\lambda_{n+1})\gamma, \\ 0 &= g(u_{n+1}) - v_{n+1} \end{aligned} \tag{3.8}$$

is used to compute u_{n+1}.

With $\tilde{A}$ from (2.8), $\tilde{U}^n := (U_2^n, \dots, U_s^n, u_{n+1})$, and $\tilde{V}^n = (V_2^n, \dots, V_s^n, v_{n+1})$, the application of half-explicit Runge–Kutta methods to (3.1) reads as follows:

$$\begin{aligned} \tilde{U}^n - \mathbb{I} \otimes u_n &= h(\tilde{A} \otimes I)\bar{f}((u_n, \tilde{U}_1^n, \dots, \tilde{U}_{s-1}^n), \Lambda^n), \\ \tilde{V}^n - \mathbb{I} \otimes g(u_n) &= -h(\tilde{A} \otimes I)\bar{B}((u_n, \tilde{U}_1^n, \dots, \tilde{U}_{s-1}^n))(g(u_n), \tilde{V}_1^n, \dots, \tilde{V}_{s-1}^n), \\ 0 &= \bar{g}(\tilde{U}^n) - \tilde{V}^n. \end{aligned} \tag{3.9}$$

This has to be completed by $\lambda_{n+1} = \psi(u_{n+1}, v_{n+1})$ or, more efficiently, by $\lambda_{n+1} = \Lambda_s^n$, provided $c_s = 1$ holds.

The reader may notice that (3.6)–(3.8) and (3.9) reduce to (2.4)–(2.6) and (2.9), respectively, when initialized with $g(u_0) = v_0 = 0$.

In this section we will guarantee the existence and uniqueness of the discrete iterates generated by a projected or a half-explicit Runge–Kutta method for all $n \in \mathbb{N}$. If one identifies the two state variables (u, v), then the solvability of the discrete systems (3.6)–(3.9) for $n \in \mathbb{N}$ with $0 \le nh \le t_{end}$ is guaranteed by the standard theory; see, e.g., Hairer and Wanner [11, Chaps. VII.4 and VII.6]. But in the process of proving Theorem 2.1 a refined stability inequality which distinguishes the two variables u and v is needed. To establish inequalities of that type we work with the concept of vector norms (see (3.2)).

LEMMA 3.3. *Let the assumptions of Theorem* 2.1 *hold and let* $u_0 \in D_{\tau_0}$, $v_0 = g(u_0)$, $\lambda_0 = \psi(u_0, v_0)$ *be a consistent initial value for the DAE* (3.1).

Then for $0 < h \le h_0$, $h_0 > 0$ *sufficiently small the projected and half-explicit Runge–Kutta iterates* (u_n, v_n, λ_n) *exist for* $n \in \mathbb{N}$. *For the stages* (U, V, Λ) *of the projected* (*respectively*, $(\tilde{U}, \tilde{V}, \Lambda)$ *of the half-explicit*) *Runge–Kutta dynamics, we have with* $v_0(u) := (\mathbb{I} \otimes u, \mathbb{I} \otimes g(u), \mathbb{I} \otimes \psi(u, g(u)))$, $u \in D_{\tau_0}$, *and* $k(u) := f(u, \psi(u, g(u)))$ *the inequality*

$$\begin{aligned} |(U, V, \Lambda) - v_0(u)| &\le O(h)(\|k(u)\| + \|g(u)\|)\ (1, 1, 1), \\ (|(\tilde{U}, \tilde{V}, \Lambda) - v_0(u)| &\le O(h)(\|k(u)\| + \|g(u)\|)\ (1, 1, 1)). \end{aligned} \tag{3.10}$$

Moreover, the coefficient $\gamma = \gamma(h, u, \lambda)$ *from the projection step* (3.8) *satisfies*

$$\|\gamma\| = O(h^{q+1}) \quad \text{for } u \in D_{\tau_0}, \quad \|\lambda - \psi(u, g(u))\| \text{ sufficiently small.}$$

Remark. The existence result for the half-explicit and projected Runge–Kutta methods applied to DAE (2.1) follows from Lemma 3.3 by restriction to initial values satisfying $g(u_0) = v_0 = 0$.

Now we give a proof of Lemma 3.3.

Proof. The first step of a projected Runge–Kutta method is a classical Runge–Kutta step. Following Hairer and Wanner [11, p. 493], we replace (3.7) for $h > 0$ by the equivalent system

$$\begin{aligned} U - (\mathbb{I} \otimes u_n) &= h(A \otimes I)\bar{f}(U, \Lambda), \\ V - (\mathbb{I} \otimes v_n) &= -h(A \otimes I)\bar{B}(U)V, \\ 0 &= \int_0^1 \operatorname{diag}(Dg(u_n + \tau(U_i^n - u_n)),\ i = 1, \dots, s)d\tau\ (A \otimes I)\bar{f}(U^n, \Lambda^n) \\ &\quad + (A \otimes I)\bar{B}(U)V + \frac{1}{h}(\bar{g}(\mathbb{I} \otimes u_n) - \mathbb{I} \otimes v_n). \end{aligned} \tag{3.11}$$

We use $v_n = g(u_n)$ and prove Lemma 3.3 by applying Lemma 3.1 to the equation

$$T_1(h,u,U,V,\Lambda) := \begin{pmatrix} U - (\mathbb{I}\otimes u) - h(A\otimes I)\bar{f}(U,\Lambda) \\ V - (\mathbb{I}\otimes g(u)) + h(A\otimes I)\bar{B}(U)V \\ \int_0^1 \operatorname{diag}(Dg(u+\tau(U_i-u)), i=1,\dots,s)d\tau\ (A\otimes I)\bar{f}(U,\Lambda) \\ +(A\otimes I)\bar{B}(U)V \end{pmatrix}$$

$$= 0, \quad u \in D_{\tau_0}. \tag{3.12}$$

Introducing the generalized norm $|(U,V,\Lambda)| = (\|U\|,\|V\|,\|\Lambda\|) \in \mathbb{R}^3$ and the central point $v_0(u) := (\mathbb{I}\otimes u, \mathbb{I}\otimes g(u), \mathbb{I}\otimes\psi(u,g(u)))$, we calculate

$$T_1(h,u,v_0(u)) = (O(h),O(h),0). \tag{3.13}$$

For the derivative of T_1 with respect to (U,V,Λ), we find with $\hat{\gamma}(u) := (u,\psi(u,g(u)))$ the representation

$$\frac{\partial}{\partial(U,V,\Lambda)}T_1(h,u,v_0(u)) = \begin{pmatrix} I+O(h) & 0 & O(h) \\ O(h) & I+O(h) & 0 \\ O(1) & O(1) & A\otimes Dg(u)\frac{\partial f}{\partial\lambda}(\hat{\gamma}(u)) \end{pmatrix}.$$

Now (A3′) and (B1) imply that $(A\otimes Dg(u)\frac{\partial f}{\partial\lambda}(\hat{\gamma}(u)))$ is nonsingular. Hence, the matrix $\frac{\partial}{\partial(U,V,\Lambda)}T_1(h,u,v_0(u))$ is invertible for $0 < h \le h_0$, $h_0 > 0$, sufficiently small and the inverse is of the form

$$\frac{\partial}{\partial(U,V,\Lambda)}T_1(h,u,v_0(u))^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ O(1) & O(1) & (A\otimes(Dg(u)\frac{\partial f}{\partial\lambda}(\hat{\gamma}(u))))^{-1} \end{pmatrix} + O(h).$$

In terms of vector norms this leads to $|\frac{\partial}{\partial(U,V,\Lambda)}T_1(h,u,v_0(u))^{-1}| \le P_h$ with

$$P_h := \begin{pmatrix} 1+O(h) & O(h) & O(h) \\ O(h) & 1+O(h) & O(h) \\ O(1) & O(1) & O(1) \end{pmatrix} \in \mathbb{R}^{3,3}.$$

Then, following along the lines of the proof of Lemma 4.1 in Beyn and Schropp [4], we obtain the unique solvability of (3.12) in $B_r(v_0) := \{(U,V,\Lambda) \in \mathbb{R}^{(N+l+l)s} \mid |(U,V,\Lambda) - (\mathbb{I}\otimes u, \mathbb{I}\otimes g(u), \mathbb{I}\otimes\psi(u,g(u)))| \le r\}$, $r = (r_1,r_2,r_3) > 0$ for $0 < h \le h_0$, $h_0 > 0$, sufficiently small. We remark that an application of the implicit function theorem ensures the smooth dependency of the solution (U,V,Λ) from (h,u). In addition, the claimed stability inequality (3.10) holds.

The second step is the projection of the classical Runge–Kutta iterates onto the constrained manifold $g(u) - v = 0$. We define the function $\lambda_p = \lambda_p(h,u,\lambda) := R(\infty)\lambda + (b^TA^{-1}\otimes I)\Lambda(h,u)$ and consider the equation

$$T_2(h,u,\lambda,u_p,v_p,\gamma) = \begin{pmatrix} u_p - u - h(b^T\otimes I)\bar{f}(U(h,u),\Lambda(h,u)) - \frac{\partial}{\partial\lambda}f(u_p,\lambda_p)\gamma \\ v_p - g(u) + h(b^T\otimes I)\bar{B}(U(h,u))V(h,u) \\ g(u_p) - v_p \end{pmatrix}$$

$$= 0 \quad \text{for } 0 < h < h_0, \quad u \in D_{\tau_0}. \tag{3.14}$$

With the central point $v_0(u) = (v_0(u)_1, v_0(u)_2, v_0(u)_3)$,

$$\begin{aligned} v_0(u)_1 &= u + h(b^T\otimes I)\bar{f}(U(h,u),\Lambda(h,u)), \\ v_0(u)_2 &= g(u) - h(b^T\otimes I)\bar{B}(U(h,u))V(h,u), \\ v_0(u)_3 &= 0, \end{aligned}$$

we can compute

$$T_2(h,u,\lambda,v_0(u)) = (0,0,g(v_0(u)_1) - v_0(u)_2) =: (0,0,r(h,u)).$$

Obviously, $r(h,u) = O(h^{q+1})$ holds from the local error analysis of the underlying classical Runge–Kutta map (see, e.g., Hairer and Wanner [11, Chap. VII, Lem. 4.4]).

Next, using $\lambda_p(h,u,\lambda) = R(\infty)\lambda + (b^T A^{-1} \otimes I)\Lambda(h,u) = \psi(u,g(u)) + O(h) + O(\epsilon)$ with $\epsilon = \|\lambda - \psi(u,g(u))\|$ and $v_0(u)_1 = u + O(h)$, we obtain

$$\frac{\partial}{\partial(u_p,v_p,\gamma)} T_2(h,u,\lambda,v_0(u)) = \begin{pmatrix} I & 0 & -\frac{\partial}{\partial\lambda} f(\hat\gamma(u)) \\ 0 & I & 0 \\ Dg(u) & -I & 0 \end{pmatrix} + O(h) + O(\epsilon). \tag{3.15}$$

Thus, $\frac{\partial}{\partial(u_p,v_p,\gamma)} T_2(h,u,\lambda,v_0(u))$ is invertible for h and ϵ sufficiently small.

Moreover, $T_2(h,u,\lambda,.,.,.) = 0$ possesses a unique solution in $B_r(v_0)$ for $r > 0$ appropriate and h, ϵ sufficiently small. Finally, the stability inequality of Lemma 3.1 in the γ-component reads

$$\|\gamma\| = O(h^{q+1}) \quad \text{for } u \in D_{\tau_0}, \qquad \|\lambda - \psi(u,g(u))\| < \epsilon.$$

It remains to show that the sequence (u_n, v_n, λ_n) generated by a projected Runge–Kutta method with consistent initial value u_0, $v_0 = g(u_0)$, $\lambda_0 = \psi(u_0, g(u_0))$ satisfies

$$\|\lambda_n - \psi(u_n,g(u_n))\| < \epsilon, \quad n \in \mathbb{N}, \quad 0 < h \le h_0, \quad h_0 > 0 \text{ sufficiently small.} \tag{3.16}$$

To that purpose we define $\eta_n := \lambda_n - \psi(u_n, g(u_n))$. The iteration scheme of this sequence reads

$$\begin{aligned} \eta_{n+1} &= R(\infty)\eta_n + \psi(u_n,g(u_n)) - \psi(u_{n+1},g(u_{n+1})) \\ &\quad + (b^T A^{-1} \otimes I)(\Lambda(h,u_n) - \mathbb{I} \otimes \psi(u_n,g(u_n))) \\ &=: R(\infty)\eta_n + \beta_n, \quad \eta_0 = 0, \end{aligned} \tag{3.17}$$

with $\beta_n := \beta(h,u_n,\eta_n) = O(h)$. Here, due to the construction of the method, we have u_{n+1} as a function of (h,u_n,η_n). Using $|R(\infty)| < 1$, the theory of difference equations yields

$$\|\eta_n\| \le \|\eta_0\| + \frac{1}{1 - R(\infty)} \sup\{\|\beta_n\| \mid n \in \mathbb{N}\} = O(h) \quad \forall n \in \mathbb{N}, \tag{3.18}$$

and (3.16) is verified.

Finally, we prove the existence of the iterates (u_n, v_n, λ_n) for half-explicit Runge–Kutta methods. We define $\tilde U = (U_2, \dots, U_s, u_p)$, $\tilde V = (V_2, \dots, V_s, v_p)$, $\Lambda = (\Lambda_1, \dots, \Lambda_s)$ as well as $U_1 = u$, $V_1 = g(u)$. Then, we rewrite (3.9) in the form

$$\begin{aligned} T(h,u,\tilde U,\tilde V,\Lambda) &= \begin{pmatrix} \tilde U - \mathbb{I} \otimes u - h(\tilde A \otimes I)\bar f((u,\tilde U_1,\dots,\tilde U_{s-1}),\Lambda) \\ \tilde V - \mathbb{I} \otimes g(u) + h(\tilde A \otimes I)\bar B((u,\tilde U_1,\dots,\tilde U_{s-1}))(g(u),\tilde V_1,\dots,\tilde V_{s-1}) \\ \bar g(\tilde U) - \tilde V \end{pmatrix} \\ &= 0 \end{aligned} \tag{3.19}$$

and apply Lemma 3.1. Except for the shift of (U,V) to $(\tilde U, \tilde V)$, this is equivalent to a classical Runge–Kutta step with invertible matrix $\tilde A$. Thus, we can adapt the first step of the projected Runge–Kutta proof with the central point $v_0(u) = (\mathbb{I} \otimes u, \mathbb{I} \otimes$

$g(u), \mathbb{I} \otimes \psi(u, g(u)))$ to work for half-explicit methods too. Moreover, for the stages $(\tilde{U}, \tilde{V}, \Lambda)$, the stability inequality

$$|(\tilde{U}, \tilde{V}, \Lambda) - v_0(u)| \leq Ch \left(\|k(u)\| + \|g(u)\|\right) (1, 1, 1) \tag{3.20}$$

is valid. This finishes the proof of Lemma 3.3. ∎

Next we complete the existence results of the projected and half-explicit Runge–Kutta methods with some qualitative properties. This is done in the following.

LEMMA 3.4. *Let the conditions of Lemma* 3.2 *hold for* (3.1). *By* (u_n, v_n, λ_n) *we denote the sequences generated with a projected Runge–Kutta method satisfying* (B1)–(B3) *or a half-explicit Runge–Kutta method fulfilling* (B1′)–(B3′) *when applied to* (3.1) *with initial values* $u_0 \in D_{\tau_0}$, $v_0 = g(u_0)$, *and* $\lambda_0 = \psi(u_0, v_0)$.

Then h_0, $\epsilon > 0$ *exist such that for* $0 < h < h_0$, $n \in \mathbb{N}$, *the projected or half-explicit Runge–Kutta scheme correctly reproduces the phase portrait of* (3.1) *in the state variables* (u, v); *that is,* $M_{0,e,\epsilon} := \{(u, v, \lambda) \in D_{\tau_0} \times \mathbb{R}^{2l} \mid g(u) = v = 0, \|\lambda - \psi(u, v)\| < \epsilon\}$ *is a positive invariant and globally attractive subset for the discrete dynamics.*

Proof. We consider the v-component of the discrete scheme (3.6)–(3.8) in more detail. We extract V from the second line in (3.7) explicitly, insert this representation into (3.6), and obtain

$$v_{n+1} = v_n - h(b^T \otimes I)\bar{B}(U(h, u_n))[I + h(A \otimes I)\bar{B}(U(h, u_n))]^{-1}(\mathbb{I} \otimes v_n). \tag{3.21}$$

Moreover, we know

$$\bar{B}(U(h, u_n)) = \bar{B}(\mathbb{I} \otimes u_n) + O(h) = (I \otimes B(u_n)) + O(h) \tag{3.22}$$

from Lemma 3.3. Combining (3.21) and (3.22) yields

$$\begin{aligned} v_{n+1} &= v_n - h(b^T \otimes I)(I \otimes B(u_n))(\mathbb{I} \otimes v_n) + O(h^2)v_n \\ &= (I - hB(u_n) + O(h^2))v_n. \end{aligned} \tag{3.23}$$

Next, with $\mu_2(-B(u)) \leq -\eta$, $u \in D_{\tau_0}$, we can compute

$$\begin{aligned} \|I - hB(u) + O(h^2)\|_2 &= \max\{\lambda \in \sigma(I - (h/2)\,(B(u) + B(u)^T) + O(h^2))\} \\ &\leq 1 - h\eta/4, \quad 0 < h \leq h_0, \quad u \in D_{\tau_0} \text{ uniformly.} \end{aligned} \tag{3.24}$$

Then, with $u_n \in D_{\tau_0}$ for $n \in \mathbb{N}$ and formula (3.24), the inequality

$$\|g(u_n)\|_2 \leq \prod_{j=0}^{n-1} \|I - hB(u_j) + O(h^2)\|_2 \cdot \|g(u_0)\|_2 \quad \to \quad 0 \quad \text{as } n \to \infty$$

follows. This shows our result for the projected Runge–Kutta scheme.

The proof also works for half-explicit Runge–Kutta methods, since (3.21) holds for these methods too. ∎

4. Embedded index 2 systems under discretization. In this section we give a proof of Theorem 2.1. We show the assertions (i)–(iv) of Theorem 2.1 simultaneously by applying Theorem 5 of Nipp and Stoffer [14] on the discrete Runge–Kutta dynamics of the DAE (3.1). First, let us analyze the projected Runge–Kutta methods. Using

the structure of the operator T_2 (see (3.14)) and the scheme (3.6), the projected Runge–Kutta iteration can be written in the explicit form

$$\begin{aligned} u_{n+1} &= \left[I - \frac{\partial}{\partial\lambda} f(\cdot, \lambda_p(h,u_n,\lambda_n))\gamma(h,u_n,\lambda_n)\right]^{-1} \\ &\quad \cdot (u_n + h(b^T \otimes I)\bar{f}(U(h,u_n), \Lambda(h,u_n))), \\ v_{n+1} &= g(u_n) - h(b^T \otimes I)\bar{B}(U(h,u_n))V(h,u_n), \\ \lambda_{n+1} &= R(\infty)\lambda_n + (b^T A^{-1} \otimes I)\Lambda(h,u_n). \end{aligned} \tag{4.1}$$

Moreover, the stability inequality of Lemma 3.3 implies $\gamma(h,u,\lambda) = O(h^{q+1})$. Then, using $(I - O(h^{q+1}))^{-1} = I + O(h^{q+1})$ (see, e.g., Söderlind [18, Cor. 2.3]) and neglecting the v-component, iteration (4.1) can be written in the form

$$\begin{aligned} u_{n+1} &= u_n + h[(b^T \otimes I)\bar{f}(U(h,u_n), \Lambda(h,u_n)) + h^q \hat{f}(h,u_n,\lambda_n)], \\ \lambda_{n+1} &= R(\infty)\lambda_n + (b^T A^{-1} \otimes I)\Lambda(h,u_n) \end{aligned} \tag{4.2}$$

with a smooth and bounded function $\hat{f}$.

Introducing $\eta_n := \lambda_n - \psi(u_n, g(u_n))$ and rewriting (4.2) yields

$$\begin{aligned} u_{n+1} &= u_n + h[(b^T \otimes I)\bar{f}(U(h,u_n), \Lambda(h,u_n)) + h^q \hat{f}(h,u_n,\eta_n + \psi(u_n, g(u_n)))] \\ &=: u_n + G_1(h,u_n,\eta_n), \\ \eta_{n+1} &= R(\infty)\eta_n + (b^T A^{-1} \otimes I)(\Lambda(h,u_n) - \mathbb{I} \otimes \psi(u_n, g(u_n))) \\ &\quad + \psi(u_n, g(u_n)) - \psi(u_n + G_1(h,u_n,\eta_n), g(u_n + G_1(h,u_n,\eta_n))) \\ &=: G_2(h,u_n,\eta_n). \end{aligned} \tag{4.3}$$

The functions G_1, G_2 are Lipschitzian with the constants

$$\begin{aligned} L_{G_1,u} &= O(h), \qquad L_{G_1,\eta} = O(h^{q+1}), \\ L_{G_2,u} &= O(1), \qquad L_{G_2,\eta} = |R(\infty)| + O(h^{q+1}) < 1. \end{aligned} \tag{4.4}$$

Obviously, for a fixed number $r \in \mathbb{N}$, the conditions

$$\begin{aligned} 2\sqrt{L_{G_1,\eta} L_{G_2,u}} &< 1 - L_{G_1,u} - L_{G_2,\eta}, \\ L_{G_2,\eta} + L_{G_1,\eta}\alpha &< (1 - L_{G_1,u} - L_{G_1,\eta}\alpha)^r, \end{aligned}$$

with

$$\alpha := \frac{2L_{G_2,u}}{1 - L_{G_1,u} - L_{G_2,\eta} + \sqrt{(1 - L_{G_1,u} - L_{G_2,\eta})^2 - 4L_{G_1,\eta}L_{G_2,u}}},$$

are satisfied for $h > 0$ sufficiently small. Now, Theorem 5 of Nipp and Stoffer [14] guarantees the existence of a C_b^ν-function η_h which defines the discrete invariant manifold by $\eta = \eta_h(u)$. In the (u,λ)-coordinates we obtain the following result with $\psi_h(u) := \psi(u, g(u)) + \eta_h(u)$, $u \in D_{\tau_0}$, for $0 < h \le h_0$, $h_0 > 0$, sufficiently small:

(i) The set $M_h = \{(u,\lambda) \in D_{\tau_0} \times \mathbb{R}^l \mid \lambda = \psi_h(u)\}$ is invariant for the projected Runge–Kutta map (4.2).

(ii) The manifold M_h is uniformly attractive with attractivity constant $\chi_h = |R(\infty)| + O(h^{q+1})$.

(iii) For every initial value (u_0, λ_0) with $\|\lambda_0 - \psi(u_0, g(u_0))\|$ sufficiently small, there is $(\tilde{u}_0, \tilde{\lambda}_0) \in M_h$ and $c, \hat{c} > 0$ such that the corresponding evolutions (u_n, λ_n) and $(\tilde{u}_n, \tilde{\lambda}_n)$ satisfy

$$\begin{aligned} \|u_i - \tilde{u}_i\| &\le c\chi_h^i \|\lambda_0 - \psi(u_0, g(u_0))\|, \quad i \in \mathbb{N}, \\ \|\lambda_i - \tilde{\lambda}_i\| &\le \hat{c}\chi_h^i \|\lambda_0 - \psi(u_0, g(u_0))\|, \quad i \in \mathbb{N}. \end{aligned}$$

(iv) $\|\psi(u, g(u)) - \psi_h(u)\| \le C \sup\{\|\beta(h, u, \lambda - \psi(u, g(u)))\| \mid (h, u, \lambda) \in \Gamma_{h_0,\epsilon,\tau_0}\}$ with $\Gamma_{h_0,\epsilon,\tau_0} := \{(h, u, \lambda) \in]0, h_0] \times D_{\tau_0} \times \mathbb{R}^l \mid \|\lambda - \psi(u, g(u))\| < \epsilon\}$.

Here the reader may notice that to apply Theorem 5 of Nipp and Stoffer [14] formally, we have to enlarge the domain of G_1, G_2 for $u \in \mathbb{R}^N$ as C_b^ν-maps which satisfy the Lipschitz conditions (4.4).

Reduced to the invariant manifold M_h, the u-component of a projected Runge–Kutta method reads

$$u_{n+1} = u_n + h[(b^T \otimes I)\bar{f}(U(h, u_n), \Lambda(h, u_n)) + h^q \hat{f}(h, u_n, \psi_h(u_n))]. \tag{4.5}$$

Obviously, the iteration scheme (4.5) can be regarded as a qth order one-step method applied to the initial value problem

$$\dot{u} = f(u, \psi(u, g(u))), \quad u(0) = u_0. \tag{4.6}$$

Next we estimate the distance between M_h and

$$M = \{(u, \lambda) \in D_{\tau_0} \times \mathbb{R}^l \mid \lambda = \psi(u, g(u))\}.$$

Since $\beta_n = O(h)$ holds with β from (3.17) we directly obtain $M_h - M = O(h)$.

The next step in the proof of Theorem 2.1 for the projected Runge–Kutta methods is to show $\beta_n := \beta(h, u_n, \eta_n) = O(h^q)$. Due to formula (3.17), the representation

$$\begin{aligned} \beta_n = {} & (b^T A^{-1} \otimes I)(\Lambda(h, u_n) - \bar{\psi}(U(h, u_n), \bar{g}(U(h, u_n)))) \\ & + (b^T A^{-1} \otimes I)(\bar{\psi}(U(h, u_n), \bar{g}(U(h, u_n))) - \mathbb{I} \otimes \psi(u_n, g(u_n))) \\ & + \psi(u_n, g(u_n)) - \psi(u_{n+1}, g(u_{n+1})) \end{aligned} \tag{4.7}$$

is valid. Now, we analyze the first term on the right-hand side of (4.7). With the solution flow $(\tilde{u}(t, u_0), \tilde{v}(t, u_0), \tilde{\lambda}(t, u_0))$, $u_0 \in D_{\tau_0}$, $\tilde{v}(t, u_0) = g(\tilde{u}(t, u_0))$, $\tilde{\lambda}(t, u_0) = \psi(\tilde{u}(t, u_0), \tilde{v}(t, u_0))$ of the underlying DAE (3.1), this gives

$$\begin{aligned} & (b^T A^{-1} \otimes I)(\Lambda(h, u_n) - \bar{\psi}(U(h, u_n), \bar{g}(U(h, u_n)))) \\ & \quad = O(1)[\Lambda_i(h, u_n) + \psi(\tilde{u}(c_i h, u_n), g(\tilde{u}(c_i h, u_n))) - \tilde{\lambda}(c_i h, u_n) \\ & \qquad - \psi(U_i(h, u_n), g(U_i(h, u_n))),\ i = 1, \dots, s] \\ & \quad = O(h^q), \end{aligned} \tag{4.8}$$

since

$$\begin{aligned} U_i(h, u_n) - \tilde{u}(c_i h, u_n) &= O(h^{q+1}), \quad i = 1, \dots, s, \\ \Lambda_i(h, u_n) - \tilde{\lambda}(c_i h, u_n) &= O(h^q), \qquad i = 1, \dots, s, \end{aligned} \tag{4.9}$$

hold (see, e.g., Hairer and Wanner [11, Chap. VII, Lem. 4.4]).

To examine the second term on the right-hand side of (4.7) appropriately we have to do some preparation. Let

$$\begin{aligned} \hat{u}_{n+1} &= \hat{u}_n + h(b^T \otimes I)\bar{f}(\hat{U}(h,\hat{u}_n), \hat{\Lambda}(h,\hat{u}_n)), \\ \hat{\lambda}_{n+1} &= R(\infty)\hat{\lambda}_n + (b^T A^{-1} \otimes I)\hat{\Lambda}(h,\hat{u}_n), \end{aligned} \tag{4.10}$$

with $(\hat{U}(h,\hat{u}_n), \hat{\Lambda}(h,\hat{u}_n))$ being a solution of

$$S(h,\hat{u}_n,\hat{U},\hat{\Lambda}) = \begin{pmatrix} \hat{U} - \mathbb{I}\otimes\hat{u}_n - h(A\otimes I)\bar{f}(\hat{U},\hat{\Lambda}) \\ D\bar{g}(\hat{U})\bar{f}(\hat{U},\hat{\Lambda}) + \bar{B}(\hat{U})\bar{g}(\hat{U}) \end{pmatrix} = 0, \tag{4.11}$$

stand for the Runge–Kutta method with tableau (2.3) applied to the DAE

$$\begin{aligned} \dot{u} &= f(u,\lambda), \\ 0 &= Dg(u)f(u,\lambda) + B(u)g(u). \end{aligned} \tag{4.12}$$

Equation (4.12) is the index 1 problem with eliminated v-variables corresponding to (3.1). From (4.11) and the fact that the continuous solution $(\tilde{u}(t,u), \tilde{\lambda}(t,u))$, $u \in D_{\tau_0}$, satisfies (4.12), we can conclude with (4.9) that

$$\begin{aligned} S(h,\hat{u}_n,U(h,\hat{u}_n),\Lambda(h,\hat{u}_n)) &= (0, D\bar{g}(U(h,\hat{u}_n))\bar{f}(U(h,\hat{u}_n),\Lambda(h,\hat{u}_n)) \\ &\quad - \bar{B}(U(h,\hat{u}_n))\bar{g}(U(h,\hat{u}_n))) \\ &= (0, O(1)[(U_i(h,\hat{u}_n) - \tilde{u}(c_i h,\hat{u}_n),\ i=1,\dots,s), \\ &\quad (\Lambda_i(h,\hat{u}_n) - \tilde{\lambda}(c_i h,\hat{u}_n),\ i=1,\dots,s)]) = (0, O(h^q)). \end{aligned}$$

Here, $(U,\Lambda) = (U(h,\hat{u}_n),\Lambda(h,\hat{u}_n))$ denotes the solution of (3.12). Moreover, we have

$$\frac{\partial}{\partial(\hat{U},\hat{\Lambda})} S(h,\hat{u}_n,\hat{U}(h,\hat{u}_n),\hat{\Lambda}(h,\hat{u}_n)) = \begin{pmatrix} I & 0 \\ O(1) & I\otimes Dg(\hat{u}_n)\frac{\partial f}{\partial\lambda}(\hat{\gamma}(\hat{u}_n)) \end{pmatrix} + O(h).$$

Thus, for $h>0$ sufficiently small, $\frac{\partial}{\partial(\hat{U},\hat{\Lambda})} S(h,\hat{u}_n,\hat{U}(h,\hat{u}_n),\hat{\Lambda}(h,\hat{u}_n))$ is invertible and the stability inequality of Lemma 3.1 shows

$$(U-\hat{U},\Lambda-\hat{\Lambda})(h,\hat{u}_n) = O(h^q). \tag{4.13}$$

We insert (4.13) into formula (4.10) and obtain with $u_n = \hat{u}_n$ and (4.2) the relation

$$\hat{u}_{n+1} = u_{n+1} + O(h^{q+1}). \tag{4.14}$$

With the formulae (4.13) and (4.14) we manipulate the second term on the right-hand side of (4.7) as follows:

(4.15)

$$\begin{aligned} &(b^T A^{-1}\otimes I)(\bar{\psi}(U(h,u_n),\bar{g}(U(h,u_n))) - \mathbb{I}\otimes\psi(u_n,g(u_n))) \\ &\qquad + \psi(u_n,g(u_n)) - \psi(u_{n+1},g(u_{n+1})) \\ &= (b^T A^{-1}\otimes I)(\bar{\psi}(\hat{U}(h,\hat{u}_n),\bar{g}(\hat{U}(h,\hat{u}_n))) - \mathbb{I}\otimes\psi(\hat{u}_n,g(\hat{u}_n))) \\ &\qquad + \psi(\hat{u}_n,g(\hat{u}_n)) - \psi(\hat{u}_{n+1},g(\hat{u}_{n+1})) + O(h^q). \end{aligned}$$

Next, we embed the index 1 problem (4.12) into the singular perturbed problem

$$\begin{aligned} \dot{u} &= f(u,\lambda), \\ \epsilon\dot{\lambda} &= Dg(u)f(u,\lambda) + B(u)g(u). \end{aligned} \tag{4.16}$$

We define $H(u,\lambda) := Dg(u)f(u,\lambda) + B(u)g(u)$ and consider the discrete scheme corresponding to (4.16). This reads

$$\hat{u}^\epsilon_{n+1} = \hat{u}^\epsilon_n + h(b^T \otimes I)\bar{f}(\hat{U}^\epsilon(h,\hat{u}_n), \hat{\Lambda}^\epsilon(h,\hat{u}_n)),$$
$$\hat{\lambda}^\epsilon_{n+1} = \hat{\lambda}^\epsilon_n + \frac{h}{\epsilon}(b^T \otimes I)\bar{H}(\hat{U}^\epsilon(h,\hat{u}_n), \hat{\Lambda}^\epsilon(h,\hat{u}_n)),$$

with $(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \hat{\Lambda}^\epsilon(h,\hat{u}^\epsilon_n))$ being the solution of

$$\hat{U}^\epsilon - \mathbb{I} \otimes \hat{u}^\epsilon_n = h(A \otimes I)\bar{f}(\hat{U}^\epsilon, \hat{\Lambda}^\epsilon),$$
$$\epsilon(\hat{\Lambda}^\epsilon - \mathbb{I} \otimes \hat{\lambda}^\epsilon_n) = h(A \otimes I)\bar{H}(\hat{U}^\epsilon, \hat{\Lambda}^\epsilon.$$

Following Nipp and Stoffer [15, formulae (6) and (7)], we define the functions

$$\begin{aligned} E(\hat{u}^\epsilon_n, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) &:= \bar{\psi}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{g}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n))) - \mathbb{I} \otimes \psi(\hat{u}^\epsilon_n, g(\hat{u}^\epsilon_n)) \\ &\quad - \frac{h}{\epsilon}(A \otimes I)\bar{H}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{\psi}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{g}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n)))), \\ e(\hat{u}^\epsilon_n, \hat{u}^\epsilon_{n+1}, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) &:= \psi(\hat{u}^\epsilon_n, g(\hat{u}^\epsilon_n)) - \psi(\hat{u}^\epsilon_{n+1}, g(\hat{u}^\epsilon_{n+1})) \\ &\quad - \frac{h}{\epsilon}(b^T \otimes I)\bar{H}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{\psi}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{g}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n)))). \end{aligned}$$

Then with $E(\hat{u}^\epsilon_n, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) = O(h^{q+1}) + O(\epsilon)$, $e(\hat{u}^\epsilon_n, \hat{u}^\epsilon_{n+1}, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) = O(h^{q+1}) + O(\epsilon)$ (see formula (18) in Nipp and Stoffer [15]), the relation

$$\begin{aligned} &(b^T A^{-1} \otimes I)E(\hat{u}^\epsilon_n, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) - e(\hat{u}^\epsilon_n, \hat{u}^\epsilon_{n+1}, \hat{U}^\epsilon(h,\hat{u}^\epsilon_n)) \\ &\quad = (b^T A^{-1} \otimes I) \cdot (\bar{\psi}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n), \bar{g}(\hat{U}^\epsilon(h,\hat{u}^\epsilon_n))) - \mathbb{I} \otimes \psi(\hat{u}^\epsilon_n, g(\hat{u}^\epsilon_n))) \\ &\qquad + \psi(\hat{u}^\epsilon_n, g(\hat{u}^\epsilon_n)) - \psi(\hat{u}^\epsilon_{n+1}, g(\hat{u}^\epsilon_{n+1})) = O(h^{q+1}) + O(\epsilon) \end{aligned} \tag{4.17}$$

follows. Letting $\epsilon \to 0$ in (4.17), noticing $\hat{U}^0 = \hat{U}$, $\hat{\Lambda}^0 = \hat{\Lambda}$, $\hat{u}^0_n = \hat{u}_n$, $\hat{\lambda}^0_n = \hat{\lambda}$, and inserting (4.17) with $\epsilon = 0$ into (4.15) and finally in (4.7) gives $\beta_n = O(h^q)$. This shows $\|\psi(u,g(u)) - \psi_h(u)\| \le Ch^q$, $u \in D_{\tau_0}$.

Now, we complete the proof of Theorem 2.1. Since the projected Runge–Kutta methods applied to the original DAE (2.1) can be regarded as the same method applied to the embedded DAE (3.1) with $v(0) = v_0 = 0$, we can draw back the results derived for (3.1) to (2.1). We restrict (4.5) to the invariant set S and define $\psi_{0,h} := \psi_h|_S$ by $\psi_{0,h}(u) = \psi_h(u)$, $u \in S$, as well as $M_{0,h} = \{(u,\lambda) \in D_{\tau_0} \times \mathbb{R}^l \mid g(u) = 0,\ \lambda = \psi_{0,h}(u)\}$.

In the case of half-explicit Runge–Kutta methods, we obtain the iteration scheme

$$\begin{aligned} u_{n+1} &= u_n + h(b^T \otimes I)\bar{f}((u_n, \tilde{U}_1(h,u_n), \dots, \tilde{U}_{s-1}(h,u_n)), \Lambda(h,u_n)), \\ \lambda_{n+1} &= \Lambda_s(h,u_n) \end{aligned} \tag{4.18}$$

from (3.19) and (2.10). Again, introducing $\eta_n = \lambda_n - \psi(u_n, g(u_n))$ yields

$$\begin{aligned} u_{n+1} &= u_n + h(b^T \otimes I)\bar{f}((u_n, \tilde{U}_1(h,u_n), \dots, \tilde{U}_{s-1}(h,u_n)), \Lambda(h,u_n)) \\ &= u_n + \tilde{G}_1(h,u_n,\eta_n), \\ \eta_{n+1} &= \Lambda_s(h,u_n) - \psi(u_n + \tilde{G}_1(h,u_n,\eta_n), g(u_n + \tilde{G}_1(h,u_n,\eta_n))) = \tilde{G}_2(h,u_n,\eta_n). \end{aligned}$$

Then we can adapt the proof of the projected Runge–Kutta method to the half-explicit Runge–Kutta scheme. With (B3′) and $c_s = 1$ we can estimate

$$\begin{aligned} \beta_n &= \Lambda_s(h,u_n) - \psi(u_{n+1}, g(u_{n+1})) \\ &= \Lambda_s(h,u_n) - \tilde{\lambda}(h,u_n) + \psi(\tilde{u}(h,u_n), g(\tilde{u}(h,u_n))) - \psi(u_{n+1}, g(u_{n+1})) = O(h^r). \end{aligned}$$

Finally, the attraction constant $\chi_h = 0$ follows from the fact that $L_{\tilde{G}_{1,\eta}} = L_{\tilde{G}_{2,\eta}} = 0$. This finishes the proof of Theorem 2.1.

5. Discretization near periodic orbits. In this section we prove Theorem 2.2. Let $(\bar{u}(t,u_0),\psi_0(\bar{u}(t,u_0)))$ be a hyperbolic T-periodic orbit of the DAE (2.1). By definition this means that $\bar{u}(t,u_0)$ is a hyperbolic T-periodic orbit of the corresponding index 0 equation (2.2) with phase space $S=\{u\in D_{\tau_0}\mid g(u)=0\}$. For fixed t we have the linearized flow

$$\frac{\partial}{\partial u}\bar{u}(t,u_0):N(Dg(u_0))\to N(Dg(\bar{u}(t,u_0)))$$

of (2.1), and $\frac{\partial}{\partial u}\bar{u}(t,u_0)v$, $v\in N(Dg(u_0))$, can be computed via solving

$$\begin{aligned}\dot{z}&=\left[\frac{\partial f}{\partial u}(\Gamma(t,u_0))+\frac{\partial f}{\partial\lambda}(\Gamma(t,u_0))D\psi_0(\bar{u}(t,u_0))\right]z,\\ z(0)&=v\in N(Dg(u_0)).\end{aligned}$$

Here $\Gamma(t,u_0)$ stands for $(\bar{u}(t,u_0),\psi_0(\bar{u}(t,u_0)))$. The reader may recall that the image space of $\frac{\partial}{\partial u}\bar{u}(t,u_0)$ is $N(Dg(\bar{u}(t,u_0)))$ since $g(\bar{u}(t,u_0))=0$, $t\in\mathbb{R}$.

For $t=T$ we obtain $\frac{\partial}{\partial u}\bar{u}(T,u_0):N(Dg(u_0))\to N(Dg(u_0))$. The hyperbolicity says that the linearized T-flow has the simple eigenvalue 1 and all other eigenvalues have absolute value different from 1.

Now, let $V\in\mathbb{R}^{N,N-l}$ denote a basis of $N(Dg(u_0))$. Then, with the monodromy matrix $X(T)=(X^1(T),\dots,X^{N-l}(T))\in\mathbb{R}^{N-l,N-l}$ in the basis $V=(v_1,\dots,v_{N-l})$, we obtain

$$\frac{\partial}{\partial u}\bar{u}(T,u_0)v_i=VX^i(T),\quad i=1,\dots,N-l. \tag{5.1}$$

Here, $X^i(T)$, v_i, $i=1,\dots,N-l$, denote the ith column of $X(T)$, V, respectively. Our first goal in this section is to show that $\bar{u}(t,u_0)$ is a hyperbolic periodic orbit of (3.5). This is the content of the following lemma.

LEMMA 5.1. *Let* (A1)–(A3) *hold and let* $(\bar{u}(t,u_0),\psi_0(\bar{u}(t,u_0)))$, $\bar{u}(t,u_0)=\bar{u}(t+T,u_0)$ *denote a hyperbolic* T*-periodic orbit of the DAE* (2.1). *Then* $\bar{u}(t,u_0)$ *is a hyperbolic* T*-periodic orbit of the ODE* (3.5).

Proof. Let $\tilde{u}(t,u)$ denote the solution flow of (3.5) and let $\bar{u}(t,u)$ stand for the solution flow of (2.2). We know $\tilde{u}(t,u)=\bar{u}(t,u)$ for $u\in S$. Obviously, $\bar{u}(t,u_0)$ is a T-periodic orbit for (3.5). It remains to show the hyperbolicity.

Linearizing (3.5) at the T-periodic orbit $\bar{u}(t,u_0)$ we obtain that $\frac{\partial}{\partial u}\tilde{u}(t,u_0)$ solves the variational equation

$$\dot{Y}(t)=Dk(\bar{u}(t,u_0))Y(t),\quad Y(0)=I.$$

We split the whole space according to $\mathbb{R}^N=N(Dg(u_0))\oplus R(\frac{\partial f}{\partial\lambda}(u_0,\psi_0(u_0)))=:R(V)\oplus R(W)$. Here the reader may recall that the split of $\mathbb{R}^N$ is possible due to the index 2 condition (A3). We analyze the behavior of $Y(T)$ for initial values in V and W separately. We claim

$$\frac{\partial}{\partial u}\tilde{u}(T,u_0)V\alpha=\frac{\partial}{\partial u}\bar{u}(T,u_0)V\alpha,\quad \alpha\in\mathbb{R}^{N-l}, \tag{5.2}$$

for initial values in $R(V)$. In what follows, we first assume (5.2) and finish the proof. Combining the formulae (5.1), (5.2) yields $Y(T)V=VX(T)$ with the monodromy matrices $X(T)$, $Y(T)$ of the periodic orbits $\bar{u}(t,u_0)$ and $\tilde{u}(t,u_0)$.

Now, let $Z(t) := Dg(\bar{u}(t,u_0))Y(t)\frac{\partial f}{\partial \lambda}(\Gamma(0,u_0))S(u_0)^{-1}$, $S(u) := Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi_0(u))$. A straightforward computation shows that Z is the fundamental matrix of

$$\dot{w} = -B(\bar{u}(t,u_0))w. \tag{5.3}$$

To analyze the behavior of $Y(T)$ for initial values in $R(W)$ we make the ansatz

$$Y(T)W = V\Lambda_{12}(T) + W\Lambda_{22}(T)$$

with matrices $\Lambda_{12}(T) \in \mathbb{R}^{N-l,l}$ and $\Lambda_{22}(T) \in \mathbb{R}^{l,l}$. Using $Dg(u_0)V \equiv 0$ we can calculate

$$Y(T)(V,W) = (V,W)\cdot\begin{pmatrix} X(T) & \Lambda_{12}(T) \\ 0 & S(u_0)^{-1}Z(T)S(u_0)\end{pmatrix}. \tag{5.4}$$

Finally, with $\mu_2(-B(u)) \le -\eta$, $\eta > 0$, we obtain

$$\|w(t)\|_2 = \|Z(t)w(0)\|_2 \le \exp(-\eta t)\|w(0)\|_2$$

for the solution of (5.3) (see Dekker and Verwer [6, Thm. 1.5.2]). Hence, $\|Z(t)\|_2 \le \exp(-\eta t)$ and $\rho(S(u_0)^{-1}Z(T)S(u_0)) = \rho(Z(T)) \le \exp(-\eta T)$ follows. Together with formula (5.4) this shows the eigenvalue condition for the monodromy matrix $Y(T)$ of the periodic orbit $\bar{u}(t,u_0)$ of (3.5).

It remains to show formula (5.2). To this purpose we need some additional relations, which we now provide. First, a straightforward calculation with $\hat{\gamma}(u) = (u,\psi(u,g(u)))$ shows

$$Dk(u) = \frac{\partial f}{\partial u}(\hat{\gamma}(u)) + \frac{\partial f}{\partial \lambda}(\hat{\gamma}(u))\left(\frac{\partial \psi}{\partial u}(u,g(u)) + \frac{\partial \psi}{\partial v}(u,g(u))Dg(u)\right). \tag{5.5}$$

Moreover, implicit differentiation of relation $Dg(u)f(u,\psi(u,v)) + B(u)v = 0$ with respect to u and v yields

$$\begin{aligned}\frac{\partial \psi}{\partial u}(u,v) = &-\left(Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi(u,v))\right)^{-1}\left[Dg(u)\frac{\partial f}{\partial u}(u,\psi(u,v)) + DB(u)v\right.\\ &\left.+ D^2g(u)f(u,\psi(u,v))\right],\\ \frac{\partial \psi}{\partial v}(u,v) = &-\left(Dg(u)\frac{\partial f}{\partial \lambda}(u,\psi(u,v))\right)^{-1}B(u).\end{aligned} \tag{5.6}$$

Thus, with $v = g(u)$ and the projector $Q(u) := \frac{\partial f}{\partial \lambda}(\hat{\gamma}(u))(Dg(u)\frac{\partial f}{\partial \lambda}(\hat{\gamma}(u)))^{-1}Dg(u)$ onto $R(\frac{\partial f}{\partial \lambda}(\hat{\gamma}(u)))$ we insert (5.6) into (5.5) and obtain

$$\begin{aligned}Dk(u) = (I - Q(u))\frac{\partial f}{\partial u}(\hat{\gamma}(u)) - \frac{\partial f}{\partial \lambda}(\hat{\gamma}(u))\left(Dg(u)\frac{\partial f}{\partial \lambda}(\hat{\gamma}(u))\right)^{-1}\\ \cdot [B(u)Dg(u) + DB(u)g(u) + D^2g(u)k(u)].\end{aligned}$$

Using $(I - Q(u))\frac{\partial f}{\partial \lambda}(\hat{\gamma}(u)) = 0$, we can compute

$$(I - Q(u))Dk(u) = (I - Q(u))\frac{\partial f}{\partial u}(\hat{\gamma}(u)). \tag{5.7}$$

To link the vectorfields $k(u) = f(u, \psi(u, g(u)))$ and $f(u, \psi_0(u))$, we differentiate the relation $Dg(u)f(u, \psi_0(u)) = 0$ (compare with (A2)) and obtain

$$D\psi_0(u) = -\left(Dg(u)\frac{\partial f}{\partial \lambda}(\hat{\gamma}(u))\right)^{-1}\left[Dg(u)\frac{\partial f}{\partial u}(\hat{\gamma}(u)) + D^2g(u)f(u, \psi_0(u))\right]. \tag{5.8}$$

Finally, we need

$$Dg(\tilde{u}(t, u_0))\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha = 0. \tag{5.9}$$

Relation (5.9) follows from the fact that $\tilde{w}(t) := Dg(\tilde{u}(t, u_0))\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha$ solves (5.3) with initial condition $\tilde{w}(0) = 0$.

Then, with (5.7)–(5.9) and $S(u) = Dg(u)\frac{\partial f}{\partial \lambda}(u, \psi_0(u))$, we can compute

$$\begin{aligned}
\frac{\partial^2}{\partial t \partial u}\tilde{u}(t, u_0)V\alpha &= Dk(\tilde{u}(t, u_0))\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha \\
&= \Bigg[(I - Q(\tilde{u}(t, u_0)))\frac{\partial f}{\partial u}(\Gamma(t, u_0)) \\
&\quad + Q(\tilde{u}(t, u_0))Dk(\tilde{u}(t, u_0))\Bigg]\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha \\
&= \Bigg[\frac{\partial f}{\partial u}(\Gamma(t, u_0)) - \frac{\partial f}{\partial \lambda}(\Gamma(t, u_0))S(\tilde{u}(t, u_0))^{-1}\Bigg[Dg(\tilde{u}(t, u_0)) \\
&\quad \cdot \frac{\partial f}{\partial u}(\Gamma(t, u_0)) + D^2g(\tilde{u}(t, u_0))k(\tilde{u}(t, u_0))\Bigg]\Bigg]\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha \\
&= \left[\frac{\partial f}{\partial u}(\Gamma(t, u_0)) + \frac{\partial f}{\partial \lambda}(\Gamma(t, u_0))D\psi_0(\bar{u}(t, u_0))\right]\frac{\partial}{\partial u}\tilde{u}(t, u_0)V\alpha,
\end{aligned} \tag{5.10}$$

and (5.2) is shown. This finishes the proof. □

In the remainder of this section we give a proof of Theorem 2.2. By virtue of Lemma 5.1, $\bar{u}(t, u_0)$ is a hyperbolic T-periodic orbit of (4.6) and, in section 4, we have seen that (4.5) is a smooth qth order one-step method applied to (4.6). Thus, Beyn [2, Thm. 1] is applicable and ensures for sufficiently small step size h the existence of an invariant curve $\bar{u}^h(\mathbb{R})$, $\bar{u}^h(t) = \bar{u}^h(t + T)$, for the one-step method

$$\begin{aligned}
u_{n+1} &= u_n + h(b^T \otimes I)\bar{f}(U(h, u_n), \Lambda(h, u_n)) + h^{q+1}\hat{f}(h, u_n, \psi_h(u_n)) \\
&=: G_h(u_n), \quad u_0 \in D_{\tau_0}
\end{aligned} \tag{5.11}$$

(compare with (4.5)) having properties

$$\begin{aligned}
\bar{u}^h(t + h + O(h^{q+1})) &= G_h(\bar{u}^h(t)), \\
\max\{\|\bar{u}(t, u_0) - \bar{u}^h(t)\| \mid t \in \mathbb{R}\} &\le Ch^q.
\end{aligned} \tag{5.12}$$

The last step in our proof is to draw back the results (5.12) to S. Obviously, we have $g(\bar{u}^h(\mathbb{R})) = 0$. This is a consequence of the fact that $\bar{u}^h(\mathbb{R})$ is an invariant set and every invariant set is located in the maximal invariant set S.

On the phase space S the iteration scheme (5.11) coincides with the u-component of the projected Runge–Kutta method (2.4)–(2.6) applied to $\dot{u} = f(u, \lambda)$, $g(u) = 0$. Thus, the discrete iteration scheme (2.4)–(2.6) possesses an invariant curve which is $O(h^q)$ close to the periodic orbit.

An analogous argument works for the half-explicit Runge–Kutta methods too and ensures the existence of an $O(h^r)$ invariant curve close to the periodic orbit.

REFERENCES

[1] U. M. Ascher and L. R. Petzold, *Projected implicit Runge–Kutta methods for differential-algebraic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1097–1120.

[2] W.-J. Beyn, *On invariant closed curves for one-step methods*, Numer. Math., 51 (1987), pp. 103–122.

[3] W.-J. Beyn, *On the numerical approximation of phase portraits near stationary points*, SIAM J. Numer. Anal., 24 (1987), pp. 1095–1113.

[4] W.-J. Beyn and J. Schropp, *Runge-Kutta discretizations of singularly perturbed gradient equations*, BIT, 40 (2000), pp. 415–433.

[5] V. Brasey and E. Hairer, *Half-explicit Runge–Kutta methods for differential-algebraic systems of index* 2, SIAM J. Numer. Anal., 30 (1993), pp. 538–552.

[6] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, CWI Monographs 2, North-Holland, Amsterdam, 1984.

[7] B. Garay, *Discretization and some qualitative properties of ordinary differential equations about equilibria*, Acta Math. Univ. Comenian. (N.S.), 62 (1993), pp. 249–275.

[8] C. W. Gear, G. K. Gupta, and B. Leimkuhler, *Automatic integration of Euler-Lagrange equations with constraints*, J. Comput. Math., 12/13 (1985), pp. 77–90.

[9] E. Hairer, Ch. Lubich, and M. Roche, *Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations*, BIT, 28 (1988), pp. 678–700.

[10] E. Hairer, Ch. Lubich, and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Lecture Notes in Math. 1409, Springer-Verlag, Berlin, 1989.

[11] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, 2nd ed., Springer-Verlag, Berlin, 1996.

[12] P. E. Kloeden and J. Lorenz, *Stable attracting sets in dynamical systems and in their one-step discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 986–995.

[13] K. Nipp, *Numerical Integration of Differential Algebraic Systems and Invariant Manifolds*, SAM Research Report 99-12, ETH, Zürich, Switzerland, 1999.

[14] K. Nipp and D. Stoffer, *Attractive Invariant Manifolds for Maps*, SAM Research Report 92-11, ETH, Zürich, Switzerland, 1992.

[15] K. Nipp and D. Stoffer, *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type - Part* I: *RK-methods*, Numer. Math., 70 (1995), pp. 245–257.

[16] J. Schropp, *A dynamical systems approach to constrained minimization*, Numer. Funct. Anal. Optim., 21 (2000), pp. 537–551.

[17] J. Schropp, *Behavior of Runge-Kutta discretizations near equilibria of index* 2 *differential algebraic systems*, Appl. Numer. Math., to appear.

[18] G. Söderlind, *Bounds on nonlinear operators in finite-dimensional Banach spaces*, Numer. Math., 50 (1984), pp. 27–44.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 891–913

UNIFORM STABILITY OF A FINITE DIFFERENCE SCHEME FOR TRANSPORT EQUATIONS IN DIFFUSIVE REGIMES*

A. KLAR† AND A. UNTERREITER‡

Abstract. An asymptotic preserving numerical scheme (with respect to diffusion scalings) for a linear transport equation is investigated. The scheme is adopted from a class of schemes developed in [S. Jin, L. Pareschi, and G. Toscani, *SIAM J. Numer. Anal.*, 38 (2000), pp. 913–936] and [A. Klar, *SIAM J. Numer. Anal.*, 35 (1998), pp. 1073–1094]. Stability is proven uniformly in the mean free path under a CFL-type condition turning into a parabolic CFL condition in the diffusion limit.

Key words. transport equations, relaxation methods, stability analysis

AMS subject classifications. 82C70, 65M06, 35B25

PII. S0036142900375700

1. Introduction.

Transport equations and kinetic equations are used for a variety of applications, for example, to simulate radiative heat transfer processes or rarefied gas flows. Near to the continuum regimes the equations are approximated by macroscopic equations like diffusion equations or fluid dynamic equations. In recent years, asymptotic preserving schemes for kinetic equations and transport equations have gained considerable attention in the literature. These schemes are used to treat singularly perturbed transport equations in situations with small mean free paths, i.e., in the above-mentioned macroscopic limits. Schemes for in-stationary transport equations in the diffusion limit can be found, for example, in [6, 7, 8, 12] and references therein. Schemes for other types of transport equations with diffusive macroscopic limits have been developed in [5, 4, 1, 9, 11, 10].

Concerning the numerical analysis of these schemes, proofs of uniform consistency with respect to a small mean free path ϵ can be found in [7, 1, 8]. Furthermore, using the homogenization theory for transport equations, a proof of uniform convergence (as $\epsilon \to 0$) for time-continuous equations discretized spatially and in velocity is given in [3, 2].

The aim of our paper is to prove a uniform stability result for semidiscrete (time- and space-discrete) numerical schemes for transport equations as developed in [7, 8]. (Numerical investigations of these schemes and proofs of uniform consistency can be found in [7, 8].)

A time-and-space discretization is performed and linear stability is proved uniformly in ϵ using a careful direct analysis of the iterative scheme. First, the problem is tackled by a von Neumann analysis of the system in a continuous function space setting. This gives explicit and accurate estimates.

Under an ϵ-dependent CFL-type restriction, the iterations are proved to be uniformly bounded. As ϵ tends to 0, the CFL-type condition turns into a parabolic CFL condition which can be satisfied by ϵ-independent grids.

*Received by the editors July 26, 2000; accepted for publication (in revised form) January 30, 2002; published electronically August 1, 2002. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grants KL 1105/7 and SFB 568.

http://www.siam.org/journals/sinum/40-3/37570.html

†Fachbereich Mathematik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany (klar@mathematik.tu-darmstadt.de).

‡Institut für Mathematik, MA 6-3 Technische Universität Berlin, D-10623 Berlin, Germany (unterreiter@math.tu-berlin.de).

For a large mean free path, the CFL condition is adapted to the transport equation.

In a further step, it is shown that the stability analysis carries over to discrete function spaces with piecewise linear finite element method (FEM) for the spatial discretization.

The paper is organized as follows. In section 2 equations and the semidiscrete scheme are introduced. Section 3 contains some definitions and the statement of the main result. In section 4 several preliminary results are established, and section 5 contains the proof of the main result. Section 6 is concerned with the recursion on a discrete function space.

2. Equations and numerical scheme. Our model problem is the one-dimensional linear transport equation with isotropic scattering,

$$\epsilon^2\,\partial_t F + \epsilon\, v\,\partial_x F = \frac{1}{2}\int_{-1}^{1} F dv - F, \tag{2.1}$$

with density $F = F(x, v, t)$, $x \in \mathbb{R}$, $v \in [-1, 1]$, and $t \in [0, \infty)$. We pass to the even-odd parity formulation by introducing, for $v > 0$, the even and odd functions

$$f(v) = \frac{1}{2}(F(v) + F(-v)),$$
$$g(v) = \frac{1}{2\epsilon}(F(v) - F(-v)).$$

This defines

$$F(v) = f(v) + \epsilon g(v), \qquad v > 0,$$
$$F(v) = f(-v) - \epsilon g(-v), \qquad v < 0,$$

such that (2.1) becomes, for $v > 0$,

$$\partial_t f + v\,\partial_x g = \frac{1}{\epsilon^2}\left([f] - f\right), \qquad [f] := \int_0^1 f dv, \tag{2.2}$$

$$\partial_t g + \frac{v}{\epsilon^2}\,\partial_x f = -\frac{1}{\epsilon^2} g. \tag{2.3}$$

Remark. Concerning the limit $\epsilon \to 0$ in (2.2), (2.3) we obtain, from a formal asymptotic expansion,

$$f = \rho = [f], \qquad g = -v\,\partial_x f,$$

where $\rho = \rho(x, t)$ fulfills the diffusion equation

$$\partial_t \rho = \frac{1}{3}\partial_{xx}\rho. \tag{2.4}$$

We describe a semidiscrete scheme which is taken from a general class of schemes developed in [8, 6]. For the time discretization we use the time step $\Delta t \in \mathbb{R}^+$. The spatial step size is Δx. The time iterations approximating $f(x, v, n\Delta t)$ and $g(x, v, n\Delta t)$ are $f^n(x, v)$ and $g^n(x, v)$ (for $n \in \mathbb{N}$ or $n = 0$), respectively.

Given f^n, g^n we calculate f^{n+1}, g^{n+1} by a fractional step scheme, where the spatial discretization is a first order discretization.

Remark. The reader is invited to consult [6] for more sophisticated approaches.

Notation. In what follows, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

ALGORITHM.

Step 1. *Approximate the solution of the system*

$$\begin{aligned} \partial_t f + v\partial_x g &= 0, \\ \partial_t g &= 0 \end{aligned}$$

by an explicit discretization; i.e., determine $f^{n+\frac{1}{2}}, g^{n+\frac{1}{2}}$ *via*

$$\left.\begin{aligned} f^{n+\frac{1}{2}} &= f^n - \Delta t\, v\, D_+ g^n \\ g^{n+\frac{1}{2}} &= g^n \end{aligned}\right\} \quad n \in \mathbb{N}_0,$$

where D_+ *is the forward difference with step size* Δx,

$$D_+ f(x) = \frac{f(x+\Delta x) - f(x)}{\Delta x}.$$

Step 2. *Approximate the solution of the system*

$$\left.\begin{aligned} \partial_t f &= \frac{1}{\epsilon^2}([f] - f) \\ \partial_t g &= \frac{1}{\epsilon^2}(-v\,\partial_x f - g) \end{aligned}\right\} \quad n \in \mathbb{N}_0$$

by a semi-implicit discretization to treat the stiffness of the equations correctly; i.e., determine f^{n+1}, g^{n+1} *from* $f^{n+\frac{1}{2}}, g^{n+\frac{1}{2}}$ *via*

$$\left.\begin{aligned} f^{n+1} &= f^{n+\frac{1}{2}} + \frac{\Delta t}{\epsilon^2}\left(\left[f^{n+\frac{1}{2}}\right] - f^{n+1}\right) \\ g^{n+1} &= g^{n+\frac{1}{2}} + \frac{\Delta t}{\epsilon^2}\left[-v\, D_- f^{n+1} - g^{n+1}\right] \end{aligned}\right\} \quad n \in \mathbb{N}_0,$$

where D_- *denotes the backward difference with step size* Δx.

We rewrite the recursion formula of Step 2 as

$$\left.\begin{aligned} f^{n+1} &= A f^{n+\frac{1}{2}} + B\left[f^{n+\frac{1}{2}}\right] \\ g^{n+1} &= A g^{n+\frac{1}{2}} - B\, v\, D_- f^{n+1} \end{aligned}\right\} \quad n \in \mathbb{N}_0, \tag{2.5}$$

with

$$\begin{aligned} A &:= \left(1 + \frac{\Delta t}{\epsilon^2}\right)^{-1}, \\ B &:= \frac{\Delta t}{\epsilon^2} A = 1 - A = \left(\frac{\epsilon^2}{\Delta t} + 1\right)^{-1}. \end{aligned}$$

For the numerical analysis, it is convenient to combine both steps into a single step,

$$\left.\begin{aligned} f^{n+1} &= A(f^n - \Delta t\, v\, D_+ g^n) + B\left[f^n - \Delta t\, v\, D_+ g^n\right] \\ g^{n+1} &= A g^n - v\, A\, B\, D_- f^{n+1} - v\, B^2\, D_-\left[f^{n+1}\right] \end{aligned}\right\} \quad n \in \mathbb{N}_0, \tag{2.6}$$

or

$$g^{n+1} = A g^n - v\, A\, B\, D_-(f^n - \Delta t\, v\, D_+ g^n) - v\, B^2\, D_-\left[f^n - \Delta t\, v\, D_+ g^n\right], \quad n \in \mathbb{N}_0.$$

We are concerned mainly with investigations of scheme (2.6). Uniform consistency of similar schemes has been considered in [8] and [7]. We will prove a *uniform (in ϵ)* stability result.

Remark. Keeping Δt fixed and considering the limit $\epsilon \to 0$ of (2.6), we have $A \to 0, B \to 1$ as $\epsilon \to 0$, and we obtain the scheme

$$f^{n+1} = \left[f^{n+1}\right] = \left[f^n - \Delta t\, v\, D_+ g^n\right],$$
$$g^{n+1} = -v\, D_- \left[f^{n+1}\right];$$

i.e., in terms of $\rho^n = [f^n]$,

$$\rho^{n+1} = \rho^n + \frac{1}{3}\Delta t\, D_+ D_- \rho^n,$$

which is a straightforward explicit discretization of the diffusion equation (2.4).

3. The main result. In this section we state a theorem on uniform stability for (2.6). The proof is settled on a von Neumann stability analysis.

The recursion scheme (2.6) involves two positive discretization parameters Δt, Δx (which enter via $D_\pm$) and the scaled mean free path $\epsilon \in (0, \infty)$. We assume that $\Delta t, \Delta x, \epsilon$ satisfy the following condition.

DEFINITION. *$\Delta t, \Delta x$, and ϵ fulfill the transport CFL condition iff*

$$\frac{\Delta t}{(\Delta x)^2}\frac{\Delta t}{\epsilon^2 + \Delta t} < \frac{1}{2}. \tag{3.1}$$

Remark. Condition (3.1) is equivalent to

$$\frac{\Delta t}{(\Delta x)^2} < \frac{\epsilon^2 + \Delta t}{2\Delta t} \tag{3.2}$$

or

$$\frac{\Delta t}{\Delta x} < \sqrt{\frac{\epsilon^2 + \Delta t}{2}}. \tag{3.3}$$

For $\epsilon^2 \ll \Delta t$, condition (3.2) reduces to a *parabolic CFL condition*,

$$\frac{\Delta t}{(\Delta x)^2} < \frac{1}{2}, \tag{3.4}$$

related to the diffusion equation and, in case $\epsilon^2 \gg \Delta t$, condition (3.3) reduces to

$$\frac{\Delta t}{\Delta x} < \frac{\epsilon}{\sqrt{2}} \tag{3.5}$$

which is, for fixed ϵ, a *hyperbolic CFL condition* related to the transport equation.

Remark. Introducing

$$\rho := \frac{\Delta t}{\epsilon^2}, \quad \text{i.e., } \Delta t = \rho\epsilon^2,$$

the transport CFL condition (3.1) holds iff

$$\exists \delta \in \mathbb{R}^+ : \quad (\Delta x)^2 = 2\rho^2 \frac{1+\delta}{1+\rho} \cdot \epsilon^2.$$

Here, $\rho \ll 1$ corresponds to the well-resolved case $\Delta t \ll \epsilon^2$, and $1 \ll \rho$ corresponds to the underresolved case $\epsilon^2 \ll \Delta t$.

In the following, a von Neumann stability analysis of the semidiscrete scheme (2.6) will be performed.

First, we shall give the recursion (2.6) a well-defined meaning by introducing sets of functions on which the recursion operator of (2.6) acts. The recursion scheme will be viewed as an operator acting on a continuous function space. Remarks on the action of the recursion on discrete functions given by the values at the grid points are included in section 6.

We introduce the spaces

$$M := \{\phi : \mathbb{R} \times [0,1] \to \mathbb{C} : \ \phi \text{ is measurable}\},$$

$$\mathcal{L}^2(dx) := \left\{\varphi : \mathbb{R} \to \mathbb{C} : \ \varphi \text{ is measurable and } \int_{\mathbb{R}} |\varphi|^2 \, dx < \infty\right\},$$

$$\mathcal{L}^1(dv) := \left\{\varphi : [0,1] \to \mathbb{C} : \ \varphi \text{ is measurable and } \int_{[0,1]} |\varphi| \, dv < \infty\right\},$$

$$\mathcal{L}^2(d(x,v)) := \left\{\varphi \in M : \int_{\mathbb{R}\times[0,1]} |\varphi|^2 \, d(x,v) < \infty\right\}$$

equipped with the standard seminorms

$$\forall \varphi \in \mathcal{L}^2(dx): \quad \|\varphi\|_{\mathcal{L}^2(dx)} = \sqrt{\int_{\mathbb{R}} |\varphi|^2 \, dx},$$

$$\forall \varphi \in \mathcal{L}^1(dv): \quad \|\varphi\|_{\mathcal{L}^1(dv)} = \int_{[0,1]} |\varphi| \, dv,$$

$$\forall \varphi \in \mathcal{L}^2(d(x,v)): \quad \|\varphi\|_{\mathcal{L}^2(d(x,v))} = \sqrt{\int_{\mathbb{R}\times[0,1]} |\varphi|^2 \, d(x,v)}.$$

We choose the anisotropic Lebesgue space

$$\mathcal{L}^2(dx, \mathcal{L}^1(dv)) = \left\{\phi \in M : \ \left(\forall x \in \mathbb{R} : \ \phi(x,.) \in \mathcal{L}^1(dv)\right) \wedge \left(\|\phi\|_{\mathcal{L}^1(dv)} \in \mathcal{L}^2(dx)\right)\right\}$$

equipped with the canonical seminorm

$$\|\phi\|_{\mathcal{L}^2(dx,\mathcal{L}^1(dv))} = \sqrt{\int_{\mathbb{R}} \|\phi(x,.)\|^2_{\mathcal{L}^1(dv)} \, dx}, \qquad \phi \in \mathcal{L}^2(dx, \mathcal{L}^1(dv)),$$

as the domain of the integration operator [.].

Remark. Obviously, $\mathcal{L}^2(d(x,v)) \subseteq \mathcal{L}^2(dx, \mathcal{L}^1(dv))$ and

$$\|\varphi\|_{\mathcal{L}^2(dx,\mathcal{L}^1(dv))} \le \|\varphi\|_{\mathcal{L}^2(d(x,v))}, \qquad \phi \in \mathcal{L}^2(d(x,v)).$$

Then we have, in operator notation,

$$[.] : \mathcal{L}^2(dx, \mathcal{L}^1(dv)) \to \mathcal{L}^2(dx), \quad [\phi](x) = \int_{[0,1]} \phi(x,.)\, dv, \quad x \in \mathbb{R}.$$

Obviously, [.] is linear and

$$|[\phi]|(x) \le \|\phi(x,.)\|_{\mathcal{L}^1(dv)}, \qquad (\phi, x) \in \mathcal{L}^2(dx, \mathcal{L}^1(dv)) \times \mathbb{R}.$$

The Fourier transform of $\phi \in \mathcal{L}^2(dx, \mathcal{L}^1(dv))$ with respect to x is

$$\hat{\phi}(\xi, v) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \phi(x,v)\, \exp(-ix\xi)\, dx, \qquad (\xi, v) \in \mathbb{R} \times [0,1].$$

Certainly,

$$\hat{\phi} \in \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)), \qquad \phi \in \mathcal{L}^2(dx, \mathcal{L}^1(dv)).$$

For later reference we introduce the Bessel potential space

$$\mathcal{H}^1\left(dx, \mathcal{L}^1(dv)\right) = \left\{ \phi \in \mathcal{L}^2(d(x,v)) : \int_{\mathbb{R}} |\xi|^2 \left\| \hat{\phi}(\xi,.) \right\|^2_{\mathcal{L}^1(dv)} d\xi < \infty \right\}$$

with canonical seminorm

$$\|\phi\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} = \sqrt{\|\phi\|^2_{\mathcal{L}^2(d(x,v))} + \int_{\mathbb{R}} |\xi|^2 \left\| \hat{\phi}(\xi,.) \right\|^2_{\mathcal{L}^1(dv)} d\xi}.$$

Remark. Obviously, $\mathcal{H}^1\left(dx, \mathcal{L}^1(dv)\right) \subset \mathcal{L}^2(d(x,v)) \subset \mathcal{L}^2(d\xi, \mathcal{L}^1(dv))$.

Remark. If $\phi \in \mathcal{L}^2(d(x,v))$ and if $\phi(.,v) \in H^1(\mathbb{R})$ for almost all $v \in [0,1]$, with square integrable (with respect to v) seminorm $\|\phi(.,v)\|_{H^1(\mathbb{R})}$, then by the theory of Sobolev spaces, $\phi \in \mathcal{H}^1\left(dx, \mathcal{L}^1(dv)\right)$.

Furthermore, let

$$\mathcal{L}^\infty(dv, \mathcal{L}^2(dx)) = \left\{ \varphi \in M : \sup_{v \in [0,1]} \|\varphi(.,v)\|_{\mathcal{L}^2(dx)} < \infty \right\}$$

equipped with the canonical seminorm

$$\|\varphi\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} = \sup_{v \in [0,1]} \|\varphi(.,v)\|_{\mathcal{L}^2(dx)}, \qquad \varphi \in \mathcal{L}^\infty(dv, \mathcal{L}^2(dx)).$$

Applying the Fourier transform (with respect to x) on (2.6) gives

$$\left.\begin{aligned} \hat{f}^{n+1} &= A\left(\hat{f}^n + v\,\alpha\,\hat{g}^n\right) + B\left[\hat{f}^n + \alpha\, v\, \hat{g}^n\right] \\ \hat{g}^{n+1} &= A\left(\hat{g}^n + v\,\beta\,\hat{f}^n - \Theta\, v^2\, \hat{g}^n\right) + B\, v\, \beta\left[\hat{f}^n + \alpha\, v\, \hat{g}^n\right] \end{aligned}\right\} \quad n \in \mathbb{N}_0, \tag{3.6}$$

where

$$\alpha = \alpha(\xi, \Delta t, \Delta x) = \frac{\Delta t}{\Delta x}\left(1 - \mathrm{e}^{i\xi\Delta x}\right) = \xi \Delta t H(\xi \Delta x), \tag{3.7}$$

$$\beta = \beta(\xi, \Delta t, \Delta x, \epsilon) = \frac{B}{\Delta x}\left(\mathrm{e}^{-i\xi\Delta x} - 1\right) = -\frac{\xi \Delta t}{\epsilon^2 + \Delta t} \overline{H}(\xi \Delta x), \tag{3.8}$$

$$\Theta = \Theta(\xi, \Delta t, \Delta x, \epsilon) = -\alpha\,\beta = \frac{\Delta t}{(\Delta x)^2}\frac{\Delta t}{\epsilon^2 + \Delta t}\,(2 - 2\cos(\xi \Delta x)), \tag{3.9}$$

and the holomorphic function H,

$$H : \mathbb{C} \to \mathbb{C}, \quad H(z) = \begin{cases} \dfrac{1 - \exp(iz)}{z}, & z \neq 0, \\ -i, & z = 0, \end{cases}$$

is bounded on $\mathbb{R}$ with $|H|(\sigma) < 1$ for all $\sigma \in \mathbb{R}$ with $\sigma \neq 0$, and $|H|(0) = 1$. Furthermore, $|\sigma H(\sigma)| \leq 2$ for all $\sigma \in \mathbb{R}$.

Remark. We have

$$0 = \inf_{\xi \in \mathbb{R}} \Theta(\xi, \Delta t, \Delta x, \epsilon) < \sup_{\xi \in \mathbb{R}} \Theta(\xi, \Delta t, \Delta x, \epsilon) = \frac{4\Delta t}{(\Delta x)^2}\frac{\Delta t}{\epsilon^2 + \Delta t}$$

for all positive $\Delta t, \Delta x, \epsilon$, which highlights the importance of the value of Θ.

Remark. The transport CFL condition is equivalent to

$$\sup_{\xi \in \mathbb{R}} \Theta(\xi, \Delta t, \Delta x, \epsilon) < 2.$$

Remark. Concerning α, β we have, on the one hand, estimates which are independent of ξ and of ϵ but depend on the grid sizes Δt and Δx,

$$|\alpha|(\xi, \Delta t, \Delta x) \leq \frac{2\Delta t}{\Delta x}, \qquad |\beta|(\xi, \Delta t, \Delta x) \leq \frac{2}{\Delta x}; \tag{3.10}$$

on the other hand, there are estimates which are "almost linear" in ξ, which depend in an acceptable way on Δt, and which are independent of ϵ,

$$|\alpha|(\xi, \Delta t, \Delta x) \leq \Delta t\,|\xi|, \qquad |\beta|(\xi, \Delta t, \Delta x) \leq |\xi|. \tag{3.11}$$

Remark. Obviously, (3.11) is suitable for uniform stability analysis, however, due to the "linear" dependence of the estimates on $|\xi|$ for the price of obtaining estimates involving different kinds of seminorms.

It is convenient to introduce for $(f, g) \in \mathcal{L}^2(dx, \mathcal{L}^1(dv)) \times \mathcal{L}^2(dx, \mathcal{L}^1(dv))$ the notation

$$\mathfrak{f} = (f, g), \qquad \hat{\mathfrak{f}} = (\hat{f}, \hat{g}).$$

We rewrite recursion (2.6) as

$$\mathfrak{f}^{n+1} = S\mathfrak{f}^n \tag{3.12}$$

and the Fourier transformed recursion (3.6) as

$$\hat{\mathfrak{f}}^{n+1} = (AT + BT_0)\,\hat{\mathfrak{f}}^n \tag{3.13}$$

with $A, B \in (0, 1), A + B = 1$ as above and $n \in \mathbb{N}_0$. The linear operators T and T_0 depend on the parameters α, β. Due to (3.10) we have

$$\begin{aligned} &T : \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \times \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \to \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \times \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)), \\ &T(\hat{\mathfrak{f}})(\xi, v) = \Big(\hat{f}(\xi, v) + \alpha\, v\, \hat{g}(\xi, v), \beta\, v\, \hat{f}(\xi, v) + (1 - \Theta\, v^2)\, \hat{g}(\xi, v)\Big), \\ &\qquad\qquad (\xi, v) \in \mathbb{R} \times [0, 1], \end{aligned} \tag{3.14}$$

and

$$T_0 : \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \times \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \to \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \times \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)), \tag{3.15}$$
$$T_0(\hat{\mathfrak{f}})(\xi, v) = \left(\left[\hat{f} + \alpha\, v\, \hat{g} \right](\xi), \beta\, v \left[\hat{f} + \alpha\, v\, \hat{g} \right](\xi) \right), \quad (\xi, v) \in \mathbb{R} \times [0,1].$$

Remark. Due to (3.10) it is seemingly impossible to obtain uniform estimates (with respect to ϵ and the grid sizes) on the operator seminorms of T_0, T_1.

Our aim is to prove uniform bounds in *suitable* seminorms of the iterations $\mathfrak{f}^n = (f^n, g^n)$ for all $n \in \mathbb{N}_0$ and for all $\epsilon \geq 0$.

The results will depend on *pointwise* estimates of $\hat{\mathfrak{f}}^n$. Let us consider the formal limiting problem when ϵ is set to zero. In this situation, recursion (3.12) reduces to $\hat{\mathfrak{f}}^{n+1} = (\hat{f}^{n+1}, \hat{g}^{n+1}) = T_0 \hat{\mathfrak{f}}^n$. This means

$$\left.\begin{aligned} \hat{f}^{n+1} &= \left[\hat{f}^n + \alpha\, v\, \hat{g}^n \right] \\ \hat{g}^{n+1} &= v\, \beta \left[\hat{f}^n + \alpha\, v\, \hat{g}^n \right] = v\, \beta \left[\hat{f}^{n+1} \right] \end{aligned}\right\} \quad n \in \mathbb{N}_0,$$

with

$$\beta = \beta(\xi, \Delta t, \Delta x, \epsilon = 0) = -\xi\, \overline{H}(\xi \Delta x),$$

and therefore,

$$\hat{f}^{n+1} = \left(1 - \Theta \left[v^2\right]\right) \left[\hat{f}^n\right] = \left(1 - \frac{\Theta}{3}\right) \left[\hat{f}^n\right], \qquad n \in \mathbb{N}_0,$$

where

$$\Theta = \Theta(\xi, \Delta t, \Delta x, \epsilon = 0) = \frac{\Delta t}{(\Delta x)^2}(2 - 2\cos(\xi \Delta x)).$$

Thus, we have *pointwise* estimates

$$\left.\begin{aligned} \left|\hat{f}^n\right|(\xi, v) &= \left|1 - \frac{\Theta}{3}\right|^n \left|\left[\hat{f}^0\right]\right|(\xi) \\ \left|\hat{g}^{n+1}\right|(\xi, v) &\leq |v|\, |\xi| \left|1 - \frac{\Theta}{3}\right|^n \left|\left[\hat{f}^0\right]\right|(\xi) \end{aligned}\right\} \quad n \in \mathbb{N}_0. \tag{3.16}$$

In particular, whenever

$$\sup_{\xi \in \mathbb{R}} \Theta(\xi, \Delta t, \Delta x, \epsilon = 0) \leq 6,$$

which is the case iff the usual parabolic CFL condition for the diffusion equation (2.4),

$$\frac{\Delta t}{(\Delta x)^2} \leq \frac{3}{2}, \tag{3.17}$$

holds, then we obtain from (3.16)

$$\left|\hat{f}^n\right|(\xi, v) \leq \left\|\hat{f}^0(\xi, .)\right\|_{\mathcal{L}^1(dv)}, \qquad (n, \xi, v) \in \mathbb{N}_0 \times \mathbb{R} \times [0,1], \tag{3.18}$$

$$\left|\hat{g}^{n+1}\right|(\xi, v) \leq |\xi| \left\|\hat{f}^0(\xi, .)\right\|_{\mathcal{L}^1(dv)}, \qquad (n, \xi, v) \in \mathbb{N}_0 \times \mathbb{R} \times [0,1], \tag{3.19}$$

and therefore, since the Fourier transform is an isometry between $\mathcal{L}^2(dx)$ and $\mathcal{L}^2(d\xi)$,

$$\|f^n\|_{\mathcal{L}^2(d(x,v))} \le \left\|f^0\right\|_{\mathcal{L}^2(d(x,v))}, \quad \|f^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le \left\|f^0\right\|_{\mathcal{L}^2(d(x,v))}, \qquad n \in \mathbb{N}_0,$$

and, furthermore,

$$\left\|g^{n+1}\right\|_{\mathcal{L}^2(d(x,v))} \le \left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^2(dv))},$$

$$\left\|g^{n+1}\right\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le \left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^2(dv))}, \qquad n \in \mathbb{N}_0,$$

which indicates that the introduced spaces are canonical.

Our main result is the following theorem.

THEOREM 3.1. *Assume* $\Delta x, \Delta t, \epsilon \in \mathbb{R}^+$ *satisfy* (3.1). *With initial values* $\mathfrak{f}^0 \in \mathcal{L}^2(d(x,v)) \times \mathcal{L}^2(d(x,v))$, *define the sequence* $(\mathfrak{f}^n)_{n\in\mathbb{N}} = ((f^n, g^n))_{n\in\mathbb{N}}$ *by* (2.6).

Then, for all $n \in \mathbb{N}_0$,

(a) $\|f^n\|_{\mathcal{L}^2(d(x,v))} \le 21. \left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + 33.\sqrt{\epsilon^2+\Delta t}. \left\|g^0\right\|_{\mathcal{L}^2(d(x,v))}$

and

$$\|g^n\|_{\mathcal{L}^2(d(x,v))} \le \left(30 + \frac{3\epsilon^2}{(\epsilon^2+\Delta t)^{3/2}}\right) \cdot \left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left(3 + 48.\sqrt{\epsilon^2+\Delta t}\right) \cdot \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))},$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^2(d(x,v))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$;

(b) *if* $\mathfrak{f}^0 \in \mathcal{L}^\infty(dv,\mathcal{L}^2(dx))$, *then*

$$\|f^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le 24 \left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + 35\sqrt{\epsilon^2+\Delta t} \left\|g^0\right\|_{\mathcal{L}^2(d(x,v))}$$

and

$$\|g^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le \left(30 + \frac{9\epsilon^2}{(\epsilon^2+\Delta t)^{3/2}}\right) \left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left(4 + 48.\sqrt{\epsilon^2+\Delta t}\right) \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))},$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^2(d(x,v))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$.

Theorem 3.1 allows for the derivation of several stability results for (2.6) independently of ϵ. As examples, we deduce the following.

COROLLARY 3.2. *Let* M, ϵ_0 *be positive constants. Then there is a positive constant* $C_0 = C_0(M,\epsilon_0)$ *such that, for all* $\Delta x, \Delta t, \epsilon \in \mathbb{R}^+$, *if* $\Delta t, \Delta x, \epsilon$ *satisfy* (3.1), $\Delta t + \epsilon \le M$, *and* $\epsilon \le \epsilon_0 \Delta t$, *then the following estimates hold for any sequence* $(\mathfrak{f}^n)_{n\in\mathbb{N}_0} = ((f^n, g^n))_{n\in\mathbb{N}}$ *defined by* (2.6) *with initial value* $\mathfrak{f}^0 \in \mathcal{L}^2(d(x,v)) \times \mathcal{L}^2(d(x,v))$:

$$\|f^n\|_{\mathcal{L}^2(d(x,v))} \le C_0. \left(\left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + \left\|g^0\right\|_{\mathcal{L}^2(d(x,v))}\right)$$

and

$$\|g^n\|_{\mathcal{L}^2(d(x,v))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} \right),$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^2(d(x,v))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$.

COROLLARY 3.3. *Let* M, ϵ_1 *be positive constants. Then there is a positive constant* $C_1 = C_1(M,\epsilon_1)$ *such that, for all* $\Delta x, \Delta t, \epsilon \in \mathbb{R}^+$, *if* $\Delta t, \Delta x, \epsilon$ *satisfy* (3.1), $\Delta t + \epsilon \le M$, *and* $\epsilon_1 \le \epsilon$, *then the following estimates hold for any sequence* $(\mathfrak{f}^n)_{n\in\mathbb{N}} = ((f^n, g^n))_{n\in\mathbb{N}}$ *defined by* (2.6) *with initial value* $\mathfrak{f}^0 \in \mathcal{L}^\infty(dv, \mathcal{L}^2(dx))$:

$$\|f^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + \left\|g^0\right\|_{\mathcal{L}^2(d(x,v))} \right)$$

and

$$\|g^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} \right),$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$.

Remark. We notice that although the scheme is developed based on consideration of the diffusive limit ϵ tending to 0, the transport CFL condition (3.1) is sufficient to guarantee stability also for large ϵ. In particular, for large mean free paths, the time step is no longer restricted by a parabolic CFL condition related to the limiting diffusion equation (2.4), but it is restricted by the hyperbolic CFL condition (3.5) related to the transport equation (2.1).

Remark. The transport CFL condition (3.1) is seemingly not optimal. For example, for ϵ tending to 0 we have

$$\frac{\Delta t}{(\Delta x)^2} < \frac{1}{2}.$$

However, as the direct analysis for $\epsilon = 0$ shows, actually, the correct restriction is the parabolic CFL (3.17)

$$\frac{\Delta t}{(\Delta x)^2} < \frac{3}{2}.$$

Remark. The conditions $\epsilon_0 \le \epsilon \le M$ and $\epsilon \le \epsilon_1 \Delta t$ cover the well-resolved and underresolved cases, respectively.

4. Preliminaries. The main ingredients of the proof of Theorem 3.1 are investigations of recursion formulae. These investigations require several preliminary estimates.

LEMMA 4.1. *We define*

$$\psi : [0,2] \times [0,1] \to [0,1], \qquad \psi(\sigma, v) = \arccos\left(1 - \frac{\sigma v^2}{2}\right).$$

Then for all $(\sigma, n) \in [0,2] \times \mathbb{N}_0$,

$$\left| \int_0^1 \left(\cos(n\psi(\sigma,v)) - \sin(n\psi(\sigma,v)) \frac{1-\cos(\psi(\sigma,v))}{\sin(\psi(\sigma,v))} \right) dv \right| \le 1.$$

Proof. Introducing for fixed $\sigma \in (0,2)$ the new variable $t := \psi(\sigma, v)$, we obtain

$$\begin{aligned}
&\int_0^1 \left(\cos(n\psi(\sigma,v)) - \sin(n\psi(\sigma,v)) \frac{1-\cos(\psi(\sigma,v))}{\sin(\psi(\sigma,v))} \right) \, dv \\
&= \frac{1}{\sqrt{2\sigma}} \int_0^{\psi(\sigma,1)} \left(\sqrt{1+\cos(t)}.\cos(nt) - \sqrt{1-\cos(t)}.\sin(nt) \right) \, dt \\
&= \frac{1}{\sqrt{\sigma}} \int_0^{\psi(\sigma,1)} \left(\cos(t/2).\cos(nt) - \sin(t/2).\sin(nt) \right) \, dt \\
&= \frac{1}{\sqrt{\sigma}} \frac{\sin((n+(1/2)).\psi(\sigma,1))}{n+(1/2)} = \frac{\sin((n+(1/2)).\psi(\sigma,1))}{(n+(1/2))\sqrt{2(1-\cos(\psi(\sigma,1)))}} \\
&= \frac{\sin((n+(1/2)).\psi(\sigma,1))}{(n+(1/2)).\psi(\sigma,1)} \cdot \frac{\psi(\sigma,1)/2}{\sin(\psi(\sigma,1)/2)} \in [-1,+1]. \qquad \square
\end{aligned}$$

LEMMA 4.2. *Let $(c_n)_{n\in\mathbb{N}_0}$ and $(\gamma_n)_{n\in\mathbb{N}_0}$ be complex sequences. Define a complex sequence $(\kappa_n)_{n\in\mathbb{N}_0}$ by recursion via*

$$\kappa_0 = c \in \mathbb{C} \quad \forall n \in \mathbb{N}_0 : \quad \kappa_{n+1} = c_n + \left(\kappa_0.\gamma_{n-1} + \kappa_1.\gamma_{n-2} + \cdots + \kappa_{n-1}.\gamma_0 \right).$$

Assume

$$\sum_{k=0}^{\infty} |c_k| < \infty, \qquad \sum_{k=0}^{\infty} |\gamma_k| \le 1.$$

Then the sequence $(\kappa_n)_{n\in\mathbb{N}_0}$ is bounded; more precisely,

$$\forall n \in \mathbb{N}_0 : \quad |\kappa_n| \le |\kappa_0| + \left(|c_0| + \cdots + |c_{n-1}| \right). \tag{4.1}$$

Proof. We prove (4.1) by induction. There is nothing to do in case $n = 0$. To pass from n to $n+1$, we calculate

$$\begin{aligned}
|\kappa_{n+1}| &= \left| c_n + \sum_{j=0}^{n-1} \kappa_j.\gamma_{n-1-j} \right| \le |c_n| + \sum_{j=0}^{n-1} |\kappa_j|.|\gamma_{n-1-j}| \\
&\le |c_n| + \max\{|\kappa_0|, \ldots, |\kappa_{n-1}|\} \sum_{j=0}^{n-1} |\gamma_j| \le |c_n| + |\kappa_0| + |c_0| + \cdots + |c_{n-1}|. \qquad \square
\end{aligned}$$

Furthermore, we require the following result about the recursion scheme (2.6) when A is set to 1 (or equivalently, when B is set to 0).

LEMMA 4.3. *Let $\xi \in \mathbb{R}$ and let $\Delta t, \Delta x, \epsilon$ be positive real numbers. Let α, β, Θ be as in* (3.7), (3.8), (3.9), *respectively. For $(\hat{f}_0, \hat{g_0}) \in \mathcal{L}^2(d\xi, \mathcal{L}^1(dv)) \times \mathcal{L}^2(d\xi, \mathcal{L}^1(dv))$ and $(\xi, v) \in \mathbb{R} \times [0,1]$, we write T from* (3.14) *in vector notation:*

$$T \begin{pmatrix} \hat{f}_0 \\ \hat{g_0} \end{pmatrix} (\xi, v) = \begin{pmatrix} 1 & \alpha\, v \\ \beta\, v & 1 - \Theta\, v^2 \end{pmatrix} \begin{pmatrix} \hat{f}_0(\xi, v) \\ \hat{g_0}(\xi, v) \end{pmatrix}.$$

For $n \in \mathbb{N}$ let

$$\begin{pmatrix} \hat{f_n} \\ \hat{g_n} \end{pmatrix} = T^n \begin{pmatrix} \hat{f}_0 \\ \hat{g}_0 \end{pmatrix}.$$

Assume $\Delta t, \Delta x, \epsilon$ satisfy the transport CFL condition (3.1). *Then for all $n \in \mathbb{N}_0$ and for all $(\xi, v) \in \mathbb{R} \times (0,1)$,*

(a) $|\hat{f}_n|(\xi,v) \le \left(2|\hat{f}_0| + \sqrt{2(\epsilon^2+\Delta t)}\,|\hat{g}_0|\right)(\xi,v)$.

(b) $|\hat{g_n}|(\xi,v) \le \left(\frac{\sqrt{2}}{\sqrt{\epsilon^2+\Delta t}}\,|\hat{f}_0| + 2|\hat{g}_0|\right)(\xi,v)$.

(c) $|\alpha\, v\, \hat{g_n}|(\xi,v) \le \left(2|\hat{f}_0| + 4\sqrt{\epsilon^2+\Delta t}\,|\hat{g}_0|\right)(\xi,v)$.

(d) *If* $\hat{f}_0 = 1$ *and* $\hat{g}_0(\xi,v) = \beta\, v$, *then*

$$|\hat{f_n}|(\xi,v) \le 2, \quad |\hat{g_n}|(\xi,v) \le 3\,|\xi|, \quad \left|\int_0^1 (\hat{f_n}(\xi,s) + \alpha\, s\ \hat{g_n}(\xi,s))\, ds\right| \le 1.$$

Proof. First, we prove statements (a) and (b). We recall that if $\Delta t, \Delta x, \epsilon$ satisfy (3.1), then $\sup_{\xi\in\mathbb{R}} \Theta(\xi,\Delta t,\Delta x,\epsilon) < 2$. We shall use this estimate frequently.

We keep $(\xi, v) \in \mathbb{R}\times[0,1]$ fixed and introduce the 2×2 matrix

$$R := R(\alpha,\beta,\Theta,v) := \begin{pmatrix} 1 & \alpha\, v \\ \beta\, v & 1-\Theta\, v^2 \end{pmatrix}.$$

Then we have for all $n \in \mathbb{N}_0$,

$$T^n \begin{pmatrix} \hat{f}_0 \\ \hat{g}_0 \end{pmatrix}(\xi,v) = R^n \cdot \begin{pmatrix} \hat{f}_0(\xi,v) \\ \hat{g}_0(\xi,v) \end{pmatrix}.$$

If $\Theta = 0$, then $\xi = 0$ and therefore $\alpha = \beta = 0$ as well. In this case, R is the identity matrix and the proof of the lemma is straightforward.

Let us assume $\Theta > 0$ henceforth. The eigenvalues of R are

$$\lambda_{1,2}(v) = \left(1 - \frac{\Theta\, v^2}{2}\right) \pm \sqrt{\left(\frac{\Theta\, v^2}{2}\right)^2 - \Theta\, v^2}.$$

Since $a := \Theta\, v^2/2 < 1$, we have $2\,a - a^2 > 0$ such that

$$\lambda_{1,2} = (1-a) \pm i\,\sqrt{2\,a - a^2};$$

i.e., R has two distinct, nonreal, complex conjugate eigenvalues

$$\lambda_1 = \lambda := (1-a) + i\,\sqrt{2\,a-a^2}, \qquad \lambda_2 = \overline{\lambda}.$$

Hence there is a regular 2×2 matrix $B = B(\alpha,\beta,\Theta,v)$ with

$$R^n = B \begin{pmatrix} \lambda^n & 0 \\ 0 & \overline{\lambda}^n \end{pmatrix} B^{-1}, \qquad n \in \mathbb{N}_0.$$

Since $|\lambda| = 1$, we have $\lambda := e^{i\,\theta}$ for some $\theta \in (0, 2\pi)$. Since $\cos(\theta) = 1 - a > 0$ and $\sin(\theta) = \sqrt{2\,a - a^2} > 0$, we have $\theta \in (0, \pi/2)$. Furthermore,

$$v = \sqrt{\frac{2a}{\Theta}} = \sqrt{\frac{2}{\Theta}\cdot(1-\cos(\theta))} = \frac{2}{\sqrt{\Theta}}\sin(\theta/2) \tag{4.2}$$

and

$$B = \begin{pmatrix} -\alpha\, v & -\alpha\, v \\ 1-e^{i\theta} & 1-e^{-i\theta} \end{pmatrix}$$

such that for all $n \in \mathbb{N}_0$

$$
\begin{aligned}
R^n &= \frac{i}{2\alpha v \sin(\theta)} \begin{pmatrix} -\alpha\, v & -\alpha\, v \\ 1-e^{i\theta} & 1-e^{-i\theta} \end{pmatrix} \begin{pmatrix} e^{ni\theta} & 0 \\ 0 & e^{-ni\theta} \end{pmatrix} \begin{pmatrix} 1-e^{-i\theta} & \alpha\, v \\ -1+e^{i\theta} & -\alpha\, v \end{pmatrix} \\
&= \frac{i}{2\alpha v \sin(\theta)} \begin{pmatrix} -\alpha\, v & -\alpha\, v \\ 1-e^{i\theta} & 1-e^{-i\theta} \end{pmatrix} \begin{pmatrix} e^{ni\theta}-e^{(n-1)i\theta} & \alpha\, v\, e^{ni\theta} \\ -e^{-ni\theta}+e^{-(n-1)i\theta} & -\alpha\, v e^{-ni\theta} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sin(n\theta)-\sin((n-1)\theta)}{\sin(\theta)} & \alpha v \frac{\sin(n\theta)}{\sin(\theta)} \\ -\frac{2\sin(n\theta)-\sin((n-1)\theta)-\sin((n+1)\theta)}{\alpha v \sin(\theta)} & -\frac{\sin(n\theta)-\sin((n+1)\theta)}{\sin(\theta)} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sin(n\theta)-\sin((n-1)\theta)}{\sin(\theta)} & \frac{2\alpha}{\sqrt{\Theta}} \frac{\sin(\theta/2)\sin(n\theta)}{\sin(\theta)} \\ -\frac{\sqrt{\Theta}}{2\alpha} \frac{2\sin(n\theta)-\sin((n-1)\theta)-\sin((n+1)\theta)}{\sin(\theta/2)\sin(\theta)} & \frac{\sin(n\theta)-\sin((n+1)\theta)}{\sin(\theta)} \end{pmatrix} \\
&= \begin{pmatrix} \cos(n\theta)+\sin(n\theta)\frac{1-\cos(\theta)}{\sin(\theta)} & \frac{2\alpha}{\sqrt{\Theta}}\, \sin(n\theta)\, \frac{\sin(\theta/2)}{\sin(\theta)} \\ -\frac{\sqrt{\Theta}}{\alpha}\, \sin(n\theta)\, \frac{1-\cos(\theta)}{\sin(\theta/2)\, \sin(\theta)} & \cos(n\theta)-\sin(n\theta)\frac{1-\cos(\theta)}{\sin(\theta)} \end{pmatrix} \\
&=: \begin{pmatrix} R_{n;11} & R_{n;12} \\ R_{n;21} & R_{n;22} \end{pmatrix}.
\end{aligned}
$$

Since $\theta \in (0, \pi/2)$, we have

$$
|R_{n;11}|, |R_{n;22}| \le 2, \qquad n \in \mathbb{N}_0. \tag{4.3}
$$

Furthermore, we have for all $n \in \mathbb{N}_0$,

$$
\begin{aligned}
R_{n;12} &= \frac{2\alpha}{\sqrt{\Theta}}\, \sin(n\theta)\, \frac{\sin(\theta/2)}{\sin(\theta)} = \alpha v \frac{\sin(n\theta)}{\sin(\theta)} = \alpha v \frac{\sin(n\theta)}{\sqrt{\Theta} v \sqrt{1-\frac{\Theta v^2}{4}}} \\
&= \frac{\alpha}{\sqrt{\Theta}} \frac{\sin(n\theta)}{\sqrt{1-\frac{\Theta v^2}{4}}} = \frac{(\xi \Delta t H(\xi \Delta x))(\Delta x \sqrt{\epsilon^2+\Delta t})}{\Delta t \sqrt{2(1-\cos(\xi \Delta x))}} \frac{\sin(n\theta)}{\sqrt{1-\frac{\Theta v^2}{4}}} \\
&= \sqrt{\epsilon^2+\Delta t} \frac{\xi \Delta x H(\xi \Delta x)}{\sqrt{2(1-\cos(\xi \Delta x))}} \frac{\sin(n\theta)}{\sqrt{1-\frac{\Theta v^2}{4}}} \\
&= \sqrt{\epsilon^2+\Delta t} \frac{H(\xi \Delta x)}{|H(\xi \Delta x)|} \frac{\sin(n\theta)}{\sqrt{1-\frac{\Theta v^2}{4}}},
\end{aligned} \tag{4.4}
$$

and analogously,

$$
R_{n;21} = -\frac{1}{\sqrt{\epsilon^2+\Delta t}} \frac{|H(\xi \Delta x)|}{H(\xi \Delta x)} \frac{\sin(n\theta)}{\sqrt{1-\frac{\Theta v^2}{4}}}; \tag{4.5}
$$

hence

$$
\begin{aligned}
|R_{n;12}| &\le \sqrt{2(\epsilon^2+\Delta t)}, \\
|R_{n;21}| &\le \tfrac{\sqrt{2}}{\sqrt{\epsilon^2+\Delta t}}.
\end{aligned} \tag{4.6}
$$

Since for all $n \in \mathbb{N}_0$,

$$\begin{aligned} \hat{f}_n(\xi, v) &= R_{n;11}\, \hat{f}_0(\xi, v) + R_{n;12}\, \hat{g}_0(\xi, v), \\ \hat{g_n}(\xi, v) &= R_{n;21}\, \hat{f}_0(\xi, v) + R_{n;22}\, \hat{g}_0(\xi, v), \end{aligned} \tag{4.7}$$

statements (a) and (b) of the lemma follow from (4.3), (4.6), and (4.7).

Now we prove statement (c). We calculate for all $n \in \mathbb{N}_0$,

$$\alpha v R_{n;21} = -v\sqrt{\Theta}\, \sin(n\theta)\, \frac{1-\cos(\theta)}{\sin(\theta/2)\, \sin(\theta)} = -2 \sin(n\theta) \cdot \frac{1-\cos(\theta)}{\sin(\theta)}$$

and

$$\alpha v R_{n;22} = 2 \sin(\theta/2) \frac{\alpha}{\sqrt{\Theta}} R_{n;22}.$$

Hence for all $n \in \mathbb{N}_0$,

$$|\alpha v R_{n;21}| \leq 2 \tag{4.8}$$

and

$$|\alpha v R_{n;22}| \leq 4\, \sqrt{\epsilon^2 + \Delta t}. \tag{4.9}$$

Statement (c) of the lemma follows from (4.7), (4.8), and (4.9).

For the proof of statement (d), we finally turn our attention to $\hat{f}_0 = 1$ and $\hat{g}_0(\xi, v) = \beta\, v$. We set

$$P_0 = 1, \qquad Q_0 = 1$$

and easily verify

$$\hat{f}_n = P_n, \quad \hat{g_n} = \beta\, v\, Q_n, \quad n \in \mathbb{N}_0,$$

where, for $n \in \mathbb{N}_0$,

$$\begin{pmatrix} P_{n+1} \\ \\ Q_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & -\Theta\, v^2 \\ \\ 1 & 1 - \Theta\, v^2 \end{pmatrix} \begin{pmatrix} P_n \\ \\ Q_n \end{pmatrix},$$

from which we obtain, after some elementary manipulations for all $n \in \mathbb{N}_0$ (we recall $\Theta < 2$),

$$\begin{aligned} P_n &= \mathrm{Re}\left(\left((1-a) - i.\sqrt{2a - a^2} \right)^n . \left(1 - \frac{i.(2a - a^2)}{\sqrt{2a - a^2}} \right) \right), \\ Q_n &= \mathrm{Re}\left(\left((1-a) - i.\sqrt{2a - a^2} \right)^n . \left(1 + \frac{i.(2 - 2a)}{\sqrt{2a - a^2}} \right) \right), \end{aligned}$$

where we recall $a = \frac{\Theta v^2}{2}$. Then we have for all $n \in \mathbb{N}_0$,

$$\left[(1, \alpha v) \cdot \begin{pmatrix} \hat{f}_n \\ \hat{g_n} \end{pmatrix} \right] = \left[(1, \alpha v) \cdot \begin{pmatrix} P_n \\ \beta\, v\, Q_n \end{pmatrix} \right] = \left[P_n - \Theta\, v^2\, Q_n \right] = [P_{n+1}]. \tag{4.10}$$

Writing as above,

$$(1-a)+i\sqrt{2a-a^2}=e^{i\theta},$$

we have $\theta\in(0,\pi/2)$, $\cos(\theta)=(1-a)$, $\sin(\theta)=\sqrt{2a-a^2}$. Hence for all $n\in\mathbb{N}_0$,

$$P_n=\cos(n\theta)-\sin(n\theta)\frac{1-\cos(\theta)}{\sin(\theta)},$$

$$Q_n=\cos(n\theta)+2\sin(n\theta)\frac{1-\cos(\theta)}{\sin(\theta)}$$

such that for all $n\in\mathbb{N}_0$,

$$|\hat{f_n}|(\xi,v)=|P_n|(\xi,v)\le 2,\ |\hat{g_n}|(\xi,v)=|\beta|\,|v|\,|Q_n|(\xi,v)\le 3\,|\beta|\le 3\,|\xi|,$$

and due to Lemma 4.1 and (4.10) for all $n\in\mathbb{N}_0$,

$$\begin{aligned}(4.11)\quad &\int_0^1 (1,\alpha\,s)\cdot\begin{pmatrix}\hat{f_n}(\xi,s)\\ \hat{g_n}(\xi,s)\end{pmatrix}\,ds=[P_{n+1}](\xi)\\ &=\int_0^1\left(\cos((n+1)\theta(s))-\sin((n+1)\theta(s))\frac{1-\cos(\theta(s))}{\sin(\theta(s))}\right)\,ds\in[-1,+1],\end{aligned}$$

where $\cos(\theta(s))=1-\frac{\Theta\,s^2}{2}$, $s\in(0,1)$. ∎

5. Proof of Theorem 3.1. For $n\in\mathbb{N}$ let $\hat{f}^n,\hat{g}^n$ be as in (3.6) and let $(\xi,v)\in\mathbb{R}\times(0,1)$ be fixed. We now introduce for $n\in\mathbb{N}$ or $n=0$ the complex number

$$\kappa_n:=B\left[\hat{f}^n+\alpha\,v\,\hat{g}^n\right](\xi). \tag{5.1}$$

Then it is easy to see that for all $n\in\mathbb{N}$,

$$\begin{pmatrix}\hat{f}^n\\ \hat{g}^n\end{pmatrix}=A^n\,T^n\begin{pmatrix}\hat{f}^0\\ \hat{g}^0\end{pmatrix}+\sum_{j=0}^{n-1}\kappa_j\,A^{n-1-j}\,T^{n-1-j}\begin{pmatrix}1\\ \beta v\end{pmatrix}. \tag{5.2}$$

We derive from (5.2) a recursion formula for $(\kappa_n)_{n\in\mathbb{N}}$,

$$\begin{aligned}\kappa_0&=B.\left[\hat{f}^0+\alpha\,v\,\hat{g}^0\right](\xi),\\ \kappa_n&=B.\left[(1,\alpha\,v)\cdot\begin{pmatrix}\hat{f}^n\\ \hat{g}^n\end{pmatrix}\right]\\ &=A^nB.\left[(1,\alpha v)\cdot\left(T^n\begin{pmatrix}\hat{f}^0\\ \hat{g}^0\end{pmatrix}\right)\right]\\ &\quad+\sum_{j=0}^{n-1}\kappa_jBA^{n-1-j}\left[(1,\alpha v)\cdot\left(T^{n-1-j}\begin{pmatrix}1\\ \beta v\end{pmatrix}\right)\right]\\ &=c_n+(\kappa_0\gamma_{n-1}+\cdots+\kappa_{n-1}\gamma_0),\end{aligned}$$

where for $n\in\mathbb{N}$,

$$c_n:=A^nB\left[(1,\alpha v)\cdot\left(T^n\begin{pmatrix}\hat{f}^0\\ \hat{g}^0\end{pmatrix}\right)\right](\xi),$$

$$\gamma_n = BA^n \left[(1, \alpha v) \cdot \left(T^n \begin{pmatrix} 1 \\ \beta v \end{pmatrix} \right) \right] (\xi).$$

By part (d) of Lemma 4.3 we have for all $n \in \mathbb{N}$,

$$|\gamma_n| \leq BA^n;$$

hence,

$$\sum_{n=0}^{\infty} |\gamma_n| \leq 1. \tag{5.3}$$

Furthermore, due to parts (a) and (c) of Lemma 4.3, we have for all $n \in \mathbb{N}$,

$$\begin{aligned} &\left| \left[(1, \alpha v) \cdot \left(T^n \begin{pmatrix} \hat{f}^0 \\ \hat{g}^0 \end{pmatrix} \right) \right] \right| (\xi) \\ &\qquad \leq \left[2 \left| \hat{f}^0 \right| + \sqrt{2(\epsilon^2 + \Delta t)} \, \left| \hat{g}^0 \right| \right] (\xi) + \left[2 \left| \hat{f}^0 \right| + 4\sqrt{\epsilon^2 + \Delta t} \, \left| \hat{g}^0 \right| \right] (\xi) \\ &\qquad\qquad \leq 4 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 6\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)}; \end{aligned}$$

hence for all $n \in \mathbb{N}$,

$$|c_n| \leq A^n B \left(4 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 6\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} \right); \tag{5.4}$$

in particular, $(c_n)_{n \in \mathbb{N}}$ is in $\ell^1(\mathbb{C})$.

We can therefore apply Lemma 4.2 to deduce for all $n \in \mathbb{N}$ the estimate

$$\begin{aligned} &|\kappa_n| \leq |\kappa_0| + (|c_0| + \cdots + |c_{n-1}|) \qquad (5.5) \\ &\leq \left| B \left[\hat{f}^0 + \alpha \, v \, \hat{g}^0 \right] \right| + \left(4 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 6\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} \right) \sum_{j=0}^{n-1} A^n B \\ &\leq [|\hat{f}^0|](\xi) + 2\sqrt{\epsilon^2 + \Delta t} \, [|\hat{g}^0|](\xi) + 4 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 6\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} \\ &\qquad = 5 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 8\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)}, \end{aligned}$$

where we made use of $|\alpha \, v| \leq 2\sqrt{\epsilon^2 + \Delta t}$; see (4.2).

From (5.2), (5.5), and parts (a) and (d) of Lemma 4.3, we deduce the estimate

$$\begin{aligned} |\hat{f}^n|(\xi, v) &\leq \left(2 \, |\hat{f}^0| + \sqrt{2 \, (\epsilon^2 + \Delta t)} \, |\hat{g}^0| \right) (\xi, v) \qquad (5.6) \\ &\quad + 10 \left\| \hat{f}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)} + 16\sqrt{\epsilon^2 + \Delta t} \, \left\| \hat{g}^0(\xi, .) \right\|_{\mathcal{L}^1(dv)}, \\ &\qquad\qquad (n, \xi, v) \in \mathbb{N}_0 \times \mathbb{R} \times [0, 1], \end{aligned}$$

and therefore due to

$$\forall (\delta_1, \delta_2, \delta_3, \delta_4) \in \mathbb{R}^4 : \quad (|\delta_1| + |\delta_2| + |\delta_3| + \delta_4|)^2 \leq 4. \left(|\delta_1|^2 + |\delta_2|^2 + |\delta_3|^2 + \delta_4|^2 \right),$$

we obtain the estimate

$$\left\|\hat{f}^n\right\|^2_{\mathcal{L}^2(d(\xi,v))} \leq 416.\left\|\hat{f}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))} + 1032.\left(\epsilon^2+\Delta t\right).\left\|\hat{g}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))}, \qquad n\in\mathbb{N}_0.$$

Hence, via

$$\forall(\delta_1,\delta_2)\in\mathbb{R}^2: \quad \sqrt{|\delta_1|^2+|\delta_2|^2}\leq|\delta_1|+|\delta_2|,$$

we deduce

$$\begin{aligned}\|f^n\|_{\mathcal{L}^2(d(x,v))} = \left\|\hat{f}^n\right\|_{\mathcal{L}^2(d(\xi,v))} &\leq 21.\left\|\hat{f}^0\right\|_{\mathcal{L}^2(d(\xi,v))} + 33.\sqrt{\epsilon^2+\Delta t}.\left\|\hat{g}^0\right\|_{\mathcal{L}^2(d(\xi,v))}\\ &= 21.\left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + 33.\sqrt{\epsilon^2+\Delta t}.\left\|g^0\right\|_{\mathcal{L}^2(d(x,v))}, \qquad n\in\mathbb{N}_0.\end{aligned}$$

In a similar way we deduce from (5.2), (5.5), and parts (b) and (d) of Lemma 4.3 the estimate

$$\begin{aligned}(5.7)\quad |\hat{g}^n|(\xi,v) &\leq A\left(\frac{\sqrt{2}}{\sqrt{\epsilon^2+\Delta t}}\,|\hat{f}^0|+2\,|\hat{g}^0|\right)(\xi,v)\\ &\quad+|\xi|\left(15\left\|\hat{f}^0(\xi,.)\right\|_{\mathcal{L}^1(dv)}+24\sqrt{\epsilon^2+\Delta t}\,\left\|\hat{f}^0(\xi,.)\right\|_{\mathcal{L}^1(dv)}\right)\\ &\leq\left(\frac{\sqrt{2}\,\epsilon^2}{(\epsilon^2+\Delta t)^{3/2}}\,|\hat{f}^0|+2\,|\hat{g}^0|\right)(\xi,v)\\ &\quad+|\xi|\left(15\left\|\hat{f}^0(\xi,.)\right\|_{\mathcal{L}^1(dv)}+24\sqrt{\epsilon^2+\Delta t}\,\left\|\hat{g}^0(\xi,.)\right\|_{\mathcal{L}^1(dv)}\right),\\ &\qquad\qquad (n,\xi,v)\in\mathbb{N}_0\times\mathbb{R}\times[0,1],\end{aligned}$$

and therefore

$$\begin{aligned}|\hat{g}^n|^2(\xi,v) &\leq \frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}\cdot|\hat{f}^0|^2(\xi,v)+900.|\xi|^2.\left\|\hat{f}^0(\xi,.)\right\|^2_{\mathcal{L}^1(dv)}\\ &\quad+8|\hat{g}^0|^2(\xi,v)+2304.|\xi|^2.\left(\epsilon^2+\Delta t\right).\left\|\hat{g}^0(\xi,.)\right\|^2_{\mathcal{L}^1(dv)},\\ &\leq\left(\frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}+900\right).\left(|\hat{f}^0|^2(\xi,v)+|\xi|^2.\left\|\hat{f}^0(\xi,.)\right\|^2_{\mathcal{L}^1(dv)}\right)\\ &\quad+\left(8+2304.\left(\epsilon^2+\Delta t\right)\right).\left(|\hat{g}^0|^2(\xi,v)+|\xi|^2.\left\|\hat{g}^0(\xi,.)\right\|^2_{\mathcal{L}^1(dv)}\right),\\ &\qquad\qquad (n,\xi,v)\in\mathbb{N}_0\times\mathbb{R}\times[0,1],\end{aligned}$$

which implies

$$\begin{aligned}&\|g^n\|_{\mathcal{L}^2(d(x,v))}\\ &\leq\sqrt{\frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}+900.}\,\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(v))}\sqrt{8+2304.(\epsilon^2+\Delta t)}.\left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(v))}\\ &\leq\left(30+\frac{3\epsilon^2}{(\epsilon^2+\Delta t)^{3/2}}\right).\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))}+\left(3+48.\sqrt{\epsilon^2+\Delta t}\right).\left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))},\\ &\qquad\qquad n\in\mathbb{N}_0.\end{aligned}$$

This establishes part (a) of Theorem 3.1. On the other hand, we deduce from (5.6) that due to the fact that the Fourier transform is an isometry between $\mathcal{L}^2(dx)$ and $\mathcal{L}^2(d\xi)$,

$$\begin{aligned}\left\|f^n(.,w)\right\|^2_{\mathcal{L}^2(dx)} &\leq 16\left\|f^0(.,w)\right\|^2_{\mathcal{L}^2(dx)} + 8\left(\epsilon^2+\Delta t\right)\left\|g^0(.,w)\right\|^2_{\mathcal{L}^2(dx)} \\ &\quad + 400\left\|f^0\right\|^2_{\mathcal{L}^2(d(x,v))} + 1024.\left(\epsilon^2+\Delta t\right).\left\|g^0\right\|^2_{\mathcal{L}^2(d(x,v))}, \\ &\qquad\qquad (n,w)\in\mathbb{N}_0\times[0,1],\end{aligned}$$

and therefore

$$\begin{aligned}\left\|f^n\right\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} &\leq 4\left\|f^0\right\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} + 3\sqrt{\epsilon^2+\Delta t}\left\|g^0\right\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \\ &\quad + 20\left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + 32.\sqrt{\epsilon^2+\Delta t}.\left\|g^0\right\|_{\mathcal{L}^2(d(x,v))}, \qquad n\in\mathbb{N}_0,\end{aligned}$$

which yields, via

$$\left\|\varphi\right\|_{\mathcal{L}^2(d(x,v))} \leq \left\|\varphi\right\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))}, \qquad \varphi\in\mathcal{L}^\infty(dv,\mathcal{L}^2(dx)),$$

the first estimate of (b).

Finally, we deduce from (5.7),

$$\begin{aligned}\left\|g^n(.,w)\right\|^2_{\mathcal{L}^2(dx)} &\leq \frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}\cdot\left\|f^0\right\|^2_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} + 16\left\|g^0\right\|^2_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \\ &\quad + 900\left\|\xi.\hat{f}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))} + 2304.\left(\epsilon^2+\Delta t\right).\left\|\xi.\hat{g}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))}, \\ &\leq \frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}\cdot\left\|f^0\right\|^2_{\mathcal{L}^2(d(x,v))} + 16\left\|g^0\right\|^2_{\mathcal{L}^2(d(x,v))} \\ &\quad + 900\left\|\xi.\hat{f}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))} + 2304.\left(\epsilon^2+\Delta t\right).\left\|\xi.\hat{g}^0\right\|^2_{\mathcal{L}^2(d(\xi,v))}, \\ &\leq \left(\frac{8\epsilon^4}{(\epsilon^2+\Delta t)^3}+900\right)\left\|f^0\right\|^2_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left(16+2304.\left(\epsilon^2+\Delta t\right)\right)\left\|g^0\right\|^2_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))}, \\ &\qquad\qquad (n,w)\in\mathbb{N}_0\times[0,1].\end{aligned}$$

The second estimate of (b) follows by elementary manipulations. ∎

6. Discrete function spaces. In this section we investigate the action of the recursion (3.12), (2.6) on a discrete space. The discretization considered here is a piecewise affine FEM on a uniform grid $(j.\Delta x)_{j\in\mathbb{Z}}$; i.e., the values at the spatial gridpoints given by the recursion define a continuous function on the whole space via linear interpolation.

We introduce some notation. The identity function on $\mathbb{R}$ is $\mathrm{id}:\mathbb{R}\to\mathbb{R}$, $\mathrm{id}(x)=x$ and if $A\subseteq\mathbb{R}$, then the indicator function of A is $\mathrm{ind}_A:\mathbb{R}\to\{0,1\}$, $\mathrm{ind}_A(x)=1$ iff $x\in A$. If $\varphi:\mathbb{R}\to\mathbb{R}$ and if $A\subseteq\mathbb{R}$, then the restriction of φ to A is

$$\varphi\restriction A : A\to\mathbb{R}, \qquad (\varphi\restriction A)(x)=\varphi(x).$$

We introduce the sequence space

$$\ell^2(\Delta x,dv)=\left\{(\gamma_j)_{j\in\mathbb{Z}}:\ \left(\forall j\in\mathbb{Z}:\ \gamma_j\in\mathcal{L}^2(dv)\right)\right\}$$

equipped with the Δx-dependent norm

$$\|(\gamma_j)_{j\in\mathbb{Z}}\|_{\ell^2(\Delta x,dv)} = \Delta x \cdot \sqrt{\sum_{j\in\mathbb{Z}} \|\gamma_j\|^2_{\mathcal{L}^2(dv)}}.$$

Given $\Delta x \in \mathbb{R}^+$, a function $\varphi : \mathbb{R}\times[0,1] \to \mathbb{C}$ is a square integrable affine Δx finite element function iff $\varphi \in \mathcal{L}^2(d(x,v))$ and if there is $(\gamma_j)_{j\in\mathbb{Z}} \in \ell^2(\Delta x, dv)$ such that

$$\varphi = \sum_{j\in\mathbb{Z}} \left(\gamma_{j+1} \cdot \frac{\mathrm{id} - j.\Delta x}{\Delta x} + \gamma_j \cdot \frac{(j+1).\Delta x - \mathrm{id}}{\Delta x} \right) \cdot \mathrm{ind}_{[j.\Delta x,(j+1).\Delta x)};$$

i.e., $\varphi \in \mathcal{L}^2(d(x,v))$ is an affine Δx finite element function iff for all $v \in [0,1]$, $\varphi(.,v)$ is continuous and for all $(j,v) \in \mathbb{Z}\times[0,1]$, $\varphi(.,v) \restriction [j.\Delta x,(j+1).\Delta x)$ is affine.

The set of all square integrable affine Δx finite element functions is $\mathcal{L}^2(\Delta x, dv)$.

We observe that the mapping

$$\iota_{\Delta x} : \mathcal{L}^2(\Delta x, dv) \to \ell^2(\Delta x, dv), \qquad \iota_{\Delta x}(\phi) = (\phi(j\Delta x, .))_{j\in\mathbb{Z}}$$

is a linear isometry.

Next we shall prove the following lemma.

LEMMA 6.1. *For all $(\Delta x, \Delta t, \epsilon) \in \mathbb{R}^+\times\mathbb{R}^+\times\mathbb{R}^+$, the operator S of algorithm (2.6) maps $\mathcal{L}^2(\Delta x, v)\times\mathcal{L}^2(\Delta x, v)$ into itself. Furthermore, if $(f,g) \in \mathcal{L}^2(\Delta x, dv)\times\mathcal{L}^2(\Delta x, dv)$, and if we set*

$$(f^*, g^*) = S(f,g),$$

then for all $(j,v) \in \mathbb{Z}\times[0,1]$,

(6.1)
$$f^*(j\Delta x, v) = A(f(j\Delta x, v) - \Delta t\, v\, D_+ g(j\Delta x, v)) + B\left[f(j\Delta x, .) - \Delta t\, v\, D_+ g(j\Delta x, .)\right],$$

(6.2)
$$g^*(j\Delta x, v) = Ag(j\Delta x, v) - v\, A\, B\, D_-(f(j\Delta x, v) - \Delta t\, v\, D_+ g(j\Delta x, v)) - v\, B^2\, D_-\left[f(j\Delta x, .) - \Delta t\, v\, D_+ g(j\Delta x, .)\right].$$

Proof. We recall that if $(f,g) \in \mathcal{L}^2(\Delta x, dv)\times\mathcal{L}^2(\Delta x, dv)$, and if $(f^*, g^*) = S(f,g)$, then

$$f^* = A(f - \Delta t\, v\, D_+ g) + B\left[f - \Delta t\, v\, D_+ g\right], \tag{6.3}$$

$$g^* = Ag - v\, A\, B\, D_-(f - \Delta t\, v\, D_+ g) - v\, B^2\, D_-\left[f - \Delta t\, v\, D_+ g\right]. \tag{6.4}$$

Furthermore, for all $\varphi \in \mathcal{L}^2(\Delta x, dv)$—in particular for f and for g—we have

$$\varphi = \sum_{j\in\mathbb{Z}} \left(\varphi((j+1)\Delta x, .)\frac{\mathrm{id} - j.\Delta x}{\Delta x} + \varphi(j\Delta x, .)\frac{(j+1).\Delta x - \mathrm{id}}{\Delta x} \right) \mathrm{ind}_{[j.\Delta x,(j+1).\Delta x)}.$$

We proceed stepwise.

Step 1. Since $(f,g) \in \mathcal{L}^2(\Delta x, dv) \times \mathcal{L}^2(\Delta x, dv)$, and since $\mathcal{L}^2(\Delta x, dv) \subset \mathcal{L}^2(d(x,v))$, we have $(f,g) \in \mathcal{L}^2(d(x,v)) \times \mathcal{L}^2(d(x,v))$, and thus $(f^*, g^*) = S(f,g) \in \mathcal{L}^2(d(x,v)) \times \mathcal{L}^2(d(x,v))$.

Step 2. The function $\varphi(. + \Delta x, .) : \mathbb{R} \times [0,1] \to \mathbb{C}$, $(x,v) \mapsto \varphi(x + \Delta x, v)$ is in $\mathcal{L}^2(\Delta x, dv)$. Indeed, we have $\varphi(. + \Delta x, .) \in \mathcal{L}^2(d(x,v))$ and, for all $(x,v) \in \mathbb{R} \times [0,1]$,

$$
\begin{aligned}
&\varphi\left(x + \Delta x, v\right) \\
&= \sum_{j \in \mathbb{Z}} \left(\varphi((j+1).\Delta x, v) \cdot \frac{x + \Delta x - j.\Delta x}{\Delta x} + \varphi(j.\Delta x, v) \cdot \frac{(j+1).\Delta x - (x + \Delta x)}{\Delta x} \right) \\
&\qquad \cdot \operatorname{ind}_{[j.\Delta x, (j+1).\Delta x)}(x + \Delta x) \\
&= \sum_{j \in \mathbb{Z}} \left(\varphi((j+2).\Delta x, v) \cdot \frac{x + \Delta x - (j+1).\Delta x}{\Delta x} \right. \\
&\qquad \left. + \varphi((j+1).\Delta x, v) \cdot \frac{(j+2).\Delta x - (x + \Delta x)}{\Delta x} \right) \cdot \operatorname{ind}_{[j.\Delta x, (j+1).\Delta x)}(x) \\
&= \sum_{j \in \mathbb{Z}} \left(\varphi((j+2).\Delta x, v) \cdot \frac{x - j.\Delta x}{\Delta x} + \varphi((j+1).\Delta x, v) \cdot \frac{(j+1).\Delta x - x}{\Delta x} \right) \\
&\qquad \cdot \operatorname{ind}_{[j.\Delta x, (j+1).\Delta x)}(x);
\end{aligned}
$$

i.e., for all $(j,v) \in \mathbb{Z} \times [0,1]$, $\varphi(. + \Delta x, v) \restriction [j.\Delta x, (j+1).\Delta x)$ is affine, and $\varphi(. + \Delta x, v)$ is continuous.

Step 3. The proof of Step 2 can be easily modified to prove that the function $\varphi(. - \Delta x, .) : \mathbb{R} \times [0,1] \to \mathbb{C}$, $(x,v) \mapsto \varphi(x - \Delta x, v)$ is in $\mathcal{L}^2(\Delta x, dv)$.

Step 4. $[\varphi] \in \mathcal{L}^2(\Delta x, dv)$. Certainly, $[\varphi] \in \mathcal{L}^2(d(x,v))$. Furthermore, for all $(x,v) \in \mathbb{R} \times [0,1]$,

$$
\begin{aligned}
[\varphi]\left(x\right) = \sum_{j \in \mathbb{Z}} \left([\varphi((j+1).\Delta x, .)] \cdot \frac{x - j.\Delta x}{\Delta x} + [\varphi(j.\Delta x, .)] \cdot \frac{(j+1).\Delta x - x}{\Delta x} \right) \\
\cdot \operatorname{ind}_{[j.\Delta x, (j+1).\Delta x)}(x);
\end{aligned}
$$

i.e., for all $(j,v) \in \mathbb{Z} \times [0,1]$, $[\varphi] \restriction [j.\Delta x, (j+1).\Delta x)$ is affine, and $[\varphi]$ is continuous.

Step 5. By Steps 1 and 2, $D_+ g \in \mathcal{L}^2(\Delta x, dv)$. Hence $\Delta t\, v\, D_+ g \in \mathcal{L}^2(\Delta x, dv)$, and therefore $f - \Delta t\, v\, D_+ g \in \mathcal{L}^2(\Delta x, dv)$, which implies, via Step 4, $[f - \Delta t\, v\, D_+ g] \in \mathcal{L}^2(\Delta x, dv)$, such that $f^* = A.\,(f - \Delta t\, v\, D_+ g) + B.[f - \Delta t\, v\, D_+ g] \in \mathcal{L}^2(\Delta x, dv)$. In a similiar way we prove $g^* \in \mathcal{L}^2(\Delta x, dv)$.

Step 6. Formulae (6.1) and (6.2) follow from evaluations (6.3) and (6.4) at $(j\Delta x, v)$, $(j,v) \in \mathbb{Z} \times [0,1]$.

This finishes the proof of Lemma 6.1. ∎

Since a function $\varphi \in \mathcal{L}^2(\Delta x, dv)$ is entirely determined by $\varphi(j\Delta, v)$, $(j,v) \in \mathbb{Z} \times [0,1]$, we deduce the following corollary from Lemma 6.1.

COROLLARY 6.2. *Let* $(f_j^0)_{j \in \mathbb{Z}} \in \ell^2(\Delta x, dv)$, $(g_j^0)_{j \in \mathbb{Z}} \in \ell^2(\Delta x, dv)$. *For* $n \in \mathbb{N}_0$ *let*

$(f_j^{n+1})_{j\in\mathbb{Z}}$, $(g_j^{n+1})_{j\in\mathbb{Z}}$ be pointwise defined via the following version of (2.6):

$$f_j^{n+1}(v) = A.\left(f_j^{n+1}(v) - \Delta t\, v\,\frac{g_{j+1}^n(v) - g_j^n(v)}{\Delta x}\right) + B.\left[f_j^{n+1}(.) - \Delta t\, v\,\frac{g_{j+1}^n(.) - g_j^n(.)}{\Delta x}\right], \tag{6.5}$$

$$\begin{aligned} g_j^{n+1}(v) &= A\, g_j^n \\ &\quad - v\, A\, B\,\left(\frac{f_{j-1}^n(v) - f_j^n(v)}{\Delta x}\right) + \Delta t\, v^2\,\left(\frac{g_{j+1}^n(v) - 2g_j^n(v) + g_{j-1}^n(v)}{(\Delta x)^2}\right) \\ &\quad - v\, B^2\,\left[\frac{f_{j-1}^n(.) - f_j^n(.)}{\Delta x}\right] + \Delta t\, v\, B^2\,\left[v\cdot\frac{g_{j+1}^n(.) - 2g_j^n(.) + g_{j-1}^n(.)}{(\Delta x)^2}\right]. \end{aligned} \tag{6.6}$$

Then

$$(f_j^n)_{j\in\mathbb{Z}} \in \ell^2(\Delta x, dv), \quad (g_j^n)_{j\in\mathbb{Z}} \in \ell^2(\Delta x, dv), \qquad n \in \mathbb{N}_0,$$

and if we set

$$f^n = \sum_{j\in\mathbb{Z}}\left(f_{j+1}^n\cdot\frac{\mathrm{id} - j.\Delta x}{\Delta x} + f_j^n\cdot\frac{(j+1).\Delta x - \mathrm{id}}{\Delta x}\right)\cdot \mathrm{ind}_{[j.\Delta x,(j+1).\Delta x)}, \quad n\in\mathbb{N}_0, \tag{6.7}$$

and

$$g^n = \sum_{j\in\mathbb{Z}}\left(g_{j+1}^n\cdot\frac{\mathrm{id} - j.\Delta x}{\Delta x} + g_j^n\cdot\frac{(j+1).\Delta x - \mathrm{id}}{\Delta x}\right)\cdot \mathrm{ind}_{[j.\Delta x,(j+1).\Delta x)}, \quad n\in\mathbb{N}_0, \tag{6.8}$$

then

$$\forall n\in\mathbb{N}_0: \quad (f^{n+1}, g^{n+1}) = S(f^n, g^n);$$

i.e., the sequence $((f^n, g^n))_{n\in\mathbb{N}_0}$ is obtained via recursion (2.6) *from the initial pair (f^0, g^0).*

Now we are in the position to finally derive two results which show that for the affine Δx FEM recursion (6.5), (6.6) the same error estimates hold as in the continuous case.

COROLLARY 6.3. *Given $(f^0, g^0) \in \mathcal{L}^2(\Delta x, dv)\times\mathcal{L}^2(\Delta x, dv)$, let*

$$f_j^0 : [0,1]\to\mathbb{C}, \quad f_j^0(v) = f^0(j\Delta x, v), \qquad j\in\mathbb{Z},$$

$$g_j^0 : [0,1]\to\mathbb{C}, \quad g_j^0(v) = g^0(j\Delta x, v), \qquad j\in\mathbb{Z}.$$

Furthermore, let M, ϵ_0 be positive constants. Then there is a positive constant $C_0 = C_0(M,\epsilon_0)$ such that, for all $\Delta x, \Delta t, \epsilon\in\mathbb{R}^+$, if $\Delta t, \Delta x, \epsilon$ satisfy (3.1), *$\Delta t+\epsilon\le M$, and $\epsilon\le\epsilon_0\Delta t$, then the following estimates hold for the sequence $(\mathfrak{f}^n)_{n\in\mathbb{N}_0} = ((f^n, g^n))_{n\in\mathbb{N}}$ defined by* (6.5), (6.6), (6.7), *and* (6.8):

$$\|f^n\|_{\mathcal{L}^2(d(x,v))} \le C_0.\left(\|f^0\|_{\mathcal{L}^2(d(x,v))} + \|g^0\|_{\mathcal{L}^2(d(x,v))}\right)$$

and

$$\|g^n\|_{\mathcal{L}^2(d(x,v))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} \right),$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^2(d(x,v))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$.

COROLLARY 6.4. *Given* $(f^0, g^0) \in \mathcal{L}^2(\Delta x, dv) \times \mathcal{L}^2(\Delta x, dv)$, *let*

$$f_j^0 : [0,1] \to \mathbb{C}, \quad f_j^0(v) = f^0(j\Delta x, v), \qquad j \in \mathbb{Z},$$

$$g_j^0 : [0,1] \to \mathbb{C}, \quad g_j^0(v) = g^0(j\Delta x, v), \qquad j \in \mathbb{Z}.$$

Furthermore, let M, ϵ_1 *be positive constants. Then there is a positive constant* $C_1 = C_1(M, \epsilon_1)$ *such that, for all* $\Delta x, \Delta t, \epsilon \in \mathbb{R}^+$, *if* $\Delta t, \Delta x, \epsilon$ *satisfy* (3.1), $\Delta t + \epsilon \le M$, *and* $\epsilon_1 \le \epsilon$, *then the following estimates hold for the sequence* $(\mathfrak{f}^n)_{n\in\mathbb{N}_0} = ((f^n, g^n))_{n\in\mathbb{N}}$ *defined by* (6.5), (6.6), (6.7), *and* (6.8):

$$\|f^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{L}^2(d(x,v))} + \left\|g^0\right\|_{\mathcal{L}^2(d(x,v))} \right)$$

and

$$\|g^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))} \le C_0 \cdot \left(\left\|f^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} + \left\|g^0\right\|_{\mathcal{H}^1(dx,\mathcal{L}^1(dv))} \right),$$

where the right-hand side of the estimate on $\|g^n\|_{\mathcal{L}^\infty(dv,\mathcal{L}^2(dx))}$ *is set to* ∞ *whenever* $\mathfrak{f}^0 \notin \mathcal{H}^1(dx,\mathcal{L}^1(dv)) \times \mathcal{H}^1(dx,\mathcal{L}^1(dv))$.

7. Conclusions. We have proved uniform stability of the iterative scheme under the following two restrictions:

- Uniform bounds of the iterative scheme could be proven for underresolved numerical computations $\epsilon \le \epsilon_0 \Delta t$ or a bounded mean free path $\epsilon_1 \le \epsilon \le M$.
- The necessary CFL restriction is, in the diffusive limit, a parabolic CFL condition, as was expected. However, for finite values of ϵ the parabolic restriction can be relaxed. One obtains a CFL condition adapted to the hyperbolic part of the original kinetic equation.

Acknowledgment. We are grateful to Ed Larsen for interesting discussions and suggestions.

REFERENCES

[1] R. E. CAFLISCH, S. JIN, AND G. RUSSO, *Uniformly accurate schemes for hyperbolic systems with relaxation*, SIAM J. Numer. Anal., 34 (1997), pp. 246–281.

[2] F. GOLSE, S. JIN, AND D. LEVERMORE, *The convergence of numerical transfer schemes in diffusive regimes* II: *The cell edge even parity scheme*, in preparation.

[3] F. GOLSE, S. JIN, AND C. D. LEVERMORE, *The convergence of numerical transfer schemes in diffusive regimes* I: *Discrete-ordinate method*, SIAM J. Numer. Anal., 36 (1999), pp. 1333–1369.

[4] S. JIN, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454.

[5] S. JIN AND D. LEVERMORE, *Fully-discrete numerical transfer in diffusive regimes*, Transport Theory Statist. Phys., 22 (1993), pp. 739–791.

[6] S. JIN, L. PARESCHI, AND G. TOSCANI, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.

[7] S. Jin, L. Pareschi, and G. Toscani, *Uniformly accurate diffusive relaxation schemes for multiscale transport equations*, SIAM J. Numer. Anal., 38 (2000), pp. 913–936.
[8] A. Klar, *An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit*, SIAM J. Numer. Anal., 35 (1998), pp. 1073–1094.
[9] A. Klar, *An asymptotic-preserving numerical scheme for kinetic equations in the low mach number limit*, SIAM J. Numer. Anal., 36 (1999), pp. 1507–1527.
[10] A. Klar and C. Schmeiser, *Numerical passage from radiative heat transfer to nonlinear diffusion models*, Math. Models Methods Appl. Sci., 11 (2001), pp. 749–767.
[11] E. Larsen, J. Morel, and W. Miller, *Asymptotic solution of numerical transport problems in optically thick, diffusive regimes*, J. Comput. Phys., 69 (1987), pp. 283–324.
[12] C. Ringhofer, C. Schmeiser, and A. Zwirchmayr, *Moment methods for the semiconductor Boltzmann equation on bounded position domains*, SIAM J. Numer. Anal., 39 (2001), pp. 1078–1095.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 914–927

SYMMETRIC ERROR ESTIMATES FOR MOVING MESH GALERKIN METHODS FOR ADVECTION-DIFFUSION EQUATIONS*

TODD F. DUPONT† AND YINGJIE LIU‡

Abstract. This work tries to increase our understanding of why moving mesh methods often work very well. It combines techniques from the symmetric error estimates (SEEs) of Dupont [*Math. Comp.,* 39 (1982), pp. 85–107] and Bank and Santos [*SIAM J. Numer. Anal.,* 30 (1993), pp. 1–18] with ideas that motivated the analysis of a modified method of characteristics by Douglas and Russell [*SIAM J. Numer. Anal.,* 19 (1982), pp. 871–885]. By changing the usual time derivative to a time derivative along approximate characteristics in the SEE norm, the symmetric error estimate of Bank and Santos can be improved. In addition, by introducing yet another SEE norm which is more strongly mesh dependent, we provide another SEE which provides different insights into the convergence of these methods; one symmetric error estimate that is presented can be used to derive optimal order L^2 convergence in certain settings.

Key words. Galerkin methods, parabolic equations, finite element, moving mesh

AMS subject classifications. 65M60, 65M12

PII. S0036142900380431

1. Introduction.

Moving mesh finite element methods have been known for quite a while [6, 5] and they are increasingly used in practice, but the analytical understanding of these methods is far from complete.

A symmetric error estimate (SEE) is, roughly speaking, a statement of the following form: If the error *can be* small in a certain norm, then it *is* small in that same norm. Somewhat more precisely, there is a norm, $\|\!| \cdot |\!\|$, and a constant, C, such that

$$\|\!|error|\!\| \leq C \|\!|\text{best approximation error}|\!\|,$$

where the left-hand side measures the error in the method at hand and the right-hand side reflects the distance between the true solution and the function spaces used in the method. Of course, we need control on C if such an estimate is to be informative.

The results in section 2 of [4], section 3 of [1], and section 3 of this work give bounds of this type. There are two things that distinguish the bounds given here from earlier work. The first is that here the constant, C, does not increase as the advective term increases in size, provided the mesh movement approximates the advective term well in a sense that is made precise. Hence, these results make it more clear that the mesh movement is actually modeling the advection. The second is that the norm in section 3 involves the convective derivative instead of the partial with respect to time and, as Douglas and Russell pointed out in [3], for advection dominated problems the convective derivative will be much smoother, and therefore easier to approximate well. To give credit where it is due, Bank and Santos noted in [1] that in part of their analysis the constants could be made independent of the size of the advective term,

*Received by the editors November 2, 2000; accepted for publication (in revised form) November 24, 2001; published electronically August 1, 2002. This research was supported by the ASCI Flash Center at the University of Chicago under DOE contract B341495 and by the MRSEC Program of the National Science Foundation under award DMR-9808595.

http://www.siam.org/journals/sinum/40-3/38043.html

†Department of Computer Science, The University of Chicago, Chicago, IL 60637 (t-dupont@uchicago.edu).

‡Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, NY 11794 (yingjie@ams.sunysb.edu).

and they also noted the similarity between the difference equations and the modified method of characteristics [3].

While symmetric error estimates for parabolic equations have a certain attractiveness in the simplicity of the statement that they make, it is sometimes hard to see the precise meaning of the result. In the case of Galerkin methods for elliptic equations, one has a symmetric error estimate in the H^1-norm, a statement that is relatively easy to understand. In the case of parabolic equations, symmetric error estimates [2, 4, 1] involve combining several norms and seminorms; in the case of [4], for example, the $\|\!| \cdot |\!\|$ is constructed from two norms and a seminorm: the maximum in time of the L^2-norm in space, the L^2 in time norm of the H^1-norm in space, and the L^2 in time norm of the discrete H^{-1} seminorm in space of the time derivative. In one of the analogous results here, the H^1-norm is replaced by the "discrete H^1"-norm, i.e., the H^1-norm of the H^1-projection into the space. It might appear at first that weakening the norms is not an advantage, but it actually highlights the importance of the only remaining norm to such a degree that one can get optimal order L^2 convergence in some contexts. In a sense, the SEE that results from this norm provides a way to combine the techniques of [4] with those of Wheeler [7]. We view this as one of the most interesting results of this work.

In section 2 we give the advection-diffusion problem whose approximate solution we are studying here, and we define a continuous-time moving mesh method in terms of a "convected time derivative." In section 3 we give three symmetric error bounds for the continuous-time case. Then we present a symmetric error estimate for a discrete time case. In sections 4 and 5 we give two optimal order L^2 error bounds that follow from the results of section 3.

2. Model problem and a moving mesh Galerkin method. Consider the following advection-diffusion model problems on $Q = \Omega \times (0, T)$:

$$\begin{cases} \partial_t u - \bigtriangledown \cdot (a \bigtriangledown u) + v \cdot \bigtriangledown u + cu = f & \text{on } Q, \\ \dfrac{\partial u}{\partial \nu} = g & \text{on } \Gamma_N \times (0,T), \\ u = 0 & \text{on } \Gamma_D \times (0,T), \\ u = u_0 & \text{for } t = 0, \end{cases} \tag{2.1}$$

where $a(x,t), v(x,t), c(x,t), f(x,t)$, and $g(x,t)$ are smooth and bounded and $0 < a_0 \le a \le a_1$ for some constants $a_0, a_1 > 0$. Ω is a bounded domain in R^d. For simplicity, we assume that Ω is a fixed polyhedron. Γ_D, Γ_N are parts of the boundary $\partial\Omega$ such that $\Gamma_D \cap \Gamma_N = \emptyset$, $\Gamma_D \cup \Gamma_N = \partial\Omega$, and Γ_D is closed. Suppose that $\bar{D} = \cup D_i$ is a fixed polyhedron, where the D_i's are closed sets with nonvoid interior such that the interiors of the D_i's are disjoint. We need few restrictions on the D_i's for much of the argument, but to keep the discussion simple we suppose that each D_i is a simplex and that together they form a tessellation of D. We suppose that there is a continuous mapping $\mathcal{G}$ from $\bar{D} \times [0,T]$ onto $\bar{\Omega}$ such that (1) for each t, $\mathcal{G}(\cdot, t)$ is a one-to-one piecewise linear mapping (with respect to $\{D_j\}$) of $\bar{D}$ onto $\bar{\Omega}$; and (2) $\mathcal{G}$ is continuously differentiable on each $D_i \times [0,T]$. We also suppose that $\partial D = \gamma_D \cup \gamma_N$ and that $\Gamma_D = \mathcal{G}(\gamma_D, t)$ and $\Gamma_N = \mathcal{G}(\gamma_N, t)$. We denote by $\Omega_i = \Omega_i(t)$ the image of D_i under $\mathcal{G}(\cdot, t)$. Let $\mathcal{M}_D$ be a finite dimensional subspace of $H^1(D)$ so that each function in $\mathcal{M}_D$ vanishes on γ_D; then the finite element space on Ω is defined by $\mathcal{M}(t) = \{\phi(x,t) : \phi(\mathcal{G}(\cdot,t),t) \in \mathcal{M}_D\}$. It is sometimes convenient to think of this moving mesh as being generated by a mapping of Ω onto itself. (See Figure 2.1.)

Let $\mathcal{G}^{-1} = \mathcal{G}^{-1}(\cdot, t)$ denote the inverse of $\mathcal{G}$ as a map of $\bar{D}$ onto $\bar{\Omega}$ with t fixed;

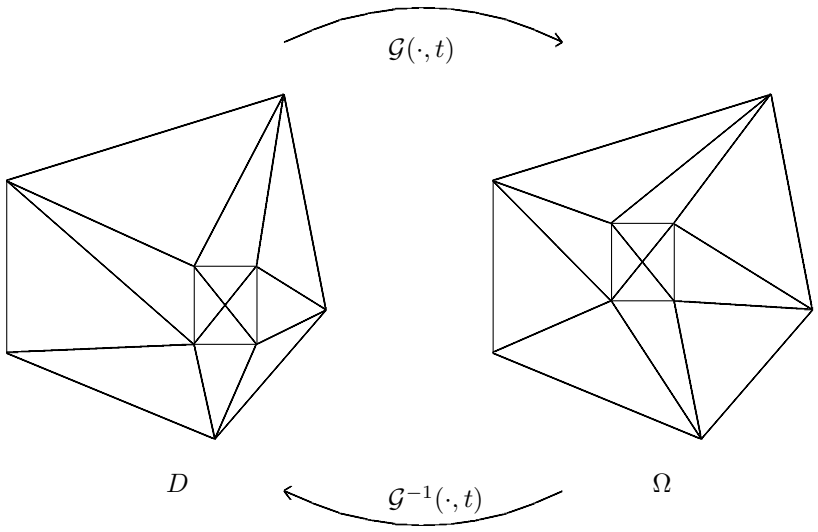


FIG. 2.1. *Moving mesh as a time dependent mapping* $\mathcal{G}$.

this function can be thought of as being defined on $\bar{Q}$. Let $\mathcal{G}_t$ be the partial derivative of $\mathcal{G}$ with respect to t. The finite element mesh is advected with a flow that is given by

$$\dot{x}(t) = \mathcal{G}_t(\mathcal{G}^{-1}(x,t),t).$$

Denote a particular directional derivative as follows:

$$\frac{D}{Dt}F(x,t) = \frac{\partial}{\partial t}F(x,t) + w \cdot \bigtriangledown_x F(x,t),$$

where $w(x,t)$ is a differentiable vector function such that $w \cdot \nu = 0$ on Γ_N for $t \geq 0$ and ν is the unit outer normal of $\partial\Omega$.

We will use $\|\cdot\|_k$ as the norm on the Sobolev space $H^k(\Omega)$; for domains R other than Ω we will use the more explicit notation $\|\cdot\|_{H^k(R)}$. The norm and inner product on $L^2(\Omega)$ will be denoted as $\|\cdot\|$ and $(\cdot,\cdot)$, respectively.

The exact solution of (2.1) will satisfy

$$\left(\frac{Du}{Dt},\psi\right) + (a \bigtriangledown u, \bigtriangledown\psi) + ((v-w)\cdot\bigtriangledown u,\psi) + (cu,\psi) = (f,\psi) + \int_{\Gamma_N} g\psi ds \tag{2.2}$$

for any $\psi \in H^1(\Omega)$. We are looking for $U \in \mathcal{M}(t)$ such that

$$\left(\frac{DU}{Dt},\phi\right) + (a \bigtriangledown U, \bigtriangledown\phi) + ((v-w)\cdot\bigtriangledown U,\phi) + (cU,\phi) = (f,\phi) + \int_{\Gamma_N} g\phi ds \tag{2.3}$$

for any $\phi \in \mathcal{M}(t)$. The inclusion of the convective derivative here is not really a change from the method discussed in [4]—we have just added and subtracted a term. However, it reflects a change in the way that we think about and analyze the method. We will take the initial value for U to be the L^2-projection of u_0 into $\mathcal{M}(0)$.

3. Symmetric error bounds. First, we get a basic relation that will be used in bounding the error. Taking $\Psi \in \mathcal{M}(t)$ and setting $\Phi = U - \Psi \in \mathcal{M}(t)$ and $\eta = u - \Psi$ give, for $\phi \in \mathcal{M}(t)$,

$$\begin{aligned}\left(\frac{D\Phi}{Dt}, \phi\right) + (a \bigtriangledown \Phi, \bigtriangledown\phi) + ((v-w)\cdot\bigtriangledown\Phi, \phi) + (c\Phi, \phi)\\ = \left(\frac{D\eta}{Dt}, \phi\right) + (a \bigtriangledown \eta, \bigtriangledown\phi) + ((v-w)\cdot\bigtriangledown\eta, \phi) + (c\eta, \phi).\end{aligned} \tag{3.1}$$

From the definition of the directional derivative we have the following equality which we use in the energy-type arguments used later.

LEMMA 1. *Suppose that $\phi(t) \in \mathcal{M}(t)$ and that ϕ is differentiable with respect to t as a map into $L^2(\Omega)$. Then*

$$\left(\frac{D\phi}{Dt}, \phi\right) = \frac{1}{2}\left\{\frac{d}{dt}\|\phi\|^2 - \int_\Omega \phi^2 \bigtriangledown_x \cdot w dx\right\}.$$

Proof.

$$\begin{aligned}\frac{d}{dt}\|\phi\|^2 &= 2\int_\Omega \phi_t\phi dx\\ &= 2\int_\Omega \frac{D\phi}{Dt}\phi dx - 2\int_\Omega (w\cdot\bigtriangledown\phi)\phi dx\\ &= 2\int_\Omega \frac{D\phi}{Dt}\phi dx + \int_\Omega \phi^2(\bigtriangledown\cdot w)dx - \int_{\Gamma_N}\phi^2 w\cdot\nu ds\\ &= 2\int_\Omega \frac{D\phi}{Dt}\phi dx + \int_\Omega \phi^2(\bigtriangledown\cdot w)dx. \qquad \square\end{aligned} \tag{3.2}$$

Define the mesh-dependent seminorm $\|\cdot\|_{(-1,\mathcal{M}(t))}$ by

$$\|u\|_{(-1,\mathcal{M}(t))} = \sup_{\phi\in\mathcal{M}(t),\phi\neq 0}\frac{|(u,\phi)|}{\|\phi\|_1}.$$

For X a normed space and v a function that maps $(0,T)$ into X, let

$$\|v\|_{L^p(0,T;X)}$$

denote the L^p-norm on the interval $(0,T)$ of the X-norm of v. The first SEE will be given in the norm $|||\cdot|||$ defined by

$$|||v|||^2 = \|v\|^2_{L^\infty(0,T;L^2(\Omega))} + \|v\|^2_{L^2(0,T;H^1(\Omega))} + \int_0^T \left\|\frac{Dv}{Dt}\right\|^2_{(-1,\mathcal{M}(t))} dt.$$

THEOREM 1. *Suppose that there exist constants c_1 and c_2 such that for all $(x,t) \in Q$,*

$$\bigtriangledown_x \cdot w(x,t) \le c_1, \qquad |w-v|(x,t) \le c_2. \tag{3.3}$$

Then there is a constant C depending only on c_1, c_2, T, and bounds on the coefficients a and c such that, for any smooth function Ψ from $[0,T]$ into $L^2(\Omega)$ with $\Psi(t) \in \mathcal{M}(t)$,

$$|||u - U||| \le C|||u - \Psi|||.$$

Proof. By using $\phi = \Phi$ in (3.1) we see then that

$$\frac{d}{dt}\|\Phi\|^2 + a_0\|\Phi\|_1^2 \le C\left\{\|\Phi\|^2 + \left\|\frac{D\eta}{Dt}\right\|^2_{(-1,\mathcal{M}(t))} + \|\eta\|_1^2\right\}. \tag{3.4}$$

This estimate and Gronwall's inequality give that

$$\|\Phi\|^2_{L^\infty(0,T;L^2(\Omega))} + a_0\|\Phi\|^2_{L^2(0,T;H^1(\Omega))} \le C\left(\|\Phi(0)\|^2_{L^2(\Omega)} + |||\eta|||^2\right). \tag{3.5}$$

Also, for any $\phi \in \mathcal{M}(t)$, (3.1) gives that

$$\left(\frac{D\Phi}{Dt}, \phi\right) \le C\left\{\|\Phi\|_1 + \left\|\frac{D\eta}{Dt}\right\|_{(-1,\mathcal{M}(t))} + \|\eta\|_1\right\}\|\phi\|_1. \tag{3.6}$$

Therefore

$$\int_0^T \left\|\frac{D\Phi}{Dt}\right\|^2_{(-1,\mathcal{M}(t))} dt \le C|||\eta|||^2. \tag{3.7}$$

Since $U(0)$ is the L^2-projection into $\mathcal{M}(0)$ of u_0, we see that $\|\Phi(0)\| \le \|\eta(0)\|$. Hence $|||\Phi||| \le C|||\eta|||$. The triangle inequality then gives that $|||u - U||| \le C|||u - \Psi|||$. ∎

In the application of Gronwall's inequality one gets exponential growth in time of the estimate of the error unless there is sufficient dissipation in the equation to counter it. If we let c_0 be a bound for the absolute value of $c(x,t)$ on Q, then the arithmetic of the proof gives that the constant C of (3.5) contains a factor $\exp(KT)$, where K can be of the form

$$K = 3c_0 + c_1 + c_2 + a_0/3 + 3c_2^2/a_0.$$

Hence, if c_2 is large and a_0 is small, this constant is very big. An interesting aspect of the above calculation is that most of it is local so that the important quantity for most parts of the estimate is the maximum of $|v - w|^2/a$. This would lead one to conjecture that in parts of the problem where diffusion is small, the directional derivatives that we bring into the estimation should be very close to the ones that point in characteristic directions.

The function w in the definition of the directional derivative should be chosen so that $|||u - \Psi|||$ in the above theorem is small. To illustrate how this might be done, we consider the case in which $\mathcal{M}(t)$ is the space of continuous piecewise linear functions over a triangular mesh given by the Ω_i's. If we take $w = \dot{x}$, then nodal interpolation commutes with the convective differentiation; i.e., $\frac{DIu}{Dt} = I\frac{Du}{Dt}$, where Iu is the nodal interpolant of u. Therefore,

$$\left\|\frac{D(u - Iu)}{Dt}\right\|_{(-1,\mathcal{M}(t))} \le \left\|\frac{D(u - Iu)}{Dt}\right\| \le C\left(\sum_i h_i^4 \left\|\frac{Du}{Dt}\right\|^2_{H^2(\Omega_i)}\right)^{\frac{1}{2}},$$

where h_i is the diameter of Ω_i. Here we emphasize that the norm involved is applied to the convective derivative, which can be a much smoother function than the usual partial time derivative.

Next we weaken the norm used in the previous theorem in two different ways to get somewhat different results.

Let $c_3 = (a_0 + c_2^2/a_0)/2$. Set

$$\mathcal{B}(\varphi, \psi) = (a \bigtriangledown (\varphi), \bigtriangledown \psi) + ((v - w) \cdot \bigtriangledown \varphi, \psi) + c_3(\varphi, \psi).$$

It is easy to check that for any $\varphi \in H^1(\Omega)$,

$$\mathcal{B}(\varphi, \varphi) \geq \frac{a_0}{2} \|\varphi\|_1^2. \tag{3.8}$$

We define a linear projection $P_1 : H^1(\Omega) \to \mathcal{M}(t)$ by

$$\mathcal{B}(v - P_1 v, \phi) = 0 \tag{3.9}$$

for all $\phi \in \mathcal{M}(t)$. Now we can define a new norm $\|\!| \cdot |\!\|_0$ in which the H^1 part of the previous norm has been weakened to be a seminorm:

$$\|\!| v |\!\|_0^2 = \|v\|^2_{L^\infty(0,T;L^2(\Omega))} + \|P_1 v\|^2_{L^2(0,T;H^1(\Omega))} + \int_0^T \left\| \frac{Dv}{Dt} \right\|^2_{(-1,\mathcal{M}(t))} dt. \tag{3.10}$$

The mnemonic for the use of the subscript 0 is that this norm emphasizes the H^0 or L^2 part of the norm.

Another norm also can be defined to put more weight on the $L^2(H^1)$ part of $\|\!| \cdot |\!\|$ by weakening the $L^\infty(L^2)$ part. Let P_0 be the L^2-projection onto $\mathcal{M}(t)$. Set

$$\|\!| v |\!\|_1^2 = \|P_0 v\|^2_{L^\infty(0,T;L^2(\Omega))} + \|v\|^2_{L^2(0,T;H^1(\Omega))} + \int_0^T \left\| \frac{Dv}{Dt} \right\|^2_{(-1,\mathcal{M}(t))} dt. \tag{3.11}$$

THEOREM 2. *If the conditions of Theorem* 1 *hold, then there is a constant* $C \geq 0$ *depending only on* c_1, c_2, T, *and bounds on the coefficients* a, c *such that for any smooth function* Ψ *from* $[0, T]$ *into* $L^2(\Omega)$ *with* $\Psi(t) \in \mathcal{M}(t)$,

$$\|\!| u - U |\!\|_0 \leq C \|\!| u - \Psi |\!\|_0,$$
$$\|\!| u - U |\!\|_1 \leq C \|\!| u - \Psi |\!\|_1.$$

Proof. Because the test function ϕ in (3.1) is in the space $\mathcal{M}(t)$, we can rewrite that relation as

$$\begin{aligned} &\left(\frac{D\Phi}{Dt}, \phi \right) + (a \bigtriangledown \Phi, \bigtriangledown \phi) + ((v - w) \cdot \bigtriangledown \Phi, \phi) + (c\Phi, \phi) \\ &= \left(\frac{D\eta}{Dt}, \phi \right) + \mathcal{B}(P_1 \eta, \phi) + ((c - c_3)\eta, \phi). \end{aligned} \tag{3.12}$$

This gives the following analogue of (3.4):

$$\frac{d}{dt} \|\Phi\|^2 + a_0 \|\Phi\|_1^2 \leq C \left\{ \|\Phi\|^2 + \left\| \frac{D\eta}{Dt} \right\|^2_{(-1,\mathcal{M}(t))} + \|P_1 \eta\|_1^2 + \|\eta\|^2 \right\}. \tag{3.13}$$

Since $P_1\Phi = \Phi$, this becomes

$$\frac{d}{dt}\|\Phi\|^2 + a_0\|P_1\Phi\|_1^2 \le C\left\{\|\Phi\|^2 + \left\|\frac{D\eta}{Dt}\right\|^2_{(-1,\mathcal{M}(t))} + \|P_1\eta\|_1^2 + \|\eta\|^2\right\}. \tag{3.14}$$

The estimate (3.6) becomes

$$\left(\frac{D\Phi}{Dt}, \phi\right) \le C\left\{\|P_1\Phi\|_1 + \left\|\frac{D\eta}{Dt}\right\|_{(-1,\mathcal{M}(t))} + \|P_1\eta\|_1 + \|\eta\|\right\}\|\phi\|_1. \tag{3.15}$$

The relations (3.14) and (3.15) give the bound for the $\|\!|\cdot|\!\|_0$ norm, just as in the proof of Theorem 1.

Examination of (3.4) shows that the η term in (3.5) can be replaced by $\|\!|\eta|\!\|_1$. The fact $P_0\Phi = \Phi$ gives

$$\|P_0\Phi\|^2_{L^\infty(0,T;L^2(\Omega))} + a_0\|\Phi\|^2_{L^2(0,T;H^1(\Omega))} \le C\left(\|P_0\Phi(0)\|^2_{L^2(\Omega)} + \|\!|\eta|\!\|_1^2\right).$$

Next, from (3.6) and the above relation we see that the analogue of (3.7) holds with $\|\!|\eta|\!\|$ replaced by $\|\!|\eta|\!\|_1$. Also, the use of $U(0) = P_0u(0)$ gives that $\|\Psi(0)\| = \|\eta(0)\|$. Combining these observations completes the proof of the second inequality in the theorem. □

Next we examine a fully discrete scheme. In this case we restrict ourselves to the case $w(x,t) = \dot{x}$. Following [1], for a given partition $P = \{t_0 = 0, t_1, \dots, t_{n-1}, t_n = T\}$ of $[0,T]$, consider $\mathcal{G}(s,t)$ to be linear in t for $t \in [t_{i-1}, t_i)$, for any i, and continuous in t on the whole of $[0,T]$. Let M be the collection of functions $\phi(x,t)$ on Q such that $\phi(\cdot,t) \in \mathcal{M}(t)$ for any $t \in [0,T]$, and such that ϕ is continuous in t and piecewise linear along the trajectory of mesh movement, i.e.,

$$\phi(\mathcal{G}(s,t),t) = \phi(\mathcal{G}(s,t_{j-1}),t_{j-1}) + \theta[\phi(\mathcal{G}(s,t_j),t_j) - \phi(\mathcal{G}(s,t_{j-1}),t_{j-1})],$$

where $t = t_{j-1} + \theta(t_j - t_{j-1})$ for any $\theta \in [0,1]$. The following relation holds for any $t \in (t_{i-1}, t_i)$, $s \in D_j$, for all i, j:

$$\frac{\phi(\mathcal{G}(s,t_i),t_i) - \phi(\mathcal{G}(s,t_{i-1}),t_{i-1})}{t_i - t_{i-1}} = \frac{D\phi}{Dt}(\mathcal{G}(s,t),t).$$

Note that $\frac{D\phi}{Dt}$ is just the same as in the continuous-time case on each (t_{i-1}, t_i) with the restriction that $w = \dot{x}$, but it also has a discrete form in this special case. It is clear that functions in M are defined by their values at the t_j's, so to define the approximate solution we need only say how it is computed at the times t_j.

The time discrete approximate solution $U \in M$ is such that $U(0)$ is the L^2-projection of $u(0)$ onto $\mathcal{M}(0)$ and, for $t = t_j-$,

$$\left(\frac{DU}{Dt}, \phi\right) + (a \bigtriangledown U, \bigtriangledown\phi) + ((v-w)\cdot\bigtriangledown U, \phi) + (cU, \phi) = (f,\phi) + \int_{\Gamma_N} g\phi ds \tag{3.16}$$

for any $\phi \in \mathcal{M}(t_j)$, $j = 1, 2, \dots, n$. Let

$$\|\!|v|\!\|^2_{0,d} = \max_{0\le j\le n}\|v(t_j)\|^2 + \sum_{j=1}^{n}(t_j - t_{j-1})\left\{\|P_1v(t_j)\|_1^2 + \left\|\frac{D}{Dt}v(t_j-)\right\|^2_{(-1,\mathcal{M}(t_j))}\right\};$$

we have the following theorem parallel to Theorem 3.1 in [1].

THEOREM 3. *Let $\mathcal{D}(t_j)$ denote the piecewise constant function $|\det(\bigtriangledown_s \mathcal{G})|$ on D. If there are constants $c_1, c_2 > 0$ independent of the mesh so that*

$$\frac{\mathcal{D}(t_j) - \mathcal{D}(t_{j-1})}{t_j - t_{j-1}} \leq c_1 \mathcal{D}(t_{j-1})$$

for $1 \leq j \leq n$, and $|w - v|(x,t) \leq c_2$ for all $(x,t) \in Q$, then there is a constant $C \geq 0$ depending only on c_1, c_2, T, and bounds of coefficients a, c such that $\|\!|u - U|\!\|_{0,d} \leq C\|\!|u - \Psi|\!\|_{0,d}$ for any $\Psi \in M$.

Proof. The fully discretized scheme yields an error similar to (3.12), with $t = t_j-$. Let $\phi = \Phi(t_j)$ in the analogue of (3.12), and use an argument like that in [1] to get

$$\begin{aligned}\left(\frac{D\Phi(t_j-)}{Dt}, \Phi(t_j)\right) \geq \quad & \frac{1}{2\Delta t_j}(\|\Phi(t_j)\|^2 - \|\Phi(t_{j-1})\|^2) \\ & -\frac{c_1}{2}\|\Phi(t_{j-1})\|^2,\end{aligned} \tag{3.17}$$

where $\Delta t_j = t_j - t_{j-1}$. We then have

$$\begin{aligned}&\frac{1}{\Delta t_j}(\|\Phi(t_j)\|^2 - \|\Phi(t_{j-1})\|^2) + \frac{1}{2}a_0\| \bigtriangledown \Phi(t_j)\|^2 \\ &\quad \leq C\left\{\left\|\frac{D\eta(t_j-)}{Dt}\right\|^2_{(-1,\mathcal{M}(t_j))} + \|P_1\eta(t_j)\|_1^2 + \|\Phi(t_j)\|^2 + \|\eta(t_j)\|^2\right\}.\end{aligned} \tag{3.18}$$

From the discrete Gronwall's inequality one obtains the following:

$$\|\Phi(t_j)\|^2 + \frac{1}{2}a_0 \sum_{i=1}^{j} \Delta t_i \| \bigtriangledown \Phi(t_i)\|^2 \leq C\|\!|\eta|\!\|^2_{0,d}. \tag{3.19}$$

Also, from the analogue to (3.12) at $t = t_j-$, for any $\phi \in \mathcal{M}(t_j)$,

$$\begin{aligned}\left(\frac{D\Phi(t_j-)}{Dt}, \phi(t_j)\right) \quad \leq \quad & C\Bigg\{\| \bigtriangledown \Phi(t_j)\| + \|\Phi(t_j)\| + \left\|\frac{D\eta(t_j-)}{Dt}\right\|_{(-1,\mathcal{M}(t_j))} \\ & + \|P_1\eta(t_j)\|_1 + \|\eta(t_j)\|\Bigg\}\|\phi(t_j)\|_1.\end{aligned} \tag{3.20}$$

With the help of (3.19), (3.20) becomes

$$\sum_{i=1}^{j} \Delta t_i \left\|\frac{D\Phi(t_i-)}{Dt}\right\|_{(-1,\mathcal{M}(t_i))} \leq C\|\!|\eta|\!\|_{0,d}. \tag{3.21}$$

Finally, combine (3.19), (3.21), and a triangle inequality to complete the proof. ∎

4. An optimal order L^2 error estimate. In this section, we prove the following optimal order error estimate for the one-space dimensional, continuous-time case: We will take $\mathcal{M}_D$ to be the space of continuous piecewise linear functions over a mesh $0 = s_0 < s_1 < \cdots < s_m = 1$ on the reference domain $D = [0,1]$, so $\mathcal{M}(t)$ is just the space of continuous functions which are polynomials of degree at most 1 on each interval $\Omega_i = [x_{i-1}, x_i]$, with $x_i(t) = \mathcal{G}(s_i, t)$. Take $w = \dot{x}$. Let h_i denote the length of Ω_i, and note that $\dot{x}$ is a continuous piecewise linear function over the mesh.

The following theorem gives an optimal order error estimate in which the error bound depends on the bounds of the difference between the growth rate of the length of each element with respect to time and the rate of "compression" of the exact solution (i.e., c_1), the difference between the convection velocity and the velocity of mesh movement (i.e., c_2), and other bounds of the coefficients of (2.1). Most importantly, the error bound does not depend on the convection velocity v, which shows an advantage of mesh movement.

THEOREM 4. *If there are constants $c_1, c_2, c_3 > 0$ so that $\|\partial_x(v - \dot{x})\|_\infty \le c_1$, $\|v - \dot{x}\|_\infty \le c_2$, and $\max_i \|\partial_x a\|_{L^\infty(\Omega_i)} \le c_3$ for all $t \in [0, T]$, then there is a constant $C(c_1, c_2, c_3, a_0, a_1, c, T; \Omega)$ such that*

$$\begin{aligned} \|u - U\|(t) \le \quad & C \left\{ \left\| \left(\sum_i h_i^4 \|u\|^2_{H^2(\Omega_i)} \right)^{1/2} \right\|_{L^\infty[0,T]} \right. \\ & \left. + \left\| \left(\sum_i h_i^4 \left\| \frac{Du}{Dt} \right\|^2_{H^2(\Omega_i)} \right)^{1/2} \right\|_{L^2[0,T]} \right\} \end{aligned} \tag{4.1}$$

for any $0 \le t \le T$.

Proof. The proof is an application of Theorem 2 using $\|\!| \cdot |\!\|_0$. Since $\|\!| u - U |\!\|_0$ dominates the term we want to bound, it suffices to show that $\|\!| u - \Psi |\!\|_0$ can be bounded by terms on the right-hand side of (4.1). We choose Ψ to be the nodal interpolant Iu of u. The estimate of $\|u - \Psi\|_{L^\infty(0,T;L^2(\Omega))}$ is straightforward. The observation that $\frac{D}{Dt}$ commutes with interpolation means that $\|\frac{D}{Dt}(u - \Psi)\|_{L^2(0,T;L^2(\Omega))}$ can be bounded by the terms on the right-hand side of (4.1); hence, the weaker semi-norm on $\frac{D}{Dt}(u - \Psi)$ is also bounded.

The $H^1(\Omega)$-norm of $P_1(u - \Psi)$ can be bounded as follows: For any $\phi \in \mathcal{M}(t)$,

$$\begin{aligned} & \mathcal{B}(P_1(u - \Psi), \phi) = \mathcal{B}(u - \Psi, \phi) \\ = \quad & \sum_i \int_{\Omega_i} a \partial_x (u - \Psi)\, \partial_x \phi dx + \sum_i \int_{\Omega_i} (v - \dot{x}) \partial_x (u - \Psi)\, \phi dx \\ & + c_3(u - \Psi, \phi) \\ = \quad & -\sum_i \int_{\Omega_i} (u - \Psi) \partial_x a\, \partial_x \phi dx - \sum_i \int_{\Omega_i} (u - \Psi) \{ \phi \partial_x (v - \dot{x}) \\ & + (v - \dot{x}) \partial_x \phi \} dx + c_3(u - \Psi, \phi) \\ \le \quad & \|\partial_x a\|_{L^\infty(\Omega)} \|u - \Psi\| \|\phi\|_1 + \|\partial_x (v - \dot{x})\|_{L^\infty(\Omega)} \|u - \Psi\| \|\phi\| \\ & + \|v - \dot{x}\|_{L^\infty(\Omega)} \|u - \Psi\| \|\phi\|_1. \end{aligned} \tag{4.2}$$

Using the coercivity of $\mathcal{B}(\cdot, \cdot)$ (see (3.8)) and taking $\phi = P_1(u - \Psi)$, we get that

$$\|P_1(u - \Psi)\|_1 \le C \|u - \Psi\|. \qquad \square$$

Note that the integration by parts was done subinterval by subinterval so a need only be locally smooth. The approximation results in this section are more local than we can prove in the general case studied in the next section.

5. Optimal order $L^2(\Omega)$ error estimate for general space dimension. In this section we return to the d-dimensional case. There will be several situations in which we need to use surface integrals on the elements Ω_i; we will use 2-dimensional terminology and refer to these as integrals over the edges. Thus an edge is the intersection of $\bar{\Omega}_i$'s with positive $(d-1)$-dimensional measure. Consider the Dirichlet problem, $\Gamma_N = \emptyset$, and take $w = \dot{x}$. Denote by e_j the edge between two adjacent elements and by n_{e_j} a normal to e_j, and define the jump operator $[\cdot]$ across the edge e_j by

$$[\mathcal{F}](x) = \lim_{\epsilon \to 0+} \{\mathcal{F}(x + \epsilon n_{e_j}) - \mathcal{F}(x - \epsilon n_{e_j})\} \quad \forall x \in e_j.$$

Assume that Ω and a are such that the Dirichlet problem has uniform H^2 regularity; i.e., there is a constant C such that for any $t \in [0,T]$, $q \in L^2(\Omega)$, there exists a $\xi \in H_0^1(\Omega) \cap H^2(\Omega)$ satisfying

$$\int_\Omega a \bigtriangledown \xi \cdot \bigtriangledown \eta dx = \int_\Omega q\eta dx \quad \forall \eta \in H_0^1(\Omega), \tag{5.1}$$

and $\|\xi\|_2 \leq C\|q\|$.

Suppose that $\mathcal{M}_D$ consists of a space of continuous piecewise polynomials of degree at most r. We assume that there is a constant $\tilde{C}$ such that for any $t \in [0,T]$ and for any $\xi \in H_0^1(\Omega) \cap \{\Pi_i H^s(\Omega_i)\}$, $s \geq 2$,

$$\inf_{\phi \in \mathcal{M}(t)} \|\xi - \phi\|_l^2 \leq \tilde{C} \sum_i h_i^{2(\min\{r+1,s\}-l)} \|\xi\|^2_{H^s(\Omega_i)}, \qquad l = 0, 1,$$

where h_i is the diameter of the element Ω_i. Let h denote $\max_i h_i$.

In this section we need bounds on $\bigtriangledown \dot{x}$, the Jacobian of the function $\dot{x}$ with respect to x. We will use the norm on matrices that is induced by the Euclidean norm on vectors. In particular, $\| \bigtriangledown \dot{x}\|_\infty$ is the $L^\infty(\Omega)$-norm of the norm of the matrix $\bigtriangledown \dot{x}$.

We have the following optimal order estimate for the $L^2(\Omega)$-norm of the error. While it looks like a generalization of Theorem 4 to higher dimensional spaces, there are differences. The hypotheses are stronger here, and the result is not quite so local.

THEOREM 5. *Suppose that there are constants $c_1, c_2, c_3, c_4, c_5 > 0$ so that, for all $t \in [0.T]$, we have $\| \bigtriangledown \dot{x}\|_\infty \leq c_1$; $\| \bigtriangledown \cdot v\|_\infty \leq c_2$; $\|v - \dot{x}\|_\infty \leq c_3$; and $\|\frac{Da}{Dt}\|_\infty, \| \bigtriangledown a\|_\infty, \| \bigtriangledown \frac{Da}{Dt}\|_\infty \leq c_4$; and the norm of the jump in $\bigtriangledown \dot{x}$ across an edge $e = \bar{\Omega}_k \cap \bar{\Omega}_m$ is bounded by $c_5 \min\{h_k, h_m\}$. Then there is a constant $C(c_1, c_2, c_3, c_4, c_5, a_0, a_1, c, T, \Omega)$ such that*

$$\begin{aligned} \|u - U\|(t) \quad \leq \quad & C \left\{ \left\| h \left(\sum_i h_i^{2(\min\{r+1,s\}-1)} \|u\|^2_{H^s(\Omega_i)} \right)^{1/2} \right\|_{L^\infty[0,T]} \right. \\ & \left. + \left\| h \left(\sum_i h_i^{2(\min\{r+1,s\}-1)} \left\| \frac{Du}{Dt} \right\|^2_{H^s(\Omega_i)} \right)^{1/2} \right\|_{L^2[0,T]} \right\} \end{aligned} \tag{5.2}$$

for any $t \in [0,T]$.

Proof. Again we will use Theorem 2 to establish this $L^2(\Omega)$ estimate. We will use an elliptic projection to give the Ψ that is in Theorem 2. The most tedious part of

the proof is bounding the time derivative part of $\|\!|\ |\!\|_0$; we do that here by estimating the $L^2(\Omega)$-norm of that term.

Set $\mathcal{B}_1(\xi,\eta) = (a \bigtriangledown \xi, \bigtriangledown\eta)$, and define a linear projection $P : H_0^1(\Omega) \to \mathcal{M}(t)$ by

$$\mathcal{B}_1(\xi - P\xi, \phi) = 0 \quad \forall \phi \in \mathcal{M}(t).$$

Denote $\eta = u - Pu$. For any given $t \in [0,T]$, let $\phi(x)$ be any function in $\mathcal{M}(t)$. Let $\psi(x,\tilde{t}) = \phi(\mathcal{G}(\mathcal{G}^{-1}(x,\tilde{t}),t))$ for any $\tilde{t} \in [0,T]$. It is easy to see that $\psi(x,t) = \phi(x)$ and $\frac{D\psi}{D\tilde{t}} = 0$ for any $\tilde{t} \in [0,T]$.

We have, at time t,

$$\begin{aligned} 0 &= \frac{d}{dt}\{\mathcal{B}_1(\eta(.,t),\psi(.,t))\} = \sum_i \left\{ \frac{d}{dt}\int_{\Omega_i} a \bigtriangledown \eta \cdot \bigtriangledown \psi dx \right\} \\ &= \sum_i \left\{ \int_{\Omega_i} a_t \bigtriangledown \eta \cdot \bigtriangledown \phi dx + \int_{\Omega_i} a \bigtriangledown \eta_t \cdot \bigtriangledown \phi dx \right. \\ &\quad \left. + \int_{\Omega_i} a \bigtriangledown \eta \cdot \bigtriangledown \psi_t dx + \int_{\partial\Omega_i} a \bigtriangledown \eta \cdot \bigtriangledown \phi (\dot{x}\cdot n) ds \right\}, \end{aligned} \tag{5.3}$$

where n is the outer norm of $\partial\Omega_i$. Note that

$$\int_{\Omega_i} a \bigtriangledown \eta_t \cdot \bigtriangledown \phi dx = \int_{\Omega_i} a \bigtriangledown \frac{D\eta}{Dt} \cdot \bigtriangledown \phi dx - \int_{\Omega_i} a \bigtriangledown (\dot{x}\cdot\bigtriangledown\eta)\cdot\bigtriangledown\phi dx$$

and

$$\begin{aligned} \dot{x}\cdot\bigtriangledown(\bigtriangledown\eta\cdot\bigtriangledown\phi) &= \sum_k \dot{x_k}\partial_{x_k}\left(\sum_j \partial_{x_j}\eta\partial_{x_j}\phi\right) \\ &= \sum_k\sum_j (\dot{x_k}\partial_{x_j}\partial_{x_k}\eta\partial_{x_j}\phi + \dot{x_k}\partial_{x_j}\partial_{x_k}\phi\partial_{x_j}\eta) \\ &= \bigtriangledown(\dot{x}\cdot\bigtriangledown\eta)\cdot\bigtriangledown\phi - (\bigtriangledown\eta)^T(\bigtriangledown\dot{x})(\bigtriangledown\phi) \\ &\quad + \bigtriangledown(\dot{x}\cdot\bigtriangledown\phi)\cdot\bigtriangledown\eta - (\bigtriangledown\phi)^T(\bigtriangledown\dot{x})(\bigtriangledown\eta). \end{aligned} \tag{5.4}$$

Using the fact that $0 = \frac{D\psi(x,t)}{Dt} = \psi_t(x,t) + \dot{x}\cdot\bigtriangledown\phi(x)$, we have

$$\begin{aligned} &\frac{d}{dt}\int_{\Omega_i} a \bigtriangledown \eta \cdot \bigtriangledown \psi dx \\ &= \int_{\Omega_i} \frac{Da}{Dt} \bigtriangledown \eta \cdot \bigtriangledown \phi dx + \int_{\Omega_i} a \bigtriangledown \left(\frac{D\eta}{Dt}\right)\cdot\bigtriangledown\phi dx + \int_{\Omega_i} a \bigtriangledown \eta\cdot\bigtriangledown\phi(\bigtriangledown\cdot\dot{x})dx \\ &\quad - \int_{\Omega_i} a(\bigtriangledown\eta)^T(\bigtriangledown\dot{x})(\bigtriangledown\phi)dx - \int_{\Omega_i} a(\bigtriangledown\phi)^T(\bigtriangledown\dot{x})(\bigtriangledown\eta)dx. \end{aligned} \tag{5.5}$$

Therefore we can write

$$\frac{d}{dt}\mathcal{B}_1(\eta,\psi) = \mathcal{B}_1\left(\frac{D\eta}{Dt},\phi\right) + E(\eta,\phi) = 0,$$

where

$$
\begin{aligned}
E(\eta,\phi) &= \int_\Omega \frac{Da}{Dt} \bigtriangledown \eta \cdot \bigtriangledown \phi dx + \int_\Omega a \bigtriangledown \eta \cdot \bigtriangledown \phi (\bigtriangledown \cdot \dot{x}) dx \\
&\quad - \int_\Omega a (\bigtriangledown \eta)^T \{ \bigtriangledown \dot{x} + (\bigtriangledown \dot{x})^T \} (\bigtriangledown \phi) dx.
\end{aligned} \tag{5.6}
$$

It is easy to see that $E(u,v) \le C\|u\|_1 \|v\|_1$, so

$$
\begin{aligned}
\left\| \frac{D\eta}{Dt} \right\|_1^2 &\le C\mathcal{B}_1 \left(\frac{D\eta}{Dt}, \frac{D\eta}{Dt} \right) \\
&= C \left\{ \mathcal{B}_1 \left(\frac{D\eta}{Dt}, \frac{D\eta}{Dt} - \phi \right) + E\left(\eta, \frac{D\eta}{Dt} - \phi \right) - E\left(\eta, \frac{D\eta}{Dt} \right) \right\} \\
&\le C \left\{ \left\| \frac{D\eta}{Dt} \right\|_1 \left\| \frac{D\eta}{Dt} - \phi \right\|_1 + \|\eta\|_1 \left\| \frac{D\eta}{Dt} - \phi \right\|_1 + \|\eta\|_1 \left\| \frac{D\eta}{Dt} \right\|_1 \right\}.
\end{aligned} \tag{5.7}
$$

It follows that

$$
\left\| \frac{D\eta}{Dt} \right\|_1 \le C \left\{ \|\eta\|_1 + \inf_{\phi \in \mathcal{M}(t)} \left\| \frac{Du}{Dt} - \phi \right\|_1 \right\}. \tag{5.8}
$$

Next we use a duality argument to get an estimate of $\|\frac{D\eta}{Dt}\|$. Let $\xi \in H_0^1(\Omega) \cap H^2(\Omega)$ satisfy

$$
\int_\Omega a \bigtriangledown \xi \cdot \bigtriangledown \zeta dx = \int_\Omega \frac{D\eta}{Dt} \zeta dx \quad \forall \zeta \in H_0^1(\Omega).
$$

For any $\phi \in \mathcal{M}(t)$,

$$
\begin{aligned}
\left(\frac{D\eta}{Dt}, \frac{D\eta}{Dt} \right) &= \mathcal{B}_1 \left(\frac{D\eta}{Dt}, \xi \right) \\
&= \mathcal{B}_1 \left(\frac{D\eta}{Dt}, \xi - \phi \right) + E(\eta, \xi - \phi) - E(\eta, \xi),
\end{aligned} \tag{5.9}
$$

and by integration by parts,

$$
\begin{aligned}
E(\eta,\xi) &= -\int_\Omega \eta \left(\bigtriangledown \frac{Da}{Dt} \cdot \bigtriangledown \xi + \frac{Da}{Dt} \bigtriangleup \xi \right) dx - \sum_i \int_{\Omega_i} \eta \bigtriangleup \xi (\bigtriangledown \cdot \dot{x}) dx \\
&\quad - \sum_j \int_{e_j} \eta \frac{\partial \xi}{\partial n_{e_j}} [\bigtriangledown \cdot \dot{x}] ds \\
&\quad + \sum_i \int_{\Omega_i} \eta \{ (\bigtriangledown a)^T (\bigtriangledown \dot{x}) (\bigtriangledown \xi) + a \bigtriangledown \cdot ((\bigtriangledown \dot{x})(\bigtriangledown \xi)) \} dx \\
&\quad + \sum_j \int_{e_j} a\, \eta ([\bigtriangledown \dot{x}](\bigtriangledown \xi)) \cdot n_{e_j} ds \\
&\quad + \sum_i \int_{\Omega_i} \eta \{ (\bigtriangledown a)^T (\bigtriangledown \dot{x})^T (\bigtriangledown \xi) + a \bigtriangledown \cdot ((\bigtriangledown \dot{x})^T (\bigtriangledown \xi)) \} dx \\
&\quad + \sum_j \int_{e_j} a\, \eta ([\bigtriangledown \dot{x}]^T (\bigtriangledown \xi)) \cdot n_{e_j} ds.
\end{aligned} \tag{5.10}
$$

We need to take a close look at the integrals over the edges. Suppose that an edge $e = \bar{\Omega}_k \cap \bar{\Omega}_m$. Let $h(e) = \min\{h_k, h_m\}$. The first boundary integral in (5.10) can be bounded as follows:

$$\begin{aligned}
\sum_j \int_{e_j} \eta \frac{\partial \xi}{\partial n_{e_j}} [\bigtriangledown \cdot \dot{x}] ds &\le C \sum_j \|h^{1/2}(e_j)\eta\|_{L^2(e_j)} \left\| h^{1/2}(e_j) \frac{\partial \xi}{\partial n_{e_j}} \right\|_{L^2(e_j)} \\
&\le C(\epsilon) \sum_j \|h^{1/2}(e_j)\eta\|^2_{L^2(e_j)} + \epsilon \sum_j \left\| h^{1/2}(e_j) \frac{\partial \xi}{\partial n_{e_j}} \right\|^2_{L^2(e_j)} \\
&\le C(\epsilon) \sum_i (\|\eta\|^2_{L^2(\Omega_i)} + h_i^2 |\eta|^2_{H^1(\Omega_i)}) + C\epsilon \sum_i (|\xi|^2_{H^1(\Omega_i)} + h_i^2 |\xi|^2_{H^2(\Omega_i)}) \\
&\le C(\epsilon) \left\{ \|\eta\|^2 + \sum_i h_i^2 |\eta|^2_{H^1(\Omega_i)} \right\} + C\epsilon \|\tfrac{D\eta}{Dt}\|^2 \quad \forall \epsilon > 0.
\end{aligned} \tag{5.11}$$

Similar results can be achieved for the other integrals over the edges e_j, so that by choosing ϵ small enough, we can conclude that

$$|E(\eta, \xi)| \le C \left\{ \|\eta\|^2 + \sum_i h_i^2 |\eta|^2_{H^1(\Omega_i)} \right\} + \frac{1}{4} \left\| \frac{D\eta}{Dt} \right\|^2.$$

Also choose $\phi \in \mathcal{M}(t)$ so that

$$\mathcal{B}_1 \left(\frac{D\eta}{Dt}, \xi - \phi \right) \le C \left\| \frac{D\eta}{Dt} \right\|_1 \|\xi - \phi\|_1 \le Ch \left\| \frac{D\eta}{Dt} \right\|_1 \left\| \frac{D\eta}{Dt} \right\|$$

and

$$E(\eta, \xi - \phi) \le Ch \|\eta\|_1 \left\| \frac{D\eta}{Dt} \right\|.$$

Therefore we have from (5.9),

$$\begin{aligned}
\left\| \frac{D\eta}{Dt} \right\|^2 &\le C \left\{ h^2 \left\| \frac{D\eta}{Dt} \right\|^2_1 + h^2 \|\eta\|^2_1 + \|\eta\|^2 + \sum_i h_i^2 |\eta|^2_{H^1(\Omega_i)} \right\} \\
&\le Ch^2 \sum_i h_i^{2(\min\{r+1,s\}-1)} \left\{ \left\| \frac{Du}{Dt} \right\|^2_{H^s(\Omega_i)} + \|u\|^2_{H^s(\Omega_i)} \right\}.
\end{aligned} \tag{5.12}$$

The rest of the proof is an application of Theorem 2 using $\|\!| \cdot |\!\|_0$. Since $\|\!| u - U |\!\|_0$ dominates the term we want to bound, it suffices to show that $\|\!| u - \Psi |\!\|_0$ can be bounded by terms on the right-hand side of (4.1).

We choose $\Psi = Pu$. The estimate of $\|u - \Psi\|_{L^\infty(0,T;L^2(\Omega))}$ is straightforward. The weaker seminorm on $\frac{D}{Dt}(u - \Psi)$ is also bounded from (5.12). The $H^1(\Omega)$-norm of $P_1(u - \Psi)$ can be bounded as follows: For any $\phi \in \mathcal{M}(t)$,

$$\begin{aligned}
& \mathcal{B}(P_1(u - \Psi), \phi) = \mathcal{B}(u - \Psi, \phi) \\
= \;& \mathcal{B}_1(u - \Psi, \phi) + ((v - \dot{x}) \cdot \bigtriangledown (u - \Psi), \phi) + c_3(u - \Psi, \phi) \\
= \;& -((u - \Psi), \phi \bigtriangledown \cdot (v - \dot{x}) + (v - \dot{x}) \cdot \bigtriangledown \phi) + c_3(u - \Psi, \phi) \\
\le \;& (\| \bigtriangledown \cdot (v - \dot{x}) \|_{L^\infty(\Omega)} + c_3) \|u - \Psi\| \|\phi\| \\
& + \|v - \dot{x}\|_{L^\infty(\Omega)} \|u - \Psi\| \|\phi\|_1.
\end{aligned} \tag{5.13}$$

Using the coercivity of $\mathcal{B}(\cdot,\cdot)$ (see (3.8)) and taking $\phi = P_1(u-\Psi)$, we get that

$$\|P_1(u-\Psi)\|_1 \le C\|u-\Psi\|. \qquad \square$$

The $\frac{D\eta}{Dt}$ term was estimated in $L^2(\Omega)$ instead of the discrete H^{-1} seminorm, so one might think that, if (5.1) satisfies an H^3-regularity bound and $\dot{x}$ was smooth enough, one might be able to weaken the norm on $\frac{D\eta}{Dt}$. We were not able to do this, except in trivial special cases.

6. Remarks. If we replace the boundary condition $w\cdot\nu = 0$ on Γ_N by $(w-\dot{x})\cdot\nu = 0$ on Γ_N, Lemma 1 holds even if the domain Ω is time dependent. Therefore it seems possible to get analogous results in this situation. However, a more interesting situation is one in which mesh elements flow into and out of the domain instead of just moving around in the domain; this will be the topic of future work.

REFERENCES

[1] R. E. Bank and R. F. Santos, *Analysis of some moving space-time finite element methods*, SIAM J. Numer. Anal., 30 (1993), pp. 1–18.

[2] J. Douglas Jr. and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.

[3] J. Douglas Jr. and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[4] T. Dupont, *Mesh modification for evolution equations*, Math. Comp., 39 (1982), pp. 85–107.

[5] K. Miller, *Moving finite elements.* II, SIAM J. Numer. Anal., 18 (1981), pp. 1033–1057.

[6] K. Miller and R. N. Miller, *Moving finite elements.* I, SIAM J. Numer. Anal., 18 (1981), pp. 1019–1032.

[7] M. F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 928–944

MULTIGRID SOLVER FOR THE INNER PROBLEM IN DOMAIN DECOMPOSITION METHODS FOR *P*-FEM*

SVEN BEUCHLER†

Abstract. From the literature it is known that the conjugate gradient method with domain decomposition preconditioners is one of the most efficient methods for solving systems of linear algebraic equations resulting from p-version finite element discretizations of elliptic boundary value problems. The ingredients of such a preconditioner are a preconditioner for the Schur complement, a preconditioner related to the Dirichlet problems in the subdomains, and an extension operator from the boundaries of the subdomains into their interior. In the case of Poisson's equation, we propose a preconditioner for the problems in the subdomains which can be interpreted as the stiffness matrix resulting from an h-version finite element discretization of a degenerate operator. For solving the corresponding systems of finite element equations a multigrid algorithm with a special line smoother is used. We prove that the convergence rate of the multigrid method is independent of the discretization parameter. The proof is based on the strengthened Cauchy inequality. The theoretical result is confirmed by numerical examples.

Key words. multigrid, FEM p-version

AMS subject classifications. 65F10, 65N22, 65N30, 65N55

PII. S0036142901393851

1. Introduction.

1.1. Origin of the problem from the *p*-version.

We consider the boundary value problem

$$\begin{array}{rcccc} -\triangle u & = & f & \text{in} & \Omega, \\ u & = & 0 & \text{on} & \partial\Omega, \end{array} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a domain which can be decomposed into (straight-line) quadrilaterals. Problem (1.1) will be discretized by means of the p-version of the finite element method using quadrilaterals R_s. Let $\mathcal{R} = (-1,1)^2$ be the reference element and $\Phi_s : \mathcal{R} \to R_s$ be the bilinear mapping to the element R_s. We define the finite element space

$$\mathbb{M} := \{u \in H_0^1(\Omega), u\mid_{R_s} = u(\Phi_s(\xi,\eta)) = \tilde{u}(\xi,\eta), \tilde{u} \in Q_p\},$$

where Q_p is the space of all polynomials $p(\xi,\eta) = p_1(\xi)p_2(\eta)$ of maximal degree p in each variable. Now we can formulate the following discretized problem: Find $u_p \in \mathbb{M}$ such that

$$a_\triangle(u_p, v_p) := \int_\Omega \nabla u_p \cdot \nabla v_p = \int_\Omega f v_p \quad \forall v_p \in \mathbb{M} \tag{1.2}$$

holds. Let $(\psi_1, \ldots, \psi_{n_p})$ be a basis of $\mathbb{M}$. Then problem (1.2) is equivalent to solving the system of algebraic finite element equations

$$A_p \underline{u}_p = \underline{f}_p, \quad \text{where} \quad A_p = [a_\triangle(\psi_j, \psi_i)]_{i,j=1}^{n_p}, \quad \underline{f}_p = \left[\int_\Omega f\psi_i\right]_{i=1}^{n_p}.$$

*Received by the editors August 16, 2001; accepted for publication (in revised form) February 14, 2002; published electronically August 8, 2002. This work was supported by the DFG Sonderforschungsbereich 393.

http://www.siam.org/journals/sinum/40-3/39385.html

†Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (sven.beuchler@mathematik.tu-chemnitz.de).

Now we specify the choice of the basis and divide the shape functions into three distinct groups:

- the vertex functions, which are the usual piecewise bilinear functions;
- the edge bubble functions;
- the interior bubbles, which are nonzero only on one element.

An edge bubble function corresponds to an edge e of the mesh. Its support is formed by those two elements which have this edge e in common. Corresponding to the division of the shape functions, we split the matrix A_p as follows:

$$A_p = \begin{pmatrix} A_{vert} & A_{vert,edg} & A_{vert,int} \\ A_{edg,vert} & A_{edg} & A_{edg,int} \\ A_{int,vert} & A_{int,edg} & A_{int} \end{pmatrix}. \tag{1.3}$$

The indices *vert*, *edg*, and *int* denote the blocks corresponding to the vertex, edge bubble, and interior bubble function, respectively. Jensen and Korneev [16] and Ivanov and Korneev [14], [15] developed preconditioners for the p-version of the finite element method in a two-dimensional domain using domain decomposition techniques [4]. They considered the preconditioning matrix

$$C_p = \begin{pmatrix} A_{vert} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{edg} & A_{edg,int} \\ \mathbf{0} & A_{int,edg} & A_{int} \end{pmatrix} \tag{1.4}$$

and proved that the condition number $\kappa(C_p^{-1}A_p)$ grows as $1+\log p$; cf. Lemma 2.3 in [14]. Therefore, the vertex unknowns can be determined separately. Computing the other unknowns, we factorize the remaining 2×2 block as follows:

$$\begin{pmatrix} A_{edg} & A_{edg,int} \\ A_{int,edg} & A_{int} \end{pmatrix} = \begin{pmatrix} I & A_{edg,int}A_{int}^{-1} \\ \mathbf{0} & I \end{pmatrix} \begin{pmatrix} S & \mathbf{0} \\ \mathbf{0} & A_{int} \end{pmatrix} \begin{pmatrix} I & \mathbf{0} \\ A_{int}^{-1}A_{int,edg} & I \end{pmatrix}$$

with the Schur complement $S := A_{edg} - A_{edg,int}A_{int}^{-1}A_{int,edg}$. The matrix A_{int} is a block diagonal matrix; one block corresponds to one element. Therefore, for computing the interior unknowns, we have to solve a Dirichlet problem on each quadrilateral. The edge unknowns are computed via the Schur complement S.

Hence, in addition to a solver for A_{vert}, we require three tools to define a preconditioner for the matrix of (1.4), namely a preconditioner for the interior problem, a preconditioner for the Schur complement, and an extension operator from the edges of a quadrilateral into its interior. Ivanov and Korneev [14], [15] derived some preconditioners C_S for the Schur complement. The condition number of $C_S^{-1}S$ is $\mathcal{O}(1+\log^2 p)$ in the worst case, where p is the polynomial degree. The solution of $C_S\underline{x} = \underline{y}$ can be done fast by solving triangular systems and fast Fourier transforms.

Jensen and Korneev [16] considered a scaled version of the integrated Legendre polynomials as the basis of the space $\mathbb{M}$. They found a spectral equivalent preconditioner C_{int} for the interior problem, which has $\mathcal{O}(p^2)$ nonzero entries. In the case of parallelogram elements, the element stiffness matrix has $\mathcal{O}(p^2)$ nonzero entries; otherwise it is a dense matrix. However, the suggested methods compute the solution of the system of equations with this preconditioning matrix in $\mathcal{O}(p^3)$ arithmetical operations. Finding a fast solver, i.e., a solver which requires $\mathcal{O}(p^2)$ arithmetical operations, was an open question. This paper is concerned with the construction of such an efficient preconditioner for the interior problem. The matrix C_{int} is a block diagonal

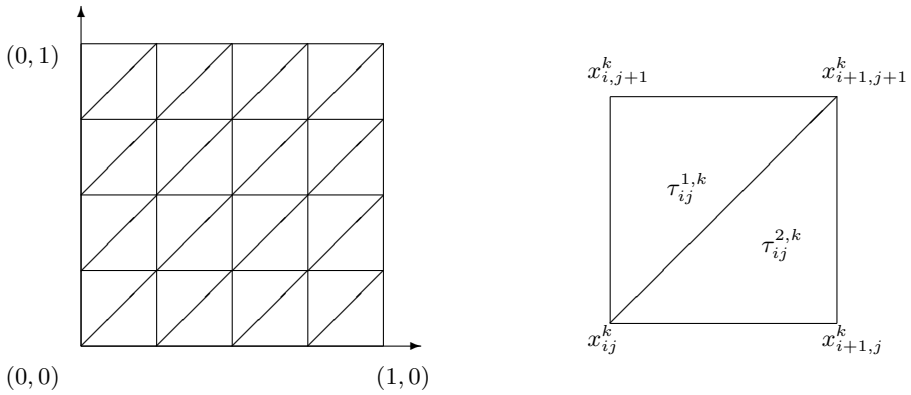


FIG. 1.1. *Mesh for h-version (left); notation within a macroelement $\mathcal{E}^k_{ij}$ (right).*

matrix of four blocks $C_{int,m}$, $m = 1, 2, 3, 4$. In [5], we derived a new preconditioner C_4 for each block of C_{int}. This matrix is defined via the matrices D_4 and T_3,

$$D_4 := \operatorname{diag}\left(4\left(i^2 + \frac{1}{6}\right)\right)_{i=1}^{n-1}, \quad T_3 := \frac{1}{2}\operatorname{tridiag}(-1, 2, -1),$$

$$C_4 := D_4 \otimes T_3 + T_3 \otimes D_4, \tag{1.5}$$

where $\otimes$ denotes the Kronecker product. We have proved in [5] that the condition number of the matrix $C_4^{-1}C_{int,m}$ grows as $(1 + \log p)$. We can interpret the matrix C_4 as the stiffness matrix arising from an h-version finite element method for an elliptic boundary value problem with a degenerate operator. This will be discussed in the following.

1.2. Formulation of the elliptic problem. We consider the following problem: Find $u \in H^1_0(\Omega_1) = \{u \in H^1(\Omega_1) : \ u = 0 \text{ on } \partial\Omega_1\}$ such that

$$a(u, v) := \int_{\Omega_1} y^2 u_x v_x + x^2 u_y v_y = \int_{\Omega_1} gv =: \langle g, v\rangle \quad \forall v \in H^1_0(\Omega_1) \tag{1.6}$$

holds. The domain $\Omega_1 = (0, 1)^2$ is the unit square. The differential operator in (1.6) is not uniformly elliptic in the Sobelev space $H^1(\Omega_1)$; an estimate of the type

$$a(u, u) \geq \gamma \parallel u \parallel^2_{H^1(\Omega_1)}$$

with a constant $\gamma > 0$ is not satisfied. We refer to the book of Kufner and Sändig [17] for such problems.

We want to find an approximate solution of (1.6) using finite elements. For this purpose, we introduce some notation. Let k be the level of approximation and $n = 2^k$. Let us denote by x^k_{ij} the nodes $x^k_{ij} = (\frac{i}{n}, \frac{j}{n})$, $i, j = 0, \ldots, n$. We triangulate Ω_1 into congruent, isosceles, right-angled triangles $\tau^{s,k}_{ij}$, where $0 \leq i, \ j < n$, and $s = 1, 2$; see Figure 1.1. The triangle $\tau^{1,k}_{ij}$ has the three vertices x^k_{ij}, $x^k_{i+1,j+1}$, and $x^k_{i,j+1}$, whereas $\tau^{2,k}_{ij}$ has the three vertices x^k_{ij}, $x^k_{i+1,j+1}$, and $x^k_{i+1,j}$; see Figure 1.1.

Furthermore, let $\mathcal{E}_{ij}^k = \overline{\tau_{ij}^{1,k} \cup \tau_{ij}^{2,k}}$ be the square (macroelement)

$$\left[\frac{i}{n}, \frac{i+1}{n}\right] \times \left[\frac{j}{n}, \frac{j+1}{n}\right].$$

Let $\mathbb{V}_k := \operatorname{span}\{\phi_{ij}^k,\ 0 < i, j < n\} \subset H_0^1(\Omega_1)$ be the subspace of piecewise linear functions. The basis functions ϕ_{ij}^k are continuous, linear on each triangle $\tau_{lm}^{s,k} \subset \Omega_1$, and fulfill the condition $\phi_{ij}^k(x_{lm}^k) = \delta_{il}\delta_{jm}$. ($\delta_{il}$ denotes the Kronecker delta.)

Now we can formulate the discrete problem. Find $u^k \in \mathbb{V}_k$ such that

$$a(u^k, v^k) = \langle g, v^k \rangle \quad \forall v^k \in \mathbb{V}_k \tag{1.7}$$

holds. Problem (1.7) is equivalent to solving the system of algebraic finite element equations $K_k \underline{u}_k = \underline{g}_k$ with

$$K_k = a(\phi_{lm}^k, \phi_{ij}^k)_{i,j;l,m=1}^{n-1}, \quad \underline{g}_k = \langle g, \phi_{lm}^k \rangle_{l,m=1}^{n-1}, \quad \underline{u}_k = (u_{ij})_{i,j=1}^{n-1}.$$

Then $u^k = \sum_{i,j=1}^{n-1} u_{ij}\phi_{ij}^k$ is the solution of (1.7). By a simple calculation one obtains

$$a(\phi_{ij}^k, \phi_{i+1,j}^k) = -\frac{1}{n^2}\left(\frac{1}{6} + j^2\right),$$

where $n > i, j$ and $j > 0$, but $i \geq 0$. By symmetry, we have ($i > 0$, $j \geq 0$)

$$a(\phi_{ij}^k, \phi_{i,j+1}^k) = -\frac{1}{n^2}\left(\frac{1}{6} + i^2\right)$$

and

$$a(\phi_{ij}^k, \phi_{ij}^k) = -\left(a(\phi_{ij}^k, \phi_{i+1,j}^k) + a(\phi_{ij}^k, \phi_{i,j+1}^k) + a(\phi_{ij}^k, \phi_{i,j-1}^k) + a(\phi_{ij}^k, \phi_{i-1,j}^k)\right).$$

All other matrix entries are zero. Inserting the boundary condition and using the definition of C_4 (1.5), we get the relation

$$K_k = \frac{1}{2n^2} C_4$$

after a proper permutation of the unknowns. In this paper, we will derive a fast solver for the degenerate problem (1.6).

For systems of finite element equations arising from the discretization of boundary value problems as, e.g., $-u_{xx} - u_{yy} = f$, efficient solution techniques are known. Examples for such solvers are the preconditioned conjugate gradient (PCG) method, with BPX preconditioners [9] or hierarchical basis preconditioners [23], and multigrid methods [12], [13]. However, we have to solve problems with variable coefficients. These coefficients tend to 0 if $x \to 0$ or $y \to 0$. Bramble and Zhang [10] consider multigrid methods in a more general case as for Laplace. They proved multigrid convergence for differential operators of the type $-(f(x,y)u_x)_x - (g(x,y)u_y)_y$, where $0 < g(x,y) \leq g_{\max}$ and $0 < f_{\min} < f(x,y) < f_{\max}$; i.e., one of the coefficients can be arbitrarily small. However, in (1.6) both coefficients can be arbitrarily small.

This paper deals with the solution of (1.7) by a multigrid method with special line smoothers. We get another efficient solver when these special line smoothers are applied within an AMLI preconditioner (see [2], [3] for the definition of the AMLI

preconditioner). In [6], we give a convergence proof for the PCG method with such an AMLI preconditioner. The analysis of both methods is based on the strengthened Cauchy inequality.

This paper is organized as follows. In section 2, we present the multigrid convergence proof for problem (1.7). In section 3 we describe and analyze the smoother used in the multigrid method and explain implementational details. Finally, in section 4 we give some numerical experiments confirming the theory.

2. Multigrid method for solving $-y^2u_{xx} - x^2u_{yy}$. For solving (1.7) approximately, we will employ the following multigrid algorithm.

2.1. Multigrid algorithm. We represent the space $\mathbb{V}_k$ as the direct sum

$$\mathbb{V}_k = \mathbb{V}_{k-1} \oplus \mathbb{W}_k, \quad \text{where} \quad \mathbb{W}_k := \operatorname{span}\{\phi_{ij}^k\}_{(i,j)\in N_k}. \tag{2.1}$$

The subset N_k contains the indices of the new nodes on level k and is given by

$$N_k := \{(i,j) \in \mathbb{N}^2, 1 \le i,j \le n-1, i = 2m+1 \text{ or } j = 2m+1, m \in \mathbb{N}\}. \tag{2.2}$$

Let u_0 be the initial guess. One step of the multigrid algorithm $u_1 = MULT(k, u_0, g)$ is defined recursively as follows.

- Set $l = k$. If $l > 1$, then do
  1. Presmoothing on $\mathbb{W}_l$: Solve
  $$a(w,v) = \langle g, v\rangle - a(u_0, v) \quad \forall v \in \mathbb{W}_l$$
  approximately by using ν steps of a simple iterative method S; the approximate solution is $\tilde{w}$. Set $u_0^1 = u_0 + \tilde{w}$.
  2. Coarse grid correction on $\mathbb{V}_{l-1}$: Find $w \in \mathbb{V}_{l-1}$ such that
  $$a(w,v) = \langle g, v\rangle - a(u_0^1, v) = \langle r, v\rangle \quad \forall v \in \mathbb{V}_{l-1}$$
  holds. Compute an approximate solution $\tilde{w}$ by using μ_{l-1} steps of the algorithm $MULT(l-1, 0, r)$. Set $u_0^2 = u_0^1 + \tilde{w}$.
  3. Postsmoothing on $\mathbb{W}_l$: Solve
  $$a(w,v) = \langle g, v\rangle - a(u_0^2, v) \quad \forall v \in \mathbb{W}_l$$
  approximately by using ν steps of a simple iterative method S; the approximate solution is $\tilde{w}$. Set $u_1 = u_0^2 + \tilde{w}$.
- else
  - Solve $a(w,v) = \langle g, v\rangle - a(u_0, v)\ \forall v \in \mathbb{V}_l$ exactly.
- endif.

Remark 2.1. In a standard multigrid algorithm the space $\mathbb{W}_l$ in steps 1 and 3 is replaced by $\mathbb{V}_l$; e.g., the smoother operates on the complete approximation space.

2.2. Algebraic multigrid proof. We prove the convergence of the multigrid algorithm for solving (1.7) using $\mu = \mu_l = 3$ and a special line smoother, which will be defined in (2.21). From [19], [20] the following convergence theorem for multigrid algorithms of the type of the algorithm $MULT$ is known.

THEOREM 2.1. *Let us assume that the following assumptions are fulfilled.*

- *Let $a(\cdot,\cdot)$ be a symmetric and positive definite bilinear form on $\mathbb{V}_k$, and let $\|\cdot\|_a^2 := a(\cdot,\cdot)$ be the energy norm.*

- *Let S be a smoother satisfying*

$$\| S^\nu w \|_a^2 \leq \rho^{2\nu} \| w \|_a^2 \quad \forall w \in \mathbb{W}_k, \tag{2.3}$$

where $0 \leq \rho < 1$ is independent of k.
- *There is a constant $0 \leq \gamma < 1$ independent of k such that the strengthened Cauchy inequality*

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall w \in \mathbb{W}_k, \forall v \in \mathbb{V}_{k-1} \tag{2.4}$$

holds.
- *Let $u_{j+1,k} = MULT(k, u_{j,k}, g)$, let u^* be the exact solution of (1.7), and let*

$$\sigma_k := \sup_{u_{j,k} - u^* \in \mathbb{V}_k} \frac{\| u_{j+1,k} - u^* \|_a}{\| u_{j,k} - u^* \|_a}$$

be the convergence rate of $MULT$ with ν smoothing operations.

Then, the following recursion formula holds:

$$\sigma_k \leq \sigma_{k-1}^{\mu_{k-1}} + (1 - \sigma_{k-1}^{\mu_{k-1}})(\rho^\nu + (1-\rho^\nu)\gamma)^2. \tag{2.5}$$

Proof. The proof is given by Theorem 2.2 of [20] with $\rho = \rho_1 = \rho_3$; see also Theorem 4 of [19]. ∎

The following lemma of the standard multigrid theory is helpful for the analysis of the recursion formula (2.5).

LEMMA 2.1. *Let $\mu_k = \mu \in \mathbb{N}$, $\mu > 1$, and*

$$\kappa := (\rho^\nu + (1-\rho^\nu)\gamma)^2 < \frac{\mu - 1}{\mu}. \tag{2.6}$$

Then the elements σ_k of the recursion

$$\sigma_0 := 0, \qquad \sigma_k := \sigma_{k-1}^\mu + (1 - \sigma_{k-1}^\mu)\kappa$$

are contained in the interval $[0, \sigma)$. The equation $\sigma = \kappa + \sigma^\mu(1-\kappa)$ has a solution $\sigma \in (0,1)$. More precisely, the sequence $\{\sigma_k\}_{k=0}^\infty$ is monotonically increasing and bounded from above by $\sigma < 1$ for $0 < \kappa < \frac{\mu-1}{\mu}$.

Proof. The proof can be found in several papers; see, e.g., Lemma 3 of [19] or Lemma 3.2 of [20]. ∎

Using Theorem 2.1 and Lemma 2.1, we can prove a mesh-size independent convergence rate in the case $\mu = 2$, i.e., the W-cycle, if $\kappa < \frac{1}{2}$.

If $\mu = 3$ and $\kappa < \frac{2}{3}$, from (2.6) it can be concluded that $\nu > \frac{\ln(\sqrt{\frac{2}{3}} - \gamma) - \ln(1-\gamma)}{\ln \rho}$ smoothing steps are needed if $\gamma^2 < \frac{2}{3}$. We want to prove multigrid convergence for system (1.7) via Theorem 2.1. For this aim, we have to determine bounds for ρ in (2.3) and γ^2 in (2.4). In the next subsection we summarize some lemmas which are helpful for our aim. Most of them are standard or trivial, and we refer to [1], [8], [11], [18], [21] or the preprint [7] for the proofs.

2.2.1. Basic definitions and helpful lemmas of the linear algebra. For proving the strengthened Cauchy inequality

$$(a(v,w))^2 \leq \gamma^2 a(v,v)a(w,w) \quad \forall v \in \mathbb{V}_{k-1}, w \in \mathbb{W}_k$$

with $\gamma^2 < 1$, we split $a(v,w)$ into

$$a(v,w) = \int_\Omega y^2 v_x w_x + x^2 v_y w_y = \sum_{i,j} \int_{\mathcal{E}^k_{i,j}} y^2 v_x w_x + x^2 v_y w_y =: \sum_{i,j} a^{\mathcal{E}^k_{i,j}}(v,w).$$

The angle between the spaces $\mathbb{V}_{k-1}$ and $\mathbb{W}_k$ can be estimated using the local angles.

LEMMA 2.2. *Let $a(\cdot,\cdot)$ be a symmetric, positive definite bilinear form. Under the assumption that*

$$\left(a^{\mathcal{E}^k_{ij}}(v,w)\right)^2 \leq \gamma^2 a^{\mathcal{E}^k_{ij}}(v,v) a^{\mathcal{E}^k_{ij}}(w,w)$$

for all $v \in \mathbb{V}_{k-1}\,|_{\mathcal{E}^k_{ij}}$ and $w \in \mathbb{W}_k\,|_{\mathcal{E}^k_{ij}}$ holds, we have

$$\left(a(v,w)\right)^2 \leq \gamma^2 a(v,v) a(w,w) \quad \forall v \in \mathbb{V}_{k-1}, w \in \mathbb{W}_k,$$

where $\mathbb{V}\,|_{\mathcal{E}^k_{ij}}$ denotes the restriction of $\mathbb{V}$ on $\mathcal{E}^k_{ij}$.

Proof. The proof is standard [8], [18]. ∎

The following lemma (see [11], [21]) relates the constant of the strengthened Cauchy inequality to the largest eigenvalue of a generalized eigenvalue problem. In order to formulate it, we need the following definition. Let $\mathbb{X}$ be a linear (finite-dimensional) space, and let $\mathbb{Y}$ be a subspace of $\mathbb{X}$. We define the difference $\mathbb{X} - \mathbb{Y}$ as any linear subspace satisfying

$$\mathbb{X} = \mathbb{Y} \oplus (\mathbb{X} - \mathbb{Y}).$$

Note that the choice of $\mathbb{X} - \mathbb{Y}$ is not unique.

LEMMA 2.3. *Consider the splitting $\mathbb{V} \oplus \mathbb{W}$. Let*

$$\mathbb{V} = \operatorname{span}\{\phi_i\}_{i=1}^n, \qquad \mathbb{W} = \operatorname{span}\{\psi_i\}_{i=1}^m,$$

$$G = [a(\phi_j,\phi_i)]_{i,j=1}^n, \quad H^t = [a(\phi_j,\psi_i)]_{i,j=1}^{n,m}, \quad K = [a(\psi_j,\psi_i)]_{i,j=1}^m.$$

Furthermore, let $\mathbb{V} \cap \mathbb{W} = \{\mathbf{0}\}$ and $\ker a \subset \mathbb{V}$, where $\ker a = \{v \in \mathbb{V} : a(v,w) = 0 \quad \forall w \in \mathbb{V}\}$ is the kernel of the bilinear form a. The bilinear form $a(\cdot,\cdot)$ is symmetric and positive semidefinite. Then the minimal constant γ^2 with

$$a(v,w)^2 \leq \gamma^2 a(v,v) a(w,w) \quad \forall v \in \mathbb{V}, w \in \mathbb{W}$$

is equal to the largest eigenvalue λ of

$$V^t H^t K^{-1} H V \underline{w} = \lambda V^t G V \underline{w},$$

where $V \in \mathbb{R}^{n,n-q}$ is chosen such that $\operatorname{im} V = \mathbb{R}^n - \ker G$ and $\ker V^t = \mathbf{0}$. The parameter q denotes the dimension of $\ker G$.

Proof. See [11], [21].

For the proof of the strengthened Cauchy inequality we need in our case an estimate for the eigenvalues of a 2×2 matrix. A useful tool is the next lemma.

LEMMA 2.4. *Let $M \in \mathbb{R}^{2,2}$ be a matrix with real eigenvalues and ϑ a real number with*

$$p = 2\vartheta - \operatorname{trace}(M) \geq 0 \quad \text{and} \quad q = \det M + \vartheta^2 - \vartheta \operatorname{trace}(M) \geq 0.$$

Then we have $\lambda_{max}(M) \leq \vartheta$.

Proof. The proof is trivial; cf. [7]. ∎

The following lemma (see [1], [22]) of the finite element analysis is helpful for the proof of the smoothing property (2.3). It analyzes the eigenvalue bounds of an assembled matrix by the eigenvalue bounds of the element matrices.

LEMMA 2.5. *Let* $\{A_i \in \mathbb{R}^{m_i,m_i}\}_{i=1}^s$ *be a finite set of symmetric positive definite matrices. Let* $A = \sum_{i=1}^s L_i^t A_i L_i$, *where* $L_i \in \mathbb{R}^{m_i,m}$ *and* $A \in \mathbb{R}^{m,m}$. *Furthermore, let* C_i *be a preconditioner for the matrix* A_i; *i.e., for all* $\underline{w} \in \mathbb{R}^{m_i}$ *the relations*

$$\underline{\lambda}_i(C_i\underline{w},\underline{w}) \leq (A_i\underline{w},\underline{w}) \leq \overline{\lambda}^i(C_i\underline{w},\underline{w})$$

with $0 < \overline{\lambda^i}$ *and* $0 \leq \underline{\lambda_i}$ *hold. Let* $C = \sum_{i=1}^s L_i^t C_i L_i$. *Then, for all* $\underline{v} \in \mathbb{R}^m$

$$\underline{\lambda}(C\underline{v},\underline{v}) \leq (A\underline{v},\underline{v}) \leq \overline{\lambda}(C\underline{v},\underline{v})$$

is valid with

$$\underline{\lambda} = \min_i \underline{\lambda}_i, \quad \overline{\lambda} = \max_i \overline{\lambda}^i.$$

2.2.2. Discussion of the strengthened Cauchy inequality on the macroelements $\mathcal{E}_{ij}^k$. We prove the strengthened Cauchy inequality (2.4) with the bilinear form $a(\cdot,\cdot)$ restricted on $\tau_{ij}^{1,k}$ and $\tau_{ij}^{2,k}$ if $i,j > 0$, and on the macroelements $\mathcal{E}_{ij}^k$ if $i = 0$ or $j = 0$.

We want to obtain the stiffness matrix on the macroelements $\mathcal{E}_{ij}^k$ with respect to the two level basis built by the basis functions of $\mathbb{V}_k \mid_{\mathcal{E}_{ij}^k}$ and $\mathbb{W}_{k+1} \mid_{\mathcal{E}_{ij}^k}$. We start with the introduction of the basis functions on $\mathcal{E}_{ij}^k$. Note that the triangle $\tau_{ij}^{2,k}$ is the union of the triangles $\tau_{2i,2j}^{2,k+1}$, $\tau_{2i+1,2j}^{1,k+1}$, $\tau_{2i+1,2j}^{2,k+1}$, and $\tau_{2i+1,2j+1}^{2,k+1}$, and the triangle $\tau_{ij}^{1,k}$ is the union of the triangles $\tau_{2i,2j}^{1,k+1}$, $\tau_{2i,2j+1}^{1,k+1}$, $\tau_{2i,2j+1}^{2,k+1}$, and $\tau_{2i+1,2j+1}^{1,k+1}$. The nodes x_{ij}^k, $x_{i,j+1}^k$, $x_{i+1,j}^k$, and $x_{i+1,j+1}^k$ are the coarse grid nodes, and the nodes $x_{2i+1,2j}^{k+1}$, $x_{2i,2j+1}^{k+1}$, $x_{2i+2,2j+1}^{k+1}$, $x_{2i+1,2j+2}^{k+1}$, and $x_{2i+1,2j+1}^{k+1}$ are new in the level $k+1$; compare Figure 2.1.

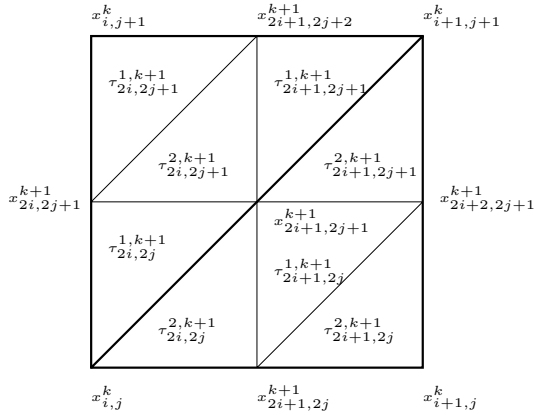


FIG. 2.1. *Local numbering of the nodes and subtriangles of* $\mathcal{E}_{ij}^k$.

Using this notation, we have

$$\mathbb{V}_k \mid_{\mathcal{E}_{ij}^k} = \operatorname{span}\{\phi_{lm}^k\}_{(l,m)\in N_{ij}^{\mathbb{V}_k}} \quad \text{and} \quad \mathbb{W}_{k+1} \mid_{\mathcal{E}_{ij}^k} = \operatorname{span}\{\phi_{lm}^{k+1}\}_{(l,m)\in N_{ij}^{\mathbb{W}_{k+1}}},$$

where

$$N_{ij}^{\mathbb{V}_k} := \{(l,m) \in \mathbb{N}^2, i \le l \le i+1, j \le m \le j+1\}$$

and with the help of N_{k+1} defined in (2.2),

$$N_{ij}^{\mathbb{W}_{k+1}} := N_{k+1} \cap \{(l,m) \in \mathbb{N}^2, 2i \le l \le 2i+2, 2j \le m \le 2j+2\}.$$

Note that for boundary macroelements $\mathcal{E}_{ij}^k$, i.e., with $i = 0, j = 0, i = n-1, j = n-1$, some modifications are necessary because of $\mathbb{V}_k \subset H_0^1(\Omega)$.

We define the matrices

$$\begin{aligned} G &:= \left[a^{\tau_{ij}^{2,k}}(\phi_{lm}^k, \phi_{rs}^k)\right]_{(r,s),(l,m)\in N_{ij}^{2,\mathbb{V}_k}}, \\ H^t &:= \left[a^{\tau_{ij}^{2,k}}(\phi_{lm}^k, \phi_{rs}^{k+1})\right]_{(r,s)\in N_{ij}^{2,\mathbb{V}_k},(l,m)\in N_{ij}^{2,\mathbb{W}_{k+1}}}, \\ K &:= \left[a^{\tau_{ij}^{2,k}}(\phi_{lm}^{k+1}, \phi_{rs}^{k+1})\right]_{(r,s),(l,m)\in N_{ij}^{2,\mathbb{W}_{k+1}}}, \end{aligned}$$

with $N_{ij}^{2,\mathbb{V}_k} := T_{ij}^2 \cap N_{ij}^{\mathbb{V}_k}$ and $N_{ij}^{2,\mathbb{W}_{k+1}} := T_{ij}^2 \cap N_{ij}^{\mathbb{W}_{k+1}}$, where $T_{ij}^2 := \{(l,m) \in \mathbb{N}^2, l-m \ge i-j\}$. The ordering of the rows and columns in the matrices G, H, and K corresponds to the ordering of the coarse grid nodes and of the new nodes introduced above. The matrices G, H, and K can be determined by a straightforward calculation. We start with the case $0 < i, j < n-1$. With

$$\begin{aligned} &a := \frac{48i^2+48i+14}{192n^2}, \quad b := \frac{48i^2+16i+2}{192n^2}, \quad c := \frac{48i^2+80i+34}{192n^2}, \\ &d := \frac{48j^2+48j+14}{192n^2}, \quad e := \frac{48j^2+16j+2}{192n^2}, \quad f := \frac{48j^2+80j+34}{192n^2} \end{aligned} \tag{2.7}$$

one obtains

$$\begin{aligned} H^t &= 2\begin{pmatrix} 0 & -d & d \\ a & d & -a-d \\ -a & 0 & a \end{pmatrix}, \qquad K = 4\begin{pmatrix} a+e & 0 & -a \\ 0 & c+d & -d \\ -a & -d & a+d \end{pmatrix}, \\ G &= \begin{pmatrix} d+e & -d-e & 0 \\ -d-e & a+c+d+e & -a-c \\ 0 & -a-c & a+c \end{pmatrix}. \end{aligned} \tag{2.8}$$

We note that in the case of elements laying on the boundary of the domain Ω_1, the matrices G, H, and K (2.8) are similarly defined, but we have only to cancel all rows and columns in G, H, and K which correspond to boundary nodes.

For choosing the matrix V of Lemma 2.3 we need $\ker G$. From the last relations it follows $\ker G = \operatorname{span}\{(1,1,1)^t\}$ and $\ker H = \operatorname{span}\{(1,1,1)^t\}$.

COROLLARY 2.1. *We have* $\ker G \subset \ker H$.

Now we determine the constant $\gamma_{\tau_{ij}^{2,k}}$, the constant of the strengthened Cauchy inequality on the triangle $\tau_{ij}^{2,k}$, by solving a generalized eigenvalue problem of the type (2.3).

LEMMA 2.6. *For* $1 \le i, j \le n-2$ *one obtains*

$$\left(a^{\tau_{ij}^{2,k}}(v,w)\right)^2 \le \gamma_{\tau_{ij}^{2,k}}^2 a^{\tau_{ij}^{2,k}}(v,v) a^{\tau_{ij}^{2,k}}(w,w) \quad \forall v \in \mathbb{V}_k\,|_{\tau_{ij}^{2,k}}, w \in \mathbb{W}_{k+1}\,|_{\tau_{ij}^{2,k}} \tag{2.9}$$

with $\gamma^2_{\tau^{2,k}_{ij}} = \frac{95}{176}$.

Proof. Corollary 2.1 states that $\ker G \subset \ker H$, and relation (2.8) states that $\ker K$ is trivial. Hence, we can apply Lemma 2.3. We know $\ker G = \operatorname{span}\{(1,1,1)^t\}$. Thus, we can choose

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{2.10}$$

The matrix V^tGV is symmetric and positive definite; the matrix $V^tH^tK^{-1}HV$ is symmetric. Therefore, the generalized 2×2 eigenvalue problem has real eigenvalues and is equivalent to the eigenvalue problem

$$(V^tGV)^{-1}V^tH^tK^{-1}HV\underline{x} = \lambda\underline{x}.$$

This is a 2×2 eigenvalue problem for which we can apply Lemma 2.4. With the help of a computer algebra system we computed the matrix

$$M := (V^tGV)^{-1}V^tH^tK^{-1}HV.$$

The choice $\gamma^2_{\tau^{2,k}_{ij}} = \frac{95}{176}$ yields

$$p = 2\gamma^2_{\tau^{2,k}_{ij}} - \operatorname{trace}(M) \geq 0 \text{ and } q = \det M + \gamma^4_{\tau^{2,k}_{ij}} - \gamma^2_{\tau^{2,k}_{ij}}\operatorname{trace}(M) \geq 0. \tag{2.11}$$

Using Lemmas 2.3 and 2.4 we infer (2.9). ∎

Note that the detailed proof of the lemma is very technical. The terms p and q of (2.11) are rational functions in i and j of the type

$$\frac{\sum_{s,r=0}^{6} b_{rs}(i-1)^r(j-1)^s}{\sum_{s,r=0}^{6} c_{rs}(i-1)^r(j-1)^s}$$

with $b_{rs} \geq 0$ and $c_{rs} > 0$ for all $r, s = 0, \ldots, 6$ and $i, j > 0$. Our aim was finding a constant $\gamma^2_{\tau^{2,k}_{ij}}$ as small as possible. The direct computation of the eigenvalues for the element $\tau^{2,k}_{1,1}$ leads to the value $\gamma^2_{\tau^{2,k}_{1,1}} = \frac{95}{176}$. Therefore, this constant cannot be improved.

The estimates presented above are valid for all $i, j > 0$. Therefore, the theory of Lemma 2.6 can be extended to the boundary elements at $x = 1$ or $y = 1$.

COROLLARY 2.2. *Let $i > 0$ and $j > 0$. The inequality*

$$\left(a^{\tau^{2,k}_{ij}}(v,w)\right)^2 \leq \gamma^2_{\tau^{2,k}_{ij}} a^{\tau^{2,k}_{ij}}(v,v)a^{\tau^{2,k}_{ij}}(w,w) \quad \forall v \in \mathbb{V}_k\,|_{\tau^{2,k}_{ij}}, w \in \mathbb{W}_{k+1}\,|_{\tau^{2,k}_{ij}} \tag{2.12}$$

is valid for $i = n-1$ or $j = n-1$ with $\gamma^2_{\tau^{2,k}_{ij}} \leq \frac{95}{176}$.

Proof. We start with $j = n-1$ and $0 < i < n-1$. We omit the unknown corresponding to $\phi^k_{i+1,j+1}$. Thus, we have to cancel the last row and column in the matrix G (2.8), which is the same as choosing the matrix V^tGV (2.10). In the case $i = n-1$ and $0 < j < n-1$, the unknowns corresponding to the second and last row and column in G and the third in K are omitted. ∎

By symmetry of the differential operator, the relation (2.12) is valid for $\tau_{ij}^{1,k}$, $0 < i, j \leq n-1$. Hence, the constant in the strengthened Cauchy inequality for the macroelements $\mathcal{E}_{ij}^k$, $0 < i, j \leq n-1$, fulfills (cf. Lemma 2.2)

$$\left(a^{\mathcal{E}_{ij}^k}(v,w)\right)^2 \leq \gamma_{\mathcal{E}_{ij}^k}^2 a^{\mathcal{E}_{ij}^k}(v,v) a^{\mathcal{E}_{ij}^k}(w,w) \quad \forall v \in \mathbb{V}_k \mid_{\mathcal{E}_{ij}^k}, w \in \mathbb{W}_k \mid_{\mathcal{E}_{ij}^k} \tag{2.13}$$

with $\gamma_{\mathcal{E}_{ij}^k}^2 = \frac{95}{176}$.

It remains to consider the case $i = 0$ or $j = 0$. The proof is similar and uses explicitly the Dirichlet boundary conditions at $x = 0$ and $y = 0$. We compute the constant of the strengthened Cauchy inequality directly on the macroelements $\mathcal{E}_{ij}^k$. We obtain

$$\gamma_{\mathcal{E}_{ij}^k}^2 < \frac{95}{176} \tag{2.14}$$

for $i = 0$ or $j = 0$. For details, see [7, Lemma 5.20 and Lemma 5.22].

Now we are able to formulate the main result of this subsection.

THEOREM 2.2. *The inequality*

$$(a(v,w))^2 \leq \gamma^2 a(v,v) a(w,w) \quad \forall v \in \mathbb{V}_k, w \in \mathbb{W}_{k+1}$$

is valid with $\gamma^2 = \frac{95}{176}$.

Proof. We apply Lemma 2.2. By inequalities (2.13) and (2.14) the assertion follows. ∎

2.2.3. Construction of the smoother. We need a good smoother for applying a multigrid solver to the linear system (1.7). This smoother will be constructed according to the local behavior of the differential operator. An idea presented in [1] for the construction of smoothers in the case of anisotropic problems is extended to the problem (1.6). This smoother operates only on the space $\mathbb{W}_{k+1}$. Consider the triangle $\tau_{ij}^{2,k}$. For our discussion only the submatrix K is needed, which corresponds to the nodal basis functions of $\mathbb{W}_{k+1}$. We discuss the two cases $i < j$ and $i \geq j$. Recall from (2.8) that

$$K_{2,ij} = 4 \begin{pmatrix} a+e & 0 & -a \\ 0 & c+d & -d \\ -a & -d & a+d \end{pmatrix}.$$

The indices i, j and 2 induce that this is the stiffness matrix with respect to $\mathbb{W}_k$ on the triangle $\tau_{ij}^{2,k}$; the index k is omitted. We now define a local preconditioner $C_{2,ij}$ for this matrix by omitting all those off-diagonal entries whose absolute values are small with respect to the main diagonals. We start with the case $i < j$. By (2.7), $a < d$ holds. Hence, we set

$$C_{2,ij} := 4 \begin{pmatrix} a+e & 0 & 0 \\ 0 & c+d & -d \\ 0 & -d & a+d \end{pmatrix}. \tag{2.15}$$

This matrix is a good preconditioner for $K_{2,ij}$ as the following lemma shows.

LEMMA 2.7. *For* $0 \leq i < j < n$ *one has*

$$\lambda_{min}(C_{2,ij}^{-1} K_{2,ij}) \geq 1 - \sqrt{\frac{1}{3}} \quad \textit{and} \quad \lambda_{max}(C_{2,ij}^{-1} K_{2,ij}) \leq 1 + \sqrt{\frac{1}{3}}.$$

Proof. The roots of the characteristical polynomial of $C_{2,ij}^{-1}K_{2,ij}$ are

$$\lambda_1 = 1, \quad \lambda_{2,3} = 1 \pm \sqrt{\rho}, \quad \text{where} \quad \rho = \frac{a}{a+e}\frac{ac+ad}{ac+ad+cd}.$$

We have for $i \le j-1$ that $\frac{c}{d} \le 1$ and $\frac{e}{a} \ge 1$. Therefore, using $a < c$, we obtain

$$\frac{ac+ad}{ac+ad+cd} \le \frac{2}{3} \quad \text{and} \quad \frac{a}{a+e} \le \frac{1}{2}. \tag{2.16}$$

Inserting the estimates (2.16), we obtain $1 - \sqrt{\frac{1}{3}} \le \lambda_3 \le \lambda_2 \le 1 + \sqrt{\frac{1}{3}}$. Hence, the assertion follows immediately. ∎

We now consider $i \ge j$. Then, $a > d$ holds, and we define

$$C_{2,ij} := 4 \begin{pmatrix} a+e & 0 & -a \\ 0 & c+d & 0 \\ -a & 0 & a+d \end{pmatrix}. \tag{2.17}$$

LEMMA 2.8. *It holds that*

$$\lambda_{min}(C_{2,ij}^{-1}K_{2,ij}) \ge 1 - \frac{1}{10}\sqrt{35} \quad \textit{and} \quad \lambda_{max}(C_{2,ij}^{-1}K_{2,ij}) \le 1 + \frac{1}{10}\sqrt{35}$$

for $n > i \ge j \ge 0$.

Proof. The proof is similar to the proof of Lemma 2.7. However, we have to distinguish the cases

- $i < n-1$ and $j > 0$, the same case as in Lemma 2.7;
- $i = n-1$ and $j > 0$, where $C_{2,ij}^{-1}K_{2,ij} = I$;
- $j = 0$, where we have to solve an eigenvalue problem with 2×2 matrices.

For details, see [7]. ∎

We define matrices $C_{1,ij}$ corresponding to the triangle $\tau_{ij}^{1,k}$ in the same way:

$$\begin{aligned} C_{1,ij} &:= 4 \begin{pmatrix} b+d & 0 & -d \\ 0 & a+f & 0 \\ -d & 0 & a+d \end{pmatrix} \quad \text{for} \quad i \le j, \\ C_{1,ij} &:= 4 \begin{pmatrix} b+d & 0 & 0 \\ 0 & a+f & -a \\ 0 & -a & a+d \end{pmatrix} \quad \text{for} \quad i > j. \end{aligned} \tag{2.18}$$

By the symmetry of the differential operator, we obtain the same results as in Lemmas 2.7 and 2.8.

Now we define a global preconditioner $C_{\mathbb{W}_{k+1}}$ using the local matrices $C_{s,ij}$. We know that

$$K_{\mathbb{W}_{k+1}} = a(\phi_{ij}^{k+1}, \phi_{lm}^{k+1})_{(i,j),(l,m)\in N_{k+1}}$$

is the stiffness matrix K_k restricted to the space $\mathbb{W}_{k+1}$ (compare (2.1), (2.2)). The matrix $K_{\mathbb{W}_{k+1}}$ is the result of assembling the local stiffness matrices $K_{s,ij}$ (2.8), $s = 1, 2$ and $i, j = 0, \ldots, n-1$, i.e.,

$$K_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^t K_{s,ij} L_{s,ij}. \tag{2.19}$$

The matrices $L_{s,ij} \in \mathbb{R}^{3\cdot 4^{k-1}-2^k,3}$ are the usual finite element assembling matrices. We define the matrix $C_{\mathbb{W}_{k+1}}$ by

$$C_{\mathbb{W}_{k+1}} = \sum_{s=1}^{2} \sum_{i,j=0}^{n-1} L_{s,ij}^t C_{s,ij} L_{s,ij}. \tag{2.20}$$

Because of the properties of the local preconditioners $C_{s,ij}$, the matrix $C_{\mathbb{W}_{k+1}}$ is a good preconditioner for $K_{\mathbb{W}_{k+1}}$. This result is stated as the main theorem of this subsection.

THEOREM 2.3. *It holds that*

$$\lambda_{min}(C_{\mathbb{W}_{k+1}}^{-1} K_{\mathbb{W}_{k+1}}) \geq 1 - \frac{1}{10}\sqrt{35}, \qquad \lambda_{max}(C_{\mathbb{W}_{k+1}}^{-1} K_{\mathbb{W}_{k+1}}) \leq 1 + \frac{1}{10}\sqrt{35}.$$

Proof. Use Lemmas 2.5, 2.7, and 2.8, and relations (2.19) and (2.20). ∎

Applying Theorem 2.3 an efficient smoother in the multigrid algorithm $MULT$ can be built as a preconditioned simple iteration method. The iteration operator of this method is defined by

$$S := I - \omega C_{\mathbb{W}_{k+1}}^{-1} K_{\mathbb{W}_{k+1}}. \tag{2.21}$$

COROLLARY 2.3. *Then, for all* $w \in \mathbb{W}_{k+1}$, $\nu \geq 1$, *and* $\omega = \omega_{opt} = 1$

$$\| S^\nu w \|_a \leq \rho^\nu \| w \|_a$$

holds with $\rho = \frac{1}{10}\sqrt{35}$.

Proof. The assertion follows by the theory of Jacobi smoothers; for details, see [7]. ∎

2.3. Application of the multigrid theory to $-x^2u_{yy} - y^2u_{xx} = g$. We now apply the theory of subsection 2.1 to problem (1.7). We state the main theorem of this paper.

THEOREM 2.4. *Consider* (1.7) *with the exact solution* u^*. *This linear system is solved by the multigrid algorithm* $MULT(k, u_{j,k}, g)$ *with* $\mu = 3$ *and* $\nu \geq 3$ *smoothing steps. Then the rate of convergence* σ_k *on level* k *can be bounded by*

$$\sigma_k \leq \sigma < 1.$$

Proof. We check the assumptions of Theorem 2.1. From Theorem 2.2 we have the estimate $\gamma^2 \leq \frac{95}{176}$ for the constant in the strengthened Cauchy inequality (2.4). The second assumption (2.3) is fulfilled for the smoother S defined in (2.21); cf. Corollary 2.3. Hence, we can prove a convergence rate $0 \leq \sigma < 1$ of the multigrid algorithm for $\mu \geq 3$.

Using Lemma 2.1, we can analyze the number of smoothing steps ν which is necessary for a convergence rate $\sigma < 1$. We have to show

$$\kappa = \rho^\nu + (1 - \rho^\nu)\gamma^2 < \frac{2}{3}$$

with $\gamma^2 = \frac{95}{176}$ and $\rho = \frac{1}{10}\sqrt{35}$ (Corollary 2.3). Therefore, we have for $\nu \geq 3$ a mesh-size independent convergence rate $\sigma < 1$. ∎

3. Implementational details. For applying the smoother S (2.21), the linear system

$$C_{\mathbb{W}_{k+1}} \underline{w} = \underline{r} \tag{3.1}$$

has to be solved. In this section, we explain an algorithm for solving (3.1). We want to show that this matrix is a block diagonal matrix of tridiagonal blocks. Furthermore, we show that the smoother S is a line smoother operating on lines L_{2m-1} which will be defined below. According to (2.20), (2.15), (2.17), and (2.18), the matrix $C_{\mathbb{W}_{k+1}}$ has the structure

$$C_{\mathbb{W}_{k+1}} = \operatorname{diag}(K_{\mathbb{W}_{k+1}}) + R,$$

where $\operatorname{diag}(K_{\mathbb{W}_{k+1}})$ is the diagonal part of the matrix $K_{\mathbb{W}_{k+1}}$ defined in (2.19). The matrix R will be defined below. Let $b : \mathbb{W}_k \times \mathbb{W}_k \to \mathbb{R}$ be the following bilinear form uniquely determined by the values of the basis functions $\{\phi_{ij}^k\}_{(i,j)\in N_k} \in \mathbb{W}_k$:

$$b(\phi_{ij}^k, \phi_{lm}^k) := \begin{cases} a(\phi_{ij}^k, \phi_{lm}^k) & \text{if} \quad \begin{array}{llll} i = l = 2r-1, & j = 2, \dots, i, & m = j-1, \\ j = m = 2r-1, & i = 2, \dots, j, & l = i-1, \end{array} \\ 0 & \text{otherwise} \end{cases}$$

for $r = 1, \dots, \frac{n}{2}$. Note that $a(\phi_{ij}^k, \phi_{lm}^k)$ is equal to the element $(i,j),(l,m)$ of the matrix K_k. The matrix R is defined via the bilinear form b. More precisely

$$R := \left[b(\phi_{ij}^k, \phi_{lm}^k) + b(\phi_{lm}^k, \phi_{ij}^k) \right]_{(i,j),(l,m)\in N_k};$$

i.e., the entries of the matrix R are those entries of the matrix $C_{\mathbb{W}_{k+1}}$ which correspond to edges marked by a bold line in Figure 3.1.

After a proper permutation the matrix $C_{\mathbb{W}_{k+1}}$ is a block diagonal matrix with diagonal and tridiagonal blocks. Therefore, we can solve the system (3.1) using Cholesky decomposition in $\mathcal{O}(n^2)$ flops. Hence, the operation $S\underline{w}$ is arithmetically optimal. We can choose $\nu = 3$ on each level. The number of unknowns of the system (1.7) increase per level to the factor 4. Therefore, using Theorem 2.4 the multigrid algorithm $MULT$ for $\mu = 3$ is an arithmetically optimal method [13].

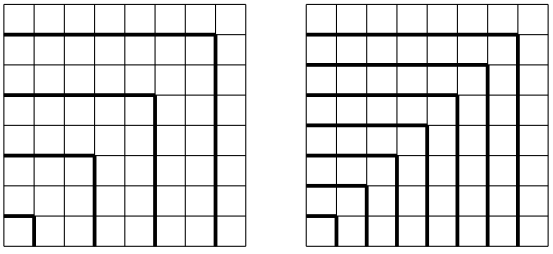


FIG. 3.1. *Nonzero entries of the matrices R (left) and $\tilde{R}$ (right).*

Additionally, a smoother S_1 is built, which uses the ideas of (2.21). This smoother operates on the space $\mathbb{V}_k$. Let $\tilde{C}_k := \operatorname{diag}(K_k) + \tilde{R}$, where

$$\tilde{R} := \left[\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) + \tilde{b}(\phi_{lm}^k, \phi_{ij}^k) \right]_{i,j;l,m=1}^{n-1},$$

with the bilinear form $\tilde{b} : \mathbb{V}_k \times \mathbb{V}_k \to \mathbb{R}$, where

$$\tilde{b}(\phi_{ij}^k, \phi_{lm}^k) := \begin{cases} a(\phi_{ij}^k, \phi_{lm}^k) & \text{if} \quad \begin{array}{rlll} & i = l = r, & j = 2, \dots, i, & m = j - 1, \\ \text{or} & j = m = r, & i = 2, \dots, j, & l = i - 1, \end{array} \\ 0 & \text{otherwise} \end{cases}$$

for $r = 1, \dots, n-1$. We now define the smoother

$$S_1 = I - \omega \tilde{C}_k^{-1} K_k. \tag{3.2}$$

In contrast to the smoother S (2.21) this smoother is a line smoother on each bold line of Figure 3.1. Therefore, we expect better convergence rates of a standard multigrid algorithm (cf. Remark 2.1) in contrast to S. The matrix $\tilde{C}_k$ is block tridiagonal. Therefore, $S_1 \underline{w} = \underline{r}$ can be solved arithmetically optimal using Cholesky decompostion in $\mathcal{O}(n^2)$ flops.

4. Numerical results.

TABLE 4.1
Convergence rates and number of iterations of multigrid algorithm MULT using smoother S ($\nu = 3$).

Level	$\mu = 1$		$\mu = 2$		$\mu = 3$		$\mu = 4$	
k	It	σ_k	It	σ_k	It	σ_k	It	σ_k
2	18	0.4070	18	0.4070	18	0.4070	18	0.4070
3	32	0.6017	24	0.4997	22	0.4778	22	0.4722
4	50	0.7239	25	0.5221	22	0.4698	21	0.4583
5	72	0.7974	27	0.5449	22	0.4770	21	0.4582
6	97	0.8463	30	0.5755	24	0.5035	22	0.4719
7	128	0.8814	34	0.6201	25	0.5156	22	0.4788
8	176	0.9123	37	0.6432	26	0.5282	23	0.4838
9	247	0.9373	41	0.6724	26	0.5339	23	0.4847
10	300	0.9545	44	0.6901	26	0.5380	23	0.4841

4.1. Convergence rate of multigrid. Table 4.1 shows the convergence rates of the multigrid algorithm $MULT$ for solving (1.7) with $g(x, y) \equiv 1$. We use several kinds of cycles. We stop the algorithm when the relative error measured in the $K_k^T K_k$ norm is less than 10^{-7}.

The V-cycle has clearly growing numbers of iterations, but for $\mu \geq 3$ we have mesh-independent convergence rates. It is not clear if the convergence rates for the W-cycle are bounded from above by $\sigma < 1$. The reason for the bad convergence of the V-cycle is the smoother S which operates only on the nodes on $\mathbb{W}_k$.

Additionally, we perform numerical experiments with multigrid algorithms using the smoother S_1, which we have defined in (3.2). Table 4.2 displays the convergence rates for this algorithm using the smoother S_1. We solve (1.7) with $g(x, y) \equiv 1$ and stop the algorithm when the error in the energy norm is less than 10^{-7}. We choose $\omega = 0.8$, which shows the best convergence rates. We obtain for the V- and W-cycle mesh-independent convergence rates.

4.2. Number of iterations for multigrid as preconditioner in the PCG method. We expect to obtain better convergence properties by using a PCG method with one multigrid cycle as preconditioner. Table 4.3 shows the number of iterations to reduce the error in the preconditioned energy norm by a factor 10^{-9}. We choose $g(x, y) \equiv 1$. We see constant numbers of iterations in two cases, V-cycle with smoother

TABLE 4.2

Convergence rates and number of iterations of a standard multigrid algorithm using smoother S_1 ($\nu = 1$, $\omega = 0.8$).

Level	$\mu = 1$		$\mu = 2$	
k	It	σ_k	It	σ_k
2	9	0.1611	9	0.1611
3	11	0.2290	10	0.1951
4	13	0.2723	12	0.2522
5	15	0.3250	14	0.2941
6	16	0.3517	15	0.3192
7	16	0.3619	15	0.3331
8	17	0.3680	15	0.3392
9	17	0.3720	16	0.3429
10	17	0.3750	16	0.3442

TABLE 4.3

Number of iterations of the PCG method using a multigrid preconditioner with smoother S ($\omega = 1$) *and* S_1 ($\omega = 0.8$) *and* $\nu = 1$.

Level	S			S_1
	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 1$
2	7	8	7	7
3	12	12	11	9
4	15	13	13	10
5	16	14	13	10
6	18	14	13	11
7	21	15	13	11
8	23	16	14	11
9	25	16	14	11

S_1 and for $\mu = 3$ with smoother S, but with a growing number of iterations for the V-cycle and smoother S. The results for the W-cycle ($\mu = 2$) give no precise answer as to whether the number of iterations are bounded or not.

5. Concluding remarks. Finally, we restate the two main results of this paper. A degenerate boundary value problem in the unit square was discretized by piecewise linear finite elements on a simple mesh. We constructed an arithmetically optimal multigrid method for the solution of the system of algebraic finite element equations arising from this discretization.

The second result concerns the application of this problem to the p-version of the finite element method. Consider (1.1) in $\Omega = (-1, 1)^2$. Discretize this problem by the p-version of the finite element method and use only one element. Then the blocks corresponding to the vertex functions and edge bubbles do not exist, i.e., $A_p = A_{int}$ in (1.3). We choose as preconditioner

$$\hat{C}_p := \text{blockdiag}\left[C_4[I - M]^{-1}\right]_{i=1}^{4},$$

where M denotes the iteration operator of our multigrid method. Then it follows from Theorem 3.4 in [7] and Theorem 2.4 (see also Theorem 7.1 in [7]) that the condition number of $\hat{C}_p^{-1}K_p$ grows as $1 + \log p$. Numerical experiments confirm the theory in this case [7]. Hence, we have found a preconditioner $\hat{C}_p$ for the interior problem of the p-version of the finite element method which is nearly optimal. Furthermore, the operation $\underline{r} = \hat{C}_p^{-1}\underline{w}$ can be done in $\mathcal{O}(p^2)$ arithmetical operations.

Acknowledgment. I thank Michael Jung for helpful comments.

REFERENCES

[1] O. Axelsson and A. Padiy, *On the additive version of the algebraic multilevel iteration for anisotropic elliptic problems*, SIAM J. Sci. Comput., 20 (1999), pp. 1807–1830.

[2] O. Axelsson and P.S. Vassilevski, *Algebraic multilevel preconditioning methods.* I, Numer. Math., 56 (1989), pp. 157–177.

[3] O. Axelsson and P.S. Vassilevski, *Algebraic multilevel preconditioning methods.* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[4] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[5] S. Beuchler, *A Preconditioner for Solving the Inner Problem of the p-Version of the FEM*, Technical Report SFB393 00-25, Technische Universität Chemnitz, Chemnitz, Germany, 2000.

[6] S. Beuchler, *AMLI preconditioner for the p-version of the FEM*, Numer. Linear Algebra Appl., submitted.

[7] S. Beuchler, *A Preconditioner for Solving the Inner Problem of the p-Version of the FEM, part* II—*Algebraic Multi-grid Proof*, Technical Report SFB393 01-07, Technische Universität Chemnitz, Chemnitz, Germany, 2001.

[8] D. Braess, *The contraction number of a multigrid method for solving the Poisson equation*, Numer. Math, 37 (1981), pp. 387–404.

[9] J. Bramble, J. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1991), pp. 1–22.

[10] J. Bramble and X. Zhang, *Uniform convergence of the multigrid v-cycle for an anisotropic problem*, Math. Comp., 70 (2001), pp. 453–470.

[11] G. Haase, U. Langer, and A. Meyer, *The approximate Dirichlet domain decomposition method. part* I: *An algebraic approach*, Computing, 47 (1991), pp. 137–151.

[12] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Heidelberg, 1985.

[13] W. Hackbusch and U. Trottenberg, *Multigrid Methods*, Lecture Notes in Math. 960 Springer-Verlag, Berlin, Heidelberg, New York, 1982.

[14] S.A. Ivanov and V.G. Korneev, *On the Preconditioning in the Domain Decomposition Technique for the p-Version Finite Element Method. Part* I, Technical Report SPC 95-35, Technische Universität Chemnitz-Zwickau, Chemnitz, Germany, 1995.

[15] S.A. Ivanov and V.G. Korneev, *On the Preconditioning in the Domain Decomposition Technique for the p-Version Finite Element Method. Part* II, Technical Report SPC 95-36, Technische Universität Chemnitz-Zwickau, Chemnitz, Germany, 1995.

[16] S. Jensen and V.G. Korneev, *On domain decomposition preconditioning in the hierarchical p−version of the finite element method*, Comput. Methods. Appl. Mech. Engrg., 150 (1997), pp. 215–238.

[17] A. Kufner and A.M. Sändig, *Some applications of weighted Sobolev spaces*, B.G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[18] J.F. Maitre and F. Musy, *The contraction number of a class of two-level methods; an exact evaluation for some finite element subspaces and model problems*, in Multigrid Methods, Lecture Notes in Math. 960, W. Hackbusch and U. Trottenberg, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1982, pp. 535–544.

[19] Ch. Pflaum, *Fast and Robust Multilevel Algorithms*, Habilitationsschrift, Universität Würzburg, Würzburg, Germany, 1998.

[20] N. Schieweck, *A multi-grid convergence proof by a strengthened Cauchy-inequality for symmetric elliptic boundary value problems*, in Second Multigrid Seminar, Rep. R-MATH 86-8, G. Telschow, ed., Akad. Wiss. DDR, Berlin, 1986.

[21] C.A. Thole, *Beiträge zur Fourieranalyse von Mehrgittermethoden: V-cycle, ILU-Glättung, anisotrope Operatoren*, Diplomarbeit, Universität Bonn, Bonn, Germany, 1983.

[22] A.J. Wathen, *An analysis of some element-by-element techniques*, Comput. Methods Appl. Mech. Engrg., 74 (1989), pp. 271–287.

[23] H. Yserentant, *On the multi-level-splitting of the finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 945–964

A FINITE DIFFERENCE SCHEME FOR SOME NONLINEAR DIFFUSION EQUATIONS IN AN ABSORBING MEDIUM: SUPPORT SPLITTING PHENOMENA*

TATSUYUKI NAKAKI† AND KENJI TOMOEDA‡

Abstract. In the porous media equation $v_t = (v^m)_{xx}$ $(m > 1)$, it is well known that there appears a finite propagation of the initial support and that, if the initial support is connected, so is $\operatorname{supp} v(t,\cdot)$ for $t > 0$. When the effect of the absorption is considered as the additional lower order term $-cv^p$ $(c > 0,\ p > 0)$, the possibility that the support will split is expected. Rosenau and Kamin [*Phys. D*, 8 (1983), pp. 273–283] tried the numerical computations and suggested the support splitting phenomena. But the theoretical justification is not discussed. In this paper, such phenomena are investigated by use of finite difference schemes, and the sufficient conditions under which the support begins to split are obtained in the specific case where $m + p = 2$ and $0 < p < 1$.

Key words. nonlinear diffusion, finite extinction, support splitting, interface, difference scheme

AMS subject classifications. 65M12, 35K65, 35B99

PII. S0036142900380303

1. Introduction.

We are concerned with the propagation of thermal waves in an absorbing medium occupying all of $\boldsymbol{R}^1$ in which there is an interaction between diffusion and absorption. The equation which describes this process is written in the form of the following initial value problem for the nonlinear diffusion equation with absorption which is used to describe the flow of liquids through porous media:

$$v_t = (v^m)_{xx} - cv^p, \qquad x \in \boldsymbol{R}^1, \quad t > 0, \tag{1.1}$$

$$v(0,x) = v^0(x), \qquad x \in \boldsymbol{R}^1, \tag{1.2}$$

where $m(> 1)$, $p(> 0)$, and $c(\geq 0)$ are constants, and $v^0(x) \in C^0(\boldsymbol{R}^1)$ is nonnegative and has compact support. In a heated plasma, v denotes the temperature and $-cv^p$ describes the losses caused by radiation. We may take $p = 0.5$ for bremsstrahlung radiation and $0.5 \leq p \leq 2$ for synchrotron radiation [24].

The existence and uniqueness of the nonnegative weak solution v of (1.1)–(1.2) was established by Oleinik, Kalashnikov, and Chzou [23], Kalashnikov [13], [14], Kersner [16], and Herrero and Vázquez [11]. The smoothness of v in the open set $\{(x,t); v(t,x) > 0 \text{ and } t > 0\}$ is also proved. Moreover, Kalashnikov proved that the solution becomes extinct in a finite time for $c > 0$ and $0 < p < 1$.

Since the diffusion rate mv^{m-1} vanishes at points where $v = 0$, the initial support propagates at finite speed; that is, there appear interface curves between the region where $v > 0$ and the region where $v = 0$. It is shown in the following papers that $\operatorname{supp} v(t,\cdot)$ exhibits three properties:

(i) **Positivity.** $\operatorname{supp} v(t,\cdot)$ expands as t increases and $\lim_{t\to\infty} \operatorname{supp} v(t,\cdot) = \boldsymbol{R}^1$, when $c = 0$, or $c > 0$ and $m \leq p$ [1], [3], [11], [13], [14], [18];

*Received by the editors November 1, 2000; accepted for publication (in revised form) January 4, 2002; published electronically August 8, 2002. This work was partially supported by the Japan Society for the Promotion of Science through Grant-in-Aid 11440035 and 13440038 for Scientific Research (B).

http://www.siam.org/journals/sinum/40-3/38030.html

†Faculty of Mathematics, Kyushu University, Higashi-ku, Fukuoka 812–8581, Japan (nakaki@math.kyushu-u.ac.jp).

‡Osaka Institute of Technology, Asahi-ku, Osaka 535–8585, Japan (tomoeda@ge.oit.ac.jp).

(ii) **Localization.** $\operatorname{supp} v(t,\cdot)$ expands as t increases and is uniformly bounded with respect to t, when $c > 0$ and $1 \le p < m$ [3], [10], [13], [14], [17], [18]. There exist constants M_j $(j = 1, 2)$ such that $\operatorname{supp} v(t,\cdot) \subset [M_1, M_2]$ for all $t \ge 0$;

(iii) **Total extinction.** $\operatorname{supp} v$ is compact in $[0,\infty) \times \boldsymbol{R}^1$, when $c > 0$, $m > 1$, and $0 < p < 1$. Thus $\operatorname{supp} v(t,\cdot)$ expands and/or shrinks and $v(t,x)$ becomes extinct in a finite time: $v(t,\cdot) \equiv 0$ for $t \ge T^*$, and $v(t,\cdot) \not\equiv 0$ for $t < T^*$, where $T^* > 0$ is some constant and is called the extinction time of v [13], [14], [15], [17].

When the solution v possesses the *positivity property* or the *localization property*, $\operatorname{supp} v(t,\cdot)$ never becomes disconnected, even if the initial function $v^0(x)$ has zeros in the interval (α_1, α_2), where $[\alpha_1, \alpha_2] = \operatorname{supp} v^0(x)$ ((A),(B)→(a) in Figure 1.1). When the solution v has the *total extinction property*, that is, absorption can cool the medium faster than diffusion supplies heat from hot area, the support shrinks and becomes disconnected ((A)→(b)). Moreover, there is the possibility of the support splitting into several disjoint sets, even if $v^0(x)$ is positive on (α_1, α_2) ((B)→(b)).

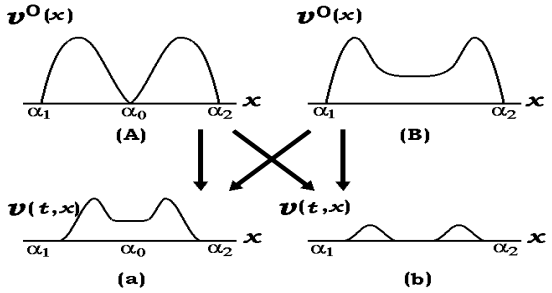


FIG. 1.1. *Support splitting and connecting.*

This motivates us to investigate the problem "How does the support vary when t varies?" To answer this problem we consider the behavior of $\operatorname{supp} v(t,\cdot)$ for the following two cases.

Case 1. $v^0(x) \in C^0(\boldsymbol{R}^1)$ is a nonnegative function with compact support $[\alpha_1, \alpha_2]$ and has only one zero $\alpha_0 \in (\alpha_1, \alpha_2)$; that is,

$$v^0(\alpha_0) = 0 \quad \text{and} \quad v^0(x) > 0 \quad \text{on} \quad (\alpha_1, \alpha_0) \cup (\alpha_0, \alpha_2). \tag{1.3}$$

Case 2. $v^0(x) \in C^0(\boldsymbol{R}^1)$ has compact support $[\alpha_1, \alpha_2]$ and satisfies

$$v^0(x) > 0 \quad \text{on } (\alpha_1, \alpha_2). \tag{1.4}$$

Rosenau and Kamin [24] tried numerical computations in Case 2, and obtained the numerical profile of v beginning to split into several subpulses, which means the splitting of the support. But they did not theoretically prove the appearance of such phenomena. From analytical points of view, Chen, Matano, and Mimura [4] constructed the solution $v(t,x)$ such that $\operatorname{supp} v(t,\cdot)$, while initially connected, splits into multiple connected components in a finite time. For Case 1 there are Kersner's results [15] which are useful to determine whether or not $\operatorname{supp} v(t,\cdot)$ becomes disconnected (see Remarks 5.1 and 5.2). In this paper, we treat such a problem from numerical points of view under the following assumption.

Assumption A. The constants m and p satisfy

$$m + p = 2 \quad \text{and} \quad 0 < p < 1. \tag{1.5}$$

To analyze the behavior of the support we have to construct the finite difference scheme of which both numerical solutions and numerical interfaces converge to the exact ones. Putting $u = v^{m-1}$, we rewrite (1.1)–(1.2) as

$$u_t = muu_{xx} + a(u_x)^2 - c', \quad a = \frac{m}{m-1}, \quad c' = (m-1)c, \tag{1.6}$$

$$u(0,x) = u^0(x) \equiv (v^0(x))^{m-1}. \tag{1.7}$$

We note that the effect of absorption is expressed as the constant $-c'$ under Assumption A.

Our difference scheme is constructed along the two important ideas which are proposed by Graveleau and Jamet [9] and DiBenedetto and Hoff [5]. Their difference schemes give the approximation to (1.6)–(1.7) with $c = 0$; that is, the porous media equation. Graveleau and Jamet's scheme is based on splitting the operator $Au = muu_{xx} + a(u_x)^2$ into two parts $u_t = Pu \equiv muu_{xx}$ and $u_t = Hu \equiv a(u_x)^2$. These two equations are discretized separately by the suitable difference schemes and solved alternately to approximate the solution of (1.6)–(1.7). They proved the convergence of numerical solutions and the finite propagation of support. However, since their difference scheme includes an artificial viscosity in approximating the nonlinear hyperbolic equation $u_t = Hu$, it does not give good approximations in realizing the interfaces. DiBenedetto and Hoff's scheme approximates (1.6)–(1.7) with the interface equation $\dot{\zeta}(t) = -\frac{m}{m-1}u_x(t,\zeta(t))$. They succeeded in proving the convergence of numerical solutions and numerical interfaces, and obtained error bounds. We note that their idea is applied to the approximation of the interfaces for a doubly degenerate parabolic equation [12].

Along the idea of Graveleau and Jamet, we construct the difference scheme consisting of three approximations to $u_t = Pu$, $u_t = Hu$, and $u_t = Du \equiv -c'$ in section 2. By solving Riemann problems included in $u_t = Hu$ and using the Rankine–Hugoniot jump equation, we propose the improved scheme so that an artificial viscosity is excluded. In sections 3 and 4 we succeed in proving the convergence of numerical solutions and numerical interfaces by using DiBenedetto and Hoff's argument. However, we are unable to obtain error bounds. In section 5, we partially improve Kersner's results in Case 1. Moreover, we obtain the sufficient condition under which the support begins to split into at least two disjoint sets in Case 2. Unfortunately, when $m + p \neq 2$, $m > 1$, and $0 < p < 1$, we are unable to construct the difference scheme satisfying the basic inequalities (Theorem 2.1) in approximating $u_t = Du$. However, our numerical computations show support splitting phenomena, which seems true. We need the improvement of our mathematical proofs to justify such phenomena. Numerical examples will be submitted for publication elsewhere.

On the other hand, Galaktionov [6], Galaktionov and Samarskii [7], [8], Kurdyumov, Kurkina, Malinetskii, and Samarskii [19], Kurdyumov, Gurevich, and Tel'kovskaya [20], and Samarskii, Galaktionov, Kurdyumov, and Mikhailov [25] established the existence of an unbounded solution for the equation with the effect of heat sources:

$$v_t = \nabla \cdot (v^m \nabla v) + v^p, \quad t > 0, \quad x \in \boldsymbol{R}^n.$$

They constructed self-similar solutions and difference solutions, and obtained the interesting properties of the behavior of the solution.

2. Difference schemes.

2.1. Difference operators. Our difference scheme approximates the problem (1.6)–(1.7) instead of (1.1)–(1.2), and is described as follows: Find the sequence $\{u_h^n\}_{n=1,2,\dots} \subset V_h$ for each $u_h^0 \in V_h$ such that

$$u_h^{n+1} = S_{h,k}u_h^n \equiv (P_{h,\kappa})^\mu \cdot \prod_{j=1}^{\nu} H_{h,\tau_j} \cdot D_{h,k}u_h^n \quad \text{for } n = 0, 1, 2, \dots. \tag{2.1}$$

Here $k \equiv k_{n+1} \equiv t_{n+1} - t_n$ ($t_0 = 0$), $\tau_j \equiv \tau_{n+1,j}$, and $\kappa \equiv \kappa_{n+1}$ are variable time steps, h is a space mesh width, $\mu \equiv \mu_{n+1}$ and $\nu \equiv \nu_{n+1}$ are positive integers satisfying

$$\sum_{j=1}^{\nu_{n+1}} \tau_{n+1,j} = \mu_{n+1}\kappa_{n+1} = k_{n+1}, \tag{2.2}$$

$D_{h,k}$, H_{h,τ_j}, and $P_{h,\kappa}$ are difference operators approximating $u_t = -c'$, $u_t = Hu$, and $u_t = Pu$, respectively, which are stated later, and V_h is the set of the nonnegative continuous functions $u_h = u_h(x)$ with the following properties:

(i) u_h has compact support with the left and right interfaces $\ell(u_h)$ and $r(u_h)$, respectively, which are defined by

$$\ell(u_h) = \sup\{\xi;\ u_h(x) = 0 \quad \text{on} \quad (-\infty, \xi]\}, \tag{2.3}$$

$$r(u_h) = \inf\ \{\xi;\ u_h(x) = 0 \quad \text{on} \quad [\xi, \infty)\}; \tag{2.4}$$

(ii) u_h is linear on each interval $[x_i, x_{i+1}]$ ($i \in \boldsymbol{Z}$), where $x_i = x_i(\ell, r)$ ($i \in \boldsymbol{Z}$),

$$x_i(\ell, r) = \begin{cases} ih & \text{for } i \in \boldsymbol{Z} \setminus \{L-1, R+1\}, \\ \ell & \text{for } i = L-1, \\ r & \text{for } i = R+1, \end{cases} \tag{2.5}$$

$$L \equiv L(\ell), \quad L(\ell) = \min\{i \in \boldsymbol{Z};\ ih > \ell\}, \quad \ell = \ell(u_h), \tag{2.6}$$

$$R \equiv R(r), \quad R(r) = \max\{i \in \boldsymbol{Z};\ ih < r\}, \quad r = r(u_h). \tag{2.7}$$

The determination of the variable time steps k, τ_j, and κ are stated in the description of $D_{h,k}$, H_{h,τ_j}, and $P_{h,\kappa}$, respectively. When $D_{h,k}u_h^{n^*} \equiv 0$ holds for some integer $n^* > 0$, we denote the numerical extinction time by $T_h^* = t_{n^*+1} \equiv t_{n^*} + k_{n^*+1}$ and stop the numerical computation by putting

$$t_n = T_h^* + (n - n^* - 1)k_{n^*+1} \ \text{ and } \ u_h^n(x) = 0 \quad \text{for all} \ \ n \geq n^* + 1. \tag{2.8}$$

We put $X(\ell, r) = \{x_i(\ell,\ r)\ ;\ i \in \boldsymbol{Z}\}$ and

$$\ell_n = \ell(u_h^n), \quad r_n = r(u_h^n), \quad L_n = L(\ell_n), \quad R_n = R(r_n) \quad (n = 0, 1, \dots, n^*). \tag{2.9}$$

For simplicity, we introduce the following notations, which will be used in the proof of Theorem 3.1 and the descriptions of Theorem 4.1 and Lemma 4.2:

$$\begin{cases} u_h^{n+1,s} = \left(\prod_{j=1}^{s} H_{h,\tau_{n+1,j}}\right) u_h^{n+1,0} \quad (s = 1, 2, \dots, \nu_{n+1}), \\ u_h^{n+1,0} = D_{h,k_{n+1}}u_h^n, \end{cases} \tag{2.10}$$

$$L_{n+1,s} = L(\ell_{n+1,s}), \quad \ell_{n+1,s} = \ell(u_h^{n+1,s}) \quad (s = 0, 1, \dots, \nu_{n+1}), \tag{2.11}$$

$$\begin{cases} u_h^{n+1,\cdot,q} = (P_{h,\kappa_{n+1}})^q u_h^{n+1,\cdot,0}\ (q = 1, 2, \dots, \mu_{n+1}), \\ u_h^{n+1,\cdot,0} = u_h^{n+1,\nu_{n+1}}, \end{cases} \tag{2.12}$$

$$L_{n+1,\cdot,q} = L(\ell_{n+1,\cdot,q}), \quad \ell_{n+1,\cdot,q} = \ell(u_h^{n+1,\cdot,q})\ (q = 0, 1, \dots, \mu_{n+1}). \tag{2.13}$$

Now we describe the difference operators $D_{h,k}$, $H_{h,\tau}$, and $P_{h,\kappa}$ by putting

$$\begin{array}{ll} h_i = x_{i+1} - x_i, & u_i = u_h(x_i), \\ \delta u_i = (u_{i+1} - u_i)/h_i, & \delta^2 u_i = 2(\delta u_i - \delta u_{i-1})/(h_i + h_{i-1}). \end{array}$$

2.1.1. Difference operator $D_{h,k}$. For $u_h \in V_h$ we define $D_{h,k}u_h$ by

$$D_{h,k}u_h(x_i') = \max\{u_h(x_i') - c'k, 0\} \text{ for all } x_i' \in X(\ell(D_{h,k}u_h), r(D_{h,k}u_h)), \tag{2.14}$$

$$\ell(D_{h,k}u_h) = \max\{\ell(u_h),\ (L'-1)h\}, \quad r(D_{h,k}u_h) = \min\{r(u_h),\ (R'+1)h\}, \tag{2.15}$$

where

$$L' = L(\ell(u(k,\cdot))), \quad R' = R(r(u(k,\cdot))). \tag{2.16}$$

We determine the time step k by either (2.17) with $i = L$ or (2.17) with $i = R-1$:

$$k = \frac{1}{c'}\max(u_i,\ u_{i+1}) \quad (i = L \text{ or } R-1). \tag{2.17}$$

2.1.2. Difference operator $H_{h,\tau}$. For $u_h \in V_h$ let $u(\tau, x)$ be the exact solution of $u_t = Hu$ with the initial value $u(0,x) = u_h(x)$. We define $H_{h,\kappa}u_h$ by

$$H_{h,\tau}u_h(x_i') = u(\tau, x_i') \quad \text{for all} \quad x_i' \in X(\ell(H_{h,\tau}u_h),\ r(H_{h,\tau}u_h)), \tag{2.18}$$

$$\ell(H_{h,\tau}u_h) = \ell(u(\tau,\cdot)), \qquad r(H_{h,\tau}u_h) = r(u(\tau,\cdot)). \tag{2.19}$$

The time step τ is the number satisfying

$$a\|(u_h)_x\|_\infty \tau \le \min\left\{\frac{h}{4},\ Lh - \ell(u_h),\ r(u_h) - Rh\right\}, \tag{2.20}$$

where $\|\cdot\|_\infty$ denotes $\|\cdot\|_{L^\infty(\boldsymbol{R}^1)}$. We put $\ell' = \ell(H_{h,\tau}u_h)$, $r' = r(H_{h,\tau}u_h)$, $L' = L(\ell')$, and $R' = R(r')$. Then, for $x_i' \in X(\ell',\ r')$, (2.18) and (2.19) are rewritten as

$$H_{h,\tau}u_h(x_i') = \begin{cases} u_i + a(\delta u_i)^2\tau & \text{if} \quad i \in S^+ = S_S^+ \cup S_R^+, \\ u_i + a(\delta u_{i-1})^2\tau & \text{if} \quad i \in S^- = S_S^- \cup S_R^-, \\ u_i & \text{if} \quad i \in S^0, \\ (L'h - \ell')\delta u_{L-1} & \text{if} \quad i = L' = L-1, \\ (R'h - r')\delta u_R & \text{if} \quad i = R' = R+1, \\ 0 & \text{if} \quad i \in \boldsymbol{Z} \setminus \{L', \dots, R'\}, \end{cases} \tag{2.21}$$

$$\ell' = \ell(u_h) - a\delta u_{L-1}\tau, \quad r' = r(u_h) - a\delta u_R\tau, \tag{2.22}$$

where

$$\begin{aligned} S_S^+ &= \Big\{i \in \{L, \dots, R\};\ \delta u_{i-1} < \delta u_i \quad \text{and} \quad \delta u_{i-1} > -\delta u_i\Big\}, \\ S_S^- &= \Big\{i \in \{L, \dots, R\};\ \delta u_{i-1} < \delta u_i \quad \text{and} \quad \delta u_{i-1} \le -\delta u_i\Big\}, \\ S_R^+ &= \Big\{i \in \{L, \dots, R\};\ \delta u_{i-1} \ge \delta u_i > 0\Big\}, \\ S_R^- &= \Big\{i \in \{L, \dots, R\};\ 0 > \delta u_{i-1} \ge \delta u_i\Big\}, \\ S^0 &= \Big\{i \in \{L, \dots, R\};\ \delta u_{i-1} \ge 0 \ge \delta u_i\Big\}. \end{aligned}$$

We apply the expression (2.21) to numerical computations, which we can show by integrating the solution $w(t,x)$ of the following Riemann initial value problem:

$$w_t = a(w^2)_x \quad \text{with the initial value} \quad w(0,x) = (u_h(x))_x. \tag{2.23}$$

Here (2.23) is derived from $u_t = Hu$ by putting $w = u_x$. The numerical interface equations (2.22) are well known as the Rankine–Hugoniot jump equation determining the line on which the solution w has a shock.

2.1.3. Difference operator $P_{h,\kappa}$. For $u_h \in V_h$ we define $P_{h,\kappa}u_h$ by the usual explicit difference operator

$$P_{h,\kappa}u_h(x_i') = u_i + \kappa m u_i \delta^2 u_i \quad \text{for all } x_i' \in X(\ell(P_{h,\kappa}u_h),\ r(P_{h,\kappa}u_h)), \tag{2.24}$$

$$\ell(P_{h,\kappa}u_h) = \ell(u_h), \quad r(P_{h,\kappa}u_h) = r(u_h). \tag{2.25}$$

The time step κ is the largest number satisfying

$$m\|u_h\|_\infty \kappa \Big\{ \frac{1}{h^2} + \frac{2}{h(h+h_j)} \Big\} \le 1 \quad \text{and} \quad \frac{4m\|(u_h)_x\|_\infty \kappa}{h+h_j} \le 1 \tag{2.26}$$

for $j = L-1$ and R.

2.2. Basic inequalities.

THEOREM 2.1 (basic inequalities). *Let $\{u_h^n\}_{n=1,2,\dots}$ be given by the scheme (2.1) for $u_h^0 \in V_h$. Then u_h^n either becomes extinct or belongs to V_h for each $n \ge 0$, and the following estimates hold for all $n \ge 0$:*

$$\ell_0 - a\|(u_h^0)_x\|_\infty t_n \le \ell_n \le r_n \le r_0 + a\|(u_h^0)_x\|_\infty t_n \text{ if } u_h^n \not\equiv 0, \tag{2.27}$$

$$0 \le u_h^n(x) \le \max(\|u_h^0\|_\infty - c' t_n,\ 0) \quad \text{on } \boldsymbol{R}^1, \tag{2.28}$$

$$\|(u_h^n)_x\|_\infty \le \|(u_h^0)_x\|_\infty, \tag{2.29}$$

$$TV((u_h^n)_x) \le TV((u_h^0)_x), \tag{2.30}$$

$$\|(u_h^{n+1} - u_h^n)/k_{n+1}\|_{L^1(\boldsymbol{R}^1)} \le (m+a)\|u_h^0\|_\infty TV((u_h^0)_x) + c' \left\{ r_0 - \ell_0 + 2a\|(u_h^0)_x\|_\infty t_n \right\}, \tag{2.31}$$

$$\inf_{i\in\boldsymbol{Z}} \delta^2 u_i^0 \le \inf_{i\in\boldsymbol{Z}} \delta^2 u_i^n, \tag{2.32}$$

where $TV(f)$ denotes the total variation of f on $\boldsymbol{R}^1$.

Proof. The inequalities (2.27)–(2.32) with $c = 0$ immediately follow from Lemmas 3.1, 3.2, and 5.1 in [21]. Put $u_h' = D_{h,k}u_h$ ($u_h \in V_h$). By simple calculations, it can be shown that

$$\|u_h'\|_\infty \le \max(\|u_h\|_\infty - c'k, 0),\ \|((u_h'))_x\|_\infty \le \|(u_h)_x\|_\infty,\ TV((u_h')_x) \le TV((u_h)_x),$$
$$\|(u_h' - u_h)/k\|_{L^1(\boldsymbol{R}^1)} \le c' \{r(u_h) - \ell(u_h)\},\ \inf_{i\in\boldsymbol{Z}} \delta^2 u_i \le \inf_{i\in\boldsymbol{Z}} \delta^2 u_i'.$$

Hence (2.27)–(2.32) hold for $c > 0$ by these inequalities and (2.15). ∎

2.3. Numerical extinction time. We show the existence of the numerical extinction time.

THEOREM 2.2 (existence of numerical extinction time). *Let $\{u_h^n\}_{n=1,2,\dots}$ be given by the scheme (2.1) for $u_h^0 \in V_h$. Then there exists an integer $N(h)$ such that $t_{N(h)}$ attains the numerical extinction time T_h^*, and the following estimate holds:*

$$T_h^* \le t_n + \frac{\|u_h^n\|_\infty}{c'} \quad \text{for all } n \ge 0. \tag{2.33}$$

Proof. We prove the existence of $N(h)$ such that $t_{N(h)} = T_h^*$. Assume the contrary; that is, suppose $u_h^n > 0$ for all $n \geq 0$. Then, it follows from (2.28) that

$$t_n < \frac{\|u_h^0\|_\infty}{c'} \quad \text{for all } n \geq 0. \tag{2.34}$$

Let Q_ℓ and Q_r be the sets of nonnegative integers defined by

$$Q_\ell = \{n; k = k_{n+1} \text{ satisfies (2.17) with } i = L\} \text{ and}$$
$$Q_r = \{n; k = k_{n+1} \text{ satisfies (2.17) with } i = R-1\},$$

respectively. From the determination of the numerical interfaces (see (2.15)) it follows that

$$\ell'_n - \ell_n \geq h \quad \text{for } n \in Q_\ell \quad \text{and} \quad r_n - r'_n \geq h \quad \text{for } n \in Q_r, \tag{2.35}$$

where $\ell'_n = \ell(D_{h,k_{n+1}} u_h^n)$ and $r'_n = r(D_{h,k_{n+1}} u_h^n)$. By using (2.22), (2.29), and (2.35), we have

$$\ell_{n+1} - \ell_n = \ell_{n+1} - \ell'_n + \ell'_n - \ell_n \geq \begin{cases} -a\|(u_h^0)_x\|_\infty k_{n+1} + h & \text{for } n \in Q_\ell, \\ -a\|(u_h^0)_x\|_\infty k_{n+1} & \text{for } n \notin Q_\ell, \end{cases} \tag{2.36}$$

$$r_{n+1} - r_n = r_{n+1} - r'_n + r'_n - r_n \leq \begin{cases} a\|(u_h^0)_x\|_\infty k_{n+1} - h & \text{for } n \in Q_r, \\ a\|(u_h^0)_x\|_\infty k_{n+1} & \text{for } n \notin Q_r. \end{cases} \tag{2.37}$$

Since $\{\ell_n\}$ and $\{r_n\}$ are bounded sequences by (2.27) and (2.34), we can extract convergent subsequences $\{\ell_{n_s}\}$ and $\{r_{n_s}\}$. We find from (2.36) and (2.37) that

$$\ell_{n_{s+1}} - \ell_{n_s} = \sum_{j=n_s}^{n_{s+1}-1} (\ell_{j+1} - \ell_j) \geq -a\|(u_h^0)_x\|_\infty \sum_{j=n_s}^{n_{s+1}-1} k_{j+1} + h \quad \text{for } n_s \in Q_\ell,$$
$$r_{n_{s+1}} - r_{n_s} = \sum_{j=n_s}^{n_{s+1}-1} (r_{j+1} - r_j) \leq \quad a\|(u_h^0)_x\|_\infty \sum_{j=n_s}^{n_{s+1}-1} k_{j+1} - h \quad \text{for } n_s \in Q_r,$$

which yield

$$a\|(u_h^0)_x\|_\infty \sum_{j=n_s}^{n_{s+1}-1} k_{j+1} \geq \begin{cases} \ell_{n_s} - \ell_{n_{s+1}} + h & \text{for } n_s \in Q_\ell, \\ r_{n_{s+1}} - r_{n_s} + h & \text{for } n_s \in Q_r. \end{cases}$$

Then there exists an integer s_0 such that

$$a\|(u_h^0)_x\|_\infty \sum_{j=n_s}^{n_{s+1}-1} k_{j+1} \geq \frac{h}{2} \qquad \text{for } s \geq s_0,$$

which implies

$$t_{n_q} - t_{n_{s_0}} = \sum_{s=s_0}^{q-1} \sum_{j=n_s}^{n_{s+1}-1} k_{j+1} \geq \frac{(q-s_0)h}{2a\|(u_h^0)_x\|_\infty} \qquad \text{for } q \geq s_0 + 1. \tag{2.38}$$

Hence $\lim_{q\to+\infty} t_{n_q} = +\infty$, which contradicts (2.34). Thus we have $t_{N(h)} = T_h^*$ for some integer $N(h)$.

Finally we show (2.33). From (2.28) we obtain

$$\|u_h^m\|_\infty \le \max\{\|u_h^n\|_\infty - c'(t_m - t_n),\ 0\} \quad \text{for all } m > n,$$

which immediately yields

$$\|u_h^m\|_\infty = 0 \quad \text{for} \quad t_m \ge t_n + \frac{\|u_h^n\|_\infty}{c'}.$$

Hence (2.33) holds, and the proof is complete. □

3. Convergence of numerical solutions. In this section we state the convergence of the numerical solutions. For this end, we start the scheme (2.1) with $u_h^0 \in V_h$ given by

$$\begin{cases} \ell(u_h^0) = \ell(u^0), \quad r(u_h^0) = r(u^0), \\ u_h^0(x_i) = u^0(x_i) \qquad \text{for all } i \in \boldsymbol{Z}. \end{cases} \tag{3.1}$$

We define the function $u_h(t,x)$ by

$$u_h(t,x) = u_h^n(x) \quad \text{on} \quad [t_n, t_{n+1}) \times \boldsymbol{R}^1 \qquad \text{for all } \ t_n \text{ and } h \tag{3.2}$$

and impose Condition B on v^0.

Condition B. $u^0 \equiv (v^0)^{m-1} \in C^0(\boldsymbol{R}^1)$ is a nonnegative function with compact support and $u_x^0 \in L^\infty(\boldsymbol{R}^1) \cap BV(\boldsymbol{R}^1)$.

From Theorem 2.1 we have the following convergence theorem.

THEOREM 3.1 (convergence of numerical solutions u_h). *Assume Condition* B. *Let $\{h\}$ be an arbitrary sequence which tends to zero. Then there exist a subsequence $\{h'\}$ of $\{h\}$ and a function u with the following properties:*

(i) $u \in C^0(\mathcal{H}) \cap L^\infty(\mathcal{H}), \quad u_x \in L^\infty(\mathcal{H})$;

(ii) $u(0,x) = u^0(x) \qquad$ *for all* $\quad x \in \boldsymbol{R}^1$;

(iii) *as* $h' \to 0$,

$$\|u_{h'} - u\|_{L^\infty(\mathcal{H})} \to 0, \tag{3.3}$$

$$\|(u_{h'})_x - u_x\|_{L^p(\mathcal{H})} \to 0 \quad (1 \le p < +\infty); \tag{3.4}$$

(iv) *u is a weak solution; that is, the following integral relation holds for any function $\phi(t,x) \in C^{1,1}(\mathcal{H})$ with compact support*

$$\iint_{\text{supp } u} (u\phi_t - muu_x\phi_x - (m-a)(u_x)^2\phi - c'\phi)dxdt + \int_{\boldsymbol{R}^1} u(0,x)\phi(0,x)dx = 0; \tag{3.5}$$

(v) *$u_{xx} \in \mathcal{E}'$ and $u_t \in \mathcal{E}'$, where $\mathcal{E}'$ is the dual of the space $\mathcal{E}$ consisting of all continuous functions with compact support in $(0,\infty) \times \boldsymbol{R}^1$.*

As a direct corollary of Theorem 3.1 we have the following theorem.

THEOREM 3.2 (convergence of numerical solutions v_h). *Let the same assumptions as stated in Theorem* 3.1 *be satisfied. Then $v_h \equiv (u_h)^{\frac{1}{m-1}}$ converges uniformly on $\mathcal{H}$ to the unique weak solution v of* (1.1)–(1.2) *as h tends to zero.*

Proof of Theorem 3.1. By following Graveleau and Jamet (Theorem 6.1 and Lemma 7.1 in [9]), the properties (i), (ii), (iii), and (v) are proved.

We now show the property (iv). We note that u has a compact support from (2.27), (2.33), and (3.3). Let $\{h\}$ take the value belonging to the extracted subsequence $\{h'\}$. From (2.1) we have

$$u_i^{n+1} - u_i^n = \sum_{q=0}^{\mu_{n+1}-1} \left(u_i^{n+1,\cdot,q+1} - u_i^{n+1,\cdot,q} \right) + \sum_{s=0}^{\nu_{n+1}-1} \left(u_i^{n+1,s+1} - u_i^{n+1,s} \right) + \left(u_i^{n+1,0} - u_i^n \right),$$

where $u_i^{n+1,s}$ $(s = 0, 1, \ldots, \nu_{n+1})$ and $u_i^{n+1,\cdot,q}$ $(q = 0, 1, \ldots, \mu_{n+1})$ are defined by (2.10) and (2.12), respectively. Let $\phi(t,x) \in C^{1,1}(\mathcal{H})$ have compact support and T^* be the extinction time of v. Then we have

$$\begin{aligned} &\sum_{(t_n, ih) \in \text{supp } u} h\phi_i^n \left(u_i^{n+1} - u_i^n \right) \\ &\quad = - \sum_{ih \in \text{supp } u^0} h u_i^0 \phi_i^0 - \sum_{(t_n, ih) \in \text{supp } u} h u_i^{n+1} (\phi_i^{n+1} - \phi_i^n) + \sum_{ih \in \text{supp } u(t_{n^*}, \cdot)} h u_i^{n^*+1} \phi_i^{n^*+1} \\ &\quad = A_h + B_h + C_h, \end{aligned}$$

where $\phi_i^n = \phi(t_n, ih)$, the integer n^* satisfies $t_{n^*} < T^*$ and $t_{n^*+1} \geq T^*$, and

$$\begin{aligned} A_h &= \sum_{(t_n, ih) \in \text{supp } u} h\phi_i^n \left\{ \sum_{q=0}^{\mu_{n+1}-1} (u_i^{n+1,\cdot,q+1} - u_i^{n+1,\cdot,q}) \right\}, \\ B_h &= \sum_{(t_n, ih) \in \text{supp } u} h\phi_i^n \left\{ \sum_{s=0}^{\nu_{n+1}-1} (u_i^{n+1,s+1} - u_i^{n+1,s}) \right\}, \\ C_h &= \sum_{(t_n, ih) \in \text{supp } u} h\phi_i^n (u_i^{n+1,0} - u_i^n). \end{aligned}$$

The following convergence is obvious:

$$\sum_{ih \in \text{supp } u^0} h u_i^0 \phi_i^0 \longrightarrow \int_{\boldsymbol{R}^1} u(0,x)\phi(0,x)\, dxdt \qquad \text{as} \quad h \to 0, \tag{3.6}$$

$$\sum_{ih \in \text{supp } u(t_{n^*}, \cdot)} h u_i^{n^*+1} \phi_i^{n^*+1} \longrightarrow 0 \qquad \text{as} \quad h \to 0. \tag{3.7}$$

By Lemma 4.1 in [21] we obtain

$$\sum_{(t_n, ih) \in \text{supp } u} h u_i^{n+1} (\phi_i^{n+1} - \phi_i^n) \longrightarrow \iint_{\text{supp } u} u\phi_t \, dxdt \qquad \text{as} \quad h \to 0, \tag{3.8}$$

$$A_h \longrightarrow - \iint_{\text{supp } u} \{muu_x\phi_x + m(u_x)^2\phi\} dxdt \qquad \text{as} \quad h \to 0, \tag{3.9}$$

$$B_h \longrightarrow \iint_{\text{supp } u} \{a(u_x)^2\phi\} dxdt \qquad \text{as} \quad h \to 0. \tag{3.10}$$

We prove the following convergence:

$$C_h \longrightarrow \iint_{\text{supp } u} (-c'\phi)\, dxdt \qquad \text{as} \quad h \to 0. \tag{3.11}$$

Let ε be an arbitrary positive number. Then we can choose the compact sets $G_j(\varepsilon)$ $(j = 1, 2, 3)$ such that

$$G_1(\varepsilon) \subset \overset{\circ}{G}_2(\varepsilon), \quad G_2(\varepsilon) \subset \overset{\circ}{\operatorname{supp}}\ u, \quad \operatorname{supp} u \subset \overset{\circ}{G}_3(\varepsilon), \tag{3.12}$$

$$\operatorname{meas}(G_2 \setminus G_1) < \varepsilon, \quad \operatorname{meas}(\operatorname{supp} u \setminus G_2) < \varepsilon, \quad \operatorname{meas}(G_3 \setminus \operatorname{supp} u) < \varepsilon, \tag{3.13}$$

where $\overset{\circ}{G}$ denotes the set of interior points of the set G, and $\operatorname{meas}(A)$ is the measure of the set A. Since u_h converges uniformly on $\mathcal{H}$, there exist positive numbers $\tilde{h}_1(\varepsilon)$ and $\tilde{h}_2(\eta_\varepsilon)$ for ε and the positive constant $\eta_\varepsilon \equiv \frac{1}{2}\min_{G_2(\varepsilon)} u(t, x)$, respectively, such that

$$\begin{aligned}(t + k_{n+1}, x + h) &\in G_3(\varepsilon) \setminus G_1(\varepsilon) \\ &\text{for} \quad (t, x) \in \operatorname{supp} u \setminus G_2(\varepsilon), \quad n \geq 0, \quad \text{and} \quad h < \tilde{h}_1(\varepsilon),\end{aligned} \tag{3.14}$$

$$\begin{aligned}u_h^n(x) - c' k_{n+1} &\geq u_h(t_n, x) - 2\|u_x^0\|_\infty h > \eta_\varepsilon \\ &\text{for} \quad (t_{n+1}, x) \in G_2(\varepsilon) \quad \text{and} \quad h < \tilde{h}_2(\eta_\varepsilon).\end{aligned} \tag{3.15}$$

Here we use the inequality $c' k_{n+1} \leq 2\|u_x^0\|_\infty h$ which follows from (2.17). From the definition of the difference operator $D_{h,k}$ it follows that

$$-c' \leq \frac{u_i^{n+1,0} - u_i^n}{k_{n+1}} \leq 0 \quad \text{for} \quad n \geq 0 \text{ and } i \in \boldsymbol{Z}, \tag{3.16}$$

$$\frac{u_i^{n+1,0} - u_i^n}{k_{n+1}} = -c' \quad \text{for} \quad (t_n, ih) \in G_2(\varepsilon) \text{ and } h < \min\{\tilde{h}_1(\varepsilon), \tilde{h}_2(\eta_\varepsilon)\}. \tag{3.17}$$

Since there exists a positive number $\tilde{h}_3(\varepsilon)$ such that

$$\left| \sum_{(t_n, ih) \in G_2(\varepsilon)} h k_{n+1} (-c' \phi_i^n) - \iint_{G_2(\varepsilon)} (-c' \phi)\, dxdt \right| < \varepsilon \quad \text{for} \quad h < \tilde{h}_3(\varepsilon),$$

we obtain from (3.12)–(3.17)

$$\begin{aligned}&\left| C_h - \iint_{\operatorname{supp} u} (-c' \phi)\, dxdt \right| \\ &\quad \leq \Big| \sum_{(t_n, ih) \in \operatorname{supp} u \setminus G_2(\varepsilon)} h k_{n+1} \phi_i^n \frac{u_i^{n+1,0} - u_i^n}{k_{n+1}} \Big| + \left| \iint_{\operatorname{supp} u \setminus G_2(\varepsilon)} c' \phi\, dxdt \right| \\ &\qquad + \Big| \sum_{(t_n, ih) \in G_2(\varepsilon)} h k_{n+1} \phi_i^n \frac{u_i^{n+1,0} - u_i^n}{k_{n+1}} - \iint_{G_2(\varepsilon)} (-c' \phi)\, dxdt \Big| \\ &\quad \leq (4c' \|\phi\|_\infty + 1)\varepsilon \quad \text{for} \quad h < \min\{\tilde{h}_1(\varepsilon), \tilde{h}_2(\eta_\varepsilon), \tilde{h}_3(\varepsilon)\},\end{aligned} \tag{3.18}$$

which yields (3.11). Hence, the property (iv) follows from (3.6)–(3.11), which completes the proof. □

THEOREM 3.3 (convergence of numerical extinction time). *Let the same assumptions as stated in Theorem* 3.1 *be satisfied. Then*

$$|T_h^* - T^*| \longrightarrow 0 \quad \textit{as} \quad h \to 0, \tag{3.19}$$

where T^* is the extinction time of the unique weak solution v of (1.1)–(1.2).

Proof. From Theorem 3.2 there exist $\tilde{x} \in \boldsymbol{R}^1$ and $h' > 0$ for an arbitrary fixed $t < T^*$ such that $v_h(t, \tilde{x}) \geq \frac{1}{2} v(t, \tilde{x}) > 0$ holds for $h < h'$. Then

$$\liminf_{h \to 0} T_h^* \geq T^*. \tag{3.20}$$

By (2.33) the inequality $T_h^* \leq T^* + \frac{1}{c'} \|u_h(T^*, \cdot)\|_\infty$ holds. Since $\|u(T^*, \cdot)\|_\infty = 0$ and

$$\|u_h(T^*, \cdot) - u(T^*, \cdot)\|_\infty \longrightarrow 0 \qquad \text{as } h \longrightarrow 0,$$

we have

$$\limsup_{h \to 0} T_h^* \leq T^*. \tag{3.21}$$

Hence (3.19) follows. □

4. Convergence of numerical interfaces. Let $\{u_h^n\}_{n=0,1,2,\ldots}$ be the numerical solutions given by (2.1) with (3.1). We introduce the set $\boldsymbol{W}$ consisting of all functions $\varphi(x)$ which satisfy Condition B and the following conditions:

(i) $(\varphi^{m-1})_x(x)$ is absolutely continuous on $\boldsymbol{I} \equiv \operatorname{supp} \varphi$;

(ii) $\operatorname{ess.inf}_{x \in \boldsymbol{I}} (\varphi^{m-1})_{xx}(x)$ is finite.

Since $\varphi(x) \in \boldsymbol{W}$ has compact support, it is clear that $\operatorname{ess.inf}_{x \in \boldsymbol{I}} (\varphi^{m-1})_{xx}(x)$ is negative.

For $v^0(x) \in \boldsymbol{W}$ we put

$$C_0 = \|u^0\|_\infty, \quad C_1 = \|u_x^0\|_\infty, \quad C_2 = -\operatorname*{ess.inf}_{x \in \boldsymbol{I}} u_{xx}^0(x), \tag{4.1}$$

where $u^0(x) = (v^0(x))^{m-1}$ and $\boldsymbol{I} = \operatorname{supp} v^0$, and define the left (resp., right) numerical interface $\ell_h(t)$ (resp., $r_h(t)$) by piecewise linearly interpolating (t_n, ℓ_n) (resp., (t_n, r_n)) $(0 \leq n \leq n^*)$. Then we obtain the following theorem, which plays an important role in proving the convergence of numerical interfaces and in finding the behavior of the support.

THEOREM 4.1. *For $v^0(x) \in \boldsymbol{W}$, assume that M and ε are positive constants satisfying*

$$u_x^0(x) > M \qquad \text{for } x \in [\ell(u^0), \ell(u^0) + \varepsilon). \tag{4.2}$$

Let the time step k_{n+1} $(n \geq 0)$ satisfy (2.17) *with $i = L$. Then the following estimates hold for each positive constant $M' < M$:*

$$\delta u_{L_0 - 1}^0, \quad \delta u_{L_0}^0, \quad \delta u_{L_0 + 1}^0 > M' \qquad \text{for} \quad h < \tilde{h}, \tag{4.3}$$

$$\delta u_{L_{n+1,s} - 1}^{n+1,s}, \quad \delta u_{L_{n+1,s}}^{n+1,s}, \quad \delta u_{L_{n+1,s} + 1}^{n+1,s} > M' (s = 0, 1, \ldots, \nu_{n+1}) \tag{4.4}$$

$$\text{for} \quad t_{n+1} \leq \tilde{T} \quad \text{and} \quad h < \tilde{h},$$

$$\delta u_{L_{n+1} - 1}^{n+1}, \quad \delta u_{L_{n+1}}^{n+1}, \quad \delta u_{L_{n+1} + 1}^{n+1} > M' \qquad \text{for} \quad t_{n+1} \leq \tilde{T} \quad \text{and} \quad h < \tilde{h}, \tag{4.5}$$

where $u_i^{n+1,s}$ $(s = 0, 1, \ldots, \nu_{n+1})$ are defined by (2.10), *and*

$$\tilde{T} = \frac{(M - M')M'}{2(2a + m)C_1 C_2 M' + 6c' C_2}, \qquad \tilde{h} = \min\left(\varepsilon, \frac{M - M'}{4C_2}\right). \tag{4.6}$$

Proof. In order to prove the theorem, we need Lemma 4.2, the proof of which is stated after the proof of this theorem.

LEMMA 4.2. *Let $v^0(x) \in \boldsymbol{W}$ and (2.17) with $i = L$ be satisfied. Then,*

$$\delta u^{n+1,0}_{L_{n+1,0}-1} \geq \delta u^{n}_{L_n-1} - 2C_2h, \tag{4.7}$$

$$\delta u^{n+1,s}_{L_{n+1,s}-1} \geq \delta u^{n+1,0}_{L_{n+1,0}-1} - C_2h - 2aC_1C_2\sum_{j=1}^{s}\tau_{n+1,j} \quad (s = 1, 2, \ldots, \nu_{n+1}), \tag{4.8}$$

$$\delta u^{n+1,\cdot,q}_{L_{n+1,\cdot,q}-1} \geq \delta u^{n+1,\cdot,0}_{L_{n+1,\cdot,0}-1} - mC_1C_2\,q\kappa_{n+1} \quad (q = 1, 2, \ldots, \mu_{n+1}) \tag{4.9}$$

for all $n \geq 0$, where $u_i^{n+1,\cdot,q}$ $(q = 0, 1, \ldots, \mu_{n+1})$ are given by (2.12).

Now we prove the theorem. It follows from (4.1), (4.2), and (4.6) that

$$\delta u^0_{L_0-1} > M > M' \qquad \text{for } h < \tilde{h}, \tag{4.10}$$

$$\delta u^0_{L_0} = \delta u^0_{L_0-1} + \frac{h + L_0h - \ell_0}{2}\delta^2 u^0_{L_0} > M - C_2h > M' \qquad \text{for } h < \tilde{h}, \tag{4.11}$$

$$\delta u^0_{L_0+1} = \delta u^0_{L_0} + h\delta^2 u^0_{L_0+1} > M - 2C_2h > M' \qquad \text{for } h < \tilde{h}, \tag{4.12}$$

which implies (4.3).

We note that (4.4) is shown in the proof of (4.5) and that the following estimates hold by Lemma 5.1 in [21] and the inequality $\inf_{i\in\boldsymbol{Z}}\delta^2 u_i \leq \inf_{i\in\boldsymbol{Z}}\delta^2 u'_i$ $(u'_h = D_{h,k}u_h)$ in the proof of Theorem 2.1:

$$\delta^2 u_i^n, \delta^2 u_i^{n+1,s}\ (s = 0, 1, \ldots, \nu_{n+1}), \delta^2 u_i^{n+1,\cdot,q}\ (q = 1, 2, \ldots, \mu_{n+1}) \geq -C_2 \quad \text{for all } i \in \boldsymbol{Z} \quad \text{and} \quad n \geq 0. \tag{4.13}$$

By Lemma 4.2 and (4.13) we have

$$\begin{aligned}
\delta u^{n+1,0}_{L_{n+1,0}+i} &\geq \delta u^{n}_{L_n-1} - 2C_2h - (i+1)C_2h \\
&\geq \delta u^{n,\cdot,0}_{L_{n,\cdot,0}-1} - mC_1C_2k_n - 2C_2h - (i+1)C_2h \\
&\geq \delta u^{n,0}_{L_{n,0}-1} - (2a+m)C_1C_2k_n - 3C_2h - (i+1)C_2h \\
&\geq \delta u^{1,0}_{L_{1,0}-1} - (2a+m)C_1C_2t_n - 3C_2nh - (i+1)C_2h \\
&\geq \delta u^{0}_{L_0-1} - (2a+m)C_1C_2t_n - 3C_2nh - 2C_2h - (i+1)C_2h \quad \text{for } n \geq 0,
\end{aligned} \tag{4.14}$$

$$\begin{aligned}
\delta u^{n+1,s}_{L_{n+1,s}+i} &\geq \delta u^{n+1,0}_{L_{n+1,0}-1} - C_2h - 2aC_1C_2k_{n+1} - (i+1)C_2h \\
&\geq \delta u^{0}_{L_0-1} - (2a+m)C_1C_2t_n - 2aC_1C_2k_{n+1} \\
&\quad - 3C_2(n+1)h - (i+1)C_2h \quad (s = 1, 2, \ldots, \nu_{n+1}) \quad \text{for } n \geq 0,
\end{aligned} \tag{4.15}$$

$$\begin{aligned}
\delta u^{n+1}_{L_{n+1}+i} &\geq \delta u^{n+1,\cdot,\mu_{n+1}}_{L_{n+1,\cdot,\mu_{n+1}}-1} - (i+1)C_2h \\
&\geq \delta u^{n+1,\cdot,0}_{L_{n+1,\cdot,0}-1} - mC_1C_2k_{n+1} - (i+1)C_2h \\
&\geq \delta u^{0}_{L_0-1} - (2a+m)C_1C_2t_{n+1} - 3C_2(n+1)h - (i+1)C_2h \quad \text{for } n \geq 0,
\end{aligned} \tag{4.16}$$

where $i = -1, 0, 1$.

Let us prove (4.5) by induction on n $(n \geq 0)$. Let $n = 0$. Since (4.3) and (2.17) with $i = L$ yield

$$hM' < c'k_1 = u^0_{L_0+1}, \tag{4.17}$$

we have from (4.2), (4.6), and (4.14)–(4.16)

$$\begin{aligned}
&\min(\delta u^{1,0}_{L_{1,0}+i}, \delta u^{1,1}_{L_{1,1}+i}, \ldots, \delta u^{1,\nu_1}_{L_{1,\nu_1}+i},\ \delta u^{1}_{L_1+i}) \\
&\quad > M - (2a+m)C_1C_2t_1 - \frac{3c'C_2k_1}{M'} - \frac{M-M'}{2} \geq M' \quad (i = -1, 0, 1),
\end{aligned} \tag{4.18}$$

which implies (4.4) and (4.5) with $n = 0$.

Suppose that (4.5) holds for all $n-1$ ($n \geq 1$). Then we have from (2.17) with $i = L$

$$hM' < c'k_{j+1} = u^j_{L_j+1} \quad (j = 0, 1, \ldots, n). \tag{4.19}$$

Hence, by (4.2), (4.6), and (4.14)–(4.16)

$$\begin{aligned} &\min(\delta u^{n+1,0}_{L_{n+1,0}+i}, \delta u^{n+1,1}_{L_{n+1,1}+i}, \ldots, \delta u^{n+1,\nu_{n+1}}_{L_{n+1,\nu_{n+1}}+i}, \delta u^{n+1}_{L_{n+1}+i}) \\ &\quad > M - (2a+m)C_1C_2t_{n+1} - 3C_2\sum_{j=0}^{n}\frac{c'k_{j+1}}{M'} - \frac{M-M'}{2} \\ &\quad = M - \Big\{(2a+m)C_1C_2 + \frac{3c'C_2}{M'}\Big\}t_{n+1} - \frac{M-M'}{2} \geq M' \quad (i = -1, 0, 1), \end{aligned} \tag{4.20}$$

which shows that (4.4) and (4.5) hold for n. Thus the induction on n is complete, and the theorem is proved. □

Proof of Lemma 4.2. For simplicity we put

$$\begin{aligned} &u^{\cdot,s}_h = u^{n+1,s}_h, \quad u^{\cdot,\cdot,q}_h = u^{n+1,\cdot,q}_h, \ \kappa = \kappa_{n+1}, \quad \tau_j = \tau_{n+1,j}, \\ &L_{\cdot,s} = L_{n+1,s}, \quad L_{\cdot,\cdot,q} = L_{n+1,\cdot,q}, \ell_{\cdot,s} = \ell_{n+1,s}, \quad \ell_{\cdot,\cdot,q} = \ell_{n+1,\cdot,q}. \end{aligned}$$

We first show (4.7) and (4.9). From (2.17) and (4.13) it follows that

$$\delta u^{\cdot,0}_{L_{\cdot,0}-1} \geq \begin{cases} \delta u^n_{L_n+1} \geq \delta u^n_{L_n-1} - 2C_2h & \text{if} \quad \delta u^n_{L_n} \geq 0, \\ 0 \geq \delta u^n_{L_n} \geq \delta u^n_{L_n-1} - C_2h & \text{if} \quad \delta u^n_{L_n} < 0, \end{cases}$$

which yields (4.7). From (2.24) we have

$$u^{\cdot,\cdot,q}_{L_{\cdot,\cdot,q}} = u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}} + \kappa m u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}}\delta^2 u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}}.$$

Since

$$\begin{aligned} &u^{\cdot,\cdot,q}_{L_{\cdot,\cdot,q}} = (L_{\cdot,\cdot,q}h - \ell_{\cdot,\cdot,q})\delta u^{\cdot,\cdot,q}_{L_{\cdot,\cdot,q}-1}, \quad u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}} = (L_{\cdot,\cdot,q-1}h - \ell_{\cdot,\cdot,q-1})\delta u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}-1}, \\ &\ell_{\cdot,\cdot,q} = \ell_{\cdot,\cdot,q-1} = \ell_{\cdot,\cdot,0}, \qquad\qquad L_{\cdot,\cdot,q} = L_{\cdot,\cdot,q-1} = L_{\cdot,\cdot,0}, \end{aligned}$$

we obtain

$$\delta u^{\cdot,\cdot,q}_{L_{\cdot,\cdot,q}-1} \geq \delta u^{\cdot,\cdot,q-1}_{L_{\cdot,\cdot,q-1}-1} - mC_1C_2\,\kappa \geq \delta u^{\cdot,\cdot,0}_{L_{\cdot,\cdot,0}-1} - mC_1C_2\,q\kappa, \tag{4.21}$$

which shows (4.9).

Next we show (4.8). By using the expression (2.21) we obtain

$$\delta u^{\cdot,s}_{L_{\cdot,s}-1} \geq \min(\delta u^{\cdot,s-1}_{L_{\cdot,s-1}-1},\ \delta u^{\cdot,s-1}_{L_{\cdot,s-1}}) \ (s = 1, 2, \ldots, \nu_{n+1}), \tag{4.22}$$

$$\delta u^{\cdot,s}_{L_{\cdot,s}} \geq \min(\delta u^{\cdot,s-1}_{L_{\cdot,s-1}-1},\ \delta u^{\cdot,s-1}_{L_{\cdot,s-1}} - 2aC_1C_2\tau_s) \ (s = 1, 2, \ldots, \nu_{n+1}). \tag{4.23}$$

For $s = 1$ we have from (4.22) and (4.13)

$$\delta u^{\cdot,1}_{L_{\cdot,1}-1} \geq \delta u^{\cdot,0}_{L_{\cdot,0}-1} - C_2h. \tag{4.24}$$

For $2 \le s \le \nu_{n+1}$ it follows from (4.22), (4.23), and (4.13) that

$$
\begin{aligned}
(4.25)\qquad \delta u^{\cdot,s}_{L_\cdot,s-1} &\ge \min(\delta u^{\cdot,s-1}_{L_\cdot,s-1-1},\ \delta u^{\cdot,s-2}_{L_\cdot,s-2-1},\ \delta u^{\cdot,s-2}_{L_\cdot,s-2} - 2aC_1C_2\tau_{s-1}) \\
&\ge \min\big(\delta u^{\cdot,s-1}_{L_\cdot,s-1-1},\ \delta u^{\cdot,s-2}_{L_\cdot,s-2-1},\ \delta u^{\cdot,s-3}_{L_\cdot,s-3-1} - 2aC_1C_2\tau_{s-1}, \\
&\qquad \delta u^{\cdot,s-4}_{L_\cdot,s-4-1} - 2aC_1C_2(\tau_{s-1}+\tau_{s-2}),\dots, \\
&\qquad \delta u^{\cdot,0}_{L_\cdot,0-1} - 2aC_1C_2(\tau_{s-1}+\tau_{s-2}+\cdots+\tau_2), \\
&\qquad \delta u^{\cdot,0}_{L_\cdot,0} - 2aC_1C_2(\tau_{s-1}+\tau_{s-2}+\cdots+\tau_1)\big) \\
&\ge \min(\delta u^{\cdot,0}_{L_\cdot,0-1},\ \delta u^{\cdot,0}_{L_\cdot,0}) - 2aC_1C_2(\tau_{s-1}+\tau_{s-2}+\cdots+\tau_1) \\
&\ge \delta u^{\cdot,0}_{L_\cdot,0-1} - C_2h - 2aC_1C_2(\tau_1+\tau_2+\cdots+\tau_s).
\end{aligned}
$$

Hence the desired inequality (4.8) holds by (4.24) and (4.25), which completes the proof. □

From Theorem 4.1 we have the following theorem.

THEOREM 4.3 (convergence of the left numerical interface). *Let the assumptions of Theorem* 4.1 *be satisfied. Then the left numerical interface $\ell_h(t)$ converges uniformly to the exact one on $[0,\tilde{T}]$ for each positive constant $M' < M$, where $\tilde{T}$ is given by* (4.6).

Proof. Since by Theorem 4.1

$$0 < u^n_{L_n} < u^n_{L_n+1} < u^n_{L_n+2} \qquad \text{for all } t_n \le \tilde{T},$$

it follows from Theorem 2.1 and (2.17) with $i = L$ that

$$
\begin{aligned}
\ell_{n+1} - \ell_n &= \frac{u^n_{L_n}}{\delta u^n_{L_n-1}} + \frac{c'k_{n+1} - u^n_{L_n}}{\delta u^n_{L_n}} - \sum_{s=1}^{\nu_{n+1}} a\delta u^{n+1,s-1}_{L_{n+1,s-1}-1}\tau_{n+1,s} \\
&\begin{cases} > -a\|u^0_x\|_\infty k_{n+1} & \text{for } t_{n+1} \in [0,\tilde{T}], \\ < \dfrac{c'k_{n+1}}{M'} & \text{for } t_{n+1} \in [0,\tilde{T}], \end{cases}
\end{aligned}
$$

which gives

$$|\dot{\ell}_h(t)| \le \max\Big(a\|u^0_x\|_\infty,\ \frac{c'}{M'}\Big) \qquad \text{for } t \in [0,\tilde{T}]. \tag{4.26}$$

From (2.27) we have

$$\ell_0 - a\|u^0_x\|_\infty\tilde{T} \le \ell_h(t) \le r_0 + a\|u^0_x\|_\infty\tilde{T} \qquad \text{for } t \in [0,\tilde{T}]. \tag{4.27}$$

By (4.26) and (4.27) we can apply Ascoli–Arzelà's theorem to the sequence $\{\ell_h\}$. Thus there exist a Lipschitz continuous function $\tilde{\ell}(t)$ and a subsequence $\{\ell_{h'}\}$ satisfying

$$\|\ell_{h'} - \tilde{\ell}\|_{L^\infty([0,\tilde{T}])} \longrightarrow 0 \quad \text{as} \quad h' \to 0. \tag{4.28}$$

Now we show that $\tilde{\ell}(t)$ is the left interface of the weak solution v of (1.1)–(1.2). For simplicity we use h instead of h'.

Let $t^* \in [0,\tilde{T}]$. Then there exist integers n and L_n such that $t_n \le t^* < t_{n+1}$ and $(L_n-1)h \le \ell_h(t_n) < L_nh$, which imply

$$u_h(t^*, (L_n-1)h - \xi) = 0 \qquad \text{for any positive number } \xi. \tag{4.29}$$

For each fixed $\eta \in (0, \frac{M'}{C_2})$ let p be the positive integer satisfying $(p-1)h \le \eta < ph$. Then we have

$$
\begin{aligned}
(4.30)\ u_h(t^*, (L_n+p)h) &> h \sum_{i=L_n}^{L_n+p-1} \delta u_i^n = h \sum_{j=0}^{p-1} \left\{ \delta u_{L_n-1}^n + \sum_{i=L_n}^{L_n+j} (\delta u_i^n - \delta u_{i-1}^n) \right\} \\
&\ge ph\delta u_{L_n-1}^n - h \sum_{j=0}^{p-1} (j+1) C_2 h = ph\Big\{ \delta u_{L_n-1}^n - \frac{(p-1)C_2 h}{2} - C_2 h \Big\} \\
&> \eta\Big(M' - \frac{C_2\eta}{2} - C_2 h \Big) > \frac{\eta(M' - 2C_2 h)}{2} > \frac{\eta M'}{4} \quad \text{for} \ \ h < \frac{M'}{4C_2}.
\end{aligned}
$$

Since u_h converges uniformly to $u = v^{m-1}$ on $\mathcal{H}$ by Theorems 3.1 and 3.2, we have from (4.29) and (4.30)

$$
u(t^*, \tilde{\ell}(t^*) - \xi) = 0 \quad \text{for any positive number } \xi, \tag{4.31}
$$

$$
u(t^*, \tilde{\ell}(t^*) + \eta) \ge \frac{\eta M'}{4} \quad \text{for each } \eta \in \Big(0, \frac{M'}{C_2}\Big). \tag{4.32}
$$

Hence, $\tilde{\ell}(t)$ becomes the left interface on $[0, \tilde{T}]$. Since the left interface is uniquely determined, $\ell_h(t)$ converges to it on $[0, \tilde{T}]$ as the whole sequence $\{h\}$ tends to zero. Thus the proof is complete. □

Remark 4.1. Taking the statement of Theorems 4.1 and 4.3 into consideration, we can easily obtain the convergence of the right numerical interface.

Remark 4.2. When $\operatorname{supp} u^0$ is concave downward on its support, the convergence of v_h, ℓ_h, r_h, and T_h^* is shown by Nakaki (see Theorems 4.2, 4.3, and 5.2 in [22]).

5. Behavior of the support. In this section we consider the possibility that the support will split in Cases 1 and 2 stated in section 1. We obtain the following theorem in Case 1, which improves Kersner's result [15]. To state it we put

$$
w_1^0(x) = \begin{cases} v^0(x) & \text{if } x \le \alpha_0, \\ 0 & \text{if } x > \alpha_0, \end{cases} \qquad w_2^0(x) = \begin{cases} 0 & \text{if } x \le \alpha_0, \\ v^0(x) & \text{if } x > \alpha_0. \end{cases} \tag{5.1}
$$

THEOREM 5.1. *In Case* 1, *assume* $w_j^0(x) \in \boldsymbol{W}$ $(j = 1, 2)$. *Put* $W_j(x) = (w_j^0)^{m-1}(x)$ $(j = 1, 2)$. *Suppose*

$$
a\left((W_1)_x(\alpha_0 - 0)\right)^2 \quad \text{and} \quad a\left((W_2)_x(\alpha_0 + 0)\right)^2 > c'. \tag{5.2}
$$

Then there exist a constant $\tilde{T} > 0$ *such that* $\operatorname{supp} v(t, \cdot)$ *is connected for each* $t \in [0, \tilde{T}]$.

Suppose

$$
a\|(W_j)_x\|_\infty^2 < c' \quad (j = 1, 2). \tag{5.3}
$$

Then $\operatorname{supp} v(t, \cdot)$ *is disconnected for each* $t \in (0,\ T_*)$, *where* $T_* = \min(T_1^*, T_2^*)$ *and* T_1^* *and* T_2^* *are the extinction times of the solutions of* (1.1)–(1.2) *with* $v^0(x) = w_1^0(x)$ *and* $v^0(x) = w_2^0(x)$, *respectively.*

Proof. Let $\tilde{\ell}(t)$ and $\{u_h^n\}_{n=0,1,2,\dots}$ be the left interface of the solution v of (1.1)–(1.2) with $v^0(x) = w_2^0(x)$ and the numerical solutions given by (2.1) with (3.1), respectively, where (2.17) with $i = L$ is satisfied and $\tilde{\ell}(0) = \alpha_0$.

We show the first assertion of the theorem. We take arbitrary positive constants K_1 and K_1' such that

$$
\Big(\frac{c'}{a}\Big)^{\frac{1}{2}} < K_1' < K_1 < (W_2)_x(\alpha_0 + 0). \tag{5.4}
$$

By applying Theorem 4.1 to the initial function $v^0(x) = w_2^0(x)$ with the positive constants $M = K_1$ and $M' = K_1'$, we have

$$\delta u_{L_n-1}^n, \quad \delta u_{L_n}^n, \quad \delta u_{L_n+1}^n > K_1' \quad \text{for } t_n \le \tilde{T}_1 \text{ and } h < \tilde{h}, \tag{5.5}$$

$$\delta u_{L_{n+1,s}-1}^{n+1,s}, \quad \delta u_{L_{n+1,s}}^{n+1,s}, \quad \delta u_{L_{n+1,s}+1}^{n+1,s} > K_1' \quad (s = 0, 1, \dots, \nu_{n+1}) \quad \text{for } t_{n+1} \le \tilde{T}_1 \text{ and } h < \tilde{h}, \tag{5.6}$$

where $\tilde{T}_1$ and $\tilde{h}$ are some positive constants. Hence, it follows from (2.15), (2.17) with $i = L$, and (2.22) that

$$\begin{aligned} \ell_{n+1} &= \ell_n + \frac{u_{L_n}^n}{\delta u_{L_n-1}^n} + \frac{c' k_{n+1} - u_{L_n}^n}{\delta u_{L_n}^n} - \sum_{s=1}^{\nu_{n+1}} a \delta u_{L_{n+1,s-1}-1}^{n+1,s-1} \tau_{n+1,s} \\ &< \ell_n + \frac{c' k_{n+1}}{K_1'} - a K_1' k_{n+1} \\ &< \ell_0 - \frac{1}{K_1'}(a K_1'^2 - c') t_{n+1} < \ell_0 = \alpha_0 \qquad \text{for } t_{n+1} \in (0, \tilde{T}_1], \end{aligned} \tag{5.7}$$

which gives

$$\ell_h(t) < \alpha_0 - \frac{1}{K_1'}(a K_1'^2 - c') t < \alpha_0 \qquad \text{for } t \in (0, \tilde{T}_1].$$

From the convergence of $\ell_h(t)$ we have

$$\tilde{\ell}(t) \le \alpha_0 - \frac{1}{K_1'}(a K_1'^2 - c') t < \alpha_0 \qquad \text{for } t \in (0, \tilde{T}_1]. \tag{5.8}$$

Taking arbitrary positive constants K_2 and K_2' such that

$$\left(\frac{c'}{a}\right)^{\frac{1}{2}} < K_2' < K_2 < -(W_1)_x(\alpha_0 - 0), \tag{5.9}$$

we can similarly obtain

$$\tilde{r}(t) \ge \alpha_0 + \frac{1}{K_2'}(a K_2'^2 - c') t > \alpha_0 \qquad \text{for } t \in (0, \tilde{T}_2], \tag{5.10}$$

where $\tilde{r}(t)$ is the right interface of the solution v of (1.1)–(1.2) with $v^0(x) = w_1^0(x)$, $\tilde{r}(0) = \alpha_0$, and $\tilde{T}_2$ is some positive constant. Put $T' = \min(\tilde{T}_1, \tilde{T}_2)$. Then we have from (5.8), (5.10), and the comparison theorem on the initial data (see [2])

$$v(t, \alpha_0) > 0 \quad \text{for } t \in (0, T']. \tag{5.11}$$

Hence, the continuity of $v(t, x)$ yields the existence of a positive constant $\tilde{T} (< T')$ such that $\operatorname{supp} v(t, \cdot)$ is connected for each $t \in [0, \tilde{T}]$.

We prove the second assertion of the theorem. Let K_3 be an arbitrary positive constant satisfying

$$\|(W_j)_x\|_\infty < K_3 < \left(\frac{c'}{a}\right)^{\frac{1}{2}} \quad (j = 1, 2). \tag{5.12}$$

By Lemma 3.1 in [21] and Theorem 2.1 we have

$$\|\delta u_i^n\|_\infty, \|\delta u_i^{n+1,s}\|_\infty < K_3 \ (i \in \boldsymbol{Z}, s = 0, 1, \dots, \nu_{n+1}) \text{ for } t_{n+1} \in (0, T_2^*]. \tag{5.13}$$

Then, it follows from (2.15), (2.17) with $i = L$, and (2.22) that

$$\begin{aligned}(5.14)\qquad \ell_{n+1} &\geq \ell_n + \frac{c'k_{n+1}}{\max(\delta u^n_{L_n}, \delta u^n_{L_n-1})} - \sum_{s=1}^{\nu_{n+1}} a\delta u^{n+1,s-1}_{L_{n+1,s-1}-1}\tau_{n+1,s} \\ &> \ell_n + \frac{c'k_{n+1}}{K_3} - aK_3k_{n+1} \\ &> \ell_0 + \frac{1}{K_3}(c' - aK_3^2)t_{n+1} > \ell_0 = \alpha_0 \qquad \text{for } t_{n+1} \in (0, T_2^*],\end{aligned}$$

which immediately yields

$$(5.15)\qquad \tilde{\ell}(t) \geq \alpha_0 + \frac{1}{K_3}(c' - aK_3^2)t > \alpha_0 \qquad \text{for } t \in (0, T_2^*].$$

Similarly, we have

$$(5.16)\qquad \tilde{r}(t) \leq \alpha_0 - \frac{1}{K_3}(c' - aK_3^2)t < \alpha_0 \qquad \text{for } t \in (0, T_1^*].$$

Hence $\operatorname{supp} v(t,\cdot)$ becomes disconnected for $t \in (0, T_*)$, which completes the proof. □

Remark 5.1. The first assertion of Theorem 5.1 holds, when the orders of vanishing of $(w_j^0(x))^{m-1}$ $(j = 1, 2)$ at $x = \alpha_0$ are less than 1. This result also coincides with the one given by Kersner (see Theorem 2 in [15]).

Remark 5.2. Let us consider the sufficient condition under which $v(t, \alpha_0) = 0$ holds for all $t \geq 0$. Put

$$(5.17)\qquad v^0(x) = \begin{cases} A_1\{(\alpha_1 - x)(x - \alpha_0)\}^{\frac{s}{m-1}} & \text{if } \alpha_1 \leq x \leq \alpha_0, \\ A_2\{(\alpha_0 - x)(x - \alpha_2)\}^{\frac{s}{m-1}} & \text{if } \alpha_0 \leq x \leq \alpha_2, \\ 0 & \text{otherwise,} \end{cases}$$

where $s = 1 + \varepsilon$ $(\varepsilon \geq 0)$ and A_j $(j = 1, 2)$ are positive constants. According to Theorem 1 in [15], Kersner's sufficient condition is

$$(5.18)\quad 4A_j^{2(m-1)}\left(\frac{|\alpha_0 - \alpha_j|}{2}\right)^{2+4\varepsilon}(1+\varepsilon)(1+m\varepsilon) < \frac{c(m-1)^2}{m} \quad (j = 1, 2).$$

Our sufficient condition (5.3) yields

$$(5.19)\quad 4A_j^{2(m-1)}\left(\frac{|\alpha_0 - \alpha_j|}{2}\right)^{2+4\varepsilon}\left(\frac{2\varepsilon}{1+2\varepsilon}\right)^{2\epsilon}\frac{(1+\varepsilon)^2}{1+2\varepsilon} < \frac{c(m-1)^2}{m} \quad (j = 1, 2).$$

For $\varepsilon = 0$ his sufficient condition coincides with ours. For $\varepsilon > 0$ we find that the condition imposed on A_j $(j = 1, 2)$ by (5.19) is weaker than that imposed by (5.18).

We show a sufficient condition under which the support begins to split into at least two disjoint sets.

THEOREM 5.2. *In Case* 2, *let* $v^0(x) \in \boldsymbol{W}$ *and* $\alpha_1 < \beta_1 < \gamma_1 < \gamma_2 < \beta_2 < \alpha_2$. *Assume*

$$(5.20)\qquad \frac{u^0(\beta_j)}{c' + mC_0C_2} > \frac{\|u^0\|_{L^1[\gamma_1,\gamma_2]}}{c'(\gamma_2 - \gamma_1) - (m+a)C_0TV(u_x^0)} > 0 \quad (j = 1, 2),$$

where $u^0(x) = (v^0(x))^{m-1}$ *and* C_j $(j = 0, 2)$ *are constants given by* (4.1). *Then there exist* $\tilde{t} > 0$ *and* $\tilde{x} \in [\gamma_1, \gamma_2]$ *such that*

$$v(\tilde{t}, \tilde{x}) = 0 \quad \textit{and} \quad v(\tilde{t}, \beta_j) > 0 \quad (j = 1, 2). \tag{5.21}$$

Also, if $u^0(x)$ *satisfies*

$$a\|u_x^0\|_\infty^2 < c', \tag{5.22}$$

then there exists a positive constant t^* *such that supp* $v(t, \cdot)$ *is disconnected for each* $t \in (\tilde{t}, \tilde{t} + t^*)$.

Proof. Let $\{u_h^n\}_{n=0,1,2,\dots}$ and u_h be the numerical solutions given by (2.1) with (3.1) and by (3.2), respectively. Put

$$T_j = \frac{u^0(\beta_j)}{c' + mC_0C_2} \quad (j = 1, 2), \quad \tilde{T} = \frac{\|u^0\|_{L^1[\gamma_1,\gamma_2]}}{c'(\gamma_2 - \gamma_1) - (m + a)C_0TV(u_x^0)}, \tag{5.23}$$

$$T' = \tilde{T} + \frac{1}{4}\{\min(T_1, T_2) - \tilde{T}\}, \quad T'' = \tilde{T} + \frac{1}{2}\{\min(T_1, T_2) - \tilde{T}\}, \tag{5.24}$$

$$\boldsymbol{S} = [0, T''] \times [\gamma_1, \gamma_2]. \tag{5.25}$$

We first show that $\boldsymbol{S}$ contains at least one point $(\tilde{t}, \tilde{x})$ such that $v(\tilde{t}, \tilde{x}) = 0$. For this end we assume the contrary; that is, suppose the solution $v(t, x)$ is positive on $\boldsymbol{S}$. Since $(u_h)^{1/(m-1)}$ converges uniformly to the solution v on $\boldsymbol{S}$ by Theorem 3.2, there exists a positive number $h'(\eta)$ for the constant $\eta \equiv \frac{1}{2}\min_{(t,x)\in\boldsymbol{S}} u(t, x) > 0$ such that

$$u_h^n(x) - c'k_{n+1} \ge u_h(t_n, x) - 2\|u_x^0\|_\infty h > \eta \quad \text{for} \quad (t_{n+1}, x) \in \boldsymbol{S} \quad \text{and} \quad h < h'(\eta). \tag{5.26}$$

Here we use the inequality $c'k_{n+1} \le 2\|u_x^0\|_\infty h$ which follows from (2.17). Then we have from Lemmas 3.1 and 3.2 in [21] and Theorem 2.1 that

$$\begin{aligned}
\|u_h^{n+1}\|_{L^1[\gamma_1,\gamma_2]} &\le \left\| u_h^{n+1} - \left(\prod_{j=1}^{\nu} H_{h,\tau_j} \right) D_{h,k_{n+1}} u_h^n \right\|_{L^1[\gamma_1,\gamma_2]} \\
&\quad + \left\| \left(\prod_{j=1}^{\nu} H_{h,\tau_j} \right) D_{h,k_{n+1}} u_h^n - D_{h,k_{n+1}} u_h^n \right\|_{L^1[\gamma_1,\gamma_2]} + \|D_{h,k_{n+1}} u_h^n\|_{L^1[\gamma_1,\gamma_2]} \\
&\le mC_0TV(u_x^0)k_{n+1} + aC_0TV(u_x^0)k_{n+1} + \|u_h^n\|_{L^1[\gamma_1,\gamma_2]} - c'(\gamma_2 - \gamma_1)k_{n+1} \\
&\le \|u_h^0\|_{L^1[\gamma_1,\gamma_2]} - \{c'(\gamma_2 - \gamma_1) - (m + a)C_0TV(u_x^0)\}t_{n+1} \\
&= \|u_h^0\|_{L^1[\gamma_1,\gamma_2]} - \|u^0\|_{L^1[\gamma_1,\gamma_2]} + \left(1 - \frac{t_{n+1}}{\tilde{T}}\right)\|u^0\|_{L^1[\gamma_1,\gamma_2]} \\
&\qquad \text{for} \quad t_{n+1} \in (0, T''] \quad \text{and} \quad h < h'(\eta).
\end{aligned} \tag{5.27}$$

Since $T' > \tilde{T}$, we have

$$\|u_h^{n+1}\|_{L^1[\gamma_1,\gamma_2]} < 0 \qquad \text{for } t_{n+1} \in (T', T''] \text{ and for sufficiently small } h, \tag{5.28}$$

which is a contradiction. Hence we obtain

$$u(\tilde{t},\ \tilde{x}) = 0 \quad \text{for some} \quad (\tilde{t}, \tilde{x}) \in \boldsymbol{S}. \tag{5.29}$$

Next, we have

$$
\begin{aligned}
(5.30)\quad u_h^{n+1}(\beta_j) &= \left(u_h^{n+1} - \left(\prod_{j=1}^{\nu} H_{h,\tau_j}\right) D_{h,k_{n+1}} u_h^n\right)(\beta_j) \\
&\quad + \left(\left(\prod_{j=1}^{\nu} H_{h,\tau_j}\right) D_{h,k_{n+1}} u_h^n - D_{h,k_{n+1}} u_h^n\right)(\beta_j) + (D_{h,k_{n+1}} u_h^n)(\beta_j) \\
&\geq -mC_0C_2k_{n+1} + u_h^n(\beta_j) - c'k_{n+1} \geq u_h^0(\beta_j) - (c' + mC_0C_2)t_{n+1} \\
&\geq u_h^0(\beta_j) - u^0(\beta_j) + (c' + mC_0C_2)(T_j - T'') \quad \text{for } t_{n+1} \in [0, T''] \ (j = 1, 2).
\end{aligned}
$$

Letting $h \to 0$, we obtain

$$
(5.31)\qquad u(t, \beta_j) \geq (c' + mC_0C_2)(T_j - T'') > 0 \quad \text{for} \quad t \in [0, T''] \quad (j = 1, 2),
$$

which implies

$$
(5.32)\qquad u(\tilde{t},\ \beta_j) > 0 \quad (j = 1, 2).
$$

Thus (5.21) follows from (5.29) and (5.32).

Finally, we show the last assertion of the theorem. Let $w_j(t, x)$ $(j = 1, 2)$ be the solutions of (1.1)–(1.2) with $v^0(x) = w_j^0(x)$ $(j = 1, 2)$, where

$$
(5.33)\qquad w_1^0(x) = \begin{cases} v(\tilde{t}, x) & \text{if } x \leq \tilde{x}, \\ 0 & \text{if } x > \tilde{x}, \end{cases} \qquad w_2^0(x) = \begin{cases} 0 & \text{if } x \leq \tilde{x}, \\ v(\tilde{t}, x) & \text{if } x > \tilde{x}. \end{cases}
$$

Since the solution v is smooth in the open set $\{(t, x);\ v(t, x) > 0\}$, it follows from Theorems 2.1 and 3.2 that $w_j^0(x)$ $(j = 1, 2)$ belong to $\boldsymbol{W}$. Applying the second assertion of Theorem 5.1 to $w_j^0(x)$ $(j = 1, 2)$, we have

$$
(5.34)\qquad r(w_1(t, x)) < \tilde{x} < \ell(w_2(t, x)) \quad \text{for } t \in (0, t^*),
$$

where $t^* = \min(T_1^*, T_2^*)$ and T_1^* and T_2^* are the extinction times of $w_1(t, x)$ and $w_2(t, x)$, respectively. Thus $\operatorname{supp} v(t, \cdot)$ is disconnected for each $t \in (\tilde{t}, \tilde{t} + t^*)$, and the proof is complete. □

Remark 5.3. We briefly explain the construction of the initial function for which (5.20) is satisfied. In Theorem 5.2 we assume

$$
(5.35)\qquad \frac{u^0(\beta_j)}{c' + mC_0C_2} > \frac{\varepsilon}{c'} \quad (j = 1, 2) \quad \text{and} \quad u^0(x) = \varepsilon \text{ on } [\gamma_1, \gamma_2]
$$

instead of (5.20), where ε is a positive constant. For an arbitrary positive number d let

$$
(5.36)\qquad u_d^0(x) = \begin{cases} u^0(x + d) & \text{if} \quad x < \gamma_1 - d, \\ \varepsilon & \text{if} \quad \gamma_1 - d \leq x \leq \gamma_2 + d, \\ u^0(x - d) & \text{if} \quad x > \gamma_2 + d, \end{cases}
$$

and put $C_0(d) = \|u_d^0\|_\infty$ and $C_2(d) = -\operatorname{ess.inf}_{x \in \boldsymbol{I}} (u_d^0)_{xx}(x)$. Since $C_j(d) = C_j$ $(j = 0, 2)$ and $TV((u_d^0)_x) = TV(u_x^0)$, we can choose d sufficiently large so that

$$
(5.37)\qquad \frac{u_d^0(\beta_j + (-1)^j d)}{c' + mC_0(d)C_2(d)} > \frac{\varepsilon}{c' - \frac{(m+a)C_0(d)TV((u_d^0)_x)}{\gamma_2 - \gamma_1 + 2d}} > \frac{\varepsilon}{c'} \quad (j = 1, 2).
$$

Then, by using $\|u_d^0\|_{L^1[\gamma_1-d,\gamma_2+d]} = (\gamma_2-\gamma_1+2d)\varepsilon$, we find that $u_d^0(x)$ satisfies (5.20), where β_j and γ_j $(j=1,2)$ are replaced by $\beta_j+(-1)^j d$ and $\gamma_j+(-1)^j d$ $(j=1,2)$, respectively.

REFERENCES

[1] D. G. Aronson, *The porous medium equation*, in Nonlinear Diffusion Problems, Lecture Notes in Math. 1224, A. Fasano and M. Primicerio, eds., Springer-Verlag, Berlin, 1986.

[2] M. Bertsch, *A class of degenerate diffusion equations with a singular nonlinear term*, Nonlinear Anal., 7 (1983), pp. 117–127.

[3] M. Bertsch, R. Kersner, and L. A. Peletier, *Positivity versus localization in degenerate diffusion equations*, Nonlinear Anal., 9 (1985), pp. 987–1008.

[4] X.-Y. Chen, H. Matano, and M. Mimura, *Finite-point extinction and continuity of interfaces in a nonlinear diffusion equation with strong absorption*, J. Reine Angew. Math., 459 (1995), pp. 1–36.

[5] E. DiBenedetto and D. Hoff, *An interface tracking algorithm for the porous medium equation*, Trans. Amer. Math. Soc., 284 (1984), pp. 463–500.

[6] V. A. Galaktionov, *On some properties of the progressing waves in a medium with non-linear heat conductivity and heat sources*, Zh. Vychisl. Mat. Mat. Fiz., 21 (1981), pp. 980–989.

[7] V. A. Galaktionov and A. A. Samarskii, *On difference solutions to one class of quasi-linear parabolic equations* I, Zh. Vychisl. Mat. Mat. Fiz., 23 (1983), pp. 646–659.

[8] V. A. Galaktionov and A. A. Samarskii, *On difference solutions to one class of quasi-linear parabolic equations* II, Zh. Vychisl. Mat. Mat. Fiz., 23 (1983), pp. 831–838.

[9] J. L. Graveleau and P. Jamet, *A finite difference approach to some degenerate nonlinear parabolic equations*, SIAM J. Appl. Math., 20 (1971), pp. 199–223.

[10] M. E. Gurtin and R. C. MacCamy, *On the diffusion of biological populations*, Math. Biosci., 33 (1977), pp. 35–49.

[11] M. A. Herrero and J. L. Vázquez, *The one-dimensional nonlinear heat equation with absorption: Regularity of solutions and interfaces*, SIAM J. Math. Anal., 18 (1987), pp. 149–167.

[12] D. Hoff, *A scheme for computing solutions and interface curves for a doubly-degenerate parabolic equation*, SIAM J. Numer. Anal., 22 (1985), pp. 687–712.

[13] A. S. Kalashnikov, *The propagation of disturbances in problems of non-linear heat conduction with absorption*, Zh. Vychisl. Mat. Mat. Fiz., 14 (1974), pp. 891–905.

[14] A. S. Kalashnikov, *Some problems of the qualitative theory of non-linear degenerate second-order parabolic equations*, Russian Math. Surveys, 42 (1987), pp. 169–222.

[15] R. Kersner, *The behavior of temperature fronts in media with nonlinear thermal conductivity under absorption*, Vestnik. Mosk. Univ. Ser. I Mat. Mekh., 33 (1978), pp. 44–51.

[16] R. Kersner, *Degenerate parabolic equations with general nonlinearities*, Nonlinear Anal., 4 (1980), pp. 1043–1062.

[17] R. Kersner, *Nonlinear heat conduction with absorption: Space localization and extinction in finite time*, SIAM J. Appl. Math., 43 (1983), pp. 1274–1285.

[18] B. F. Knerr, *The behavior of the support of solutions of the equation of nonlinear heat conduction with absorption in one dimension*, Trans. Amer. Math. Soc., 249 (1979), pp. 409–424.

[19] S. P. Kurdyumov, E. S. Kurkina, H. H. Malinetskii, and A. A. Samarskii, *Non-stationary dissipative structures in non-linear two-component media with volumetric sources*, Dokl. Acad. Nauk SSSR, 258 (1981), pp. 1084–1088.

[20] S. P. Kurdyumov, M. I. Gurevich, and O. V. Tel'kovskaya, *Automodel solutions to the quasi-linear heat equation with spaced density and non-linear volumetric sources*, Differ. Uravn., 31 (1995), pp. 1722–1733.

[21] M. Mimura, T. Nakaki, and K. Tomoeda, *A numerical approach to interface curves for some nonlinear diffusion equations*, Japan J. Appl. Math., 1 (1984), pp. 93–139.

[22] T. Nakaki, *Numerical interfaces in nonlinear diffusion equations with finite extinction phenomena*, Hiroshima Math. J., 18 (1988), pp. 373–397.

[23] O. A. Oleinik, A. S. Kalashnikov, and Y.-L. Chzou, *The Cauchy problem and boundary value problems for equations of the type of nonstationary filtration*, Izv. Acad. Nauk SSSR Ser. Mat., 22 (1958), pp. 667–704.

[24] P. Rosenau and S. Kamin, *Thermal waves in an absorbing and convecting medium*, Phys. D, 8 (1983), pp. 273–283.

[25] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Peaking Modes in Problems for Quasi-Linear Parabolic Equations*, Nauka, Moscow, 1987.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 965–994

MINIMIZERS OF COST-FUNCTIONS INVOLVING NONSMOOTH DATA-FIDELITY TERMS. APPLICATION TO THE PROCESSING OF OUTLIERS*

MILA NIKOLOVA†

Abstract. We present a theoretical study of the recovery of an unknown vector $x \in \mathbb{R}^p$ (such as a signal or an image) from noisy data $y \in \mathbb{R}^q$ by minimizing with respect to x a regularized cost-function $\mathcal{F}(x, y) = \Psi(x, y) + \alpha\Phi(x)$, where Ψ is a data-fidelity term, Φ is a smooth regularization term, and $\alpha > 0$ is a parameter. Typically, $\Psi(x, y) = \|Ax - y\|^2$, where A is a linear operator. The data-fidelity terms Ψ involved in regularized cost-functions are generally smooth functions; only a few papers make an exception to this and they consider restricted situations. Nonsmooth data-fidelity terms are avoided in image processing. In spite of this, we consider both smooth and nonsmooth data-fidelity terms. Our goal is to capture essential features exhibited by the local minimizers of regularized cost-functions in relation to the smoothness of the data-fidelity term.

In order to fix the context of our study, we consider $\Psi(x, y) = \sum_i \psi(a_i^T x - y_i)$, where a_i^T are the rows of A and ψ is $\mathcal{C}^m$ on $\mathbb{R} \setminus \{0\}$. We show that if $\psi'(0^-) < \psi'(0^+)$, then typical data y give rise to local minimizers $\hat{x}$ of $\mathcal{F}(., y)$ which fit exactly a certain number of the data entries: there is a possibly large set $\hat{h}$ of indexes such that $a_i^T \hat{x} = y_i$ for every $i \in \hat{h}$. In contrast, if ψ is smooth on $\mathbb{R}$, for almost every y, the local minimizers of $\mathcal{F}(., y)$ do not fit any entry of y. Thus, the possibility that a local minimizer fits some data entries is due to the nonsmoothness of the data-fidelity term. This is a strong mathematical property which is useful in practice. By way of application, we construct a cost-function allowing aberrant data (outliers) to be detected and to be selectively smoothed. Our numerical experiments advocate the use of nonsmooth data-fidelity terms in regularized cost-functions for special purposes in image and signal processing.

Key words. inverse problems, MAP estimation, nonsmooth analysis, perturbation analysis, proximal analysis, reconstruction, regularization, stabilization, outliers, total variation, variational methods

AMS subject classifications. 49N45, 62H12, 49J52, 49N60, 94A12, 94A08, 35A15, 68U10, 26B10

PII. S0036142901389165

1. Introduction.

We consider the general problem where a sought vector (e.g., an image or a signal) $\hat{x} \in \mathbb{R}^p$ is obtained from noisy data $y \in \mathbb{R}^q$ by minimizing a regularized cost-function $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the form

$$\mathcal{F}(x, y) = \Psi(x, y) + \alpha\Phi(x), \tag{1}$$

where typically $\Psi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is a data-fidelity term and $\Phi : \mathbb{R}^p \to \mathbb{R}$ is a regularization term, with $\alpha > 0$ a parameter. In many applications, the relation between x and y is modeled by $y_i = a_i^T x + n_i$ for $i = 1, \ldots, q$, where $a_i^T : \mathbb{R}^p \to \mathbb{R}$ are linear operators and n_i accounts for perturbations. We focus on such situations and assume that a_i^T, $i = 1, \ldots, q$, are known and non-null. The relevant data-fidelity term assumes the form

$$\Psi(x, y) = \sum_{i=1}^{q} \psi_i(a_i^T x - y_i), \tag{2}$$

*Received by the editors May 9, 2001; accepted for publication (in revised form) December 28, 2001; published electronically August 8, 2002.
http://www.siam.org/journals/sinum/40-3/38916.html
†CNRS URA820–ENST Dpt. TSI, ENST, 46 rue Barrault, 75013 Paris, France (nikolova@tsi.enst.fr).

where $\psi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, q$, are continuous functions which decrease on $(-\infty, 0]$ and increase on $[0, +\infty)$. Usually, $\psi_i = \psi$ for all i. One usual choice is $\psi(t) = |t|^\rho$, for $\rho > 0$, which yields [31, 4]

$$\Psi(x, y) = \sum_{i=1}^{q} |a_i^T x - y_i|^\rho. \tag{3}$$

Let $A \in \mathbb{R}^{q \times p}$ be the matrix whose rows are a_i^T for $i = 1, \ldots, q$. This matrix can be ill-posed, or singular, or invertible. Most often, $\Psi(x, y) = \|Ax - y\|^2$, that is, $\psi(t) = t^2$. Such data-fidelity terms are currently used in denoising, in deblurring, and in numerous inverse problems [37, 35, 13, 33, 1, 14, 38]. In a statistical framework, Ψ accounts for both the distortion and the noise intervening between the original x and the device recording the data y. The above quadratic form of Ψ corresponds to white Gaussian noise $\{n_i\}$. Recall that many papers are dedicated to the minimization of $\Psi(., y)$ *alone* and of the form (3), i.e., $\mathcal{F} = \Psi$, mainly for $\psi(t) = t^2$ [22], in some cases for $\psi(t) = |t|$ [8], but functions $\psi(t) = |t|^\rho$ for different values for ρ in the range $(0, \infty]$ also have been considered [31, 30]. Specific data-fidelity terms arise in applications such as emission and transmission computed tomography, X-ray radiography, eddy-currents evaluation, and many others [23, 20, 34, 10]. In general, for every y, the data-fidelity term $\Psi(., y)$ is a function which is smooth and usually convex. The introduction of nonsmooth data-fidelity terms in regularized cost-functions (1) remains very unusual. Only a few papers make an exception to this; we cite [2, 3], where Ψ corresponds to $\psi(t) = |t|$ and $a_i^T x = x_i$ for all i. Nonsmooth data-fidelity terms Ψ are avoided in image processing, for instance. In spite of this, we analyze the effects produced by both smooth and nonsmooth data-fidelity terms Ψ. In the latter case we suppose that $\{\psi_i\}$ *are any functions which are* $\mathcal{C}^m$*-smooth on* $\mathbb{R} \setminus \{0\}$*,* $m \geq 2$*, whereas at zero they admit finite side derivatives which satisfy* $\psi_i'(0^-) < \psi_i'(0^+)$.

The regularization term Φ usually takes the form

$$\Phi(x) = \sum_{i=1}^{r} \varphi(\|G_i^T x\|), \tag{4}$$

where $G_i^T : \mathbb{R}^p \to \mathbb{R}^s$ for $s \in \mathbb{N}^*$ are linear operators, e.g., operators yielding the differences between neighboring samples; $\|.\|$ stands for a norm on $\mathbb{R}^s$; and $\varphi : \mathbb{R} \to \mathbb{R}$ is a potential function. In a Bayesian estimation framework, Φ is the prior energy of the unknown x modeled using a Markov random field [6, 17, 24]. Several customarily used potential functions φ are [20, 29, 21, 33, 9, 7, 39, 36]

$$\begin{array}{ll}
\mathrm{L}^\nu & \varphi(t) = |t|^\nu,\ 1 \leq \nu \leq 2, \\
\text{Lorentzian} & \varphi(t) = \nu t^2/(1 + \nu t^2), \\
\text{Concave} & \varphi(t) = \nu|t|/(1 + \nu|t|), \\
\text{Gaussian} & \varphi(t) = 1 - \exp(-\nu t^2), \\
\text{Huber} & \varphi(t) = t^2 \text{ if } |t| \leq \nu, \quad \varphi(t) = \nu(\nu + 2|t - \nu|) \text{ if } |t| > \nu, \\
\text{Mean-field} & \varphi(t) = -\log\big(\exp(-\nu t^2) + 1\big),
\end{array} \tag{5}$$

where $\nu > 0$ is a parameter. Being convex and differentiable, the function L^ν for $1 < \nu \leq 2$ is preferred in many applications requiring intensive computation [9, 10]. In our paper, Φ *in* (1) *is any* $\mathcal{C}^m$*-smooth function*, with $m \geq 2$.

The visual aspect of a minimizer of a cost-function is determined on the one hand by the data and on the other hand by the shape of the cost-function. Our goal is to

capture essential features expressed by the local minimizers of cost-functions of the form (1)–(2) in relation to the smoothness of the data-fidelity term Ψ. Note that all our results hold for local minimizers, and hence for global minimizers as well, so we systematically speak of local minimizers. There is a striking distinction in the behavior of the local minimizers relevant to smooth and nonsmooth data-fidelity terms. It concerns the possibility of fitting *exactly* a certain number of the data entries, i.e., that for y given, a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ satisfies $a_i^T\hat{x} = y_i$ for some, or even for many, indexes i (see section 2). Intuitively, one is unlikely to obtain such minimizers, especially when data are noisy. *Our main result states that for $\mathcal{F}$ of the form* (1)–(2)*, with Ψ nonsmooth as specified, typical data y give rise to local minimizers $\hat{x}$ which fit a certain number of the data entries; i.e., there is a nonempty set $\hat{h}$ of indexes such that $a_i^T\hat{x} = y_i$ for every $i \in \hat{h}$* (see sections 3 and 4). This effect is due to the nondifferentiability of Ψ since it cannot occur when $\mathcal{F}$ is differentiable (see section 5). The obtained result is a strong mathematical property which can be used in different ways. Based on it, we construct a cost-function allowing aberrant data (outliers) to be detected and to be selectively smoothed from signals, or from images, or from noisy data, while preserving efficiently all the nonaberrant entries (see section 7). This is illustrated using numerical experiments.

Readers may associate cost-functions where Ψ is nonsmooth (e.g., $\psi(t) = |t|$) with cost-functions where Ψ is smooth and Φ is nonsmooth, e.g., $\Psi(x,y) = \|Ax - y\|^2$ and $\varphi(t) = |t|$ in (4), as in total-variation methods [33, 1, 14, 12]. Since the latter methods arouse an increasing interest in the area of image and signal restoration, we compare in section 6 nonsmooth regularization to the cost-functions considered in this paper. To this end, we use some previous results [26, 27] and illustrate the strikingly different visual effects they produce (see section 7).

2. The problem of an exact fit for some data entries. We shall use the symbol $\|.\|$ to denote the ℓ_2-norm of vectors. Next, we denote by $\mathbb{N}^*$ the positive integers and $\mathbb{R}_+ = \{t \in \mathbb{R} : t \geq 0\}$. The letter S will systematically denote the centered, unit sphere in $\mathbb{R}^n$, say $S := \{x \in \mathbb{R}^n : \|x\| = 1\}$, for whatever dimension n is appropriate in the context. For $x \in \mathbb{R}^n$ and $\rho > 0$, we put $B(x,\rho) := \{x' \in \mathbb{R}^n : \|x' - x\| < \rho\}$. For any $i = 1, \ldots, n$ the letter e_i represents the ith vector of the canonical basis of $\mathbb{R}^n$ (i.e., $e_i = e_i[i] = 1$ and $e_i[j] = 0$ for all $j \neq i$). The closure of a set N will be denoted $\overline{N}$. For a subspace T; its orthogonal complement is denoted $T^\perp$. If a function $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ depends on two variables, its kth differential with respect to the jth variable is denoted $D_j^k f$. The notation $f \in \mathcal{C}^m(N)$ means that the function f is $\mathcal{C}^m$-smooth on the set N. For a discrete, finite set $h \subset \{1, \ldots, n\}$, with $n \in \mathbb{N}^*$, the symbol $\#h$ is the cardinality of h and h^c is the complementary of h. Next we introduce a set-valued function which is constantly evoked in what follows.

DEFINITION 1. *Let $\mathcal{H}$ be the function which for every $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ yields the following set:*

$$(x,y) \quad \to \quad \mathcal{H}(x,y) = \left\{ i \in \{1, \ldots, q\} : a_i^T x = y_i \right\}. \tag{6}$$

Given y and a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$, the set of all data entries which are fitted exactly by $\hat{x}$ reads $\hat{h} := \mathcal{H}(\hat{x},y)$. Furthermore, with every $h \subseteq \{1, \ldots, q\}$ we associate the following sets:

$$(h,y) \to \Theta_h(y) := \{x \in \mathbb{R}^p : \ a_i^T x = y_i \ \forall i \in h \text{ and } a_i^T x \neq y_i \ \forall i \in h^c\}, \tag{7}$$

$$h \quad \to \quad T_h := \{u \in \mathbb{R}^p : \ a_i^T u = 0 \ \ \forall\, i \in h\}, \tag{8}$$

$$h \quad \to \quad M_h := \{(x,y) \in \mathbb{R}^p \times \mathbb{R}^q : \ a_i^T x = y_i \ \forall i \in h \text{ and } a_i^T x \neq y_i \ \forall i \in h^c\}. \tag{9}$$

Note that for every y and $h \neq \emptyset$, the sets $\Theta_h(y)$ and M_h are composed of a finite number of connected components, whereas their closures $\overline{\Theta_h(y)}$ and $\overline{M_h}$, respectively, are affine subspaces. The family of all Θ_h, when h ranges over all the subsets of $\{1,\ldots,q\}$, forms a partition of $\mathbb{R}^p$. Observe that for $y \in \mathbb{R}^q$ fixed, $\{x \in \mathbb{R}^p : (x,y) \in M_h\} = \Theta_h(y)$. Notice also the equivalences

$$\mathcal{H}(x',y') = h \;\Leftrightarrow\; x' \in \Theta_h(y') \;\Leftrightarrow\; (x',y') \in M_h. \tag{10}$$

The theory in this paper is developed by analyzing how the local minimizers of every $\mathcal{F}(.,y)$ behave under small variations of the data y. We thus consider local minimizer functions.

Definition 2. *Let $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ and $N \subseteq \mathbb{R}^q$. The family $f(.,N) := \{f(.,y) : y \in N\}$ is said to admit a local minimizer function $\mathcal{X} : N \to \mathbb{R}^p$ if for any $y \in N$ the function $f(.,y)$ has a strict local minimum at $\mathcal{X}(y)$.*

The next lemma addresses local minimizer functions relevant to smooth cost-functions.

Lemma 1. *Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q$ be a $\mathcal{C}^m$-function with $m \geq 2$. For $y \in \mathbb{R}^q$, assume that $\hat{x} \in \mathbb{R}^p$ is such that $D_1\mathcal{F}(\hat{x},y) = 0$, and $D_1^2\mathcal{F}(\hat{x},y)$ is positive definite.*

Then there exists a neighborhood $N \subset \mathbb{R}^q$ containing y and a $\mathcal{C}^{m-1}$-function $\mathcal{X} : N \to \mathbb{R}^p$ such that for every $y' \in N$ we have $D_1\mathcal{F}(\mathcal{X}(y'),y') = 0$, and $D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is positive definite. In particular, $\hat{x} = \mathcal{X}(y)$.

Equivalently, $\mathcal{X} : N \to \mathbb{R}^p$ is a local minimizer function relevant to $\mathcal{F}(.,N)$ such that $D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is positive definite for every $y' \in N$.

Proof. Being a local minimizer of $\mathcal{F}(.,y)$, $\hat{x}$ satisfies $D_1\mathcal{F}(\hat{x},y) = 0$. We focus on the equation $D_1\mathcal{F}(x',y') = 0$ in the vicinity of $(\hat{x},y)$ and notice that $D_1^2\mathcal{F}(\hat{x},y)$ determines an isomorphism from $\mathbb{R}^p$ to itself. From the implicit functions theorem [5], there exist $\rho_1 > 0$ and a unique $\mathcal{C}^{m-1}$-function $\mathcal{X} : B(y,\rho_1) \to \mathbb{R}^p$ such that $D_1\mathcal{F}\left(\mathcal{X}(y'),y'\right) = 0$ for all $y' \in B(y,\rho_1)$. Furthermore, since $y' \to \det D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is continuous and $\det D_1^2\mathcal{F}(\hat{x},y) > 0$, there is $\rho_2 \in (0,\rho_1]$ such that $\det D_1^2\mathcal{F}(\mathcal{X}(y'),y') > 0$ for all $y' \in B(y,\rho_2)$. □

Remark 1 (on the conditions required in Lemma 1). The minimizers of $\mathcal{C}^m$-functions of the form

$$\mathcal{F}(x,y) = \|Ax - y\|^2 + \alpha\Phi(x)$$

are extensively studied in [16]. It is shown there that if $\mathrm{rank}A = p$, and under some assumptions ensuring that $\mathcal{F}(.,y)$ admits local minimizers for every $y \in \mathbb{R}^q$, the data domain $\mathbb{R}^q$ contains a subset N whose interior is dense in $\mathbb{R}^q$ such that for every $y \in N$, then every local minimizer $\hat{x}$ of the corresponding $\mathcal{F}(.,y)$ is strict and $D_1^2\mathcal{F}(\hat{x},y)$ is positive definite. Reciprocally, all data leading to minimizers at which the conditions of Lemma 1 fail belong to a closed negligible subset of $\mathbb{R}^q$: the chance of acquiring data placed in such subsets is null.

The central question of this paper is how the shape of a cost-function $\mathcal{F}$ favors, or inhibits, the possibility that a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$, for $y \in \mathbb{R}^q$, fits a certain number of the entries of this same y, i.e., that the set $\hat{h} := \mathcal{H}(\hat{x},y)$ is nonempty. It will appear that this possibility is closely related to the smoothness of Ψ. We recall some facts about nonsmooth functions [32].

Definition 3. *Let $E_0 \subseteq \mathbb{R}^p$ be an affine subspace and E be the relevant vector space. Consider a function $f : E_0 \to \mathbb{R}$, and let $x \in E_0$ and $u \in E$. The function f admits a one-sided derivative at x in the direction of $u \neq 0$, denoted by $\delta g(x)(u)$, if*

the following (possibly infinite) limit exists:

$$\delta f(x)(u) := \lim_{t \downarrow 0} \frac{f(x+tu) - f(x)}{t}.$$

If $u = 0$, *put* $\delta f(x)(0) = 0$.

The downward pointing arrow above means that $t \in \mathbb{R}_+$ converges to zero by positive values. If f is differentiable at x, then $\delta f(x)(u) = Df(x).u$. If $f : \mathbb{R} \to \mathbb{R}$, we have $\delta f(x)(1) = f'(x^+)$. The left derivative of f at x for u is $-\delta f(x)(-u)$. In the following, $\delta_1 \mathcal{F}$ will address one-sided derivatives of $\mathcal{F}$ with respect to its first argument.

3. Cost-functions with nonsmooth data-fidelity terms. Here and in section 4 we focus on cost-functions which read

$$\mathcal{F}(x, y) = \Psi(x, y) + \alpha \Phi(x, y), \tag{11}$$

$$\Psi(x, y) = \sum_{i=1}^{q} \psi(a_i^T x - y_i), \tag{12}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is $\mathcal{C}^m$ on $\mathbb{R} \setminus \{0\}$, with $m \geq 2$, whereas at zero it admits finite side derivatives satisfying $\psi'(0^-) < \psi'(0^+)$. The term $\Phi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is any $\mathcal{C}^m$-function. This formulation allows us to address data-fidelity terms composed of a nonsmooth function Ψ and of a smooth function $\tilde{\Psi}$, since we can write $\Phi(x, y) = \tilde{\Psi}(x, y) + \tilde{\Phi}(x)$ with $\tilde{\Phi}$ a regularization term. For example, we can have $\Phi(x, y) = \sum_i \big(\phi_i(B_i^T x - y_{q_i}) + \varphi_i(G_i^T x)\big)$, where $\phi_i : \mathbb{R}^{q_i} \to \mathbb{R}$ and $\varphi_i : \mathbb{R}^{p_i} \to \mathbb{R}$ are $\mathcal{C}^m$-functions, $y_{q_i} \in \mathbb{R}^{q_i}$ are data, and $B_i^T \in \mathbb{R}^{q_i \times p}$ and $G_i^T \in \mathbb{R}^{p_i \times p}$, with $p_i \in \mathbb{N}^*$ and $q_i \in \mathbb{N}^*$.

Remark 2. The results presented in sections 3 and 4 are developed for Ψ of the form (12), that is, $\psi_i = \psi$ for all i, but we should emphasize that they remain true for Ψ of the form (2) provided that all ψ_i, for $i = 1, \ldots, q$, have finite side derivatives at zero satisfying $\psi_i'(0^-) < \psi_i'(0^+)$. The proofs are straightforward to extend to this situation but at the expense of complicated notation which may cloud the presentation.

We start by providing a sufficient condition for a strict local minimum.

PROPOSITION 1. *For* $y \in \mathbb{R}^q$, *let* $\mathcal{F}(., y) : \mathbb{R}^p \to \mathbb{R}$ *be of the form* (11)–(12), *where* $\Phi \in \mathcal{C}^m(\mathbb{R}^p \times \mathbb{R}^q)$ *for* $m \geq 1$ *and* $\psi \in \mathcal{C}^m(\mathbb{R} \setminus \{0\})$ *satisfies* $-\infty < \psi'(0^-) < \psi'(0^+) < +\infty$. *Let* $\hat{x} \in \mathbb{R}^p$ *be such that*

1. *the restricted function* $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}}(., y) : \overline{\Theta_{\hat{h}}(y)} \to \mathbb{R}$ *reaches a strict local minimum at* $\hat{x}$,
2. $\delta_1 \mathcal{F}(\hat{x}, y)(u) > 0$ *for all* $u \in T_{\hat{h}}^{\perp} \cap S$,

where $\hat{h} := \mathcal{H}(\hat{x}, y)$, $\Theta_{\hat{h}}(y)$, *and* $T_{\hat{h}}$ *are determined according to* (6), (7), *and* (8), *respectively.*

Then $\mathcal{F}(., y)$ *reaches a strict local minimum at* $\hat{x}$.

Proof. The result is a tautology if $\hat{h} = \emptyset$ since then $\Theta_{\hat{h}}(y) = \mathbb{R}^p$. So consider that $\hat{h}$ is nonempty. First of all, we put $\mathcal{F}$ into a more convenient form. Define

$$\tilde{\psi}(t) := \psi(t) - \frac{t}{2}\big(\psi'(0^-) + \psi'(0^+)\big) - \psi(0). \tag{13}$$

Now we have

$$\tilde{\psi}'(0^+) = -\tilde{\psi}'(0^-) > 0 \quad \text{and} \quad \tilde{\psi}(0) = 0, \tag{14}$$

which will allow important simplifications. By means of $\tilde{\psi}$, the cost-function $\mathcal{F}$ assumes the form

$$\mathcal{F}(x,y) = \tilde{\Psi}(x,y) + \tilde{\Phi}(x,y), \tag{15}$$

$$\text{where } \tilde{\Psi}(x,y) = \sum_{i=1}^{q} \tilde{\psi}(a_i^T x - y_i)$$

$$\text{and } \tilde{\Phi}(x,y) = \sum_{i=1}^{q} \frac{\psi'(0^-) + \psi'(0^+)}{2}(a_i^T x - y_i) + q\psi(0) + \alpha\Phi(x,y).$$

Both $\tilde{\Psi}$ and $\tilde{\Phi}$ satisfy the assumptions about Ψ and Φ, respectively. Henceforth, we deal with the formulation of $\mathcal{F}$ given in (15). For notational convenience, we systematically write ψ for $\tilde{\psi}$, Ψ for $\tilde{\Psi}$, and Φ for $\tilde{\Phi}$.

Let us consider the altitude increment of $\mathcal{F}(.,y)$ at $\hat{x}$ in the direction of an arbitrary $u \in S$,

$$\mathcal{F}(\hat{x} + tu, y) - \mathcal{F}(\hat{x}, y) \quad \text{for } t \in \mathbb{R}_+.$$

In order to avoid misunderstandings, u_0 will denote a vector of $T_{\hat{h}}$ and $u_{\perp}$ a vector of $T_{\hat{h}}^{\perp}$. Using the fact that every $u \in S$ has a unique decomposition into

$$u = u_0 + u_{\perp} \text{ with } u_0 \in T_{\hat{h}} \cap \overline{B(0,1)} \text{ and } u_{\perp} \in T_{\hat{h}}^{\perp} \cap \overline{B(0,1)}, \tag{16}$$

we decompose the altitude increment of $\mathcal{F}(.,y)$ accordingly:

$$\mathcal{F}(\hat{x} + tu, y) - \mathcal{F}(\hat{x}, y) = \mathcal{F}(\hat{x} + tu_0 + tu_{\perp}, y) - \mathcal{F}(\hat{x} + tu_0, y) \tag{17}$$

$$+ \mathcal{F}(\hat{x} + tu_0, y) - \mathcal{F}(\hat{x}, y). \tag{18}$$

The term on the right-hand side of (17) is analyzed with the aid of assumption 2. In order to calculate the side derivative $\delta_1\mathcal{F}(\hat{x},y)$, we decompose $\mathcal{F}$ into

$$\mathcal{F}(x',y') = \Psi_{\hat{h}}(x',y') + \mathcal{F}_{\hat{h}}(x',y'), \tag{19}$$

$$\text{where } \Psi_{\hat{h}}(x',y') := \sum_{i \in \hat{h}} \psi(a_i^T x' - y_i')$$

$$\text{and } \mathcal{F}_{\hat{h}}(x',y') = \sum_{i \in \hat{h}^c} \psi(a_i^T x - y_i') + \alpha\Phi(x',y').$$

This decomposition is used recurrently in the following.

Remark 3. The function $\mathcal{F}_{\hat{h}}$ is $\mathcal{C}^m$ on a neighborhood of $(\hat{x},y)$ which contains $B(\hat{x},\sigma) \times B(y,\sigma)$ for

$$\sigma := \frac{1}{2(\|a\|_\infty + 1)} \min_{i \in \hat{h}^c} |a_i^T \hat{x} - y_i|, \tag{20}$$

$$\|a\|_\infty := \max_{i=1}^{q} \|a_i\|. \tag{21}$$

Indeed, for every $(x',y') \in B(\hat{x},\sigma) \times B(y,\sigma)$ we have

$$i \in \hat{h}^c \quad \Rightarrow \quad |a_i^T x' - y_i'| = \left|(a_i^T \hat{x} - y_i) + a_i^T(x' - \hat{x}) + (y_i - y_i')\right| \tag{22}$$

$$\geq \left|a_i^T \hat{x} - y_i\right| - \left|a_i^T(x' - \hat{x})\right| - |y_i - y_i'|$$

$$\geq \min_{i \in \hat{h}^c} |a_i^T \hat{x} - y_i| - \|a\|_\infty \sigma - \sigma = (\|a\|_\infty + 1)\sigma > 0,$$

since clearly $\|a\|_\infty > 0$ and $\sigma > 0$.

In contrast, $\Psi_{\hat{h}}$ is nonsmooth at $(\hat{x}, y)$. Using Definition 3 we calculate that for every $u \in \mathbb{R}^p$,

$$\delta_1 \mathcal{F}(x,y)(u) = \delta_1 \Psi_{\hat{h}}(\hat{x},y)(u) + D\mathcal{F}_{\hat{h}}(\hat{x},y).u, \tag{23}$$

$$\text{where } \ \delta_1 \Psi_{\hat{h}}(\hat{x},y)(u) = \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u|, \tag{24}$$

since $\delta\psi(a_i^T\hat{x} - y_i)(u) = \lim_{t\downarrow 0} \psi(ta_i^T u)/t = \psi'(0^+)|a_i^T u|$, for every $i \in \hat{h}$, which accounts for (14). Notice that $\delta_1\Psi_{\hat{h}}(\hat{x},y)(u) = \delta_1\Psi_{\hat{h}}(\hat{x},y)(-u) \geq 0$ for every $u \in \mathbb{R}^p$. Applying assumption 2 to both $u_\perp \in T_{\hat{h}}^\perp$ and $-u_\perp$ yields

$$|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp| < \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_\perp| \ \ \forall u_\perp \in T_{\hat{h}}^\perp. \tag{25}$$

Now consider the function

$$f : T_{\hat{h}}^\perp \cap S \to \mathbb{R},$$
$$u_\perp \to f(u_\perp) := \frac{|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp|}{\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_\perp|}.$$

Since for every $u_\perp \in T_{\hat{h}}^\perp \cap S$ there is at least one index $i \in \hat{h}$ such that $a_i^T u_\perp \neq 0$, this function is well defined and continuous. If $u_\perp \to D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp$ is not identically null on $T_{\hat{h}}^\perp$, put

$$c_0 := \sup_{u_\perp \in T_{\hat{h}}^\perp \cap S} f(u_\perp). \tag{26}$$

Since $T_{\hat{h}}^\perp \cap S$ is compact, f reaches the maximum value c_0. By (25) we see that $0 < c_0 < 1$. If $D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp = 0$ for all $u_\perp \in T_{\hat{h}}^\perp$, we put $c_0 := 1/2$. In both cases,

$$|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp| \leq c_0\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_\perp| \ \ \forall u_\perp \in T_{\hat{h}}^\perp. \tag{27}$$

Using (19), the right-hand side of (17) takes the form

$$\mathcal{F}(\hat{x}+tu_0+tu_\perp, y) - \mathcal{F}(\hat{x}+tu_0, y) = \Psi_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) - \Psi_{\hat{h}}(\hat{x}+tu_0, y) \tag{28}$$
$$+ \mathcal{F}_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) - \mathcal{F}_{\hat{h}}(\hat{x}+tu_0, y). \tag{29}$$

First, we focus on the right-hand side of (28). From the definition of $\hat{h}$ and (16),

$$\Psi_{\hat{h}}(\hat{x}+tu_0, y) = 0,$$
$$\Psi_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) = \sum_{i\in\hat{h}} \psi\left(a_i^T(\hat{x}+tu_\perp+tu_0) - y_i\right) = \sum_{i\in\hat{h}} \psi(ta_i^T u_\perp).$$

Applying Definition 3 to $\psi'(0^+)$ shows that there is $\eta_0 \in (0, \sigma]$ such that

$$\frac{\psi(t)}{t} \geq \psi'(0^+) - \frac{1-c_0}{2}\psi'(0^+) \ \ \forall t \in (0, \|a\|_\infty \eta_0)\,,$$

since $(1-c_0)/2 \in (0,1)$. On the other hand, $|a_i^T u| \leq \|a_i\|\|u\| \leq \|a\|_\infty$ for all $u \in \overline{B(0,1)}$ and for all $i \in \{1,\ldots,q\}$. Then

$$t \in (0,\eta_0) \;\Rightarrow\; \psi(ta_i^T u_{\perp\!\!\!\perp}) \geq \frac{c_0+1}{2}\psi'(0^+)\, t\, |a_i^T u_{\perp\!\!\!\perp}| \;\; \forall u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp} \cap \overline{B(0,1)}.$$

Hence, taking $t \in (0,\eta_0)$ ensures that for all $u \in S$, decomposed into $u = u_0 + u_{\perp\!\!\!\perp}$ as in (16), we have

$$\Psi_{\hat{h}}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) \geq \frac{c_0+1}{2}\, t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}|. \tag{30}$$

Second, we consider (29). Define the constants

$$c_1 := \min_{u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp} \cap S} \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}|, \tag{31}$$

$$c_2 := c_1 \psi'(0^+) \frac{1-c_0}{4}, \tag{32}$$

and notice that $c_1 > 0$ and $c_2 > 0$, and that (31) implies

$$\sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| \geq c_1 \|u_{\perp\!\!\!\perp}\| \quad \forall u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp}. \tag{33}$$

Since $\mathcal{F}_{\hat{h}}(.,y) \in \mathcal{C}^1\left(B(\hat{x},\sigma)\right)$ (see Remark 3), the mean-value theorem [5] shows that for every $u \in S$ and for every $t \in [0,\sigma)$ there exists $\theta \in (0,1)$ such that

$$\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) - \mathcal{F}_{\hat{h}}(\hat{x} + tu_0, y) = tD_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}, \tag{34}$$

where $u = u_0 + u_{\perp\!\!\!\perp}$ is decomposed as in (16). Moreover, there is $\eta_1 \in (0,\eta_0)$ such that for every $t \in (0,\eta_1)$,

$$\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp} - D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| \leq c_2\|u_{\perp\!\!\!\perp}\| \quad \forall u \in S, \;\; \forall \theta \in (0,1),$$

and hence

$$\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}\right| \leq \left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| + c_2\|u_{\perp\!\!\!\perp}\| \;\; \forall u \in S, \;\; \forall \theta \in (0,1). \tag{35}$$

Starting with (28)–(29), we derive

$$\begin{aligned}
&\mathcal{F}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) - \mathcal{F}(\hat{x} + tu_0, y) \qquad (36)\\
&\geq \frac{c_0+1}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - t\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}\right| \quad \text{[by (30) and (34)]}\\
&\geq \frac{c_0+1}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - t\left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (35)]}\\
&\geq \frac{1-c_0}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (27)]}\\
&\geq \frac{1-c_0}{2} \psi'(0^+) tc_1\|u_{\perp\!\!\!\perp}\| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (33)]}\\
&= \frac{1-c_0}{4} \psi'(0^+) tc_1\|u_{\perp\!\!\!\perp}\|. \quad \text{[by (32)]} \qquad (37)
\end{aligned}$$

Consequently,

$$(38)\ t \in (0,\eta_1) \quad \Rightarrow \quad \mathcal{F}(\hat{x}+tu_0+tu_\perp,y)-\mathcal{F}(\hat{x}+tu_0,y)>0 \quad \forall u\in S \text{ with } u_\perp \neq 0.$$

From assumption 1, there exists $\eta_2 \in (0,\eta_1]$ such that

$$(39)\quad t\in(0,\eta_2) \quad \Rightarrow \quad \mathcal{F}(\hat{x}+tu_0,y)-\mathcal{F}(\hat{x},y)>0 \quad \forall u_0 \in T_{\hat{h}} \cap \overline{B(0,1)} \setminus \{0\}.$$

If $u_0=0$, then (38) holds since $\|u_\perp\|=1$, whereas if $u_\perp=0$, then (39) is true since $\|u_0\|=1$. Introducing (38) and (39) into (17)–(18) shows that if $t\in(0,\eta_2)$, then $\mathcal{F}(\hat{x}+tu,y)-\mathcal{F}(\hat{x},y)>0$ for every $u\in S$. ∎

Remark 4. The conditions required in Proposition 1 are pretty weak. Indeed, if an arbitrary function $\mathcal{F}(.,y):\mathbb{R}^p\to\mathbb{R}$ has a strict minimum at $\hat{x}$, then assumption 1 is trivially true and necessarily $\delta_1\mathcal{F}(\hat{x},y)(u)\geq 0$ for all $u\in T_{\hat{h}}^{\perp}\cap S$ [32]. In comparison, assumption 2 requires only that the latter inequality be strict.

Observe that the above sufficient condition for strict minimum concerns the behavior of $\mathcal{F}(.,y)$ on two orthogonal subspaces *separately*. This occurs because of the nonsmoothness of ψ.

4. Minimizers that fit exactly some data entries. The theorem below states the main contribution of this work.

THEOREM 1. *Consider $\mathcal{F}$ as given in* (11)–(12)*, where $\Phi\in\mathcal{C}^m(\mathbb{R}^p\times\mathbb{R}^q)$ for $m\geq 2$, and $\psi\in\mathcal{C}^m(\mathbb{R}\setminus\{0\})$ has finite side derivatives at zero such that $\psi'(0^-)<\psi'(0^+)$. Given $y\in\mathbb{R}^q$ and $\hat{x}\in\mathbb{R}^p$, let $\hat{h}:=\mathcal{H}(\hat{x},y)$, $\Theta_{\hat{h}}(y)$, and $T_{\hat{h}}$ be obtained by* (6)*,* (7)*, and* (8)*, respectively. Suppose the following:*

1. *The set $\{a_i : i\in\hat{h}\}$ is linearly independent;*
2. *for every $u\in T_{\hat{h}}\cap S$ we have $D_1(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}})(\hat{x},y).u=0$ and $D_1^2(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}})(\hat{x},y)(u,u)>0$;*
3. *for every $u\in T_{\hat{h}}^{\perp}\cap S$ we have $\delta_1\mathcal{F}(\hat{x},y)(u)>0$.*

Then there is a neighborhood $N\subset\mathbb{R}^q$ containing y and a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}:N\to\mathbb{R}^p$ relevant to $\mathcal{F}(.,N)$ (see Definition 2) yielding, in particular, $\hat{x}=\mathcal{X}(y)$, whereas for every $y'\in N$,

$$(40)\qquad \begin{aligned} a_i^T\mathcal{X}(y') &= y'_i \quad \text{if} \quad i\in\hat{h},\\ a_i^T\mathcal{X}(y') &\neq y'_i \quad \text{if} \quad i\in\hat{h}^c. \end{aligned}$$

The latter means that $\mathcal{H}(\mathcal{X}(y'),y')=\hat{h}$ is constant on N.

Proof. If $\hat{h}=\emptyset$, then $\Theta_{\hat{h}}(y')=\mathbb{R}^p$ for all y'. Applying Lemma 1 shows the existence of $\tilde{N}\subset\mathbb{R}^q$ and of a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}$ relevant to $\mathcal{F}(.,\tilde{N})$. By the continuity of $\mathcal{X}$, there is $N\subset\tilde{N}$ where (40) holds, in which case (40) is reduced to $a_i^T\mathcal{X}(y')\neq y'_i$ for all $i\in\{1,\dots,q\}$.

In the following we consider that $\hat{h}$ is nonempty. As in the proof of Proposition 1, we use the formulation of $\mathcal{F}$ given in (13)–(15) and write ψ for $\tilde{\psi}$ and Φ for $\tilde{\Phi}$. This proof is based on two lemmas given next.

LEMMA 2. *Let assumptions* 1 *and* 2 *of Theorem* 1 *be satisfied. Then there exist $\nu>0$ and a $\mathcal{C}^{m-1}$-function $\mathcal{X}:B(y,\nu)\to\mathbb{R}^p$ so that for every $y'\in B(y,\nu)$ the point $\hat{x}':=\mathcal{X}(y')$ belongs to $\Theta_{\hat{h}}(y')$ and satisfies*

$$(41)\qquad D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\hat{x}',y').u=0 \ \text{ and } \ D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\hat{x}',y')(u,u)>0 \quad \forall u\in T_{\hat{h}}\setminus\{0\}.$$

In particular, $\hat{x} = \mathcal{X}(y)$.

Proof of Lemma 2. We start by commenting on the restricted functions in (41).

Remark 5. For σ as in (20), the inequality reached in (22) shows that for all $(x', y') \in B(\hat{x}, \sigma) \times B(y, \sigma)$ we have $\mathcal{H}(x', y') \subseteq \hat{h}$. On the other hand, if $x' \in \overline{\Theta_{\hat{h}}(y')}$, then $\mathcal{H}(x', y') \supseteq \hat{h}$. If we put

$$B_{\hat{h}}\left((\hat{x}, y), \sigma\right) := \left(B(\hat{x}, \sigma) \times B(y, \sigma)\right) \cap \overline{M_{\hat{h}}}, \tag{42}$$

where $M_{\hat{h}}$ is given in (9), we have

$$(x', y') \in B_{\hat{h}}\left((\hat{x}, y), \sigma\right) \quad \Rightarrow \quad \mathcal{H}(x', y') = \hat{h},$$

and $B_{\hat{h}}\left((\hat{x}, y), \sigma\right) \subset M_{\hat{h}}$. By (7) and (10), for every $(x', y') \in M_{\hat{h}}$ we find $\Psi_{\hat{h}}(x', y') = 0$ and hence $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y') = \mathcal{F}_{\hat{h}}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y')$. Since $\mathcal{F}_{\hat{h}} \in \mathcal{C}^m\left(B(\hat{x}, \sigma) \times B(y, \sigma)\right)$ (see Remark 3), we get

$$\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}} \in \mathcal{C}^m\left(B_{\hat{h}}\left((\hat{x}, y), \sigma\right)\right) \text{ and } \mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y') = \mathcal{F}_{\hat{h}}(x', y') \ \forall\ (x', y') \in B_{\hat{h}}\left((\hat{x}, y), \sigma\right).$$

We now pursue the proof of the lemma. Let the indexes contained in $\hat{h}$ read $\hat{h} = \{j_1, \ldots, j_{\#\hat{h}}\}$. Let $I_{\hat{h}}$ be the $\#\hat{h} \times q$ matrix with entries $I_{\hat{h}}[i, j_i] = 1$ for $i = 1, \ldots, \#\hat{h}$, the remaining entries being null. Thus $y_{\hat{h}} := I_{\hat{h}} y \in \mathbb{R}^{\#\hat{h}}$ is composed of only those entries of y whose indexes are in $\hat{h}$. Similarly, put $A_{\hat{h}} := I_{\hat{h}} A$; then $A_{\hat{h}} \in \mathbb{R}^{\#\hat{h} \times p}$ and $A_{\hat{h}} \hat{x} = y_{\hat{h}}$. With this notation, $\overline{M_{\hat{h}}} = \left\{(x', y') \in \mathbb{R}^p \times \mathbb{R}^q : A_{\hat{h}} x' - I_{\hat{h}} y' = 0\right\}$. By assumption 1, $\mathrm{rank} A_{\hat{h}} = \#\hat{h}$. Then for every y' we have the following dimensions: $\dim \overline{\Theta_{\hat{h}}(y')} = \dim T_{\hat{h}} = p - \#\hat{h}$ while $\dim \overline{M_{\hat{h}}} = p - \#\hat{h} + q$. Recalling that $A_{\hat{h}} A_{\hat{h}}^T$ is invertible, put

$$P_{\hat{h}} := A_{\hat{h}}^T \left(A_{\hat{h}} A_{\hat{h}}^T\right)^{-1} I_{\hat{h}}. \tag{43}$$

Let $C_{\hat{h}} : T_{\hat{h}} \to \mathbb{R}^{p - \#\hat{h}}$ be an isomorphism. The affine mapping

$$\begin{aligned} \Gamma : \quad & \overline{M_{\hat{h}}} \to \mathbb{R}^{p - \#\hat{h}}, \\ & (x', y') \to \Gamma(x', y') = C_{\hat{h}}\left(x' - \hat{x} - P_{\hat{h}}(y' - y)\right) \end{aligned} \tag{44}$$

is well defined for every $y' \in \mathbb{R}^q$ since on the one hand $\hat{x} + P_{\hat{h}}\left(y' - y\right)$ is the orthogonal projection[1] of $\hat{x}$ onto $\overline{\Theta_{\hat{h}}(y')}$, whereas on the other hand $x' \in \overline{\Theta_{\hat{h}}(y')}$ by (10). Consider also the conjugate mapping

$$\begin{aligned} \Gamma^\dagger : \quad & \mathbb{R}^{p - \#\hat{h}} \times \mathbb{R}^q \to \overline{\Theta_{\hat{h}}(y')}, \\ & (z, y') \to \Gamma^\dagger(z, y') = C_{\hat{h}}^{-1} z + \hat{x} + P_{\hat{h}}(y' - y), \end{aligned} \tag{45}$$

[1] The orthogonal projection of $\hat{x}$ onto $\overline{\Theta_{\hat{h}}(y')}$, denoted by $\hat{x}_{y'}$, is unique and is determined by solving the problem

$$\text{minimize } \ \|\hat{x}_{y'} - \hat{x}\| \quad \text{subject to } \ \hat{x}_{y'} \in \overline{\Theta_{\hat{h}}(y')}.$$

The latter constraint also reads $A_{\hat{h}} \hat{x}_{y'} = y'_{\hat{h}}$ if we denote $y'_{\hat{h}} = I_{\hat{h}} y'$. It is easily calculated that the solution to this problem reads

$$\hat{x}_{y'} = \hat{x} - A_{\hat{h}}^T \left(A_{\hat{h}} A_{\hat{h}}^T\right)^{-1} \left(A_{\hat{h}} \hat{x} - y'_{\hat{h}}\right).$$

Recalling that $A_{\hat{h}} \hat{x} = I_{\hat{h}} y$ from the definition of $\hat{h}$, we obtain that $\hat{x}_{y'} = \hat{x} + P_{\hat{h}}\left(y' - y\right)$.

which is also well defined. Let

$$\nu_0 := \frac{\sigma}{2}\,\min\left\{1,\,\left(\sup_{z\in S}\|C_{\hat{h}}^{-1}z\| + \sup_{y'\in S}\|P_{\hat{h}}y'\|\right)^{-1}\right\}. \tag{46}$$

Clearly, $0 < \nu_0 < \sigma$. It is worth noticing that

$$\Gamma^\dagger(z,y') \in \overline{\Theta_{\hat{h}}(y')} \cap B(\hat{x},\sigma) \subset \Theta_{\hat{h}}(y') \quad \forall (z,y') \in B(0,\nu_0)\times B(y,\nu_0), \tag{47}$$

since on the one hand (45) shows that $\Gamma^\dagger(z,y') \in \overline{\Theta_{\hat{h}}(y')}$ whereas, on the other hand,

$$\|\Gamma^\dagger(z,y') - \hat{x}\| \le \|C_{\hat{h}}^{-1}\|\,\|z\| + \|P_{\hat{h}}\|\,\|y'-y\| \le (\|C_{\hat{h}}^{-1}\| + \|P_{\hat{h}}\|)\,\nu_0 < \sigma.$$

Now we introduce the function

$$\begin{aligned} &\mathcal{G} : \mathbb{R}^{p-\#\hat{h}} \times \mathbb{R}^q \to \mathbb{R}, \\ &\qquad (z,y') \to \mathcal{G}(z,y') := \mathcal{F}_{\hat{h}}\left(\Gamma^\dagger(z,y'),\ y'\right). \end{aligned} \tag{48}$$

Since for every $y' \in \mathbb{R}^q$ we have

$$z = \Gamma(x',y') \quad \Leftrightarrow \quad x' = \Gamma^\dagger(z,y'),$$

then

$$\mathcal{G}\left(\Gamma(x',y'),y'\right) = \mathcal{F}_{\hat{h}}(x',y') = \mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x',y') \quad \forall (x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right),$$

where the last equality comes from Remark 5. Now for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right)$, the derivatives of $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}$, mentioned in (41), can be calculated in terms of $\mathcal{G}$ and Γ as follows:

$$D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(x',y').u_0 = D_1\mathcal{G}\left(\Gamma(x',y'),y'\right).C_{\hat{h}}u_0 \qquad \forall u_0 \in T_{\hat{h}}, \tag{49}$$

$$D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}}\right)(x',y')(u_0,u_0) = D_1^2\mathcal{G}\left(\Gamma(x',y'),y'\right).\left(C_{\hat{h}}u_0, C_{\hat{h}}u_0\right) \qquad \forall u_0 \in T_{\hat{h}}. \tag{50}$$

Since $C_{\hat{h}}$ is an isomorphism, $D_1\Gamma(x',y').u_0 = C_{\hat{h}}.u_0 \neq 0$ for every $u_0 \in T_{\hat{h}} \setminus \{0\}$, whereas $C_{\hat{h}}.T_{\hat{h}} = \mathbb{R}^{p-\#\hat{h}}$. Then assumption 2, combined with the fact that $\Gamma(\hat{x},y) = 0$ by construction, yields

$$\begin{aligned} D_1\mathcal{G}(0,y) &= 0, \\ D_1^2\mathcal{G}(0,y)(u,u) &> 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}. \end{aligned}$$

By Lemma 1, there exist $\nu \in (0,\nu_0]$ and a unique $\mathcal{C}^{m-1}$-function $\mathcal{Z} : B(y,\nu) \to B(0,\nu_0)$ such that

$$D_1\mathcal{G}\left(\mathcal{Z}(y'),\ y'\right) = 0 \ \text{ and } \ D_1^2\mathcal{G}\left(\mathcal{Z}(y'),\ y'\right) \text{ is positive definite } \quad \forall y' \in B(y,\nu), \tag{51}$$

with, in particular, $\mathcal{Z}(y) = 0$. Next we express the derivatives in (51) in terms of $\mathcal{F}_{\hat{h}}$ and $\Gamma^\dagger$. From (47) and Remark 5 it follows that $\mathcal{F}_{\hat{h}}$ is $\mathcal{C}^m$ at every $\left(\Gamma^\dagger(z,y'),y'\right)$ relevant to $(z,y') \in B(0,\nu_0)\times B(y,\nu)$, in which case (48) gives rise to

$$D_1\mathcal{G}(z,y').u = D_1\mathcal{F}_{\hat{h}}(\Gamma^\dagger(z,y'),y').C_{\hat{h}}^{-1}u, \tag{52}$$

$$D_1^2\mathcal{G}(z,y')(u,u) = D_1^2\mathcal{F}_{\hat{h}}(\Gamma^\dagger(z,y'),y')\left(C_{\hat{h}}^{-1}u, C_{\hat{h}}^{-1}u\right). \tag{53}$$

Put

$$\mathcal{X}(y') := \Gamma^{\dagger}\left(\mathcal{Z}(y'), y'\right) \quad \forall y' \in B(y,\nu), \tag{54}$$

and notice that $\mathcal{X}(y') \in \Theta_{\hat{h}}(y')$. Then (51) implies that for every $y' \in B(y,\nu)$,

$$D_1\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y').C_{\hat{h}}^{-1}u = 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}},$$
$$D_1^2\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y')\left(C_{\hat{h}}^{-1}u, C_{\hat{h}}^{-1}u\right) > 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}.$$

Since $C_{\hat{h}}^{-1}u \neq 0$ for all $u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}$ and $C_{\hat{h}}^{-1}.\mathbb{R}^{p-\#\hat{h}} = T_{\hat{h}}$, it follows that for every $y' \in B(y,\nu)$,

$$D_1\mathcal{F}_{\hat{h}}\left(\mathcal{X}(y'),\ y'\right).u_0 = 0 \ \text{ and } \ D_1^2\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y').(u_0,u_0) > 0 \ \ \forall u_0 \in T_{\hat{h}} \setminus \{0\}.$$

Again applying Remark 5 allows us to write that if $y' \in B(y,\nu)$, then

$$D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)\left(\mathcal{X}(y'),\ y'\right).u_0 = 0 \ \text{ and } \ D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\mathcal{X}(y'), y')(u_0,u_0) > 0$$
$$\forall u_0 \in T_{\hat{h}} \setminus \{0\}.$$

The proof of Lemma 2 is complete. □

The next lemma addresses assumption 3 of the theorem.

LEMMA 3. *Given $\hat{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, let $\hat{h} = \mathcal{H}(\hat{x}, y) \neq \emptyset$. Let assumption 3 of Theorem 1 hold.*

Then there exists $\mu > 0$ such that

$$y' \in B(\hat{x},\mu) \ \text{ and } \ x' \in \Theta_{\hat{h}}(y') \cap B(\hat{x},\mu) \ \Rightarrow \ \delta_1\mathcal{F}(x',y')(u_{\perp}) > 0 \ \ \forall u_{\perp} \in T_{\hat{h}}^{\perp} \cap S. \tag{55}$$

Proof of Lemma 3. We decompose $\mathcal{F}$ according to (19). Let σ and $B_{\hat{h}}\left((\hat{x},y),\sigma\right)$ be defined according to (20) and (42), respectively. Remark 5 applies to $B_{\hat{h}}\left((\hat{x},y),\sigma\right)$. Similarly to (23)–(24), for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right)$ we have

$$\delta_1\mathcal{F}(x',y')(u) = \psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u| + D_1\mathcal{F}_{\hat{h}}(x',y').u \quad \forall u \in \mathbb{R}^p. \tag{56}$$

By the continuity of $D_1\mathcal{F}_{\hat{h}}$, there is $\mu \in (0,\sigma]$ such that for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\mu\right)$,

$$\left|D_1\mathcal{F}_{\hat{h}}(x',y').u_{\perp} - D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp}\right| \leq \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| \quad \forall u_{\perp} \in T_{\hat{h}}^{\perp}, \tag{57}$$

where $c_0 \in (0,1)$ and $c_1 > 0$ are the constants given in (26) and (31), respectively. We derive the following inequality chain which holds for all $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\mu\right)$ and for all $u_{\perp} \in T_{\hat{h}}^{\perp}$:

$$\begin{aligned}
&\left|D_1\mathcal{F}_{\hat{h}}(x',y').u_{\perp}\right| \\
&\leq \left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp}\right| + \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| && \text{[by (57)]} \\
&\leq c_0\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| + \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| && \text{[by (27)]} \qquad (58)\\
&\leq c_0\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| + \frac{1-c_0}{2}\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| && \text{[by (33)]} \\
&= \frac{c_0+1}{2}\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}|. && (59)
\end{aligned}$$

On the other hand, (56) shows that for every $(x', y') \in B_{\hat{h}}((\hat{x}, y), \mu)$ and for all $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$, we have

$$\begin{aligned}\delta_1 \mathcal{F}(x', y')(u_{\perp}) &\geq \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u_{\perp}| - |D_1 \mathcal{F}_{\hat{h}}(x', y').u_{\perp}| \\ &\geq \left(1 - \frac{c_0 + 1}{2}\right) \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u_{\perp}| > 0. \qquad \text{[by (59)]}\end{aligned}$$

The last inequality is strict since for every $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$, there is at least one index $i \in \hat{h}$ for which $a_i^T u_{\perp} \neq 0$. □

We now complete the proof of Theorem 1. Consider $\nu > 0$ and $\mu > 0$ the radii found in Lemmas 2 and 3 and $\mathcal{X}$ the function exhibited in Lemma 2. By the continuity of $\mathcal{X}$, there exists $\xi \in (0, \min\{\mu, \nu\}]$ such that $\mathcal{X}(y') \in B(\hat{x}, \mu)$ for every $y' \in B(y, \xi)$. For any $y' \in B(y, \xi)$, consider the point $\hat{x}' := \mathcal{X}(y')$. From Lemma 2, $\hat{x}' \in \Theta_{\hat{h}}(y')$ and $\hat{x}'$ is a strict local minimizer of $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(., y')$. From Lemma 3, $\delta_1 \mathcal{F}(\hat{x}', y')(u_{\perp}) > 0$ for all $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$. All the conditions of Proposition 1 being satisfied, $\mathcal{F}(., y')$ reaches a strict local minimum at $\hat{x}'$. It follows that $\mathcal{X} : B(y, \xi) \to \mathbb{R}^p$ is the sought-after $\mathcal{C}^{m-1}$ minimizer function. □

We now focus on the assumptions involved in this theorem. Assumption 2 is nothing else but the very classical sufficient condition for a strict local minimum of a smooth function over an affine subspace. Assumption 3 was used in Proposition 1 and was discussed therein.

Remark 6 (on assumption 1). The subset $\{a_i : i \in \hat{h}\}$ in assumption 1 is determined by (6). With the notation introduced in the beginning of Lemma 2, $y_{\hat{h}} := I_{\hat{h}} y \in \mathbb{R}^{\#\hat{h}}$ belongs to the range of $A_{\hat{h}}$, denoted by $\mathcal{R}(A_{\hat{h}})$. Since $\dim \mathcal{R}(A_{\hat{h}}) = \text{rank} A_{\hat{h}}$, it follows that if $\text{rank} A_{\hat{h}} < \#\hat{h}$, then all $y'_{\hat{h}}$ belonging to $\mathcal{R}(A_{\hat{h}})$ belong to a subspace of dimension strictly smaller than $\#\hat{h}$. Thus, assumption 1 fails to hold only if y is included in a subspace of dimension smaller than q. But the chance that noisy data y belong to such a subspace is null. Hence, assumption 1 is satisfied for almost all $y \in \mathbb{R}^q$.

It is worth emphasizing that the independence of the whole set $\{a_i : i \in \{1, \ldots, q\}\}$ is not required. Thus, Theorem 1 addresses any matrix A whether it be ill conditioned, or singular, or invertible.

Theorem 1 entails some important consequences which are discussed next.

Remark 7 (stability of minimizers). The fact that there is a $\mathcal{C}^{m-1}$ local minimizer function shows that, in spite of the nonsmoothness of $\mathcal{F}$, for any y, all the strict local minimizers of $\mathcal{F}(., y)$ which satisfy the conditions of the theorem are *stable under weak perturbations of data* y. This result extends Lemma 1 to nonsmooth functions of the form (11)–(12). Moreover, if for every $y \in \mathbb{R}^q$ the function $\mathcal{F}(., y)$ is strictly convex, then the unique minimizer function $\mathcal{X} : \mathbb{R}^q \to \mathbb{R}^p$, relevant to $\mathcal{F}(., \mathbb{R}^q)$, is $\mathcal{C}^{m-1}$ on $\mathbb{R}^q$.

Remark 8 (stability of $\hat{h}$). The result formulated in (40) means that *the set-valued function* $y' \to \mathcal{H}(\mathcal{X}(y'), y')$ *is constant on* N, *i.e., that* $\mathcal{H}$ *is constant under small perturbations of* y. Equivalently, *all residuals* $(a_i^T \mathcal{X}(y') - y'_i)$ *for* $i \in \hat{h}$ *are null on* N.

Remark 9 (data domain). Theorem 1 reveals that the data domain $\mathbb{R}^q$ contains *volumes of positive measure* composed of data that lead to local minimizers which

fit exactly the data entries belonging to the same set (e.g., for A invertible, $\alpha = 0$ yields $\hat{h} = \{1, \ldots, q\}$ and the data volume relevant to this $\hat{h}$ is $\mathbb{R}^q$). For a meaningful choice of ψ, Φ, and α, there are volumes corresponding to various $\hat{h}$, and they are large enough so that noisy data come across them. That is why in practice, nonsmooth data-fidelity terms yield minimizers fitting exactly a certain number of the data entries. The resultant numerical effect is observed in section 7.

Next we present a simple example which illustrates Theorem 1.

Example 1 (nonsmooth data-fidelity term). Consider the function

$$\mathcal{F}(x, y) = \sum_{i=1}^{q} |x_i - y_i| + \alpha \sum_{i=1}^{q} \frac{x_i^2}{2},$$

where $\alpha > 0$. For every $y \in \mathbb{R}^q$, the function $\mathcal{F}(., y)$ is strictly convex, so it has a unique minimizer and the latter is strict. Moreover,

$$\min_x \mathcal{F}(x, y) = \sum_{i=1}^{q} \min_{x_i} f(x_i, y_i),$$
$$\text{where} \quad f(x_i, y_i) = |x_i - y_i| + \frac{\alpha x_i^2}{2} \quad \text{for} \quad i = 1, \ldots, q.$$

For $y \in \mathbb{R}^q$, let $\hat{x}$ be the minimizer of $\mathcal{F}(., y)$. Now $\hat{h} = \{i : \hat{x}_i = y_i\}$. For every i, the fact that $f(., y_i)$ has a minimum at $\hat{x}_i$ means that $\delta_1 f(\hat{x}_i, y_i)(u) \geq 0$ for every $u \in \mathbb{R}$. Then for every $u \in \mathbb{R}$ we have

$$\text{if} \ \ (i \in \hat{h}^c \ \Leftrightarrow \ \hat{x}_i \neq y_i), \ \text{then} \ \delta_1 f(x_i, y_i)(u) = Df(x_i, y_i).u = (\operatorname{sign}(x_i - y_i) + \alpha x_i) .u \geq 0;$$
$$\text{if} \ \ (i \in \hat{h} \ \Leftrightarrow \ \hat{x}_i = y_i), \ \text{then} \ \delta_1 f(\hat{x}_i, y_i)(u) = |u| + (\alpha y_i) .u \geq 0.$$

From Proposition 1, the entries of the minimizer function $\mathcal{X}$ are

$$\text{if} \ \ |y_i| > \frac{1}{\alpha}, \quad \text{then} \quad \mathcal{X}_i(y) = \frac{1}{\alpha} \operatorname{sign}(y_i);$$
$$\text{if} \ \ |y_i| \leq \frac{1}{\alpha}, \quad \text{then} \quad \mathcal{X}_i(y) = y_i.$$

Theorem 1 applies, provided that $|y_i| \neq 1/\alpha$ for every $i \in \hat{h}$, which corresponds to assumption 3. In such a case, we can take for the neighborhood exhibited in Theorem 1

$$N = B(y, \xi) \ \text{ with } \ \xi = \min_{i=1}^{q} \left| \ |y_i| - \frac{1}{\alpha} \right|.$$

We see that $y' \to \mathcal{H}(\mathcal{X}(y'), y')$ reads

$$\mathcal{H}(\mathcal{X}(y'), y') = \left\{ i \in \{1, \ldots, q\} : |y'_i| \leq \frac{1}{\alpha} \right\}$$

and is constant on N. The above expression shows also that the cardinality of $\hat{h}$ increases when α decreases.

We now illustrate Remark 9. For $h \subset \{1, \ldots, q\}$, put

$$V_h := \left\{ y \in \mathbb{R}^q : \ |y_i| \leq \frac{1}{\alpha} \ \forall i \in h \ \text{ and } \ |y_i| > \frac{1}{\alpha} \ \forall i \in h^c \right\}.$$

Obviously, every $y' \in V_h$ gives rise to a minimizer $\hat{x}'$ of $\mathcal{F}(.,y')$ satisfying $\mathcal{H}(\hat{x}',y') = h$. That is, the function $y' \to \mathcal{H}(\mathcal{X}(y'),y')$ is constant on V_h. Note that $V_\emptyset = \{y \in \mathbb{R}^q : |y_i| > 1/\alpha \text{ for all } i\}$ and that $V_\emptyset = \emptyset$ if $\alpha = 0$. Moreover, *for every* $h \subset \{1,\dots,q\}$, *the set* V_h *has a positive volume in* $\mathbb{R}^q$, whereas the family of all V_h, when h ranges over the family of all the subsets of $\{1,\dots,q\}$ (including the empty set), is a *partition* of $\mathbb{R}^q$.

5. Smooth data-fidelity terms. In this section we focus on smooth cost-functions with the goal of checking whether we can get minimizers which fit exactly a certain number of data entries. We start with an illuminating example.

Example 2 (smooth cost-function). For $A \in \mathbb{R}^{q\times p}$ and $G \in \mathbb{R}^{r\times p}$ with $r \in \mathbb{N}^*$, consider the cost-function $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$,

$$\mathcal{F}(x,y) = \|Ax - y\|^2 + \alpha\|Gx\|^2. \tag{60}$$

Recall that since the publication of [37], cost-functions of this form are among the most widely used tools in signal and image estimations [25, 22, 35, 13]. Under the classical assumption $\ker A^TA \cap \ker G^TG = \emptyset$, it is seen that for every $y \in \mathbb{R}^q$, $\mathcal{F}(.,y)$ is strictly convex and its unique minimizer $\hat{x}$ is determined by solving the equation

$$D_1\mathcal{F}(\hat{x},y) = 0 \quad \text{where} \quad D_1\mathcal{F}(\hat{x},y) = 2(A\hat{x} - y)^TA + 2\alpha\hat{x}^TG^TG.$$

The relevant minimizer function $\mathcal{X} : \mathbb{R}^q \to \mathbb{R}^p$ reads

$$\mathcal{X}(y) = (A^TA + \alpha\, G^TG)^{-1}A^T.\, y. \tag{61}$$

We now determine the set of *all* data points $y \in \mathbb{R}^q$ for which $\hat{x} := \mathcal{X}(y)$ fits exactly the ith data entry y_i. To this end, we have to solve with respect to y the equation

$$a_i^T\,\mathcal{X}(y) = y_i. \tag{62}$$

Using (61), this is equivalent to solving the equation

$$\begin{aligned} &p_i(\alpha).y = 0, \\ \text{where} \quad &p_i(\alpha) \quad = a_i^T(A^TA + \alpha\, G^TG)^{-1}A^T - e_i^T. \end{aligned} \tag{63}$$

We can have $p_i(\alpha) = 0$ only if α belongs to the discrete set of several values which satisfy a data-independent system of q polynomials of degree p. However, α will almost never belong to such a set so, in general, $p_i(\alpha) \neq 0$. Then (63) implies $y \in \{p_i(\alpha)\}^\perp$. More generally, we have the implication

$$\exists i \in \{1,\dots,q\} \ \text{ such that } \ \mathcal{X}_i(y) = y_i \quad \Rightarrow \quad y \in \bigcup_{j=1}^{q} \{p_j(\alpha)\}^\perp.$$

Since every $\{p_i(\alpha)\}^\perp$ is a subspace of $\mathbb{R}^q$ of dimension $q-1$, the union on the right-hand side above is a *closed, negligible subset of* $\mathbb{R}^q$. The chance that noisy data come across this union is null. Hence, the chance that noisy data y yield a minimizer $\mathcal{X}(y)$ which fits even one data entry, i.e., that there is at least one index i such that (62) holds, is null.

The theorem stated below generalizes this example.

THEOREM 2. *Consider a* $\mathcal{C}^m$*-function* $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$, *with* $m \geq 2$, *of the form* (1)–(2), *and let* $h \subset \{1\dots,q\}$ *be nonempty. Assume the following:*

1. *For all* $i = 1, \dots, q$, *the functions* $\psi_i : \mathbb{R} \to \mathbb{R}$ *satisfy* $\psi_i''(t) > 0$ *for all* $t \in \mathbb{R}$;
2. A *is invertible* (*recall that for every* $i = 1, \dots, q$, *the* i*th row of* A *is* a_i^T);
3. *there is an open domain* $N_0 \subset \mathbb{R}^q$ *so that* $\mathcal{F}(., N_0)$ *admits a* $\mathcal{C}^{m-1}$ *local minimizer function* $\mathcal{X} : N_0 \to \mathbb{R}^p$, *such that* $D_1^2\mathcal{F}(\mathcal{X}(y), y)$ *is positive definite, for all* $y \in N_0$;
4. *for every* $x \in \mathcal{X}(N_0) \subset \mathbb{R}^p$ *and for every* $i \in h$ *we have* $D^2\Phi(x).[A^{-1}]_i \neq 0$, *where* $[A^{-1}]_i$ *denotes the* i*th column of* A^{-1}, *for* $i = 1, \dots, q$.

For a given set of constants $\{\theta_i, i \in h\}$, *and for any* $N \subset N_0$ *a closed subset of* $\mathbb{R}^q$, *put*

$$\Upsilon_h := \left\{ y \in N : a_i^T \mathcal{X}(y) = y_i + \theta_i \ \forall i \in h \right\}. \tag{64}$$

Then Υ_h *is a closed subset of* $\mathbb{R}^q$ *whose interior is empty.*

Proof. For every h nonempty we have

$$\Upsilon_h = \bigcap_{i \in h} \Upsilon_{\{i\}}.$$

It is hence sufficient to consider $\Upsilon_{\{i\}}$ for some $i \in h$. For simplicity, in the following we write Υ_i for $\Upsilon_{\{i\}}$. Since $\mathcal{X}$ is continuous on N, every Υ_i is closed in N and hence in $\mathbb{R}^q$. Our reasoning below is developed *ad absurdum*. So suppose that Υ_i contains an open, connected subset of $\mathbb{R}^q$, say $\tilde{N} \subset \Upsilon_i \subset N$. We can hence write

$$a_i^T \mathcal{X}(y) = y_i + \theta_i \quad \forall y \in \tilde{N}. \tag{65}$$

Differentiating both sides of this identity with respect to y yields

$$a_i^T D\mathcal{X}(y) = e_i^T \quad \forall y \in \tilde{N}. \tag{66}$$

We next determine the form of $D\mathcal{X}$. Since for every $y \in \tilde{N}$ the point $\mathcal{X}(y)$ is a local minimizer of $\mathcal{F}(., y)$, it satisfies $D_1\mathcal{F}(\mathcal{X}(y), y) = 0$. Differentiating both sides of the latter identity leads to

$$D_1^2\mathcal{F}\left(\mathcal{X}(y), y\right) D\mathcal{X}(y) + D_{1,2}\mathcal{F}\left(\mathcal{X}(y), y\right) = 0 \quad \forall y \in \tilde{N}. \tag{67}$$

The Hessian of $x \to \mathcal{F}(x, y)$, denoted $H(x, y) := D_1^2\mathcal{F}(x, y)$, reads

$$\begin{aligned} H(x, y) &= D_1^2\Psi(x, y) + \alpha D^2\Phi(x) \\ &= A^T \operatorname{Diag}\left(\ddot{\psi}(x, y)\right) A + \alpha D^2\Phi(x), \end{aligned} \tag{68}$$

where for every x and y, $\ddot{\psi}(x, y) \in \mathbb{R}^q$ is the vector whose entries read

$$[\ddot{\psi}(x, y)]_i = \psi_i''(a_i^T x - y_i) \ \text{ for } \ i = 1, \dots, q.$$

By assumption 3, $H\left(\mathcal{X}(y), y\right)$ is an invertible matrix for every $y \in \tilde{N}$. Furthermore,

$$D_{1,2}\mathcal{F}(x, y) = -A^T \operatorname{Diag}\left(\ddot{\psi}(x, y)\right).$$

Inserting the last expression and (68) into (67) shows that

$$D\mathcal{X}(y) = \left(H(\mathcal{X}(y), y)\right)^{-1} A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y), y)\right) \quad \forall y \in \tilde{N}. \tag{69}$$

Now introducing (69) into (66) yields

$$a_i^T \left(H(\mathcal{X}(y),y)\right)^{-1} A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right) = e_i^T \quad \forall y \in \tilde{N}. \tag{70}$$

By assumption 1, $\operatorname{Diag}\big(\ddot{\psi}(\mathcal{X}(y),y)\big)$ is invertible for every $y \in \tilde{N}$. Its inverse is a diagonal matrix whose diagonal terms are $\left(\psi_i''(a_i^T\mathcal{X}(y)-y_i)\right)^{-1}$ for $i=1\ldots,q$. Noticing that

$$e_i^T \left(\operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right)\right)^{-1} = \frac{e_i^T}{\psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)},$$

we find that (70) equivalently reads

$$\psi_i''(a_i^T\mathcal{X}(y)-y_i)\,.a_i^T \left(H(\mathcal{X}(y),y)\right)^{-1} = e_i^T A^{-T} \quad \forall y \in \tilde{N},$$

where $A^{-T} := \left(A^T\right)^{-1}$. Then, taking into account (68),

$$\psi_i''(a_i^T\mathcal{X}(y)-y_i)\,.a_i^T = e_i^T A^{-T}\left(A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right) A + \alpha D^2\Phi(\mathcal{X}(y))\right) \quad \forall y \in \tilde{N}.$$

By the invertibility of A (assumption 2), and noticing that $e_i^T A = a_i^T$, the latter expression is simplified to

$$\psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)\,.a_i^T = \psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)\,.a_i^T + \alpha e_i^T A^{-T} D^2\Phi(\mathcal{X}(y)) \quad \forall y \in \tilde{N},$$

and finally to

$$D^2\Phi(\mathcal{X}(y)).A^{-1}e_i = 0 \quad \forall y \in \tilde{N}.$$

However, the obtained identity contradicts assumption 4. Hence the conclusion. □

Let us comment on the assumptions taken in this theorem. Recall first that assumption 3 was discussed in Lemma 1 and Remark 1. In the typical case when Ψ is a data-fidelity measure, every ψ_i is a strictly convex function satisfying $\psi_i(0)=0$ and $\psi_i(t)=\psi_i(-t)$.

Remark 10 (on assumption 2). This proposition also addresses the case when

$$\mathcal{F}(x,y) = \|Ax-y\|^2 + \alpha\Phi(x) \;\text{ with }\; \operatorname{rank}A = p \le q.$$

Indeed, for $p<q$, $\mathcal{F}$ can equivalently be expressed in terms of an invertible $p\times p$ matrix $\tilde{A}$ with $\tilde{A}^T\tilde{A} = A^TA$ in place of A.

Remark 11 (on assumption 4). By the invertibility of A (assumption 2), we see that $[A^{-1}]_i = A^{-1}e_i \neq 0$ for every $i=1,\ldots,q$. It would be a "pathological" situation to have some of the columns of A^{-1} in $\ker D^2\Phi(x)$ for some x. For instance, focus on the classical case given in (4) with $G_i^T : \mathbb{R}^p \to \mathbb{R}$. Let G denote the $r\times p$ matrix whose rows are G_i^T for $i=1,\ldots,r$. Then $D^2\Phi(x) = G^T\operatorname{Diag}\left(\ddot{\varphi}(Gx)\right)G$, where $\ddot{\varphi}(Gx)\in\mathbb{R}^r$ is the vector with entries $[\ddot{\varphi}(Gx)]_i = \varphi''(G_i^Tx)$ for $i=1,\ldots,r$. Focus on the case when $\varphi''(t)>0$ for all $t\in\mathbb{R}$ (e.g., φ is strictly convex) and G yields first-order differences between neighboring samples. Then $\operatorname{Ker}D^2\Phi(x)$ is composed of the constant vectors $\kappa[1,\ldots,1]^T$, $\kappa\in\mathbb{R}$. Then assumption 4 is satisfied provided that A^{-1} does not involve constant columns.

Remark 12 (meaning of the theorem). If for some $y\in\mathbb{R}^q$ a minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ satisfies an affine equation of the form $a_i^T\hat{x} = y_i + \theta_i$, then Theorem 2 asserts that

y belongs to a closed subset of $\mathbb{R}^q$ whose interior is empty. There is no chance that noisy data y yield local minimizers of a smooth cost-function $\mathcal{F}(.,y)$ satisfying such an equation.

The next proposition states the same conclusions but under different assumptions.

PROPOSITION 2. *Consider a $\mathcal{C}^m$-function $\mathcal{F}:\mathbb{R}^p\times\mathbb{R}^q\to\mathbb{R}$, with $m\geq 2$, of the form (1)–(2) and let $h\subset\{1\ldots,q\}$ be nonempty. Assume the following:*

1. *There is a domain $N_0\subset\mathbb{R}^q$ so that $\mathcal{F}(.,N_0)$ admits a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}:N_0\to\mathbb{R}^p$ such that $D_1^2\mathcal{F}(\mathcal{X}(y),y)$ is positive definite for all $y\in N_0$;*
2. *for every $y\in N_0$ and for every $i\in h$ there exists $j\in\{1,\ldots,q\}$ such that the function $\mathcal{K}_{i,j}$,*

$$\mathcal{K}_{i,j}(y'):=\psi_i''\left(a_j^T\mathcal{X}(y')-e_j^Ty'\right).a_i^T\left(H(\mathcal{X}(y'),y')\right)^{-1}.a_j,$$

where H was given in (68), is nonconstant on any neighborhood of y.

For $\{\theta_i\in\mathbb{R}:i\in h\}$ given, and for every $N\subset N_0$ a closed subset of $\mathbb{R}^q$, put

$$\Upsilon_h:=\left\{y\in N:a_i^T\mathcal{X}(y)=y_i+\theta_i\ \forall i\in h\right\}. \tag{71}$$

Then Υ_h is a closed subset of $\mathbb{R}^q$ whose interior is empty.

Proof. As in the proof of Theorem 2, we focus on Υ_i for $i\in h$ and develop our reasoning by contradiction. So suppose that Υ_i contains an open ball $\tilde{N}$. Then (65) and (66) are true. In particular, comparing (66) for $y'\neq y$ with the same equality for y yields

$$a_i^TD\mathcal{X}(y')=a_i^TD\mathcal{X}(y)\quad\forall y'\in\tilde{N}. \tag{72}$$

Notice that $A^T\operatorname{Diag}\big(\ddot{\psi}(x,y')\big)$ is a matrix whose jth column reads $\psi''(a_j^Tx-y_j').a_j$. Introducing (69) into (72) shows that the latter is equivalent to the system

$$\mathcal{K}_{i,j}(y')=\mathcal{K}_{i,j}(y)\ \ \forall j\in\{1,\ldots,q\},\ \ \forall y'\in\tilde{N}.$$

The obtained result contradicts assumption 2. ∎

Remark 13 (on assumption 2). Although a general proof of the validity of this assumption appears to be more intricate than important, we conjecture that it is usually satisfied. The intuitive arguments are the following. Let us focus on the classical case when Φ is as in (4). The entries of $H(x',y')$ read

$$[H(x',y')]_{m,n}=\sum_{j=1}^q\eta_{j,m}^2\psi''(a_jx'-y_j')+\sum_{j=1}^r\kappa_{j,n}^2\varphi''(G_jx')\ \text{ for }\ (m,n)\in\{1,\ldots,p\}^2, \tag{73}$$

where $\eta_{j,m}$, $j=1,\ldots,q$, and $\kappa_{j,n}$, $j=1,\ldots,r$, are constants that are calculated from G and A. From Cramer's rule for matrix inversion, for every j, the term $a_i^T\left(H(x',y')\right)^{-1}a_j$ is the fraction of two polynomials. The entries of the numerator read $\beta_{s,m,n}([H(x',y')]_{m,n})^s$ for all $(m,n)\in\{1,\ldots,p\}^2$ with $\beta_{s,m,n}\in\mathbb{R}$ for $s=0,\ldots,p-1$. In the denominator we have $\gamma_{s,m,n}([H(x',y')]_{m,n})^s$ for all $(m,n)\in\{1,\ldots,p\}^2$ with $\gamma_{s,m,n}\in\mathbb{R}$ for $s=0,\ldots,p$. For $\mathcal{X}$ a minimizer function and j and i given, $\mathcal{K}_{i,j}$ has the form

$$\mathcal{K}_{i,j}(y')=\psi''\left(a_j^T\mathcal{X}(y')-y_j'\right).\frac{\sum_{s=1}^{p-1}\sum_{(m,n)}\beta_{s,m,n}([H(\mathcal{X}(y'),y')]_{m,n})^s}{\sum_{s=1}^{p}\sum_{(m,n)}\gamma_{s,m,n}([H(\mathcal{X}(y'),y')]_{m,n})^s}. \tag{74}$$

Assumption 2 requires that for $i\in h$, there is at least one index $j\in\{1,\ldots,q\}$ for which the relevant function $\mathcal{K}_{i,j}$ does not remain constant on any neighborhood of y.

6. Nonsmooth regularization versus nonsmooth data-fidelity. In this section we compare cost-functions involving nonsmooth data-fidelity terms to cost-functions involving nonsmooth regularization terms. The visual effects produced by these classes of cost-functions can be seen in section 7.

Cost-functions with *nonsmooth regularization* typically have the form (1), where Ψ is a $\mathcal{C}^m$-function, $m \geq 2$, whereas Φ is as in (4) with φ nonsmooth at zero. Most often, $\Psi(x, y) = \|Ax - y\|^2$. Nonsmooth functions φ are, for instance, the L^1- and concave functions in (5). Since the publication of [33, 18], such cost-functions are customarily used in signal and image restoration [18, 1, 14, 11, 12, 38]. Visually, the obtained minimizers exhibit a *staircasing effect* since they typically involve many constant regions—see, for instance, Figures 6 and 10 in section 8. This effect is discussed by many authors [18, 15, 14, 12]. In particular, the ability of the L^1-function to recover noncorrelated "nearly black" images in the simplest case when $G_i = e_i$ for all i was interpreted in [15] using mini-max decision theory. Total-variation methods, corresponding to $\varphi(t) = |t|$ also, were observed to yield "blocky images" [14, 12]. The concave function was shown to transform ramp-shaped data into a step-shaped minimizer [19].

A theoretical explanation of staircasing was given in [26, 27, 28]. It was shown there that regularization of the form (4) with φ nonsmooth at zero yields local minimizers $\hat{x}$ which satisfy $G_i^T \hat{x} = 0$ *exactly* for many indexes i. For instance, if G_i^T, $i = 1, \ldots, r$, yield first-order differences between neighboring samples (if x is a signal of $\mathbb{R}^p$, $G_i^T x = x_i - x_{i+1}$ for $i = 1, \ldots, p-1$), the relevant minimizers $\hat{x}$ are constant over many zones. If G_i^T, $i = 1, \ldots, r$, yield second-order differences, then $\hat{x}$ involves many zones over which it is affine, etc. More generally, the sets of indexes i for which $G_i^T \hat{x} = 0$ determine zones which can be said to be *strongly homogeneous* [27]. Staircasing is due to a special form of stability property which is explained next. Let a data point y give rise to a local minimizer $\hat{x}$ which satisfies $G_i^T \hat{x} = 0$ for all $i \in \hat{h}$, where $\hat{h} \neq \emptyset$. It is shown in [26, 27, 28] that y is in fact contained in a neighborhood $N \in \mathbb{R}^q$ whose elements $y' \in N$ (noisy data) give rise to local minimizers $\hat{x}'$ of $\mathcal{F}(., y')$, placed near $\hat{x}$, which satisfy $G_i^T \hat{x}' = 0$ for all $i \in \hat{h}$. Since every such N is a volume of positive measure, noisy data come across these volumes and yield minimizers satisfying $G_i^T \hat{x}' = 0$ for many indexes i. Notice that this behavior is due to the nonsmoothness of φ at zero since it cannot occur with differentiable cost-functions [27, 28].

The behavior of the minimizers of cost-functions with *nonsmooth data-fidelity*, as considered in Theorem 1, is opposite. If y leads to a minimizer $\hat{x}$ which fits exactly a set $\hat{h}$ of entries of y, Theorem 1 shows that y is contained in a neighborhood N such that the relevant minimizer function $\mathcal{X}$ follows closely every small variation of all data entries y'_i for $i \in \hat{h}$ when y' ranges over N. Thus $a_i^T \mathcal{X}(y')$ is never constant in the vicinity of y for $i \in \hat{h}$.

7. Nonsmooth data-fidelity to detect and smooth outliers. Our objective now is to process data in order to detect, and possibly to smooth, outliers and impulsive noise. To this end, take $a_i = e_i$ for every $i \in \{1, \ldots, q\}$ in (2). Focus on

$$\mathcal{F}(x, y) = \sum_{i=1}^{q} \psi(x_i - y_i) + \alpha \sum_{i=1}^{r} \varphi(G_i^T x), \tag{75}$$

where $G_i^T : \mathbb{R}^p \to \mathbb{R}$ for $i = 1, \ldots, r$ yield differences between neighboring samples (e.g., $G_i^T x = x_i - x_{i+1}$ if x is a signal); ψ and φ are even and strictly increasing on $[0, \infty)$, with $\psi'(0^+) > 0$ and φ smooth on $\mathbb{R}$. Suppose that $\hat{x}$ is a strict minimizer

of $\mathcal{F}(.,y)$ and put $\hat{h}=\mathcal{H}(\hat{x},y)$. Based on the results in section 4, we naturally come to the following method for the detection of outliers. Since every y_i corresponding to $i\in\hat{h}$ is kept intact in the minimizer $\hat{x}$, that is, $\hat{x}_i=y_i$, every such y_i can be considered as a *faithful data entry*. In contrast, every y_i with $i\in\hat{h}^c$ corresponds to $\hat{x}_i\neq y_i$ which can indicate that this y_i is aberrant. In other words, *given $y\in\mathbb{R}^q$, we posit that $\hat{h}^c$, the complementary of $\hat{h}=\mathcal{H}(\mathcal{X}(y),y)$, provides an estimate of the locations of the outliers in y*. The possibility of keeping intact all faithful data entries is both spectacular and valuable from a practical point of view, e.g., to preprocess data.

Remark 14 (stability of the detection of outliers). If a minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ for $y\in\mathbb{R}^q$ gives rise to $\hat{h}=\mathcal{H}(\hat{x},y)$, then Theorem 1 ensures that all data y' placed near y yield minimizers $\hat{x}'$ which recover exactly the same set of outlier positions $\hat{h}^c$. Hence, the suggested method for detection of outliers is stable under small data variations.

We also can envisage *smoothing* outliers since the value of every $\hat{x}_i$ for $i\in\hat{h}^c$ is obtained from the values of neighboring data samples through the terms $\alpha\varphi(G_j^T\hat{x})$ for all j neighbor of i. Small values of α make the weight of Ψ more important, so the relevant minimizers $\hat{x}$ fit larger sets of data entries, i.e., $\hat{h}$ is larger. At the same time, all samples $\hat{x}_i$ for $i\in\hat{h}^c$ incur an only-weak smoothing and may remain close to y_i. In contrast, large values of α improve smoothing since they increase the weight of Φ. To resume, small values of α are better adapted for the detection of outliers while large values of α are better suited for smoothing of outliers. We are hence faced with a compromise between efficiency of detection and quality of smoothing. The next example, as well as the experiments presented below, corroborate this conjecture.

Example 3. Consider the following cost-function:

$$\mathcal{F}(x,y)=\sum_{i=1}^{q}|x_i-y_i|+\alpha\sum_{i=1}^{p-1}(x_i-x_{i+1})^2.$$

Let $\hat{x}$ be a minimizer of $\mathcal{F}(.,y)$ for which $\hat{h}:=\mathcal{H}(\hat{x},y)$ is nonempty. Focus on $i\in\hat{h}^c$. Since $\hat{x}_i\neq y_i$, then

$$0=\frac{\partial\mathcal{F}(\hat{x},y)}{\partial\hat{x}_i}=\operatorname{sign}(\hat{x}_i-y_i)+2\alpha\left((\hat{x}_i-\hat{x}_{i+1})-(\hat{x}_{i-1}-\hat{x}_i)\right),$$

which yields

$$\hat{x}_i=\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{\operatorname{sign}(\hat{x}_i-y_i)}{4\alpha}. \tag{76}$$

Hence, $\hat{x}_i$ takes the form (76) only if we have

$$\text{either}\quad y_i>\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}+\frac{1}{4\alpha}\quad\text{or}\quad y_i<\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{1}{4\alpha}.$$

We remark that (76) does not involve y_i but only the sign of $(\hat{x}_i-y_i)$. Thus, if y_i is an outlier, the value of $\hat{x}_i$ relies only on faithful data entries y_j for $j\in\hat{h}$ by means of $\hat{x}_{i-1}$ and $\hat{x}_{i+1}$. Moreover, the smoothing incurred by $\hat{x}_i$ is stronger for large values of α, since then $\hat{x}_i$ is closer to the mean of $\hat{x}_{i-1}$ and $\hat{x}_{i+1}$. Otherwise, if $i\in\hat{h}$, we have $\delta_1\mathcal{F}(\hat{x},y)(e_i)\geq 0$, which yields

$$\hat{x}_i=y_i\quad\Leftrightarrow\quad\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{1}{4\alpha}\leq y_i\leq\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}+\frac{1}{4\alpha}.$$

This inequality is easier to satisfy if α is small, in which case numerous data samples are fitted exactly, whereas only a few samples are detected as outliers.

Concrete results depend on the shape of ψ, φ, $\{G_i^T\}$, and α. We leave this crucial question for future work. In order to recover and smooth outliers, we take the following cost-function:

$$\mathcal{F}(x,y) = \sum_{i=1}^{q} |x_i - y_i| + \alpha \sum_{i=1}^{p} \sum_{j \in \mathcal{N}(i)} |x_i - x_j|^\nu \quad \text{for} \quad \nu \in (1,2], \tag{77}$$

where for every $i = 1, \ldots, p$ the set $\mathcal{N}(i)$ contains the indexes of all samples j which are neighbors to i. In all the restorations presented below, $\mathcal{N}(i)$ is composed of the eight nearest neighbors. Since the publication of [9], we can expect that $\nu > 1$ but close to 1 allow edges to be better preserved when outliers are smoothed. Based on this, all the experiments with (77) in the following correspond to $\nu = 1.1$.

The minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ for $y \in \mathbb{R}^q$ is calculated by continuation. Using that the Huber function (5),

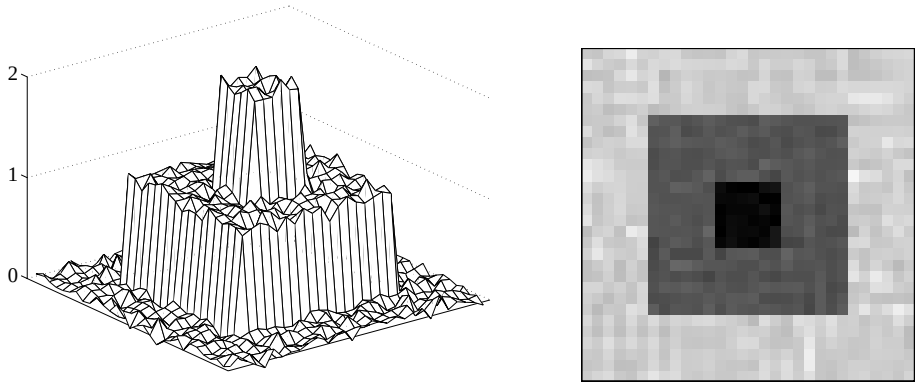
$$\psi_\nu(t) = \begin{cases} t^2 & \text{if} \quad |t| \leq \nu, \\ \nu(\nu + 2|t - \nu|) & \text{if} \quad |t| > \nu, \end{cases} \quad \text{where} \quad \nu > 0,$$

satisfies $\psi_\nu(t) \to |t|$ when $\nu \downarrow 0$, we construct a family of functions $\mathcal{F}_\nu(.,y)$ indexed by $\nu > 0$:

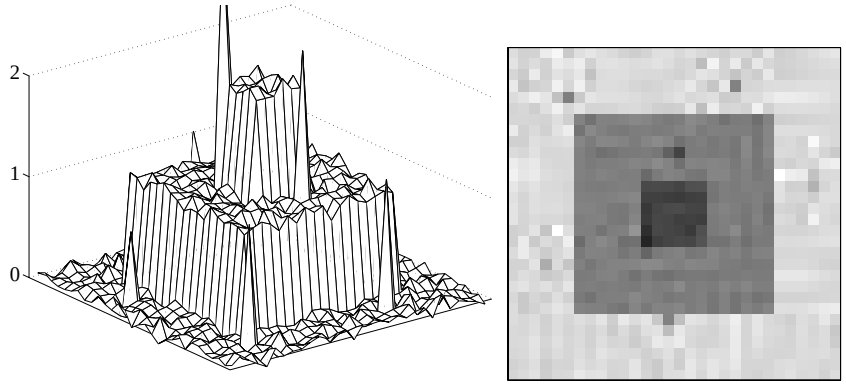
$$\mathcal{F}_\nu(x,y) := \sum_{i=1}^{q} \psi_\nu(a^T x - y_i) + \Phi(x).$$

Being strictly convex and differentiable, every $\mathcal{F}_\nu(.,y)$ has a unique minimizer, denoted by $\hat{x}_\nu$, which is calculated by a gradient descent. Since by construction having $\nu > \nu'$ entails $\mathcal{F}_\nu(x,y) \geq \mathcal{F}_{\nu'}(x,y)$ for all $x \in \mathbb{R}^p$, we see that $\mathcal{F}_\nu(\hat{x}_\nu, y)$ decreases monotonically when ν decreases to 0. It is easy to check that, moreover, as $\nu \downarrow 0$, we have $\mathcal{F}_\nu(\hat{x}_\nu, y) \to \mathcal{F}(\hat{x}, y)$, and hence $\hat{x}_\nu \to \hat{x}$, since every $\mathcal{F}_\nu(.,y)$ has a unique minimizer and the latter is strict. Total-variation methods are similar from a numerical point of view since they involve $\varphi(t) = |t|$. Many authors used smooth approximations [33, 38], e.g., $\varphi_\nu = \sqrt{t^2 + \nu}$. However, approximation using the Huber function has the numerical advantage of involving only quadratic and affine segments. At the same time, the fact that ψ'_ν is discontinuous at $\pm\nu$ is of no practical importance since the chance of obtaining a minimizer $\hat{x}_\nu$ involving a difference whose modulus is exactly ν is null [27].

First experiment. The original image x in Figure 1(a) can be assumed to be a noisy version of an ideal piecewise constant image. Data y in Figure 1(b) are obtained by adding aberrant impulsions to x whose locations are seen in Figure 4, left. Recall that our goal is to detect, and possibly smooth, the outliers in y, while preserving all the remaining entries of y.



(a) Original x.

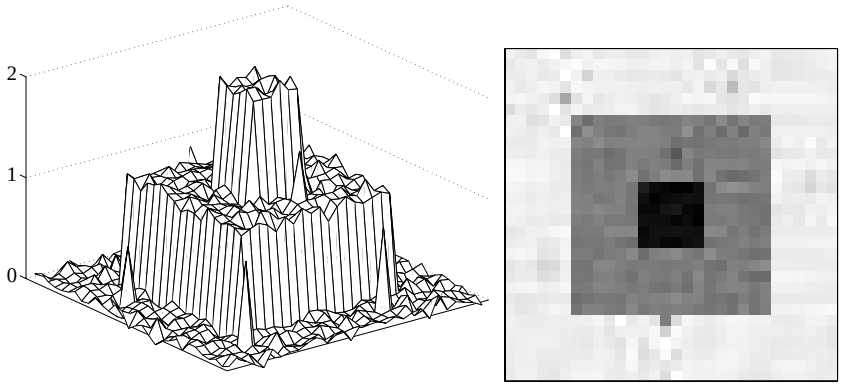


(b) Data $y = x +$ outliers.

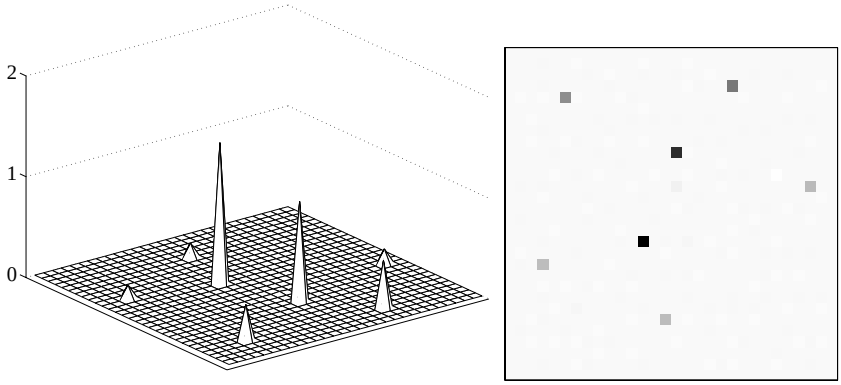
FIG. 1. *Original x and data y degraded by outliers.*

The image in Figure 2(a) is the minimizer $\hat{x}$ of the cost-function $\mathcal{F}(., y)$ proposed in (77), with $\nu = 1.1$ and $\alpha = 0.14$. The outliers are clearly visible although their amplitudes are considerably reduced. The image of the residuals $y - \hat{x}$, shown in Figure 2(b), is null everywhere except at the positions of the outliers in y. Reciprocally, the pixels corresponding to nonzero residuals (i.e., the elements of $\hat{h}^c$) provide a faithful estimate of the locations of the outliers in y, as seen in Figure 4, middle. Next, in Figure 3(a) we show a minimizer $\hat{x}$ of the same $\mathcal{F}(., y)$ obtained for $\alpha = 0.25$. This minimizer does not contain visible outliers and is very close to the original image x. The image of the residuals $y - \hat{x}$ in Figure 3(b) is null only on restricted areas but has a very small magnitude everywhere beyond the positions of the outliers. However, applying the above detection rule now leads to numerous false detections, as seen in Figure 4, right. These experiments confirm our conjecture about the role of α.

The issue of the minimization of a smooth cost-function, namely, $\mathcal{F}$ in (75) with $\psi(t) = \varphi(t) = t^2$ and $\alpha = 0.2$, is shown in Figure 5(a). As expected, edges are blurred, whereas outliers are clearly seen. The residuals in Figure 5(b) are large everywhere, which shows that $\hat{x}$ does not fit any data entry. The minimizer in Figure 6(a) is obtained using nonsmooth regularization, where $\mathcal{F}$ is of the form (75) with $\psi(t) = t^2$, $\varphi(t) = |t|$, and $\alpha = 0.2$. In accordance with our discussion in section 6, $\hat{x}$ exhibits



(a) Restoration $\hat{x}$ for $\alpha = 0.14$.



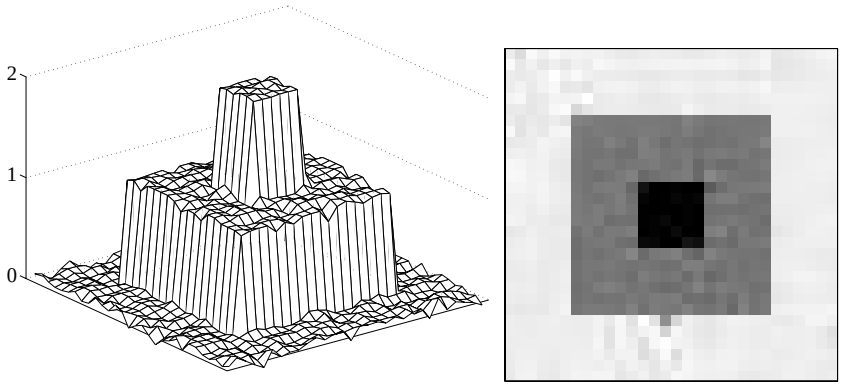
(b) Residuals $y - \hat{x}$.

FIG. 2. *Restoration using the proposed cost-function $\mathcal{F}$ with nonsmooth data-fidelity in* (77) *for $\nu = 1.1$ and $\alpha = 0.14$. The residuals provide a faithful estimator for the locations of outliers.*

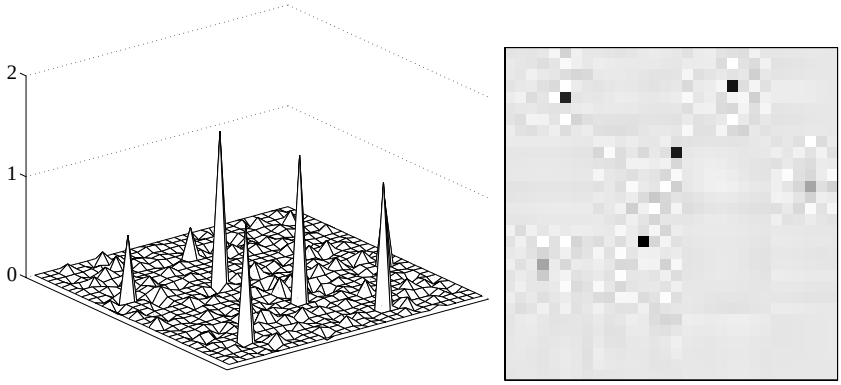
staircasing since it is constant on very large regions.

Second experiment. The original, clean image x is shown in Figure 7(a). The data y, shown in Figure 7(b), are obtained by adding to x 770 impulsions with random locations and random amplitudes in the interval $(0, 1.2)$.

In Figure 8(a) we show a zoom of the histograms of x (up) and of y (down). Figure 8(b) shows the result from applying to y two iterations of median filtering. The obtained image contains only a few outliers with weak amplitude but the entire image is degraded and, in particular, the edges are blurred. The ℓ_1-norm of the error $\|\hat{x} - x\|_1 = \sum_i |\hat{x}_i - x_i|$ is 523. The next two restorations in Figure 9 are obtained by minimizing the cost-function $\mathcal{F}$ with nonsmooth data-fidelity proposed in (77), where $\nu = 1.1$. The minimizer in Figure 9(a) corresponds to $\alpha = 0.2$ and it fits exactly the data everywhere except for several hundred pixels, where it detects outliers. This detection gives rise to 50 erroneous nondetections and to 15 false alarms, the remaining detections being correct. Figure 9(b) is obtained for $\alpha = 0.55$. The minimizer $\hat{x}$ does not contain outliers any longer but it fits exactly only a restricted number of the data entries. Nevertheless, it remains very close to all data entries which are not outliers, since the ℓ_1-norm of the error is 126. This minimizer provides a very clean restoration,



(a) Restoration $\hat{x}$ for $\alpha = 0.25$.



(b) Residuals $y - \hat{x}$.

FIG. 3. *Restoration using the proposed cost-function $\mathcal{F}$ in* (77) *for $\nu = 1.1$ and $\alpha = 0.25$. The outliers are well smoothed in $\hat{x}$, whereas the residuals remain small everywhere beyond the outlier locations.*

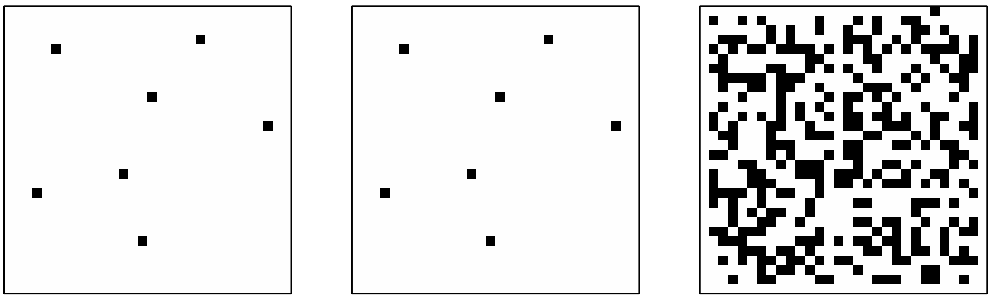
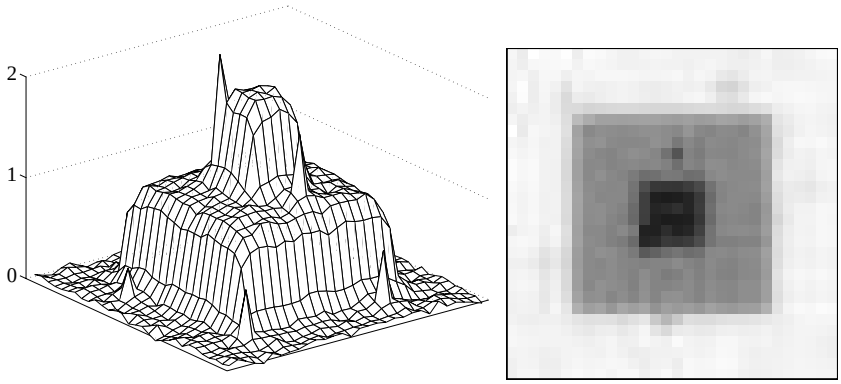
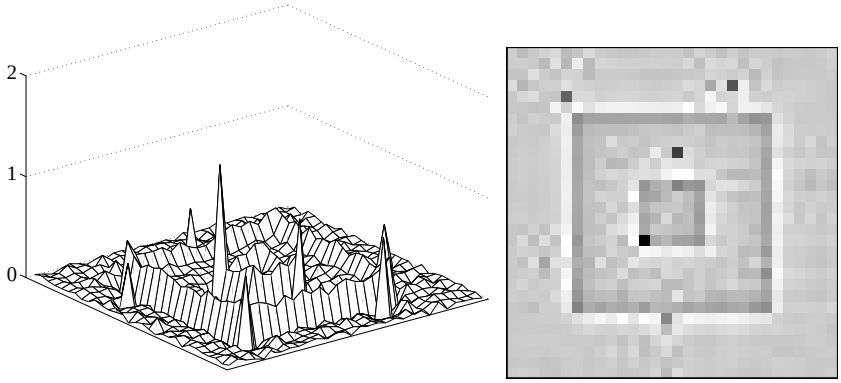


FIG. 4. *Left: The locations of the outliers in y. Middle: The locations of the pixels i of $\hat{x}$ at which $\hat{x}_i \neq y_i$, where $\hat{x}$ is the minimizer obtained for $\alpha = 0.14$ given in Figure* 2. *Right: The same locations for $\hat{x}$ the minimizer relevant to $\alpha = 0.25$, shown in Figure* 3.

where both edges and smoothly varying areas are nicely preserved. The restoration in Figure 10(a) results from a smooth cost-function $\mathcal{F}$, as in (75) with $\psi(t) = \varphi(t) = t^2$



(a) Restoration from $yo\ \hat{x}$ for $\alpha = 0.2$.

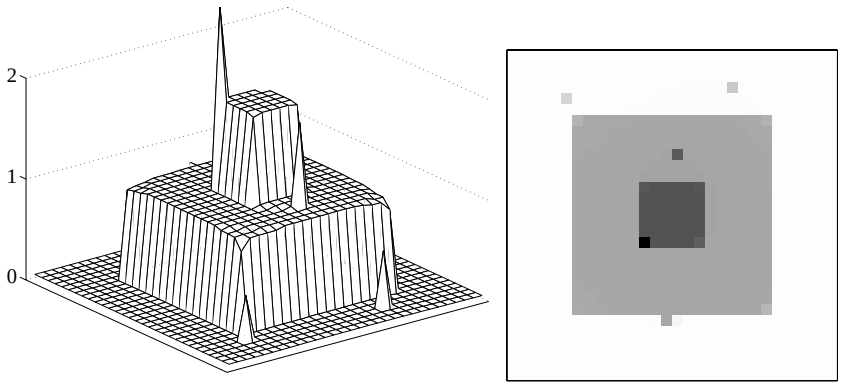


(b) Residuals $y - \hat{x}$.

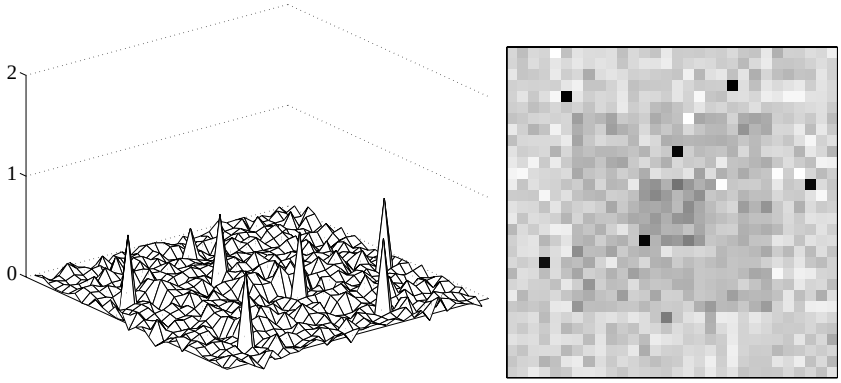
FIG. 5. *Restoration using a smooth cost-function, namely,* $\mathcal{F}$ *in* (75) *with* $\psi(t) = \varphi(t) = t^2$ *and* $\alpha = 0.2$.

and $\alpha = 0.2$. This image fits no data entry while edges are smooth. Figure 10(b) illustrates the staircasing effect induced by nonsmooth regularization. This minimizer corresponds to $\mathcal{F}$, of the form (75) with $\psi(t) = t^2$ and $\varphi(t) = |t|$, for $\alpha = 0.4$ and it still contains several outliers.

8. Conclusion. We showed that taking nonsmooth data-fidelity terms in a regularized cost-function yields minimizers which fit exactly a certain number of the data entries. In contrast, this cannot occur for a smooth cost-function. These are strong properties which can be used in different ways. We proposed a cost-function with a nonsmooth data-fidelity term in order to process outliers. The obtained results advocate the use of nonsmooth data-fidelity terms in image processing.

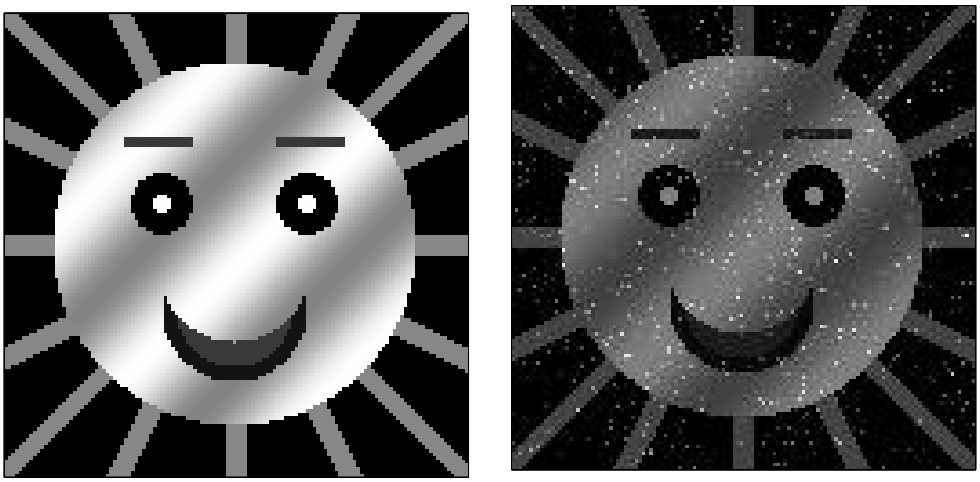


(a) Restoration $\hat{x}$ for $\alpha = 0.2$.



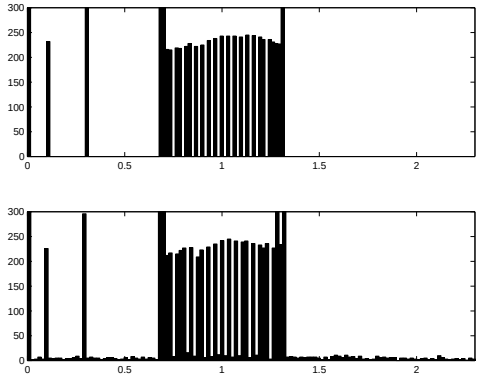
(b) Residuals $y - \hat{x}$.

FIG. 6. *Restoration involving nonsmooth regularization:* $\mathcal{F}$ *is as in* (75) *with* $\psi(t) = t^2$ *and* $\varphi(t) = |t|$ *for* $\alpha = 0.2$. *The minimizer* $\hat{x}$ *is constant over large regions.*

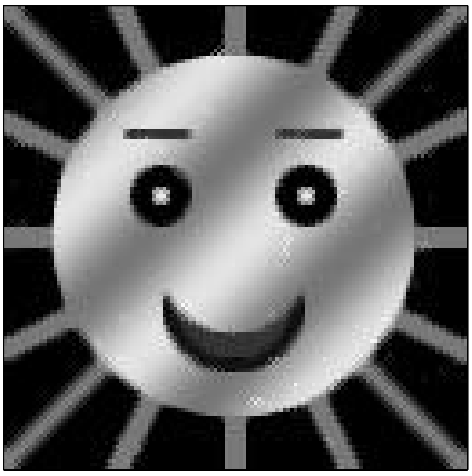


(a) Original image x. (b) Data $y = x + 770$ outliers.

FIG. 7. *Original image x and data y obtained by adding to x 770 outliers with random location and random amplitude.*

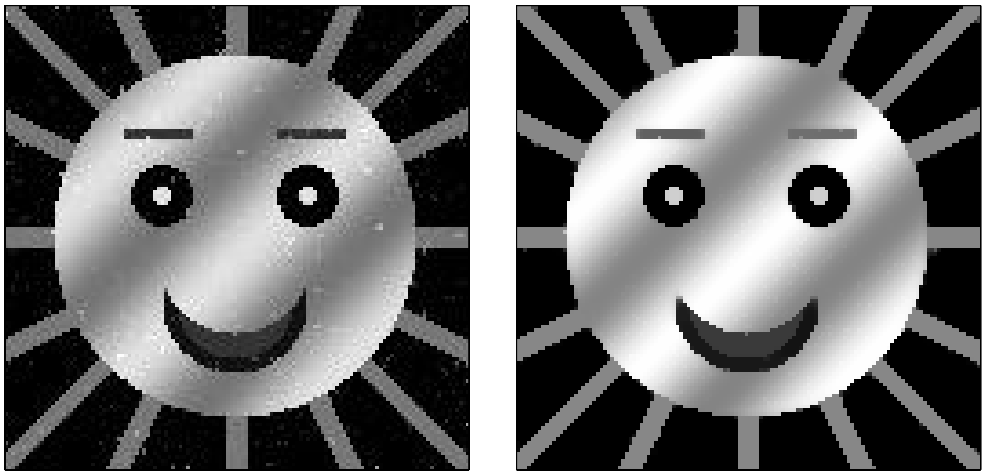


(a) Histograms: x (up), y (down).



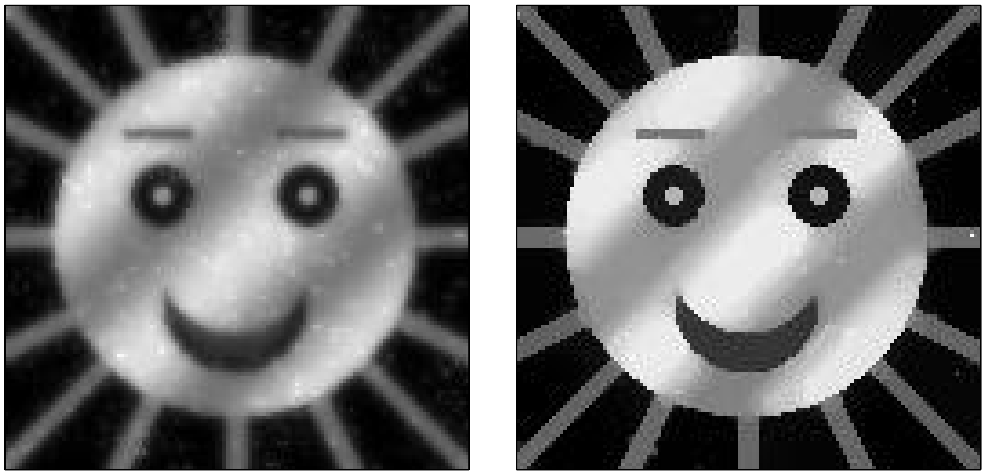
(b) Restoration by median filtering.

FIG. 8. (a) *Zoom of the histograms of the original x (up) and of the data y (down).* (b) *Restoration using two iterations of median filtering.*



(a) Minimizer obtained for $\alpha = 0.2$. (b) Minimizer calculated for $\alpha = 0.55$.

FIG. 9. *Minimizers obtained using the proposed cost-function $\mathcal{F}$ in* (77) *involving a nonsmooth data-fidelity term.* (a) *For $\alpha = 0.2$ there are* 720 *correct and* 65 *erroneous detections of outliers. Outliers are only weakly smoothed.* (b) *For $\alpha = 0.55$, outliers are well smoothed and the error is weak.*



(a) Smooth cost-function. (b) Nonsmooth regularization.

FIG. 10. *Minimizers obtained by minimizing $\mathcal{F}$ of the form* (75). (a) *For $\psi(t) = t^2 = \varphi(t)$ and $\alpha = 0.2$. Outliers are clearly seen, whereas edges are degraded.* (b) *For $\psi(t) = t^2$, $\varphi(t) = |t|$, and $\alpha = 0.4$. Only several outliers remain visible. Staircasing is clearly present.*

REFERENCES

[1] R. Acar and C. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, IEEE Trans. Image Process., 10 (1994), pp. 1217–1229.

[2] S. Alliney, *Digital filters as absolute norm regularizers*, IEEE Trans. Medical Imaging, 12 (1993), pp. 173–181.

[3] S. Alliney, *A property of the minimum vectors of a regularizing functional defined by means of absolute norm*, IEEE Trans. Signal Process., 45 (1997), pp. 913–917.

[4] S. Alliney and S. A. Ruzinsky, *An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation*, IEEE Trans. Signal Process., 42 (1994), pp. 618–627.

[5] A. Avez, *Calcul différentiel*, Masson, Paris, 1991.

[6] J. E. Besag, *On the statistical analysis of dirty pictures (with discussion)*, J. Roy. Statist. Soc. Ser. B, 48 (1986), pp. 259–302.

[7] M. Black and A. Rangarajan, *On the unification of line processes, outlier rejection, and robust statistics with applications to early vision*, Internat. J. Computer Vision, 19 (1996), pp. 57–91.

[8] B. Bloomfield and W. L. Steiger, *Least Absolute Deviations: Theory, Applications and Algorithms*, Birkhäuser, Boston, 1983.

[9] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Trans. Image Process., 2 (1993), pp. 296–310.

[10] C. Bouman and K. Sauer, *A unified approach to statistical tomography using coordinate descent optimization*, IEEE Trans. Image Process., 5 (1996), pp. 480–492.

[11] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[12] T. F. Chan and C. K. Wong, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.

[13] G. Demoment, *Image reconstruction and restoration: Overview of common estimation structure and problems*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2024–2036.

[14] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1199.

[15] D. Donoho, I. Johnstone, J. Hoch, and A. Stern, *Maximum entropy and the nearly black object*, J. Roy. Statist. Soc. Ser. B, 54 (1992), pp. 41–81.

[16] S. Durand and M. Nikolova, *Stability of image restoration by minimizing regularized objective functions*, in Proceedings of the IEEE Int. Conf. on Computer Vision/Workshop on Variational and Level-Set Methods, Vancouver, Canada, 2001, pp. 73–80.

[17] D. Geman, *Random fields and inverse problems in imaging*, in École d'Été de Probabilités de Saint-Flour XVIII 1988, Lecture Notes in Math. 1427, Springer-Verlag, Berlin, 1990, pp. 117–193.

[18] D. Geman and G. Reynolds, *Constrained restoration and recovery of discontinuities*, IEEE Trans. Pattern Anal. Machine Intelligence, 14 (1992), pp. 367–383.

[19] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Trans. Image Process., 4 (1995), pp. 932–946.

[20] S. Geman and D. McClure, *Statistical methods for tomographic image reconstruction*, in Proceedings of the 46th Session of the International Statistical Institute, Vol. 4 (Tokyo, 1987), Bull. Inst. Internat. Statist., 52 (1987), pp. 5–21.

[21] P. J. Green, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Trans. Medical Imaging, 9 (1990), pp. 84–93.

[22] T. Kailath, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, 20 (1974), pp. 146–181.

[23] A. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, NY, 1987.

[24] S. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, New York, 1995.

[25] K. S. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.

[26] M. Nikolova, *Estimées localement fortement homogènes*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 665–670.

[27] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.

[28] M. Nikolova, *Weakly Constrained Minimization. Application to the Estimation of Images and Signals Involving Constant Regions*, Tech. report, CMLA—ENS de Cachan, France, 2001. Available online at http://www.cmla.ens-cachan.fr/Cmla/index.html

[29] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intelligence, 12 (1990), pp. 629–639.
[30] T. T. Pham and R. J. P. De Figueiredo, *Maximum likelihood estimation of a class of non-Gaussian densities with application to l_p deconvolution*, IEEE Trans. Signal Process., 37 (1989), pp. 73–82.
[31] J. R. Rice and J. S. White, *Norms for smoothing and estimation*, SIAM Rev., 6 (1964), pp. 243–256.
[32] R. T. Rockafellar and J. B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[33] L. Rudin, S. Osher, and C. Fatemi, *Nonlinear total variation based noise removal algorithm*, Phys. D, 60 (1992), pp. 259–268.
[34] K. Sauer and C. Bouman, *A local update strategy for iterative reconstruction from projections*, IEEE Trans. Signal Process., 41 (1993), pp. 534–548.
[35] A. Tarantola, *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier Science Publishers, Amsterdam, 1987.
[36] S. Teboul, L. Blanc-Féraud, G. Aubert, and M. Barlaud, *Variational approach for edge-preserving regularization using coupled PDE's*, IEEE Trans. Image Process., 7 (1998), pp. 387–397.
[37] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*, Winston, Washington, DC, 1977.
[38] C. R. Vogel and M. E. Oman, *Fast, robust total variation-based reconstruction of noisy, blurred images*, IEEE Trans. Image Process., 7 (1998), pp. 813–824.
[39] J. Weickert, *Anisotropic Diffusion in Image Processing*, B. G. Teubner, Stuttgart, 1998.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 995–1020

ANALYSIS OF NUMERICAL ERRORS IN LARGE EDDY SIMULATION*

V. JOHN† AND W. J. LAYTON‡

Abstract. We consider the question of "numerical errors" in large eddy simulation. It is often claimed that straightforward discretization and solution using centered methods of models for large eddy motion can simulate the motion of turbulent flows with complexity independent of the Reynolds number and dependence only on the resolution "δ" of the eddies sought. This report considers this question analytically: Is it possible to prove error estimates for discretizations of *actually used* large eddy models whose error constants depend only on δ but not Re? We consider the most common, simplest, and most mathematically tractable model and the most mathematically clear discretization. In two cases, we prove such an error estimate and try to explain why our technique of proof fails in the most general case. Our analysis aims to assume as little time regularity on the true solution as possible.

Key words. large eddy simulation, Navier–Stokes equations, turbulence, finite element methods

AMS subject classifications. 76F65, 65M60

PII. S0036142900375554

1. Introduction.

The laminar or turbulent flow of an incompressible fluid is modeled by solutions (u, p) of the incompressible Navier–Stokes equations:

$$\begin{aligned} u_t + u \cdot \nabla u + \nabla p - Re^{-1}\Delta u &= f && \text{in } \Omega \times (0,T], \\ \nabla \cdot u &= 0 && \text{in } \Omega \times [0,T], \\ u(x,0) &= u_0(x) && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma \times [0,T], \\ \int_\Omega p\, dx &= 0 && \text{in } (0,T]. \end{aligned} \tag{1.1}$$

Here $\Omega \subset \mathbb{R}^d (d = 2, 3)$ is a bounded, simply connected domain with polygonal boundary Γ, $u : \Omega \times [0,T] \to \mathbb{R}^d$ is the fluid velocity, $p : \Omega \times (0,T] \to \mathbb{R}$ is the fluid pressure, $f(x,t)$ is the (known) body force, $u_0(x)$ is the initial flow field, and Re is the Reynolds number. Unfortunately, when Re is large the resulting turbulent flow is typically so complex that so-called direct numerical simulation of (u, p) is not practically feasible.

One conjecture of Leray is that "turbulence" in nature is associated with a breakdown of uniqueness of weak solutions to (1.1). It is known that, for example, weak solutions to (1.1) are unique for $d = 3$ and for very small time intervals, e.g., $0 \le t \le O(Re^{-3})$, and, more importantly, over $O(1)$ time intervals $0 \le t \le T$ if

$$\int_0^T \|\nabla u\|^4_{L^2(\Omega)} dt < \infty.$$

*Received by the editors July 21, 2000; accepted for publication (in revised form) January 30, 2002; published electronically August 8, 2002.

http://www.siam.org/journals/sinum/40-3/37555.html

†Faculty of Mathematics, Otto-von-Guericke-University, Postfach 4120, 39016 Magdeburg, Germany (john@mathematik.uni-magdeburg.de). The research of this author was partially funded by the DAAD (Deutsche Akademische Austauschdienst).

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (wjl@pitt.edu, http://www.math.pitt.edu/∼wjl). The research of this author was partially supported by NSF grants INT-9805563, INT 9814115, and DMS 99-72622.

There are numerous generalizations of this basic result [13, 28]. With this in mind, solutions u to (1.1) with $\|\nabla u\|_{L^2(\Omega)} \in L^4(0,T)$ are frequently described as "laminar." Thus, the L^p-regularity in time which can be reasonably assumed is of critical importance.

There are numerous approaches to the simulation of turbulent flows in practical settings. One of the most promising current approaches is large eddy simulation (LES) in which approximations to local spatial averages of u are calculated. A spatial length scale δ is selected. The large eddies are considered to be those of size greater than or equal to $O(\delta)$ and the small eddies are considered to be those of size less than $O(\delta)$. The large eddies are approximated directly while the effects of the small eddies on the large eddies are modeled. In computational turbulence studies using LES it is often reported that *the resulting computational complexity is independent of the Reynolds number* (but dependent on the resolution sought, δ). There has been little or no analytical support for this observation, however. The goal of this report is to begin numerical analysis in support of this claim.

To be more specific, a smooth, nonnegative function $g(x)$ with $g(0) = 1$ and $\int_{\mathbb{R}^d} g\, dx = 1$ is selected and the mollifer $g_\delta(x)$ is defined in the usual way:

$$g_\delta(x) = \delta^{-d} g(x/\delta).$$

One common example is a Gaussian, $g(x) = (6/\pi)^{d/2} \exp(-6x_j x_j)$, where the summation convention is used. The spatial averaging/filtering operation is now defined by convolution:

$$\overline{u}(x,t) = g_\delta * u(x,t), \quad \overline{p} = g_\delta * p, \quad \overline{f} = g_\delta * f, \text{ etc.}$$

In LES, approximations to $(\overline{u}, \overline{p})$ are sought rather than to (u,p). The usual procedure is to first filter the Navier–Stokes equations:

$$\begin{aligned} \overline{u}_t + \nabla \cdot (\overline{u}\ \overline{u}) + \nabla \overline{p} - Re^{-1} \Delta \overline{u} &= \overline{f} + \nabla \cdot \mathbb{T} && \text{in } \Omega, \\ \nabla \cdot \overline{u} &= 0 && \text{in } \Omega, \end{aligned}$$

where the "Reynolds stress tensor" $\mathbb{T}$ is

$$\mathbb{T} = \mathbb{T}(\overline{u}, u) = \overline{u}\ \overline{u} - \overline{u\ u}.$$

Closure is addressed by a modeling step in which $\mathbb{T}$ is written in terms of $\overline{u}$. The resulting (closed) space filtered Navier–Stokes equations are solved numerically. In this procedure, there are three essential issues:

1. The "modeling error" committed in approximating $\mathbb{T}$.
2. The "numerical error" in solving the resulting system.
3. Correct boundary conditions for the flow averages.

In this report, we study the numerical error analytically. Since there are many models in LES (see, e.g., [25, 14, 23, 9, 2, 31, 35, 34]) and few analytical studies, we take herein the simplest model commonly in use, presented, for example, in Ferziger and Peric [9, section 9.3].

To describe the model, let $\mathbb{D}(u)$ be the deformation tensor associated with the indicated velocity field by

$$\mathbb{D}(u) = \frac{1}{2}(\nabla u + \nabla u^t) = \frac{1}{2}(u_{i,x_j} + u_{j,x_i}).$$

The Reynolds stresses are thought of as a turbulent diffusion process based upon the Boussinesq assumption or eddy viscosity hypothesis that "turbulent fluctuations are dissipative in the mean," [25, 11, 32, 34]. We will accordingly consider a model of the form

$$\nabla \cdot \mathbb{T} \sim \nabla \cdot (\nu_{\text{turb}}(\overline{u}, \delta) \;\; \mathbb{D}(\overline{u})),$$

where $\nu_{\text{turb}} \doteq \nu_{\text{turb}}(\overline{u}, \delta)$ is the so-called turbulent viscosity or eddy viscosity. This turbulent viscosity's determination can be very complex, involving even solutions of accompanying systems of nonlinear partial differential equations. In the simplest case, the turbulent viscosity depends on the mean flow $\overline{u}$ through the magnitude of the deformation of $\overline{u}, \nu_{\text{turb}} = \nu_{\text{turb}}(\mathbb{D}(\overline{u}))$, with a functional dependence. Under the Boussinesq assumption, $\nabla \cdot \mathbb{T}$ should act like a physical viscosity. Following the reasoning of Ladyzhenskaya [29], thermodynamic considerations imply that the Taylor series of $\nu_{\text{turb}}(\mathbb{D})$ should be dominated by odd degree terms. The simplest case is of linear dependence upon $|\mathbb{D}|$:

$$\nu_{\text{turb}} = \nu_{\text{turb}}(|\mathbb{D}(\overline{u})|) = a_0(\delta) + a_1(\delta)|\mathbb{D}(\overline{u})|, \tag{1.2}$$

where $|\mathbb{D}(\overline{u})|$ denotes the Frobenius norm of $\mathbb{D}(\overline{u})$. For specificity and for accord with the most commonly used Smagorinsky [37] model, we take the bulk turbulent viscosity $a_0(\delta) \geq 0$ and $a_1(\delta) = C_s\delta^2$. Other scalings are possible [30] though less tested, as are many other subgridscale models [25, 35]. Here C_s is typically either chosen to be around 0.1 or taken to be a function $C_s = C_s(x, t)$ and extrapolated as in the "dynamic subgridscale model" of Germano et al. [15].

With the model (1.2), the resulting system of equations for the approximations (w, q) to $(\overline{u}, \overline{p})$ is

$$\begin{aligned} w_t + \nabla \cdot (w\ w) + \nabla q - Re^{-1}\Delta w - \nabla \cdot (\nu_{\text{turb}}\mathbb{D}(w)) &= \overline{f} && \text{in } \Omega \times (0, T], \\ \nabla \cdot w &= 0 && \text{in } \Omega \times [0, T], \\ w(x, 0) &= w_0(x) && \text{in } \Omega, \\ \int_\Omega q\ dx &= 0 && \text{in } [0, T]. \end{aligned} \tag{1.3}$$

Boundary conditions must be supplied for the large eddies. It is physically clear that large eddies do not adhere to solid walls. (For example, tornadoes and hurricanes move while touching the earth and lose energy as they move.) Therefore, in [14, 26] (see also [34] for the use of similar boundary conditions in a conventional turbulence model), it was proposed that the large eddies should satisfy a no-penetration condition and a slip with friction condition on $\partial\Omega$:

$$\begin{aligned} w \cdot \hat{n} &= 0 && \text{on } \Gamma, \\ w \cdot \hat{\tau} &= 0 && \text{on } \Gamma_0,\ \text{meas}\ (\Gamma_0) > 0, \\ \beta w \cdot \hat{\tau} + \vec{t} \cdot \hat{\tau} &= 0 && \text{on } \ \Gamma \setminus \Gamma_0, \end{aligned} \tag{1.4}$$

where $\vec{t}$ is the Cauchy stress vector on Γ (for background information, see Serrin [36]), $\beta = \beta(\delta, Re)$ is the friction coefficient (calculated explicitly in [26]), $\hat{n}$ is the outward unit normal, and $\hat{\tau}$ is an orthonormal system of tangent vectors on each face of Γ. The friction coefficient β can be calculated once a specific filter is chosen [26]. It has the property [26] that no slip conditions are recovered as $\delta \to 0$:

$$\beta(Re,\ \delta) \to \infty \quad \text{as } \delta \to 0.$$

A Dirichlet boundary condition $w = w_{\text{inflow}}$ on Γ_0 is appropriate if Γ_0 is an inflow boundary upon which $\overline{u}$ can be calculated by extending the known, inflow velocity field upstream. We take $w_{\text{inflow}} = 0$ for simplicity.

The Cauchy stress vector $\vec{t}$ includes the action of both the viscous stresses and Reynolds stresses and is given by

$$\vec{t}(w) := \hat{n} \cdot [-q\mathbb{I} + 2Re^{-1}\mathbb{D}(w) + a_0(\delta)\mathbb{D}(w) + C_s\delta^2|\mathbb{D}(w)|\mathbb{D}(w)].$$

Standard properties of convolution operators imply that the flow averages $(\overline{u}, \overline{p})$ are $C^\infty(\Omega)$ in space, have bounded kinetic energy

$$\int_\Omega |\overline{u}|^2 dx \le \int_\Omega |u|^2 dx \le C(\Omega, f, u_0),$$

have no solution scales smaller than $O(\delta)$, and converge to u as $\delta \to 0$ [24]. On the other hand, it is *not* obvious, nor has it been proven yet, that solutions (w, q) to the large eddy model approximating $(\overline{u}, \overline{p})$ share *any* of these properties! Nevertheless, the spatial regularity of solutions (w, q) we shall consider to be a *modeling* issue (beyond the scope of this report studying *numerical errors* in LES). The *time regularity* of solutions (w, q) is still an important consideration. For example, we shall show that solutions of this model satisfy

$$\int_0^T \|\nabla w\|_{L^3}^3 dt < \infty$$

uniformly in Re. One goal is thus to assume no greater time regularity than this. The fundamental error analysis of Heywood and Rannacher [22] for the Navier–Stokes equations is based, in part, on a laminar-type assumption $\nabla u \in L^\infty(0, T; L^2(\Omega))$. Weakening this to an assumption of the form $\nabla u \in L^3(0, T; L^3(\Omega))$ (as we seek to do herein) is nontrivial.

2. Preliminaries. This section sets the notation used in the report, describes the function spaces employed, and collects several useful inequalities. The notation used is standard for the most part. The $L^p(\Omega)$ norms, for $p \neq 2$, are explicitly denoted as $\|f\|_{L^p}$. Sobolev spaces $W^{k,p}(\Omega)$ are defined in the usual way [1]. The associated norm is denoted by $\|\cdot\|_{k,p}$. If the domain in question is *not* Ω (e.g., $\Omega \times (0, T)$), then it will be explicitly indicated. If $p = 2$, these norms will be written $\|\cdot\|_k$ for the $W^{k,2}(\Omega)$ norm and $\|\cdot\|_{k,\Gamma}$ for the $W^{k,2}(\Gamma)$ norm and $\|\cdot\|$ and $\|\cdot\|_\Gamma$, respectively, for the $L^2(\Omega)$ and $L^2(\Gamma)$ norms. We suppose the polygonal boundary Γ is composed of faces $\Gamma_0, \Gamma_1, \ldots, \Gamma_J$, where (with some abuse of notation) Γ_0 consists of the face(s) upon which $v = 0$ is strongly imposed.

The spaces associated with the boundary conditions (1.4) are

$$X := \{v : v \in \left(W^{1,3}(\Omega)\right)^d, v = 0 \text{ on } \Gamma_0 \text{ and } v \cdot \hat{n} = 0 \text{ on } \Gamma_j, j = 1, \ldots, J\},$$
$$Q := L_0^2(\Omega) := \left\{q \in L^2(\Omega) : \int_\Omega q \, dx = 0\right\}.$$

The boundary condition in X is defined to hold in the sense of the trace theorem on each Γ_j, and $\hat{n}$ is the outward unit normal to Γ. The $L^2(\Omega)$ and $L^2(\Gamma)$ inner products are denoted by $(\cdot, \cdot)$ and $(\cdot, \cdot)_\Gamma$, respectively.

If $v \in X$, $\mathbb{D}(v)$ denotes the usual deformation tensor, defined in the introduction. The unit vector $\hat{\tau}$ denotes an orthonormal system of tangent vectors on Γ. Whenever

$\hat{\tau}$ occurs, it will be understood that the term is to be summed over the two tangent vectors if $d = 3$; for example,

$$\|v \cdot \hat{\tau}\|^2_{\Gamma_j} \text{ if } d = 3 \text{ means } (\|v \cdot \hat{\tau}_1\|^2_{\Gamma_j} + \|v \cdot \hat{\tau}_2\|^2_{\Gamma_j}).$$

LEMMA 2.1 (inf-sup condition). *Let* $\tilde{X} := \{v : v \in \left(W^{1,2}(\Omega)\right)^d$, $v = 0$ *on* Γ_0 *and* $v \cdot \hat{n} = 0$ *on* $\Gamma_j, j = 1, \ldots, J\}$. *The velocity-pressure spaces* $(\tilde{X}, Q)$ *satisfy the inf-sup condition*

$$\inf_{\lambda \in Q} \sup_{v \in \tilde{X}} \frac{(\lambda, \nabla \cdot v)}{\|\lambda\| \left[\|\mathbb{D}(v)\|^2 + \sum_{j=1}^J \|v \cdot \hat{\tau}\|^2_{1/2,\Gamma_j}\right]^{1/2}} \geq C > 0. \tag{2.1}$$

Proof. Since $\|\nabla v\| \geq \|\mathbb{D}(v)\|$, the trace theorem [20] shows that (2.1) is implied by the usual inf-sup condition

$$\inf_{\lambda \in Q} \sup_{v \in \tilde{X} \cap H^1_0(\Omega)^d} \frac{(\lambda, \nabla \cdot v)}{\|\lambda\| \, \|\nabla v\|} \geq C > 0. \qquad \square$$

Lemma 2.1 implies that the space of weakly divergence free functions V,

$$V := \{v \in \tilde{X} : (\lambda, \nabla \cdot v) = 0 \text{ for all } \lambda \in Q\},$$

is a well defined, nontrivial, closed subspace of $\tilde{X}$.

Remark 2.1. Since Γ is *not* C^1, discontinuities in $\hat{\tau}$ and $\hat{n}_j$ have forced modifications in the norms to piecewise definition. For example, $v \cdot \hat{\tau} \notin H^{1/2}(\Gamma)$ for $v \in H^1(\Omega)$ but $v \cdot \hat{\tau} \in H^{1/2}(\Gamma_j), j = 0, \ldots, J$.

The conforming finite element method for this problem begins by selecting finite element spaces $X^h \subset X$ and $Q^h \subset Q$, where h denotes as usual a representative mesh width for (X^h, Q^h), satisfying the usual approximation theoretic conditions required of finite element spaces. The condition that $X^h \subset X$ imposes the restriction that $v^h \cdot \hat{n}_{|\Gamma_j} = 0$ for all $v^h \in X^h$. For intricate boundaries, this could possibly be onerous so it is interesting to consider imposing $v^h \cdot \hat{n}_{|\Gamma} = 0$ with penalty or Lagrange multiplier methods, following, e.g., the work in [31]. Nevertheless, there is already considerable computational experience with imposing this condition in finite element methods (see, e.g., [19, 8]), so we shall not focus on the interesting detail of the treatment of corners. Without these additional regularizations in the numerical method, it is useful in the analysis to assume that (X^h, Q^h) satisfies the discrete analogue of (2.1),

$$\inf_{\lambda^h \in Q^h} \sup_{v^h \in X^h} \frac{(\lambda^h, \nabla \cdot v^h)}{\|\lambda^h\| \left[\|\mathbb{D}(v^h)\|^2 + \sum_{j=1}^J \|v^h \cdot \hat{\tau}\|^2_{1/2,\Gamma_j}\right]^{1/2}} \geq C > 0, \tag{2.2}$$

where $C > 0$ is independent of h. The next lemma shows, in essence, that if the computational mesh follows the boundary and if the velocity space, restricted to no slip boundary conditions, and the pressure space satisfy the usual inf-sup condition, then (2.2) holds.

LEMMA 2.2 (discrete inf-sup condition). *If* (X^h, Q^h) *satisfies*

$$\inf_{\lambda_h \in Q^h} \sup_{v_h \in X^h \cap H^1_0(\Omega)^d} \frac{(\lambda^h, \nabla \cdot v^h)}{\|\lambda^h\| \, \|\nabla v^h\|} \geq C_1 > 0,$$

then (2.2) *holds.*

Proof. By trace theorem [20] and the Poincaré–Friedrichs inequality, for any $\lambda_h \neq 0$, $v^h (\neq 0) \in X^h$,

$$\frac{(\lambda^h, \nabla \cdot v^h)}{\|\lambda^h\| \left[\|\mathbb{D}(v^h)\|^2 + \sum_{j=1}^J \|v^h \cdot \hat{\tau}\|^2_{1/2,\Gamma_j} \right]^{1/2}} \geq C \frac{(\lambda^h, \nabla \cdot v^h)}{\|\lambda^h\| \, \|v^h\|_1} \geq C \frac{(\lambda^h, \nabla \cdot v^h)}{\|\lambda^h\| \, \|\nabla v^h\|}. \qquad \square$$

Thus, (2.2) will be assumed throughout this report. Under (2.2), the space of discretely divergence free functions

$$V^h := \{v^h \in X^h : (\lambda^h, \nabla \cdot v^h) = 0 \text{ for all } \lambda^h \in Q^h\}$$

is a nontrivial closed subspace of X^h [16, 21].

We shall frequently use Young's inequality in the form

$$ab \leq \frac{\epsilon}{q} a^q + \frac{\epsilon^{-q'/q}}{q'} b^{q'}, \quad 1 < q, q' < \infty, \quad \frac{1}{q} + \frac{1}{q'} = 1.$$

The generalization of Hölder's inequality

$$\int_\Omega |u| \, |v| \, |w| dx \leq \|u\|_{L^p} \|v\|_{L^q} \|w\|_{L^r}, \quad \frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1, \quad 1 \leq p, q, r \leq \infty,$$

is also useful. We shall frequently use the Sobolev embedding theorem, often, but not always, in the form that in three dimensions $W^{1,3}(\Omega) \hookrightarrow L^p(\Omega)$ for $1 \leq p < \infty$.

The nonlinear form in the subgridscale term, for $v, w \in (W^{1,3}(\Omega))^d$

$$(|\mathbb{D}(w)|\mathbb{D}(w), \mathbb{D}(v)),$$

is of p-Laplacian type (with $p = 3$). Thus, it is strongly monotone and locally Lipschitz continuous in the sense made precise in the following well-known lemma; see, e.g., [30, 7].

LEMMA 2.3 (strong monotonicity and local Lipschitz-continuity). *There are constants* $\underline{C}$ *and* $\overline{C}$ *such that for all* $u_1, u_2, v \in (W^{1,3}(\Omega))^d$ *and* $d = 2$ *or* 3, *with* $r = \max\{\|\mathbb{D}(u_1)\|_{L^3}, \|\mathbb{D}(u_2)\|_{L^3}\}$,

$$(|\mathbb{D}(u_1)|\mathbb{D}(u_1) - |\mathbb{D}(u_2)|\mathbb{D}(u_2), \mathbb{D}(u_1 - u_2)) \geq \underline{C} \|\mathbb{D}(u_1 - u_2)\|^3_{L^3},$$
$$(|\mathbb{D}(u_1)|\mathbb{D}(u_1) - |\mathbb{D}(u_2)|\mathbb{D}(u_2), \mathbb{D}(v)) \leq \overline{C} r \|\mathbb{D}(u_1 - u_2)\|_{L^3} \|\mathbb{D}(v)\|_{L^3}.$$

Korn's inequalities relate L^p norms of the deformation tensor $\mathbb{D}(v)$ to those same norms of the gradient for $1 < p < \infty$ (see Galdi, Heywood, and Rannacher [13], Gobert [17, 18], Temam [39], or Fichera [10]) and fail if $p = 1$.

THEOREM 2.4 (Korn's inequalities). *There is a* $C > 0$ *such that for* $1 < p < \infty$

$$\|v\|^p_{W^{1,p}} \leq C(\Omega)[\|v\|^p_{L^p} + \|\mathbb{D}(v)\|^p_{L^p}]$$

for all $v \in (W^{1,p}(\Omega))^d$.

Further, if $\gamma(v)$ *is a seminorm on* $L^p(\Omega)$ *which is a norm on the constants, then*

$$\|\nabla v\|_{L^p} \leq C(\Omega)[\gamma(v) + \|\mathbb{D}(v)\|_{L^p}]$$

holds for $1 < p < \infty$ *and for all* $v \in (W^{1,p}(\Omega))^d$.

As a consequence of Korn's inequality it follows that, taking $\gamma(v) = \|v\|_{L^p(\Gamma_0)}$, if $\text{meas}(\Gamma_0) > 0$, then

$$\|\nabla v\|_{L^p}^p \le \|v\|_{1,p}^p \le C_K \|\mathbb{D}(v)\|_{L^p}^p$$

for all $v \in \{v \in W^{1,p}(\Omega)^d : v_{|\Gamma_0} = 0\}$.

We will often use Poincaré's inequality, which holds since $v \cdot \hat{n} = 0$ on Γ (Galdi [12, p. 56]),

$$\|v\| \le C(\Omega)\|\nabla v\| \text{ for all } v \in X.$$

We shall use the Gagliardo–Nirenberg inequality in $W^{1,p}(\Omega) \cap X$. This inequality [1, 33, 13, 6] states that, provided Γ satisfies a weak regularity condition (holding in particular for polygonal domains) and $\text{meas}(\Gamma_0) > 0$ for all $v \in W^{1,p}(\Omega) \cap X, 1 \le q, s \le \infty$,

$$\|v\|_{L^q} \le C\|\nabla v\|_{L^p}^a \|v\|_{L^s}^{1-a} \text{ for all } v \in \left(W^{1,p}(\Omega)\right)^d \cap X,$$

where, for $\Omega \subset \mathbb{R}^3$ (improvable if $\Omega \subset \mathbb{R}^2$), $p \ge 3$, $q \ge s$, $0 \le a < 1$, and

$$a = \left(\frac{1}{s} - \frac{1}{q}\right)\left(\frac{1}{3} - \frac{1}{p} + \frac{1}{s}\right)^{-1}.$$

In particular, note that taking $q = 6$, $p = 3$, and $s = 2$ gives

$$\|v\|_{L^6(\Omega)} \le C\|\nabla v\|_{L^3(\Omega)}^{2/3} \|v\|^{1/3}. \tag{2.3}$$

The following combination of this and Korn's inequality will be useful in section 4.

LEMMA 2.5. *Let* $\text{meas}(\Gamma_0) > 0$ *and* $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. *Then,*

$$\|v\|_{L^6} \le C\|v\|^{1/3}\|\mathbb{D}(v)\|_{L^3}^{2/3}, \; C = C(\Omega).$$

Proof. This follows immediately from (2.3) and Korn's inequality. ∎

The following dual norms are defined in an equivalent but slightly nonstandard way: for $\frac{1}{q} + \frac{1}{q'} = 1, 1 < q, q' < \infty$,

$$\|f\|_* := \sup_{v \in X} \frac{(f, v)}{\|\mathbb{D}(v)\|},$$

$$\|f\|_{W^{-1,3/2}} := \sup_{v \in X} \frac{(f, v)}{\|\mathbb{D}(v)\|_{L^3}},$$

$$\|f\|_{W^{-1,q'}(\Omega \times (0,t))} := \sup_{v \in L^q(0,T;X)} \frac{\int_0^t (f, v)dt'}{(\int_0^t \|\mathbb{D}(v)\|_{L^q}^q dt')^{1/q}}.$$

Note that $\|\mathbb{D}(\cdot)\|_{L^3}$ defines a norm in X as a consequence of Poincaré's and Korn's inequality.

3. The finite element formulation. This section develops the finite element method for the LES model. The stability of the model is also studied. In particular, we show w and $w^h \in L^\infty(0, T; L^2(\Omega)) \cap L^3(0, T; H^1(\Omega))$ uniformly in Re. Lastly, the error in an equilibrium projection is considered.

The variational formulation is derived in the usual way by multiplication of (1.3) by $(v, q) \in (X, Q)$ and applying the divergence theorem. The boundary integral terms

require careful treatment (following, e.g., [31]) on account of the slip with friction condition on Γ. Let $\alpha \geq 0$ be a constant. The formulation which results is to find $w : [0,T] \to X, q : (0,T] \to Q$ satisfying

$$\begin{aligned}(w_t, v) + \beta(\delta, Re) \sum_{j=1}^{J} (w \cdot \hat{\tau}, v \cdot \hat{\tau})_{\Gamma_j} + \left((2Re^{-1} + a_0(\delta) + C_s \delta^2 |\mathbb{D}(w)|)\mathbb{D}(w), \mathbb{D}(v)\right) \\ + (w \cdot \nabla w, v) - (q, \nabla \cdot v) + \alpha(\nabla \cdot w, \nabla \cdot v) = (\overline{f}, v) \quad \text{for all } v \in X, \\ (\lambda, \nabla \cdot w) = 0 \quad \text{for all } \lambda \in Q,\end{aligned} \tag{3.1}$$

and $w(x,0) = \overline{u}_0(x) \in X$. For compactness, define the nonlinear and trilinear form:

$$\begin{aligned}a(u,w,v) &:= \alpha(\nabla \cdot w, \nabla \cdot v) + \sum_{j=1}^{J} \beta(w \cdot \hat{\tau}, v \cdot \hat{\tau})_{\Gamma_j} \\ &\quad + ((2Re^{-1} + a_0(\delta) + C_s \delta^2 |\mathbb{D}(u)|)\mathbb{D}(w), \mathbb{D}(v)), \\ b(u,w,v) &:= \frac{1}{2}(u \cdot \nabla w, v) - \frac{1}{2}(u \cdot \nabla v, w).\end{aligned}$$

It is a simple index calculation to check that for $v \in X, w \in V$ (since such functions have zero normal components on Γ) $(w \cdot \nabla w, v) = b(w,w,v)$. Thus, the variational formulation can be rewritten as follows: find $(w,q) : [0,T] \to (X,Q)$ satisfying $w(x,0) = \overline{u}_0(x)$ and

$$(w_t, v) + a(w,w,v) + b(w,w,v) + (\lambda, \nabla \cdot w) - (q, \nabla \cdot v) = (\overline{f}, v) \tag{3.2}$$

for all $(v, \lambda) \in (X, Q)$.

Using Lemma 2.3, it is easy to prove that the LES model (1.3), (1.4) satisfies the analogue of Leray's inequality for the Navier–Stokes equations.

LEMMA 3.1 (Leray's inequality for the LES model). *A solution of* (3.2) *satisfies*

$$\begin{aligned}\frac{1}{2}\|w(t)\|^2 + \int_0^t \left[\sum_{j=1}^{J} \beta \|w \cdot \hat{\tau}\|_{\Gamma_j}^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w)\|^2 + \underline{C} C_s \delta^2 \|\mathbb{D}(w)\|_{L^3}^3 \right] dt' \\ \leq \frac{1}{2}\|w(0)\|^2 + \int_0^t (\overline{f}, w) dt'.\end{aligned}$$

Proof. Set $v = w, \lambda = q$ in (3.2) and use Lemma 2.3. ∎

Remark 3.1.

1. Because of the slip with friction boundary conditions (1.4), it is important to choose the formulation of the viscous terms, as in (3.1), (3.2), involving the deformation tensor.

2. Leray's inequality immediately implies stability in various norms (which we will develop) and is the key, first step in proving existence of weak solutions to (1.3), (1.4). This last question is fully investigated (under different boundary conditions) in remarkable papers by Ladyzhenskaya [27], Parés [34], and Du and Gunzburger [7].

LEMMA 3.2. *Let (w,q) be the solution of* (1.3). *Then, there is a constant C independent of Re such that for almost all $t \in (0,T)$ with $0 < T < \infty$*

$$\begin{aligned}\|w_t\|_{W^{-1,3/2}} \leq C \left(\|w\|_{L^3}^2 + \|q\|_{L^{3/2}} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w)\|_{L^{3/2}} \right. \\ \left. + C_s \delta^2 \|\mathbb{D}(w)\|_{L^3}^2 + \|\overline{f}\|_{W^{-1,3/2}} \right),\end{aligned}$$

$$\|w_t\|^{3/2}_{L^{3/2}(0,T;W^{-1,3/2})} \le C\Big(\|w\|^3_{L^3(0,T;L^3)} + \|q\|^{3/2}_{L^{3/2}(0,T;L^{3/2})} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w)\|^{3/2}_{L^{3/2}(0,T;L^{3/2})} + C_s\delta^2\|\mathbb{D}(w)\|^3_{L^3(0,T;L^3)} + \|\overline{f}\|^{3/2}_{L^{3/2}(0,T;W^{-1,3/2})}\Big).$$

Proof. From the momentum equation in (1.3) it follows that for almost all $t \in (0,T)$ (alternately, dividing (3.1) by $\|v\|_{W^{1,3}}$ and taking the supremum over v)

$$\|w_t\|_{W^{-1,3/2}} \le \|\nabla\cdot(ww)\|_{W^{-1,3/2}} + \|\nabla q\|_{W^{-1,3/2}} + C_s\delta^2\|\nabla\cdot(|\mathbb{D}(w)|\mathbb{D}(w))\|_{W^{-1,3/2}} + (2Re^{-1} + a_0(\delta))\|\nabla\cdot\mathbb{D}(w)\|_{W^{-1,3/2}} + \|\overline{f}\|_{W^{-1,3/2}}.$$

The definition of the norm, integration by parts, using $v\cdot n = 0$ on Γ for $v \in X$, Hölder's inequality, and Korn's inequality give, e.g.,

$$\|\nabla q\|_{W^{-1,3/2}} = \sup_{v\in X}\frac{\int_\Omega -q(\nabla\cdot v)dx}{\|\mathbb{D}(v)\|_{L^3}} \le \sup_{v\in X}\frac{\|q\|_{L^{3/2}}\|\nabla\cdot v\|_{L^3}}{\|\mathbb{D}(v)\|_{L^3}} \le C\|q\|_{L^{3/2}}.$$

The other terms are estimated in the same way also using $\|ww\|_{L^{3/2}} = \|w\|^2_{L^3}$ and $\||\mathbb{D}(w)|\mathbb{D}(w)\|_{L^{3/2}} = \|\mathbb{D}(w)\|^2_{L^3}$.

The second inequality follows raising both sides to the power 3/2 and integrating in time. ∎

The continuous-in-time finite element method for (1.3) uses the variational formulation (3.2) as follows. First, velocity-pressure finite element spaces $X^h \subset X \cap (W^{1,3}(\Omega))^d$, $Q^h \subset Q$ satisfying (2.2), and the parameter $\alpha \ge 0$ are selected.

The finite element approximations to (w,q) are maps $(w^h, q^h) : [0,T] \to (X^h, Q^h)$ satisfying

$$(w^h_t, v^h) + a(w^h, w^h, v^h) + b(w^h, w^h, v^h) + (\lambda^h, \nabla\cdot w^h) - (q^h, \nabla\cdot v^h) = (\overline{f}, v^h) \tag{3.3}$$

for all $(v^h, \lambda^h) \in (X^h, Q^h)$ where $w^h(x,0) \in X^h$ is an approximation to $w(x,0) = \overline{u}_0$.

It is straightforward to verify that Leray's inequality holds for w^h as well as w.

LEMMA 3.3 (Leray's inequality for w^h). *For $\alpha \ge 0$, any solution of* (3.3) *satisfies*

$$\frac{1}{2}\|w^h(t)\|^2 + \int_0^t\Big[\beta\sum_{j=1}^J\|w^h\cdot\hat{\tau}\|^2_{\Gamma_j} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w^h)\|^2 + \alpha\|\nabla\cdot w^h\|^2 + \underline{C}\ C_s\delta^2\|\mathbb{D}(w^h)\|^3_{L^3}\Big]dt' \le \frac{1}{2}\|w^h(0)\|^2 + \int_0^t(\overline{f}, w^h)dt'.$$

Using various inequalities in the right-hand side, stability bounds for w^h follow from Lemma 3.3.

PROPOSITION 3.4 (stability of w^h). *The solution w^h of* (3.3) *satisfies*

$$\frac{1}{2}\|w^h(t)\|^2 + \int_0^t\Big[\sum_{j=1}^J\beta\|w^h\cdot\hat{\tau}\|^2_{\Gamma_j} + (Re^{-1} + a_0(\delta))\|\mathbb{D}(w^h)\|^2 + \alpha\|\nabla\cdot w^h\|^2 + \underline{C}C_s\delta^2\|\mathbb{D}(w^h)\|^3_{L^3}\Big]dt' \le \frac{1}{2}\|w^h(0)\|^2 + \frac{Re}{4}\int_0^t\|\overline{f}\|^2_*dt', \tag{3.4}$$

$$\frac{1}{2}\|w^h(t)\|^2 + \int_0^t \left[\sum_{j=1}^{J} \beta \|w^h \cdot \hat{\tau}\|^2_{\Gamma_j} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w^h)\|^2 + \alpha \|\nabla \cdot w^h\|^2 \right.$$
$$\left. + \frac{2}{3}\underline{C}C_s\delta^2 \|\mathbb{D}(w^h)\|^3_{L^3} \right] dt' \leq \frac{1}{2}\|w^h(0)\|^2$$
$$+ \frac{2}{3}(\underline{C}C_s)^{-1/2}\delta^{-1}\|\overline{f}\|^{3/2}_{W^{-1,3/2}(\Omega\times(0,t))}, \tag{3.5}$$

$$\|w^h(t)\|^2 + 2\int_0^t e^{t-t'} \left[\sum_{j=1}^{J} \beta \|w^h \cdot \hat{\tau}\|^2_{\Gamma_j} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w^h)\|^2 + \alpha \|\nabla \cdot w^h\|^2 \right.$$
$$\left. + \underline{C}C_s\delta^2 \|\mathbb{D}(w^h)\|^3_{L^3} \right] dt' \leq e^t\|w^h(0)\|^2 + \int_0^t e^{t-t'}\|\overline{f}\|^2 dt'. \tag{3.6}$$

Proof. Inequality (3.4) follows by applying Young's inequality to Lemma 3.3. The bound (3.5) follows from the definition of the dual norm and $ab \leq \frac{\epsilon}{3}a^3 + \frac{2}{3}\epsilon^{-1/2}b^{3/2}$ applied in the same manner.

For (3.6), set $v^h = w^h$ and $\lambda^h = q^h$ in (3.3), use Lemma 2.3, and apply Young's inequality on the right-hand side. This gives

$$\frac{d}{dt}\|w^h\|^2 - \|w^h\|^2 + 2\left[\sum_{j=1}^{J} \beta \|w^h \cdot \hat{\tau}\|^2_{\Gamma_j} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w^h)\|^2 + \alpha \|\nabla \cdot w^h\|^2 \right.$$
$$\left. + \underline{C}C_s\delta^2 \|\mathbb{D}(w^h)\|^3_{L^3} \right] \leq \|\overline{f}\|^2.$$

Inequality (3.6) now follows by using an integrating factor. ∎

In the analysis of the error in the approximation of the time dependent problem, it is useful to have a clear description of the error in the Stokes projection under slip with friction boundary conditions [31]. It is also necessary that any dependence on Re, δ, and β be made explicit.

Under the discrete inf-sup condition, the Stokes projection $\Pi : (X, Q) \to (X^h, Q^h)$ is defined as follows. Let $\Pi(w, q) = (\tilde{w}, \tilde{q})$, where $(\tilde{w}, \tilde{q})$ satisfies

$$\alpha(\nabla \cdot (w - \tilde{w}), \nabla \cdot v^h) + (2Re^{-1} + a_0(\delta))(\mathbb{D}(w - \tilde{w}), \mathbb{D}(v^h))$$
$$+ \sum_{j=1}^{J} \beta((w - \tilde{w}) \cdot \hat{\tau}, v^h \cdot \hat{\tau})_{\Gamma_j} - (q - \tilde{q}, \nabla \cdot v^h) = 0 \quad \text{for all } v^h \in X^h,$$
$$(\nabla \cdot (w - \tilde{w}), \lambda^h) = 0 \quad \text{for all } \lambda^h \in Q^h.$$

This is equivalent to the following formulation provided $w \in V$ and $v^h \in V^h$. Given (w, q), find $\tilde{w} \in V^h$ satisfying

$$\alpha(\nabla \cdot (w - \tilde{w}), \nabla \cdot v^h) + (2Re^{-1} + a_0(\delta))(\mathbb{D}(w - \tilde{w}), \mathbb{D}(v^h))$$
$$+ \sum_{j=1}^{J} \beta((w - \tilde{w}) \cdot \hat{\tau}, v^h \cdot \hat{\tau})_{\Gamma_j} - (q - \lambda^h, \nabla \cdot v^h) = 0$$

for all $v^h \in V^h$ and $\lambda^h \in Q^h$. Under the discrete inf-sup condition, it is well known that $(\tilde{w}, \tilde{q})$ is a quasi-optimal approximation of (w, q). The dependence of the stability

and error constants upon Re and $\beta = \beta(Re, \delta)$ is important to the error analysis. That dependence is described in the next lemma and proposition.

LEMMA 3.5 (stability of the projection $\tilde{w}$). *Let $w \in V$ be given. Then if $\alpha > 0$, $\tilde{w}$ satisfies*

$$\alpha\|\nabla\cdot\tilde{w}\|^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(\tilde{w})\|^2 + \sum_{j=1}^{J}\beta\|\tilde{w}\cdot\hat{\tau}\|_{\Gamma_j}^2$$
$$\leq \alpha^{-1}\|q\|^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w)\|^2 + \sum_{j=1}^{J}\beta\|w\cdot\hat{\tau}\|_{\Gamma_j}^2.$$

If $\alpha = 0$, then

$$\frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\tilde{w})\|^2 + \sum_{j=1}^{J}\beta\|\tilde{w}\cdot\hat{\tau}\|_{\Gamma_j}^2$$
$$\leq 2(2Re^{-1} + a_0(\delta))^{-1}\|q\|^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w)\|^2 + \sum_{j=1}^{J}\beta\|w\cdot\hat{\tau}\|_{\Gamma_j}^2.$$

Proof. Set $v^h = \tilde{w} \in V^h$ in the second formulation of the Stokes projection. This immediately gives

$$\alpha\|\nabla\cdot\tilde{w}\|^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(\tilde{w})\|^2 + \sum_{j=1}^{J}\beta\|\tilde{w}\cdot\hat{\tau}\|_{\Gamma_j}^2$$
$$= (2Re^{-1} + a_0(\delta))(\mathbb{D}(w), \mathbb{D}(\tilde{w})) + \sum_{j=1}^{J}\beta(w\cdot\hat{\tau}, \tilde{w}\cdot\hat{\tau})_{\Gamma_j} + (q - \lambda^h, \nabla\cdot\tilde{w})$$
$$\leq \frac{1}{2}(2Re^{-1} + a_0(\delta))[\|\mathbb{D}(w)\|^2 + \|\mathbb{D}(\tilde{w})\|^2] + \sum_{j=1}^{J}\frac{\beta}{2}[\|w\cdot\hat{\tau}\|_{\Gamma_j}^2 + \|\tilde{w}\cdot\hat{\tau}\|_{\Gamma_j}^2]$$
$$+ \frac{\alpha}{2}\|\nabla\cdot\tilde{w}\|^2 + \frac{1}{2\alpha}\|q\|^2,$$

from which the first result follows. If $\alpha = 0$, the term $(q, \nabla\cdot\tilde{w})$ is bounded by noting that $\nabla\cdot\tilde{w} = \text{trace}\,(\mathbb{D}(\tilde{w}))$ so that

$$(q, \nabla\cdot\tilde{w}) \leq \|q\|\,\|\mathbb{D}(\tilde{w})\| \leq \frac{1}{4}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\tilde{w})\|^2 + (2Re^{-1} + a_0(\delta))^{-1}\|q\|^2. \qquad \square$$

PROPOSITION 3.6. *Suppose the discrete inf-sup condition (2.2) holds. Then, $(\tilde{w}, \tilde{q})$ exists uniquely in (X^h, Q^h) and satisfies*

$$\alpha\|\nabla\cdot(w - \tilde{w})\|^2 + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w - \tilde{w})\|^2 + \sum_{j=1}^{J}\beta\|(w - \tilde{w})\cdot\hat{\tau}\|_{\Gamma_j}^2$$
$$\leq C \inf_{v^h\in X^h, \lambda^h\in Q^h}\Bigg\{(2Re^{-1} + a_0(\delta))\|\mathbb{D}(w - v^h)\|^2$$
$$+ \sum_{j=1}^{J}\beta\|(w - v^h)\cdot\hat{\tau}\|_{\Gamma_j}^2 + \min\{\alpha^{-1}, (2Re^{-1} + a_0(\delta))^{-1}\}\|q - \lambda^h\|^2\Bigg\}.$$

Proof. The proof follows standard arguments, carefully tracking the dependence of the constants upon Re and β. ∎

Note that the use of least squares penalization of incompressibility allows an error estimate for the Stokes projection whose constants are essentially independent of the Reynolds number in a suitably weighted norm.

4. The convergence theorem. Let us first note that for standard piecewise polynomial finite element spaces it is known that the L^2-projection of a function in $L^p, p \geq 2$, is in L^p itself and the L^2-projection operator is stable in $L^p, p \geq 2$ [5].

Let $e = w - w^h$ and let $\tilde{w}$ denote a stable approximation of w in $V^h \cap (W^{1,3}(\Omega))^d$, for example, the L^2-projection under the conditions of [5]. This stability in $W^{1,p}$ follows for many piecewise polynomial finite element spaces using [5].

The error is decomposed as $e = (w - \tilde{w}) - (w^h - \tilde{w}) = \eta - \phi^h$, where $\eta = w - \tilde{w}$ and $\phi^h = w^h - \tilde{w} \in V^h$. An error equation is obtained by subtracting (3.2) from (3.3) and using the fact that $w \in V$. This gives, for any $v^h \in V^h \cap (W^{1,3}(\Omega))^d$ and $\lambda^h \in Q^h$,

$$\begin{aligned}(e_t, v^h) + a(w, w, v^h) - a(w^h, w^h, v^h) \\ + b(w, w, v^h) - b(w^h, w^h, v^h) - (q - \lambda^h, \nabla \cdot v^h) = 0.\end{aligned} \tag{4.1}$$

This is rewritten, adding and subtracting terms and setting $v^h = \phi^h$, as follows:

$$\begin{aligned}(\phi_t^h, \phi^h) + a(w^h, w^h, \phi^h) - a(\tilde{w}, \tilde{w}, \phi^h) = (\eta_t, \phi^h) \\ + a(w, w, \phi^h) - a(\tilde{w}, \tilde{w}, \phi^h) + b(w, w, \phi^h) - b(w^h, w^h, \phi^h) - (q - \lambda^h, \nabla \cdot \phi^h).\end{aligned} \tag{4.2}$$

The monotonicity lemma (Lemma 2.3) implies that

$$\begin{aligned}&a(w^h, w^h, \phi^h) - a(\tilde{w}, \tilde{w}, \phi^h) \\ &\geq \underline{C} C_s \delta^2 \|\mathbb{D}(\phi^h)\|_{L^3}^3 + \alpha \|\nabla \cdot \phi^h\|^2 + (2Re^{-1} + a_0(\delta)) \|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J} \beta \|\phi^h \cdot \hat{\tau}\|_{\Gamma_j}^2,\end{aligned}$$

and with $r := \max\{\|\mathbb{D}(w)\|_{L^3}, \|\mathbb{D}(\tilde{w})\|_{L^3}\}$

$$\begin{aligned}&a(w, w, \phi^h) - a(\tilde{w}, \tilde{w}, \phi^h) \\ &\leq (2Re^{-1} + a_0(\delta)) \|\mathbb{D}(\phi^h)\| \, \|\mathbb{D}(\eta)\| + \sum_{j=1}^{J} \beta \|\phi^h \cdot \hat{\tau}\|_{\Gamma_j} \, \|\eta \cdot \hat{\tau}\|_{\Gamma_j} \\ &\quad + C_s \overline{C} \delta^2 r \|\mathbb{D}(\eta)\|_{L^3} \|\mathbb{D}(\phi^h)\|_{L^3} + \alpha \|\nabla \cdot \eta\| \, \|\nabla \cdot \phi^h\|.\end{aligned}$$

Remark 4.1. If $\tilde{w}$ is taken to be the Stokes projection of (w, q) into V^h, then, e.g., the term "$Re^{-1} \|\mathbb{D}(\phi^h)\| \, \|\mathbb{D}(\eta)\|$" on this last right-hand side does not occur.

Inserting these two bounds in (4.2) and using the Cauchy–Schwarz and Young's inequalities gives

$$\begin{aligned}&\frac{1}{2} \frac{d}{dt} \|\phi^h\|^2 + \underline{C} C_s \delta^2 \|\mathbb{D}(\phi^h)\|_{L^3}^3 + \alpha \|\nabla \cdot \phi^h\|^2 \\ &\quad + (2Re^{-1} + a_0(\delta)) \|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J} \beta \|\phi^h \cdot \hat{\tau}\|_{\Gamma_j}^2 \\ &\leq |b(w, w, \phi^h) - b(w^h, w^h, \phi^h)| + \frac{\epsilon_1}{3} \delta^2 \|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{2}{3} \epsilon_1^{-1/2} \delta^{-1} \|\eta_t\|_{W^{-1,3/2}}^{3/2}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}(2Re^{-1}+a_0(\delta))\|\mathbb{D}(\phi^h)\|^2\frac{1}{2}(2Re^{-1}+a_0(\delta))\|\mathbb{D}(\eta)\|^2\\
&+\sum_{j=1}^{J}\left(\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|^2_{\Gamma_j}+\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|^2_{\Gamma_j}\right)+\frac{\epsilon_1}{3}\delta^2\|\mathbb{D}(\phi^h)\|^3_{L^3}\\
&+\frac{2}{3}\epsilon_1^{-1/2}\delta^2\overline{C}^{3/2}C_s^{3/2}r^{3/2}\|\mathbb{D}(\eta)\|^{3/2}_{L^3}+\frac{\alpha}{2}\|\nabla\cdot\phi^h\|^2+\frac{1}{\alpha}\|q-\lambda^h\|^2+\alpha\|\nabla\cdot\eta\|^2.
\end{aligned}
$$

Picking $\epsilon_1=\underline{C}C_s$ and collecting terms gives

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2+\frac{1}{3}\underline{C}C_s\delta^2\|\mathbb{D}(\phi^h)\|^3_{L^3}+\frac{\alpha}{2}\|\nabla\cdot\phi^h\|^2\\
&\quad+\frac{1}{2}(2Re^{-1}+a_0(\delta))\|\mathbb{D}(\phi^h)\|^2+\sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|^2_{\Gamma_j}\\
&\le|b(w,w,\phi^h)-b(w^h,w^h,\phi^h)|+\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|^{3/2}_{W^{-1,3/2}}\\
&\quad+\frac{1}{2}(2Re^{-1}+a_0(\delta))\|\mathbb{D}(\eta)\|^2+\sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|^2_{\Gamma_j}\\
&\quad+\frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|^{3/2}_{L^3}+\alpha^{-1}\|q-\lambda^h\|^2+\alpha\|\nabla\cdot\eta\|^2.
\end{aligned}
\tag{4.3}
$$

This is the basic differential inequality for the error. Three cases will be considered, revolving around the treatment of the first term on the right-hand side of (4.3).

Remark 4.2. If $\alpha=0$, an estimate which is uniform in Re can still be obtained by using Korn's inequality and Young's inequality on the term $(q-\lambda^h,\nabla\cdot\phi^h)$ as follows:

$$
\begin{aligned}
(q-\lambda^h,\nabla\cdot\phi^h)&\le\|q-\lambda^h\|_{L^{3/2}}\|\nabla\cdot\phi\|_{L^3}\le C\|\mathbb{D}(\phi^h)\|_{L^3}\|q-\lambda^h\|_{L^{3/2}}\\
&\le\frac{1}{3}\underline{C}C_s\delta^2\|\mathbb{D}(\phi^h)\|^3_{L^3}+C\delta^{-1}\|q-\lambda^h\|^{3/2}_{L^{3/2}}.
\end{aligned}
$$

However, an estimate of the nonlinear convective term which is uniform in Re fails in the case $\alpha=0$; see Remark 4.7.

Consider the convection terms

$$
b(w,w,\phi^h)-b(w^h,w^h,\phi^h)=b(w,\eta-\phi^h,\phi^h)+b(\eta-\phi^h,w^h,\phi^h).
\tag{4.4}
$$

The terms containing η shall be bounded first. Consider $b(w,\eta,\phi^h)$ and $b(\eta,w^h,\phi^h)$. Using the inequalities in section 2 appropriately gives

$$
\begin{aligned}
|b(w,\eta,\phi^h)|&=\left|\frac{1}{2}[(w\cdot\nabla\eta,\phi^h)-(w\cdot\nabla\phi^h,\eta)]\right|\\
&\le\frac{1}{2}\left[\|\phi^h\|\,\|\nabla\eta\|_{L^{s'}}\|w\|_{L^s}+\|\nabla\phi^h\|_{L^3}\|w\|_{L^q}\|\eta\|_{L^{q'}}\right],
\end{aligned}
$$

where $\frac{1}{2}+\frac{1}{s'}+\frac{1}{s}=1$ and $\frac{1}{3}+\frac{1}{q}+\frac{1}{q'}=1$. Picking $s'=3$, $s=6$, $q=2$, and $q'=6$ gives

$$
\begin{aligned}
|b(w,\eta,\phi^h)|&\le\frac{1}{2}\left[\|\phi^h\|\,\|\nabla\eta\|_{L^3}\|w\|_{L^6}+\|\nabla\phi^h\|_{L^3}\|w\|\,\|\eta\|_{L^6}\right]\\
&\le\frac{1}{4}\|\phi^h\|^2\|w\|^2_{L^6}+\frac{1}{4}\|\nabla\eta\|^2_{L^3}+\frac{\epsilon_3}{6}\|\mathbb{D}(\phi^h)\|^3_{L^3}+\frac{C}{3}\epsilon_3^{-1/2}\|w\|^{3/2}\|\eta\|^{3/2}_{L^6}.
\end{aligned}
\tag{4.5}
$$

The term $b(\eta, w^h, \phi^h)$ is similarly bounded as follows:

$$
\begin{aligned}
|b(\eta, w^h, \phi^h)| &= \left| \frac{1}{2}[(\eta \cdot \nabla w^h, \phi^h) - (\eta \cdot \nabla \phi^h, w^h)] \right| \\
&\leq \frac{1}{2} \|\nabla w^h\|_{L^3} \|\eta\|_{L^6} \|\phi^h\| + \frac{1}{2} \|\nabla \phi^h\|_{L^3} \|\eta\|_{L^6} \|w^h\| \\
&\leq \frac{1}{4} \|\nabla w^h\|_{L^3}^2 \|\phi^h\|^2 + \frac{1}{4} \|\eta\|_{L^6}^2 + \frac{\epsilon_3}{6} \|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{C}{3} \epsilon_3^{-1/2} \|\eta\|_{L^6}^{3/2} \|w^h\|^{3/2}. \qquad (4.6)
\end{aligned}
$$

Korn's inequality and the stability bounds (3.5) and (3.6) immediately imply that $\mathbb{D}(w^h) \in L^3(0,T;L^3)$ uniformly in Re so that $\|\nabla w^h\|_{L^3}^2 \in L^1(0,T)$, uniformly in Re. The Sobolev imbedding theorem and Korn's inequality also imply $\|w\|_{L^6}^2 \in L^1(0,T)$ uniformly in Re. Thus, these bounds suffice for a later application of Gronwall's inequality.

The first term containing only $\phi^h, b(w, \phi^h, \phi^h)$, is zero due to skew symmetry. Thus, there only remains the term $b(\phi^h, w^h, \phi^h)$. Estimating the term $b(\phi^h, w^h, \phi^h)$ is the essential, core difficulty in obtaining an error bound which is uniform in Re. There are only a few natural ways to bound this using Hölder's inequality and the Sobolev embedding theorem. There are two cases in which the analysis is successful:

(i) $a_0(\delta) \neq 0$ and $\nabla w \in L^3(0,T;L^3(\Omega))$,

(ii) $a_0(\delta) = 0$ and ∇w very regular, $\nabla w \in L^2(0,T;L^\infty(\Omega))$.

There is one important case in which the analysis fails:

(iii) $a_0(\delta) = 0$ and $\nabla w \in L^3(0,T;L^3(\Omega))$.

To highlight subsequent analysis and, hopefully, spur further study, we shall first present the case (iii) and explain the failure of the analysis.

Remark 4.3. On the condition $a_0(\delta) > 0$ in part (i), if $a_1(\delta) > 0$ and $a_0(\delta) \geq 0$, then it is known that the difference between two weak solutions of (1.3) can be bounded (nonuniformly in Re) by the change in the problem data [7, 29, 34]. These bounds imply uniqueness over $O(1)$ time intervals. On the other hand, if $a_1(\delta) \equiv 0$ and $a_0(\delta) > 0$, weak solutions are then only known to be unique over very small time intervals $0 \leq t \leq T^*(\delta)$, where (loosely speaking) $T^*(\delta) \sim (a_0(\delta) + Re^{-1})^3$.

4.1. The case $\nabla w \in L^3(0,T;L^3(\Omega))$ and $a_0(\delta) = 0$. If we assume only that $\nabla w \in L^3(0,T;L^3(\Omega))$, there is no need to add and subtract terms since a priori bounds on $\|\nabla w^h\|_{L^3(0,T;L^3)}$ have been proven which are uniform in Re. Thus, we can use Hölder's inequality to write

$$
\begin{aligned}
|b(\phi^h, w^h, \phi^h)| &= \left| \frac{1}{2}(\phi^h \cdot \nabla w^h, \phi^h) - \frac{1}{2}(\phi^h \cdot \nabla \phi^h, w^h) \right| \\
&\leq \frac{1}{2} \|\nabla w^h\|_{L^3} \|\phi^h\|_{L^s} \|\phi^h\|_{L^{s'}} + \frac{1}{2} \|\nabla \phi^h\|_{L^3} \|\phi^h\| \, \|w^h\|_{L^6},
\end{aligned}
$$

where $\frac{1}{3} + \frac{1}{s'} + \frac{1}{s} = 1$ and $1 \leq s', s \leq \infty$. Thus, picking $s' = 2$, $s = 6$, using the embedding $W^{1,3}(\Omega) \to L^6(\Omega)$ and Poincaré's inequality gives

$$
\begin{aligned}
|b(\phi^h, w^h, \phi^h)| &\leq \frac{C(\Omega)}{2} \|\nabla w^h\|_{L^3} \|\phi^h\| \, \|\phi^h\|_{1,3} + \frac{C(\Omega)}{2} \|\nabla \phi^h\|_{L^3} \|\phi^h\| \, \|w^h\|_{1,3} \\
&\leq \frac{\epsilon}{6} \|\phi^h\|_{1,3}^3 + C\epsilon^{-1/2} \|\nabla w^h\|_{L^3}^{3/2} \|\phi^h\|^{3/2} + \frac{\epsilon}{6} \|\mathbb{D}(\phi^h)\|_{L^3}^3. \qquad (4.7)
\end{aligned}
$$

Remark 4.4. Using Lemma 2.5 instead of the embedding of $W^{1,3} \to L^6$ changes the critical exponent on $\|\phi^h\|$ "3/2" to 12/7 in the first term of (4.7) but not the final conclusion.

Combining (4.5), (4.6), (4.7) with $\epsilon_3 = \epsilon$ gives an initial bound on the convection term's difference:

$$
\begin{aligned}
&|b(w,w,\phi^h) - b(w^h,w^h,\phi^h)| \\
&\le \left[\frac{1}{4}\|\nabla\eta\|_{L^3}^2 + \frac{1}{4}\|\eta\|_{L^6}^2 + C\epsilon^{-1/2}\left(\|w\|^{3/2} + \|w^h\|^{3/2}\right)\|\eta\|_{L^6}^{3/2}\right] \\
&\quad + \frac{2\epsilon}{3}\|\nabla\phi^h\|_{L^3}^3 + C\epsilon^{-1/2}\|\nabla w^h\|_{L^3}^{3/2}\|\phi^h\|^{3/2} + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2\right]\|\phi^h\|^2. \qquad (4.8)
\end{aligned}
$$

Inserting (4.8) into (4.3), applying Korn's inequality, and collecting terms gives

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2 + \left(\frac{1}{3}\underline{C}C_s\delta^2 - \frac{2\epsilon}{3}\right)\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{\alpha}{2}\|\nabla\cdot\phi^h\|^2 \\
&\quad + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|_{\Gamma_j}^2 \\
&\le \left[\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|_{W^{-1,3/2}}^{3/2} + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\eta)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|_{\Gamma_j}^2\right. \\
&\quad + \frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|_{L^3}^{3/2} + \alpha^{-1}\|q-\lambda^h\|^2 + \alpha\|\nabla\cdot\eta\|^2 + \frac{1}{4}\|\nabla\eta\|_{L^3}^2 \\
&\quad \left. + C\epsilon^{-1/2}\|w\|^{3/2}\|\eta\|_{L^6}^{3/2} + \frac{1}{4}\|\eta\|_{L^6}^2 + C\epsilon^{-1/2}\|\eta\|_{L^6}^{3/2}\|w^h\|^{3/2}\right] \\
&\quad + C\epsilon^{-1/2}\|\nabla w^h\|_{L^3}^{3/2}\|\phi^h\|^{3/2} + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2\right]\|\phi^h\|^2.
\end{aligned}
$$

Thus, pick ϵ such that

$$\frac{2\epsilon}{3} = \frac{1}{6}\underline{C}C_s\delta^2,$$

i.e., $\epsilon = O(\delta^2)$. This gives

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2 + \frac{1}{6}\underline{C}C_s\delta^2\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{\alpha}{2}\|\nabla\cdot\phi^h\|^2 + Re^{-1}\|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|_{\Gamma_j}^2 \\
&\le \left[\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|_{W^{-1,3/2}}^{3/2} + Re^{-1}\|\mathbb{D}(\eta)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|_{\Gamma_j}^2\right. \\
&\quad + \frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|_{L^3}^{3/2} + \alpha^{-1}\|q-\lambda^h\|^2 + \alpha\|\nabla\cdot\eta\|^2 + \frac{1}{4}\|\nabla\eta\|_{L^3}^2 \\
&\quad \left. + \frac{C}{\delta}\left(\|w\|^{3/2} + \|w^h\|^{3/2}\right)\|\eta\|_{L^6}^{3/2} + \frac{1}{4}\|\eta\|_{L^6}^2\right] \\
&\quad + \left[C\delta^{-1}\|\nabla w^h\|_{L^3}^{3/2}\right]\|\phi^h\|^{3/2} + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2\right]\|\phi^h\|^2.
\end{aligned}
$$

Consider the bracketed terms on this right-hand side. The first is approximation theoretic; the second is an L^1 function multiplying $\|\phi^h(t)\|^{3/2}$; the third is an L^1

function multiplying $\|\phi^h(t)\|^2$. Let $y(t) := \|\phi^h(t)\|^2$. This inequality may then be written as

$$\frac{d}{dt}y(t) + (\text{nonnegative terms}) \leq C(t)h^\gamma + a(t)y(t) + b(t)\delta^{-1}y^{3/4}(t),$$

where $a(t), b(t) \in L^1(0,T)$.

The final step would normally be to apply Gronwall's inequality to deduce $y(t) = \frac{1}{2}\|\phi^h(t)\|^2$ to be bounded by its initial values and approximation theoretic terms. Unfortunately, the term $y^{3/4}$ is not Lipschitz, so the argument fails at this last step.

Tracing the inequalities backward, the problem term arises from the steps used to bound $b(\phi^h, w^h, \phi^h)$ to obtain Re independence. The error analysis in the successful cases (i) and (ii) centers therefore on alternate bounds for this term. We shall first consider case (i).

Remark 4.5. If the estimate in (4.7) is improved as noted in Remark 4.3, the term $y(t)^{3/4}$ is changed to $y(t)^{6/7}$ but the final conclusion still holds.

4.2. The case $\nabla w \in L^3(0,T;L^3(\Omega))$ and $a_0(\delta) > 0$. The main result of this section is the following theorem.

THEOREM 4.1. *Assume $\alpha > 0$ and $a_0(\delta) > 0$. Let*

$$a(t) = \frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2 + \frac{C}{a_0(\delta)}\|\nabla w^h\|_{L^3}^2 + Ca_0(\delta)^{-1/2}\alpha^{-3/2}\|\mathbb{D}(w^h)\|_{L^3}^3.$$

Then, there is a $C_1 = C_1(\delta)$, independent of Re and h, such that

$$\|a(t)\|_{L^1(0,T)} \leq C_1(\delta).$$

Further, there is a $C_2 = C_2(\delta)$, independent of Re and h, such that

$$\frac{C}{\delta}\left(\|w\|^{3/2} + \|w^h\|^{3/2}\right) \leq C_2(\delta).$$

Then, the error $w - w^h$ satisfies for $T > 0$

$$\begin{aligned}
&\|w - w^h\|_{L^\infty(0,T;L^2)}^2 + \delta^2\|\mathbb{D}(w - w^h)\|_{L^3(0,T;L^3)}^3 + \alpha\|\nabla\cdot(w - w^h)\|_{L^2(0,T;L^2)}^2 \\
&\quad + \left(Re^{-1} + Ca_0(\delta)\right)\|\mathbb{D}(w - w^h)\|_{L^2(0,T;L^2)}^2 + \sum_{j=1}^{J}\beta\|(w - w^h)\cdot\hat{\tau}\|_{L^2(0,T;L^2(\Gamma_j))}^2 \\
&\leq C\exp(C_1(\delta))\|(w - w^h)(x,0)\|^2 + C\inf_{\tilde{w}\in V^h\cap(W^{1,3}(\Omega))^d, \lambda^h\in Q^h}\mathcal{F}(w - \tilde{w}, q - \lambda^h, \delta)
\end{aligned}$$

with

$$\begin{aligned}
&\mathcal{F}(w - \tilde{w}, r - q^h, \delta) \\
&= \|w - \tilde{w}\|_{L^\infty(0,T;L^2)}^2 + \delta^2\|\mathbb{D}(w - \tilde{w})\|_{L^3(0,T;L^3)}^3 \\
&\quad + \exp(C_1(\delta))\Bigg[\|(w - \tilde{w})(x,0)\|^2 + \delta^{-1}\|(w - \tilde{w})_t\|_{L^{3/2}(0,T;W^{-1,3/2})}^{3/2} \\
&\quad + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w - \tilde{w})\|_{L^2(0,T;L^2)}^2 + \sum_{j=1}^{J}\beta\|(w - \tilde{w})\cdot\hat{\tau}\|_{L^2(0,T;L^2(\Gamma_j))}^2 \\
&\quad + C(\delta)\|\mathbb{D}(w - \tilde{w})\|_{L^3(0,T;L^3)}^{3/2} + \alpha^{-1}\|q - \lambda^h\|_{L^2(0,T;L^2)}^2 + \alpha\|\nabla\cdot(w - \tilde{w})\|_{L^2(0,T;L^2)}^2 \\
&\quad + \|\nabla(w - \tilde{w})\|_{L^2(0,T;L^3)}^2 + \|w - \tilde{w}\|_{L^2(0,T;L^6)}^2 + C_2(\delta)\|w - \tilde{w}\|_{L^{3/2}(0,T;L^6)}^{3/2}\Bigg].
\end{aligned}$$

Proof. This analysis follows the previous discussion closely except for the treatment of the $b(\phi^h, w^h, \phi^h)$ term and the final application of Gronwall's inequality.

Consider, therefore, $b(\phi^h, w^h, \phi^h)$. Integration by parts and using the fact that $\phi^h \cdot \hat{n} = 0$ on Γ give

$$\begin{aligned} b(\phi^h, w^h, \phi^h) &= \frac{1}{2}(\phi^h \cdot \nabla w^h, \phi^h) - \frac{1}{2}(\phi^h \cdot \nabla \phi^h, w^h) \\ &= (\phi^h \cdot \nabla w^h, \phi^h) + \frac{1}{2}(\nabla \cdot \phi^h, \phi^h \cdot w^h) \\ &\le \|\nabla w^h\|_{L^3} \|\phi^h\|_{L^3}^2 + \frac{1}{2}|(\nabla \cdot \phi^h, \phi^h \cdot w^h)|. \end{aligned} \tag{4.9}$$

Using the embedding $H^{1/2} \hookrightarrow L^3$ in $d = 2, 3$ and Young's inequality give

$$\left|b(\phi^h, w^h, \phi^h)\right| \le \frac{\epsilon_1}{2}\|\mathbb{D}(\phi^h)\|^2 + \frac{C}{2\epsilon_1}\|\nabla w^h\|_{L^3}^2 \|\phi^h\|^2 + \frac{1}{2}|(\nabla \cdot \phi^h, \phi^h \cdot w^h)|. \tag{4.10}$$

Consider now the last term on the above right-hand side. By Hölder's inequality, we obtain

$$|(\nabla \cdot \phi^h, \phi^h \cdot w^h)| \le \|\nabla \cdot \phi^h\| \, \|\phi^h\|_{L^{r'}} \|w^h\|_{L^s},$$

where $\frac{1}{r'} + \frac{1}{s} = \frac{1}{2}$. Thus,

$$|(\nabla \cdot \phi^h, w^h \cdot \phi^h)| \le \frac{\alpha}{4}\|\nabla \cdot \phi^h\|^2 + \alpha^{-1}\|\phi^h\|_{L^{r'}}^2 \|w^h\|_{L^s}^2. \tag{4.11}$$

The Sobolev embedding theorem implies that for any $s, 1 \le s < \infty$ in two or three dimensions, $W^{1,3}(\Omega) \hookrightarrow L^s(\Omega)$. Thus,

$$\|w^h\|_{L^s}^2 \le C(s, \Omega) \, \|w^h\|_{W^{1,3}(\Omega)}^2 \le C(s, \Omega)\|\mathbb{D}(w^h)\|_{L^3}^2.$$

This implies that for any $r' > 2$

$$|(\nabla \cdot \phi^h, w^h \cdot \phi^h)| \le \frac{\alpha}{4}\|\nabla \cdot \phi^h\|^2 + C(r', \Omega)\alpha^{-1}\|\phi^h\|_{L^{r'}}^2 \|\mathbb{D}(w^h)\|_{L^3}^2.$$

Consider the last term on the above right-hand side. The Sobolev embedding theorem also implies

$$\|\phi^h\|_{L^{r'}} \le C(r', \Omega)\|\phi^h\|_{W^{t,2}(\Omega)} \text{ for } t \ge \frac{3}{2} - \frac{3}{r'}.$$

(The final result is not improved by applying here instead the Gagliardo–Nirenberg inequality.) As $r' \to 2, t \to 0$ in this inequality. Thus, picking $r' = r'(t) > 2$ close enough to 2 implies that, using an embedding inequality and Korn's inequality,

$$\|\phi^h\|_{L^{r'}}^2 \le C(t, \Omega)\|\phi^h\|_t^2 \le C(t, \Omega)\|\phi^h\|^{2(1-t)} \, \|\mathbb{D}(\phi^h)\|^{2t}$$

for any $t > 0$. Thus, for these values of r' and s

$$\frac{1}{\alpha}\|\phi^h\|_{L^{r'}}^2 \|w^h\|_{L^s}^2 \le \frac{C}{\alpha}\|\mathbb{D}(\phi^h)\|^{2t}\|\phi^h\|^{2(1-t)}\|\mathbb{D}(w^h)\|_{L^3}^2$$

for any $t > 0$. For conjugate exponents $q = 3$ and $q' = \frac{3}{2}$ in Young's inequality, we then have

$$\frac{1}{\alpha}\|\phi^h\|_{L^{r'}}^2 \|w^h\|_{L^s}^2 \le \frac{\epsilon}{3}\|\mathbb{D}(\phi^h)\|^{6t} + \epsilon^{-1/2}\alpha^{-3/2}C\|\phi^h\|^{3(1-t)}\|\mathbb{D}(w^h)\|_{L^3}^3.$$

Picking $t = \frac{1}{3} > 0$ gives for these values of r' and s

$$\frac{1}{\alpha}\|\phi^h\|_{L^{r'}}^2\|w^h\|_{L^s}^2 \leq \frac{\epsilon}{3}\|\mathbb{D}(\phi^h)\|^2 + C(r',s,t,\Omega)\epsilon^{-1/2}\alpha^{-3/2}\|\phi^h\|^2\|\mathbb{D}(w^h)\|_{L^3}^3.$$

Using this bound, (4.10) and (4.11) finally give

$$\begin{aligned}|b(\phi^h, w^h, \phi^h)| \leq{}& \frac{\epsilon_1}{2}\|\mathbb{D}(\phi^h)\|^2 + \frac{C}{2\epsilon_1}\|\nabla w^h\|_{L^3}^2\|\phi^h\|^2 \\ &+ \frac{\alpha}{8}\|\nabla\cdot\phi^h\|^2 + \frac{\epsilon_2}{6}\|\mathbb{D}(\phi^h)\|^2 + C\epsilon_2^{-1/2}\alpha^{-3/2}\|\mathbb{D}(w^h)\|_{L^3}^3\|\phi^h\|^2.\end{aligned}$$

Remark 4.6. It appears on first consideration that this last term $(\nabla\cdot\phi^h, w^h\cdot\phi^h)$ can be agreeably bounded more directly and easily by

$$\begin{aligned}|(\nabla\cdot\phi^h, w^h\cdot\phi^h)| &\leq C\|\nabla\cdot\phi^h\|\ \|\nabla w^h\|\ \|\phi^h\|^{1/2}\ \|\nabla\phi^h\|^{1/2} \\ &\leq C\|\nabla\phi^h\|^{3/2}\|\phi^h\|^{1/2}\|\nabla w^h\| \leq \epsilon\|\nabla\phi^h\|^2 + C(\epsilon)\|\nabla w^h\|^4\|\phi^h\|^2.\end{aligned}$$

This bound, while certainly true, is not sufficient because of the condition that inevitably arises from using it that w^h or $w \in L^4(0,T;H^1(\Omega))$. The extra work in the bound we use reduces the time regularity requirements arising from this term to $w^h \in L^3(0,T;W^{1,3}(\Omega))$ (which is bounded uniformly in Re by problem data in section 3).

Substituting this bound for $b(\phi^h, w^h, \phi^h)$ in the derivation of the upper estimate (4.8) for the difference of the convection terms gives

$$\begin{aligned}&|b(w,w,\phi^h) - b(w^h,w^h,\phi^h)| \\ &\leq \left[\frac{1}{4}\|\nabla\eta\|_{L^3}^2 + \frac{C}{3}\epsilon_3^{-1/2}\|w\|^{3/2}\|\eta\|_{L^6}^{3/2} + \frac{1}{4}\|\eta\|_{L^6}^2 + \frac{C}{3}\epsilon_3^{-1/2}\|\eta\|_{L^6}^{3/2}\|w^h\|^{3/2}\right] \\ &\quad + \left[\frac{\epsilon_3}{3}\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{\epsilon_1}{2}\|\mathbb{D}(\phi^h)\|^2 + \frac{\alpha}{8}\|\nabla\cdot\phi^h\|^2 + \frac{\epsilon_2}{6}\|\mathbb{D}(\phi^h)\|^2\right] \\ &\quad + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2 + \frac{C}{2\epsilon_1}\|\nabla w^h\|_{L^3}^2 + C\epsilon_2^{-1/2}\alpha^{-3/2}\|\mathbb{D}(w^h)\|_{L^3}^3\right]\|\phi^h\|^2.\end{aligned} \tag{4.12}$$

To proceed further, (4.12) is inserted in the right-hand side of (4.3). This yields the differential inequality

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2 + \left(\frac{1}{3}\underline{C}C_s\delta^2 - \frac{\epsilon_3}{3}\right)\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{3}{8}\alpha\|\nabla\cdot\phi^h\|^2 \\ &\quad + \left(\frac{1}{2}(2Re^{-1} + a_0(\delta)) - \frac{\epsilon_1}{2} - \frac{\epsilon_2}{6}\right)\|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|_{\Gamma_j}^2 \\ &\leq \left[\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|_{W^{-1,3/2}}^{3/2} + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\eta)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|_{\Gamma_j}^2\right. \\ &\quad + \frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|_{L^3}^{3/2} + \alpha^{-1}\|q-\lambda^h\|^2 + \alpha\|\nabla\cdot\eta\|^2 + \frac{1}{4}\|\nabla\eta\|_{L^3}^2 \\ &\quad \left. + \frac{C}{3}\epsilon_3^{-1/2}\left(\|w\|^{3/2} + \|w^h\|^{3/2}\right)\|\eta\|_{L^6}^{3/2} + \frac{1}{4}\|\eta\|_{L^6}^2\right] \\ &\quad + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2 + \frac{C}{\epsilon_1}\|\nabla w^h\|_{L^3}^2 + C\epsilon_2^{-1/2}\alpha^{-3/2}\|\mathbb{D}(w^h)\|_{L^3}^3\right]\|\phi^h\|^2.\end{aligned} \tag{4.13}$$

Pick $\epsilon_3 = C\underline{C}C_s\delta^2, C < 1/3$, $\epsilon_1 = a_0(\delta)/3$, and $\epsilon_2 = a_0(\delta)$. These choices simplify (4.13) to

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2 + C\underline{C}C_s\delta^2\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{3}{8}\alpha\|\nabla\cdot\phi^h\|^2 \\
&\quad + \left(Re^{-1} + \frac{a_0(\delta)}{6}\right)|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|_{\Gamma_j}^2 \\
&\le \left[\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|_{W^{-1,3/2}}^{3/2} + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\eta)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|_{\Gamma_j}^2\right. \\
&(4.14)\quad + \frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|_{L^3}^{3/2} + \alpha^{-1}\|q-\lambda^h\|^2 + \alpha\|\nabla\cdot\eta\|^2 \\
&\qquad \left. + \frac{1}{4}\|\nabla\eta\|_{L^3}^2 + \frac{C}{\delta}(\|w\|^{3/2} + \|w^h\|^{3/2})\|\eta\|_{L^6}^{3/2} + \frac{1}{4}\|\eta\|_{L^6}^2\right] \\
&\qquad + \left[\frac{1}{4}\|w\|_{L^6}^2 + \frac{1}{4}\|\nabla w^h\|_{L^3}^2 + \frac{C}{a_0(\delta)}\|\nabla w^h\|_{L^3}^2 + \frac{C}{a_0(\delta)^{1/2}\alpha^{3/2}}\|\mathbb{D}(w^h)\|_{L^3}^3\right]\|\phi^h\|^2.
\end{aligned}$$

Before applying Gronwall's inequality, let us first verify that it will indeed give us an error bound that is uniform in the Reynolds number by considering the coefficients on the right-hand side of (4.14).

First, note that $r \le C\|D(w)\|_{L^3}$. By the stability estimates $\|w\| \in L^\infty(0,T)$ and $\|w^h\| \in L^\infty(0,T)$ uniformly in Re. Thus,

$$\frac{C}{\delta}\|w\|^{3/2}\|\eta\|_{L^6}^{3/2} + \frac{C}{\delta}\|\eta\|_{L^6}^{3/2}\|w^h\|^{3/2} \le \frac{C}{\delta}(\|w\|^{3/2} + \|w^h\|^{3/2})\|\eta\|_{L^6}^{3/2} \le C_2(\delta)\|\eta\|_{L^6}^{3/2}.$$

Consider the (critical) bracketed coefficient of the last term on the right-hand side. We must show this coefficient is in $L^1(0,T)$ uniformly in Re. Indeed, by the stability estimates and the Sobolev imbedding $\|w\|_{L^6}$, $\|\mathbb{D}(w^h)\|_{L^3}$, $\|\mathbb{D}(w)\|_{L^3} \in L^3(0,T)$ uniformly in Re. Since $T < \infty$, $L^3(0,T) \subset L^2(0,T)$, and thus the first factor of the last term is in $L^1(0,T)$ uniformly in Re.

Hiding all constants in generic C's, Gronwall's lemma now implies for almost all $t \in [0,T]$ that

$$\begin{aligned}
&\|\phi^h(x,t)\|^2 + \delta^2\|\mathbb{D}(\phi^h)\|_{L^3(0,t;L^3)}^3 + \alpha\|\nabla\cdot\phi^h\|_{L^2(0,t;L^2)}^2 \\
&\quad + \left(Re^{-1} + Ca_0(\delta)\right)\|\mathbb{D}(\phi^h)\|_{L^2(0,t;L^2)}^2 + \sum_{j=1}^{J}\beta\|\phi^h\cdot\hat{\tau}\|_{L^2(0,t;L^2(\Gamma_j))}^2 \\
&\le C\exp\left(\|a(t)\|_{L^1(0,t)}\right)\|\phi^h(x,0)\|^2 \\
&\quad + C\exp\left(\|a(t)\|_{L^1(0,t)}\right)\Bigg[\delta^{-1}\|\eta_t\|_{L^{3/2}(0,T;W^{-1,3/2})}^{3/2} + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(\eta)\|_{L^2(0,t;L^2)}^2 \\
&\quad + \sum_{j=1}^{J}\beta\|\eta\cdot\hat{\tau}\|_{L^2(0,t;L^2(\Gamma_j))}^2 + \delta^2\int_0^t\|\mathbb{D}(w)\|_{L^3}^{3/2}\|\mathbb{D}(\eta)\|_{L^3}^{3/2}dt' \\
&\quad + \alpha^{-1}\|q-\lambda^h\|_{L^2(0,t;L^2)}^2 + \alpha\|\nabla\cdot\eta\|_{L^2(0,t;L^2)}^2 + \|\nabla\eta\|_{L^2(0,t;L^3)}^2 + \|\eta\|_{L^2(0,t;L^6)}^2 \\
&\quad + \int_0^t\frac{1}{\delta}(\|w\|^{3/2} + \|w^h\|^{3/2})\|\eta\|_{L^6}^{3/2}dt'\Bigg].
\end{aligned}$$

Note that by the Cauchy–Schwarz inequality in $L^2(0,t)$, $t \in [0,T]$, and the stability estimates

$$\int_0^t \|\mathbb{D}(w)\|_{L^3}^{3/2} \|\mathbb{D}(\eta)\|_{L^3}^{3/2} dt' \leq \|\mathbb{D}(w)\|_{L^3(0,t;L^3)}^{3/2} \|\mathbb{D}(\eta)\|_{L^3(0,t;L^3)}^{3/2} \leq C(\delta) \|\mathbb{D}(\eta)\|_{L^3(0,t;L^3)}^{3/2}.$$

Now, the essential supremum of $t \in [0,T]$ is applied on both sides of the inequality. As $w - w^h = \eta - \phi^h$, the triangle inequality completes the proof of Theorem 4.1. ∎

Remark 4.7. On the condition $\alpha > 0$, the least squares control of $\nabla \cdot u$ seems to be essential to get an estimate uniform in Re. Consider (4.9) in the proof of Theorem 4.1. There are two important nonlinear terms in the error equation corresponding loosely to convection and reaction. The reaction term is controlled by the subgrid model. The convection term can be converted into a reaction-like term. It is controllable provided that $\nabla \cdot \phi^h$ is controllable, which $\alpha > 0$ accomplishes.

Another promising approach is to use a variational formulation, such as SUPG developed by Brooks and Hughes [3], which will control the convection term directly. We note that both SUPG and least squares control of $\nabla \cdot u$ are consistent: they work on the error and do not change the solution.

4.3. The case $\nabla w \in L^2(0,T;L^\infty(\Omega))$ and $a_0(\delta) \geq 0$. We now consider the case of smoother w, i.e.,

$$w \in L^2(0,T;W^{1,\infty}(\Omega)) \text{ uniformly in } Re,$$

allowing for the case $a_0(\delta) \equiv 0$. This case is primarily of interest because many tests involve "academic" flow fields given in closed form (as in section 5). These are typically smooth and bounded. In this case Theorem 4.2 gives an error estimate with constants independent of Re (but depending on δ and α). It is noteworthy in this estimate that *multiplicative constants* depend on δ but the *rate constant* in the (inevitable) exponential term takes the form

$$\exp(C_3(w)), \quad C_3 = C_3(\|w\|_{L^2(0,T;W^{1,\infty}(\Omega))}),$$

with no *explicit* dependence on δ.

THEOREM 4.2. *Suppose $a_0(\delta) \geq 0$, $\alpha > 0$, and $w \in L^2(0,T;W^{1,\infty}(\Omega))$ uniformly in Re. Let*

$$a(t) := \frac{3}{4} + \|\nabla w\|_{L^\infty} + \left(\frac{1}{4} + \frac{1}{4\alpha}\right) \|w\|_{L^\infty}^2 + \frac{1}{2} \|\nabla w\|_{L^\infty}^2;$$

then there is a $C_3 = C_3(w)$ such that

$$\|a(t)\|_{L^1(0,T)} \leq C_3(w).$$

Let $C_4 = C_4(\delta)$ be such that

$$\|\mathbb{D}(w^h)\|_{L^3(0,T;L^3)} \leq C_4(\delta).$$

Then, the error $w - w^h$ satisfies

$$\begin{aligned}
&\|w - w^h\|_{L^\infty(0,T;L^2)}^2 + \delta^2 \|\mathbb{D}(w - w^h)\|_{L^3(0,T;L^3)}^3 + \alpha \|\nabla \cdot (w - w^h)\|_{L^2(0,T;L^2)}^2 \\
&\quad + \left(Re^{-1} + C a_0(\delta)\right) \|\mathbb{D}(w - w^h)\|_{L^2(0,T;L^2)}^2 + \sum_{j=1}^{J} \beta \|(w - w^h) \cdot \hat{\tau}\|_{L^2(0,T;L^2(\Gamma_j))}^2 \\
&\leq C \exp(C_3(w)) \|(w - w^h)(x,0)\|^2 + C \inf_{\tilde{w} \in V^h \cap (W^{1,3}(\Omega))^d, \lambda^h \in Q^h} \mathcal{F}(w - \tilde{w}, q - \lambda^h, \delta)
\end{aligned}$$

with

$$
\begin{aligned}
&\mathcal{F}(w-\tilde{w}, r-q^h, \delta)\\
&= \|w-\tilde{w}\|^2_{L^\infty(0,T;L^2)} + \delta^2\|\mathbb{D}(w-\tilde{w})\|^3_{L^3(0,T;L^3)}\\
&\quad + \exp(C_3(w))\Bigg[\|(w-\tilde{w})(x,0)\|^2 + \delta^{-1}\|(w-\tilde{w})_t\|^{3/2}_{L^{3/2}(0,T;W^{-1,3/2})}\\
&\quad + (2Re^{-1} + a_0(\delta))\|\mathbb{D}(w-\tilde{w})\|^2_{L^2(0,T;L^2)} + \sum_{j=1}^{J}\beta\|(w-\tilde{w})\cdot\hat{\tau}\|^2_{L^2(0,T;L^2(\Gamma_j))}\\
&\quad + C(\delta)\|\mathbb{D}(w-\tilde{w})\|^{3/2}_{L^3(0,T;L^3)} + \alpha^{-1}\|q-\lambda^h\|^2_{L^2(0,T;L^2)}\\
&\quad + \left(\frac{1}{4}+\alpha\right)\|\nabla\cdot(w-\tilde{w})\|^2_{L^2(0,T;L^2)} + \|w-\tilde{w}\|^2_{L^2(0,T;L^2)}\\
&\quad + C_4(\delta)\left(\|\mathbb{D}(w-\tilde{w})\|^2_{L^{18/5}(0,T;L^3)} + \|w-\tilde{w}\|^2_{L^6(0,T;L^6)}\right)\Bigg].
\end{aligned}
$$

Proof. In this case, the difference in the nonlinear terms is decomposed a bit differently as

$$
\begin{aligned}
|b(w,w,\phi^h) - b(w^h,w^h,\phi^h)| &= |b(\eta-\phi^h,w,\phi^h) + b(w^h,\eta-\phi^h,\phi^h)|\\
&= |b(\eta,w,\phi^h) - b(\phi^h,w,\phi^h) + b(w^h,\eta,\phi^h)|.
\end{aligned}
\tag{4.15}
$$

Consider the individual terms on the right-hand side of (4.15):

$$
\begin{aligned}
|b(\eta,w,\phi^h)| &= \left|\frac{1}{2}(\eta\cdot\nabla w,\phi^h) - \frac{1}{2}(\eta\cdot\nabla\phi^h,w)\right|\\
&= \left|(\eta\cdot\nabla w,\phi^h) + \frac{1}{2}(\nabla\cdot\eta,\phi^h\cdot w)\right|\\
&\le \frac{1}{2}\|\eta\|^2 + \frac{1}{2}\|\nabla w\|^2_{L^\infty}\|\phi^h\|^2 + \frac{1}{4}\|\nabla\cdot\eta\|^2 + \frac{1}{4}\|w\|^2_{L^\infty}\|\phi^h\|^2,\\
|b(\phi^h,w,\phi^h)| &= \left|(\phi^h\cdot\nabla w,\phi^h) + \frac{1}{2}(\nabla\cdot\phi^h,w\cdot\phi^h)\right|\\
&\le \|\nabla w\|_{L^\infty}\|\phi^h\|^2 + \frac{\alpha}{4}\|\nabla\cdot\phi^h\|^2 + \frac{1}{4\alpha}\|w\|^2_{L^\infty}\|\phi^h\|^2,\\
|b(w^h,\eta,\phi^h)| &= \left|(w^h\cdot\nabla\eta,\phi^h) + \frac{1}{2}(\nabla\cdot w^h,\eta\cdot\phi^h)\right|\\
&\le \|w^h\|_{L^6}\|\nabla\eta\|_{L^3}\|\phi^h\| + \frac{1}{2}\|\nabla\cdot w^h\|_{L^3}\|\eta\|_{L^6}\|\phi^h\|\\
&\le C\left(\|w^h\|^2_{L^6}\|\nabla\eta\|^2_{L^3} + \|\mathbb{D}(w^h)\|^2_{L^3}\|\eta\|^2_{L^6}\right) + \frac{3}{4}\|\phi^h\|^2.
\end{aligned}
$$

Combining these three estimates gives

$$
\begin{aligned}
&|b(w,w,\phi^h) - b(w^h,w^h,\phi^h)|\\
&\le \frac{1}{2}\|\eta\|^2 + \frac{1}{4}\|\nabla\cdot\eta\|^2 + C\|w^h\|^2_{L^6}\|\nabla\eta\|^2_{L^3} + C\|\mathbb{D}(w^h)\|^2_{L^3}\|\eta\|^2_{L^6} + \frac{\alpha}{4}\|\nabla\cdot\phi^h\|^2\\
&\quad + \left(\frac{3}{4} + \frac{1}{2}\|\nabla w\|^2_{L^\infty} + \frac{1}{4}\|w\|^2_{L^\infty} + \|\nabla w\|_{L^\infty} + \frac{1}{4\alpha}\|w\|^2_{L^\infty}\right)\|\phi^h\|^2.
\end{aligned}
\tag{4.16}
$$

The term $\|w^h\|_{L^6}$ is bounded using the Gagliardo–Nirenberg inequality (Lemma 2.5)

$$\|w^h\|_{L^6}^2 \leq C\|w^h\|^{2/3}\|\mathbb{D}(w^h)\|_{L^3}^{4/3}.$$

Since $\|w^h\|$ is bounded uniformly in ν and h by (3.5) or (3.6), it follows that

$$\|w^h\|_{L^6}^2 \leq C\|\mathbb{D}(w^h)\|_{L^3}^{4/3}.$$

This bound, together with (4.16), is now inserted in the right-hand side of (4.3) giving

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\phi^h\|^2 + \frac{1}{3}\underline{C}C_s\delta^2\|\mathbb{D}(\phi^h)\|_{L^3}^3 + \frac{\alpha}{2}\|\nabla\cdot\phi^h\|^2 \\
&\quad + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\phi^h)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\phi^h\cdot\hat{\tau}\|_{\Gamma_j}^2 \\
&\leq \left[\frac{2}{3}(CC_s)^{-1/2}\delta^{-1}\|\eta_t\|_{W^{-1,3/2}}^{3/2} + \frac{1}{2}(2Re^{-1} + a_0(\delta))\|\mathbb{D}(\eta)\|^2 + \sum_{j=1}^{J}\frac{\beta}{2}\|\eta\cdot\hat{\tau}\|_{\Gamma_j}^2\right. \\
&\quad + \frac{2}{3}\underline{C}^{-1/2}C_s\overline{C}^{3/2}r^{3/2}\delta^2\|\mathbb{D}(\eta)\|_{L^3}^{3/2} + \alpha^{-1}\|q-\lambda^h\|^2 + \alpha\|\nabla\cdot\eta\|^2 + \frac{1}{2}\|\eta\|^2 \\
&\quad \left. + \frac{1}{4}\|\nabla\cdot\eta\|^2 + C\|\mathbb{D}(w^h)\|_{L^3}^{4/3}\|\nabla\eta\|_{L^3}^2 + C\|\mathbb{D}(w^h)\|_{L^3}^2\|\eta\|_{L^6}^2\right] + \left[\frac{\alpha}{4}\|\nabla\cdot\phi^h\|^2\right] \\
&\quad + \left(\frac{3}{4} + \|\nabla w\|_{L^\infty} + \left(\frac{1}{4}+\frac{1}{4\alpha}\right)\|w\|_{L^\infty}^2 + \frac{1}{2}\|\nabla w\|_{L^\infty}^2\right)\|\phi^h\|^2.
\end{aligned}$$

To apply Gronwall's inequality we need

$$\frac{3}{4} + \|\nabla w\|_{L^\infty} + \left(\frac{1}{4}+\frac{1}{4\alpha}\right)\|w\|_{L^\infty}^2 + \frac{1}{2}\|\nabla w\|_{L^\infty}^2 \in L^1(0,T),$$

in other words $w \in L^2(0,T;W^{1,\infty}(\Omega))$. The term on the right-hand side of this inequality containing $r^{3/2}$ is treated as in the proof of Theorem 4.1. In the final result of Gronwall's lemma, we must also verify that the resulting terms containing $\|\mathbb{D}(w^h)\|_{L^3}$ are bounded uniformly in Re. To this end, apply Hölder's inequality

$$\int_0^T \|\mathbb{D}(w^h)\|_{L^3}^{4/3}\|\mathbb{D}(\eta)\|_{L^3}^2 dt \leq \|\mathbb{D}(w^h)\|_{L^{4q/3}(0,T;L^3)}^{4/3}\|\mathbb{D}(\eta)\|_{L^{2q'}(0,T;L^3)}^2,$$

where $\frac{1}{q} + \frac{1}{q'} = 1$. From the stability estimates, we clearly must take q such that $4q/3 \leq 3$. Accordingly, take $q = \frac{9}{4}, q' = \frac{9}{5}$. This gives

$$\begin{aligned}
\int_0^T \|\mathbb{D}(w^h)\|_{L^3}^{4/3}\|\mathbb{D}(\eta)\|_{L^3}^2 dt &\leq C\|\mathbb{D}(w^h)\|_{L^3(0,T;L^3)}^{4/3}\|\mathbb{D}(\eta)\|_{L^{18/5}(0,T;L^3)}^2 \\
&\leq CC_4(\delta)\|\mathbb{D}(\eta)\|_{L^{18/5}(0,T;L^3)}^2.
\end{aligned}$$

Similarly, for q and q' conjugate exponents take $q = \frac{3}{2}$, $q' = 3$,

$$\begin{aligned}
\int_0^T \|\mathbb{D}(w^h)\|_{L^3}^2\|\eta\|_{L^6}^2 dt &\leq \|\mathbb{D}(w^h)\|_{L^{2q}(0,T;L^3)}^2\|\eta\|_{L^{2q'}(0,T;L^6)}^2 \\
&\leq \|\mathbb{D}(w^h)\|_{L^3(0,T;L^3)}^2\|\eta\|_{L^6(0,T;L^6)}^2 \leq C_4(\delta)\|\eta\|_{L^6(0,T;L^6)}^2.
\end{aligned}$$

The stated error estimate now follows from Gronwall's inequality and the triangle inequality as in the proof of Theorem 4.1. ∎

5. A numerical example. To give a numerical illustration several decisions must be made, mainly whether to work on an "academic" flow problem with a known exact solution or to work on a more realistic flow problem containing the accompanying uncertainties. Since our aim is to illustrate a convergence theorem, we have chosen the former. (To assess a model or study the limitations of an algorithm, we would naturally have chosen the latter.) Accordingly, we have selected the vortex decay problem of Chorin [4], used also by others, e.g., Tafti [38]. The domain is $\Omega = (0,1)^2$ and we choose

$$\begin{aligned} w_1 &= -\cos(n\pi x)\sin(n\pi y)\exp(-2n^2\pi^2 t/\tau), \\ w_2 &= \sin(n\pi x)\cos(n\pi y)\exp(-2n^2\pi^2 t/\tau), \\ q &= -\tfrac{1}{4}(\cos(2n\pi x)+\cos(2n\pi y))\exp(-4n^2\pi^2 t/\tau). \end{aligned} \tag{5.1}$$

For the relaxation time $\tau = Re$ this is a solution of the Navier–Stokes equation consisting of an array of opposite signed vortices which decay as $t \to \infty$. The right-hand side f, initial condition, and nonhomogeneous Dirichlet boundary conditions are chosen so that (w_1, w_2, q) is the closed form solution of (1.3).

Since we are studying convergence as $h \to 0$ for δ fixed and Re varying we have accordingly chosen a 4×4 array of the vortices (so $n = 4$) and

$$\begin{aligned} \tau &= 1000, \\ \text{final time } T &= 8, \\ \text{eddy scale } \delta &= 0.1, \\ \text{Smagorinski constant } C_s &= 0.05, \\ a_0(\delta) &= 0. \end{aligned}$$

It is significant that $\delta = 0.1 \le \frac{1}{4} = \frac{1}{n}$ so that the vortices are larger than $O(\delta)$ and hence should be "visible" to the model.

The fractional—step θ—scheme with an equal distant time step $\Delta t_n = 0.001$ is used as discretization in time. The time discretization error should be kept small by using this very small time step. In space, the Q_2/P_1^{disc} and the Q_3/P_2^{disc} finite element discretizations are applied; see Table 1 for the number of degrees of freedom for different mesh sizes. The unit square was divided into an $h \times h$ mesh with $h = 1/2$ on level 0. Both the Smagorinsky subgridscale model and the convection term are treated implicitly. The viscous term is treated not as $(\nabla w^h, \nabla v^h)$ but as using the deformation tensor formulation, $(\mathbb{D}(w^h), \mathbb{D}(v^h))$, as analyzed herein. The least squares constant α is chosen to be zero and we used the convective form of the nonlinear convection term. The nonlinear system in each time step is solved up to a Euclidean norm of the residual vector less than 10^{-10}.

The numbers of degrees of freedom in space are certainly not extremely large. However, their importance is only relative to the Reynolds number, ranging from 10^2

TABLE 1
Mesh widths and degrees of freedom in space.

	Q_2/P_1^{disc}			Q_3/P_2^{disc}		
Mesh width	Velocity	Pressure	Total	Velocity	Pressure	Total
1/4	-	-	-	338	96	434
1/8	578	192	770	1 250	384	1 634
1/16	2 178	768	2 946	4 802	1 536	6 338
1/32	8 450	3 072	11 522	18 818	6 144	24 962
1/64	33 282	12 288	45 570	-	-	-
1/128	132 098	49 152	181 250	-	-	-

TABLE 2
Q_2/P_1^{disc} *finite element discretization,* $\|e\|_{L^\infty(0,T;L^2)}$.

$Re \setminus h$	1/8	1/16	1/32	1/64	1/128
10^2	2.20176e-2	2.76780e-3	3.47796e-4	4.35185e-5	5.43988e-6
10^3	3.19389e-2	3.50372e-3	4.81015e-4	4.86864e-5	5.50381e-6
10^4	5.97051e-2	7.01100e-3	1.00294e-3	1.39466e-4	1.44707e-5
10^5	7.67057e-2	7.73782e-3	1.09801e-3	1.62252e-4	1.92555e-5
10^6	7.86394e-2	7.81755e-3	1.10830e-3	1.64891e-4	1.98666e-5
10^7	7.88349e-2	7.82560e-3	1.10934e-3	1.65161e-4	1.99290e-5
10^8	7.88545e-2	7.82641e-3	1.10945e-3	1.65188e-4	1.99373e-5
10^9	7.88564e-2	7.82649e-3	1.10946e-3	1.65190e-4	1.99379e-5
10^{10}	7.88566e-2	7.82650e-3	1.10946e-3	1.65191e-4	1.99380e-5

TABLE 3
Q_3/P_2^{disc} *finite element discretization,* $\|e\|_{L^\infty(0,T,L^2)}$.

$Re \setminus h$	1/4	1/8	1/16	1/32
10^2	3.09237e-2	2.14568e-3	1.39746e-4	8.96881e-6
10^3	7.61050e-2	2.53153e-3	1.40003e-4	8.85819e-6
10^4	1.09160e-1	2.79077e-3	1.43102e-4	8.81959e-6
10^5	1.13716e-1	2.82963e-3	1.43815e-4	8.81915e-6
10^6	1.14186e-1	2.83371e-3	1.43896e-4	8.81835e-6
10^7	1.14234e-1	2.83412e-3	1.43904e-4	8.81929e-6
10^8	1.14238e-1	2.83416e-3	1.43905e-4	8.81853e-6
10^9	1.14239e-1	2.83417e-3	1.43905e-4	8.81961e-6
10^{10}	1.14239e-1	2.83417e-3	1.43905e-4	8.81754e-6

to 10^{10}, and the resolution sought, $\delta = 0.1$. Again, LES is focused on situations in which the number of degrees of freedom is small relative to Re. Thus, the chosen values of h and Re seem appropriate.

Tables 2 and 3 present the $L^\infty(0,T;L^2)$ norm of the error for both discretizations in space. Note that the behavior is exactly as anticipated by the theory: the error in this norm is clearly independent of Re.

Tables 4 and 5 present the errors in $L^2(0,T;H^1)$. These errors are not predicted to be in general uniform in Re. But in the particular example which we have chosen, one can observe uniformity in Re.

6. Conclusions. Reynolds number dependence in finite element error analysis arises in three basic places: multiplicative error constants (Re), time scale constants ($\exp(C(Re)T)$), and time regularity assumptions on the true solution (needed even to prove continuous dependence on the initial data) which might fail for turbulent flows. In the error analysis of a large eddy model all three sources must be addressed. The idea of our error analysis herein for the Smagorinsky model has been that the greater spatial regularity of the large eddies must be used to compensate for the reduced time regularity of the underlying turbulent flow. The execution of this idea is necessarily technical since it entails using, in so far as possible, L^3 bounds (the natural norm arising from the model) for the nonlinear error terms. For different models, this same idea can be possibly applied; its execution will vary with the particular features of the model.

The error equation contains nonlinear terms resembling both convection and reaction. Our analysis suggests that uniformity in Re can be accomplished by the control of both effects. The second is controlled by the subgrid model while the first seems to need a correctly adapted numerical method; see Remark 4.7.

TABLE 4
Q_2/P_1^{disc} *finite element discretization,* $\|\mathbb{D}(e)\|_{L^2(0,T,L^2)}$.

$Re \setminus h$	1/8	1/16	1/32	1/64	1/128
10^2	1.24827	3.13720e-1	7.84736e-2	1.96114e-2	4.90234e-3
10^3	1.56935	3.60470e-1	8.42787e-2	2.00913e-2	4.93406e-3
10^4	2.35100	4.66554e-1	1.05387e-1	2.34301e-2	5.28506e-3
10^5	2.68127	4.98844e-1	1.14609e-1	2.61063e-2	5.79700e-3
10^6	2.72037	5.02793e-1	1.15920e-1	2.66473e-2	5.96091e-3
10^7	2.72434	5.03197e-1	1.16058e-1	2.67093e-2	5.98435e-3
10^8	2.72474	5.03237e-1	1.16072e-1	2.67156e-2	5.98686e-3
10^9	2.72478	5.03241e-1	1.16073e-1	2.67162e-2	5.98711e-3
10^{10}	2.72478	5.03242e-1	1.16073e-1	2.67163e-2	5.98714e-3

TABLE 5
Q_3/P_2^{disc} *finite element discretization,* $\|\mathbb{D}(e)\|_{L^2(0,T,L^2)}$.

$Re \setminus h$	1/4	1/8	1/16	1/32
10^2	1.15300	1.65587e-1	2.07562e-2	2.60050e-3
10^3	2.87920	1.93565e-1	2.15627e-2	2.62325e-3
10^4	4.79996	2.24195e-1	2.27512e-2	2.67749e-3
10^5	5.07949	2.31037e-1	2.31277e-2	2.71302e-3
10^6	5.10843	2.31820e-1	2.31768e-2	2.72525e-3
10^7	5.11134	2.31899e-1	2.31819e-2	2.72679e-3
10^8	5.11163	2.31907e-1	2.31824e-2	2.72674e-3
10^9	5.11166	2.31908e-1	2.31825e-2	2.72698e-3
10^{10}	5.11166	2.31908e-1	2.31825e-2	2.72759e-3

We note that replacing multipliers like $\exp(C(Re)T)$ in the error estimate by $\exp(C(\delta)T)$ establishes that a LES will be valid over a much longer time interval than a DNS, although still not over $0 < t \leq \infty$. It would certainly be interesting to know which flow statistics could be accurately approximated over $0 < t \leq \infty$, but this requires a different analysis.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] J. BARDINA, J. FERZIGER, AND W. REYNOLDS, *Improved Subgrid Models for Large Eddy Simulation*, AIAA Paper 80-1357, Reston, VA, 1980.

[3] A. BROOKS AND T. HUGHES, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations.*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[4] A. CHORIN, *Numerical solution for the Navier–Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[5] M. CROUZEIX AND V. THOMÉE, *The stability in L^p and $W^{1,p}$ of the L^2-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–523.

[6] E. DIBENEDETTO, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[7] Q. DU AND M. GUNZBURGER, *Analysis of a Ladyzhenskaya model for incompressible viscous flow*, J. Math. Anal. Appl., 155 (1991), pp. 21–45.

[8] M. ENGLEMAN, R. SANI, AND P. GRESHO, *The implementation of normal and/or tangential boundary conditions in finite element codes for incompressible fluid flows*, Internat. J. Numer. Methods Fluids, 2 (1982), pp. 225–238.

[9] J. FERZIGER AND M. PERIC, *Computational Methods for Fluid Dynamics*, 2nd ed., Springer-Verlag, Berlin, 1999.

[10] G. FICHERA, *Existence theorems in elasticity*, in Handbuch der Physik, Springer-Verlag, Berlin, 1972.

[11] U. FRISCH, *Turbulence. The Legacy of A. N. Kolmogorov*, Cambridge University Press, Cambridge, UK, 1995.

[12] G. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations, Vol.* I: *Linearized Theory*, Springer Tracts Nat. Philos. 38, Springer-Verlag, New York, 1994.
[13] G. Galdi, J. Heywood, and R. Rannacher, eds., *Fundamental Directions in Mathematical Fluid Mechanics*, Adv. Math. Fluid Mech., Birkhäuser, Basel, 2000.
[14] P. Galdi and W. Layton, *Approximation of the larger eddies in fluid motion* II: *A model for space filtered flow*, Math. Models Methods Appl. Sci., 10 (2000), pp. 343–350.
[15] M. Germano, U. Piomelli, P. Moin, and W. Cabot, *A dynamic subgrid–scale eddy viscosity model*, Phys. Fluids A, 3 (1991), pp. 1760–1765.
[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1986.
[17] J. Gobert, *Une inequation fondametale de la théorie de lélasticité*, Bull. Soc. Roy. Sci. Liège, 31 (1962), pp. 182–191.
[18] J. Gobert, *Sur une in égalité de coercivité*, J. Math. Anal. Appl., 36 (1971), pp. 518–528.
[19] P. Gresho and R. Sani, *Incompressible Flow and the Finite Element Method*, Vol. 1, John Wiley, Chichester, UK, 2000.
[20] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Springer-Verlag, Paris, 1992.
[21] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, 1989.
[22] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[23] T. Hughes, L. Mazzei, and K. Jansen, *Large eddy simulation and the variational multiscale method*, Comput. Visual. Sci., 3 (2000), pp. 47–59.
[24] T. Iliescu, V. John, W. Layton, G. Matthies, and L. Tobiska, *A numerical study of a class of LES models*, Internat. J. Comput. Fluid Dynam., to appear.
[25] T. Iliescu and W. Layton, *Approximating the larger eddies in fluid motion* III: *The Boussinesq model for turbulent fluctuations*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi Mat. (N.S.), 44 (1998), pp. 245–261.
[26] V. John, W. Layton, and N. Sahin, *Derivation and analysis of near wall models for channel and recirculating flows*, J. Engrg. Math., submitted.
[27] O. Ladyshenskaya, *New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary value problems for them*, Proc. Steklov Inst. Math., 102 (1967), pp. 95–118.
[28] O. Ladyzhenskaya, *The Mathemetical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, London, Paris, 1969.
[29] O. Ladyzhenskaya, *Modification of the Navier-Stokes equations for large velocity gradients*, in Boundary Value Problems of Mathematical Physics and Related Aspects of Function Theory, Consultants Bureau, New York, 1970, pp. 57–69.
[30] W. J. Layton, *A nonlinear, subgridscale model for incompressible viscous flow problems*, SIAM J. Sci. Comput., 17 (1996), pp. 347–357.
[31] A. Liakos, *Weak Imposition of Boundary Conditions in the Stokes and Navier–Stokes Equation*, Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, 1999.
[32] B. Mohammadi and O. Pironneau, *Analysis of the K-Epsilon Turbulence Model*, John Wiley, Chichester, UK, 1994.
[33] L. Nirenberg, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa (3), 13 (1959), pp. 115–162.
[34] C. Parés, *Existence, uniqueness and regularity of solutions of the equations of a turbulence model for incompressible fluids*, Appl. Anal., 43 (1992), pp. 245–296.
[35] P. Sagaut, *Large Eddy Simulation for Incompressible Flows*, Springer-Verlag, Berlin, Heidelberg, New York, 2001.
[36] J. Serrin, *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1959, pp. 125–263.
[37] J. Smagorinsky, *General circulation experiments with the primitive equations*, Mon. Weather Rev., 91 (1963), pp. 99–164.
[38] D. Tafti, *Comparison of some upwind–biased high–order formulations with a second–order central–difference scheme for time integration of the incompressible Navier–Stokes equations*, Comput. & Fluids, 25 (1996), pp. 647–665.
[39] R. Temam, *Problemes Mathematiques en Plasticite*, Gauthier–Villars, Paris, 1983.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1021–1040

A SINGULAR FIELD METHOD FOR MAXWELL'S EQUATIONS: NUMERICAL ASPECTS FOR 2D MAGNETOSTATICS*

CHRISTOPHE HAZARD† AND STEPHANIE LOHRENGEL‡

Abstract. The present paper deals with the solution of Maxwell-type problems by means of nodal H^1-conforming finite elements. In a nonconvex piecewise regular domain surrounded by a perfect conductor, such a discretization cannot in general approximate the singular behavior of the electromagnetic field near "reentrant" corners or edges. The *singular field method* consists of adding to the finite element discretization space some particular fields which take into account the singular behavior. The latter are deduced from the singular functions associated with the scalar Laplace operator.

The theoretical justification of this approach as well as the analysis of the convergence of the approximation are presented for a very simple model problem arising from magnetostatics in a translation invariant setting, but the study can be easily extended to numerous Maxwell-type problems. The numerical implementation of both variants is studied for a domain containing a single reentrant corner.

Key words. Maxwell's equations, singularities of solutions, finite element method, singular function methods, error analysis

AMS subject classifications. 35A40, 35B65, 78M10

PII. S0036142900375761

1. Introduction.

In a previous paper [8], we described the theoretical aspects of a method for solving Maxwell's equations in a polyhedron of $\mathbb{R}^3$ by means of nodal finite elements. This approach is based on the fact that a H^1-conforming discretization cannot in general approximate the singular behavior of the electromagnetic field near nonconvex edges or corners of a perfect conductor (which explains the imperfection of the attempts in engineering to make nodal elements work; see, e.g., [6] and the remarks in [9]). However, this behavior can be expressed explicitly in terms of the so-called *singularities* of the scalar Laplace operator. The idea of the *singular field method* consists of using this explicit knowledge by splitting the electromagnetic field into two parts: a *regular* part which can be approximated by a H^1-conforming finite element method and a *singular* part which is taken into account explicitly.

The initial motivation of this work was to find a cure for the failure of classical Lagrange elements for solving Maxwell-type problems in nonconvex piecewise regular domains. To a certain extent the *singular field method* proposes a way to adapt usual codes for those who cling to H^1-conforming elements! The question is to know whether this approach represents a good alternative to the popular edge elements. The answer depends on the problem to be solved. Of course, edge elements take into account the singular behavior of the field near reentrant corners, provided the mesh is properly refined. However, in the case of the time-dependent Maxwell equations, refinement in space induces refinement in time and thus heavy computations. Nodal

*Received by the editors July 24, 2000; accepted for publication (in revised form) February 14, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/sinum/40-3/37576.html

†Laboratoire de Simulation et Modélisation des Phénomènes de Propagation (URA CNRS 853), Ecole Nationale Supérieure de Techniques Avancées: ENSTA / SMP, 32 Boulevard Victor, 75739 Paris Cedex 15, France (hazard@ensta.fr).

‡Laboratoire J.A. Dieudonné (UMR CNRS 6621), Université de Nice–Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France (lohrenge@math.unice.fr).

elements are also advantageous when coupled with other codes requiring continuous fields, for instance in the resolution of the Vlasov–Maxwell system.

The aim of the present paper is not to compare both methods (which would require a precise investigation of convergence rates) but rather to describe some practical aspects of the *singular field method* in the two-dimensional case.

For the sake of simplicity, we consider a simple model problem which describes magnetostatics in a translation invariant setting and holds for transversal components of the field. Such a particular problem is hardly of practical interest. However, as mentioned in [8], there are numerous straightforward extensions of the method for Maxwell-type problems: time-harmonic full Maxwell's equations, eddy current models, inhomogeneous media, exterior scattering by obstacles or screens, etc. A more interesting but more complex application is presented in [25]: it deals with the determination of guided modes in an inhomogeneous cylindrical optical waveguide. More generally, the method applies for two-dimensional elliptic variational problems (or compact perturbations of such problems) in some piecewise $\mathcal{H}(\mathrm{curl}, \mathrm{div})$ space (i.e., for fields $\boldsymbol{E} \in L^2$ such that $\mathrm{curl}\, \boldsymbol{E} \in L^2$ and $\mathrm{div}\, \boldsymbol{E} \in L^2$) containing proper boundary or transmission conditions along piecewise regular curves. To extend the *singular field method* to such a problem, the main difficulty consists of identifying the singular fields related to the variational space, which amounts to the determination of scalar singular functions (by means of the techniques described for instance in Grisvard [21, 22] or Dauge [18]).

The paper is organized as follows. In section 2, we define our model problem and exhibit an equivalent *regularized* formulation, as well as a "spurious" regularized problem: the latter consists in the same variational equation as the former, set in a smaller functional space (which contains only regular fields). The key to the method lies in the fact that the latter space can be completed by a finite dimensional space of *singular fields* in order to solve the proper regularized problem: two such decompositions are given in section 3. They lead to the *singular field method* (SFM) and its orthogonal variant, the orthogonal singular field method (OSFM) which are described in section 4. The analysis of the convergence of these methods is the object of section 5. Both numerical schemes seem to have the same rate of convergence, but the numerical applications presented in section 6 clearly show that the OSFM yields far better results: we shall try to explain why.

Let us mention a similar approach developed by Assous, Ciarlet, and Segré [3] and Assous, Ciarlet, and Garcia [2] in the context of time-dependent Maxwell's equations (see also [23] for a unified presentation of both approaches): their method is based on an alternative decomposition of the electromagnetic field into regular and singular components. This approach has been extended to three dimensions in the case of axisymmetric geometries [1, 11]. Also worth mentioning is the procedure proposed by Costabel, Dauge, and Martin [17]. Contrary to our approach, the idea is to penalize the homogeneous boundary condition near the perfect conductor by an impedance condition simulating an imperfect conductor with high conductivity. From a theoretical point of view, the addition of singular fields is no longer necessary. Indeed, this is due to the density of regular fields in the functional space associated with the penalized problem (see [12, 16]). In practice, however, the choice of the penalization parameter is not obvious. Notice that for the unpenalized problem the density result fails: this is precisely the underlying cause of the failure of a direct use of nodal elements.

2. Classical and regularized formulations. Let $\Omega \subset \mathbb{R}^2$ be a bounded, simply connected, nonconvex polygonal domain with boundary Γ. We denote by $\boldsymbol{n} = (n_1, n_2)$ the unit normal vector to Γ. We are looking for a numerical approximation of the solution $\boldsymbol{E} = (E_1, E_2) \in \mathbb{R}^2$ to

$$\mathbf{curl}\,\mathrm{curl}\,\boldsymbol{E} = \boldsymbol{J} \ \text{ in } \Omega, \tag{1}$$

$$\mathrm{div}\,\boldsymbol{E} = 0 \ \text{ in } \Omega, \tag{2}$$

$$\boldsymbol{E} \times \boldsymbol{n} = E_1 n_2 - E_2 n_1 = 0 \ \text{ on } \Gamma, \tag{3}$$

where the datum $\boldsymbol{J}$ is assumed divergence-free:

$$\mathrm{div}\,\boldsymbol{J} = 0 \ \text{ in } \Omega. \tag{4}$$

The notations curl and **curl** distinguish between the scalar and vector curl operators:

$$\mathrm{curl}\,\boldsymbol{E} = \frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2} \quad \text{and} \quad \mathbf{curl}\,\varphi = \left(\frac{\partial \varphi}{\partial x_2}, -\frac{\partial \varphi}{\partial x_1}\right).$$

Let us mention that (1)–(3) does not define a "real vector problem" since it is known to reduce to a scalar problem. Indeed, the assumption (4) amounts to saying that $\boldsymbol{J} = \mathbf{curl}\, f$ for some scalar potential f (which can be chosen such that $\int_\Omega f = 0$; see, e.g., [20]). Hence the solution to (1)–(3) is nothing but $\boldsymbol{E} = \mathbf{curl}\,\psi$, where ψ satisfies

$$\begin{aligned} -\Delta\psi &= f \ \text{ in } \Omega, \\ \frac{\partial \psi}{\partial \boldsymbol{n}} &= 0 \ \text{ on } \Gamma \end{aligned}$$

since $\mathrm{curl}\,\mathbf{curl} = -\Delta$. Nevertheless, the vector formulation (1)–(3) here plays the role of a model problem whose interest lies in its simplicity rather than in its physical significance. The reader interested in the full Maxwell equations may replace operator $\mathbf{curl}\,\mathrm{curl}$ by $\mathbf{curl}\,\mathrm{curl} - k^2$ in (1) (with k^2 real or complex). Except for well-posedness results, the same method applies. From a numerical point of view, the only difference induced by this perturbation is the coupling between the regular and singular parts of the field which appears in the OSFM variant (see section 4).

The SFM is based on the fact that the solution to (1)–(3) can be found by solving an equivalent *regularized* problem involving the vector Laplace operator $\boldsymbol{\Delta}$ instead of $\mathbf{curl}\,\mathrm{curl}$ and a divergence-free boundary condition instead of the volume condition (2). For the time being, we write the regularized problem in an imprecise form:

$$-\boldsymbol{\Delta}\boldsymbol{E} = \boldsymbol{J} \ \text{ in } \Omega, \tag{5}$$

$$\mathrm{div}\,\boldsymbol{E} = 0 \ \text{ on } \Gamma, \tag{6}$$

$$\boldsymbol{E} \times \boldsymbol{n} = 0 \ \text{ on } \Gamma. \tag{7}$$

Formally, a solution to (1)–(3) satisfies (5)–(7) (since $-\boldsymbol{\Delta} = \mathbf{curl}\,\mathrm{curl} - \mathbf{grad}\,\mathrm{div}$), and, conversely, a solution to (5)–(7) is divergence-free (which shows that it satisfies (1)–(3)). Indeed, by denoting $\varphi = \mathrm{div}\,\boldsymbol{E}$, we deduce from (5) and (6) that

$$\begin{aligned} -\Delta\varphi &= 0 \ \text{ in } \Omega, \\ \varphi &= 0 \ \text{ on } \Gamma, \end{aligned}$$

which yields $\varphi \equiv 0$ provided φ is regular enough.

In order to clarify the real meaning of the equivalence between the initial and regularized problems, we have to make precise the functional framework which leads to their variational interpretations. Following the notations used in [8], we introduce the spaces

$$\begin{aligned}
\mathcal{H}_N(\mathrm{curl}) &= \left\{ \boldsymbol{E} \in L^2(\Omega)^2 \;\middle|\; \mathrm{curl}\, \boldsymbol{E} \in L^2(\Omega) \text{ and } (\boldsymbol{E} \times \boldsymbol{n})_{|\Gamma} = 0 \right\}, \\
\mathcal{H}_N(\mathrm{curl}, \mathrm{div}) &= \left\{ \boldsymbol{E} \in \mathcal{H}_N(\mathrm{curl}) \;\middle|\; \mathrm{div}\, \boldsymbol{E} \in L^2(\Omega) \right\}, \\
\mathcal{H}_N(\mathrm{curl}, \mathrm{div}\, 0) &= \left\{ \boldsymbol{E} \in \mathcal{H}_N(\mathrm{curl}) \;\middle|\; \mathrm{div}\, \boldsymbol{E} = 0 \text{ in } \Omega \right\}, \\
\mathcal{H}_N(\mathrm{grad}) &= \left\{ \boldsymbol{E} \in H^1(\Omega)^2 \;\middle|\; (\boldsymbol{E} \times \boldsymbol{n})_{|\Gamma} = 0 \right\}
\end{aligned}$$

(where the index N indicates that the fields are normal to Γ). We denote by $(\cdot\,,\cdot)$ the usual scalar product in $L^2(\Omega)^2$. Notice that the sesquilinear form

$$a(\boldsymbol{E}, \boldsymbol{E}') = (\mathrm{curl}\, \boldsymbol{E}, \mathrm{curl}\, \boldsymbol{E}') + (\mathrm{div}\, \boldsymbol{E}, \mathrm{div}\, \boldsymbol{E}')$$

is continuous and coercive in the last three spaces: for each of them, it defines actually a scalar product whose associated norm is equivalent to the graph norm. For $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$, this derives from the compactness of the embedding of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$ in $L^2(\Omega)^2$ (see, e.g., [28]). The case of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div}\, 0)$ follows since it is a closed subspace of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$. (Note that $a(\cdot\,,\cdot) = (\mathrm{curl}\,\cdot\,, \mathrm{curl}\,\cdot)$ in this case.) Finally, for $\mathcal{H}_N(\mathrm{grad})$, the result is not obvious (see Costabel [13]): $\mathcal{H}_N(\mathrm{grad})$ appears as a closed subspace of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$ with finite codimension (see section 3).

In what follows, we denote by $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$, $\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$, and $\mathcal{P}_N(\mathrm{grad})$ the following problems:

$$(\mathcal{P}_N(\cdots)) \qquad \begin{array}{l} \text{Find } \boldsymbol{E} \in \mathcal{H}_N(\cdots) \text{ such that} \\ a(\boldsymbol{E}, \boldsymbol{E}') = (\boldsymbol{J}, \boldsymbol{E}') \quad \forall \boldsymbol{E}' \in \mathcal{H}_N(\cdots) \end{array}$$

for a given $\boldsymbol{J} \in L^2(\Omega)^2$. The coerciveness of $a(\cdot\,,\cdot)$ shows that these problems are well posed. (The assumption (4) is necessary only for $\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$.) The question is to understand the role of each of them.

$\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$ is the variational interpretation of our initial problem (1)–(3): it is easy to see that a field $\boldsymbol{E} \in \mathcal{H}_N(\mathrm{curl}, \mathrm{div}\, 0)$ satisfies (1) in the sense of distributions (in $\mathcal{D}'(\Omega)^2$) if and only if it is a solution to $\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$ (see [24]).

$\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$ is actually the "good" interpretation which makes the *regularized* problem (5)–(7) equivalent to (1)–(2): its solution coincide with that to $\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$, provided $\mathrm{div}\, \boldsymbol{J} = 0$. Indeed, if $\boldsymbol{E} \in \mathcal{H}_N(\mathrm{curl}, \mathrm{div}\, 0)$ satisfies (1), then $(\mathrm{curl}\, \boldsymbol{E}, \mathrm{curl}\, \boldsymbol{E}') = (\boldsymbol{J}, \boldsymbol{E}')$ for every $\boldsymbol{E}' \in \mathcal{H}_N(\mathrm{curl})$; hence it is also a solution to $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$.

$\mathcal{P}_N(\mathrm{grad})$ is a sort of "spurious" interpretation of the regularized problem, where the boundary condition (6) has to be understood in a "weak" sense (see [8]). In general, its solution differs from that to $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$, for $\mathcal{H}_N(\mathrm{grad})$ is strictly contained in $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$.

3. Decomposition of $\boldsymbol{\mathcal{H}_N(\mathrm{curl}, \mathrm{div})}$. With regard to the numerical approximation of $\mathcal{P}_N(\mathrm{curl}, \mathrm{div}\, 0)$, the idea of solving its regularized formulation $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$ seems attractive since it involves the elliptic operator $-\boldsymbol{\Delta}$ and avoids taking into account the divergence-free condition in Ω. However, one has to be careful: a H^1-conforming finite element discretization can provide only an approximation of the "spurious" solution to $\mathcal{P}_N(\mathrm{grad})$ but not of the "physical" solution. The SFM consists of adding to $\mathcal{H}_N(\mathrm{grad})$ a complementary subspace $\mathcal{H}_{\mathrm{sing}}$ such that

$$\mathcal{H}_N(\mathrm{curl}, \mathrm{div}) = \mathcal{H}_N(\mathrm{grad}) \oplus \mathcal{H}_{\mathrm{sing}}. \tag{8}$$

In order to understand how to construct such a *singular subspace*, let us first recall some classical results about the *singularities* of the (scalar) Laplace operator in a nonconvex polygon (see Grisvard [22]). We denote by $\boldsymbol{x}_\ell$, for $\ell = 1$ to L, the vertices of the polygon Ω, numbered according to the positive orientation. At each vertex, we define local polar coordinates (r_ℓ, θ_ℓ), where $r_\ell(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_\ell\|$, $\theta_\ell(\boldsymbol{x}_{\ell+1}) = 0$, and $\theta_\ell(\boldsymbol{x}_{\ell-1}) = \omega_\ell > 0$. (Here we denote $\boldsymbol{x}_{L+1} = \boldsymbol{x}_1$ and $\boldsymbol{x}_0 = \boldsymbol{x}_L$.) For every ℓ, consider a regular cut-off function $\eta_\ell = \eta_\ell(r_\ell)$ such that $\eta_\ell \equiv 1$ near $\boldsymbol{x}_\ell$ and η_ℓ vanishes if $r_\ell > c_\ell$ with $0 < c_\ell < \min_{\ell' \neq \ell} \|\boldsymbol{x}_{\ell'} - \boldsymbol{x}_\ell\|$. For the reentrant corners, i.e., for ℓ such that $\omega_\ell > \pi$, let us finally introduce the *singular functions*

$$s_\ell(r_\ell, \theta_\ell) = r_\ell^{\pi/\omega_\ell} \sin\left(\frac{\pi}{\omega_\ell}\,\theta_\ell\right). \tag{9}$$

Function s_ℓ satisfies $\Delta s_\ell = 0$ in the sector $\{(r, \theta_\ell) \mid r > 0, \theta_\ell \in]0, \omega_\ell[\}$, vanishes on its boundary, and belongs to $H^1(\Omega) \setminus H^2(\Omega)$. Hence, the truncated singular function $\eta_\ell s_\ell$ belongs to $H_0^1(\Omega) \setminus H^2(\Omega)$ and satisfies $\Delta(\eta_\ell s_\ell) = 0$ in a neighborhood of the reentrant corner.

Then consider the subspace $\mathcal{S}$ of $H_0^1(\Omega)$ defined by

$$\mathcal{S} = \operatorname{span}\{\eta_\ell s_\ell \mid \omega_\ell > \pi\}. \tag{10}$$

This space, whose dimension is exactly the number of "reentrant corners,"gives us the first level of singularity of the solution $\varphi \in H_0^1(\Omega)$ of the variational equation

$$(\mathbf{grad}\,\varphi, \mathbf{grad}\,\psi) = (f, \psi) \quad \forall \psi \in H_0^1(\Omega)$$

for a given $f \in L^2(\Omega)$. Indeed, φ can be uniquely decomposed as

$$\varphi = \varphi_{\text{reg}} + \varphi_{\text{sing}}, \quad \text{where} \quad \varphi_{\text{reg}} \in H_0^1(\Omega) \cap H^2(\Omega) \text{ and } \varphi_{\text{sing}} \in \mathcal{S}.$$

The space $\mathcal{S}$ offers a first possible choice for the singular subspace $\mathcal{H}_{\text{sing}}$ in (8). Indeed, notice that $\mathbf{grad}\,\mathcal{S} \subset \mathcal{H}_N(\text{curl}, \text{div})$ since $\text{curl}\,\mathbf{grad}\,\varphi = 0$, $(\mathbf{grad}\,\varphi \times \boldsymbol{n})_{|\Gamma} = 0$ for all $\varphi \in H_0^1(\Omega)$ and $\text{div}(\mathbf{grad}(\eta_\ell s_\ell)) = \Delta(\eta_\ell s_\ell)$ is a regular function in Ω. As a consequence, we infer that

$$\mathbf{grad}\,\mathcal{S} \subset \mathcal{H}_N(\text{curl}, \text{div}) \setminus \mathcal{H}_N(\text{grad}).$$

The point is that $\mathbf{grad}\,\mathcal{S}$ describes all the possible singularities of fields in the space $\mathcal{H}_N(\text{curl}, \text{div})$, which is expressed by the following direct decomposition of type (8) (see [5] or [8] for the proof):

$$\mathcal{H}_N(\text{curl}, \text{div}) = \mathcal{H}_N(\text{grad}) \oplus \mathbf{grad}\,\mathcal{S}. \tag{11}$$

This sum is not orthogonal, but it can be used to construct an orthogonal complementary subspace of $\mathcal{H}_N(\text{grad})$. Indeed, for each ℓ such that $\omega_\ell > \pi$, the field

$$\boldsymbol{E}_\ell = \mathbf{grad}(\eta_\ell s_\ell) + \boldsymbol{G}_\ell \quad \text{with} \quad \boldsymbol{G}_\ell \in \mathcal{H}_N(\text{grad}) \tag{12}$$

is orthogonal to $\mathcal{H}_N(\text{grad})$ (for the scalar product $a(\cdot\,,\cdot)$) if and only if

$$a(\boldsymbol{G}_\ell, \boldsymbol{E}') = -(\Delta(\eta_\ell s_\ell), \text{div}\,\boldsymbol{E}') \quad \forall \boldsymbol{E}' \in \mathcal{H}_N(\text{grad}),$$

which amounts to solving $\mathcal{P}_N(\mathrm{grad})$ for $\boldsymbol{J} = \mathbf{grad}(\Delta(\eta_\ell s_\ell))$. Following the idea proposed by Moussaoui [27] for the scalar Laplace operator, another way of writing the decomposition (12) is

$$\boldsymbol{E}_\ell = \mathbf{grad}(s_\ell) + \boldsymbol{F}_\ell, \tag{13}$$

where $\boldsymbol{F}_\ell = \boldsymbol{G}_\ell + \mathbf{grad}((\eta_\ell - 1)s_\ell)$. Noticing that $\Delta s_\ell = 0$, we see that $\boldsymbol{F}_\ell$ is the solution to the following inhomogeneous variational problem (similar to $\mathcal{P}_N(\mathrm{grad})$) :

$$\begin{array}{l} \text{Find } \boldsymbol{F}_\ell \in H^1(\Omega)^2 \text{ such that} \\ a(\boldsymbol{F}_\ell, \boldsymbol{E}') = 0 \quad \forall \boldsymbol{E}' \in \mathcal{H}_N(\mathrm{grad}), \quad \text{and} \\ \boldsymbol{F}_\ell \times \boldsymbol{n} = -\,\mathbf{grad}\, s_\ell \times \boldsymbol{n} \ \text{ on } \Gamma. \end{array} \tag{14}$$

Hence the decomposition (8) becomes orthogonal if we choose

$$\mathcal{H}_{\mathrm{sing}} = \mathrm{span}\,\{\boldsymbol{E}_\ell \mid \omega_\ell > \pi\}\,. \tag{15}$$

4. Numerical implementation. We shall illustrate the numerical implementation of the SFM in the case of a simple model domain. To this end, let us consider a domain Ω which has only one reentrant corner situated at the origin $\mathbf{0}$. We thus can omit the subscript ℓ in the notation of the angle ω, the local polar coordinates (r, θ), the singular function s, and the cut-off function η introduced in the previous section. We shall indicate the complications that arise from the numerical implementation for more than one reentrant corner.

Let $\mathcal{T}_h, 0 < h \leq h_0$ be a family of regular triangulations of the domain Ω; that means that there exist two constants $c_0, c_1 > 0$ such that every triangle $T_h \in \mathcal{T}_h$ contains a circular disc with radius $c_0 h$ and is contained within another disc with radius $c_1 h$. The associated discretization space of finite elements of type P1 is given by

$$Y_h = \left\{ \boldsymbol{E}_h \in H^1(\Omega)^2 \;\middle|\; \boldsymbol{E}_{h|T_h} \text{ is affine } \forall\, T_h \in \mathcal{T}_h \right\}. \tag{16}$$

Let $\{M_I\}_{I=1,\ldots,\dim Y_h}$ be the set of nodal points of the triangulation. Then

$$V_h = \{\boldsymbol{E}_h \in Y_h \mid \text{“}(\boldsymbol{E}_h \times \boldsymbol{n})(M_I) = 0\text{”} \ \forall M_I \in \Gamma\} \tag{17}$$

is a finite dimensional subspace of $\mathcal{H}_N(\mathrm{grad})$, and we denote $(\boldsymbol{w}_I)_{I=1,\ldots,N_h}$ its basis. Note that the discrete boundary condition “$(\boldsymbol{E}_h \times \boldsymbol{n})(M_I) = 0$” is ambiguous at a vertex of the polygon: the use of the quotes means that here $\boldsymbol{E}_h(M_I) = 0$.

As mentioned in [8], the idea of the SFM consists of discretizing the regular part of the solution by means of nodal finite elements, whereas the singular part is taken into account explicitly. The question is now, Which choice of the complementary space $\mathcal{H}_{\mathrm{sing}}$ is the best? In what follows, we shall illustrate the approach for the two spaces introduced in section 3.

4.1. The SFM. The nonorthogonal decomposition (11) of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$ naturally leads to the following discretization space:

$$X_h = V_h \oplus \mathbf{grad}\,\mathcal{S}, \tag{18}$$

where $\mathcal{S} = \mathrm{span}\{\eta s\}$ with the notations introduced at the beginning of this section. The corresponding discrete problem then reads as follows:

$$\begin{array}{l} \text{Find } \boldsymbol{E}_h \in X_h \ \text{ such that} \\ a(\boldsymbol{E}_h, \boldsymbol{E}'_h) = (\boldsymbol{J}, \boldsymbol{E}'_h) \quad \forall \boldsymbol{E}'_h \in X_h. \end{array} \tag{19}$$

It is clear that X_h is of dimension N_h+1, and (19) may be written in an equivalent manner in matrix form:

$$\begin{pmatrix} A_{\mathrm{FE}} & C \\ C^T & a_s \end{pmatrix} \begin{pmatrix} E_R \\ \alpha_h \end{pmatrix} = \begin{pmatrix} J_{\mathrm{FE}} \\ j_s \end{pmatrix}, \tag{20}$$

where

- A_{FE} and J_{FE}, respectively, denote the stiffness matrix and the right-hand side corresponding to the space of finite elements, V_h,

$$A_{\mathrm{FE}} = (a(\boldsymbol{w}_J, \boldsymbol{w}_I))_{I,J=1,\dots,N_h}, \text{ and } J_{\mathrm{FE}} = ((\boldsymbol{J}, \boldsymbol{w}_I))^T_{I=1,\dots,N_h}; \tag{21}$$

- a_s and j_s denote the matrix and the right-hand side of order 1 corresponding to the singular field,

$$a_s = a(\mathbf{grad}(\eta s), \mathbf{grad}(\eta s)) \text{ and } j_s = (\boldsymbol{J}, \mathbf{grad}(\eta s)); \tag{22}$$

- C is finally a matrix of order $N_h \times 1$ coupling the basis functions of FE-type and the singular field,

$$C = (a(\mathbf{grad}(\eta s), \boldsymbol{w}_I))^T_{I=1,\dots,N_h}. \tag{23}$$

The decomposition (18) appears in the block structure of the stiffness matrix, and we may rewrite (20) in order to solve two subproblems involving only the standard FE-matrix A_{FE}. The algorithm of the SFM then consists of the following three steps:

1. Solve two linear systems

$$\begin{cases} A_{\mathrm{FE}} E_{\mathrm{FE}} = J_{\mathrm{FE}}, \\ A_{\mathrm{FE}} S = C. \end{cases} \tag{24}$$

2. Calculate the approximate singular coefficient α_h through the identity

$$\alpha_h = \frac{j_s - C^T E_{\mathrm{FE}}}{a_s - C^T S}. \tag{25}$$

3. Finally, set

$$\boldsymbol{E}_h = \sum_{I=1}^{N_h} (E_{\mathrm{FE},I} - \alpha_h S_I) \boldsymbol{w}_I + \alpha_h \, \mathbf{grad}(\eta s) \tag{26}$$

in order to obtain the approximate solution.

Remark 4.1. In the case of more than one reentrant corner, say N_s, the coefficient a_s becomes a $N_s \times N_s$ matrix A_s whose generic term is simply given by

$$(A_s)_{\ell,\ell'} = a(\mathbf{grad}(\eta_\ell s_\ell), \mathbf{grad}(\eta_{\ell'} s_{\ell'})).$$

Notice that this is a diagonal matrix since the supports of the different cut-off functions η_ℓ do not intersect. Similarly, j_s becomes a $N_s \times 1$ matrix, denoted J_s, and the coupling matrix $C = (C_1 \dots C_{N_s})$ is of order $N_h \times N_s$. The adaptation of the algorithm described above then reads as follows: in step 1, we have to solve $N_s + 1$ linear systems of order N_h involving the same *sparse* matrix A_{FE},

$$\begin{cases} A_{\mathrm{FE}} E_{\mathrm{FE}} = J_{\mathrm{FE}}, \\ A_{\mathrm{FE}} S_l = C_l, \ l \in \{1, \dots, N_s\}. \end{cases} \tag{27}$$

In step 2, the computation of the N_s singular coefficients $\alpha_{h,l}$, may be achieved by solving the linear system of order N_s,

$$(A_s - C^T S)\alpha_h = J_s - C^T E_{\mathrm{FE}}, \tag{28}$$

where $\alpha_h = (\alpha_{h,1} \dots \alpha_{h,N_s})^T$ and $S = (S_1 \dots S_{N_s}) \in \mathbb{R}^{N_h \times N_s}$. The formula corresponding to (26) is straightforward.

4.2. The OSFM. Let us notice that the SFM presented in the previous section depends on the choice of the cut-off function η. It is well known from the implementation of *singular function methods* that the use of a cut-off function may cause numerical difficulties since its derivatives take very high values. Therefore we consider here a singular subspace, $\mathcal{H}_{\mathrm{sing}}$, that does not use any cut-off function. Indeed, it has been shown in section 3 that this may be done by choosing the following orthogonal decomposition of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$:

$$\mathcal{H}_N(\mathrm{curl}, \mathrm{div}) = \mathcal{H}_N(\mathrm{grad}) \oplus \mathrm{span}\{\mathbf{grad}(s) + \boldsymbol{F}\}, \tag{29}$$

where $\boldsymbol{F}$ is the solution to the inhomogeneous problem (14).

In order to use (29) for an approximation of the solution to $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$, we thus have to discretize separately $\mathcal{H}_N(\mathrm{grad})$ and $\mathrm{span}\{\mathbf{grad}(s) + \boldsymbol{F}\}$. With regard to $\mathcal{H}_N(\mathrm{grad})$, this will be done as before by means of nodal finite elements. However, contrary to the SFM, the singular field $\mathbf{grad}(s) + \boldsymbol{F}$ is not known exactly because of the contribution of $\boldsymbol{F}$. We are thus led to introduce an approximate singular field

$$\mathbf{grad}(s) + \boldsymbol{F}_h, \tag{30}$$

where $\boldsymbol{F}_h$ is the FE-approximation of $\boldsymbol{F}$, that is, the solution to the following problem:

$$\begin{array}{l} \text{Find } \boldsymbol{F}_h \in Y_h \ \text{ such that} \\ a(\boldsymbol{F}_h, \boldsymbol{E}'_h) = 0 \quad \forall \boldsymbol{E}'_h \in V_h, \ \text{and} \\ \text{“}(\boldsymbol{F}_h \times \boldsymbol{n})(M_J) = -(\mathbf{grad}(s) \times \boldsymbol{n})(M_J)\text{”} \quad \forall M_J \in \Gamma. \end{array} \tag{31}$$

If $\widetilde{X_h}$ denotes the finite dimensional subspace of $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$ given by

$$\widetilde{X_h} = V_h \oplus \mathrm{span}\{\mathbf{grad}(s) + \boldsymbol{F}_h\}, \tag{32}$$

the discrete formulation of $\mathcal{P}_N(\mathrm{curl}, \mathrm{div})$ reads as follows:

$$\begin{array}{l} \text{Find } \widetilde{\boldsymbol{E}_h} \in \widetilde{X_h} \ \text{ such that} \\ a(\widetilde{\boldsymbol{E}_h}, \boldsymbol{E}'_h) = (\boldsymbol{J}, \boldsymbol{E}'_h) \quad \forall \boldsymbol{E}'_h \in \widetilde{X_h}, \end{array} \tag{33}$$

or, equivalently, in matrix form

$$\begin{pmatrix} A_{\mathrm{FE}} & 0 \\ 0 & \tilde{a}_s \end{pmatrix} \begin{pmatrix} E_{\mathrm{FE}} \\ \tilde{\alpha}_h \end{pmatrix} = \begin{pmatrix} J_{\mathrm{FE}} \\ \tilde{\jmath}_s \end{pmatrix}. \tag{34}$$

As before, A_{FE} and J_{FE} denote the stiffness matrix and the right-hand side of the FE-discretization, and $\tilde{a}_s$ and $\tilde{\jmath}_s$ are, respectively, given by

$$\tilde{a}_s = a(\boldsymbol{F}_h, \boldsymbol{F}_h) \ \text{ and } \tilde{\jmath}_s = (\boldsymbol{J}, \mathbf{grad}(s) + \boldsymbol{F}_h) \tag{35}$$

since $\mathbf{grad}(s)$ is curl- and divergence-free. Note as well that no coupling terms occur in (34) because of the discrete orthogonality relation

$$a(\mathbf{grad}(s) + \boldsymbol{F}_h, \boldsymbol{E}'_h) = 0 \quad \forall \boldsymbol{E}'_h \in V_h. \tag{36}$$

The algorithm of the OSFM then reads as follows:

1. Solve (31) and the linear system
$$A_{\mathrm{FE}} E_{\mathrm{FE}} = J_{\mathrm{FE}}.$$
2. Calculate the approximate singular coefficient $\tilde{\alpha}_h$ through the identity
$$\tilde{\alpha}_h = \frac{\tilde{j}_s}{\tilde{a}_s}.$$
3. Set
$$\widetilde{\boldsymbol{E}_h} = \sum_{I=1}^{N_h} E_{\mathrm{FE},I}\, \boldsymbol{w}_I + \tilde{\alpha}_h \left(\mathbf{grad}(s) + \boldsymbol{F}_h\right).$$

Remark 4.2. As for the SFM, the algorithm can be easily adapted for $N_s > 1$ reentrant corners. Step 1 will consist of solving $N_s + 1$ linear systems (one for E_{FE} and one for each problem of type (31) corresponding to each singular function s_ℓ). The determination of the N_s singular coefficients in step 2 will be achieved by solving a linear system of order N_s which involves a full matrix $\widetilde{A_s}$ describing the coupling between the N_s discrete singular fields. Finally, note that for more involved situations (such as the full Maxwell equations), the inhomogeneous problem of type (31) has to be replaced by the following problem:

$$\begin{array}{l} \text{Find } \boldsymbol{F}_{h,\ell} \in Y_h \ \text{ such that} \\ a(\boldsymbol{F}_h, \boldsymbol{E}'_h) = -a(\mathbf{grad}(s_\ell), \boldsymbol{E}'_h) \quad \forall \boldsymbol{E}'_h \in V_h, \text{ and} \\ \text{“}(\boldsymbol{F}_{h,\ell} \times \boldsymbol{n})(M_J) = -(\mathbf{grad}(s_\ell) \times \boldsymbol{n})(M_J)\text{”} \quad \forall M_J \in \Gamma, \end{array} \tag{37}$$

introducing a coupling of singular fields and basis functions of FE-type.

4.3. Computational cost. In this section we will compare the computational cost of both SFM and OSFM in the general case of the full Maxwell equations and multiple corners.

Notice that in both methods we have to solve $N_s + 1$ linear systems of order N_h involving the standard (sparse) FE-matrix and *one* linear system of order $N_s \ll N_h$ involving a full symmetric matrix. From this point of view, SFM and OSFM thus have the same complexity.

What about the computational cost and the implementation of the respective matrices and right-hand sides? Besides the standard FE-matrix A_{FE} and the right-hand side J_{FE}, which may be calculated using a standard finite element code, there are three kinds of terms:

- pure FE-terms obtained by matrix-vector or vector-vector products of order N_h (such as $C^T S$ in the SFM and $a(\boldsymbol{F}_{h,\ell}, \boldsymbol{F}_{h,\ell'}) = F_\ell^T A_{\mathrm{FE}} F_{\ell'}$ in the OSFM),
- coupling terms of order N_h which need high order quadrature rules at least near the corners, and
- singular terms which may be calculated analytically or again by means of high order quadrature rules.

Notice that coupling and singular terms in OSFM occur only for the full Maxwell equations and are reduced to L^2-scalar products (i.e., containing no derivatives of the singular fields) since the function s_ℓ is harmonic for all ℓ, and $\mathbf{grad}(s_\ell)$ is thus curl- and divergence-free in Ω.

Summing up, the number of FE-terms, singular and coupling terms is of order $\mathcal{O}(N_s)$ in SFM and of order $\mathcal{O}(N_s^2)$ for OSFM. However, the number of reentrant corners is not significant compared to the complexity of the mesh, and thus there is no real difference between the computational complexity of the two methods.

5. Error analysis. In this section, we shall investigate error estimates for the respective solutions obtained by the SFM and the OSFM based on finite elements of type P1. The generalization of the results to higher order elements is straightforward. We study the global error in the energy norm and in the L^2-norm as well as the error on the singularity coefficient. It turns out that the SFM and the OSFM are of the same order.

PROPOSITION 5.1 (energy norm estimates for the SFM). *Let $\boldsymbol{E}$ and $\boldsymbol{E}_h$ be the respective solutions of $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ and* (19)*. There exists a constant $C > 0$ such that*

$$||\boldsymbol{E}-\boldsymbol{E}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq C \inf_{\boldsymbol{E'}_h \in V_h} ||\boldsymbol{E}_{\mathrm{reg}}-\boldsymbol{E'}_h||_{1,\Omega}\,, \tag{38}$$

where $\boldsymbol{E}_{\mathrm{reg}}$ denotes the regular part of $\boldsymbol{E}$ corresponding to the decomposition (11)*. If in addition $\boldsymbol{E}_{\mathrm{reg}}$ belongs to $H^{s+1}(\Omega)^2$ with $s > 0$, there is a constant $C' > 0$ such that*

$$||\boldsymbol{E}-\boldsymbol{E}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq C'\, h^{\min(s,1)}\, ||\boldsymbol{E}_{\mathrm{reg}}||_{s+1,\Omega}\,. \tag{39}$$

Remark 5.2. Estimate (38) implies that the SFM will converge but does not give any convergence order. Nevertheless, the regular part $\boldsymbol{E}_{\mathrm{reg}}$ of the solution to $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ always belongs to $H^{s+1}(\Omega)^2$ for all $s < \frac{2\pi}{\omega} - 1$ (see [15] and Lemma 5.5). Hence, (39) implies that the SFM is at least of order $\frac{2\pi}{\omega} - 1 - \varepsilon$ (with $\varepsilon > 0$ arbitrarily small).

Proof of Proposition 5.1. Since the bilinear form $a(\cdot,\cdot)$ is continuous and coercive on $\mathcal{H}_N(\mathrm{curl},\mathrm{div})$, Cea's lemma implies

$$||\boldsymbol{E}-\boldsymbol{E}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq C \inf_{\boldsymbol{E}'_h \in X_h} ||\boldsymbol{E}-\boldsymbol{E}'_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}\,, \tag{40}$$

where $C > 0$ depends only on the continuity and coercivity constants of $a(\cdot,\cdot)$. Decomposition (11) implies that there is $\boldsymbol{E}_{\mathrm{reg}} \in \mathcal{H}_N(\mathrm{grad})$ and $\alpha \in \mathbb{R}$ such that

$$\boldsymbol{E} = \boldsymbol{E}_{\mathrm{reg}} + \alpha\, \mathbf{grad}(\eta s).$$

Now note that for any field $\boldsymbol{G}_h$ belonging to the FE-space V_h the field

$$\boldsymbol{E'}_h = \boldsymbol{G}_h + \alpha\, \mathbf{grad}(\eta s)$$

belongs to X_h, and hence

$$\inf_{\boldsymbol{E}'_h \in X_h} ||\boldsymbol{E}-\boldsymbol{E}'_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq \inf_{\boldsymbol{G}_h \in V_h} ||\boldsymbol{E}_{\mathrm{reg}}-\boldsymbol{G}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}\,,$$

and we obtain (38) since, on $\mathcal{H}_N(\mathrm{grad})$, the norms $||\cdot||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}$ and $||\cdot||_{1,\Omega}$ are equivalent.

If, in addition, the regular part $\boldsymbol{E}_{\mathrm{reg}}$ of the solution belongs to $H^{s+1}(\Omega)^2$ with $s > 0$, the standard error analysis of the FE-method yields (39) (see [10, 21]). ∎

Remark 5.3. Note that the error on the singular coefficient of the SFM is governed by the same term as the global error, that is, the interpolation error of the finite element method. Indeed, the decomposition of $\mathcal{H}_N(\mathrm{curl},\mathrm{div})$ into the *direct* sum of $\mathcal{H}_N(\mathrm{grad})$ and $\mathcal{H}_{\mathrm{sing}}$ yields

$$|\alpha - \alpha_h| \leq C\, ||\boldsymbol{E}-\boldsymbol{E}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}$$

and thus estimates similar to (38) and (39) for the error on the singular coefficient. Hence, there is a fundamental difference compared to *singular function methods* known to *improve* the convergence order of a nodal finite element method: these methods are of order 1 (for finite elements of type P1) with regard to the global error in the energy norm but of order less than 1 (depending on the angle of the reentrant corner) with regard to the convergence of the singular coefficient (see, e.g., [7]). In our context the fact that $\mathcal{H}_N(\mathrm{grad})$ is a *closed* subspace of $\mathcal{H}_N(\mathrm{curl},\mathrm{div})$ forbids any nodal discretization to approach the singular fields. In other words, the "angle" between $\mathbf{grad}\,\mathcal{S}$ and V_h (see (18)) cannot shrink to zero.

The next proposition gives error estimates in the L^2-norm.

PROPOSITION 5.4 (L^2-estimates for the SFM). *Let $\boldsymbol{E}$ and $\boldsymbol{E}_h$ be the respective solutions of $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ and (19), and let $\boldsymbol{E}_{\mathrm{reg}}$ denote the regular part of $\boldsymbol{E}$ corresponding to the decomposition (11). Assume that $\boldsymbol{E}_{\mathrm{reg}}$ belongs to $H^{s+1}(\Omega)^2$, where $s>0$. There is a constant $C>0$ such that*

$$||\boldsymbol{E}-\boldsymbol{E}_h||_{0,\Omega} \le C\,h^{\lambda}\,||\boldsymbol{E}_{\mathrm{reg}}||_{s+1,\Omega} \quad \forall \lambda < \min(s,1)+\frac{2\pi}{\omega}-1. \tag{41}$$

Proof. In order to get the L^2-error estimates for the SFM, we apply the Aubin–Nitsche trick which yields

$$\begin{aligned}&||\boldsymbol{E}-\boldsymbol{E}_h||_{0,\Omega}\\ &\le C\,||\boldsymbol{E}-\boldsymbol{E}_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}\left(\sup_{\boldsymbol{K}\in L^2(\Omega)^2}\left\{\frac{1}{||\boldsymbol{K}||_{0,\Omega}}\inf_{\boldsymbol{E}'_h\in X_h}||\boldsymbol{E}_K-\boldsymbol{E}'_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})}\right\}\right),\end{aligned}$$

where $\boldsymbol{E}_K\in\mathcal{H}_N(\mathrm{curl},\mathrm{div})$ denotes the unique solution to $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ corresponding to a datum $\boldsymbol{K}$ in $L^2(\Omega)^2$. According to Proposition 5.1, the first factor in the preceding inequality can be bounded by

$$C\,h^{\min(s,1)}\,||\boldsymbol{E}_{\mathrm{reg}}||_{s+1,\Omega}.$$

In order to get an estimate for the second factor, we consider the regular part of $\boldsymbol{E}_K$ which belongs to $H^{t+1}(\Omega)^2$ for all $t<\frac{2\pi}{\omega}-1$ (see [15] and Lemma 5.5 below). We thus infer once more from Proposition 5.1 that

$$\inf_{\boldsymbol{E}'_h\in X_h}||\boldsymbol{E}_K-\boldsymbol{E}'_h||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \le C\,h^t\,||\boldsymbol{E}_{K,\mathrm{reg}}||_{t+1,\Omega} \quad \forall t<\frac{2\pi}{\omega}-1.$$

Now notice that $||\boldsymbol{E}_{K,\mathrm{reg}}||_{t+1,\Omega}\le C\,||\boldsymbol{K}||_{0,\Omega}$ according to the a priori inequality (42) of Lemma 5.5. This completes the proof. ∎

For the sake of completeness, we now prove the a priori estimate that was used in the proof of Proposition 5.4.

LEMMA 5.5. *Let $\boldsymbol{E}$ be the solution to $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ with a right-hand side $\boldsymbol{J}$ in $L^2(\Omega)^2$. Let $\boldsymbol{E}_{\mathrm{reg}}$ denote the regular part of $\boldsymbol{E}$ according to decomposition (11). Then $\boldsymbol{E}_{\mathrm{reg}}\in H^{t+1}(\Omega)^2$ for all $t<\frac{2\pi}{\omega}-1$, and there exists a constant $C>0$ such that the following a priori estimate holds:*

$$||\boldsymbol{E}_{\mathrm{reg}}||_{t+1,\Omega}\le C\,||\boldsymbol{J}||_{0,\Omega}. \tag{42}$$

Proof. According to (11) we get the following decomposition of $\boldsymbol{E}$:

$$\boldsymbol{E}=\boldsymbol{E}_{\mathrm{reg}}+\alpha\,\mathbf{grad}(\eta s).$$

The regularity result has been proven in [15]. The definition of $\mathcal{P}_N(\mathrm{curl},\mathrm{div})$ implies

$$a(\boldsymbol{E}_{\mathrm{reg}},\boldsymbol{E}') = (\boldsymbol{J},\boldsymbol{E}') - \alpha(\Delta(\eta s),\mathrm{div}\,\boldsymbol{E}') \quad \forall \boldsymbol{E}' \in \mathcal{H}_N(\mathrm{grad}). \tag{43}$$

$\boldsymbol{E}_{\mathrm{reg}}$ thus coincides with the solution to $\mathcal{P}_N(\mathrm{grad})$ corresponding to the datum $\widetilde{\boldsymbol{J}} = \boldsymbol{J} + \alpha\,\mathbf{grad}(\Delta(\eta s))$. (Note that no boundary terms occur in the integration by parts of the right-hand side of (43) since $\Delta(\eta s) \in H_0^1(\Omega)$.) We infer that

$$||\boldsymbol{E}_{\mathrm{reg}}||_{t+1,\Omega} \leq C_1(||\boldsymbol{J}||_{0,\Omega} + |\alpha|\,||\Delta(\eta s)||_{1,\Omega}) \tag{44}$$

from the *closed graph theorem*. Indeed, consider the linear operator $\mathbb{T} : L^2(\Omega)^2 \to H^s(\Omega)^2$ defined by $\mathbb{T}\boldsymbol{J} = \boldsymbol{E}$, where $\boldsymbol{E} \in \mathcal{H}_N(\mathrm{grad})$ is the unique solution to $\mathcal{P}_N(\mathrm{grad})$ with right-hand side $\boldsymbol{J}$. Inequality (44) is nothing but the continuity of $\mathbb{T}$ whose graph is a closed subset of $L^2(\Omega)^2 \times H^s(\Omega)^2$.

As in Remark 5.3, we have

$$|\alpha| \leq C_2\,||\boldsymbol{E}||_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq C_3\,||\boldsymbol{J}||_{0,\Omega}$$

and (42) follows with $C = C_1(1 + C_3\,||\Delta(\eta s)||_{1,\Omega})$. ∎

Remark 5.6. The weak point of the SFM is hidden among the above estimates. It is actually related to the cut-off function η involved in the definition of the singular field $\mathbf{grad}(\eta s)$. Such a function introduces artificially high variations of the regular part $\boldsymbol{E}_{\mathrm{reg}}$ in the region where η varies from 0 to 1. Inequality (44) yields a quantified representation of this effect: the right-hand side depends on the H^1-norm of $\Delta(\eta s)$ which involves the third derivatives of the cut-off function η. Hence the constant in the estimate (42) may have a very high value. We shall come back to this problem in the next section.

The error analysis for the OSFM is very similar to the one for the SFM, the main difference lying in an appropriate definition of the interpolation error.

For a given real parameter α let us introduce the space $Y_{h,\alpha}$ of those fields $\boldsymbol{G}_h$ of Y_h which satisfy the inhomogeneous boundary condition "$(\boldsymbol{G}_h \times \boldsymbol{n})(M_I) = -\alpha(\mathbf{grad}(s) \times \boldsymbol{n})(M_I)$" for all $M_I \in \Gamma$. Note that $Y_{h,\alpha}$ is an affine subspace of Y_h of the form $\alpha\boldsymbol{R} + V_h$, where $\boldsymbol{R}$ is some fixed FE-element lifting of the tangential trace of $-\mathbf{grad}(s)$.

One gets the analogous error estimates when substituting $\boldsymbol{E}_{\mathrm{reg}}$ by $\widetilde{\boldsymbol{E}_{\mathrm{reg}}} + \alpha\boldsymbol{F}$, where $\widetilde{\boldsymbol{E}_{\mathrm{reg}}}$ and α denote, respectively, the regular part and the singular coefficient of $\boldsymbol{E}$ corresponding to the decomposition (29) and $\boldsymbol{F}$ is the solution in $H^1(\Omega)^2$ to (14).

PROPOSITION 5.7 (energy norm estimates for the OSFM). *Let $\widetilde{\boldsymbol{E}_h}$ denote the solution of* (31). *There exists a constant $C > 0$ such that*

$$\left|\left|\boldsymbol{E} - \widetilde{\boldsymbol{E}_h}\right|\right|_{\mathcal{H}_N(\mathrm{curl},\mathrm{div})} \leq C \inf_{\boldsymbol{E}'_h \in Y_{h,\alpha}} \left|\left|(\widetilde{\boldsymbol{E}_{\mathrm{reg}}} + \alpha\boldsymbol{F}) - \boldsymbol{E}'_h\right|\right|_{1,\Omega} \tag{45}$$

with the notations introduced just before.

Proof. We use the following characterization of $\widetilde{X_h}$:

$$\begin{aligned} &\boldsymbol{E}_h \in \widetilde{X_h} \text{ if and only if } \exists \boldsymbol{G}_h \in Y_h \text{ and } \alpha \in \mathbb{R} \text{ such that} \\ &\boldsymbol{E}_h = \boldsymbol{G}_h + \alpha\,\mathbf{grad}(s), \text{ and} \\ &\text{“}(\boldsymbol{G}_h \times \boldsymbol{n})(M_I) = -\alpha(\mathbf{grad}(s) \times \boldsymbol{n})(M_I)\text{”} \quad \forall M_I \in \Gamma. \end{aligned} \tag{46}$$

It may be easily seen that (32) and (46) indeed define the same space.

As for the SFM, Cea's lemma leads us to look for an estimation of the interpolation error on $\widetilde{X_h}$. With a given field $\boldsymbol{G}_h$ belonging to $Y_{h,\alpha}$ we associate the field $\boldsymbol{E}'_h =$

$\boldsymbol{G}_h + \alpha\,\mathbf{grad}(s)$, where α is the singular coefficient of $\boldsymbol{E}$. $\boldsymbol{E}'_h$ clearly belongs to $\widetilde{X_h}$ and (45) follows as before. ∎

The estimates corresponding to (39) and (41) follow in the same way.

Remark 5.8. It may easily be seen that the fields $\boldsymbol{E}_{\rm reg}$ and $\widetilde{\boldsymbol{E}_{\rm reg}} + \alpha\boldsymbol{F}$ (corresponding to decompositions (11) and (29), respectively) have the same regularity: they differ by the field $\alpha\,\mathbf{grad}((1-\eta)s)$ which is of class $\mathcal{C}^\infty$. The convergence order of both methods is thus apparently the same.

Remark 5.9. The generalization of the previous error analysis to the case of multiple corners is straightforward. Indeed, the energy norm estimates (38) and (39) keep unchanged. Notice that the minimal regularity of the regular part is now H^{s+1} with

$$s < \min_{1\le\ell\le N_s} \frac{2\pi}{\omega_\ell}.$$

Consequently, the convergence rate of the error in the L^2-norm is $\mathcal{O}(h^\lambda)$ with $\lambda < \min(s,1) + \min_l \frac{2\pi}{\omega_\ell} - 1$.

6. Numerical results. In this section, we present numerical tests of the SFM and the OSFM for several examples where the exact solutions are known. The model domain is formed by the three quarters of a disc with center $\mathbf{0}$ and radius 2; the only reentrant corner is thus of measure $3\pi/2$. Such a "camembert-like"domain is polygonal near the reentrant corner and convex in the neighborhood of any other irregular point of the boundary. The results obtained in the previous sections thus carry over to this curvilinear polygon.

Notice that both methods have been implemented in the case of one reentrant corner only, but the generalization to several corners is straightforward and has been described in section 4 (Remarks 4.1 and 4.2). The additional complexity has been investigated in section 4.3.

We consider two classes of exact solution fields which are defined with the help of the respective eigenfunctions of the Laplace–Beltrami operator on $]0,\omega[$ with Dirichlet and Neumann boundary conditions:

$$\boldsymbol{G}_n(r,\theta) = \mathbf{grad}(\eta(r) r^{na} \sin(na\theta)) \tag{47}$$

and

$$\boldsymbol{H}_n(r,\theta) = \mathbf{curl}(\eta(r) r^{na} \cos(na\theta)), \quad \text{where } a = \pi/\omega = 2/3 \text{ and } n \in \mathbb{N}. \tag{48}$$

Note that

$$\boldsymbol{H}_n = -\boldsymbol{G}_n + \text{regular field},$$

where the regular field vanishes near the corners. The regularity of the fields $\boldsymbol{H}_n$ and $\boldsymbol{G}_n$ depends on n:

$$\boldsymbol{G}_n, \boldsymbol{H}_n \in H^s(\Omega)^2 \quad \forall s < na.$$

In particular, $\boldsymbol{G}_n$ and $\boldsymbol{H}_n$ belong to $\mathcal{H}_N(\mathrm{grad})$ for $n > 1$, whereas $\boldsymbol{G}_1$ and $\boldsymbol{H}_1$ are singular fields in the sense that they have a nonzero component in $\mathcal{H}_{\rm sing}$. ($\boldsymbol{G}_1$ does indeed coincide with the singular field that spans $\mathcal{H}_{\rm sing}$ in the case of the SFM.)

The "numerical proof" that the FE-method fails for singular fields appears clearly in Figure 6.1. We represent the x-component of the FE-element approximation of $\boldsymbol{G}_1$

on a mesh with $h = 2^{-4}$ and compare it to the analogous approximation obtained by the SFM (Figure 6.2). It turns out that nodal finite elements are not at all able to reproduce the singular behavior at the reentrant corner. Indeed, the condition $(\boldsymbol{E}_h \times \boldsymbol{n}) = 0$ on the boundary nodes forces the FE-approximation to vanish at $\mathbf{0}$, whereas the exact solution tends to ∞ at the corner. Thus, we do not have to deal with an accuracy problem but, as it has been mentioned in section 3, with the choice of the appropriate functional frame: the FE-approximation converges to the solution to $\mathcal{P}_N(\mathrm{grad})$.

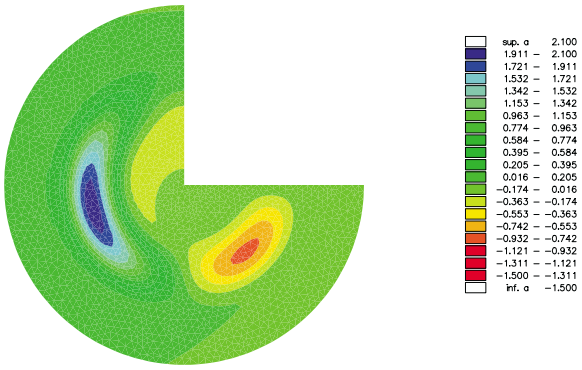


FIG. 6.1. *FE-approximation of* $\boldsymbol{G}_1$.

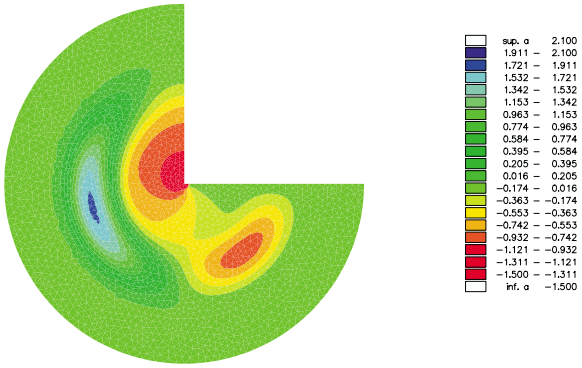


FIG. 6.2. *SFM-approximation of* $\boldsymbol{G}_1$.

Both the SFM and the OSFM have been tested on unstructured triangulations $\mathcal{T}_h$ *without mesh refinement* near the corner. (See Figure 6.3 for an example.) We use Lagrange finite elements of type P1. Notice that in this case the error induced by the approximation of the curvilinear domain Ω by a polygon does not affect the convergence rate of the method.

The mesh parameter $h = \sup_{T_l \in \mathcal{T}_h} \operatorname{diam} T_l$ varies from $h = 2^{-2}$ to $h = 2^{-6}$, the latter corresponding to roughly 2×31900 degrees of freedom.

The cut-off function η is a piecewise polynomial function of class $\mathcal{C}^3$: $\eta \equiv 1$ for $0 \le r < 0.5$, $\eta = p_\eta$ for $0.5 \le r < 1.5$ with a polynomial p_η of degree 7, and $\eta \equiv 0$ for

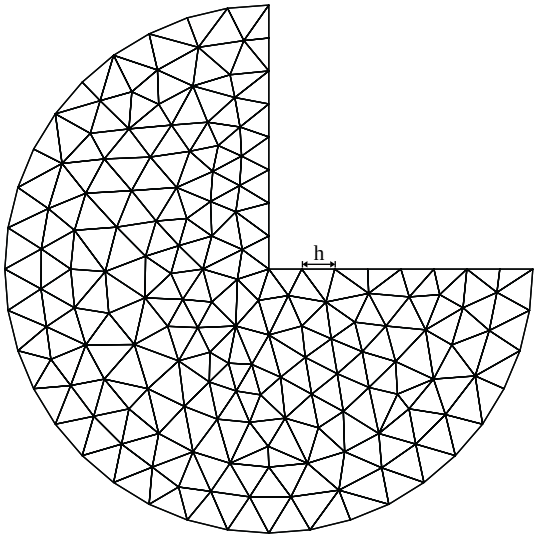


FIG. 6.3. *The coarse mesh* ($h = 1/4$).

$1.5 \leq r \leq 2$. The coefficients of the coupling term C are computed using a 7-point-quadrature formula which is exact for polynomials up to order 5. The coefficients a_s and j_s are calculated analytically.

The implementation of the boundary condition is realized via a rotation which maps the canonical basis on a local basis of the normal and tangential vectors; in the latter basis the vector boundary condition becomes decoupled and standard techniques apply. In order to implement the inhomogeneous boundary condition in the algorithm of the OSFM (see (31)), we use a FE-lifting $\boldsymbol{R}$ of the tangential trace of $-\mathbf{grad}(s)$ on Γ. Again, the only difficulty comes from the coupling of the two degrees of freedom at each boundary node and we proceed in the same way as in the homogeneous case.

The linear systems occurring in the algorithms are solved by a direct method based on Cholesky factorization. All tests have been realized with the FE-code MELINA.[1]

We represent the error on the exact solution in logarithmic scale for a discrete L^2-norm which is given by

$$||\boldsymbol{E} - \boldsymbol{E}_h||_h = \left(\frac{1}{N_h} \sum_{I \in \overset{\circ}{\Omega}} |\boldsymbol{E}(M_I) - \boldsymbol{E}_h(M_I)|^2 \right)^{1/2} . \tag{49}$$

We could have chosen instead a discrete energy norm, which may appear more relevant from a physical point of view. Our aim, however, is simply to compare the numerical results with the theoretical predictions, in spite of their poor practical interest.

Figures 6.4–6.7 show the convergence rates of the SFM and the OSFM for the fields $\boldsymbol{G}_n$ and $\boldsymbol{H}_n$ with $n = 2$ and $n = 5$. This illustrates the L^2-error estimate of Proposition 5.4. For $n = 2$, the numerical values are in good concordance with the theory which yields an order of 2/3. The influence of the reentrant corner may be seen even if the exact solution belongs to H^2 (cf. the tests for $\boldsymbol{G}_5$ and $\boldsymbol{H}_5$): the observed

[1] Developed by D. Martin (IRMAR, University of Rennes 1, and ENSTA/SMP, Paris, France) [26].

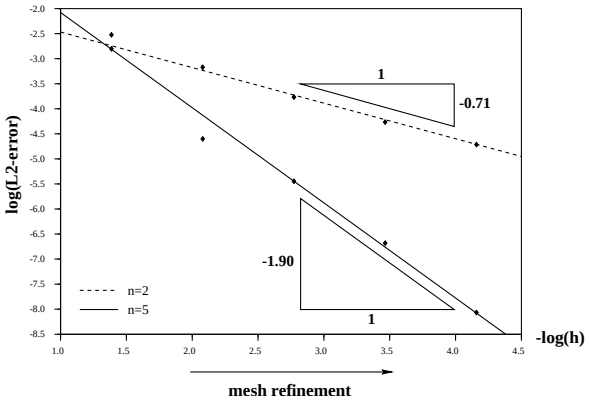


FIG. 6.4. *The SFM for* $\boldsymbol{G}_n$, $n = 2, n = 5$.

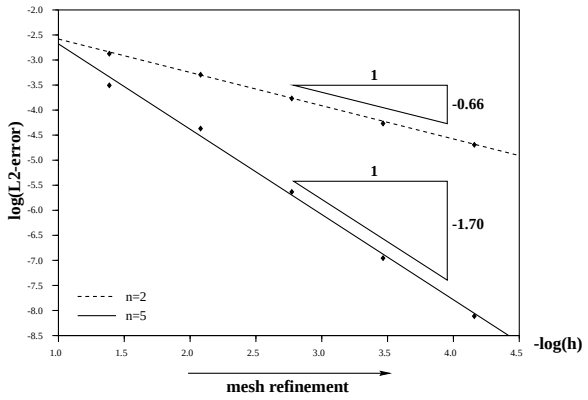


FIG. 6.5. *The SFM for* $\boldsymbol{H}_n$, $n = 2, n = 5$.

convergence rates still keep below the "optimal" rate $\mathcal{O}(h^2)$ for finite elements of type P1 in regular domains. It turns out, however, that the numerical values for $n = 5$ are higher than the order given by Proposition 5.4, $\mathcal{O}(h^{4/3})$.

Next, we represent the approximation of the singular fields $\boldsymbol{G}_1$ and $\boldsymbol{H}_1$. Figure 6.8 shows once again that a nodal finite element method does fail: the relative error on the approximation of $\boldsymbol{G}_1$ is of 52% for the finest mesh (and there is no reason to believe in a convergence as poor as it may be ...), compared to 0.1% for the OSFM. The points representing the error of the finite element method seem to have a horizontal asymptote (corresponding to the "gap" between $\mathcal{H}_N(\mathrm{grad})$ and $\mathcal{H}_N(\mathrm{curl}, \mathrm{div})$).

At last, Figure 6.9 compares the SFM- and the OSFM-approximation of $\boldsymbol{H}_1$, and it is obvious that the OSFM yields much better results. As mentioned in Remark 5.6, this is due to the cut-off function η. Indeed, the performance of the SFM depends on the implementation of the singular field involving the cut-off function and its derivatives. The strong variations of the latter lead to high values of the constant in the error estimations and hence to poor accuracy: in the case of the two coarsest grids we cannot even speak of an approximation since the relative errors are too high (about 70% for $h = 1/8$ compared to 1% for the OSFM) and even the finest mesh still yields an error of 7%. This means that much more mesh refinement has to be done when using the SFM in order to get results comparable with those of the OSFM.

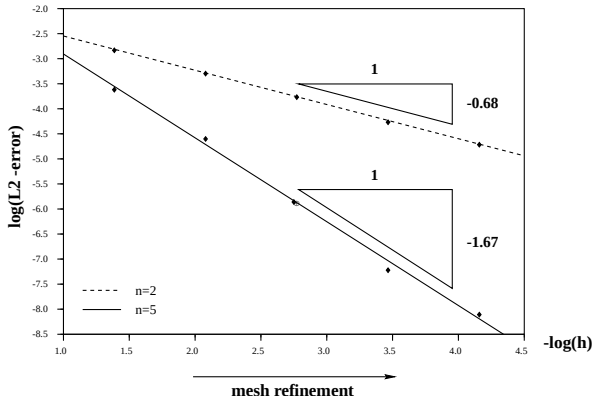


FIG. 6.6. *The OSFM for* $\boldsymbol{G}_n$, $n = 2, n = 5$.

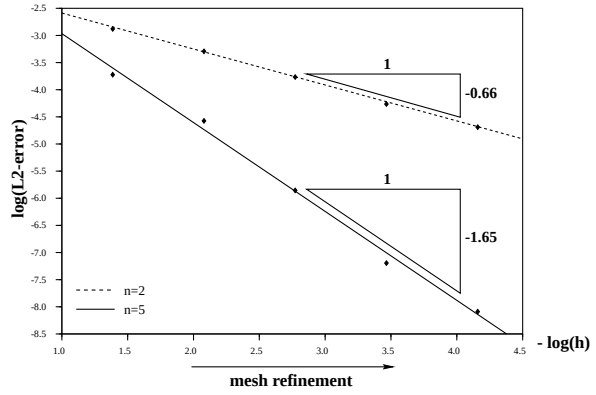


FIG. 6.7. *The OSFM for* $\boldsymbol{H}_n$, $n = 2, n = 5$.

Nevertheless, the numerical results confirm the theoretical order of convergence for the SFM: taking into account the finest three meshes only, the observed convergence rate is $\mathcal{O}(h^{1.24})$ compared to $\mathcal{O}(h^{4/3})$. For the OSFM, a "superconvergence" is observed. We were not able to find a theoretical justification of this phenomena. Maybe it is due to some particular symmetry of the exact solution. Notice as well that we omitted the error of the OSFM for $h = 1/64$ since it is slightly higher than the one for $h = 1/32$. This is probably due to the influence of rounding errors since the FE-solver is implemented with simple precision only.

One question may arise in comparing the SFM and the OSFM: why do accuracy problems occur only for the field $\boldsymbol{H}_1$? To understand this, consider the fields $\boldsymbol{G}_n$ and $\boldsymbol{H}_n$ for $n > 1$. The FE-method will converge since the fields belong to $\mathcal{H}_N(\mathrm{grad})$. It seems plausible that the convergence in the variable θ is much better than in r due to the influence of the cut-off function which does not depend on θ. If we further assume that the FE-method keeps approximately the separation in variables of the exact solution, that is,

$$\boldsymbol{G}_{n,\mathrm{FE}} \approx \mathbf{grad}(u_n(r)\sin(na\theta)) \quad \forall n > 1,$$
$$\boldsymbol{H}_{n,\mathrm{FE}} \approx \mathbf{curl}(v_n(r)\cos(na\theta)) \quad \forall n > 1,$$

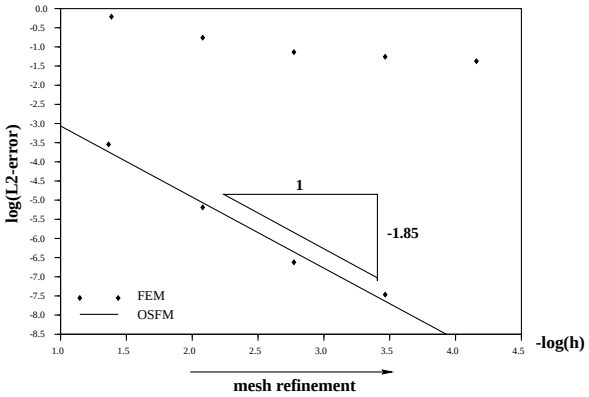


FIG. 6.8. *FEM/OSFM for* $\boldsymbol{G}_1$.

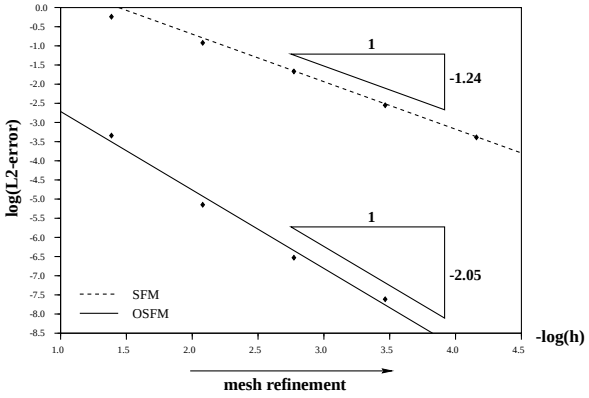


FIG. 6.9. *SFM/OSFM for* $\boldsymbol{H}_1$.

we get

$$(\boldsymbol{G}_{n,\mathrm{FE}}, \mathbf{grad}(\eta s)) \approx \int_\Omega u_n(r)\sin(na\theta)w(r)\sin(a\theta)\, r\, dr d\theta = 0 \tag{50}$$

and

$$a(\boldsymbol{G}_{n,\mathrm{FE}}, \mathbf{grad}(\eta s)) \approx 0, \tag{51}$$

since $\sin(na\theta)$ and $\sin(a\theta)$ are orthogonal on $(0,\omega)$, and the corresponding terms for $\boldsymbol{H}_{n,\mathrm{FE}}$ vanish as well. Now, if we introduce these identities into (25) (note that $C^T E_{\mathrm{FE}}$ is nothing but the left-hand side of (51)) and take into account that $j_s = 0$, we get a good approximation of the singular coefficient $\alpha = 0$, in spite of an eventual pollution effect due to the cut-off function.

On the other hand, consider the field $\boldsymbol{H}_1$. The corresponding FE-solution $\boldsymbol{H}_{1,\mathrm{FE}}$ yields an approximation of the solution to the "spurious" problem $\mathcal{P}_N(\mathrm{grad})$ which is different from $\boldsymbol{H}_1$. In particular, there is no reason to get the "quasi orthogonality"of (50) and (51). The computation of the singular coefficient α_h thus involves the coupling terms $C^T E_{\mathrm{FE}}$ and $C^T S$, and the above-mentioned difficulties due to the cut-off function appear. Unfortunately, this is the general situation, the "superconvergence"observed for $\boldsymbol{G}_n$ and $\boldsymbol{H}_n (n > 1)$ being a lucky exception.

7. Conclusion. We proposed two variants, SFM and OSFM, of a *singular field method* in order to solve Maxwell equations in two-dimensional regions with corners. Both methods were implemented for a simple model problem on a curvilinear polygon with one reentrant corner. The SFM is based on singular fields *localized* to a neighborhood of the corners, whereas its orthogonal variant, OSFM, makes use of approximate singular fields which are *orthogonal* (for an appropriate scalar product) to the discretization space of standard finite elements. With regard to computational complexity, the methods are equivalent, and the error analysis suggests that they are of the same order. In practice, however, the OSFM yields better results, the principal difficulty of the SFM being the implementation of the cut-off function.

Notice that for more involved situations as the one of the model problem studied in this paper, coupling and singular terms occur in the OSFM which have to be calculated on the whole domain. This increases slightly the computational cost of the method. However, the singular fields are regular away from the corners. Hence, special attention has to be paid only near the geometric singularities. If the domain does contain large regions where the boundary is regular, a hybrid method, similar to domain decomposition techniques, may apply which uses the OSFM only near the corners and a standard FE-method (without singular fields) anywhere else.

To finish, let us have a glance at the three-dimensional situation. Whereas the theory is now well understood (see, for example, [8, 14, 15]), the implementation of the method is less simple. This is essentially due to the infinite dimension of the singular subspace. Roughly speaking, the singular coefficient corresponding to the reentrant edge of a polyhedron is now a function belonging to some weighted Sobolev space, and its discretization is far from being obvious. However, some results have been obtained recently in the particular case of an axisymmetric conical point [19]: in this case the space of singular fields is still of finite dimension.

Acknowledgment. The authors would like to thank the anonymous referees for many valuable comments and suggestions.

REFERENCES

[1] F. Assous, P. Ciarlet, Jr., and S. Labrunie, *Theoretical tools to solve the axisymmetric Maxwell equations*, Math. Methods Appl. Sci., 25 (2002), pp. 49–78.

[2] F. Assous, P. Ciarlet, Jr., and E. Garcia, *Numerical solution to Maxwell equations in singular domains: The singular complement method*, in Proceedings of the Fifth International Conference on Mathematical and Numerical Aspects of Wave Propagation, A. Bermúdez, D. Gomez, C. Hazard, P. Joly, and J.E. Roberts, eds., SIAM, Philadelphia, 2000, pp. 714–718.

[3] F. Assous, P. Ciarlet, Jr., and J. Segré, *Numerical solution to the time dependent Maxwell equations in two dimensional singular domains: The singular complement method*, J. Comput. Phys., 161 (2000), pp. 218–249.

[4] F. Assous, P. Degond, E. Heintzé, P.A. Raviart, and J. Segré, *On a finite element method for solving the three-dimensional Maxwell equations*, J. Comput. Phys., 109 (1993), pp. 222–237.

[5] M. Sh. Birman and M.Z. Solomyak, *L_2-theory of the Maxwell operator in arbitrary domains*, Uspekhi Mat. Nauk, 42 (1987), pp. 61–76 (in Russian); Russian Math. Surveys, 42 (1987), pp. 75–96 (in English).

[6] O. Biro and K. Richter, *CAD in electromagnetism*, Adv. Electron. Electron Phys. 82, P. Hawkes, ed., Academic Press, New York, 1991, pp. 1–96.

[7] H. Blum and M. Dobrowolski, *On finite element methods for elliptic equations on domains with corners*, Computing, 28 (1982), pp. 53–63.

[8] A.-S. Bonnet-Ben Dhia, C. Hazard, and S. Lohrengel, *A singular field method for the solution of Maxwell's equations in polyhedral domains*, SIAM J. Appl. Math., 59 (1999), pp. 2028–2044.

[9] A. Bossavit, *On the Lorenz gauge*, COMPEL, 18 (1999), pp. 323–336.
[10] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[11] P. Ciarlet, Jr., N. Filonov, and S. Labrunie, *Un résultat de fermeture pour les équations de Maxwell en géométrie axisymétrique*, C. R. Acad. Sci. Paris, Sér. I Math., 331 (2000), pp. 293–398.
[12] P. Ciarlet, Jr., C. Hazard, and S. Lohrengel, *Les équations de Maxwell dans un polyèdre: un résultat de densité*, C. R. Acad. Sci. Paris, Sér. I Math., 326 (1998), pp. 1305–1310.
[13] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.
[14] M. Costabel and M. Dauge, *Singularités des équations de Maxwell dans un polyèdre*, C. R. Acad. Sci. Paris, Sér. I Math., 324 (1997), pp. 1005–1010.
[15] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.
[16] M. Costabel and M. Dauge, *Un résultat de densité pour les équations de Maxwell régularisées dans un domaine lipschitzien*, C. R. Acad. Sci. Paris, Sér. I Math., 327 (1998), pp. 849–854.
[17] M. Costabel, M. Dauge, and D. Martin, *Numerical investigation of a boundary penalization method for Maxwell equations*, in Proceedings of the Third European Conference on Numerical Mathematics and Advanced Applications, Jyväskylä, Finland, 1999, pp. 198–205.
[18] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math., 1341, Springer-Verlag, Berlin, 1988.
[19] E. Garcia, *Résolution des équations de Maxwell instationnaires dans des domaines non convexes, la méthode du complément singulier*, Thesis, Université de Paris VI, Paris, France, 2002.
[20] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[21] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.
[22] P. Grisvard, *Singularities in Boundary Value Problems*, Masson, Paris, 1992.
[23] C. Hazard, *Numerical simulation of corner singularities: a Paradox in Maxwell-like problems*, C. R. Mécanique, 330 (2002), pp. 57–68.
[24] C. Hazard and M. Lenoir, *On the solution of time-harmonic scattering problems for Maxwell's equation*, SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.
[25] S. Lohrengel, *Etude mathématique et résolution numérique des équations de Maxwell dans un domaine non régulier*, Thesis, University of Paris 6, Paris, France, 1998.
[26] D. Martin, *Documentation* MELINA, Université de Rennes 1, Rennes Cedex, France, 1997, http: //www.maths.univ-rennes1.fr/∼dmartin/melina/www/homepage.html.
[27] M.A. Moussaoui, *Sur l'approximation des solutions du problème de Dirichlet dans un ouvert avec coins*, in Singularities and Constructive Methods for Their Treatment, P. Grisvard, W. Wendland, and J.R. Whiteman, eds., Springer-Verlag, Berlin, 1985.
[28] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1041–1063

STRONG CONVERGENCE OF EULER-TYPE METHODS FOR NONLINEAR STOCHASTIC DIFFERENTIAL EQUATIONS*

DESMOND J. HIGHAM†, XUERONG MAO‡, AND ANDREW M. STUART§

Abstract. Traditional finite-time convergence theory for numerical methods applied to stochastic differential equations (SDEs) requires a global Lipschitz assumption on the drift and diffusion coefficients. In practice, many important SDE models satisfy only a local Lipschitz property and, since Brownian paths can make arbitrarily large excursions, the global Lipschitz-based theory is not directly relevant. In this work we prove strong convergence results under less restrictive conditions. First, we give a convergence result for Euler–Maruyama requiring only that the SDE is locally Lipschitz and that the pth moments of the exact and numerical solution are bounded for some $p > 2$. As an application of this general theory we show that an implicit variant of Euler–Maruyama converges if the diffusion coefficient is globally Lipschitz, but the drift coefficient satisfies only a one-sided Lipschitz condition; this is achieved by showing that the implicit method has bounded moments and may be viewed as an Euler–Maruyama approximation to a perturbed SDE of the same form. Second, we show that the optimal *rate* of convergence can be recovered if the drift coefficient is also assumed to behave like a polynomial.

Key words. backward Euler, Euler–Maruyama, finite-time convergence, implicit, moment bounds, nonlinearity, one-sided Lipschitz condition, split-step

AMS subject classifications. 65C30, 65C20, 65L20

PII. S0036142901389530

1. Introduction.

In this paper we study the numerical solution of the stochastic differential equation (SDE)

$$dy(t) = f(y(t))dt + g(y(t))dW(t), \quad 0 \le t \le T, \quad y(0) = y_0. \tag{1.1}$$

Here $y(t) \in \mathbb{R}^m$ for each t, and $W(t)$ is a d-dimensional Brownian motion. Thus $f : \mathbb{R}^m \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^{m\times d}$. We assume that the initial condition is chosen independently of the Wiener measure driving the equation and that all pth moments of y_0, $p > 0$, are finite. Our primary objective is to study strong convergence questions for numerical approximations in the case where f and g are not necessarily globally Lipschitz functions. Most of the existing convergence theory for numerical methods requires f and g to be globally Lipschitz; see [12, 14], for example. Recent work has studied convergence in probability [6, 8] and almost sure convergence [7], under more relaxed conditions on f and g. We focus here on strong mean square convergence, in the sense of [12, Theorem 10.6.3], which implies convergence in probability. The main result of [11] is directly related to our work; we summarize the connections at the end of this section. We also note that the work of Schurz [16] contains a number

*Received by the editors May 18, 2001; accepted for publication (in revised form) March 1, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/sinum/40-3/38953.html

†Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK (djh@maths.strath.ac.uk).

‡Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK (xuerong@stams.strath.ac.uk). The research of this author was supported by the Biotechnology and Biological Sciences Research Council of the UK under grant 78/MMI09712.

§Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (stuart@maths.warwick.ac.uk). The research of this author was supported by the Engineering and Physical Sciences Research Council of the UK under grant GR/N00340.

of a priori bounds on the numerical solutions of SDEs from a nonlinear stability perspective, under structural assumptions similar to those we employ here.

In section 2, we prove that the Euler–Maruyama method converges strongly if f and g are locally Lipschitz (e.g., for $f, g \in C^1$) and the exact and numerical solution have a bounded pth moment for some $p > 2$ (Theorem 2.2). The bounded moment assumption will not, of course, hold, in general, as solutions to the SDE may explode in a finite time. In section 3 we therefore impose further assumptions on f and g to ensure that $y(t)$ has bounded moments: we assume that g is globally Lipschitz and that f satisfies a one-sided Lipschitz condition. The one-sided Lipschitz condition has proved effective in the analysis of numerical methods for deterministic problems and occurs naturally in a variety of applications [1, 3, 4, 5, 17]. For a suitably constructed "split-step" implicit variant of Euler–Maruyama we establish strong convergence (Theorem 3.3) by (a) showing that the method corresponds to Euler–Maruyama on a perturbed SDE and (b) showing that all moments of the numerical solution are bounded. We are unable to establish moment bounds for the forward Euler method and, indeed, it may not be possible to do so; the work of [7] is of importance for nonimplicit methods since, being an almost sure convergence study, it does not require moment bounds.

In section 4 we turn our attention to establishing an optimal *rate* of convergence for the split-step implicit method. We show that if f satisfies a one-sided Lipschitz condition and behaves polynomially, then Euler–Maruyama converges strongly at the optimal rate (Theorem 4.4), provided moment bounds hold. We use this to study the split-step implicit method, for which moment bounds can be found (Theorem 4.7), again by showing that the method corresponds to Euler–Maruyama on a perturbed SDE. In section 5 this result is extended to a more widely used implicit variant of Euler–Maruyama by relating the two implicit methods (Theorem 5.3). A summary is given in section 6.

It is worth mentioning at this point how our work compares with that of Hu [11]. Theorem 2.4 of [11] is a very important contribution to numerical SDE theory, being, to our knowledge, the first strong convergence result without global Lipschitz assumptions. Hu considered only the backward Euler method and derived a result with the optimal rate of convergence, and hence his work may be compared with Theorem 5.3 below. Both results assume C^1 coefficients in the SDE, a one-sided Lipschitz condition for the drift, and a global Lipschitz condition for the diffusion. Theorem 5.3 below imposes polynomial-type growth on the drift (Assumption 4.1), whereas Hu allows for a more general exponential growth. On the other hand, Theorem 5.3 and all the other results in our work deal with a very strong error measure, $\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Z}(t)-y(t)|^2\right]$, whereas [11] uses the less stringent measure $\int_0^T \mathbb{E}|\overline{Z}(t)-y(t)|^2\,dt$. We also note that [11] uses a different continuous-time extension. Overall, Hu's result for backward Euler applies to a wider class of SDEs but controls a weaker measure of the error. The techniques of analysis are significantly different, although both hinge on establishing moment bounds for the exact and numerical solutions.

2. General result for Euler–Maruyama. Given a stepsize $\Delta t > 0$, the Euler–Maruyama (EM) method applied to (1.1) computes approximations $X_k \approx y(t_k)$, where $t_k = k\Delta t$, by setting $X_0 = y_0$ and forming

$$X_{k+1} = X_k + \Delta t f(X_k) + g(X_k)\Delta W_k, \tag{2.1}$$

where $\Delta W_k = W(t_{k+1}) - W(t_k)$. We find it convenient to use continuous-time approximations, and hence we define $\overline{X}(t)$ by

$$\overline{X}(t) := X_k + (t - t_k)f(X_k) + g(X_k)(W(t) - W(t_k)) \quad \text{for } t \in [t_k, t_{k+1}).$$

In our analysis it will be more natural to work with the equivalent definition

$$\overline{X}(t) := X_0 + \int_0^t f(X(s))ds + \int_0^t g(X(s))dW(s), \tag{2.2}$$

where $X(t)$ is defined by

$$X(t) := X_k \quad \text{for } t \in [t_k, t_{k+1}). \tag{2.3}$$

Note that $\overline{X}(t_k) = X(t_k) = X_k$; that is, $\overline{X}(t)$ and $X(t)$ coincide with the discrete solution at the gridpoints. We refer to $X(t)$ and $\overline{X}(t)$ as *continuous-time extensions* of the discrete approximation $\{X_k\}$. We will study the error in $\overline{X}(t)$ in the supremum norm; this will, of course, give an immediate bound for the error in the discrete approximation.

Our first result makes the following assumption on the SDE (1.1) and the exact and numerical solutions. Here, and throughout the paper, $|\cdot|$ is used to denote both the Euclidean vector norm and the Frobenius (or trace) matrix norm.

ASSUMPTION 2.1. *For each $R > 0$ there exists a constant C_R, depending only on R, such that*

$$|f(a) - f(b)|^2 \vee |g(a) - g(b)|^2 \leq C_R|a - b|^2 \ \forall a, b \in \mathbb{R}^m \text{ with } |a| \vee |b| \leq R. \tag{2.4}$$

For some $p > 2$ there is a constant A such that

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |\overline{X}(t)|^p\right] \vee \mathbb{E}\left[\sup_{0\leq t\leq T} |y(t)|^p\right] \leq A. \tag{2.5}$$

Inequality (2.4) is a local Lipschitz assumption. From the mean value theorem, any f and g in C^1 will satisfy (2.4). The inequality (2.5) states that the pth moments of the exact and numerical solution are bounded for some $p > 2$. We now prove that Assumption 2.1 is sufficient to ensure strong convergence of EM.

THEOREM 2.2. *Under Assumption* 2.1, *the EM solution* (2.1) *with continuous-time extension* (2.2) *satisfies*

$$\lim_{\Delta t\to 0} \mathbb{E}\left[\sup_{0\leq t\leq T} |\overline{X}(t) - y(t)|^2\right] = 0. \tag{2.6}$$

Proof. First, we define

$$\tau_R := \inf\{t \geq 0 : |\overline{X}(t)| \geq R\}, \quad \rho_R := \inf\{t \geq 0 : |y(t)| \geq R\}, \quad \theta_R := \tau_R \wedge \rho_R$$

and

$$e(t) := \overline{X}(t) - y(t).$$

Recall the Young inequality: for $r^{-1} + q^{-1} = 1$

$$ab \leq \frac{\delta}{r}a^r + \frac{1}{q\delta^{q/r}}b^q \quad \forall a, b, \delta > 0.$$

We thus have for any $\delta > 0$

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\right] &= \mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\mathbf{1}_{\{\tau_R>T,\rho_R>T\}}\right] + \mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\mathbf{1}_{\{\tau_R\le T \text{ or } \rho_R\le T\}}\right]\\ &\le \mathbb{E}\left[\sup_{0\le t\le T}|e(t\wedge\theta_R)|^2\mathbf{1}_{\{\theta_R>T\}}\right] + \frac{2\delta}{p}\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^p\right]\\ &\quad + \frac{1-\frac{2}{p}}{\delta^{2/(p-2)}}\mathbb{P}\big(\tau_R\le T \text{ or } \rho_R\le T\big). \qquad (2.7)\end{aligned}$$

Now

$$\mathbb{P}\left(\tau_R\le T\right) = \mathbb{E}\left[\mathbf{1}_{\{\tau_R\le T\}}\frac{|\overline{X}(\tau_R)|^p}{R^p}\right] \le \frac{1}{R^p}\mathbb{E}\left[\sup_{0\le t\le T}|\overline{X}(t)|^p\right] \le \frac{A}{R^p},$$

using (2.5). A similar result can be derived for ρ_R so that

$$\mathbb{P}\big(\tau_R\le T \text{ or } \rho_R\le T\big) \le \mathbb{P}\big(\tau_R\le T\big) + \mathbb{P}\big(\rho_R\le T\big) \le \frac{2A}{R^p}.$$

Using these bounds along with

$$\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^p\right] \le 2^{p-1}\mathbb{E}\left[\sup_{0\le t\le T}\big(|\overline{X}(t)|^p + |y(t)|^p\big)\right] \le 2^pA$$

in (2.7) gives

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\right] &\le \mathbb{E}\left[\sup_{0\le t\le T}|\overline{X}(t\wedge\theta_R) - y(t\wedge\theta_R)|^2\right]\\ &\quad + \frac{2^{p+1}\delta A}{p} + \frac{(p-2)2A}{p\delta^{2/(p-2)}R^p}. \qquad (2.8)\end{aligned}$$

Now we bound the first term on the right-hand side of (2.8) using an approach similar to a finite-time convergence proof for the globally Lipschitz case. Using

$$y(t\wedge\theta_R) := y_0 + \int_0^{t\wedge\theta_R} f(y(s))ds + \int_0^{t\wedge\theta_R} g(y(s))dW(s),$$

(2.2), and Cauchy–Schwarz, we have

$$\begin{aligned}|\overline{X}(t\wedge\theta_R) - y(t\wedge\theta_R)|^2 &= \left|\int_0^{t\wedge\theta_R} f(X(s)) - f(y(s))ds\right.\\ &\qquad \left. + \int_0^{t\wedge\theta_R} g(X(s)) - g(y(s))dW(s)\right|^2\\ &\le 2\left[T\int_0^{t\wedge\theta_R}|f(X(s)) - f(y(s))|^2ds\right.\\ &\qquad \left. + \left|\int_0^{t\wedge\theta_R} g(X(s)) - g(y(s))dW(s)\right|^2\right].\end{aligned}$$

So, from the local Lipschitz condition (2.4) and Doob's martingale inequality [14] we have for any $\tau \le T$

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le\tau}|\overline{X}(t\wedge\theta_R)-y(t\wedge\theta_R)|^2\right]\\
&\le 2C_R(T+4)\mathbb{E}\int_0^{\tau\wedge\theta_R}|X(s)-y(s)|^2ds\\
&\le 4C_R(T+4)\mathbb{E}\int_0^{\tau\wedge\theta_R}\left[|X(s)-\overline{X}(s)|^2+|\overline{X}(s)-y(s)|^2\right]ds\\
&\le 4C_R(T+4)\left[\mathbb{E}\int_0^{\tau\wedge\theta_R}|X(s)-\overline{X}(s)|^2ds+\mathbb{E}\int_0^{\tau}|\overline{X}(s\wedge\theta_R)-y(s\wedge\theta_R)|^2ds\right]\\
&\le 4C_R(T+4)\left[\mathbb{E}\int_0^{\tau\wedge\theta_R}|X(s)-\overline{X}(s)|^2ds+\int_0^{\tau}\mathbb{E}\sup_{0\le r\le s}|\overline{X}(r\wedge\theta_R)-y(r\wedge\theta_R)|^2ds\right].
\end{aligned}\tag{2.9}$$

To bound the first term in the parentheses on the right-hand side of (2.9), given $s\in[0,T\wedge\theta_R)$, let k_s be the integer for which $s\in[t_{k_s},t_{k_s+1})$. Then

$$\begin{aligned}
X(s)-\overline{X}(s)&=X_{k_s}-\left(X_{k_s}+\int_{t_{k_s}}^{s}f(X(s))ds+\int_{t_{k_s}}^{s}g(X(s))dW(s)\right)\\
&=-f(X_{k_s})(s-t_{k_s})-g(X_{k_s})\left(W(s)-W(t_{k_s})\right).
\end{aligned}$$

Hence,

$$|X(s)-\overline{X}(s)|^2\le 2\left[|f(X_{k_s})|^2\Delta t^2+|g(X_{k_s})|^2|W(s)-W(t_{k_s})|^2\right].\tag{2.10}$$

Now, from the local Lipschitz condition (2.4), for $|y|\le R$ we have

$$|f(y)|^2\le 2\left(|f(y)-f(0)|^2+|f(0)|^2\right)\le 2\left(C_R|y|^2+|f(0)|^2\right),$$

and, similarly,

$$|g(y)|^2\le 2\left(C_R|y|^2+|g(0)|^2\right).$$

Hence, in (2.10),

$$|X(s)-\overline{X}(s)|^2\le 4(C_R|X_{k_s}|^2+|f(0)|^2\vee|g(0)|^2)(\Delta t^2+|W(s)-W(t_{k_s})|^2).$$

Thus, using (2.5) and the Lyapunov inequality [12]

$$\begin{aligned}
&\mathbb{E}\int_0^{\tau\wedge\theta_R}|X(s)-\overline{X}(s)|^2ds\\
&\le\mathbb{E}\int_0^{\tau\wedge\theta_R}4(C_R|X_{k_s}|^2+|f(0)|^2\vee|g(0)|^2)(\Delta t^2+|W(s)-W(t_{k_s})|^2)ds\\
&\le\int_0^{\tau}4\mathbb{E}\left[C_R|X_{k_s}|^2+|f(0)|^2\vee|g(0)|^2)(\Delta t^2+|W(s)-W(t_{k_s})|^2\right]ds\\
&\le\int_0^{T}4(C_R\mathbb{E}[|X_{k_s}|^2]+|f(0)|^2\vee|g(0)|^2)(\Delta t^2+m\Delta t)ds\\
&\le 4T(C_RA^{2/p}+|f(0)|^2\vee|g(0)|^2)\Delta t(\Delta t+m).
\end{aligned}$$

In (2.9) we then have

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le\tau}\left|\overline{X}(t\wedge\theta_R)-y(t\wedge\theta_R)\right|^2\right]\\&\le 16C_R(T+4)T\Delta t(\Delta t+m)(C_RA^{2/p}+|f(0)|^2\vee|g(0)|^2)\\&\quad+4C_R(T+4)\int_0^\tau\mathbb{E}\sup_{0\le r\le s}\left[\left|\overline{X}(r\wedge\theta_R)-y(r\wedge\theta_R)\right|^2\right]ds.\end{aligned}$$

Applying the Gronwall inequality [14] we obtain

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\overline{X}(t\wedge\theta_R)-y(t\wedge\theta_R)\right|^2\right]\le C\Delta t(C_R^2+1)e^{4C_R(T+4)},$$

where here, and in the following, C is a universal constant independent of Δt, R, and δ. Inserting this into (2.8) gives

$$\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\right]\le C\Delta t(C_R^2+1)e^{4C_R(T+4)}+\frac{2^{p+1}\delta A}{p}+\frac{(1-\frac{2}{p})2A}{\delta^{2/(p-2)}R^p}. \tag{2.11}$$

Given any $\epsilon>0$, we can choose δ so that $(2^{p+1}\delta A)/p<\epsilon/3$, then choose R so that

$$\frac{(1-\frac{2}{p})2A}{\delta^{2/(p-2)}R^p}<\frac{\epsilon}{3},$$

and then choose Δt sufficiently small for

$$C\Delta t(C_R^2+1)e^{4C_R(T+4)}<\frac{\epsilon}{3}$$

so that, in (2.11), $\mathbb{E}[\sup_{0\le t\le T}|e(t)|^2]<\epsilon$, as required. □

We remark that the proof of Theorem 2.2 is optimal in the sense that in the globally Lipschitz case ($C_R\le C$ for all R) we may take $\delta=\Delta t$ and $R=\Delta t^{-1/(p-2)}$ in (2.11) to recover the classical finite-time convergence result

$$\mathbb{E}\left[\sup_{0\le t\le T}|e(t)|^2\right]=O(\Delta t),$$

found, for example, in [12, 14].

3. Convergence with a one-sided Lipschitz condition.

3.1. The one-sided Lipschitz condition. In this section we impose further assumptions on the SDE. In section 3.2 we show that these assumptions guarantee moment bounds for $y(t)$. Theorem 2.2 requires bounds on the pth moment of the exact *and* numerical solution—a condition that is difficult to verify in practice for the method (2.1) and indeed may fail to hold. In section 3.3 we introduce an implicit version of the EM method for which moment bounds, and hence a convergence result, can be obtained.

We make the following assumptions on the SDE.

ASSUMPTION 3.1. *The functions f and g in* (1.1) *are C^1, and there exist constants $\mu, c>0$ such that*

$$\langle a-b,f(a)-f(b)\rangle\le\mu|a-b|^2\quad\forall a,b\in\mathbb{R}^m, \tag{3.1}$$

$$|g(a)-g(b)|^2\le c|a-b|^2\quad\forall a,b\in\mathbb{R}^m. \tag{3.2}$$

Note that we work with the case $\mu > 0$. In the deterministic setting there is a lot of attention paid to the contractive case $\mu < 0$. This case is of less interest here because, for most diffusion coefficients g, contractivity is destroyed. Hence $\mu > 0$ is a natural assumption.

It follows from Assumption 3.1 that

$$\langle f(a), a\rangle = \langle f(a) - f(0), a\rangle + \langle f(0), a\rangle \le \mu|a|^2 + |f(0)||a| \le \tfrac{1}{2}|f(0)|^2 + (\mu + \tfrac{1}{2})|a|^2$$

and

$$|g(a)|^2 \le 2|g(0)|^2 + 2|g(a) - g(0)|^2 \le 2|g(0)|^2 + 2c|a|^2.$$

This gives

$$\langle f(a), a\rangle \vee |g(a)|^2 \le \alpha + \beta|a|^2 \quad \forall a \in \mathbb{R}^m, \tag{3.3}$$

where

$$\alpha := \tfrac{1}{2}|f(0)|^2 \vee 2|g(0)|^2 \quad \text{and} \quad \beta := (\mu + \tfrac{1}{2}) \vee 2c. \tag{3.4}$$

The inequality (3.3) will prove very useful in what follows. We note that from [14, Theorem 2.3.5] $f, g \in C^1$ and (3.3) ensure the existence of a unique solution to the SDE (1.1).

The inequality (3.1) in Assumption 3.1, which is known as a *one-sided Lipschitz condition*, has been exploited successfully in the deterministic numerical analysis literature [1, 3, 4, 5, 17] and in the case of SDEs has been used in [11, 15, 16]. The condition (3.3) is closely related to the *monotone* condition in [14, section 2.4]. Any f of the form $f(y) = -y^p + y$, where the integer $p \ge 3$ is odd, satisfies (3.1), and further examples can be found in [17].

3.2. Moment bounds for the SDE. We now show that under Assumption 3.1 the SDE solution has a bounded pth moment for each $p > 2$.

LEMMA 3.2. *Under Assumption* 3.1*, for each* $p > 2$ *there is* $C = C(p, T) > 0$ *such that*

$$\mathbb{E}\left[\sup_{0\le t\le T} |y(t)|^p\right] \le C(1 + \mathbb{E}|y_0|^p). \tag{3.5}$$

Proof. Theorem 2.4.1 of [14] shows that, for $p \ge 2$, there is $C = C(p, T)$ such that

$$\mathbb{E}|y(t)|^p \le C[1 + \mathbb{E}|y_0|^p] \quad \forall t \in [0, T]. \tag{3.6}$$

By the Itô formula

$$|y(t)|^2 = |y_0|^2 + 2\int_0^t \langle f(y(s)), y(s)\rangle ds + \int_0^t |g(y(s))|^2 ds + 2\int_0^t \langle y(s), g(y(s))dB(s)\rangle.$$

By (3.3) we have that, for some $K = K(p)$ and $t_1 \in [0, T]$,

$$\sup_{0\le t\le t_1} |y(t)|^p \le K\left(|y_0|^p + \left\{\int_0^{t_1} [\alpha + \beta|y(s)|^2]ds\right\}^{p/2} + \sup_{0\le t\le t_1}\left|\int_0^t \langle y(s), g(y(s))dB(s)\rangle\right|^{p/2}\right).$$

By property (3.6) we can take expectations to give, for a possibly different $K = K(p,T)$,

$$\mathbb{E}\left[\sup_{0\le t\le t_1}|y(t)|^p\right] \le K\left(1+\mathbb{E}|y_0|^p + \int_0^{t_1}\mathbb{E}|y(s)|^p ds \right. \\ \left. + \mathbb{E}\left[\sup_{0\le t\le t_1}\left|\int_0^t \langle y(s), g(y(s))dB(s)\rangle\right|^{p/2}\right]\right).$$

By the Burkholder–Davis–Gundy inequality [14], we compute that, again for redefined $K = K(p,T)$,

$$\mathbb{E}\left[\sup_{0\le t\le t_1}|y(t)|^p\right] \le K\left(1+E|y_0|^p + \int_0^{t_1}\mathbb{E}|y(s)|^p ds \right. \\ \left. + \mathbb{E}\left[\int_0^{t_1}|y(s)|^2|g(y(s))|^2 ds\right]^{p/4}\right). \tag{3.7}$$

Next, note that, by Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}\left[\int_0^{t_1}|y(s)|^2|g(y(s))|^2 ds\right]^{p/4} &\le \mathbb{E}\left[\sup_{0\le s\le t_1}|y(s)|^{p/2}\left(\int_0^{t_1}|g(y(s))|^2 ds\right)^{p/4}\right] \\ &\le \frac{1}{2K}\mathbb{E}\left[\sup_{0\le s\le t_1}|y(s)|^p\right] + \frac{K}{2}\mathbb{E}\left[\int_0^{t_1}|g(y(s))|^2 ds\right]^{p/2} \\ &\le \frac{1}{2K}\mathbb{E}\left[\sup_{0\le s\le t_1}|y(s)|^p\right] \\ &\quad + \frac{K}{2}T^{(p-2)/2}\mathbb{E}\int_0^{t_1}(\alpha+\beta|y(s)|^2)^{p/2}ds.\end{aligned}$$

Substituting this into (3.7) yields, again for a possibly different $K = K(p,T)$,

$$\mathbb{E}\left[\sup_{0\le t\le t_1}|y(t)|^p\right] \le K\left(1+\mathbb{E}|y_0|^p + \int_0^{t_1}\mathbb{E}|y(s)|^p ds\right).$$

The required assertion now follows from property (3.6). ∎

3.3. Split-step backward Euler. We now consider the split-step backward Euler (SSBE) method, which is defined by taking $Y_0 = y_0$ and, generally,

$$Y_k^\star = Y_k + \Delta t f(Y_k^\star), \tag{3.8}$$

$$Y_{k+1} = Y_k^\star + g(Y_k^\star)\Delta W_k. \tag{3.9}$$

We state our convergence theorem here and then give a sequence of results that lead to a proof.

THEOREM 3.3. *Consider the SSBE* (3.8)–(3.9) *applied to the SDE* (1.1) *under Assumption* 3.1. *There exists a continuous-time extension* $\overline{Y}(t)$ *of the numerical solution (so that* $\overline{Y}(t_k) = Y_k$*) for which*

$$\lim_{\Delta t\to 0}\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Y}(t)-y(t)|^2\right] = 0.$$

Proof. See the end of this subsection. ∎

Note that (3.8) is an implicit equation that must be solved in order to obtain the intermediate approximation $Y_k^\star$. Having obtained $Y_k^\star$, adding the appropriate stochastic increment $g(Y_k^\star)\Delta W_k$ produces the next approximation Y_{k+1} in (3.9). The SSBE method reduces to the deterministic backward Euler method [5, 9] when $g \equiv 0$ and y_0 is nonrandom. The method is also studied in [15], where it is effective for inheriting ergodicity; for related reasons it is effective here in enabling the derivation of moment bounds. Another stochastic extension of the deterministic backward Euler method is considered in section 5.

Our proof of Theorem 3.3 relies on showing that SSBE has two key properties under Assumption 3.1: (a) it may be regarded as EM applied to a modified SDE of a similar form, and (b) it produces solutions with all moments bounded. The first property is established in the next lemma and corollary.

LEMMA 3.4. *Let Assumption* 3.1 *hold and suppose* $\Delta t \in (0, \Delta t_c), \Delta t_c < 1/(2\beta)$, *where* β *is defined in* (3.4). *Given* $d \in \mathbb{R}^m$ *the implicit equation*

$$c = d + \Delta t f(c) \tag{3.10}$$

has a unique solution c. *If we define the functions* $F_{\Delta t}(\cdot)$, $f_{\Delta t}(\cdot)$ *and* $g_{\Delta t}(\cdot)$ *by*

$$F_{\Delta t}(d) = c, \quad f_{\Delta t}(d) = f(F_{\Delta t}(d)), \quad \text{and} \quad g_{\Delta t}(d) = g(F_{\Delta t}(d)), \tag{3.11}$$

then $F_{\Delta t}, f_{\Delta t}, g_{\Delta t} \in C^1$, $g_{\Delta t}(\cdot) \to g(\cdot)$ *and* $f_{\Delta t}(\cdot) \to f(\cdot)$ *as* $\Delta t \to 0$ *in* C^1 *uniformly on compact sets and, for any* $a, b \in \mathbb{R}^m$,

$$|f_{\Delta t}(a)| \le \frac{|f(a)|}{1 - \Delta t\mu}, \tag{3.12}$$

$$|F_{\Delta t}(d) - F_{\Delta t}(e)|^2 \le \frac{1}{1 - 2\Delta t\mu}|d - e|^2, \tag{3.13}$$

$$\langle a - b, f_{\Delta t}(a) - f_{\Delta t}(b)\rangle \le \frac{\mu}{1 - 2\mu\Delta t}|a - b|^2. \tag{3.14}$$

Further, $g_{\Delta t}$ *is globally Lipschitz, and there exist* $\alpha', \beta' > 0$ *such that*

$$\langle f_{\Delta t}(a), a\rangle \vee |g_{\Delta t}(a)|^2 \le \alpha' + \beta'|a|^2 \quad \forall a \in \mathbb{R}^m. \tag{3.15}$$

Proof. See Appendix A. ∎

COROLLARY 3.5. *Let Assumption* 3.1 *hold and suppose* $\Delta t \in (0, \Delta t_c), \Delta t_c < 1/(2\beta)$, *where* β *is defined in* (3.4). *Then SSBE applied to* (1.1) *is equivalent to EM applied to the modified SDE*

$$dy_{\Delta t}(t) = f_{\Delta t}(y_{\Delta t}(t))dt + g_{\Delta t}(y_{\Delta t}(t))dW(t), \qquad 0 \le t \le T,\ y_{\Delta t}(0) = y_0, \tag{3.16}$$

where $f_{\Delta t}, g_{\Delta t}$ *are defined in Lemma* 3.4.

Proof. Lemma 3.4 allows us to express the SSBE method (3.8)–(3.9) in the form

$$Y_{k+1} = Y_k + \Delta t f_{\Delta t}(Y_k) + g_{\Delta t}(Y_k)\Delta W_k, \tag{3.17}$$

and the result is then immediate. ∎

Next, we show that the solution of the modified SDE (3.16) has bounded moments and converges strongly to $y(t)$.

LEMMA 3.6. *Under Assumption* 3.1, *for each* $p > 2$, *there is* $C = C(p,T) > 0$ *such that*

$$\mathbb{E}\left[\sup_{0\le t\le T}|y_{\Delta t}(t)|^p\right] \le C(1+\mathbb{E}|y_0|^p), \tag{3.18}$$

provided Δt *is sufficiently small. In addition,*

$$\lim_{\Delta t\to 0}\mathbb{E}\left[\sup_{0\le t\le T}|y(t)-y_{\Delta t}(t)|^2\right] = 0. \tag{3.19}$$

Proof. It follows from Lemma 3.4 that for sufficiently small Δt the functions $f_{\Delta t}$ and $g_{\Delta t}$ satisfy (3.3) with α and β replaced by 2α and 2β. Following through the proof of Lemma 3.2, which is based entirely on (3.3), we obtain (3.18).

Now to prove (3.19) we note from Lemma 3.4 that given $R > 0$ there is a function $K_R : (0,\infty) \to (0,\infty)$ such that $K_R(\Delta t) \to 0$ as $\Delta t \to 0$ and

$$|f_{\Delta t}(u) - f(u)|^2 \vee |g_{\Delta t}(u) - g(u)|^2 \le K_R(\Delta t) \quad \forall u \in \mathbb{R}^m, |u| \le R, \tag{3.20}$$

provided Δt is sufficiently small. Also, since $f, g \in C^1$, there is a constant H_R such that

$$|f(u) - f(v)|^2 \vee |g(u) - g(v)|^2 \le H_R|u-v|^2 \quad \forall u, v \in \mathbb{R}^m, |u| \vee |v| \le R. \tag{3.21}$$

From Lemma 3.2 and (3.18) we have

$$\mathbb{E}\left[\sup_{0\le t\le T}|y(t)|^p\right] \vee \mathbb{E}\left[\sup_{0\le t\le T}|y_{\Delta t}(t)|^p\right] \le K := C(1+\mathbb{E}|y_0|^p). \tag{3.22}$$

The remainder of the proof follows in a similar manner to that of Theorem 2.2. Define

$$\tau_R = \inf\{t \ge 0 : |y(t)| \ge R\}, \quad \rho_R = \inf\{t \ge 0 : |y_{\Delta t}(t)| \ge R\}, \quad \theta_R = \tau_R \wedge \rho_R.$$

For any $\delta > 0$, in the same way that (2.8) was obtained, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}|y(t)-y_{\Delta t}(t)|^2\right] &\le \mathbb{E}\left[\sup_{0\le t\le T}|y(t\wedge\theta_R)-y_{\Delta t}(t\wedge\theta_R)|^2\right] \\ &\quad + \frac{2^{p+1}\delta K}{p} + \frac{(p-2)2K}{p\delta^{2/(p-2)}R^p}.\end{aligned} \tag{3.23}$$

To bound the first term on the right-hand side of (3.23), we observe that

$$\begin{aligned}&y(t\wedge\theta_R) - y_{\Delta t}(t\wedge\theta_R) \\ &= \int_0^{t\wedge\theta_R}[f(y(s)) - f(y_{\Delta t}(s)) + f(y_{\Delta t}(s)) - f_{\Delta t}(y_{\Delta t}(s))]ds \\ &\quad + \int_0^{t\wedge\theta_R}[g(y(s)) - g(y_{\Delta t}(s)) + g(y_{\Delta t}(s)) - g_{\Delta t}(y_{\Delta t}(s))]dW(s).\end{aligned}$$

Using (3.20), (3.21), Cauchy–Schwarz, and the Doob martingale inequality, we have that, for any $\tau \le T$,

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le\tau}|y(t\wedge\theta_R)-y_{\Delta t}(t\wedge\theta_R)|^2\right] \\ &\le 4H_R(T+4)\int_0^\tau \mathbb{E}\left[\sup_{0\le t\le s}|y(t\wedge\theta_R)-y_{\Delta t}(t\wedge\theta_R)|^2\right]ds \\ &\quad + 4T(T+4)K_R(\Delta t).\end{aligned}$$

So the Gronwall inequality yields

$$\mathbb{E}\left[\sup_{0\le t\le T}|y(t\wedge\theta_R)-y_{\Delta t}(t\wedge\theta_R)|^2\right]\le 4T(T+4)K_R(\Delta t)e^{4H_R(T+4)T}.$$

Inserting this into (3.23) gives

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}|y(t)-y_{\Delta t}(t)|^2\right]&\le 4T(T+4)K_R(\Delta t)e^{4H_R(T+4)T}\\&\quad+\frac{2^{p+1}\delta K}{p}+\frac{(p-2)2K}{p\delta^{2/(p-2)}R^p}.\end{aligned}\tag{3.24}$$

The final step of the proof follows that of Theorem 2.2. □

Now we show that the special structure of SSBE makes it possible for us to bound all moments of the numerical solution, under Assumption 3.1. We deal first with the discrete approximation and then with a continuous-time extension.

LEMMA 3.7. *Suppose Assumption* 3.1 *holds and let* $\Delta t\le\Delta t_c<1/(2\beta)$*, where* β *is defined in* (3.4)*. Then for each* $p\ge 2$ *there exists a* $C=C(p,T)>0$ *(independent of* Δt*) such that for the SSBE method* (3.8)–(3.9)

$$\mathbb{E}\sup_{n\Delta t\in[0,T]}|Y_n|^{2p}\le C.$$

Proof. In the following we assume that N and M are positive integers such that $N\Delta t\le M\Delta t\le T$. From (3.3) and (3.8) we have

$$\begin{aligned}|Y_n^\star|^2&=\langle Y_n,Y_n^\star\rangle+\Delta t\langle f(Y_n^\star),Y_n^\star\rangle\\&\le\tfrac12|Y_n|^2+\tfrac12|Y_n^\star|^2+\Delta t(\alpha+\beta|Y_n^\star|^2).\end{aligned}$$

Thus

$$|Y_n^\star|^2\le\frac{|Y_n|^2+2\alpha\Delta t}{1-2\beta\Delta t}.\tag{3.25}$$

From (3.9) and (3.25) we have

$$\begin{aligned}|Y_{n+1}|^2&\le|Y_n|^2+\frac{2\beta\Delta t}{1-2\beta\Delta t}|Y_n|^2+\frac{2\alpha\Delta t}{1-2\beta\Delta t}\\&\quad+2\langle Y_n^\star,g(Y_n^\star)\Delta W_n\rangle+|g(Y_n^\star)\Delta W_n|^2.\end{aligned}$$

Summing, and using the notation $K=(1-2\beta\Delta t)^{-1}$, we obtain

$$\begin{aligned}|Y_N|^2&\le|Y_0|^2+2\beta\Delta tK\sum_{j=0}^{N-1}|Y_j|^2+2\alpha\Delta tNK\\&\quad+2\sum_{j=0}^{N-1}\langle Y_j^\star,g(Y_j^\star)\Delta W_j\rangle+\sum_{j=0}^{N-1}|g(Y_j^\star)\Delta W_j|^2.\end{aligned}$$

Raising both sides to the power p we have

$$\frac{1}{5^{p-1}}|Y_N|^{2p} \le |Y_0|^{2p} + (2\beta\Delta t K)^p \left(\sum_{j=0}^{N-1} |Y_j|^2\right)^p + (2\alpha T K)^p + 2^p \left|\sum_{j=0}^{N-1} \langle Y_j^\star, g(Y_j^\star)\Delta W_j\rangle\right|^p + \left(\sum_{j=0}^{N-1} |g(Y_j^\star)\Delta W_j|^2\right)^p$$

$$\le |Y_0|^{2p} + (2\beta K)^p T^{p-1}\Delta t \sum_{j=0}^{N-1} |Y_j|^{2p} + (2\alpha T K)^p + 2^p \left|\sum_{j=0}^{N-1} \langle Y_j^\star, g(Y_j^\star)\Delta W_j\rangle\right|^p + N^{p-1}\sum_{j=0}^{N-1} |g(Y_j^\star)\Delta W_j|^{2p}. \tag{3.26}$$

Now

$$\mathbb{E}\left[\sup_{0\le N\le M} \sum_{j=0}^{N-1} |Y_j|^{2p}\right] = \sum_{j=0}^{M-1} \mathbb{E}|Y_j|^{2p}. \tag{3.27}$$

Also, letting $C = C(p,T)$ be a constant that may change line by line,

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\le N\le M} \sum_{j=0}^{N-1} |g(Y_j^\star)\Delta W_j|^{2p}\right] &= \mathbb{E}\sum_{j=0}^{M-1} |g(Y_j^\star)\Delta W_j|^{2p} \\
&\le \sum_{j=0}^{M-1} \mathbb{E}|g(Y_j^\star)|^{2p}\mathbb{E}|\Delta W_j|^{2p} \\
&\le C\Delta t^p \sum_{j=0}^{M-1} \mathbb{E}[\alpha + \beta|Y_j^\star|^2]^p \\
&\le C\Delta t^p \sum_{j=0}^{M-1} \mathbb{E}[\alpha^p + \beta^p|Y_j^\star|^{2p}] \\
&\le C\Delta t^{p-1} + C\Delta t^p \sum_{j=0}^{M-1} \mathbb{E}[|Y_j|^2 + 2\alpha\Delta t|]^p \\
&\le C\Delta t^{p-1} + C\Delta t^p \sum_{j=0}^{M-1} \mathbb{E}|Y_j|^{2p}, \qquad (3.28)
\end{aligned}$$

where we have used (3.3) and (3.25). Finally, using the Burkholder–Davis–Gundy inequality [14],

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\le N\le M} \left|\sum_{j=0}^{N-1} \langle Y_j^\star, g(Y_j^\star)\Delta W_j\rangle\right|^p\right] &\le C\mathbb{E}\left[\sum_{j=0}^{M-1} |Y_j^\star|^2|g(Y_j^\star)|^2\Delta t\right]^{p/2} \\
&\le C(\Delta t)^{p/2} M^{p/2-1}\mathbb{E}\sum_{j=0}^{M-1} |Y_j^\star|^p(\alpha + \beta|Y_j^\star|^2)^{p/2} \\
&\le C\Delta t \sum_{j=0}^{M-1} [1 + \mathbb{E}|Y_j^\star|^{2p}]
\end{aligned}$$

$$\leq C\Delta t \sum_{j=0}^{M-1} [1 + \mathbb{E}(2\alpha\Delta t + |Y_j|^2)^p]$$

$$\leq C + C\Delta t \sum_{j=0}^{M-1} \mathbb{E}|Y_j|^{2p}. \tag{3.29}$$

Combining (3.26)–(3.29) we obtain

$$\mathbb{E}\left[\sup_{0\leq N\leq M} |Y_N|^{2p}\right] \leq C + C\Delta t \sum_{j=0}^{M-1} \mathbb{E}|Y_j|^{2p} \leq C + C\Delta t \sum_{j=0}^{M-1} \mathbb{E}\left[\sup_{0\leq N\leq j} \mathbb{E}|Y_N|^{2p}\right].$$

Using the discrete-type Gronwall inequality (see, for example, [13]) and noting that $M\Delta t \leq T$, we obtain

$$\mathbb{E}\left[\sup_{0\leq N\leq M} |Y_N|^{2p}\right] \leq Ce^{C\Delta tM} \leq Ce^{CT},$$

and the desired result follows. ∎

COROLLARY 3.8. *Suppose Assumption* 3.1 *holds and let* $\Delta t \in (0, \Delta t_c), \Delta t_c < 1/(2\beta)$*, where* β *is defined in* (3.4)*. Let* $p \geq 2$*. Then there exists a continuous-time extension* $\overline{Y}(t)$ *of the SSBE solution* $\{Y_k\}$ *and a constant* $C = C(p, T) > 0$ *(independent of* Δt*) such that*

$$\mathbb{E} \sup_{0\leq t\leq T} |\overline{Y}(t)|^{2p} \leq C.$$

Proof. We know that SSBE can be regarded as EM applied to the modified SDE (3.16). Hence, we may define $\overline{Y}(t)$ using (2.2)–(2.3) with f, g replaced by $f_{\Delta t}$, $g_{\Delta t}$ and X, $\overline{X}, X_k$ replaced by Y, $\overline{Y}, Y_k$. By definition we have, for $t_n = n\Delta t$,

$$\overline{Y}(t_n + s) = Y_n + sf_{\Delta t}(Y_n) + g_{\Delta t}(Y_n)\Delta W_n(s), \quad s \in [0, \Delta t),$$

where

$$\Delta W_n(s) := W(t_n + s) - W(t_n).$$

However, $Y_n^\star = Y_n + \Delta t f_{\Delta t}(Y_n)$ and so, for $a = s/\Delta t$, we have

$$\overline{Y}(t_n + s) = aY_n^\star + (1 - a)Y_n + g_{\Delta t}(Y_n)\Delta W_n(s), \qquad s \in [0, \Delta t).$$

Since $\Delta t \leq \Delta t_c < 1/(2\beta)$, it follows from (3.25) that

$$|\overline{Y}(t_n + s)|^2 \leq C[1 + |Y_n|^2 + |g_{\Delta t}(Y_n)\Delta W_n(s)|^2].$$

Thus

$$\sup_{0\leq t\leq T} |\overline{Y}(t)|^{2p} \leq \sup_{0\leq n\Delta t\leq T} \sup_{0\leq s\leq \Delta t} |\overline{Y}(t_n + s)|^{2p}$$

$$\leq \sup_{0\leq n\Delta t\leq T} \sup_{0\leq s\leq \Delta t} C[1 + |Y_n|^{2p} + |g_{\Delta t}(Y_n)\Delta W_n(s)|^{2p}]$$

$$\leq C[1 + \sup_{0\leq n\Delta t\leq T} |Y_n|^{2p} + \sup_{0\leq s\leq \Delta t} \sum_{j=0}^{N} |g_{\Delta t}(Y_j)\Delta W_j(s)|^{2p}], \tag{3.30}$$

where $0 \le N\Delta t \le T$. Now, using Doob's martingale inequality [14] and (3.15)

$$\begin{aligned}\mathbb{E}\sup_{0\le s\le \Delta t}|g_{\Delta t}(Y_j)\Delta W_j(s)|^{2p} &\le C\mathbb{E}|g_{\Delta t}(Y_j)\Delta W_j(\Delta t)|^{2p}\\ &\le C\mathbb{E}|g_{\Delta t}(Y_j)|^{2p}\mathbb{E}|\Delta W_j(\Delta t)|^{2p}\\ &\le C\Delta t^p[1+\mathbb{E}|Y_j|^{2p}]\\ &\le C\Delta t, \qquad (3.31)\end{aligned}$$

where C is a universal constant, independent of Δt. Since $N\Delta t \le T$, combining Lemma 3.7, (3.30), and (3.31) gives the desired result. □

Proof of Theorem 3.3. The proof now follows from an application of the triangle inequality: the SSBE method has a solution close to the solution of an SDE with modified vector fields, and the solution of this SDE in turn is close to that of the original SDE. More precisely, we may use Corollary 3.8 to define $\overline{Y}(t)$ and bound $\mathbb{E}\sup_{0\le t\le T}|\overline{Y}(t)|^p$ and Lemma 3.6 to bound $\mathbb{E}\sup_{0\le t\le T}|y_{\Delta t}(t)|^p$. We also know from Lemma 3.4 that $f_{\Delta t}$ and $g_{\Delta t}$ are uniformly locally Lipschitz for small Δt. Hence, we may follow the proof of Theorem 2.2 to give

$$\lim_{\Delta t\to 0}\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Y}(t)-y_{\Delta t}(t)|^2\right]=0.$$

Combining this with (3.19) in Lemma 3.6 via the triangle inequality gives the result. □

4. Convergence rates. In this section we show that by augmenting Assumption 3.1 with the condition that f behaves polynomially, it is possible to establish a rate of convergence. The rate is optimal, agreeing with the standard theory for the explicit EM scheme in the globally Lipschitz case. The work of [7] also yields optimal rates of almost sure convergence for the EM scheme, under conditions on the vector fields similar to ours.

ASSUMPTION 4.1. *There exist constants $D \in \mathbb{R}^+$ and $q \in \mathbb{Z}^+$ such that for all $a, b \in \mathbb{R}^m$,*

$$|f(a)-f(b)|^2 \le D\,(1+|a|^q+|b|^q)\,|a-b|^2. \qquad (4.1)$$

To obtain a convergence rate for EM we require the following moment bound assumption.

ASSUMPTION 4.2. *The SDE and EM solutions satisfy*

$$\mathbb{E}\sup_{0\le t\le T}|y(t)|^p, \quad \mathbb{E}\sup_{0\le t\le T}|X(t)|^p, \quad \mathbb{E}\sup_{0\le t\le T}|\overline{X}(t)|^p<\infty \qquad \forall p\ge 1.$$

Throughout the following analysis, K and u denote generic positive real and integer constants whose values may change between occurrences. Before obtaining a convergence rate for EM, we give the following lemma.

LEMMA 4.3. *Under Assumptions* 3.1, 4.1, and 4.2, *for any even integer $r \ge 2$, there exists a constant $E = E(r)$ such that*

$$\sup_{0\le t\le T}\mathbb{E}|X(t)-\overline{X}(t)|^r \le E\Delta t^{r/2}.$$

Proof. Let $t \in [k\Delta t, (k+1)\Delta t)$. Then

$$\begin{aligned}|X(t) - \overline{X}(t)|^r &= |(t-t_k)f(X_k) + g(X_k)(W(t) - W(t_k))|^r \\ &\le 2^r \left(\Delta t^r |f(X_k)|^r + |g(X_k)|^r |W(t) - W(t_k)|^r\right).\end{aligned}$$

Hence, for some $E = E(r)$,

$$\mathbb{E}|X(t) - \overline{X}(t)|^r \le E\left(\Delta t^r \left[1 + \mathbb{E}\sup_{0\le t\le T}|X(t)|^u\right] + \left[1 + \mathbb{E}\sup_{0\le t\le T}|X(t)|^u\right](t-t_k)^{r/2}\right).$$

Since $t - t_k \le \Delta t$, the result follows by redefinition of E. ∎

THEOREM 4.4. *Under Assumptions* 3.1, 4.1, *and* 4.2 *the EM solution* (2.1) *with continuous-time extension* (2.2) *satisfies*

$$\mathbb{E}\left[\sup_{0\le t\le T}|\overline{X}(t) - y(t)|^2\right] = O(\Delta t).$$

Proof. Using (2.2) and

$$y(t) = y_0 + \int_0^t f(y(s))ds + \int_0^t g(y(s))dW(s), \tag{4.2}$$

and letting $e(t) := y(t) - \overline{X}(t)$, the Itô formula gives

$$\begin{aligned}|e(t)|^2 &= \int_0^t 2\langle f(y(s)) - f(X(s)), e(s)\rangle ds + \int_0^t |g(y(s)) - g(X(s))|^2 ds \\ &\quad + M(t) \\ &\le \int_0^t \left(2\langle f(y(s)) - f(\overline{X}(s)), e(s)\rangle + c|y(s) - X(s)|^2\right) ds \\ &\quad + \int_0^t 2\langle f(\overline{X}(s)) - f(X(s)), e(s)\rangle ds \\ &\quad + M(t) \\ &\le \int_0^t 2\mu|e(s)|^2 + 2c|e(s)|^2 + 2c|X(s) - \overline{X}(s)|^2 ds \\ &\quad + \int_0^t |f(\overline{X}(s)) - f(X(s))|^2 + |e(s)|^2 ds \\ &\quad + M(t) \\ &\le (1 + 2(\mu + c))\int_0^t |e(s)|^2 ds + \int_0^t K\left(1 + |\overline{X}(s)|^q + |X(s)|^q\right)|X(s) - \overline{X}(s)|^2 ds \\ &\quad + M(t),\end{aligned}$$

where

$$M(t) = \int_0^t 2\langle e(s), (g(y(s)) - g(X(s)))dW(s)\rangle.$$

Using Cauchy–Schwarz and Lemma 4.3 with $r = 4$

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] &\le (1+2(\mu+c))\int_0^t \mathbb{E}|e(s)|^2 ds \\ &\quad + \int_0^t K\left(\mathbb{E}\left(1+|\overline{X}(s)|^q+|X(s)|^q\right)^2\mathbb{E}|X(s)-\overline{X}(s)|^4\right)^{1/2} ds + m(t) \\ &\le (1+2(\mu+c))\int_0^t \mathbb{E}|e(s)|^2 ds + K\Delta t\int_0^t \mathbb{E}\left(1+|\overline{X}(s)|^{2q}+|X(s)|^{2q}\right) ds \\ &\quad + m(t) \\ &\le (1+2(\mu+c))\int_0^t \mathbb{E}|e(s)|^2 ds + K\Delta t + m(t), \qquad (4.3)\end{aligned}$$

where

$$m(t) = \mathbb{E}\left[\sup_{0\le s\le t}|M(s)|\right].$$

From the Burkholder–Davis–Gundy inequality,

$$\begin{aligned} m(t) &\le 16\mathbb{E}\left[\int_0^t |e(s)|^2|g(y(s))-g(X(s))|^2 ds\right]^{1/2} \\ &\le 16\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\int_0^t c|y(s)-X(s)|^2 ds\right]^{1/2} \\ &\le \tfrac{1}{2}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] + 128c\mathbb{E}\int_0^t |y(s)-X(s)|^2 ds \\ &\le \tfrac{1}{2}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] + 256c\int_0^t [\mathbb{E}|e(s)|^2 + \mathbb{E}|\overline{X}(s)-X(s)|^2] ds \\ &\le \tfrac{1}{2}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] + 256c\int_0^t \mathbb{E}|e(s)|^2 ds + K\Delta t. \end{aligned}$$

Hence, in (4.3),

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] &\le 2(1+2(\mu+c)+256c)\int_0^t \mathbb{E}|e(s)|^2 ds + K\Delta t \\ &\le 2(1+2(\mu+c)+256c)\int_0^t \mathbb{E}\left[\sup_{0\le r\le s}|e(r)|^2\right] ds + K\Delta t.\end{aligned}$$

The result follows from the Gronwall inequality. ∎

Note that Theorem 4.4 requires moment bounds on the numerical solution (Assumption 4.2). We know that SSBE has bounded moments under Assumption 3.1, and hence we would expect to get an analogous convergence result for this method without requiring Assumption 4.2. To obtain such a result we first establish further properties of $f_{\Delta t}$ and $g_{\Delta t}$ under Assumptions 3.1 and 4.1.

LEMMA 4.5. *Under Assumptions* 3.1 *and* 4.1, *for* $\Delta t \le \Delta t_c < 1/(2\beta)$, *where* β *is defined in* (3.4), *there exist constants* $c', D' \in \mathbb{R}^+$ *and* $q' \in \mathbb{Z}^+$ *such that for all* $a, b \in \mathbb{R}^m$

$$|f_{\Delta t}(a) - f_{\Delta t}(b)|^2 \le D'\left(1+|a|^q+|b|^q\right)|a-b|^2, \qquad (4.4)$$

$$|f(a) - f_{\Delta t}(a)|^2 \le c'\left(1+|a|^{q'}\right)\Delta t^2, \qquad (4.5)$$

$$|g(a) - g_{\Delta t}(a)|^2 \le c'\left(1+|a|^{q'}\right)\Delta t^2. \qquad (4.6)$$

Proof. See Appendix A. ∎

LEMMA 4.6. *Under Assumptions* 3.1 *and* 4.1 *the solution* $y_{\Delta t}(t)$ *of the modified SDE* (3.16) *satisfies*

$$\mathbb{E}\left[\sup_{0\le t\le T}|y_{\Delta t}(t)-y(t)|^2\right]=O(\Delta t^2).$$

Proof. Using

$$y_{\Delta t}(t)=y_0+\int_0^t f_{\Delta t}(y_{\Delta t}(s))ds+\int_0^t g_{\Delta t}(y_{\Delta t}(s))dW(s),$$

and (4.2), and letting $e(t):=y(t)-y_{\Delta t}(t)$, the Itô formula gives

$$\begin{aligned}
|e(t)|^2 &= \int_0^t 2\langle f(y(s))-f_{\Delta t}(y_{\Delta t}(s)),e(s)\rangle ds+\int_0^t |g(y(s))-g_{\Delta t}(y_{\Delta t}(s))|^2 ds\\
&\quad + M(t)\\
&= \int_0^t 2\langle f(y(s))-f(y_{\Delta t}(s))+f(y_{\Delta t}(s))-f_{\Delta t}(y_{\Delta t}(s)),e(s)\rangle ds\\
&\quad + \int_0^t |g(y(s))-g_{\Delta t}(y_{\Delta t}(s))|^2 ds\\
&\quad + M(t)\\
&\le \int_0^t [2\mu|e(s)|^2+|f(y_{\Delta t}(s))-f_{\Delta t}(y_{\Delta t}(s))|^2+|e(s)|^2]ds\\
&\quad + 2\int_0^t |g(y(s))-g(y_{\Delta t}(s))|^2+|g(y_{\Delta t}(s))-g_{\Delta t}(y_{\Delta t}(s))|^2 ds\\
&\quad + M(t)\\
&\le K\int_0^t |e(s)|^2\,ds+K\Delta t^2\int_0^t\left(1+|y_{\Delta t}(s)|^{q'}\right)ds\\
&\quad + M(t),
\end{aligned}$$

where we used Lemma 4.5 and

$$M(t)=\int_0^t 2\langle e(s),(g(y(s))-g_{\Delta t}(y_{\Delta t}(s)))dW(s)\rangle.$$

Hence,

$$\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right]\le K\int_0^t \mathbb{E}|e(s)|^2\,ds+K\Delta t^2\int_0^t \mathbb{E}\left[1+|y_{\Delta t}(s)|^{q'}\right]ds+m(t),$$

where

$$m(t)=\mathbb{E}\left[\sup_{0\le s\le t}|M(s)|\right].$$

Since $y_{\Delta t}(s)$ has bounded moments, we have

$$\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right]\le K\int_0^t \mathbb{E}|e(s)|^2\,ds+K\Delta t^2+m(t). \tag{4.7}$$

However, in the same way as in the proof of Theorem 4.4, we can show

$$m(t) \le \tfrac{1}{2}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] + 128\mathbb{E}\int_0^t |g(y(s)) - g_{\Delta t}(y_{\Delta t}(s))|^2 ds,$$

while

$$\begin{aligned}\mathbb{E}\int_0^t |g(y(s)) - g_{\Delta t}(y_{\Delta t}(s))|^2 ds &\le 2\mathbb{E}\int_0^t |g(y(s)) - g(y_{\Delta t}(s))|^2 \\ &\quad + |g(y_{\Delta t}(s)) - g_{\Delta t}(y_{\Delta t}(s))|^2 ds \\ &\le K\int_0^t \mathbb{E}|e(s)|^2\, ds + K\Delta t^2.\end{aligned}$$

In (4.7) we therefore have

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le s\le t}|e(s)|^2\right] &\le K\int_0^t \mathbb{E}|e(s)|^2\, ds + K\Delta t^2 \\ &\le K\int_0^t \mathbb{E}\left[\sup_{0\le r\le s}|e(r)|^2\right] ds + K\Delta t^2.\end{aligned}$$

The result follows from the Gronwall inequality. □

We may now prove a convergence result for SSBE.

THEOREM 4.7. *Consider the SSBE method* (3.8)–(3.9) *applied to the SDE* (1.1) *under Assumptions* 3.1 *and* 4.1. *There exists a continuous-time extension* $\overline{Y}(t)$ *of the numerical solution (so that* $\overline{Y}(t_k) = Y_k$*) for which*

$$\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Y}(t) - y(t)|^2\right] = O(\Delta t).$$

Proof. We know that SSBE can be regarded as EM applied to the modified SDE (3.16). Lemmas 3.6 and 3.7 and Corollary 3.8 show that $y_{\Delta t}(t)$, $Y(t)$, and $\overline{Y}(t)$ have bounded moments. Hence, copying the proof of Theorem 4.4 we may conclude that

$$\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Y}(t) - y_{\Delta t}(t)|^2\right] = O(\Delta t).$$

Combining this with Lemma 4.6 via the triangle inequality gives the required result. □

5. Backward Euler. The SSBE method (3.8)–(3.9) is a stochastic extension of the deterministic backward Euler method. Another, perhaps more natural, extension of backward Euler is given by $Z_0 = y_0$ and

$$Z_{k+1} = Z_k + \Delta t f(Z_{k+1}) + g(Z_k)\Delta W_k. \tag{5.1}$$

Indeed, this implicit method has appeared frequently in the literature—it is a member of the *family of implicit Euler schemes* [12, section 12.2] or the stochastic theta method class [10] and is sometimes called the semi-implicit Euler method [2]. We will refer to the method (5.1) as simply the backward Euler (BE) method for (1.1). As mentioned at the end of section 1, our convergence result, Theorem 5.3 below, is closely related to that of [11, Theorem 2.4].

The BE method (5.1) requires an implicit equation to be solved. Under Assumption 3.1, the homotopy argument in the proof of Lemma 3.4 shows that for $2\mu\Delta t < 1$ a unique solution exists with probability one. The next lemma points out a useful connection between BE and SSBE.

LEMMA 5.1. *Let $\{Y_k\}$ and $\{Z_k\}$ denote the SSBE and BE solutions, given by* (3.8)–(3.9) *and* (5.1), *respectively. Under Assumption* 3.1, *if $Y_0 = Z_0 - \Delta t f(Z_0)$, then*

$$Z_k = Y_k + \Delta t f_{\Delta t}(Y_k) \qquad \forall k \geq 0. \tag{5.2}$$

Proof. Let $Q_k^\star = Z_k$ and $Q_k = Z_k - \Delta t f(Z_k)$, where $\{Z_k\}$ is the BE solution (5.1). Then

$$Q_k^\star = Q_k + \Delta t f(Q_k^\star)$$

and, using (5.1),

$$Q_{k+1} = Z_{k+1} - \Delta t f(Z_{k+1}) = Q_k^\star + g(Q_k^\star)\Delta W_k.$$

Hence, $\{Q_k\}$ is precisely the SSBE solution. This gives $Y_k = Z_k - \Delta t f(Z_k)$. The relation (5.2) then follows immediately from Lemma 3.4. ∎

Lemma 5.1 shows that the BE solution can be regarded as an $O(\Delta t)$ perturbation of the SSBE solution. We may use this relation between BE and SSBE in order to obtain a convergence result for BE via Theorem 3.3. We first deal with the perturbation to the initial data.

LEMMA 5.2. *Under Assumptions* 3.1 *and* 4.1, *if $y(t)$ and $z(t)$ are solutions of the SDE* (1.1) *with initial conditions such that*

$$\mathbb{E}|y(0)|^p,\ \mathbb{E}|z(0)|^p \leq \infty \qquad \forall p \geq 1,$$

then, for some constant M,

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |y(t) - z(t)|^2\right] \leq M\mathbb{E}|y(0) - z(0)|^2.$$

Proof. Letting $e(t) := y(t) - z(t)$ and applying the Itô formula to $|e(t)|^2$, the inequality can be obtained by following the process used in the proofs of Theorem 4.4 and Lemma 4.6. ∎

THEOREM 5.3. *Consider the BE method* (5.1) *applied to the SDE* (1.1) *under Assumptions* 3.1 *and* 4.1. *There exists a continuous-time extension $\overline{Z}(t)$ of the numerical solution (so that $\overline{Z}(t_k) = Z_k$) for which*

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |\overline{Z}(t) - y(t)|^2\right] = O(\Delta t).$$

Proof. Let $\overline{Y}(t)$ denote the continuous-time extension to SSBE defined in Theorem 3.3, with initial data $\overline{Y}(0) = y_0 - \Delta t f(y_0)$. Also, let $\overline{Z}(t) = \overline{Y}(t) + \Delta t f_{\Delta t}(\overline{Y}(t))$, so that, from Lemma 5.1, $\overline{Z}(t)$ is a continuous-time extension to the BE solution with $\overline{Z}(0) = y_0$. We let $\widehat{y}_{\Delta t}(t)$ denote the solution to (1.1), with initial data $y_{\Delta t}(0) = y_0 - \Delta t f(y_0)$.

From Lemma 5.2 we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |y(t) - \widehat{y}_{\Delta t}(t)|^2\right] \leq M\Delta t^2\mathbb{E}|f(y_0)|^2 = O(\Delta t^2). \tag{5.3}$$

Also, the SSBE convergence result in Theorem 4.7 shows that

$$\mathbb{E}\left[\sup_{0\le t\le T}|\widehat{y}_{\Delta t}(t)-\overline{Y}(t)|^2\right]=O(\Delta t). \tag{5.4}$$

Further,

$$\mathbb{E}\left[\sup_{0\le t\le T}|\overline{Y}(t)-\overline{Z}(t)|^2\right]\le \Delta t^2\mathbb{E}\left[\sup_{0\le t\le T}|f_{\Delta t}(\overline{Y}(t))|^2\right]=O(\Delta t^2), \tag{5.5}$$

because, by Lemma 4.5, $f_{\Delta t}$ is polynomially bounded. Combining (5.3)–(5.5) completes the proof. □

6. Summary. Our aim in this work was to extend strong mean square convergence theory for numerical SDE simulations beyond the realm of globally Lipschitz problems. The only previous published work in this area that we are aware of is [11]. We gave a strong convergence theorem for EM in the case where the vector fields are locally Lipschitz (e.g., C^1) and moment bounds are available. This style of analysis is useful whenever moment bounds can be established, both for the EM method and for other methods that can be shown to be "close" to EM. In general, it is not clear when such moment bounds can be expected to hold for explicit methods with $f, g \in C^1$. However, for an implicit variant of EM, we obtained bounds on all moments in the case where the diffusion coefficient is globally Lipschitz but the drift coefficient satisfies only a one-sided Lipschitz condition. Then, by interpreting the implicit method as EM applied to a modified SDE we were able to get a strong convergence result. We then considered the case where it is further assumed that f behaves like a polynomial. If all moments are bounded, then EM can be shown to converge strongly at the optimal rate, again assuming moment bounds. Moment bounds can be established for two different implicit variants of EM, allowing us to show that these implicit methods converge at the optimal rate. One of these convergence results is comparable to the main result in [11]—we use a stronger error measure but require a more restrictive assumption on the growth of the drift coefficient.

The methods of analysis could be extended to other implicit methods, such as the stochastic theta method with $\theta \in [1/2, 1]$. Such schemes, especially their split-step variants, may be of practical interest for Hamiltonian problems perturbed by damping and/or noise in the case $\theta = 1/2$.

Appendix A. Proofs of Lemmas 3.4 and 4.5.

Proof of Lemma 3.4. Existence and uniqueness for (3.10) can be proved via a contraction mapping theorem, which also establishes the C^1 smoothness of $f_{\Delta t}$ and $F_{\Delta t}$ and the convergence property of $f_{\Delta t}$; see [5, 17]. The smoothness and convergence properties of $g_{\Delta t}$ follow from $g_{\Delta t}(\cdot) = g(F_{\Delta t}(\cdot))$.

An alternative proof of uniqueness for (3.10) via a homotopy argument is given in [9, Theorem 14.2]. We repeat the homotopy construction here, as it will be used to obtain the bound (3.12).

Suppose $h = h(\tau)$ satisfies

$$h = d + \Delta t f(h) + (\tau - 1)\Delta t f(d), \tag{A.1}$$

where τ is our homotopy parameter. For $\tau = 1$, h solves (3.10). For $\tau = 0$ we have

$$h - d = \Delta t\,(f(h) - f(d))$$

and so, using Assumption 3.1,

$$|h-d|^2 = \Delta t\langle h-d, f(h)-f(d)\rangle \le \Delta t\mu |h-d|^2.$$

Note that $\beta > \mu$ so that $2\Delta t\mu < 1$. It follows that $h = d$ is the unique solution to (A.1) when $\tau = 0$. Differentiating (A.1) with respect to τ gives

$$\dot{h} = \Delta t \frac{\partial f}{\partial y}(h)\dot{h} + \Delta t f(d).$$

So

$$|\dot{h}|^2 - \Delta t \left\langle \dot{h}, \frac{\partial f}{\partial y}(h)\dot{h} \right\rangle = \Delta t\langle \dot{h}, f(d)\rangle. \tag{A.2}$$

Setting $a - b = \epsilon u$ in (3.1) and letting $\epsilon \to 0$, we see that

$$\left\langle u, \frac{\partial f}{\partial y}(b)u \right\rangle \le \mu |u|^2 \quad \text{for any } u, b \in \mathbb{R}^m.$$

Hence, in (A.2),

$$|\dot{h}|^2 - \Delta t\mu |\dot{h}|^2 \le \Delta t|\dot{h}|\,|f(d)|.$$

So

$$|\dot{h}| \le \Delta t \frac{|f(d)|}{1-\Delta t\mu}.$$

It follows that $h(\tau)$ exists uniquely for all $\tau > 0$ and

$$|h(1)-d| = \left| \int_0^1 \dot{h}(s)\,ds \right| \le \Delta t \frac{|f(d)|}{1-\Delta t\mu},$$

which establishes (3.12).

To obtain (3.13) we note that if $c^{(1)} = d^{(1)} + \Delta t f(c^{(1)})$ and $c^{(2)} = d^{(2)} + \Delta t f(c^{(2)})$, then

$$|c^{(1)}-c^{(2)}|^2 - \Delta t\langle f(c^{(1)}) - f(c^{(2)}), c^{(1)} - c^{(2)}\rangle = \langle d^{(1)} - d^{(2)}, c^{(2)} - c^{(2)}\rangle$$

and so, using Assumption 3.1,

$$(1-\Delta t\mu)|c^{(1)}-c^{(2)}|^2 \le \frac{1}{2}|d^{(1)}-d^{(2)}|^2 + \frac{1}{2}|c^{(1)}-c^{(2)}|^2,$$

which gives (3.13).

Next, note from the implicit definition (3.11) that $f_{\Delta t}(a)$ is equivalent to $f(a + \Delta t f_{\Delta t}(a))$. Using (3.1) we thus have

$$\langle f_{\Delta t}(a) - f_{\Delta t}(b), a + \Delta t f_{\Delta t}(a) - b - \Delta t f_{\Delta t}(b)\rangle \le \mu |a + \Delta t f_{\Delta t}(a) - b - \Delta t f_{\Delta t}(b)|^2.$$

Hence,

$$\langle f_{\Delta t}(a) - f_{\Delta t}(b), a-b\rangle + \Delta t|f_{\Delta t}(a) - f_{\Delta t}(b)|^2 \le \mu|a-b|^2 + 2\mu\langle a-b, f_{\Delta t}(a) - f_{\Delta t}(b)\rangle\Delta t + \mu\Delta t^2|f_{\Delta t}(a) - f_{\Delta t}(b)|^2,$$

and (3.14) follows.

The global Lipschitz property of $g_{\Delta t}$ follows from (3.13).

Finally, we use $f_{\Delta t}(a) = f(a + \Delta t f_{\Delta t}(a))$ and (3.3) to give

$$\langle f_{\Delta t}(a), a + \Delta t f_{\Delta t}(a)\rangle \le \alpha + \beta|a + \Delta t f_{\Delta t}(a)|^2.$$

Hence

$$(1 - 2\beta\Delta t)\langle f_{\Delta t}(a), a\rangle \le \alpha + \beta|a|^2 + [\beta\Delta t^2 - \Delta t]|f_{\Delta t}(a)|^2 \le \alpha + \beta|a|^2.$$

Since $g_{\Delta t}$ is globally Lipschitz, the inequality (3.15) follows. □

Proof of Lemma 4.5. Recall from Lemma 3.4 that $f_{\Delta t}(a) := f(F_{\Delta t}(a))$, where $F_{\Delta t}$ is globally Lipschitz. Hence,

$$\begin{aligned}|f_{\Delta t}(a) - f_{\Delta t}(b)|^2 &= |f(F_{\Delta t}(a)) - f(F_{\Delta t}(b))|^2 \\ &\le D\,(1 + |F_{\Delta t}(a)|^q + |F_{\Delta t}(b)|^q)\,|F_{\Delta t}(a) - F_{\Delta t}(b)|^2 \\ &\le D'\,(1 + |a|^q + |b|^q)\,|a-b|^2.\end{aligned}$$

Next, it follows from the equivalence of $f_{\Delta t}(a)$ in (3.11) and $f(a + \Delta t f_{\Delta t}(a))$ that

$$|f(a) - f_{\Delta t}(a)|^2 \le D\,(1 + |a|^q + |a + \Delta t f_{\Delta t}(a)|^q)\,\Delta t^2|f_{\Delta t}(a)|^2.$$

From (3.12), $|f_{\Delta t}(a)| \le 2|f(a)|$, and hence we obtain (4.5). A similar argument gives (4.6). □

REFERENCES

[1] K. Burrage and J. C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM J. Numer. Anal., 16 (1979), pp. 46–57.

[2] K. Burrage and T. Tian, *A note on the stability properties of the Euler methods for solving stochastic differential equations*, New Zealand J. Math., 29 (2000), pp. 115–127.

[3] J. C. Butcher, *A stability property of implicit Runge–Kutta methods*, BIT, 15 (1975), pp. 358–361.

[4] G. Dahlquist, *Error analysis for a class of methods for stiff non-linear initial value problems*, in Numerical Analysis, Lecture Notes in Math. 506, G. A. Watson, ed., Springer-Verlag, Berlin, 1976, pp. 60–72.

[5] K. Dekker and J. G. Verwer, *Stability of Runge–Kutta Methods for Stiff Nonlinear Equations*, North-Holland, Amsterdam, 1984.

[6] G. Fleury and P. Bernard, *Convergence of numerical schemes for stochastic differential equations*, Monte Carlo Methods Appl., 7 (2001), pp. 35–44.

[7] I. Gyöngy, *A note on Euler's approximations*, Potential Anal., 8 (1998), pp. 205–216.

[8] I. Gyöngy and N. Krylov, *Existence of strong solutions for Itô's stochastic equations via approximations*, Probab. Theory Related Fields, 105 (1996), pp. 143–158.

[9] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II*, Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.

[10] D. J. Higham, *Mean-square and asymptotic stability of the stochastic theta method*, SIAM J. Numer. Anal., 38 (2000), pp. 753–769.

[11] Y. Hu, *Semi-implicit Euler-Maruyama scheme for stiff stochastic equations*, in Stochastic Analysis and Related Topics V: The Silvri Workshop, Progr. Probab. 38, H. Koerezlioglu, ed., Birkhauser, Boston, 1996, pp. 183–202.

[12] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1999.

[13] X. Mao, *Stability of Stochastic Differential Equations with Respect to Semimartingales*, Pitman Res. Notes Math. Ser. 251, Longman Scientific and Technical, Harlow, UK, 1991.

[14] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, UK, 1997.
[15] J. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and Approximations: Locally Lipschitz Vector Fields and Degenerate Noise*, Tech. Rep. 7, University of Strathclyde, Department of Mathematics, Glasgow, UK, 2001.
[16] H. Schurz, *Stability, Stationarity, and Boundedness of some Implicit Numerical Methods for Stochastic Differential Equations and Applications*, Logos Verlag, Berlin, 1997.
[17] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1064–1076

GALERKIN METHODS IN AGE AND SPACE FOR A POPULATION MODEL WITH NONLINEAR DIFFUSION*

BRUCE P. AYATI† AND TODD F. DUPONT‡

Abstract. We present Galerkin methods in both the age and space variables for an age-dependent population undergoing nonlinear diffusion. The methods presented are a generalization of methods, where the approximation space in age is the space of piecewise constant functions. In this paper, we allow the use of discontinuous piecewise polynomial subspaces of L^2 as the approximation space in age. As in the piecewise constant case, we move the discretization along characteristic lines. The time variable has been left continuous. The methods are shown to be superconvergent in the age variable.

Key words. population dynamics, age-dependence, nonlinear diffusion, Galerkin methods, superconvergence

AMS subject classifications. Primary, 65M60; Secondary, 35Q80, 65M15, 92D25

PII. S0036142900379679

1. Introduction.

We present Galerkin methods in both the age and space variables for a model of an age-dependent population undergoing nonlinear diffusion. The methods presented are a generalization of methods presented in [2], where the approximation space in age was taken to be the space of piecewise constant functions. The use, analysis, and numerical solution of models with dependence on age and time, and of models that also include space, is discussed in [2] and references therein.

In this paper, we allow the use of discontinuous piecewise polynomial subspaces of L^2 as the approximation space in age. As in the piecewise constant case, we move the discretization along characteristic lines. This preserves the important fact that age and time advance together and that the resulting discretization will be dispersion-free.

Some previous numerical methods [8, 9, 12] for age-structured models with spatial diffusion also discretized along characteristics, but they did so simultaneously in age and time and thus imposed the often crippling constraint that the time and age steps be both constant and equal. The difficulty with this approach is twofold. First, the use of constant age and time steps prevents adaptivity of the discretization in age or, especially, time. Second, and more importantly, the coupling of the age and time meshes can cause great losses of efficiency since only rarely will the dynamics in time be on the same scale as the dynamics in age. This is particularly the case when space is involved since sharp moving fronts can require small time steps, whereas the behavior in the age variable can remain relatively smooth. A computational example illustrating the advantages of decoupling age and time is presented in [2] and for the context of *Proteus mirabilis* swarm colony development [6, 14] in Chapter 4 of [1].

The age discretization presented in [8, 9, 12] can be viewed as special cases of the methods presented here and in [2] by setting the time and age meshes to be constant

*Received by the editors October 17, 2000; accepted for publication (in revised form) March 13, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/sinum/40-3/37967.html
†Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455 (bruce@ima.umn.edu).
‡Departments of Computer Science and Mathematics, The University of Chicago, Chicago, IL 60637 (dupont@cs.uchicago.edu). This work made use of MRSEC Shared Facilities and was supported by the National Science Foundation under award DMR-9400379.

and equal and using a backward Euler discretization in time and a piecewise constant finite element space in age.

The importance of allowing different age and time discretizations is perhaps illustrated by the application of the methods of de Roos [3]. These methods have found use in the study of ecological systems such as *Daphnia* (see [4] and the references therein) as well as in theoretical population biology [10, 13]. The methods of de Roos are formulated for the case of time and a variable representing some sort of physiological structure, most simply age, and involve moving the age nodes along characteristics. However, the representation of the approximate solution is probabilistic and not functional, and birth and death are handled differently than in this paper. Even so, it would be interesting to know if an energy analysis could provide a framework for the convergence analysis sought in [5]. The main effect of de Roos's methods is to separate the age and time discretizations, while yielding an approximation that is dispersion-free in age, in order to provide a method that works in practice.

The main purpose of this paper is to provide a description and analysis of the use of higher order finite element spaces in the age variable. The time variable has been left continuous. The use of continuous time simplifies the presentation and analysis of the method as well as emphasizes the independence of the age discretization from any suitable time discretization. The methods are shown to be superconvergent in the age variable. We provide an example system that illustrates some of the benefits of using a higher order approximation space in age as well as highlights some of the interactions between the age and time discretizations in these methods.

2. A continuous model. We consider the age-dependent population model with nonlinear diffusion,

$$\partial_t u + \partial_a u = \nabla \cdot \big(k(x,p)\nabla u\big) - \mu(x,a,p)u, \qquad x \in \Omega,\ a > 0,\ t > 0, \tag{2.1}$$

where ∇ and $\nabla\cdot$ denote the gradient and the divergence, respectively, in x. The function $u(x,a,t)$ represents the distribution of individuals, $\Omega \subset \mathbb{R}^n$ represents the spatial domain, a represents age, and t represents time. The function $\mu > 0$ is the death rate. The total population density, p, is given by

$$p(x,t) = \int_0^\infty u(x,a,t)\,da, \qquad x \in \Omega,\ t > 0. \tag{2.2}$$

We have a birth condition

$$u(x,0,t) = b(x,u(x,\cdot,t)), \qquad x \in \Omega,\ t > 0, \tag{2.3}$$

that is dependent on the entire population distribution. We note that b is an operator whose second argument is a function defined on $\mathbb{R}^+$, where $\mathbb{R}^+$ denotes the nonnegative real numbers. The diffusion arises from the symmetric random motion of each individual (Fickian diffusion). We have a Neumann boundary condition, with ν denoting the outward normal to $\partial\Omega$,

$$k(x,p)\nabla u \cdot \nu = 0, \qquad x \in \partial\Omega,\ a > 0,\ t > 0, \tag{2.4}$$

that represents an isolated habitat. The initial condition is

$$u(x,a,0) = u_0(x,a), \qquad x \in \Omega,\ a > 0. \tag{2.5}$$

Langlais [11] proved the existence of unique nonnegative solutions for the case when k, μ, and β in (2.6) are independent of x, and Ω is bounded. A corresponding treatment for the system (2.1)–(2.5) is beyond the scope of this paper; we will concentrate on the numerical aspects of the problem. Thus we assume existence and uniqueness of smooth, nonnegative solutions.

We make several assumptions.

CONDITION 2.1. *There exists constants C_0 and C_1 such that, for $(x,p) \in \Omega \times \mathbb{R}$, k satisfies $0 < C_0 \le k(x,p) \le C_1$ and μ satisfies $0 < C_0 \le \mu(x,a,p) \le C_1$ for all a.*

CONDITION 2.2. *The functions $k(x,p)$ and $\mu(x,a,p)$ are uniformly Lipschitz continuous with respect to p with Lipschitz constants K_k and K_μ, respectively. The derivative $\partial_p k(x,p)$ exists. The derivative $\partial_a \mu(x,a,p)$ exists, is uniformly bounded by C_1 as a function of all its arguments, and $\|\partial_a \mu(x,\cdot,p)\|_{L^2(\mathbb{R}^+)} \le C_1$ uniformly as a function of x and p.*

CONDITION 2.3. *The birth operator, $b : \Omega \times \left(L^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)\right) \to \mathbb{R}^+$, is of the form*

$$b(x,\varphi(x,\cdot,t)) = \int_0^\infty \beta(x,a,\Phi)\varphi(x,a,t)\ da, \tag{2.6}$$

where $\beta \ge 0$ is the birth rate and Φ is the total population density, i.e., the integral of φ with respect to age. The function β is assumed to be uniformly Lipschitz continuous as a function of Φ. As a function of a, $\beta(x,a,\Phi)$ is in $H^1(\mathbb{R}^+)$, with its H^1-norm bounded independently of x and Φ; and it is also assumed that there is a positive a_{small} and a natural number k_0 such that $\beta(x,\cdot,\Phi)$ is a polynomial of degree at most k_0 on $(0, a_{small})$ with the coefficients bounded independently of x and Φ.

CONDITION 2.4. *The initial condition, $u_0(x,a)$, is bounded and nonnegative, and there exists $\tilde{a}_{max}$ such that $u_0(x,a) = 0$ for $a > \tilde{a}_{max}$.*

We note that the birth operator, b, is uniformly bounded and satisfies the Lipschitz condition

$$\begin{aligned}|b(x,\varphi(x,\cdot,t)) - b(x,\psi(x,\cdot,t))| \\ \le K_b \left(\left(1 + \|\varphi\|_{L^1(\mathbb{R}^+)}\right) \left| \int_0^\infty (\varphi - \psi)\ da \right| + \|\varphi - \psi\|_{H^{-1}(\mathbb{R}^+)} \right),\end{aligned}$$

where $H^{-1}(\mathbb{R}^+)$ is the dual to $H^1(\mathbb{R}^+)$. The condition that β be polynomial in a for small a allows us to avoid some issues associated with the introduction of a new age interval.

Condition 2.4 is technically convenient and seems mild in light of the exponential decay of u in age [2]. A consequence of this condition is that $u(x,a,t)$ is zero if $a > \tilde{a}_{\max} + t$. Since we are dealing with time in a bounded interval in this work, the fact that the age is bounded above means that the behavior of β for very large a is unimportant.

3. An age and space discrete method. Let $D = \partial_t + \partial_a$. We reuse the symbol k to denote the form

$$k(\Phi;\varphi,v) = \int_\Omega k(x,\Phi)\nabla\varphi \cdot \nabla v\ dx;$$

the distinction between the form and $k(x,\Phi)$ should be clear from context. In variational form, for every $t \in \mathbb{R}^+$ and every $v \in H^1(\Omega) \otimes L^2(\mathbb{R}^+)$, we have

$$\int_0^\infty (Du, v) + k(p; u, v) + (\mu u, v)\ da = (b(x, u(x,\cdot,t)) - u(x,0,t), v(x,0)), \tag{3.1}$$

where $(\cdot,\cdot)$ denotes the L^2-inner product over Ω. Note that no regularity in a is required on v because both sides are zero independent of the choice of v, but we find it useful to think of $v(x,0)$ being the trace of v at $a = 0$ when v is smooth.

Let $\mathcal{M}$ denote a finite dimensional subspace of $H^1(\Omega)$. Let $\{a_i\}_{i=0}^{-\infty}$ be a sequence such that $a_0 = \tilde{a}_{\max}$, $0 < a_{i+1} - a_i < \Delta a$, and $a_i \to -\infty$ as $i \to -\infty$. Let $\mathcal{J}$ be the set of a_i's and let $\check{\mathcal{J}}$ denote the set of $-a_i$'s. For a fixed nonnegative integer q, let $\mathcal{C}$ denote the space of all piecewise continuous functions over the partition of $\mathbb{R}$ defined by $\mathcal{J}$ such that $\varphi \in \mathcal{C}$ has the property that φ restricted to (a_i, a_{i+1}) is a polynomial of degree at most q for $i < 0$ and φ is zero on (a_0, ∞). We define a finite dimensional space in age that moves along the characteristic curves, $da/dt = 1$:

$$\mathcal{A}(t) = \left\{\varphi \in L^2(\mathbb{R}^+) : \varphi(\cdot) = \psi(\cdot - t)\big|_{\mathbb{R}^+},\ \psi \in \mathcal{C}\right\}.$$

Note that the dimension of $\mathcal{A}$ increases by $q+1$ as t goes from $-a_i - 0$ to $-a_i + 0$. This discretization will allow the numerical method to be free of numerical dispersion in age. We take $U(\cdot,\cdot,t) \in \mathcal{M} \otimes \mathcal{A}(t)$. For $t \notin \check{\mathcal{J}}$,

$$\begin{aligned}\int_0^\infty (DU, v) + k(P; U, v) + (\mu(x,a,P)U, v)\ da \\ = (b(x, U(x,\cdot,t)) - U(x,0,t), v(x,0,t))\end{aligned} \tag{3.2}$$

for every $v(\cdot,\cdot,t) \in \mathcal{M} \otimes \mathcal{A}(t)$. So that U is defined across points in $\mathcal{J}$, we require U to be a continuous mapping of time into $L^2(\Omega) \otimes L^2(\mathbb{R}^+)$. The total population density is approximated by

$$P(x,t) = \int_0^\infty U(x,a,t)\ da.$$

We want to emphasize that the function U is differentiable in the characteristic direction so that DU makes sense. Functions in $\mathcal{A}(t)$ are discontinuous, but those discontinuities move along characteristic directions.

The system defined by (3.2) is just a set of ordinary differential equations when $t \notin \check{\mathcal{J}}$. Consider $t \in \check{\mathcal{J}}$. Since the part that is nonstandard is related to the age variables, we will suppress the x variables. Let $\{\varphi_i\}$ denote a basis for $\mathcal{C}$, where each φ_i has support in one interval $[a_k, a_{k+1}]$. Then a natural basis for $\mathcal{A}(t)$ is of the form $\{\varphi_i(\cdot - t)\}$; the functions in the basis for $\mathcal{A}$ are restricted to $\mathbb{R}^+$, and there are only a finite number of these that are nontrivial. The function U is expressed as

$$U(a,t) = \sum_i c_i(t)\varphi_i(a-t) \text{ so that } DU = \sum_i c_i'(t)\varphi_i(a-t).$$

While it is natural to look at a set of relations of the form

$$\int_0^\infty DU(a,t)\varphi_j(a-t)da = F(t,U,\varphi_j) + (b(U) - U(0,t))\varphi_j(0-t) \tag{3.3}$$

as a set of ordinary differential equations for the c_i's, there is a difficulty because the coefficient of the vector of c_i''s is singular at a transition. Note that

$$\int_0^\infty DU\varphi_j(a-t)da + U(0,t)\varphi_j(0-t) = \frac{d}{dt}\int_0^\infty U\varphi_j(a-t)da.$$

Thus (3.3) can be written as $m_j' = F(t,U,\varphi_j) + b(U)\varphi_j(0-t)$, where

$$m_j(t) = \int_0^\infty U(a,t)\varphi_j(a-t)da;$$

these are the natural variables. The initial values for the new m_j's added to the system when t crosses a point of $\tilde{\mathcal{J}}$ are clearly zero because of the assumed continuity of U (as a map into $L^2(\mathbb{R}^+)$) at such points. The m_j's already present are continuous at these transitions. We must check whether the birth operator is Lipschitz with respect to these natural variables. This is easy to confirm in the case in which the function β is polynomial of degree at most k_0 near $a=0$, and that is why we chose to address birth operators of that form. What is needed is that we can choose a basis such that the coefficients of the L^2-projection of $\beta(a)$ into $\mathcal{A}$ are bounded. This is trivial away from $a=0$ because of the equivalence of norms on finite dimensional spaces.

4. Error analysis. Wheeler, in her analysis for parabolic equations [15], showed the value of choosing the right projection in constructing an argument; in her case it was the elliptic projection. In this paper we use a tensor product projection based on an elliptic projection in space and an L^2-projection in age.

It is convenient to use $H^{-1}(\Omega)$ as the dual to $H^1(\Omega)$. Let $\|\cdot\|$, $\|\cdot\|_{L^\infty}$, $\|\cdot\|_{H^1}$, and $\|\cdot\|_{H^{-1}}$ denote the L^2, L^∞, H^1, and H^{-1} norms over Ω, respectively. Suppose that Υ is a normed space with norm $\|\cdot\|_\Upsilon$. Then, for any sufficiently nice function $\varphi : \mathbb{R}^+ \to \Upsilon$, let

$$|\!|\!|\varphi|\!|\!|_\Upsilon^2 = \int_0^\infty \|\varphi(a)\|_\Upsilon^2 \, da.$$

A lack of a subscript indicates $\Upsilon = L^2(\Omega)$. For $\varphi : \Omega \to \Upsilon$, define

$$\|\varphi\|_{L^p(\Omega,\Upsilon)} = \Big\| \, \|\varphi(x)\|_\Upsilon \, \Big\|_{L^p(\Omega)}.$$

We show that the approximate solution U is close to a function X, which is the elliptic projection in space and the L^2-projection in age of the true solution u. Let $\mathsf{A}(t) : L^2(\mathbb{R}^+) \to \mathcal{A}(t)$ denote the L^2-projection. To construct X we first project into space. For each (a,t), we take $\tilde{X}(a,t) \in \mathcal{M}$ such that $k(p; u-\tilde{X}, v) = 0$ for all $v \in \mathcal{M}$ and such that

$$\int_\Omega |u - \tilde{X}| \, dx = 0.$$

Similarly, for each t, we take $Y(t) \in \mathcal{M}$ to satisfy $k(p; p-Y, v) = 0$ for all $v \in \mathcal{M}$ and

$$\int_\Omega |p - Y| \, dx = 0.$$

To project into age, we choose $X(t) \in \mathcal{M} \otimes \mathcal{A}(t)$ such that $X(t) = \mathsf{A}(\tilde{X}(a,t))$. We set

$$\vartheta = U - X, \quad \eta = u - \tilde{X}, \quad \tilde{\eta} = \tilde{X} - X, \quad \varpi = P - Y, \text{ and } \sigma = p - Y.$$

In the following estimate we will suppose that ∇Y and the L^2-norm in age of ∇X are uniformly bounded. We could instead add conditions on $\mathcal{M}$, Ω, and u that would imply these bounds, but this would add complexity with no benefit in understanding why the numerics work. We add the following condition.

CONDITION 4.1. *We suppose the quantities*

$$\|\!|u|\!\|_{L^\infty}, \quad \|u\|_{L^\infty(\Omega\times\mathbb{R}^+)}, \quad \|u\|_{L^\infty(\Omega,L^1(\mathbb{R}^+))}, \quad K_k\|\!|\nabla X|\!\|_{L^\infty}, \|p\|_{L^\infty}, \text{ and } K_k\|\!|\nabla Y|\!\|_{L^\infty}$$

are bounded uniformly in time.

THEOREM 4.1. *Let*

$$\theta_1(t) = \int_0^t C_0(\|\!|\vartheta|\!\|_{H^1}^2 + \|\varpi\|_{H^1}^2)(\tau)\, d\tau,$$

$$\epsilon(t) = \int_0^t (\|\!|\eta|\!\|^2 + \|\sigma\|^2 + \|\!|D\eta|\!\|_{H^{-1}}^2 + \|\partial_t\sigma\|_{H^{-1}}^2 + (\Delta a)^2\|\!|\tilde\eta|\!\|^2 + \|\eta(x,0)\|^2)(\tau)\, d\tau.$$

Assume Conditions 2.1, 2.2, 2.3, 2.4, and 4.1 *hold. There exists* $C^*(t) > 0$ *(dependent only on* K_b, K_μ, C_0, C_1, *and the bounds in Condition* 4.1*) such that*

$$(\|\!|\vartheta|\!\|^2 + \|\varpi\|^2 + \theta_1)(t) \le C^*(t)\big((\|\!|\vartheta|\!\|^2 + \|\varpi\|^2)(0) + \epsilon(t)\big).$$

Remark. This result shows superconvergence of one additional power of Δa in the age variable since only $\tilde\eta$ involves approximation in age. Hence, as a function of age, U is closer to the L^2-projection in age of u than it is to u itself, at least for Δa sufficiently small.

Proof. For this proof C will denote an arbitrary constant with dependencies not greater than those of C^*. When only a single argument is given to U, u, η, $\tilde\eta$, or ϑ, that argument denotes age.

Subtract (3.1) from (3.2) and let $v = \vartheta$ to get

$$\int_0^\infty (D(\vartheta - \eta - \tilde\eta), \vartheta) + k(P;U,\vartheta) - k(p;u,\vartheta) + (\mu(P)U,\vartheta) - (\mu(p)u,\vartheta)\, da + (U(0) - u(0), \vartheta(0)) = (b(U) - b(u), \vartheta(0)). \tag{4.1}$$

By orthogonality, for $v(\cdot,\cdot,t) \in \mathcal{M}\otimes\mathcal{A}(t)$,

$$\int_0^\infty (\tilde\eta, v)\, da = 0. \tag{4.2}$$

For $t \notin \check{\mathcal{J}}$, let $\delta > 0$ be such that $(t-\delta, t+\delta)\cap\check{\mathcal{J}} = \emptyset$. For a given $v(\cdot,\cdot,t) \in \mathcal{M}\otimes\mathcal{A}(t)$ and $-s \in (t-\delta, t+\delta)$, take $v(\cdot,\cdot,s) \in \mathcal{M}\otimes\mathcal{A}(s)$ such that v is constant along characteristics. By (4.2) we have, for $0 < \Delta t < \delta$,

$$\begin{aligned}
0 &= \frac{1}{\Delta t}\int_0^\infty \big(\tilde\eta(\cdot,a,t+\Delta t), v(\cdot,a,t+\Delta t)\big) - \big(\tilde\eta(\cdot,a,t), v(\cdot,a,t)\big)\, da \\
&= \frac{1}{\Delta t}\int_0^\infty \big(\tilde\eta(\cdot,a+\Delta t,t+\Delta t) - \tilde\eta(\cdot,a,t), v(\cdot,a+\Delta t,t+\Delta t)\big)\, da \\
&\quad + \frac{1}{\Delta t}\int_0^{\Delta t} \big(\tilde\eta(\cdot,a,t+\Delta t), v(\cdot,a,t+\Delta t)\big)\, da \\
&\quad + \frac{1}{\Delta t}\int_0^\infty \big(\tilde\eta(\cdot,a,t), v(\cdot,a+\Delta t,t+\Delta t) - v(\cdot,a,t)\big)\, da.
\end{aligned}$$

In this expression the last term is zero because v is constant along characteristics. Taking limits we see that for $v(\cdot,\cdot,t) \in \mathcal{M} \otimes \mathcal{A}(t)$,

$$\int_0^\infty \big(D\tilde{\eta}(\cdot,a,t), v(\cdot,a,t)\big)\ da + \big(\tilde{\eta}(\cdot,0,t), v(\cdot,0,t)\big) = 0.$$

Hence, with $v = \vartheta$, we note that

$$\begin{aligned}(4.3)\qquad \big(U-u,\vartheta\big)(0) - \int_0^\infty \big(D\tilde{\eta},\vartheta\big)\ da &= \big(\vartheta-\eta-\tilde{\eta},\vartheta\big)(0) - \int_0^\infty \big(D\tilde{\eta},\vartheta\big)\ da \\ &= \|\vartheta(0)\|^2 - \big(\eta,\vartheta\big)(0) \\ &\geq \frac{3}{4}\|\vartheta(0)\|^2 - \|\eta(0)\|^2.\end{aligned}$$

Rearranging terms in (4.1) and applying (4.2) and (4.3) gives

$$\begin{aligned}\int_0^\infty (D\vartheta,\vartheta) + k(P;\vartheta,\vartheta) + (\mu(P)\vartheta,\vartheta)\ da + \frac{3}{4}\|\vartheta(0)\|^2 \\ \leq \int_0^\infty (k(x,Y)-k(x,P), \nabla X\cdot\nabla\vartheta) + (k(x,p)-k(x,Y), \nabla X\cdot\nabla\vartheta)\ da \\ + \int_0^\infty ((\mu(p)-\mu(P))u,\vartheta) + (D\eta,\vartheta) + (\mu(P)(\eta+\tilde{\eta}),\vartheta)\ da \\ + \ (b(U)-b(u),\vartheta(0)) + \|\eta(0)\|^2.\end{aligned}$$

We have the equality

$$\int_0^\infty (D\vartheta,\vartheta)\ da = \frac{1}{2}\int_0^\infty D\|\vartheta\|^2\ da = \frac{1}{2}\partial_t|\!|\!|\vartheta|\!|\!|^2 - \frac{1}{2}\|\vartheta(0)\|^2.$$

Using Conditions 2.1–2.2, Hölder's inequality, and the arithmetic-geometric inequality, $yz \leq \frac{1}{2}(\varepsilon y^2 + (1/\varepsilon)z^2)$, we get the following bounds:

$$\begin{aligned}\int_0^\infty k(P;\vartheta,\vartheta) + (\mu(P)\vartheta,\vartheta)\ da &\geq C_0|\!|\!|\vartheta|\!|\!|^2_{H^1}, \\ \int_0^\infty (k(x,Y)-k(x,P),\nabla X\cdot\nabla\vartheta)\ da &\leq \int_0^\infty (K_k|\varpi|, |\nabla X\cdot\nabla\vartheta|)\ da \\ &\leq \frac{2K_k^2}{C_0}|\!|\!|\nabla X|\!|\!|^2_{L^\infty}\|\varpi\|^2 + \frac{C_0}{8}|\!|\!|\vartheta|\!|\!|^2_{H^1}, \\ \int_0^\infty (k(x,Y)-k(x,P),\nabla X\cdot\nabla\vartheta)\ da &\leq \frac{2K_k^2}{C_0}|\!|\!|\nabla X|\!|\!|^2_{L^\infty}\|\sigma\|^2 + \frac{C_0}{8}|\!|\!|\vartheta|\!|\!|^2_{H^1}, \\ \int_0^\infty ((\mu(p)-\mu(P))u,\vartheta)\ da &\leq \int_0^\infty (K_\mu|P-p|, |u\vartheta|)\ da \\ &\leq \frac{K_\mu}{2}|\!|\!|u|\!|\!|_{L^\infty}\left(2(\|\varpi\|^2+\|\sigma\|^2) + |\!|\!|\vartheta|\!|\!|^2\right), \\ \int_0^\infty (D\eta,\vartheta)\ da &\leq \frac{1}{C_0}|\!|\!|D\eta|\!|\!|^2_{H^{-1}} + \frac{C_0}{4}|\!|\!|\vartheta|\!|\!|^2_{H^1}, \\ \int_0^\infty (\mu(P)\eta,\vartheta)\ da &\leq \frac{C_1}{2}(|\!|\!|\eta|\!|\!|^2 + |\!|\!|\vartheta|\!|\!|^2).\end{aligned}$$

Let $\bar{\mu}$ denote the average of μ in age over each interval of the age discretization. Then

$$\begin{aligned}\int_0^\infty (\mu(P)\tilde{\eta},\vartheta)\ da &= \int_0^\infty \big((\mu(P)-\bar{\mu}(P))\tilde{\eta},\vartheta\big)\ da\\ &\le \frac{1}{2}\big(\|\mu(P)-\bar{\mu}(P)\|^2_{L^\infty(\Omega\times\mathbb{R}^+)}|\!|\!|\tilde{\eta}|\!|\!|^2+|\!|\!|\vartheta|\!|\!|^2\big)\\ &\le \frac{(\Delta a)^2}{8}C_1^2|\!|\!|\tilde{\eta}|\!|\!|^2+\frac{1}{2}|\!|\!|\vartheta|\!|\!|^2.\end{aligned}$$

For the birth term we make the bound

$$\begin{aligned}(b(U)-b(u),\vartheta(0)) &\le \|b(U)-b(u)\|^2+\frac{1}{4}\|\vartheta(0)\|^2\\ &\le 3K_b^2\Big((1+\|u\|^2_{L^\infty(\Omega,L^1(\mathbb{R}^+))})\|P-p\|^2\\ &\qquad +\|U-u\|^2_{L^2(\Omega,H^{-1}(\mathbb{R}^+))}\Big)+\frac{1}{4}\|\vartheta(0)\|^2.\end{aligned}$$

We combine the above inequalities and use the fact that $\|\tilde{\eta}\|_{L^2(\Omega,H^{-1}(\mathbb{R}^+))}\le\frac{\Delta a}{\pi}|\!|\!|\tilde{\eta}|\!|\!|$ (see Appendix A of [2]) to get

$$\begin{aligned}\partial_t|\!|\!|\vartheta|\!|\!|^2+C_0|\!|\!|\vartheta|\!|\!|^2_{H^1} &\le C\Big(\|\varpi\|^2+|\!|\!|\vartheta|\!|\!|^2+\|\sigma\|^2+|\!|\!|\eta|\!|\!|^2+|\!|\!|D\eta|\!|\!|^2_{H^{-1}}\\ &\qquad +(\Delta a)^2|\!|\!|\tilde{\eta}|\!|\!|^2+\|\eta(0)\|^2\Big).\end{aligned}\tag{4.4}$$

Before we can use the above evolution inequality to get bounds on the error, we need corresponding relationships for the total population density. We integrate (2.1) over a and take the inner product with a test function $v\in\mathcal{M}$ to obtain

$$(\partial_t p,v)+k(p;p,v)+\left(\int_0^\infty \mu(p)u\ da\,,v\right)=(b(u),v).\tag{4.5}$$

For the approximate total population density we have

$$(\partial_t P,v)+k(P;P,v)+\left(\int_0^\infty \mu(P)U\ da\,,v\right)=(b(U),v).\tag{4.6}$$

We subtract (4.5) from (4.6) and let $v=\varpi$ to get

$$\begin{aligned}\frac{1}{2}\partial_t\|\varpi\|^2+k(P;P,\varpi)-k(p;p,\varpi)+\left(\int_0^\infty \mu(P)U-\mu(p)u\ da,\varpi\right)\\ =(b(U)-b(u),\varpi)+(\partial_t\sigma,\varpi).\end{aligned}$$

This has a form similar to (4.1). We have the bound

$$\begin{aligned}\left(\int_0^\infty \mu(P)\tilde{\eta}\ da,\varpi\right) &= \left(\int_0^\infty (\mu(P)-\bar{\mu}(P))\tilde{\eta}\ da,\varpi\right)\\ &\le \Big(\|\mu(x,\cdot,P)-\bar{\mu}(x,\cdot,P)\|^2_{L^2(\mathbb{R}^+)}\|\tilde{\eta}(x,\cdot,t)\|^2_{L^2(\mathbb{R}^+)},\varpi\Big)\\ &\le \frac{(\Delta a)^2C_1^2}{2\pi^2}|\!|\!|\tilde{\eta}|\!|\!|^2+\frac{1}{2}\|\varpi\|^2.\end{aligned}$$

Using bounds similar to those for (4.1) for the other terms gives

$$\partial_t\|\varpi\|^2 + C_0\|\varpi\|^2_{H^1} \le C\Big(\|\varpi\|^2 + |\!|\!|\vartheta|\!|\!|^2 + \|\sigma\|^2 + |\!|\!|\eta|\!|\!|^2 + \|\partial_t\sigma\|^2_{H^{-1}} + (\Delta a)^2|\!|\!|\tilde{\eta}|\!|\!|^2\Big). \tag{4.7}$$

By adding (4.4) and (4.7) and applying a Gronwall's lemma,[1] we obtain the stated result. □

5. Computational example. In this section we provide a computational example that shows some of the benefits of using a higher order approximation space in the age variable. In particular, we are able to use a coarser age discretization for the same or better error.

The example system in [2] illustrates the importance of being able to decouple the age and time discretizations so that the age and time steps are neither uniform nor equal. The spatial dynamics of the problem require small time steps for accurate resolution. The small time steps taken in the simulation, particularly the initial step, are caused by roughness in space. The behavior in age is relatively smooth, which in turn calls for a much coarser discretization in age than in time.

The example presented in [2] was meant to illustrate the need for a method that discretizes age and time separately because of the influence of space. It does not clearly illustrate two aspects of the interaction of age and time. First, it is not clear what is needed to align the introduction of an age interval with the start of a time step. Second, it does not illustrate the level to which the age dynamics of a system will determine the size of the time step.

We present an example system that illustrates the benefits of using higher order approximation spaces in age, as well as some aspects of the interaction of age and time in these methods. In order to achieve the latter goal, we assume uniformity in space. Because the dynamics in age can be, and often are, independent of space, the benefits of using higher order polynomial spaces in age will generalize to systems that include spatial dynamics.

We consider the system (2.1)–(2.5) with $k = 0$. We use the birth term,

$$b(x, u(x,\cdot,t)) = \int_0^\infty 5a\,u\;da,$$

so that fecundity increases linearly with age. For the death modulus, we use

$$\mu(x,a,p) = \mu(a) = \frac{10e^{10(a-0.8)}}{e^{10(a-0.8)} + e^{-10(a-0.8)}} + \frac{1}{2}.$$

This represents a situation where mortality remains low until around a certain age, at which point it increases dramatically. This is the case in *Proteus mirabilis* swarm colony development [6].

For the initial condition, we use a population of older organisms,

$$u_0(x,a) = 128|a-0.5|^3 - 48(a-0.5)^2 + 1,$$

if $|a-0.5| < 0.25$, and $u_0(x,a) = 0$, otherwise.

[1] Assume $u, b, c \ge 0$ are continuous and $g \ge 0$ is differentiable. Then $g'(t) + b(t) \le c(t) + u(t)g(t)$ implies $g(t) + \int_0^t b(\tau)\;d\tau \le \exp(\int_0^t u(\tau)\;d\tau)(g(0) + \int_0^t c(\tau)\;d\tau)$.

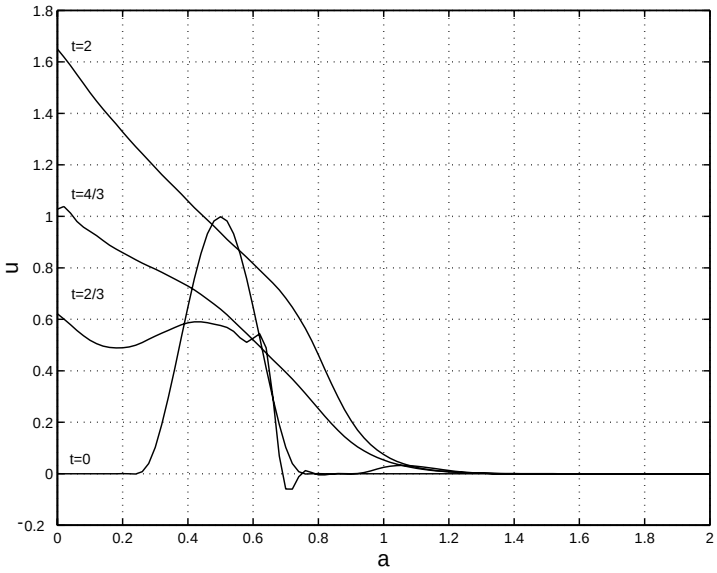


FIG. 5.1. *Profiles of the population density, u. The profiles are $t = 2/3$ apart.*

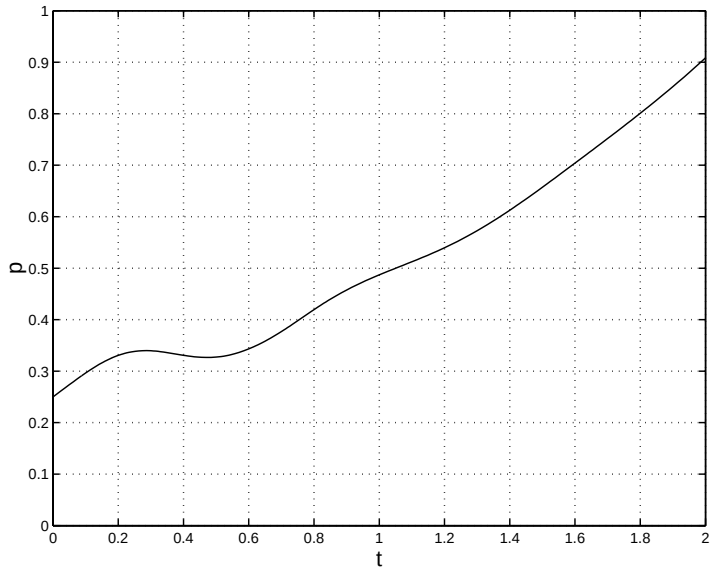


FIG. 5.2. *The total population density, p.*

We take the temporal domain to be $[0, 2]$. We find that truncating the age domain to $[0, 2]$ is sufficient. We assume uniformity of the solution over the spatial domain Ω.

We implement step-size control in time via step-doubling (without extrapolation) [1, 7]. This means that for each time step we take a step of size Δt and compare it to the solution obtained by taking two steps of size $\Delta t/2$. We adjust a parameter that limits local truncation error so that the simulation is well resolved in time.

For the age discretization, we assume $\mathcal{J}$ is uniform with age intervals of size Δa.

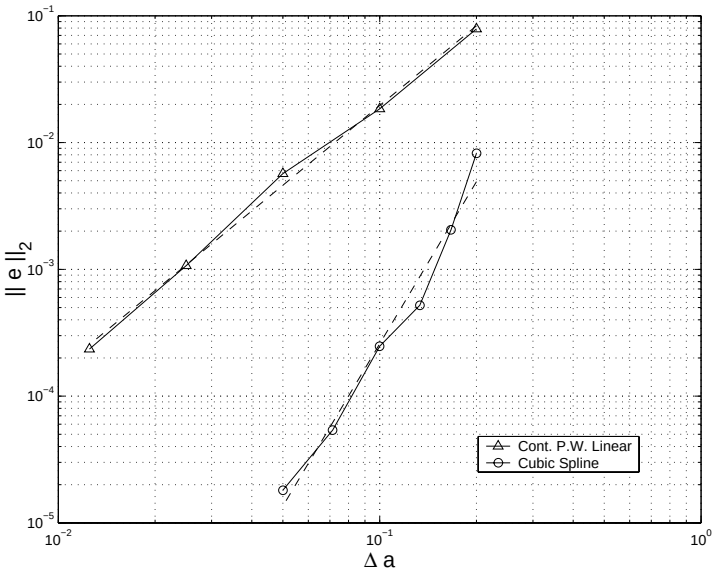


FIG. 5.3. *Convergence study for $q = 0$ and $q = 1$ showing second order and fourth order convergence of the method, respectively. The comparison is of the computed solution, u, at time $t = 2$. The slope of the least squares fit for the piecewise constant case postprocessed to continuous piecewise linear functions is approximately* 2.09. *The slope of the least squares fit for the discontinuous piecewise linear case postprocessed to cubic splines is approximately* 4.25.

In other words, all age intervals that are not the birth interval are of length Δa. We study the convergence of the method with piecewise constants postprocessed to continuous piecewise linear functions ($q = 0$) and with discontinuous piecewise linear functions postprocessed to cubic splines ($q = 1$). The postprocessing for piecewise constants was discussed in [2]; it involves using knot values that are obtained from a line connecting the midpoints of adjacent intervals. The cubic splines can be produced from the discontinuous linear functions in several ways; here we used the two Gauss points in each of two adjacent intervals to define a cubic that is used to give knot values and slopes. This is a natural choice since the L^2-projection into discontinuous piecewise linear functions is superconvergent at the two Gauss points.

Figure 5.1 shows solution profiles $t = 2/3$ apart for the simulation using $q = 1$ and $\Delta a = 0.1$. The solution at $t = 2/3$ has a discontinuity because the initial condition does not contain any newborns. However, this discontinuity dies out over time. Figure 5.2 shows the growth of the total population density. There is a period of population decline, due to the die-off of the initial population, before the population enters a stage of exponential growth.

Figure 5.3 shows the results of a convergence study using $q = 0$ and $q = 1$ with postprocessing. The error is determined by comparison with the numerical solution solved with $\Delta a = 6.25 \times 10^{-3}$ for the case of $q = 0$ and $\Delta a = 2.5 \times 10^{-2}$ for $q = 1$. We get the expected result that the use of discontinuous piecewise linear functions gives fourth order convergence with much better initial error than the second order convergence given by the use of piecewise constants.

Figure 5.4 shows the time steps taken for the simulation using $q = 1$ and $\Delta a = 0.1$. We find that the time step needed to resolve this simulation is roughly 10^{-2}, with the exception of a trough at $t \approx 1.8$ that corresponds to the die-off of the relatively

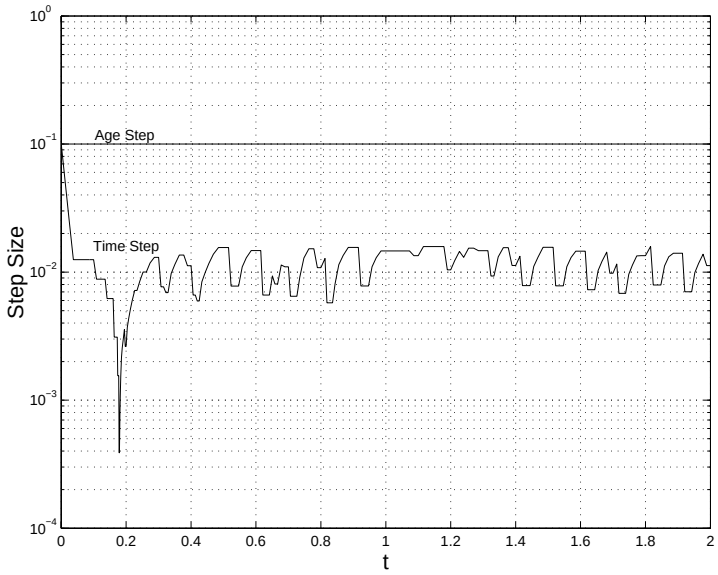


FIG. 5.4. *Step sizes for the age and time discretizations. There were* 204 *accepted steps and* 22 *rejected steps during the simulation. The smallest step size was approximately* 3.88×10^{-4} *taken at* $t \approx 1.8$. *This trough corresponds to the die-off of the relatively large initial population of older individuals.*

large initial population of older individuals. The need for this small time step is due to the increased complexity of the underlying problem at this point, not the moving age grid.

The restrictions on the time steps imposed by the age discretization are due to the need to introduce a new age interval at the start of a time step. This requires that $\Delta t \leq \Delta a$. Moreover, this restriction may require a slightly smaller time step before the introduction of a new age interval at the birth boundary. Smaller time steps may also be needed during the initial birthing into a new age interval. These cause the minor time step fluctuations we see throughout the latter part of the simulation.

REFERENCES

[1] B. P. AYATI, *Methods for Computational Population Dynamics*, Ph.D. thesis, University of Chicago, Chicago, IL, 1998.

[2] B. P. AYATI, *A variable time step method for an age-dependent population model with nonlinear diffusion*, SIAM J. Numer. Anal., 37 (2000), pp. 1571–1589.

[3] A. M. DE ROOS, *Numerical methods for structured population models: The escalator boxcar train*, Numer. Methods Partial Differential Equations, 4 (1989), pp. 173–195.

[4] A. M. DE ROOS, *A gentle introduction to physiologically structured population models*, in Structured-population Models in Marine, Terrestrial, and Freshwater Systems, Population and Community Biology Series 18, S. Tuljapurkar and H. Caswell, eds., Chapman & Hall, New York, 1997, pp. 119–204.

[5] A. M. DE ROOS AND J. A. J. METZ, *Towards a numerical analysis of the escalator boxcar train*, in Differential Equations with Applications in Biology, Physics and Engineering, Lecture Notes in Pure and Appl. Math. 133, J. A. Goldstein, F. Kappel, and W. Schappacher, eds., Marcel Dekker, New York, 1991, pp. 91–113.

[6] S. E. ESIPOV AND J. A. SHAPIRO, *Kinetic model of* Proteus mirabilis *swarm colony development*, J. Math. Biol., 36 (1998), pp. 249–268.

[7] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice–

Hall, Englewood Cliffs, NJ, 1971.

[8] M.-Y. Kim, *Galerkin methods for a model of population dynamics with nonlinear diffusion*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 59–73.

[9] M.-Y. Kim and E.-J. Park, *Mixed approximation of a population diffusion equation*, Comput. Math. Appl., 30 (1995), pp. 23–33.

[10] B. W. Kooi and S. A. L. M. Kooijman, *Discrete event versus continuous approach to reproduction in structured population dynamics*, Theoret. Population Biol., 56 (1999), pp. 91–105.

[11] M. Langlais, *A nonlinear problem in age-dependent population diffusion*, SIAM J. Math. Anal., 16 (1985), pp. 510–529.

[12] L. Lopez and D. Trigiante, *A finite difference scheme for a stiff problem arising in the numerical solution of a population dynamic model with spatial diffusion*, Nonlinear Anal., 9 (1985), pp. 1–12.

[13] M. Pascual and S. A. Levin, *Spatial scaling in a benthic population model with density-dependent disturbance*, Theoret. Population Biol., 56 (1999), pp. 106–122.

[14] O. Rauprich, M. Matsushita, K. Weijer, F. Siegert, S. E. Esipov, and J. A. Shapiro, *Periodic phenomena in* Proteus mirabilis *swarm colony development*, J. Bacteriol., 178 (1996), pp. 6525–6538.

[15] M. F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1077–1099

SQUEEZABLE ORTHOGONAL BASES: ACCURACY AND SMOOTHNESS*

GEORGE C. DONOVAN†, JEFFREY S. GERONIMO‡, AND DOUGLAS P. HARDIN§

Abstract. We present a method for generating local orthogonal bases on arbitrary partitions of $\mathbf{R}$ from a given local orthogonal shift-invariant basis via what we call a *squeeze map*. We give necessary and sufficient conditions for a squeeze map to generate a nonuniform basis that preserves any smoothness and/or accuracy (polynomial reproduction) of the shift-invariant basis. When the shift-invariant basis has sufficient smoothness or accuracy, there is a unique squeeze map associated with a given partition that preserves this property and, in this case, the squeeze map may be calculated locally in terms of the ratios of adjacent intervals. If both the smoothness and accuracy are large enough, then the resulting nonuniform space contains the nonuniform spline space characterized by that smoothness and accuracy.

Our examples include a multiresolution on nonuniform partitions such that each space has a local orthogonal basis consisting of continuous piecewise quadratic functions. We also construct a family of smooth, local, orthogonal, piecewise polynomial generators with arbitrary approximation order.

Key words. orthogonal bases, nonuniform grids, polynomial reproduction, piecewise polynomial, multiresolution

AMS subject classifications. 41A15, 41A63

PII. S0036142900380868

1. Introduction.

Finitely generated shift-invariant (FSI) spaces naturally arise in several areas of numerical analysis and approximation theory, including the theory of splines and wavelets. A major advantage of an FSI space is the existence of a convenient basis generated by a (usually) small number of functions. When the basis is local and orthogonal the process of finding the orthogonal projection Pf of $f \in L^2(\mathbf{R})$ onto the space is local so that changing f on a compact interval affects only Pf on a slightly larger interval.

In this paper we introduce and investigate a method for adapting local shift-invariant bases to nonuniform partitions via what we call a *squeeze map*. When the shift-invariant basis is orthogonal, the squeeze map may be chosen so that the nonuniform basis is also orthogonal.

The notion of squeeze maps generalizes ideas introduced in [4], where we gave examples of local orthogonal piecewise polynomial shift-invariant bases that are easily adaptable to arbitrary grids in $\mathbf{R}$. The focus of this paper is on characterizing when a squeeze map generates a nonuniform basis preserving any smoothness and/or accuracy (polynomial reproduction) of the shift-invariant basis. When the shift-invariant basis has sufficient smoothness or accuracy, there is a unique squeeze map associated with a given partition of $\mathbf{R}$ that preserves this property and, in this case, the squeeze map may be calculated locally in terms of the ratios of adjacent intervals. When both the smoothness and accuracy are large enough, we find that the resulting nonuniform

*Received by the editors November 13, 2000; accepted for publication (in revised form) February 25, 2002; published electronically August 28, 2002.
http://www.siam.org/journals/sinum/40-3/38086.html
†20491 Hazelton Way, Ashburn, VA 20147 (gdonovan@bbn.com).
‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (geronimo@math.gatech.edu)
§Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (hardin@math.vanderbilt.edu).

space contains the nonuniform spline space characterized by that smoothness and accuracy.

Two applications that provide motivation for our work are adaptive least squares and the construction of orthogonal wavelets on semiregular and irregular families of grids:

(1) Since the bases constructed here are local and orthogonal and depend locally on the given grid, it is relatively easy to calculate changes in the orthogonal projection of a given function (onto the span of this basis) resulting from changes in the grid, making them well suited for adaptive least square problems.

(2) While we do not focus on refinable spaces in this paper, it is the refinable case that provides the main motivation for our study. We remark that our methods provide a means to adapt a multiresolution on uniform grids to one on a semiuniform family of grids (that is, an arbitrary coarse grid that is uniformly subdivided). In the example in section 6.3, we start with Daubechies's famous orthogonal scaling function ${}_2\phi$. We find that, given a nonuniform grid, there is a unique squeeze map that preserves the accuracy of the space. In the example in section 6.4, we use ideas from [5] to construct a multiresolution on an arbitrary nonuniform subdivision. (The only requirement is that each interval is subdivided into two subintervals.) Each space has a local orthogonal basis consisting of continuous piecewise quadratic functions.

Finally, in section 7 we construct a family of smooth, local, orthogonal, piecewise polynomial generators with arbitrary approximation order using techniques developed in [6]. These generators have fewer components than the corresponding refinable generators constructed in [6], and so we prefer them when refinability is not required. We mention that a possible application of this family is to code division multiple access (CDMA) technology, where several users share a single channel using orthogonal decompositions.

1.1. Shift-invariant spaces. We call a compactly supported, finite-length (column) vector

$$\Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix} \in L^2(\mathbf{R})^n$$

a *generator*. Note that when it is clear from the context, we also consider a generator Φ to be the set of its components; that is, we also consider $\Phi \subset L^2(\mathbf{R})$. When we refer to the span of Φ we mean the subspace of $L^2(\mathbf{R})$ spanned by the components of Φ.

For a generator Φ, let

$$B(\Phi) := \{\phi_i(\cdot - j) \mid j \in \mathbf{Z}, i = 1, \ldots, n\}.$$

If $B(\Phi)$ is an orthogonal set, we say Φ is an *orthogonal generator*. For a generator Φ, let

$$S(\Phi) := \left\{ \sum_{j \in \mathbf{Z}} c(j)^\top \Phi(\cdot - j) \mid c(j) \in \mathbf{R}^n, j \in \mathbf{Z} \right\}.$$

If $V = S(\Phi)$ for some generator Φ, then V is called a *finitely generated shift-invariant* (FSI) space.

1.2. Minimally supported generators. Our procedure for constructing local bases on nonuniform partitions starts with generators supported on $[-1,1]$ satisfying a local linear independence condition on $[0,1]$. In particular, for $k \le n$, we say that a generator

$$\Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_k \\ \phi_{k+1} \\ \vdots \\ \phi_n \end{pmatrix} = \begin{pmatrix} \bar{\Phi} \\ \breve{\Phi} \end{pmatrix}$$

(where $\bar{\Phi}$ consists of the first k elements of Φ and $\breve{\Phi}$ consists of the last $n-k$) is a *minimally supported k-generator* (or just *minimally supported*) if

(1) supp $\Phi \subset [-1,1]$;
(2) supp $\breve{\Phi} \subset [0,1]$;
(3) the collection $\breve{\Phi} \cup \bar{\Phi}\chi_{[0,1]} \cup (\bar{\Phi}(\cdot - 1))\chi_{[0,1]}$ is linearly independent.

We denote the collection of all minimally supported k-generators with n components by $\mathcal{G}_k^n$. See section 5 for several illustrative examples of orthogonal minimally supported generators. The notion of generators minimally supported on $[-1,1]$ played a central role in the construction of orthogonal, smooth, piecewise polynomial wavelets given in [5].

For $\Phi \in \mathcal{G}_k^n$, we denote the "left" and "right" pieces of $\bar{\Phi}$ by

$$\Phi_R := \bar{\Phi}\chi_{[0,1]} \text{ and } \Phi_L := \bar{\Phi}\chi_{[-1,0)}.$$

Obviously, condition (3) can be rewritten as $\breve{\Phi} \cup \Phi_R \cup \Phi_L(\cdot - 1)$ is linearly independent. If Φ is minimally supported, then it follows from the local linear independence condition (3) above that $B(\Phi)$ is linearly independent; that is, any $f \in S(\Phi)$ has a unique representation of the form $f = \sum c_j \Phi(\cdot - j)$. In the remainder of this paper, when there is clearly some underlying minimally supported generator with k and n as above, then, for any (row or column) vector v of length n, we let $\bar{v}$ denote the subvector of the first k components of v and $\breve{v}$ the subvector of the last $n-k$ components of v.

Also, for $f, g \in L^2(\mathbf{R})^n$ we define $\langle f, g \rangle := \int_{\mathbf{R}} f(x)g(x)^\top dx \in \mathbf{R}^{n\times n}$, where $v^\top$ denotes the transpose of a (column) vector v.

2. Squeeze maps. Let $a = (a_j)_{j\in\mathbf{Z}}$ be a strictly increasing real-valued sequence with no accumulation point in $\mathbf{R}$, in which case we call a a *knot sequence*. Let $L_j := a_{j+1} - a_j$ denote the length of the jth interval $[a_j, a_{j+1}]$ and let $\tau_j = \tau_j^a$ be given by

$$\tau_j(x) = \begin{cases} (x - a_j)/L_{j-1} & \text{for } x \le a_j, \\ (x - a_j)/L_j & \text{for } x \ge a_j. \end{cases} \tag{2.1}$$

Then τ_j maps the points a_{j-1}, a_j, and a_{j+1} to -1, 0, and 1, respectively.

Suppose Φ is an orthogonal minimally supported generator. Consider

$$B_0 = \bigcup_{j\in\mathbf{Z}} \Phi \circ \tau_j.$$

If Φ is continuous and $k = 1$ (for example, see the example in section 6.1), then (because τ_j is affine on each "overlap" interval $[a_j, a_{j+1}]$ and continuous on $\mathbf{R}$) it follows that B_0 is a continuous orthogonal basis for its span.

On the other hand, if $\Phi \in C^1(\mathbf{R})$ and $\Phi'(0) \neq 0$ (for example, consider the continuously differentiable Φ with $k = 2$ in the example in section 6.2), then the components of B_0 are not in $C^1(\mathbf{R})$ for nonuniform a. In particular, $\bar{\Phi} \circ \tau_j$ is not differentiable at a_j unless $L_{j-1} = L_j$. This leads us to consider a more general construction in which linear combinations of $\bar{\Phi}_L \circ \tau_j$ are pieced together with linear combinations of $\bar{\Phi}_R \circ \tau_j$ via what we call a *squeeze map*.

More specifically, let $A_{\mathrm{L}}^{(j)}$ and $A_{\mathrm{R}}^{(j)}$ be invertible $k \times k$ matrices for $j \in Z$ and let $A_j : \mathbf{R} \to \mathbf{R}^{k\times k}$ denote the matrix-valued function on $\mathbf{R}$ defined by

$$A_j = \chi_{[-1,0)} A_{\mathrm{L}}^{(j)} + \chi_{[0,1]} A_{\mathrm{R}}^{(j)}.$$

Given A and a knot sequence a, we call the sequence of mappings $\sigma = (\sigma_j)_{j\in\mathbf{Z}}$, where $\sigma_j : \mathcal{G}_k^n \to L^2(\mathbf{R})^n$ is given by

$$\sigma_j(\Phi) = \begin{pmatrix} A_j \bar{\Phi} \circ \tau_j \\ \breve{\Phi} \circ \tau_j \end{pmatrix},$$

a *squeeze map (on $\mathcal{G}_k^n$)*.

As before, we let $\bar{\sigma}_j(\Phi)$ denote the vector of the first k components of $\sigma_j(\Phi)$ and $\breve{\sigma}_j(\Phi)$ the remaining $n - k$ components. Observe that

$$\bar{\sigma}_j(\Phi) = \left(\chi_{[-1,0)} A_{\mathrm{L}}^{(j)} \bar{\Phi} + \chi_{[0,1]} A_{\mathrm{R}}^{(j)} \bar{\Phi}\right) \circ \tau_j = \left(A_{\mathrm{L}}^{(j)} \Phi_L + A_{\mathrm{R}}^{(j)} \Phi_R\right) \circ \tau_j$$

and $\mathrm{supp}\bar{\sigma}_j(\Phi) \subset [a_{j-1}, a_{j+1}]$, while $\mathrm{supp}\breve{\sigma}_j(\Phi) \subset [a_j, a_{j+1}]$.

If σ is a squeeze map on $\mathcal{G}_k^n$ and $\Phi \in \mathcal{G}_k^n$, then we define

$$B_\sigma(\Phi) := \bigcup_{j\in\mathbf{Z}} \sigma_j(\Phi)$$

and

$$S_\sigma(\Phi) := \left\{ \sum_{j\in\mathbf{Z}} c(j)^\top \sigma_j(\Phi) \mid c(j) \in \mathbf{R}^n, j \in \mathbf{Z} \right\}.$$

The minimal support of Φ and the invertibility of $A_{\mathrm{L}}^{(j)}$ and $A_{\mathrm{R}}^{(j)}$ imply that $B_\sigma(\Phi)$ is linearly independent.

If σ is a squeeze map with matrix sequences $(A_{\mathrm{L}}^{(j)})$ and $(A_{\mathrm{R}}^{(j)})$, we define

$$R_j = R_j(\sigma) := (A_{\mathrm{L}}^{(j)})^{-1} A_{\mathrm{R}}^{(j)} \qquad (j \in \mathbf{Z}).$$

We say that two squeeze maps σ and ν on $\mathcal{G}_k^n$ are *equivalent* whenever $S_\sigma(\Phi) = S_\nu(\Phi)$ for any $\Phi \in \mathcal{G}_k^n$.

LEMMA 2.1. *Suppose σ and ν are squeeze maps on $\mathcal{G}_k^n$. Then σ and ν are equivalent if and only if*

$$R_j(\sigma) = R_j(\nu) \qquad (j \in \mathbf{Z}). \tag{2.2}$$

Proof. Suppose (2.2) holds. Then

$$\left(A_{\mathrm{L}}^{(j)}\right)^{-1}\bar{\sigma}_j(\Phi) = \left(\tilde{A_{\mathrm{L}}}^{(j)}\right)^{-1}\bar{\nu}_j(\Phi) \qquad (j \in \mathbf{Z}),$$

where σ has matrix sequences $(A_{\mathrm{L}}^{(j)})$ and $(A_{\mathrm{R}}^{(j)})$ and ν has matrix sequences $(\tilde{A_{\mathrm{L}}}^{(j)})$ and $(\tilde{A_{\mathrm{R}}}^{(j)})$. Since $A_{\mathrm{L}}^{(j)}$ and $\tilde{A_{\mathrm{L}}}^{(j)}$ are nonsingular the above shows that $\bar{\sigma}_j(\Phi)$ and $\bar{\nu}_j(\Phi)$ (considered as sets) have the same span. By definition $\breve{\sigma}_j(\Phi)$ and $\breve{\nu}_j(\Phi)$ have the same span showing that $\sigma_j(\Phi)$ and $\nu_j(\Phi)$ have the same span, and hence $S_\sigma(\Phi) = S_\nu(\Phi)$.

On the other hand, if $S_\sigma(\Phi) = S_\nu(\Phi)$, then the local linear independence of $B_\sigma(\Phi)$ and $B_\nu(\Phi)$ shows that $\bar{\sigma}_j(\Phi)$ and $\bar{\nu}_j(\Phi)$ have the same span for each $j \in \mathbf{Z}$. Thus, there must be some nonsingular matrix W_j such that

$$\bar{\nu}_j(\Phi) = W_j\bar{\sigma}_j(\Phi) \qquad (j \in \mathbf{Z}),$$

which implies that (2.2) holds. ∎

Our motivation for considering squeeze maps is that if Φ is a minimally supported orthogonal generator, then we can always find a local orthogonal basis for $S_\sigma(\Phi) \cap L^2(\mathbf{R})$. To see this, note that the elements of $\bar{\sigma}_j(\Phi)$ are orthogonal to the elements of $\bar{\sigma}_{j+1}(\Phi)$:

$$\langle\bar{\sigma}_j(\Phi), \bar{\sigma}_{j+1}(\Phi)\rangle = L_j A_{\mathrm{R}}^{(j)}\langle\Phi_R, \Phi_L(\cdot - 1)\rangle\left(A_{\mathrm{L}}^{(j+1)}\right)^{\top} = 0 \qquad (j \in \mathbf{Z}).$$

It then follows that $\sigma_j(\Phi)$ is orthogonal to $\sigma_{j'}(\Phi)$ for any $j \neq j' \in \mathbf{Z}$. Finally, for each $j \in \mathbf{Z}$, we choose some orthogonal basis for the span of $\bar{\sigma}_j(\Phi)$ (for instance, by applying the Gram–Schmidt process to $\bar{\sigma}_j(\Phi)$). This change of basis corresponds to constructing a squeeze map ν equivalent to σ such that $B_\nu(\Phi)$ is an orthogonal set and is equivalent to performing the following matrix factorization: Let $B_jB_j^{\top}$ be a Cholesky factorization of $\langle\bar{\sigma}_j, \bar{\sigma}_j\rangle$, that is,

(2.3)
$$B_jB_j^{\top} = \langle\bar{\sigma}_j(\Phi), \bar{\sigma}_j(\Phi)\rangle = L_{j-1}A_{\mathrm{L}}^{(j)}\langle\Phi_L, \Phi_L\rangle\left(A_{\mathrm{L}}^{(j)}\right)^{\top} + L_jA_{\mathrm{R}}^{(j)}\langle\Phi_R, \Phi_R\rangle\left(A_{\mathrm{R}}^{(j)}\right)^{\top}.$$

Then ν with matrix sequences $(B_j^{-1}A_{\mathrm{L}}^{(j)})$ and $(B_j^{-1}A_{\mathrm{R}}^{(j)})$ is equivalent to σ, and $B_\nu(\Phi)$ is an orthogonal basis for $S_\sigma(\Phi) \cap L^2(\mathbf{R})$. Thus we have the following lemma.

LEMMA 2.2. *Suppose Φ is a minimally supported orthogonal generator and σ is a squeeze map for Φ. Then there is some squeeze map ν equivalent to σ such that $B_\nu(\Phi)$ is an orthogonal basis for $S_\sigma(\Phi) \cap L^2(\mathbf{R})$.*

3. Polynomial reproduction and smoothness. In this section we give necessary and sufficient conditions for a squeeze map σ to preserve the accuracy (polynomial reproduction) and regularity of $S(\Phi)$. Throughout this section Φ is a generator in $\mathcal{G}_k^n$ and σ is a squeeze map on $\mathcal{G}_k^n$ with matrix sequences $(A_{\mathrm{L}}^{(j)})$ and $(A_{\mathrm{R}}^{(j)})$. Recall that $R_j = (A_{\mathrm{L}}^{(j)})^{-1}A_{\mathrm{R}}^{(j)}$ for $j \in \mathbf{Z}$.

First we address the smoothness of $S_\sigma(\Phi)$. Since $S_\sigma(\Phi)$ restricted to bounded intervals has finite dimension it follows that $S_\sigma(\Phi) \subset C^m(\mathbf{R})$ if and only if $\sigma_j(\Phi) \subset C^m(\mathbf{R})$ for all $j \in \mathbf{Z}$.

THEOREM 3.1. *Suppose $\Phi \subset C^m(\mathbf{R})$. Then, for $j \in \mathbf{Z}$, $\sigma_j(\Phi) \subset C^m(\mathbf{R})$ if and only if $\bar{\Phi}^{(q)}(0)$ is either 0 or a right eigenvector of R_j with eigenvalue $(L_j/L_{j-1})^q$ for $0 \le q \le m$, that is, if and only if*

$$R_j \bar{\Phi}^{(q)}(0) = (L_j/L_{j-1})^q \bar{\Phi}^{(q)}(0). \tag{3.1}$$

(Here $\bar{\Phi}^{(q)}$ denotes the qth derivative of $\bar{\Phi}$.)

Hence, $S_\sigma(\Phi) \subset C^m(\mathbf{R})$ if and only if (3.1) holds for all $j \in \mathbf{Z}$.

Proof. The theorem follows from

$$\sigma_j(\Phi)^{(q)}(j^-) = \begin{pmatrix} (L_{j-1})^{-q} A_{\mathrm{L}}^{(j)} \bar{\Phi}^{(q)}(0^-) \\ 0 \end{pmatrix}$$

and

$$\sigma_j(\Phi)^{(q)}(j^+) = \begin{pmatrix} (L_j)^{-q} A_{\mathrm{R}}^{(j)} \bar{\Phi}^{(q)}(0^+) \\ 0 \end{pmatrix}$$

for $0 \le q \le m$ and $j \in \mathbf{Z}$. □

Let Π_p, $p \ge 0$, denote the collection of univariate polynomials of degree at most p. A generator Φ is said to have *accuracy* $p+1$ if $\Pi_p \subset S(\Phi)$. If Φ has accuracy $p+1$, then (since $B(\Phi)$ is a linearly independent set), for each $l = 0, \dots, p$, there is a unique sequence of $1 \times n$ vectors $(\alpha_l(j))_{j \in \mathbf{Z}}$ such that

$$x^l = \sum_{j \in \mathbf{Z}} \alpha_l(j) \Phi(x-j) = \sum_j \bar{\alpha}_l(j) \bar{\Phi}(x-j) + \breve{\alpha}_l(j) \breve{\Phi}(x-j). \tag{3.2}$$

We say $S_\sigma(\Phi)$ has *accuracy* $p+1$ if $\Pi_p \subset S_\sigma(\Phi)$, in which case there exists, for each $l = 0, \dots, p$, a unique sequence $(\alpha'_l(j))_{j \in \mathbf{Z}}$, such that

$$x^l = \sum_j \alpha'_l(j) \sigma_j(\Phi)(x). \tag{3.3}$$

THEOREM 3.2. *Suppose Φ has accuracy $p+1$ and σ is a squeeze map for Φ. Then $S_\sigma(\Phi)$ has accuracy $p+1$ if and only if $\bar{\alpha}_l(0)$ is either 0 or a left eigenvector of R_j with eigenvalue $(L_j/L_{j-1})^l$ for $l = 0, \dots, p$ and all $j \in \mathbf{Z}$.*

Proof. Using (3.3) and the definition of $\sigma_j(\Phi)$, observe that $S_\sigma(\Phi)$ having accuracy $p+1$ is equivalent to the existence of sequences $(\alpha'_l(j))_{j \in \mathbf{Z}}$, $l = 0, \dots, p$, such that

$$x^l = \bar{\alpha}'_l(j) A_{\mathrm{R}}^{(j)} \bar{\Phi} \circ \tau_j(x) + \bar{\alpha}'_l(j+1) A_{\mathrm{L}}^{(j+1)} \bar{\Phi} \circ \tau_{j+1}(x) + \breve{\alpha}'_l(j) \breve{\Phi} \circ \tau_j(x),$$

for $j \in \mathbf{Z}$, and $x \in \tau_j^{-1}([0,1]) = [a_j, a_{j+1}]$. By substituting $\tau_j^{-1}(x)$ for x in the above, we obtain

$$\sum_{i=0}^{l} \binom{l}{i} L_j^i x^i a_j^{l-i} = \bar{\alpha}'_l(j) A_{\mathrm{R}}^{(j)} \bar{\Phi}(x) + \bar{\alpha}'_l(j+1) A_{\mathrm{L}}^{(j+1)} \bar{\Phi}(x-1) + \breve{\alpha}'_l(j) \breve{\Phi}(x),$$

where l and j are as above, but here $x \in [0,1]$. Now, since Φ has accuracy $p+1$, we can use (3.2) to replace x^i in the above. In particular,

$$x^i = \bar{\alpha}_i(0) \bar{\Phi}(x) + \bar{\alpha}_i(1) \bar{\Phi}(x-1) + \breve{\alpha}(0) \breve{\Phi}(x)$$

for $x \in [0,1]$. With this substitution and the minimal support properties of Φ, we find an equivalent system of equations,

$$\begin{aligned} \breve{\alpha}_l'(j) &= \sum_{i=0}^{l} \binom{l}{i} L_j^i a_j^{l-i} \breve{\alpha}_i(0), \\ \bar{\alpha}_l'(j) A_{\mathrm{R}}^{(j)} &= \sum_{i=0}^{l} \binom{l}{i} L_j^i a_j^{l-i} \bar{\alpha}_i(0), \\ \bar{\alpha}_l'(j+1) A_{\mathrm{L}}^{(j+1)} &= \sum_{i=0}^{l} \binom{l}{i} L_j^i a_j^{l-i} \bar{\alpha}_i(1). \end{aligned} \tag{3.4}$$

Now, since $A_{\mathrm{L}}^{(j)}$ and $A_{\mathrm{R}}^{(j)}$ are invertible for all j, the last two of these lead to

$$\sum_{i=0}^{l} \binom{l}{i} L_j^i a_j^{l-i} \bar{\alpha}_i(0) (A_{\mathrm{R}}^{(j)})^{-1} = \sum_{i=0}^{l} \binom{l}{i} L_{j-1}^i a_{j-1}^{l-i} \bar{\alpha}_i(1) (A_{\mathrm{L}}^{(j)})^{-1}. \tag{3.5}$$

Here, we may apply Lemma 3.6 proved at the end of this section, observing that $\alpha_i(0)$ and $\alpha_i(1)$ satisfy (3.11), as, therefore, do $\bar{\alpha}_i(0)$ and $\bar{\alpha}_i(1)$. The "only if" part of the result follows.

All steps in the above argument are reversible except the one from (3.4) to (3.5). The "if" part of the result is achieved by choosing $\breve{\alpha}'$ and $\bar{\alpha}'$ as in (3.4). The choice is consistent with (3.5) and leads to the desired accuracy of $S_\sigma(\Phi)$. □

If $\Phi \subset C^m(\mathbf{R})$ and has accuracy $p+1$, then $\breve{\Phi}^{(q)}(0) = 0$ for $0 \le q \le m$ and so

$$\bar{\alpha}_l(0) \bar{\Phi}^{(q)}(0) = \left. \frac{d^q}{dx^q} x^l \right|_{x=0} = (q!) \delta_{l,q} \tag{3.6}$$

for $0 \le q \le m$ and $0 \le l \le p$, where $\delta_{l,q}$ denotes the Kronecker delta. For $0 \le q \le m$ and $0 \le l \le p$, we define the following matrices:

$$V_l = \begin{pmatrix} \bar{\alpha}_0(0) \\ \vdots \\ \bar{\alpha}_l(0) \end{pmatrix} \text{ and } W_q = \left(\bar{\Phi}(0) \cdots \bar{\Phi}^{(q)}(0) \right). \tag{3.7}$$

Then (3.6) is equivalent to the matrix equation

$$V_p W_m = D, \tag{3.8}$$

where D is the $(p+1) \times (m+1)$ diagonal matrix whose (l,l)th component is $(l-1)!$. The rank of the right side of (3.8) is $\min(m+1, p+1)$. Also, V_p and W_m have rank at most k which gives the following bound for k.

LEMMA 3.3. *Suppose $\Phi \subset C^m(\mathbf{R})$ and has accuracy $p+1$; then*

$$k \ge \min(m+1, p+1).$$

Next we consider when accuracy or smoothness uniquely determines the squeeze map (up to equivalency) and when accuracy forces smoothness or smoothness forces accuracy.

THEOREM 3.4. *Suppose $\Phi \subset C^m(\mathbf{R})$ and has accuracy $p+1$. Let a be a given knot sequence.*

(i) *If $k \le p+1$ and the square matrix V_{k-1} is nonsingular, then there exists a unique (up to equivalence) squeeze map σ with knot sequence a such that $S_\sigma(\Phi)$ has accuracy k. In addition, $S_\sigma(\Phi) \subset C^m(\mathbf{R})$.*

(ii) *If $k \le m+1$ and the square matrix W_{k-1} is nonsingular, then there exists a unique (up to equivalence) squeeze map σ with knot sequence a such that $S_\sigma(\Phi) \subset C^{k-1}(\mathbf{R})$. Furthermore, $S_\sigma(\Phi)$ has accuracy $p+1$.*

Proof. Case (i). Suppose $k \le p+1$, $V := V_{k-1}$ is nonsingular, and σ is a squeeze map for Φ. Then $\bar\alpha_l(0) \ne 0$ for $0 \le l \le k-1$, and so Theorem 3.2 asserts that $S_\sigma(\Phi)$ has accuracy k if and only if $\bar\alpha_l(0)$ is a left eigenvector of R_j with eigenvalue $(\frac{L_j}{L_{j-1}})^l$ for $0 \le l \le k-1$ and $j \in \mathbf{Z}$. The latter condition is equivalent to

$$V\,R_j = \Lambda(L_j/L_{j-1})\,V \qquad (j \in \mathbf{Z}),$$

where $\Lambda(\lambda)$ is a $k \times k$ diagonal matrix whose (l,l)th component is λ^{l-1} for $\lambda \in \mathbf{R}_+$. Thus, $S_\sigma(\Phi)$ has accuracy k if and only if

$$R_j = V^{-1}\,\Lambda(L_j/L_{j-1})\,V \qquad (j \in \mathbf{Z}). \tag{3.9}$$

Equation (3.8) shows that $\Phi^{(q)}(0)$ is the qth column of $V^{-1}D$. Multiplying both sides of (3.9) on the right by $\Phi^{(q)}(0)$ then shows that $\bar\Phi^{(q)}(0)$ is a right eigenvector of R_j with eigenvalue $(\frac{L_j}{L_{j-1}})^q$ for $0 \le q \le m$. Hence, Theorem 3.1 shows that $S_\sigma(\Phi) \subset C^m(\mathbf{R})$.

Case (ii). Now suppose $k \le m+1$ and $W := W_{k-1}$ is nonsingular. As in case (i) we find that $S_\sigma(\Phi)$ has accuracy k if and only if

$$R_j = W\,\Lambda(L_j/L_{j-1})\,W^{-1} \tag{3.10}$$

and that $\bar\alpha_l(0)$ is a left eigenvector of R_j with eigenvalue $(\frac{L_j}{L_{j-1}})^l$ for $0 \le l \le p$. Hence Theorem 3.2 shows that $S_\sigma(\Phi)$ has accuracy $p+1$. □

If $k = \min(m+1, p+1)$, then it follows from (3.8) that V_{k-1} and W_{k-1} are both nonsingular, and so both cases in Theorem 3.4 hold. The next theorem shows that $S_\sigma(\Phi)$ contains the spline space

$$S_p^m(a) := \{f \in C^m(\mathbf{R}) \mid f|_{(a_j, a_{j+1})} \in \Pi_p,\ j \in \mathbf{Z}\}$$

when $k = \min(m+1, p+1)$. In this case, it is known from classical spline theory that the accuracy determines the approximation order of $S_\sigma(\Phi)$. Note that $S_p^m(a) = \Pi_p$ if $m \ge p$.

THEOREM 3.5. *Suppose $\Phi \subset C^m(\mathbf{R})$, Φ has accuracy $p+1$ and $k = \min(m+1, p+1)$. Let a be a given knot sequence.*

(i) *There exists a squeeze map σ with knot sequence a such that $S_\sigma(\Phi) \subset C^m(\mathbf{R})$ and has accuracy $p+1$.*

(ii) *If ν is any other squeeze map with knot sequence a such that either $S_\nu(\Phi) \subset C^{k-1}(\mathbf{R})$ or (ν, Φ) has accuracy k, then ν is equivalent to σ.*

(iii) *$S_p^m(a) \subset S_\sigma(\Phi)$. (This is nontrivial only when $m < p$, in which case $k = m+1$.)*

Proof. If $k = \min(m+1, p+1)$, then it follows from (3.8) that V_{k-1} and W_{k-1} are both nonsingular and parts (i) and (ii) follow from Theorem 3.4.

From part (i) we have $\Pi_p \subset S_\sigma(\Phi)$, and so we need only consider the case $m < p$. Since W_{k-1} is nonsingular, it follows from (3.8) that $\bar{\alpha}_l(0) = 0$ for $l = m+1, \ldots, p$.

For simplicity, first suppose that one of the knots, say a_i, is 0. Then (3.4) implies

$$\bar{\alpha}_l'(i) A_{\mathrm{R}}^{(j)} = L_i^l \bar{\alpha}_l(i) = 0 \qquad (l = m+1, \ldots, p).$$

Thus (3.3) becomes

$$x^l = \breve{\alpha}_l'(i)\breve{\sigma}_i(\Phi) + \sum_{j \neq i} \alpha_l'(j)\sigma_j(\Phi).$$

Thus the truncated powers $(x_+)^l$, $l = m+1, \ldots, p$, can be written as

$$(x_+)^l = \breve{\alpha}_l'(i)\breve{\sigma}_i(\Phi) + \sum_{j > i} \alpha_l'(j)\sigma_j(\Phi),$$

and so they are in $S_\sigma(\Phi)$ for $l = m+1, \ldots, p$. Observe that shifting the knots by a constant shift translates the basis $B_\sigma(\Phi)$ by the same amount. Hence $S_\sigma(\Phi)$ contains the truncated powers $((x - a_j)_+)^l$ for $l = m+1, \ldots, p$ and $j \in \mathbf{Z}$. The truncated powers form a basis for $S_p^m(a)$ showing that (iii) holds. □

Finally, we prove the following lemma that was used in the proof of Theorem 3.2.

LEMMA 3.6. *Suppose $a_0, a_1, L_1 \in \mathbf{R}$ and $L_0 = a_1 - a_0$. Further, suppose $\alpha(0)$ and $\alpha(1)$ are sequences of $1 \times k$ vectors such that*

$$\alpha_l(1) = \sum_{i=0}^{l} \binom{l}{i} \alpha_i(0) \tag{3.11}$$

for $l = 0, \ldots, p$. Then the $k \times k$ matrices C and D satisfy the conditions

$$\sum_{i=0}^{l} \binom{l}{i} L_1^i a_1^{l-i} \alpha_i(0) C = \sum_{i=0}^{l} \binom{l}{i} L_0^i a_0^{l-i} \alpha_i(1) D \tag{3.12}$$

for $l = 0, \ldots, p$ if and only if

$$\alpha_l(0)(L_1^l C - L_0^l D) = 0 \tag{3.13}$$

for $l = 0, \ldots, p$.

Proof. For a given l, we may use (3.11) to substitute for $\alpha_i(1)$ in (3.12). Then using routine combinatorial manipulations we find

$$\sum_{i=0}^{l} \binom{l}{i} L_1^i a_1^{l-i} \alpha_i(0) C = \sum_{j=0}^{l} \binom{l}{j} \alpha_j(0) D \sum_{i=j}^{l} \binom{l-j}{i-j} L_0^i a_0^{l-i}. \tag{3.14}$$

By shifting the index on the inner sum by j, the left-hand side becomes

$$\begin{aligned} &\sum_{j=0}^{l} \binom{l}{j} \alpha_j(0) D \sum_{i=0}^{l-j} \binom{l-j}{i} L_0^{i+j} a_0^{l-i-j} \\ &= \sum_{j=0}^{l} \binom{l}{j} \alpha_j(0) L_0^j a_1^{l-j} D, \end{aligned}$$

where the final equality follows from $a_1 = a_0 + L_0$ and the binomial theorem. Thus (3.12) is equivalent to

$$\sum_{i=0}^{l} \binom{l}{i} a_1^{l-i} \alpha_i(0)(L_1^i C - L_0^i D) = 0.$$

From here it is easy to show the equivalence with (3.13) by induction on $l = 0, \ldots, p$. □

4. Constructing the squeeze map. Suppose $\Phi \subset C^m(\mathbf{R})$, Φ has accuracy $p+1$ and $k \le \max(m+1, p+1)$. Then either case (i) or (ii) of Theorem 3.4 holds and the squeeze map preserving accuracy in case (i) or smoothness in case (ii) is unique up to equivalence. In both cases there is a full set of k eigenvectors for R_j for $j \in \mathbf{Z}$ with specified eigenvalues. These eigenvectors then uniquely determine R_j through either (3.9) or (3.10). In case (i), let $U = V_{k-1}$ and in case (ii) let $U = W_{k-1}^{-1}$, where V_{k-1} and W_{k-1} are given by (3.7). Let

$$R(\lambda) := U^{-1}\,\Lambda(\lambda)\,U \qquad (\lambda > 0). \tag{4.1}$$

Then $R_j = R(\lambda_j)$, where $\lambda_j := L_j/L_{j-1}$ for $j \in \mathbf{Z}$. Thus, the squeeze map is determined (up to equivalence) for an arbitrary knot sequence. Furthermore, each R_j is determined only by the ratio L_j/L_{j-1}.

Now suppose Φ is an orthogonal generator. Let σ be the squeeze map with matrix sequences (I, R_j). Following the proof of Lemma 2.2, an equivalent squeeze map ν so that $B_\nu(\Phi)$ is orthogonal may be found as follows. First, find a Cholesky factorization (see (2.3)):

$$B(\lambda)B(\lambda)^\top = \langle \Phi_L, \Phi_L \rangle + \lambda R(\lambda) \langle \Phi_R, \Phi_R \rangle R(\lambda)^\top. \tag{4.2}$$

Let $B_j = \sqrt{L_{j-1}} B(\lambda_j)$ for $j \in \mathbf{Z}$. Then ν with matrix sequences $A_{\mathrm{L}}^{(j)} = (B_j^{-1})$ and $A_{\mathrm{R}}^{(j)} = (B_j^{-1} R_j)$ gives an orthogonal basis. Again note that for fixed Φ, B_j depends only on L_{j-1} and L_j, and (since a Cholesky factorization is equivalent to an LU factorization using Gaussian elimination) we can find a closed form expression for ν_j in terms of the ratio $\lambda_j = L_j/L_{j-1}$. This makes it simple and quick to construct the squeeze map for an arbitrary knot sequence.

In our examples we consider only $k = 1$ or $k = 2$. When $k = 1$ it is trivial to obtain B_j. Suppose

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

is a symmetric positive definite matrix. (That is, $v^\top A v > 0$ for any nonzero 2-vector v.) Then A is positive definite if and only if both a and $\det A$ are positive. One choice for B such that $BB^\top = A$ is given by

$$B = \frac{1}{\sqrt{a}} \begin{pmatrix} a & 0 \\ b & \sqrt{\det A} \end{pmatrix}. \tag{4.3}$$

5. Orthogonal minimally supported generators.

5.1. Rescaling orthogonal generators. Any orthogonal compactly supported generator may be used to construct an orthogonal generator supported on $[-1,1]$ as we next describe. If the support of $\Phi = (\phi_1, \dots, \phi_n)^\top$ is contained in $[-1, M]$, then let Φ_M denote the generator consisting of the concatenation of the M generators $\Phi(M \cdot + k)$, $(k = 0, \dots, M-1)$. Then Φ_M is an orthogonal generator supported in $[-1,1]$ and $S(\Phi_M)$ equals $S(\Phi)(M\cdot)$ (that is, the dilation by $1/M$ of the space $S(\Phi)$). The local linear independence conditions for minimal support must then be checked separately. However, when Φ is an orthogonal scalar ($n = 1$) refinable generator it is known that Φ is locally linearly independent (that is, the nonzero restrictions of the shifts of Φ to any open interval are linearly independent), which implies the weaker type of local linear independence we require in the definition of minimal support. The example in section 6.3 is constructed in this way.

5.2. General construction. In [5] the authors developed a method for constructing orthogonal generators. For $W \subset L^2(\mathbf{R})$, let P_W denote the orthogonal projection onto W.

LEMMA 5.1 (see [5]). *Suppose Φ is a minimally supported k-generator. There exists an orthogonal minimally supported k-generator Ψ such that $S(\Psi) = S(\Phi)$ if and only if*

$$(I - P_{S(\check{\Phi})})\bar{\Phi} \perp (I - P_{S(\check{\Phi})})\bar{\Phi}(\cdot - 1). \tag{5.1}$$

(That is, $(I - P_{S(\check{\Phi})})\phi_i \perp (I - P_{S(\check{\Phi})})\phi_j(\cdot - 1)$ for $1 \le i, j \le k$.)

Proof (sketch of proof). Let $\check{\Psi}$ be an orthogonal basis for the span of $\check{\Phi}$ and choose $\bar{\Psi}$ to be an orthogonal basis for the span of $(I - P_{S(\check{\Phi})})\bar{\Phi}$. Then Ψ is an orthogonal, minimally supported k-generator for $S(\Phi)$ if Ψ and $\Psi(\cdot - 1)$ are orthogonal (or, equivalently, if (5.1) holds). The other direction relies on the observation that if Φ and Ψ are minimally supported k-generators such that $S(\Psi) = S(\Phi)$, then

$$\operatorname{span}\check{\Phi} = \operatorname{span}\check{\Psi}$$

and

$$\operatorname{span}\Phi \cup \check{\Phi}(\cdot + 1) = \operatorname{span}\check{\Psi} \cup \check{\Psi}(\cdot + 1) \qquad \square$$

The idea of the construction is to choose $\check{\Phi}$ so that (5.1) holds. The orthogonal generators in the examples in sections 6.1 and 6.2 and section 7 are constructed in this way.

6. Examples. In this section we present several examples to illustrate our methods. The examples in sections 6.1 and 6.2 first appeared in [4]. In both examples it is the smoothness condition that determines the squeeze map. Also, in these two examples, $k = \min(m+1, p+1)$, and so the resulting $S_\sigma(\Phi)$ contains $S_p^m(a)$ by Theorem 3.5.

In the example in section 6.3, we rescale Daubechies's orthogonal scaling function ${}_2\phi$ as described in section 5.1 to construct a continuous orthogonal refinable generator minimally supported on [-1,1] with $k = n = 2$. The accuracy in this case is $p + 1 = 2$ and, by Theorem 3.4 (i), the squeeze map is uniquely determined by the accuracy condition once a knot sequence is specified. In fact, it is this example that motivated our study of the accuracy of squeezed spaces $S_\sigma(\Phi)$. In the example in section 6.3 we have $m + 1 = 1 < 2 = k$, and so Theorem 3.5 does not apply in this case.

We are also interested in this example because the generator Φ is *refinable*; that is,

$$\Phi(\cdot/2) = \sum_{j\in\mathbf{Z}} c(j)\Phi(\cdot - j) \tag{6.1}$$

for some finitely supported sequence $c : \mathbf{Z} \mapsto \mathbf{R}^{n\times n}$. (In this case the support of c is $\{-2,-1,0,1\}$.)

We next remark that such a refinable minimally supported generator Φ generates a *semiregular multiresolution analysis* (that is, a multiresolution consisting of a nonuniform coarse space that is uniformly refined; see [3]) as follows: Let a^0 be an arbitrary knot sequence and let $a^1 \supset a^0$ be given by

$$a^1_{2j} = a^0_j \text{ and } a^1_{2j+1} = (a^0_j + a^0_j)/2 \qquad (j \in \mathbf{Z}).$$

Let σ^0 and σ^1 be the squeeze maps determined (up to equivalence) by the knot sequences a^0 and a^1, respectively. Then one may verify that $S_{\sigma^0}(\Phi) \subset S_{\sigma^1}(\Phi)$. *Thus we provide a way to construct orthogonal semiregular multiresolutions from orthogonal scaling functions in a way that preserves the accuracy and smoothness of the shift-invariant multiresolution.*

In the example in section 4, we construct an *irregular multiresolution analysis* (that is, a fully nonuniform multiresolution; see [3]) such that each space in the multiresolution has a compactly supported orthogonal basis consisting of continuous piecewise quadratic functions. The spaces in this irregular multiresolution are not, strictly speaking, squeezed spaces of the form $S_\sigma(\Phi)$ but instead result from a slight generalization of our notion of the squeeze map.

6.1. $k = 1$, $m = 0$, $p = 1$, $n = 2$. Let h denote the hat function defined by $h(x) = (1 - |x|)^+$ and suppose $w \in L^2(\mathbf{R})$ is nontrivial and supported in the interval $[0,1]$. Let $\Phi = (h, w)$. Then (5.1) reduces to

$$\langle h, h(\cdot - 1)\rangle = \frac{\langle h, w\rangle\langle w, h(\cdot - 1)\rangle}{\langle w, w\rangle}. \tag{6.2}$$

Thus, any $w \in L^2(\mathbf{R})$ supported in $[0,1]$ and satisfying (6.2) gives an orthogonal generator Ψ by the process described in Lemma 5.1. For example, let q be the piecewise quadratic function given by $q(x) = x(1-x)\chi_{[0,1]}(x)$. Choose $w \in \operatorname{span}\{q, q^2\}$ so that $w = c_1 q + c_2 q^2$ for some constants c_1, c_2. Substituting into (6.2) yields a quadratic equation in the variable $\alpha := c_2/c_1$:

$$\alpha^2 + 30\alpha + 105 = 0$$

or $\alpha = -15 \pm 2\sqrt{30}$. The graphs of ϕ_1 and ϕ_2 are shown in Figure 6.1 for $\alpha = -15 - 2\sqrt{30}$. (This example was first given in [4].)

For $0 \le x \le 1$, we have

$$\phi_1(x) = \sqrt{3}\,(1-x)\,\left(1 - 2\,x + \left(-3 + \sqrt{30}\right)\,x\,(1 - 5\,(1-x)\,x)\right)$$

and

$$\phi_2(x) = \sqrt{330 - 60\,\sqrt{30}}\,(1-x)\,x\,\left(-1 + \left(15 + 2\,\sqrt{30}\right)\,(1-x)\,x\right).$$

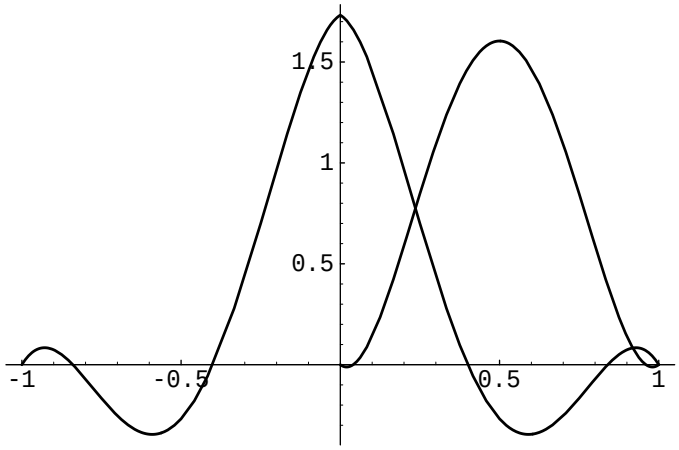


FIG. 6.1. *Continuous orthogonal generator of the example in section* 6.1.

Note that ϕ_1 is even and supported on $[-1,1]$ and that ϕ_2 has support $[0,1]$.

In the case $k=1$ and $m=0$, the squeeze maps preserving continuity are given by $R_j = 1$ for all $j \in \mathbf{Z}$. By Theorem 3.4, this squeeze map will also preserve the approximation of Φ. By the symmetry of Φ we have

$$\langle \Phi_L, \Phi_L \rangle = \langle \Phi_R, \Phi_R \rangle = 1/2.$$

Using (2.3) we get that σ given by

$$A_L^j = A_R^j = \sqrt{\frac{2}{L_{j-1} + L_j}}$$

generates an orthogonal basis $B_\sigma(\Phi)$.

6.2. $k=2$, $m=1$, $p=3$, $n=4$. We next construct a continuously differentiable orthogonal generator. We start with the C^1 cubic Hermite spline functions

$$h_1(x) = \begin{cases} (1+x)^2(1-2x), & x \in [-1,0], \\ (1-x)^2(1+2x), & x \in [0,1], \\ 0 & \text{otherwise}, \end{cases}$$

$$h_2(x) = \begin{cases} (1+x)^2 x, & x \in [-1,0], \\ (1-x)^2 x, & x \in [0,1], \\ 0 & \text{otherwise} \end{cases}$$

and add two continuously differentiable functions w_1 and w_2 supported on $[0,1]$. (In [5], it is shown that at least two w's are required in this case.) The condition (6.2) is equivalent to the following:

$$\langle h_i, h_j(\cdot - 1) \rangle = \frac{\langle h_i, w_1 \rangle \langle w_1, h_j(\cdot - 1) \rangle}{\langle w_1, w_1 \rangle} + \frac{\langle h_i, w_2 \rangle \langle w_2, h_j(\cdot - 1) \rangle}{\langle w_2, w_2 \rangle}. \tag{6.3}$$

Again let q be the piecewise quadratic function given by $q(x) = x(1-x)\chi_{[0,1]}(x)$. We choose w_1 to be of the form $(c_1 + c_2 q + c_3 q^2)q^2$ and w_2 of the form $(\cdot - 1/2)(c_4 + c_5 q)q^2$

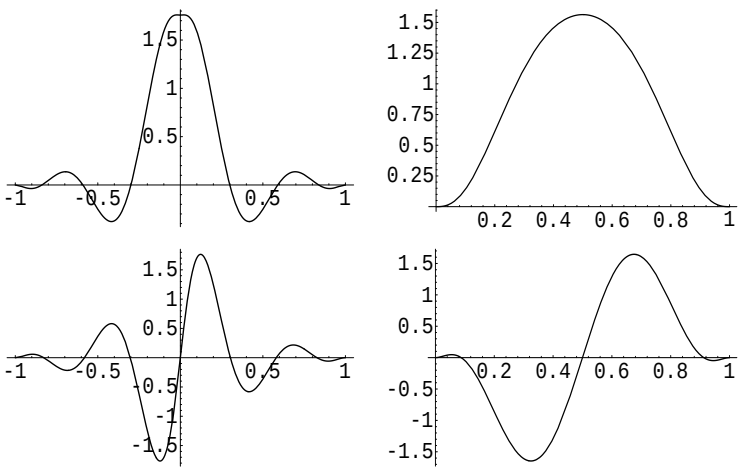


FIG. 6.2. *The C^1 orthonormal generator of the example in section* 6.2. *From upper left, going clockwise: ϕ_1, ϕ_3, ϕ_4, ϕ_2.*

so that w_1 is symmetric about $x = 1/2$ and w_2 is antisymmetric about $x = 1/2$. Substituting into (6.3) yields three quadratic equations in the three variables c_2/c_1, c_3/c_1, and c_5/c_4. Solving these equations numerically and choosing c_1 and c_4 so that $\|w_1\| = \|w_2\| = 1$ yields several solutions. One solution with good properties is given by

c_1	+2.102558692333885
c_2	+214.7707569159831
c_3	-492.4339092336308
c_4	-112.0742772596177
c_5	+1401.893433767276

The graphs of the components of the resulting orthogonal generator $(\phi_1, \phi_2, \phi_3, \phi_4)$ are shown in Figure 6.2.

From the construction of Φ we see that $W = \left(\bar{\Phi}(0)\, \bar{\Phi}'(0)\right)$ is diagonal, and so, using (4.1), we get that $S_\sigma(\Phi) \subset C^1(\mathbf{R})$ if

$$R_j = R(\lambda_j) = \begin{pmatrix} 1 & 0 \\ 0 & \lambda_j \end{pmatrix},$$

where $\lambda_j := L_j / L_{j-1}$.

Since Φ is piecewise polynomial, the inner products $\langle \Phi_L, \Phi_L \rangle$ and $\langle \Phi_R, \Phi_R \rangle$ are easily calculated. Using *Mathematica* to perform these calculations, we arrive at the squeeze maps defined by

$$A_L^{(j)} = \frac{1}{\sqrt{L_{j-1}}} \times \begin{pmatrix} \frac{1.414213}{\sqrt{1+\lambda_j}} & 0 \\ \frac{2.829115 - 2.829115\,\lambda_j^2}{(1+\lambda_j)\sqrt{(0.381634+\lambda_j)\,(1+\lambda_j)\,(2.62031+\lambda_j)}} & \frac{3.162893}{\sqrt{(0.381634+\lambda_j)\,(1+\lambda_j)\,(2.62031+\lambda_j)}} \end{pmatrix}$$

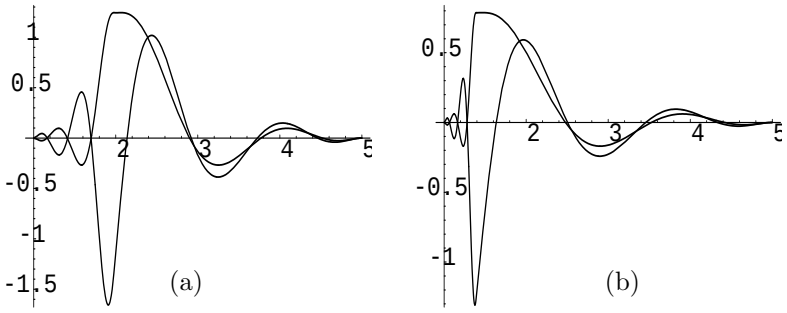


FIG. 6.3. *C^1 basis functions $\bar{\sigma}_j(\Phi)$ from the example in section* 6.2 *for* (a) $\lambda_j = 3$ *(knots at* 1, 2, *and* 5*) and* (b) $\lambda_j = 7$ *(knots at* 1, 3/2, *and* 5*).*

and

$$A_R^{(j)} = \frac{1}{\sqrt{L_{j-1}}} \times \begin{pmatrix} \frac{1.414213}{\sqrt{1+\lambda_j}} & 0 \\ \frac{2.829115-2.829115\,\lambda_j^2}{(1+\lambda_j)\sqrt{(0.381634+\lambda_j)\,(1+\lambda_j)\,(2.62031+\lambda_j)}} & \frac{3.162893\,\lambda_j}{\sqrt{(0.381634+\lambda_j)\,(1+\lambda_j)\,(2.62031+\lambda_j)}} \end{pmatrix}.$$

We show in Figure 6.3 the resulting $\bar{\sigma}_j(\Phi)$ for several values of λ_j.

6.3. Semiregular multiresolution analysis: $k = 2$, $m = 0$, $p = 1$, $n = 2$. Let ${}_2\phi$ denote the continuous orthogonal scaling function of Daubechies supported on $[0, 3]$ (see [2]) and let

$$\Phi = \sqrt{2} \begin{pmatrix} {}_2\phi(2 \cdot +2) \\ {}_2\phi(2 \cdot +1) \end{pmatrix}.$$

Then, as discussed in section 5.1, Φ is an orthogonal generator supported on $[-1, 1]$. The local linear independence condition for minimal support may be verified from the support properties of Φ and the fact that the components of Φ_R are orthogonal to the components of Φ_L, thus showing that Φ is a minimally supported generator with $k = 2$. Also, note that Φ is continuous and has accuracy 2. In this example, it is the accuracy that determines the squeeze map.

Recall that ${}_2\phi$ satisfies a refinement equation

$$ {}_2\phi = \sum_{j=0}^{3} c_j \, {}_2\phi(2 \cdot -j), \tag{6.4}$$

where

$$c_0 = \frac{1+\sqrt{3}}{4}, \quad c_1 = \frac{3+\sqrt{3}}{4}, \quad c_2 = \frac{3-\sqrt{3}}{4}, \quad c_3 = \frac{1-\sqrt{3}}{4}.$$

Using the refinement equation it is possible to calculate the following coefficients

from the zeroth and first moments of ${}_2\phi$ (see [1]):

$$\alpha_0(0) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \qquad \alpha_1(0) = \left(\frac{-1-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right)$$

and

$$\langle \Phi_R, \Phi_R \rangle = \begin{pmatrix} \frac{7}{12} + \frac{5}{7\sqrt{3}} & \frac{1}{28\sqrt{3}} \\ \frac{1}{28\sqrt{3}} & \frac{5}{12} + \frac{5}{7\sqrt{3}} \end{pmatrix}, \qquad \langle \Phi_L, \Phi_L \rangle = \begin{pmatrix} \frac{5}{12} - \frac{5}{7\sqrt{3}} & \frac{-1}{28\sqrt{3}} \\ \frac{-1}{28\sqrt{3}} & \frac{7}{12} - \frac{5}{7\sqrt{3}} \end{pmatrix}.$$

Then

$$R(\lambda) = \frac{1}{2}\begin{pmatrix} 1-\sqrt{3}+\left(1+\sqrt{3}\right)\lambda & \left(1+\sqrt{3}\right)(1-\lambda) \\ \left(1-\sqrt{3}\right)(1-\lambda) & 1+\sqrt{3}+\left(1-\sqrt{3}\right)\lambda \end{pmatrix}$$

and

$$\left(\langle \Phi_L, \Phi_L \rangle + \lambda R(\lambda) \langle \Phi_R, \Phi_R \rangle R(\lambda)^\top\right) = \frac{1}{84}\begin{pmatrix} a(\lambda) & b(\lambda) \\ b(\lambda) & c(\lambda) \end{pmatrix},$$

where

$$a(\lambda) = 35 - 20\sqrt{3} + 4\left(21+8\sqrt{3}\right)\lambda - 4\left(44+23\sqrt{3}\right)\lambda^2 + \left(141+80\sqrt{3}\right)\lambda^3,$$

$$b(\lambda) = (-1+\lambda)\left(\sqrt{3} - 84\lambda - 31\sqrt{3}\lambda + 42\lambda^2 + 19\sqrt{3}\lambda^2\right),$$

$$c(\lambda) = 49 - 20\sqrt{3} + 4\left(21+8\sqrt{3}\right)\lambda - 4\left(19+2\sqrt{3}\right)\lambda^2 + \left(27-4\sqrt{3}\right)\lambda^3.$$

The factors B_j may then be calculated from (4.3).

6.4. Irregular multiresolution analysis: $k = 1$, $m = 0$, $p = 2$, $n = 3$. Let $\left(a^\ell\right)_{\ell \in \mathbf{Z}}$ be a sequence of nested knot sequences such that $a^{\ell+1}_{2j} = a^\ell_j$ for ℓ, $j \in \mathbf{Z}$ and such that $\{a^\ell_j \mid \ell, j \in \mathbf{Z}\}$ is dense in $\mathbf{R}$. Let $V_\ell = S_{0,2}(a^\ell)$ denote the space of continuous piecewise quadratic splines with break points given by a^ℓ. From the theory of splines it follows that (V_ℓ) is a multiresolution analysis. Here we construct a multiresolution (V'_ℓ) such that

$$V_\ell \subset V'_\ell \subset V_{\ell+1}$$

and each V'_ℓ has a local orthogonal basis. The local orthogonal basis for V'_ℓ is generated with a generalization of the squeeze map idea. Our construction here extends the idea of intertwining multiresolution analyses developed in [5] to the nonuniform case.

Let $\Phi = (h, q)$, where h and q are as in the example in section 6.1. Then $V_\ell = S_{\sigma^\ell}(\Phi)$, where σ^ℓ is the squeeze map with knot sequence a^ℓ given by $R_j = 1$.

Let $I^\ell_j = [a^\ell_j, a^\ell_{j+1}]$. The idea of the construction is to add basis functions $w^\ell_j \in V_\ell$ supported on I^ℓ_j for each $j \in \mathbf{Z}$ to the basis $B_{\sigma^\ell}(\Phi)$ in such a way that the resulting space V'_ℓ has a local orthogonal basis. We first describe the construction when $I = I^\ell_j = [0,1]$; the general case will follow by rescaling. Then $a := a^{\ell+1}_{2j+1}$ is in $(0,1)$. Define $q_{1,0}$, $q_{1,1}$, and h_1 by

$$q_{1,0}(x) = q(x/a), \qquad q_{1,1} = q\left(\frac{x-a}{a}\right)$$

and

$$h_1(x) = \begin{cases} x/a & \text{for } x \in [0,a], \\ (1-x)/(1-a) & \text{for } x \in [a,1], \\ 0 & \text{otherwise.} \end{cases}$$

Observe that the space $\mathcal{A}$ of functions in $V^{\ell+1}$ whose support is contained in [0,1] is spanned by $q_{1,0}$, $q_{1,1}$, and h_1. Note that q is in this 3-dimensional space. We choose $w = w_j^\ell$ in the 2-dimensional orthogonal complement of q in $\mathcal{A}$. A basis for this space is given by (with help from *Mathematica*)

$$\begin{aligned} u_0 &= a^2(3a-5)q_{1,0} + (1-a)^2(2+3a)q_{1,1} \\ u_1 &= \left(-2+3\,(-1+a)\,a^3\right)q_{1,0} + \left(-2+3\,(-1+a)^3\,a\right)q_{1,1} \\ &\quad + \left(\frac{16}{5} - 12\,(-1+a)^2\,a^2\right)h_1. \end{aligned}$$

We choose w in $\mathcal{A}$ and orthogonal to q so that it is of the form

$$w = c_1 u_1 + c_2 u_2.$$

Define

$$\theta_R = (I - P_{\mathrm{span}(w,q)})h_R$$

and

$$\theta_L = (I - P_{\mathrm{span}(w(\cdot+1),q(\cdot+1))})h_L,$$

where $h_R = h\chi_{[0,1)}$ and $h_L = h\chi_{[-1,0)}$. In order to construct a local orthogonal basis we require

$$\langle \theta_R, \theta_L(\cdot - 1)\rangle = 0,$$

which is equivalent to the following quadratic equation in the variable $c = c_1/c_2$:

$$\begin{aligned} 0 =& 5\left(4 - 5\,(1-a)^2\,a^2\,(15 + (1-a)\,a)\right) \\ &- 20\,(2 + a\,(9 + 13\,a\,(-3 + 2\,a)))\,c + 4\,(1 + 45\,(1-a)\,a)\,c^2. \end{aligned} \tag{6.5}$$

The discriminant of this equation is

$$80\left(4 - 15\,(1-a)^2\,a^2\right)^2,$$

giving the two solutions

$$c = \frac{20(2 + a(9 + 13a(2a-3))) \pm 4\sqrt{5}(4 - 15(1-a)^2a^2)}{8(1 + 45(1-a)a)}.$$

Hence, there are two choices for w for any $a \in (0,1)$. For each $a \in (0,1)$ choose one such w and denote it by W_λ, where $\lambda = (1-a)/a$ is the ratio of the lengths of the two subintervals $[0,a]$ and $[a,1]$. Let $\theta_{R,\lambda} = \theta_R$ and $\theta_{L,\lambda} = \theta_L$ with $w = W_\lambda$. Define

$$\Phi^{\lambda_L,\lambda_R} = \begin{pmatrix} \theta_{L,\lambda_L} + \theta_{R,\lambda_R} \\ q \\ W_{\lambda_R} \end{pmatrix}$$

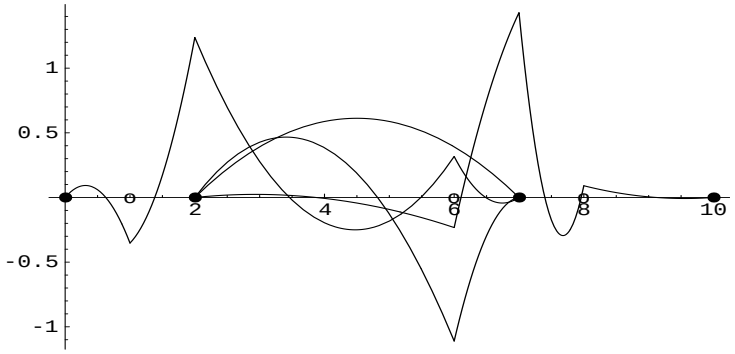


FIG. 6.4. *Continuous, orthogonal piecewise quadratic basis functions from the example in section* 6.4 *with knots* $a_{\ell+1} = \dots, 0, 1, 3, 6, 7, 8, 10, \dots$.

and note that $\Phi^{\lambda_L,\lambda_R}$ is continuous and supported on $[-1, 1]$. Given $a^{\ell+1}$ we construct basis functions supported on $[a^\ell_{j-1}, a^\ell_{j+1}] = [a^{\ell+1}_{2j-2}, a^{\ell+1}_{2j+2}]$ as follows. Let τ^ℓ_j be as in (2.1) with knot sequence a^ℓ, let $L^\ell_j = a^\ell_{j+1} - a^\ell_j$, and let

$$\lambda^\ell_j = L^{\ell+1}_{2j+1}/L^{\ell+1}_{2j}.$$

Note that the collection of functions

$$B^\ell = \bigcup_{j\in\mathbf{Z}} \Phi^{\lambda^\ell_{j-1},\lambda^\ell_j} \circ \tau^\ell_j$$

is an orthogonal system of functions. Let

$$V'_\ell = \operatorname{span}_{L^2} B^\ell$$

for $\ell \in \mathbf{Z}$. Then

$$V_\ell \subset V'_\ell \subset V_{\ell+1} \subset V'_{\ell+1},$$

from which it follows that $(V'_\ell)_{\ell\in\mathbf{Z}}$ is a multiresolution with local orthogonal basis B^ℓ.

Figure 6.4 shows several of the basis functions (we chose the minus branch of the square root) for $a_\ell = \dots, 0, 3, 7, 10, \dots$ and $a_{\ell+1} = \dots, 0, 1, 3, 6, 7, 8, 10, \dots$.

7. Higher order accuracy and smoothness. Let S^n_m be the space of polynomial splines of degree n with C^m knots at the integers. If we denote $A^{n,m} = \{g \in S^n_m : \operatorname{supp} g = [0, 1]\}$, then it is easy to see [6] that an orthogonal basis for $A^{n,m}$ is provided by $\phi^m_i(t) = t^m(1-t)^m p^{2m+5/2}_{i-2m-2}(2t-1)$, $2m+2 \le i \le n$, where $p^{2m+5/2}_j(t)$ is the monic ultraspherical polynomial of degree j with $\lambda = 2m + 5/2$. If we set $\Phi = (\phi^m_0 \cdots \phi^m_n)^T$, where ϕ^m_i, $i = 0, \dots, m$, $\operatorname{supp} \phi^m_i = [-1, 1]$ are appropriately chosen (i.e., judicious linear combinations of r^i_m and l^i_m, $i = 0, \dots, m$, with $r^i_m(t) = t^i(1+t)^{m+1} - 1 \le t \le 0$ and $l^i_m(t) = t^i(1-t)^{m+1} 0 < t \le 1$), then Φ and all its integer translates form a basis for S^n_m. This basis is not orthogonal, so Φ does not generate a local orthogonal basis. We will modify Φ in order to construct an orthogonal set of generators. We do this by adding to Φ, $m+1$ functions w_i chosen so that $W \perp A^{n,m}$ and $\langle (I - P_W)\hat\phi^m_i, (I - P_W)\hat\phi^m_j(\cdot - 1)\rangle = 0$ $i, j = 1, \dots, m+1$. Here $W = \operatorname{span}\{w_i : i = 1, \dots, m+1\}$, P_W is the orthogonal projection onto

W, and $\hat{\phi}_i^m = (I - P_{\{A^{n,m}, A^{n,m}(+1)\}})\phi_i^m$. In the examples given below we will choose w_i to be linear combinations of $\{\phi_j^m\}_{j>n}$. In this way $w_i \perp A^{n,m}$ since the $\{t^m(1-t)^m p_l^{2m+5/2}(2\cdot -1)\}_{l=0}^{\infty}$ is a set of orthogonal polynomials. Notice that the above w_i will have their knots located at the integers. This is in contrast to the construction carried out in [6] where in order to build a MRA it was necessary to use w_i with half integer knots.

7.1. C^0 example. As a first example we consider the case $m = 0$. Then $r_0(t) = (1+t)$ and $l_0(t) = (1-t)$, and we will choose $w_1^n = \phi_{n+1}^0 + \alpha_n \phi_{n+3}^0$. Since ϕ_i^0 is symmetric or antisymmetric about $1/2$ depending on whether i is even or odd, respectively, we see that w_1^n chosen above will be either symmetric or antisymmetric. With $\hat{r}_0^n(\cdot) = (I - P_{A^{n,0}})r_0(\cdot - 1)$ and $\hat{l}_0^n(t) = (I - P_{A^{n_0}})l_0$ we choose α_n so that $\langle (I - P_{w_1^n})\hat{r}_0, (I - P_{w_1^n})\hat{l}_0(t)\rangle = 0$. This gives the following quadratic equation for α_n:

$$\langle \hat{r}_0^n, \hat{l}_0^n\rangle \langle w_1^n, w_1^n\rangle = \langle w_1^n, r_0(\cdot - 1)\rangle \langle w_1^n, l_0\rangle \tag{7.1}$$

or

$$\begin{aligned} &\langle \hat{r}_0^n, \hat{l}_0^n\rangle (\langle \phi_{n+1}^0, \phi_{n+1}^0\rangle + \alpha_n^2 \langle \phi_{n+3}^0, \phi_{n+3}^0\rangle) \\ &= (\langle \phi_{n+1}^0, r_0(\cdot - 1)\rangle + \alpha_n \langle \phi_{n+3}^0, r_0(\cdot - 1)\rangle)(\langle \phi_{n+1}^0, l_0\rangle + \alpha_n \langle \phi_{n+3}^0, l_0\rangle). \end{aligned} \tag{7.2}$$

From [6] we find $\langle \hat{r}_0^n, \hat{l}_0^n\rangle = \frac{(-1)^{n+1} n!}{(n+3)!}$, $\langle \hat{r}_0^n, \hat{r}_0^n\rangle = \frac{1}{n(n+2)}$, and $\langle r_0, \phi_n^0\rangle = 2^{n-2}\frac{n!(n-2)}{2n!}$. Furthermore, since $\langle \phi_n^0, \phi_n^0\rangle = \frac{1}{32}\frac{(n+2)!(n-2)!}{(2n-1)!(2n+1)!!}$, the above equation may be solved for α_n to obtain

$$\alpha_n = -\frac{((2n+7)(2n+3)(n+1) \pm \sqrt{3(2n+7)(2n+3)(n+1)(n+3)}(n+3))(2n+5)}{(n+2)(n+1)(n^2-5n-30)},$$

and $\phi_0^{n,0}$ is given by

$$\phi_0^{n,0}(t) = (I - P_{(w_1^n, w_1^n(\cdot+1))})h(t),$$

where $h(t) = (1-|t|)^+$.

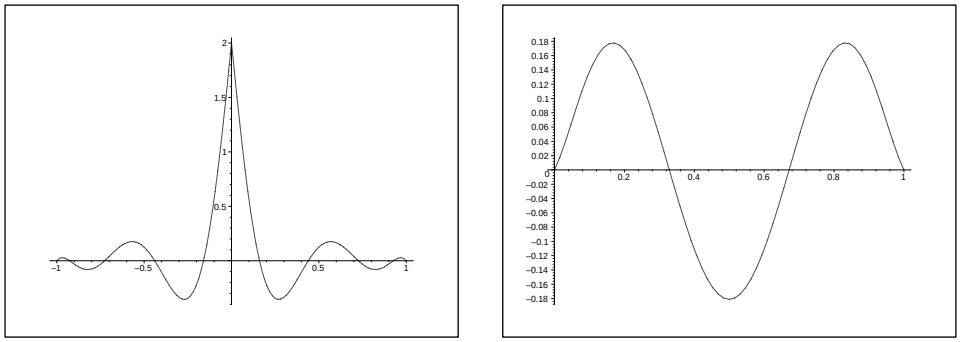
With $\phi_1^{n,0} = w_1^n$ we have the following theorem,

THEOREM 7.1. *For $n \geq 3$, $\Phi_n = (\phi_0^{n,0}, \phi_1^{n,0}, \phi_2^0 \dots, \phi_n^0)^T$ constructed as above is a continuous orthogonal generator for $B(\Phi)$. Furthermore, Φ_n has accuracy $n+1$.*

Figure 7.1 shows $\phi_0^{n,0}$ and $\phi_1^{n,0}$ for $n = 3$.

7.2. C^1 example. We now construct a family of C^1 orthogonal compactly supported generators which have varying degrees of accuracy. In this case four ramp functions, $r_1^i = t^i(1+t)^2$, $i = 0, 1$ and $l_1^i = t^i(1-t)^2$, $i = 0, 1$, are needed in the construction of the orthogonal generators with support equal to $[-1, 1]$. We set $\hat{r}_1^{n,i}(\cdot) = (I - P_{A^{n,1}})r_1^i(\cdot - 1)$ and $\hat{l}_1^{n,i}(t) = (I - P_{A^{n,1}})l_1^i$. The necessary integrals to compute the above projections can be found in [6]. In order to make the computations somewhat more tractable we biorthogonalize the above ramp functions. Utilizing the integrals [6]

$$\langle \hat{r}_0^{n,1}, \hat{l}_0^{n,1}\rangle = \frac{4(-1)^{n+1}(n^2+2n-9)(n-2)!}{(n+3)!}, \tag{7.3}$$



$\phi_0^{3,0}$ $\phi_1^{3,0}$

FIG. 7.1. *The functions* ϕ_0 *and* ϕ_3 *from section* 7.1 *for* $n = 3$.

$$\langle \hat{r}_0^{n,1}, \hat{l}_1^{n,1} \rangle = \frac{12(-1)^{n+1}(n-2)!}{(n+3)!}, \tag{7.4}$$

and

$$\langle \hat{r}_1^{n,1}, \hat{l}_1^{n,1} \rangle = \frac{36(-1)^{n+1}(n-3)!}{(n+4)!}, \tag{7.5}$$

we set $r_{n,0} = \hat{r}_0^{n,1}$, $l_{n,0} = \hat{l}_0^{n,1}$, $r_{n,1} = \hat{r}_1^{n,1} - \frac{\langle \hat{r}_1^{n,1}, l_{n,0} \rangle}{\langle r_{n,0}, l_{n,0} \rangle} r_{n,0}$, and $l_{n,1} = \hat{l}_1^{n,1} - \frac{\langle \hat{l}_1^{n,1}, r_{n,0} \rangle}{\langle r_{n,01}, l_{n,1} \rangle} l_{n,1}$. With the help of the inner products given above, we find

$$\langle r_{n,1}, l_{n,1} \rangle = \frac{(-1)^n 36(n-3)!}{(n+4)!(n^2+2n-9)} \tag{7.6}$$

and

$$\langle r_{n,1}, \phi_i^1 \rangle = -\frac{3}{8} \frac{2^{n+i}(n+i)!(n+i-4)!(i^2+i+2ni-n-3)}{(2n+2i)!(n^2+2n-9)}. \tag{7.7}$$

Two functions w_i, $i = 1, 2$ will be needed to construct orthogonal generators from the above ramp functions, and these will be symmetric and antisymmetric with respect to $1/2$ in order to construct symmetric or antisymmetric generators. To this end let $w_1 = v_0(n) + \alpha_1(n) v_2(n)$, where $v_i(n)$ linear combinations of ϕ_{n+1+i}^1 and ϕ_{n+3+i}^1 and chosen so that $\langle v_i(n), r_{n,1} \rangle = 0$. Thus $v_0(n) = -\frac{(5n+9)(n-2)}{2(2n+5)(2n+3)} \phi_{n+1}^1 + \phi_{n+3}^1$ and $v_2(n) = -9 \frac{(n+3)(n+1)n}{2(2n+9)(2n+7)(5n+9)} \phi_{n+3}^1 + \phi_{n+5}^1$. Likewise, $w_2 = v_1(n) + \alpha_2(n) v_3(n)$, where $v_i(n)$ $i = 1, 3$ are orthogonal to $r_{n,0}$. In this case $v_1(n) = -\frac{(n+8)(n+1)n}{2(2n+7)(2n+5)(n+8)} \phi_{n+2}^1 + \phi_{n+4}^1$ and $v_3(n) = v_1(n+2)$. The biorthogonality of the ramps and the construction of v_i, $i = 0, 1, 2, 3$ imply that each $\alpha_i(n)$ must be chosen as a solution to the equation

$$\langle r_{n,i}, l_{n,i} \rangle \langle w_{i+1}, w_{i+1} \rangle = \langle w_{i+1}, r_i^1(\cdot - 1) \rangle \langle w_{i+1}, l_i^1 \rangle.$$

Utilizing (7.6), (7.7), and $\langle \phi_n^1, \phi_n^1 \rangle = \frac{n!(n+8)!}{256(2n+9)!!(2n+7)!!}$ to compute the inner products needed in the above equation we find using *Maple* that

$$\alpha_1(n) = \frac{(5n+9)(2n+7)}{(n+3)} \times \frac{(2n+11)q_1(n) \pm (n+4)(n+5)(5n+9)(2n+7)\left\{\frac{5(2n+11)(n+4)}{n(n+1)(2n+3)} q_2(n)\right\}^{\frac{1}{2}}}{2q_3(n)},$$

where

$$q_1(n) = 41n^5 + 625n^4 + 3733n^3 + 11099n^2 + 17010n + 11340,$$

$$q_2(n) = 17n^5 + 131n^4 - 105n^3 - 2979n^2 - 7884n - 6804,$$

and

$$q_3(n) = 37n^7 + 1376n^6 + 18862n^5 + 139394n^4 + 502291n^3 + 1099160n^2 + 1287090n + 635040.$$

Likewise,

$$\alpha_2(n) = \frac{(2n+9)(n+8)}{(n+3)(n+6)} \times \frac{\left(-(2n+13)q_4(n) \pm (n+5)(n+6)(n+8)(2n+9)\left\{\frac{7(2n+13)(n+4)(n+1)}{(2n+5)(n+2)} q_5(n)\right\}^{\frac{1}{2}}\right)}{2q_6(n)},$$

where

$$q_4(n) = 11n^6 + 115n^5 + 323n^4 + 893n^3 + 8642n^2 + 28968n + 25200,$$

$$q_2(n) = 3n^5 + 27n^4 + 7n^3 - 503n^2 - 1486n - 1400,$$

and

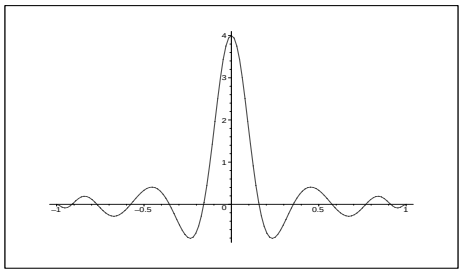
$$q_6(n) = 5n^7 + 39n^6 - 335n^5 - 5129n^4 - 29484n^3 - 112048n^2 - 242304n - 159600.$$

Knowing w_1 and w_2, we are now able to construct the orthogonal C^1 generator. Let $h_0(t) = 2|t|^3 - 3|t|^2 + 1$, if $t \in [-1, 1)$, and 0 elsewhere; $h_1(t) = (1-|t|)^2 t$, if $t \in [-1, 1)$, and 0 elsewhere; and $\phi_{i+1}^{n,1} = w_i$, $i = 1, 2$. Figure 7.2 shows $\phi_0^{n,1}$, $\phi_1^{n,1}$, $\phi_2^{n,1}$, and $\phi_1^{3,1}$ for $n = 6$. Then with

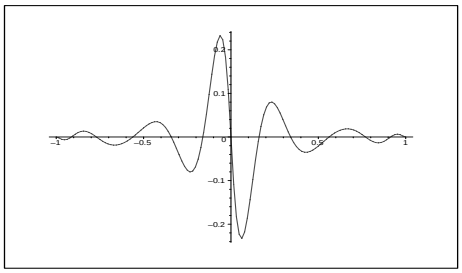
$$\phi_i^{n,1} = \left(I - P_{(\phi_2^1, \ldots, \phi_n^1, \phi_2^1(\cdot+1), \ldots, \phi_n^1(\cdot+1))}\right) h_i \qquad (i = 0, 1),$$

the above computations give the following theorem.

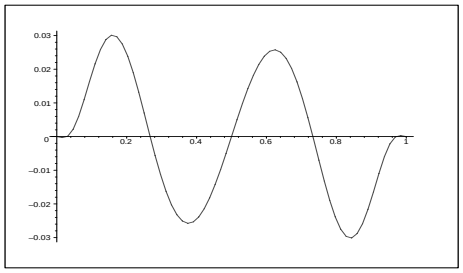
THEOREM 7.2. *For $n \geq 5$, and $\alpha_i(n)$ given above, $\Phi^1(n) = (\phi_0^{n,1}, \ldots, \phi_n^1)^T$ is a continuously differentiable orthogonal generator for $B(\Phi^1(n))$. Furthermore, the last $n-1$ functions are symmetric or antisymmetric about $1/2$. The first function $\phi_0^{n,1}$ is symmetric about 0, while $\phi_1^{n,1}$ is antisymmetric about 0.*



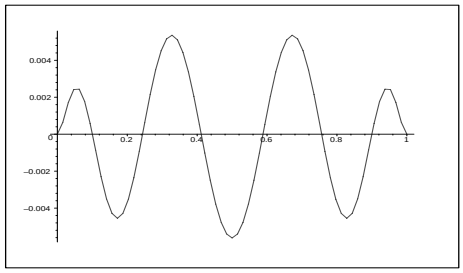
$\phi_0^{6,1}$



$\phi_1^{6,1}$



$\phi_2^{6,1}$



$\phi_3^{6,1}$

FIG. 7.2. *The functions* $\phi_i^{n,1}$ *for* $n = 6$ *and* $i = 0, \ldots, 3$.

We now construct the squeeze map associated with $\Phi^1(n)$. Since the last $n-2$ generators are supported on $[0,1]$ we need only concentrate on $\phi_0^{n,1}$ and $\phi_1^{n,1}$. Because of the definition of h_0 and h_1 and the symmetry of $\phi_0^{n,1}$ and $\phi_1^{n,1}$ it is easy to see that $W(n)$ is a diagonal matrix for all n. Therefore $R(n)$ is as in the previous C^1 example, and with $A_L^{(j)}$ a diagonal matrix $R(n)$ is equal to $A_R^{(j)}$. In order to complete the construction of the squeeze map we need to compute the inner products $\langle \Phi_L, \Phi_L \rangle$ and $\langle \Phi_R, \Phi_R \rangle$. From (3.9) in [6] (we would like to point out some errors in that equation; namely, $r_{i+1}^{n,k}$ in the first term on the right-hand side should be $r_i^{n,k}$, the factor multiplying the third term on the right-hand side should be $(n-k-1-i)$, and the factor multiplying the last term should be $(n+k+i+3)$) we find that

$$\langle r_0^{n,1}, r_0^{n,1} \rangle = 4 \frac{(n^2+2n-6)(n-2)!}{(n+3)!}, \tag{7.8}$$

$$\langle r_1^{n,1}, r_0^{n,1} \rangle = 6 \frac{(n-2)!}{(n+3)!}, \tag{7.9}$$

and

$$\langle r_1^{n,1}, r_1^{n,1} \rangle = 12 \frac{(n-3)!}{(n+4)!}. \tag{7.10}$$

To continue on we choose the minus sign in $\alpha_1(n)$ and the plus sign in $\alpha_2(n)$ to compute $\phi_i^{n,1}$ $i = 2, 3$. Then (7.7) and the norm squared of ϕ_n^1 can be employed to compute (using *Maple*) the norms of $\phi_i^{n,1}$, $i = 2, 3$ and the inner products of these functions with $\hat{r}_i^{n,1}, i = 0, 1$. With these in hand, (7.8), (7.9), and (7.10) can be used to compute $\langle \Phi_R, \Phi_R \rangle$, which is

$$\langle \Phi_R, \Phi_R \rangle = \begin{pmatrix} 4\frac{n^5+3n^4-10n^3-21n^2+27n+18}{(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} & -6\frac{n^3-9n+6}{(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} \\ -6\frac{n^3-9n+6}{(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} & 12\frac{n-3}{(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} \end{pmatrix}.$$

Since these functions are either symmetric or antisymmetric $\langle \Phi_L, \Phi_L \rangle$ is the same as the above matrix, except that the off diagonal elements take the opposite sign. Thus (2.3) becomes

$$BB^T = (L_j + L_{j-1}) \times \begin{pmatrix} 4\frac{(n^5+3n^4-10n^3-21n^2+27n+18)}{(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} & 6\frac{(L_j-L_{j-1})(n^3-9n+6)}{L_{j-1}(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} \\ 6\frac{(L_j-L_{j-1})(n^3-9n+6)}{L_{j-1}(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} & 12\frac{(L_j^2-L_jL_{j-1}+L_{j-1}^2)(n-3)}{L_{j-1}^2(n-2)(n^2+2n+9)(n+1)(n+2)(n+3)} \end{pmatrix}.$$

The determinant of the above matrix may be written as

$$\det(BB^T) = (L_j + L_{j-1})^2 \times \frac{12(n^3 - 9n - 18)(L_j^2 + L_{j-1}^2) + 6(5n^5 - 92n^3 + 54n^2 + 423n - 450)L_jL_{j-1}}{(L_{j-1}(n-2)(n^2+2n+9)(n+1)(n+2)(n+3))^2}$$

so that (4.3) may be used to compute B.

Acknowledgment. We thank the anonymous referees for their careful reading and suggestions.

REFERENCES

[1] W. Dahmen and C. A. Micchelli, *Using the refinement equation for evaluating integrals of wavelets,* SIAM J. Numer. Anal., 30 (1993), pp. 507–537.

[2] I. Daubechies, *Orthonormal bases of compactly supported wavelets,* Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[3] I. Daubechies, I. Guskov, P. Schröder, and W. Sweldens, *Wavelets on irregular point sets,* R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 2397–2413.

[4] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *Squeezable orthogonal bases and adaptive least squares,* in Wavelet Applications in Signal and Image Processing, Proc. SPIE 3169, A. Aldroubi, A. F. Laine, and M. A. Unser, eds., SPIE, Bellingham, WA, 1997, pp. 48–54.

[5] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *Intertwining multiresolution analyses and the construction of piecewise-polynomial wavelets,* SIAM J. Math. Anal., 27 (1996), pp. 1791–1815.

[6] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *Orthogonal polynomials and the construction of piecewise polynomial smooth wavelets,* SIAM J. Math. Anal., 30 (1999), pp. 1029–1056.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1100–1135

A PRECONDITIONER FOR THE ELECTRIC FIELD INTEGRAL EQUATION BASED ON CALDERON FORMULAS*

SNORRE H. CHRISTIANSEN† AND JEAN-CLAUDE NÉDÉLEC†

Abstract. We describe a preconditioning technique for the Galerkin approximation of the electric field integral equation (EFIE), which arises in the scattering theory for harmonic electromagnetic waves. It is based on a discretization of the Calderon formulas and the Helmholtz decomposition. We prove several properties of the method, in particular that it produces a variational solution on a subspace of the Galerkin space for which we have an LBB inf-sup condition. When the Krylov spaces associated with the continuous operators are nondegenerate we prove that the discrete Krylov spaces converge as the mesh refinement goes to zero; when, moreover, the EFIE is nondegenerate on the continuous Krylov spaces, the discrete Krylov iterates converge towards the continuous ones. We also argue that one might expect the continuous Krylov iterates to exhibit superlinear convergence of the form $n \mapsto C^n(n!)^{-\alpha}$ for some $C > 0$ and $\alpha > 0$. Finally, we illustrate the theory with numerical experiments.

Key words. electric field integral equation, Calderon formula, preconditioning, Krylov subspace

AMS subject classifications. 65N38, 78M15

PII. S0036142901388731

Introduction. In [51] Steinbach and Wendland described several strategies for the preconditioning of some boundary integral equations of the first kind, based on the knowledge of an operator of the opposite order. On several examples of symmetric positive definite (SPD) integral operators, they provided a discretization of the operators to construct a preconditioner such that the extreme eigenvalues of the preconditioned matrix remain bounded away from 0 and $+\infty$ independently of the mesh refinement. When iterative algorithms are used to solve the matrix equations, this in turn is well known to yield convergence estimates that are also independent of the mesh refinement, of the form $\|U^n - U^\star\| \leq C\alpha^n\|U^\star\|$ for some $C > 0$, $0 < \alpha < 1$ (so-called linear convergence).

In [17] we adapted the theory to some non-SPD problems. We showed that in these cases one can still prove that the spectral condition number of the preconditioned matrix remains bounded independently of the mesh refinement, as long as all the bilinear forms involved satisfy uniform inf-sup estimates on the Galerkin spaces. This was applied to some problems of three-dimensional acoustic scattering. Complementary results, in particular close to resonant frequencies, were exposed in [18].

We restricted our attention to preconditioners deduced from Calderon formulas—for some scalar operators—and remarked that they provide an inverse up to a compact endomorphism. One would expect this to produce a matrix with a spectrum clustered around 1. This would imply very fast, perhaps in some sense superlinear, convergence of Krylov subspace methods. However, we presented no formal proof of this intuition.

In this paper we use the Calderon formulas—for some operators on tangential vector fields—to construct a preconditioner for the electric field integral equation (EFIE). Compared with the scalar case there is a major pitfall: the involved bilinear forms do not all satisfy uniform inf-sup conditions on the standard Galerkin spaces

*Received by the editors April 27, 2001; accepted for publication (in revised form) March 3, 2002; published electronically August 28, 2002. This work was supported by Thales Airborne Systems.
http://www.siam.org/journals/sinum/40-3/38873.html
†CMAP, Ecole Polytechnique, 91128 Palaiseau, France (snorre@cmapx.polytechnique.fr, nedelec@cmapx.polytechnique.fr).

(Proposition 3.1). We nevertheless construct a numerically efficient preconditioner for which we can identify and prove satisfactory properties. In effect the Galerkin problem is solved on a subspace of the standard space, and we show that this subspace has all the properties required to ensure the well-posedness of the EFIE on it (Theorem 3.15). We also prove that the discretized preconditioner is stable and approximating in the sense of Proposition 4.6. The discrete Krylov spaces converge as the mesh refinement h decreases to 0 (Theorem 4.7) and from this we deduce that, under some natural hypotheses, the approximate solution at iteration n converges as $h \to 0$, towards a vector u^n (Corollary 4.10). We also explain why we expect (u^n) to converge superlinearly, perhaps even at the rate $C^n(n!)^{-\alpha}$, for some $C > 0$, $\alpha > 0$ (section 4.3).

We first officially detailed this discretization technique in [16] and announced the theorems justifying it in [19]. Proofs of one of them (Theorem 1.2) can be found in [15]. (This paper also contains complementary results on the behavior of the equation at resonant frequencies.) In the present paper we explicitate and prove—and sometimes improve—the remaining announced results, as well as a few others that sustain the construction.

At about the same time, in [21], another research team announced progress on what also amounts to the use of the Calderon formulas for the construction of a stable method. However, discretizations were proposed only for Nyström schemes, and it seems that for these they did not encounter a difficulty comparable to Proposition 3.1. Since this was the main problem we had to solve in the Galerkin setting, in extending the method from acoustics to electromagnetics, we believe there to be no overlap for the techniques involved.

Outline. The paper is organized as follows:

Section 1. We describe the continuous problem we are dealing with and state sufficient conditions for the discretization to satisfy uniform inf-sup estimates. We also prove the Calderon formulas.

Section 2. We define the Galerkin spaces we will use and give some useful properties: negative norm estimates, approximation properties for harmonic tangent fields, and properties of the discrete Helmholtz decomposition.

Section 3. We introduce the preconditioner we propose, after having described the main difficulty. Then we give a first interpretation of the system we solve and the projections we use. We give several characterizations of the range of the discrete rotation operator and deduce from these that the EFIE is well posed on this subspace. We also devise an intrinsic stopping criterion for the algorithm.

Section 4. We show the convergence of the Krylov subspaces and explain why this should lead to superlinear convergence of the iterates.

Section 5. We illustrate the theory with numerical results for diffraction by a sphere, a cavity, and a singular geometry.

1. The EFIE. We briefly recall the setting for exterior boundary value problems for the harmonic Maxwell equations, the integral representation of exterior electromagnetic fields, and the related integral equation known as the EFIE, as presented, for instance, in Cessenat [14] or Nédélec [43]. Then we turn to the discretization of this equation by the Galerkin method and state some new sufficient conditions (obtained in Christiansen [15]) for its well-posedness, in the sense of satisfying a uniform inf-sup condition. This in turn is well known to guarantee *quasi-optimal* convergence of the approximate solutions. Finally, we include a proof of the Calderon formulas and explain why they should lead to an efficient preconditioning technique.

1.1. The continuous problem. Let Ω_- be a smooth, bounded, and open subset of $\mathbb{R}^3$, denote by Γ its surface, and denote by Ω_+ the complement of $\Omega_- \cup \Gamma$. We refer to Ω_- as the interior domain and to Ω_+ as the exterior domain. The unit-length orthogonal vector field on Γ, pointing into Ω_+, is denoted n. We suppose throughout that Ω_+ is connected. The free-space harmonic Maxwell equations for vector fields E and H in Ω_+ are

$$\operatorname{curl} E = +i\omega\mu H, \tag{1.1}$$

$$\operatorname{curl} H = -i\omega\epsilon E, \tag{1.2}$$

where μ is the magnetic permeability, ϵ is the electric permittivity, and ω is the pulsation. Define the wavenumber k and the impedance Z by

$$k = \omega(\mu\epsilon)^{1/2}, \tag{1.3}$$

$$Z = (\mu/\epsilon)^{1/2}. \tag{1.4}$$

Then we have $+i\omega\mu = +ikZ$ and $-i\omega\epsilon = -ik/Z$.

Spaces of functions. On any smooth Riemannian manifold M which is either an open subset of an Euclidean space or is compact without boundary, the usual Sobolev spaces of scalar and tangential fields of regularity order $s \in \mathbb{R}$ are denoted $\mathrm{H}^s(M)$ and $\mathrm{H}^s_{\mathrm{T}}(M)$, respectively (see, e.g., Taylor [52, Chap. 4]), and the corresponding norms are both written

$$u \mapsto |u|_s. \tag{1.5}$$

On any open subset Ω of $\mathbb{R}^3$, define the Sobolev spaces $\mathrm{H}^s_{\mathrm{curl}}(\Omega)$ of vector fields by

$$\mathrm{H}^s_{\mathrm{curl}}(\Omega) = \{v \in \mathrm{H}^s_{\mathrm{T}}(\Omega) \ : \ \operatorname{curl} v \in \mathrm{H}^s_{\mathrm{T}}(\Omega)\}. \tag{1.6}$$

We will use also the Hilbert spaces $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$ of tangent fields on Γ defined by

$$\mathrm{H}^s_{\mathrm{div}}(\Gamma) = \{u \in \mathrm{H}^s_{\mathrm{T}}(\Gamma) \ : \ \operatorname{div} u \in \mathrm{H}^s(\Gamma)\}. \tag{1.7}$$

The spaces $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$ are equipped with the norms

$$u \mapsto \|u\|_s \ : \ \|u\|_s^2 = |u|_s^2 + |\operatorname{div} u|_s^2. \tag{1.8}$$

We define $\mathrm{H}^s_{\mathrm{rot}}(\Gamma)$ in a similar way, but we do not introduce any notation for the corresponding norm. Notice that $u \mapsto u \times n$ induces isomorphisms $\mathrm{H}^s_{\mathrm{rot}}(\Gamma) \to \mathrm{H}^s_{\mathrm{div}}(\Gamma)$ and $\mathrm{H}^s_{\mathrm{div}}(\Gamma) \to \mathrm{H}^s_{\mathrm{rot}}(\Gamma)$.

For any space of (scalar) distributions X on Γ, we denote by $X^\bullet$ the subspace of elements u such that for all v that are constant on each connected component of Γ (i.e., satisfying $\Delta v = 0$) $\langle u, v\rangle = 0$. Δ has a meaning even for vector distributions (currents). As usual, the Hilbert spaces we consider are vector spaces over $\mathbb{C}$ obtained as complexifications of real Hilbert spaces of scalar and tangent fields. In particular, they are equipped with conjugations $u \mapsto \overline{u}$, induced by the standard conjugation in $\mathbb{C}$.

Recall the result of Paquet [45] (see [43, Thm. 5.4.2, p. 209]) that we have well-defined continuous and surjective tangential trace operators for arbitrary large enough $R > 0$ (with $B_R = \{x \in \mathbb{R}^3 \ : \ |x| < R\}$):

$$\gamma_{\mathrm{T}} : \left\{ \begin{array}{rcl} \mathrm{H}^0_{\mathrm{curl}}(\Omega_+ \cap B_R) & \to & \mathrm{H}^{-1/2}_{\mathrm{rot}}(\Gamma), \\ v & \mapsto & v_{\mathrm{T}} = v - (v \cdot n)n. \end{array} \right. \tag{1.9}$$

For simplicity we denote by $\mathrm{H}^0_{\mathrm{curl}}(\Omega_+)_{loc}$ the Fréchet space of vector fields in Ω_+ whose restrictions are in $\mathrm{H}^0_{\mathrm{curl}}(\Omega_+ \cap B_R)$ for all $R > 0$. We caution the reader that this notation is sometimes (but not in this article) used for a different space consisting of the fields whose restrictions are in $\mathrm{H}^0_{\mathrm{curl}}(\Omega_+ \cap U)$ for all open U such that $\overline{U}$ is a compact subset of Ω_+. This space is larger, since it allows rather wild behavior close to Γ. In particular, there is no trace operator on this space.

A technique we shall use many times is to write tangential vector fields u in the form of their Helmholtz decomposition:

$$u = \operatorname{grad} p + \operatorname{rot} q + \alpha, \tag{1.10}$$

with $\Delta\alpha = 0$, and to use regularity theorems of the Laplacian to characterize p and q. We shall refer to this technique as the *HDRL*. It was used to study electromagnetic scattering by DeLaBourdonnaye [24].

Integral representation for exterior scattering problems. The basic property for exterior boundary value problems for the harmonic Maxwell equations is (see [43, Thm. 5.4.6, p. 220]) that for any $k > 0$, and any $v \in \mathrm{H}^{-1/2}_{\mathrm{rot}}(\Gamma)$, there is a unique $(E, H) \in \mathrm{H}^0_{\mathrm{curl}}(\Omega_+)^2_{loc}$ such that

- (E, H) solves the harmonic Maxwell equations in Ω_+,
- (E, H) satisfies the Silver–Müller radiation conditions,
- and $\gamma_{\mathrm{T}} E = v$.

Let G_k be the fundamental solution of the Helmholtz operator $-\Delta - k^2$ satisfying the Sommerfeld radiation condition

$$G_k(x, y) = \frac{e^{ik|x-y|}}{4\pi|x-y|}, \tag{1.11}$$

and let Φ_k be the potential, mapping any sufficiently smooth tangent field u on Γ to the field in $\mathbb{R}^3$ defined away from Γ by

$$(\Phi_k u)(y) = \int_\Gamma G_k(x, y) u(x) \mathrm{d}x. \tag{1.12}$$

Returning to the above boundary value problem, if k is not a resonance of the *interior* Maxwell equations there is a unique $u \in \mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$ such that for all $y \in \Omega_+$

$$E = (1 + (1/k^2) \operatorname{grad} \operatorname{div}) \Phi_k u, \tag{1.13}$$

$$H = 1/(ikZ) \operatorname{curl} \Phi_k u. \tag{1.14}$$

This formula is a special case of the Stratton–Chu integral representation (see [43, Thm. 5.5.1, p. 234]). For any $k \neq 0$ we define the electric field integral operator A_k by

$$A_k u = \gamma_{\mathrm{T}} (1 + (1/k^2) \operatorname{grad} \operatorname{div}) \Phi_k u. \tag{1.15}$$

One shows that the EFIO is continuous $A_k : \mathrm{H}^s_{\mathrm{div}} \to \mathrm{H}^s_{\mathrm{rot}}$ and that if k is not a resonance of the interior problem it is an isomorphism (see [43, Thm. 5.6.2, p. 247]).

Thus if k is not a resonant frequency the exterior problem for a given $v \in \mathrm{H}^{-1/2}_{\mathrm{rot}}(\Gamma)$ is reduced to the problem of solving the integral equation $A_k u = v$, called EFIE.

Variational formulation. From the HDRL it follows that the bilinear form on smooth tangent fields,

$$(u,v) \mapsto \langle u, v\rangle = \int_\Gamma u \cdot v, \tag{1.16}$$

extends continuously to a duality between $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$ and $\mathrm{H}^{-1-s}_{\mathrm{rot}}(\Gamma)$ (see also [43, Lem. 4.5.1, p. 208]). Thus one obtains the variational formulation of the EFIE. For a given $v \in \mathrm{H}^s_{\mathrm{rot}}(\Gamma)$, solve

$$u \in \mathrm{H}^s_{\mathrm{div}}(\Gamma) \quad \text{and} \quad \forall u' \in \mathrm{H}^{-1-s}_{\mathrm{div}}(\Gamma) \quad \langle A_k u, u'\rangle = \langle v, u'\rangle. \tag{1.17}$$

Remark that the case $s = -1/2$ is symmetric. From a practical point of view it is important that we have the following expression, with only weakly singular integrals (all integrals are on Γ):

$$\begin{aligned}\langle A_k u, u'\rangle = &\iint G_k(x,y) u(x)\cdot u'(y)\mathrm{d}x\mathrm{d}y \\ &-(1/k^2)\iint G_k(x,y)\,\mathrm{div}\, u(x)\,\mathrm{div}\, u'(y)\mathrm{d}x\mathrm{d}y.\end{aligned} \tag{1.18}$$

1.2. Discretization. Put $X = \mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$. Given some sequence (X_h) of closed (finite-dimensional) subspaces of X, the Galerkin method to solve (1.17) is to consider the problems

$$u \in X_h \quad \text{and} \quad \forall u' \in X_h \quad \langle A_k u, u'\rangle = \langle v, u'\rangle. \tag{1.19}$$

When it is uniquely solvable for each h one obtains a sequence (u_h), and it is of fundamental importance to know to which extent it converges towards $A_k^{-1}v$.

In this context we have the following fundamental theorem due to Babuska [2].

THEOREM 1.1. *Let X be a reflexive Banach space and $\mathcal{A} : X \to X^\star$ be linear and continuous. Suppose we have a closed subspace X_h of X and that for some $\alpha > 0$ we have*

$$\inf_{u\in X_h}\sup_{u'\in X_h}\frac{|(\mathcal{A}u)(u')|}{\|u\|\,\|u'\|} \geq \alpha \tag{1.20}$$

$$\forall u' \in X_h \quad \big(\forall u \in X_h\ (\mathcal{A}u)(u') = 0\big) \Rightarrow \big(u' = 0\big). \tag{1.21}$$

Then the induced map $\mathcal{A}_h : X_h \to X_h^\star$ is invertible (with an inverse of norm less than α^{-1}). Moreover, for all $l \in X^\star$, if we have a solution $u \in X$ to $\mathcal{A}u = l$, then

$$\|\mathcal{A}_h^{-1}(l|_{X_h}) - u\| \leq \big(1 + \alpha^{-1}\|\mathcal{A}\|\big)\inf\{\|u' - u\| \ :\ u' \in X_h\}. \tag{1.22}$$

Notice that finite-dimensional subspaces are closed and that for these condition (1.21) is implied by (1.20).

Now suppose we have a family (X_h) of closed subspaces of X. When there is an α that holds for all h in estimate (1.20), and (1.21) holds for all h, we say that we have a *uniform discrete inf-sup condition*. Then the only remaining point is whether the spaces X_h are *approximating*, in the sense that

$$\forall u \in X \quad \lim_h \inf\{\|u - u'\| \ :\ u' \in X_h\} = 0. \tag{1.23}$$

In general this question is well studied in the literature (with improved convergence estimates on some dense subspaces of X). However, in order to justify the preconditioning technique we shall describe in this paper, we will need to study this question for some nonstandard spaces.

Inf-sup conditions for the EFIE. The Galerkin discretization of the EFIE by Raviart–Thomas-type vector fields was studied by Bendali [4, 5]. More generally, we consider the following hypotheses for the Galerkin spaces (X_h).

(H0) The spaces X_h are finite-dimensional subspaces of $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$, which are stable under the conjugation $u \mapsto \overline{u}$ (*conj.-stable* for short).

(H1) There is $C > 0$ such that for all $u \in \mathrm{H}^1_{\mathrm{div}}(\Gamma)$

$$\inf_{u' \in X_h} \|u - u'\|_0 \leq Ch\|u\|_1. \tag{1.24}$$

(H2) There is $C > 0$ such that, for all $u \in X_h$, $\|u\|_0 \leq Ch^{-1}\|u\|_{-1}$.

(H3) Putting $W_h = \{u \in X_h \ : \ \operatorname{div} u = 0\}$, there is $C > 0$ such that for all $u \in X_h$, if

$$\forall w \in W_h \quad \langle u, w\rangle = 0, \tag{1.25}$$

then the solution ϕ of

$$\phi \in \mathrm{H}^1(\Gamma)^\bullet \quad \text{and} \quad \Delta\phi = \operatorname{div} u \tag{1.26}$$

satisfies

$$|u - \operatorname{grad}\phi|_0 \leq Ch|\operatorname{div} u|_0. \tag{1.27}$$

Notice that (H3) implies the usual inf-sup estimate: There is $C > 0$ such that

$$\inf_{q \in \operatorname{div} X_h} \sup_{u \in X_h} \frac{|\langle q, \operatorname{div} u\rangle|}{|q|_0\|u\|_0} \geq \frac{1}{C}. \tag{1.28}$$

The following theorem was proved in Christiansen [15].

THEOREM 1.2. *If k is not a resonance of the interior problem and a family (X_h) of Galerkin spaces satisfies the four conditions* (H0),...,(H3)*, then the bilinear form induced by A_k on X satisfies a uniform inf-sup condition on X_h.*

Of course these hypotheses also guarantee that in addition the (X_h) are approximating, so the approximate solution converges to the exact one (see section 2.2 for some details on this question).

The fact that the fields obtained by suitable transportation of Raviart–Thomas fields onto Γ satisfy these hypotheses is also checked in Christiansen [15], relying mostly on classical results that can be found, for instance, in Brezzi and Fortin [12]. Variants of the estimate on discrete Helmholtz decompositions appearing in (H3) have been used to study eigenvalue problems in mixed form and related discrete compactness results (see Kikuchi [37], Boffi [7], Boffi, Brezzi, and Gastaldi [8], and Demkowicz et al. [27]). Here, as already indicated, we will need to prove the hypothesis for some new spaces, in order to justify our preconditioning technique.

Solving the matrix equation. To solve the Galerkin problem one chooses a basis $e_h = (e_h(i))$ of X_h and defines the matrix $A_h(k)$ and the tuple V_h by

$$A_h(k)_{ij} = \langle A(k)e_h(j), e_h(i)\rangle, \qquad (V_h)_i = \langle v, e_h(i)\rangle. \tag{1.29}$$

In other words $A_h(k)$ is the matrix, from e_h to its dual basis, of the induced map

$$\mathcal{A}_h(k) : \left\{ \begin{array}{rcl} X_h & \to & X_h^\star, \\ u & \mapsto & \langle A(k)u, \cdot\rangle, \end{array} \right. \tag{1.30}$$

whereas V_h is the coordinate vector, in the dual basis of e_h, of the linear form $\langle v, \cdot \rangle$ restricted to X_h.

Then the discrete Galerkin problem (1.19) is stated in matrix terms as

$$A_h(k)U = V_h. \tag{1.31}$$

When this equation is solved iteratively one usually observes very slow convergence, if one observes it at all. Loosely speaking this is due to the fact that the operator A_k, via the Helmholtz decomposition, is seen to have one term of order 1 and another of order -1 acting on supplementary infinite-dimensional subspaces and with different signs. Thus, at least if the basis e_h is such that the canonical scalar product on $\mathbb{C}^{\dim X_h}$ corresponds to the $\mathrm{H}^0_{\mathrm{T}}(\Gamma)$ scalar product, the spectrum of $A_h(k)$ accumulates both at 0 and ∞. The presence of resonant frequencies further deteriorates the conditioning of the matrix.

This motivates our search for a preconditioner, that is, a matrix Z_h, such that, when Z_h is incorporated in an iterative solver, the reduction in the number of iterations required to obtain a satisfactory approximate solution outweighs the overhead of multiplying by Z_h. It is well known that this is achieved whenever Z_h is some easily computable approximate inverse of $A_h(k)$.

For ease of interpretation we will drop the matrix point of view and look instead for some easily computable $\mathcal{Z}_h : X_h^\star \to X_h$ which approximately inverts $\mathcal{A}_h(k)$. However, it should be kept in mind that the method is effective only in as far as it can be translated into a matrix manipulating algorithm.

1.3. Calderon formulas. The preconditioning technique we study in this paper is based on the Calderon formulas which we start by recalling. They are detailed in the electromagnetic setting in both Cessenat [14] and Nédélec [43]. We include a derivation of them mainly because the notations adopted here are not quite the same. Of course many of the arguments developed in this section were implicitly assumed while we introduced the EFIE and should be placed earlier in a strictly logical development.

Denote by B the operator on tangent fields on Γ defined by

$$Bu = u \times n. \tag{1.32}$$

Let $\wp$ be the orthogonal projection onto Γ, which is defined and smooth on a tubular neighborhood of Γ. Extending n to this neighborhood by $\wp$, we can define at any point x of this neighborhood, the tangential component of any vector v, by $\mathrm{T}_x v = v - (v \cdot n(x))n(x)$. Define an operator C_k on tangent fields on Γ, by taking the principal value of the tangential component in the exterior and interior domains (with respect to shrinking balls centered on Γ), of the following potential:

$$C_k u = \mathrm{PV}\,\mathrm{T}\,\mathrm{curl}\,\Phi_k u. \tag{1.33}$$

In fact, for smooth u the field $\mathrm{curl}\,\Phi_k u$ has different interior and exterior tangential traces which are both finite. More precisely, denoting γ_{T}^+ and γ_{T}^- the exterior and interior trace operators one has

$$\gamma_{\mathrm{T}}^+ \,\mathrm{curl}\,\Phi_k u = +(1/2)Bu + C_k u, \tag{1.34}$$

$$\gamma_{\mathrm{T}}^- \,\mathrm{curl}\,\Phi_k u = -(1/2)Bu + C_k u. \tag{1.35}$$

In particular, one has the familiar *jump formula*

$$u = [(\mathrm{curl}\,\Phi_k u) \times n] = B(\gamma^- \,\mathrm{curl}\,\Phi_k u - \gamma^+ \,\mathrm{curl}\,\Phi_k u). \tag{1.36}$$

We also remind the reader that for potentials of the form

$$(1+(1/k^2)\operatorname{grad}\operatorname{div})\Phi_k u, \tag{1.37}$$

the exterior and interior tangential traces are equal (and given by the EFIO), thus there is no tangential "jump" for these.

To derive the Calderon formulas the last ingredient we need is the *representation theorem*.

THEOREM 1.3. *Suppose (E,H) is a field whose restrictions to Ω_- and Ω_+ are in $\mathrm{H}^0_{\mathrm{curl}}(\Omega_-)^2$ and $\mathrm{H}^0_{\mathrm{curl}}(\Omega_+)^2_{loc}$ and solve Maxwell's equations for a given wavenumber k. Suppose also that it verifies the Silver–Müller radiation conditions. Define the electric and magnetic currents j and m on Γ by the jump formulas*

$$j=[H\times n]=(\gamma_{\mathrm{T}}^- H-\gamma_{\mathrm{T}}^+ H)\times n, \tag{1.38}$$

$$m=[E\times n]\ =(\gamma_{\mathrm{T}}^- E-\gamma_{\mathrm{T}}^+ E)\times n. \tag{1.39}$$

Then in Ω_- and Ω_+ we have

$$E=\ (+ikZ)(1+(1/k^2)\operatorname{grad}\operatorname{div})\Phi_k j\ +\operatorname{curl}\Phi_k m, \tag{1.40}$$

$$H=(-ik/Z)(1+(1/k^2)\operatorname{grad}\operatorname{div})\Phi_k m+\operatorname{curl}\Phi_k j. \tag{1.41}$$

Now the theorem we are interested in is the following.

THEOREM 1.4. *The following operator is a projector:*

$$\begin{bmatrix} 1/2-BC_k & -(-ik/Z)BA_k \\ -(+ikZ)BA_k & 1/2-BC_k \end{bmatrix}. \tag{1.42}$$

More explicitly, we have

$$BC_kBC_k+k^2BA_kBA_k=1/4, \tag{1.43}$$

$$BC_kBA_k+BA_kBC_k=0. \tag{1.44}$$

Proof. Choose two (smooth enough) tangent fields u and v on Γ. Define fields E and H by putting, *in the exterior domain*,

$$E=\ (+ikZ)(1+(1/k^2)\operatorname{grad}\operatorname{div})\Phi_k u\ +\operatorname{curl}\Phi_k v, \tag{1.45}$$

$$H=(-ik/Z)(1+(1/k^2)\operatorname{grad}\operatorname{div})\Phi_k v+\operatorname{curl}\Phi_k u. \tag{1.46}$$

Then we have

$$-\gamma_{\mathrm{T}}^+ H\times n=-(-ik/Z)BA_k v+(1/2-BC_k)u, \tag{1.47}$$

$$-\gamma_{\mathrm{T}}^+ E\times n=-(+ikZ)BA_k u+(1/2-BC_k)v. \tag{1.48}$$

In the interior domain put $E=0$ and $H=0$. Now we have

$$[H\times n]=-\gamma_{\mathrm{T}}^+ H\times n, \tag{1.49}$$

$$[\,E\times n]=-\gamma_{\mathrm{T}}^+ E\times n. \tag{1.50}$$

Using the representation theorem, these tangent fields give rise to new integral representations for E and H. Now to say that the announced operator is a projector just expresses that taking the jumps (or the appropriate exterior traces) of these new

integral representations for the same fields (E, H) should yield the same jumps (or exterior traces). □

The operator appearing in (1.42) is called the *exterior Calderon projector*.

The crucial remark is now that the operator BC_kBC_k is a compact endomorphism of $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$, thus $4k^2BA_kB$ inverts A_k up to a compact operator. The scalar coefficient $4k^2$ is unimportant for preconditioning purposes, so our aim will be to discretize the operator BA_kB. Since we deal with variational formulations we express our goal in terms of bilinear forms, for which it is preferable to have symmetric formulations, so, remarking that $B = -B^\star = B^{\star-1}$, we set out to discretize the map

$$\mathcal{Z}_k = \mathcal{B}^{\star-1}\mathcal{A}_k\mathcal{B}^{-1}, \tag{1.51}$$

where $\mathcal{B}$ is the isomorphism

$$\left\{ \begin{array}{rcl} \mathrm{H}^s_{\mathrm{div}}(\Gamma) & \to & \mathrm{H}^{-1-s}_{\mathrm{div}}(\Gamma)^\star, \\ u & \mapsto & \langle Bu, \cdot \rangle. \end{array} \right. \tag{1.52}$$

2. Some properties of some Galerkin spaces. We recall the definition and basic properties of the Galerkin spaces on Γ that we will use in this article, including the null sequences relating spaces of scalar and tangent finite elements, as well as some negative norm estimates. Of particular importance will be the approximation of harmonic tangent fields and the structure of the discrete Helmholtz decomposition that largely follows from it.

2.1. Surface finite element spaces. Recall that we denote by $\wp$ the orthogonal projection onto Γ, which is defined and smooth on a tubular neighborhood of Γ. Let $(\mathcal{T}_h)$ be a family of triangulations of Γ, where for all h the largest diameter of a triangle of $\mathcal{T}_h$ is h. We will always suppose that $(\mathcal{T}_h)$ is regular and most often that it is (globally) quasi-uniform. Let Γ_h be the affine polyhedron determined by $\mathcal{T}_h$, considered as a Lipschitz manifold. For small enough h, $\wp$ induces Lipschitz isomorphisms $\Gamma_h \to \Gamma$, and we denote by Ξ_h the inverse mappings.

Fix a nonzero $m \in \mathbb{N}$. On Γ_h we consider the space $\mathsf{S}^0(\mathcal{T}_h)$ of continuous scalar functions whose restriction to any triangle is P^m (a polynomial of degree m), the space $\mathsf{S}^1(\mathcal{T}_h)$ of Raviart–Thomas $\mathrm{H}^0_{\mathrm{div}}$-conforming vector fields of degree m, and the space $\mathsf{S}^2(\mathcal{T}_h)$ of scalar functions whose restriction to any triangle is P^{m-1}.

From these finite element spaces on Γ_h we deduce finite element spaces on Γ by the transport formulas

$$\begin{array}{rcll} \mathsf{S}^0_h & = & \{x \mapsto & u(\Xi_h(x)) : u \in \mathsf{S}^0(\mathcal{T}_h)\}, \\ \mathsf{S}^1_h & = & \{x \mapsto & \mathrm{Jac}\,\Xi_h(x)\mathrm{D}\Xi_h(x)^{-1}u(\Xi_h(x)) : u \in \mathsf{S}^1(\mathcal{T}_h)\}, \\ \mathsf{S}^2_h & = & \{x \mapsto & \mathrm{Jac}\,\Xi_h(x)u(\Xi_h(x)) : u \in \mathsf{S}^2(\mathcal{T}_h)\}. \end{array} \tag{2.1}$$

These transport formulas were chosen to make the following diagram commute. The horizontal arrows are the differential operators rot and div, whereas the vertical ones are the above transport formulas.

$$\begin{array}{ccccccccc} 0 & \to & \mathsf{S}^0(\mathcal{T}_h) & \to & \mathsf{S}^1(\mathcal{T}_h) & \to & \mathsf{S}^2(\mathcal{T}_h) & \to & 0 \\ \downarrow & 0 & \downarrow & \mathrm{rot} & \downarrow & \mathrm{div} & \downarrow & 0 & \downarrow \\ 0 & \to & \mathsf{S}^0_h & \to & \mathsf{S}^1_h & \to & \mathsf{S}^2_h & \to & 0 \end{array} \tag{2.2}$$

Of course this diagram is a realization of a corresponding diagram on differential forms, on which the exterior derivative act, transported by the standard pull-back

of differential forms determined by Ξ_h. The connection between finite elements and differential forms, especially Whitney forms, was stressed by Bossavit [9] and further explicitated in the affine case in Hiptmair [32]. While this is useful to keep in mind we stick to tangent vector fields and scalar fields on Γ, since in accordance with widespread conventions we have chosen to represent the exterior electromagnetic fields as fields of vectors, not alternate forms. The relevance of commuting diagrams to the study of finite elements is noted in Boffi [7].

Remark also that when studying the approximation of the boundary Γ by piecewise polynomial triangulations as in Nédélec [40], one is led to consider Galerkin spaces defined by pull-backs by maps that are slightly different from Ξ_h.

2.2. Basic negative norm estimates. Since negative (and noninteger) Sobolev norms and corresponding approximation results pervade this article, we now recall rather informally the results needed. Of course we do not claim any originality for these results, and we have included them mainly for the convenience of the exposition.

Let (X_h) be a family of Galerkin spaces satisfying (H0) and (H1). Let $\mathcal{Q}_h$ be the $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$-orthogonal projection onto X_h.

It is well known that

$$\|\mathcal{Q}_h u\|_0 \leq C\|u\|_0 \quad \text{and} \quad \|u - \mathcal{Q}_h u\|_0 \leq Ch\|u\|_1. \tag{2.3}$$

From the HDRL (section 1.1) it follows that the spaces $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$ for $0 \leq s \leq 1$ can be obtained by interpolation. Hence interpolation on the operator $\mathcal{I} - \mathcal{Q}_h$, for $0 \leq s \leq 1$, gives

$$\|u - \mathcal{Q}_h u\|_0 \leq Ch^s\|u\|_s. \tag{2.4}$$

Then one uses the regularity of the $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$-inner product (written $(\cdot|\cdot)_0$) on various Sobolev spaces. This technique is the familiar Aubin–Nitsche trick. That $\mathrm{H}^s_{\mathrm{div}}(\Gamma)$ and $\mathrm{H}^{-s}_{\mathrm{div}}(\Gamma)$ are dual with respect to the $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$-inner product can be deduced from the fact that the operator $I - \operatorname{grad}\operatorname{div}$ is an isomorphism $\mathrm{H}^s_{\mathrm{div}}(\Gamma) \to \mathrm{H}^{s-1}_{\mathrm{rot}}(\Gamma)$ and that this space, as already remarked, is the $\mathrm{L}^2_{\mathrm{T}}$-dual of $\mathrm{H}^{-s}_{\mathrm{div}}(\Gamma)$. Both of these facts can be proved using the HDRL. For $0 \leq s \leq 1$ we have

$$\|u - \mathcal{Q}_h u\|_{-s} \leq C \sup_{v \in \mathrm{H}^s_{\mathrm{div}}} \frac{|(u - \mathcal{Q}_h u|v)_0|}{\|v\|_s} \tag{2.5}$$

$$\leq C \sup_{v \in \mathrm{H}^s_{\mathrm{div}}} \frac{|(u - \mathcal{Q}_h u|v - \mathcal{Q}_h v)_0|}{\|v\|_s} \tag{2.6}$$

$$\leq C\|u - \mathcal{Q}_h u\|_0 \, \|\mathcal{I} - \mathcal{Q}_h\|_{0,s}. \tag{2.7}$$

Here $\|\mathcal{I} - \mathcal{Q}_h\|_{0,s}$ is of course the norm of the induced map

$$\mathcal{I} - \mathcal{Q}_h : \mathrm{H}^s_{\mathrm{div}}(\Gamma) \to \mathrm{H}^0_{\mathrm{div}}(\Gamma). \tag{2.8}$$

This gives for $0 \leq s, s' \leq 1$

$$\|u - \mathcal{Q}_h u\|_{-s} \leq Ch^{s+s'}\|u\|_{s'}. \tag{2.9}$$

If in addition we have the inverse inequality (H2) we obtain

$$\|\mathcal{Q}_h u\|_{-1} \leq \|u\|_{-1} + \|u - \mathcal{Q}_h u\|_{-1} \leq \|u\|_{-1} + Ch\|u\|_0 \leq C\|u\|_{-1}. \tag{2.10}$$

This is proved first for $u \in \mathrm{H}^0_{\mathrm{div}}(\Gamma)$ and then extended to $u \in \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)$ by a density argument. By interpolation on $\mathcal{Q}_h$ one then extends this stability result to all $-1 \leq s \leq 0$:

$$\|\mathcal{Q}_h u\|_s \leq C\|u\|_s. \tag{2.11}$$

2.3. Approximation of harmonic fields. Since we deal with surfaces which we do not require to be simply connected, a useful construct is that of the associated cohomology groups of which we give the realizations in terms of vector and scalar fields, the so-called harmonic fields. (This notion of harmonicity is only remotely related to the harmonicity of the electromagnetic waves we consider.) We denote these spaces by $\mathbb{G}^i$ for $i = 0, 1, 2$. For instance $\mathbb{G}^1$ can be characterized as the L_2-orthogonal of the range of the rot operator on smooth scalar fields (or $\mathrm{H}^1(\Gamma)$), in the kernel of the div operator on smooth tangent fields (resp., $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$).

Noticing that the two rows in the diagram (2.2) are null sequences, we consider, for each horizontal pair of consecutive arrows, the L_2-orthogonal of the range of the left arrow, in the kernel of the right arrow. For the second row we denote these vector spaces by G^0_h, G^1_h, and G^2_h.

It is a remarkable fact that these "discrete" cohomology groups have the "right" dimension, i.e., the dimension of their continuous analogues $\mathbb{G}^i$. This is either elementary or can be deduced from the *Euler–Poincaré formula*. The use of this formula should not come as a surprise, since it is one of the main reasons for the effectiveness of *simplical* triangulations in algebraic topology. It has been used in finite element theory for quite some time, even at the textbook level; see, for instance, Nédélec [42].

For each h, let N^0_h be the number of vertices, N^1_h the number of edges (segments), and N^2_h the number of faces (triangles) in $\mathcal{T}_h$. Let N^C be the number of connected components of Γ.

We leave it as an exercise to check that for G^0_h and G^2_h the dimension is the number N^C of connected components of Γ. Remark also that $\mathsf{G}^0_h = \mathbb{G}^0$, whereas $\mathsf{G}^2_h \neq \mathbb{G}^2$. However, the elements of $\mathbb{G}^2$, which are the functions that are constant on each connected component of Γ, are of course well approximated by the elements of G^2_h.

We now turn to the more interesting case of G^1_h. We have

$$\dim \mathsf{G}^1_h = \big(\dim \mathsf{S}^1_h - (\dim \mathsf{S}^2_h - N^C)\big) - \big(\dim \mathsf{S}^0_h - N^C\big) \tag{2.12}$$

$$= -\dim \mathsf{S}^0_h + \dim \mathsf{S}^1_h - \dim \mathsf{S}^2_h + 2N^C \tag{2.13}$$

$$= -\big(N^0_h + (m-1)N^1_h + (m-2)(m-1)/2N^2_h\big) \tag{2.14}$$

$$+\big(mN^1_h + m(m-1)N^2_h\big) - \big(m(m+1)/2N^2_h\big) + 2N^C \tag{2.15}$$

$$= -N^0_h + N^1_h - N^2_h + 2N^C \tag{2.16}$$

$$= \dim \mathbb{G}^1. \tag{2.17}$$

To see that G^1_h converges in some sense to $\mathbb{G}^1$, consider the map Ω_h, called a Fortin operator in Boffi [6], which to any $u_0 \in \mathrm{H}^0_{\mathrm{div}}(\Gamma)$ associates the first component u of the solution (u, q) of

$$\left\{ \begin{array}{l} u \in \mathsf{S}^1_h \\ q \in \mathsf{S}^{2\bullet}_h \end{array} \right. \quad \left\{ \begin{array}{lll} \forall u' \in \mathsf{S}^1_h & \langle u, u'\rangle + \langle q, \operatorname{div} u'\rangle & = \langle u_0, u'\rangle; \\ \forall q' \in \mathsf{S}^{2\bullet}_h & \langle q', \operatorname{div} u\rangle & = \langle q', \operatorname{div} u_0\rangle. \end{array} \right. \tag{2.18}$$

This saddle-point problem satisfies the LBB inf-sup conditions; therefore there is a $C > 0$ such that for all h and all $u \in \mathrm{H}^0_{\mathrm{div}}(\Gamma)$ we have

$$\|u - \Omega_h u\|_0 \leq C \inf\{\|u - u'\|_0 \ : \ u' \in \mathsf{S}^1_h\}. \tag{2.19}$$

Notice also that Ω_h maps divergence-free fields to divergence-free fields and that if $u \in \mathrm{H}^0_{\mathrm{div}}(\Gamma)$ is such that $\mathrm{rot}\, u = 0$ (as elements of $\mathrm{H}^{-1}(\Gamma)$), then $\Omega_h u$ is L_2-orthogonal to $\mathrm{rot}\, \mathsf{S}^0_h$. In particular, Ω_h maps $\mathbb{G}^1$ into G^1_h. Since all norms on $\mathbb{G}^1$ are equivalent, we therefore have an estimate of the form

$$\forall u \in \mathbb{G}^1 \quad \|u - \Omega_h u\|_0 \leq C h^m \|u\|_0 \tag{2.20}$$

($m = 1 > 0$ for lowest order elements[1]) so that, for sufficiently small h, Ω_h determines injections $\mathbb{G}^1 \to \mathsf{G}^1_h$ which are arbitrarily close in norm to the identity mapping on $\mathbb{G}^1$. Since the spaces have the same dimension these induced maps are in fact isomorphisms, and the inverse mappings are Ch^m-close to the identity mapping on G^1_h.

Remark. For reasons of dimension, for the above system (2.18) to satisfy the LBB inf-sup conditions it is necessary that, *among all spaces that contain* $\mathrm{div}\, \mathsf{S}^1_h$, $\mathsf{S}^{2\bullet}_h$ be minimal. (Of course the LBB condition can be verified for some smaller spaces that do not contain $\mathrm{div}\, \mathsf{S}^1_h$.) On the other hand, for the above constructed injection $\mathbb{G}^1 \to \mathsf{G}^1_h$ to be *onto* it is necessary that $\mathsf{S}^{0\bullet}_h$ be maximal *among all spaces that* rot *maps into* S^1_h. It is remarkable that these algebraic optimality conditions (which were our guide for the choice of spaces) are also sufficient for convergence purposes.

Given a tangent field u one may now ask how the field a_0 defined by

$$a_0 \in \mathbb{G}^1 \quad \text{and} \quad \forall a' \in \mathbb{G}^1 \quad \langle a_0, a' \rangle = \langle u, a' \rangle \tag{2.21}$$

relates to its discrete analogue a_h defined by

$$a_h \in \mathsf{G}^1_h \quad \text{and} \quad \forall a' \in \mathsf{G}^1_h \quad \langle a_h, a' \rangle = \langle u, a' \rangle. \tag{2.22}$$

Since G^1_h is not a subspace of $\mathbb{G}^1$ one can view this as a nonconforming Galerkin problem. We have already proved that G^1_h converges in a sense to $\mathbb{G}^1$. Later, in Proposition 4.8 we provide the necessary variant of Theorem 1.1 to deduce from this the existence of a_h and its convergence to a_0.

PROPOSITION 2.1. *For the exact and approximate harmonic tangent fields a_0 and a_h obtained as solutions of* (2.21) *and* (2.22)*, for a given tangent field u, we have the estimates*

$$\|a_h - a_0\|_0 \leq C h^m \|u\|_{\mathrm{H}^{-1}_{\mathrm{rot}}(\Gamma)}. \tag{2.23}$$

Another useful observation is that, parallel to the fact that all norms on the finite-dimensional space $\mathbb{G}^1$ are equivalent, we have easily obtained the following.

LEMMA 2.2. *There is $C > 0$ such that for all $-1 \leq s \leq 0$ and all h*

$$\forall u \in \mathsf{G}^1_h \quad \|u\|_0 \leq C \|u\|_s. \tag{2.24}$$

2.4. Discrete Helmholtz decomposition. We will frequently use the fact that each $u \in \mathsf{S}^1_h$ can be written in a unique way:

$$u = \mathrm{rot}\, p + a + g, \tag{2.25}$$

with $p \in \mathsf{S}^{0\bullet}_h$, $a \in \mathsf{G}^1_h$, and g in the L_2-orthogonal of the kernel of the divergence operator in S^1_h. Notice that this decomposition expresses that S^1_h is split into a direct

[1] In fact, to derive the inf-sup estimates and related stability results we need, it will not be necessary to know that higher order elements yield higher order estimates.

sum of three subspaces which are orthogonal both for the $\mathrm{H}^0_{\mathrm{T}}$ and the $\mathrm{H}^0_{\mathrm{div}}$ scalar products.

What makes this decomposition useful is that it has the following continuity and approximation properties, compared with the "exact" Helmholtz decomposition, which we write

$$u = \operatorname{rot} p_0 + a_0 + g_0, \tag{2.26}$$

with $p_0 \in \mathrm{H}^{1/2}(\Gamma)^\bullet$, $a_0 \in \mathbb{G}^1$, and $g_0 \in \operatorname{grad} \mathrm{H}^{3/2}(\Gamma)$.

PROPOSITION 2.3. *There is $C > 0$ such that for all h, all $u \in \mathsf{S}^1_h$, the above decompositions* (2.25) *and* (2.26) *are related by*

$$\|g - g_0\|_0 \le Ch\|u\|_0, \qquad \|g\|_{-1} \le C\|u\|_{-1}, \tag{2.27}$$

$$\|a - a_0\|_0 \le Ch\|u\|_0, \qquad \|a\|_{-1} \le C\|u\|_{-1}, \tag{2.28}$$

$$\|\operatorname{rot} p - \operatorname{rot} p_0\|_0 \le Ch\|u\|_0, \qquad \|\operatorname{rot} p\|_{-1} \le C\|u\|_{-1}. \tag{2.29}$$

Proof. (i) As already remarked, by hypothesis (H3), we have

$$|g - g_0|_0 \le Ch|\operatorname{div} u|_0. \tag{2.30}$$

This immediately gives $\|g - g_0\|_0 \le Ch\|u\|_0$. Then we remark that

$$\|g\|_{-1} \le \|g - g_0\|_{-1} + \|g_0\|_{-1} \le Ch\|u\|_0 + \|u\|_{-1} \le C\|u\|_{-1}. \tag{2.31}$$

(ii) The fact that $\|a - a_0\|_0 \le Ch\|u\|_0$ was proved in the preceding section and gives $\|a\|_{-1} \le C\|u\|_{-1}$ just as above.

(iii) The last part of the proposition is deduced from the two preceding ones writing

$$\operatorname{rot} p - \operatorname{rot} p_0 = -(g + a) + (g_0 + a_0) = (g_0 - g) + (a_0 - a). \tag{2.32}$$

□

3. Stable discretizations of the Calderon formulas. First we explain why the most natural idea (at least to us, for quite some time) is actually flawed. Then we define the discretization which we propose to use for preconditioning. It is associated with a subspace of the usual Galerkin space for which we prove an LBB inf-sup condition and some basic approximation properties.

3.1. A flawed idea. Given a family of Galerkin spaces X_h in X the most straightforward idea is to introduce the maps

$$\mathcal{B}_h : \left\{ \begin{array}{rcl} X_h & \to & X_h^\star, \\ u & \mapsto & \langle Bu, \cdot \rangle \end{array} \right. \tag{3.1}$$

and then to put

$$\mathcal{Z}_h(k) = \mathcal{B}_h^{\star -1} \mathcal{A}_h(k) \mathcal{B}_h^{-1}. \tag{3.2}$$

As remarked in Christiansen and Nédélec [17], if not only $\mathcal{A}_h(k)$ but also $\mathcal{B}_h$ satisfies a uniform discrete inf-sup condition on X_h, then the spectral condition number of $\mathcal{Z}_h(k)\mathcal{A}_h(k)$ is bounded independently of h. Of course, since the operators we deal with are not SPD, this is not enough to guarantee convergence of Krylov subspace algorithms, but it is nevertheless a significant progress compared with the lack of a

preconditioner. Unfortunately, for the standard Galerkin spaces, this last inf-sup condition does not hold. Indeed, throughout this paragraph let X_h denote the Raviart–Thomas spaces of degree m (with our conventions the minimal degree is $m = 1$) on Γ. We will use the fact that X_h satisfies the hypotheses (H0),...,(H3).

PROPOSITION 3.1. *Let $X_h = \mathsf{S}_h^1$. Let K_h be the space*

$$\{u \in \mathsf{S}_h^1 \ : \ \forall v \in \mathsf{S}_h^1 \quad \operatorname{div} v = 0 \Rightarrow \langle u, v\rangle = 0 \quad \textit{and} \\ \forall v \in \mathsf{S}_h^0 \quad \langle u, \operatorname{grad} v\rangle = 0\}. \tag{3.3}$$

Then we have

$$\liminf_h \frac{\dim K_h}{\dim X_h} \geq \frac{1}{2m+1} \tag{3.4}$$

and

$$\limsup_h \sup_{u\in K_h} \sup_{u'\in X_h} \frac{b(u,u')}{\|u\|_{-1/2}\|u'\|_{-1/2}} h^{-1/2} < +\infty. \tag{3.5}$$

Proof. (i). We have

$$\dim K_h \geq (\dim \mathsf{S}_h^2 - N^C) - (\dim \mathsf{S}_h^0 - N^C) \tag{3.6}$$

$$\geq (m(m+1)/2)N_h^2 \\ -N_h^0 + (m-1)N_h^1 + ((m-2)(m-1)/2)N_h^2 \tag{3.7}$$

$$\geq -N_h^0 - (m-1)N_h^1 + (2m-1)N_h^2. \tag{3.8}$$

Recall that since each segment is shared by exactly two triangles, $2N_h^1 = 3N_h^2$, which together with the Euler–Poincaré formula gives

$$N_h^1 \sim 3N_h^0 \quad \text{and} \quad N_h^2 \sim 2N_h^0. \tag{3.9}$$

This gives

$$-N_h^0 - (m-1)N_h^1 + (2m-1)N_h^2 \sim mN_h^0. \tag{3.10}$$

One also checks that

$$\dim X_h \sim m(2m+1)N_h^0. \tag{3.11}$$

This gives the first inequality.

(ii). To prove the second part of the theorem we use the fact that X_h satisfies (H3). For any $u \in X_h$ we denote by ϕ_u the unique $\phi \in \mathrm{H}^1(\Gamma)^\bullet$ such that $\Delta\phi = \operatorname{div} u$. Then (H3) asserts that if $u \in X_h$ is L_2-orthogonal to the kernel of the divergence operator on X_h, then we have an estimate of the form $|u - \operatorname{grad}\phi_u|_0 \leq Ch|\operatorname{div} u|_0$. From Proposition 2.3, using an inverse inequality, one can deduce that

$$\|u - \operatorname{grad}\phi_u\|_{-1/2} \leq Ch^{1/2}\|u\|_{-1/2} \tag{3.12}$$

and

$$\|\operatorname{grad}\phi_u\|_{-1/2} \leq C\|u\|_{-1/2}. \tag{3.13}$$

Choose $u \in K_h$ and $u' \in X_h$. Put $u' = \operatorname{rot} p' + a' + g'$ as in (2.25). Remark first that

$$\langle u \times n, \operatorname{rot} p' \rangle = -\langle \operatorname{div} u, p' \rangle = 0. \tag{3.14}$$

Then write

$$\langle u \times n, g' \rangle = \langle (u - \operatorname{grad} \phi_u + \operatorname{grad} \phi_u) \times n, (g' - \operatorname{grad} \phi_{g'} + \operatorname{grad} \phi_{g'}) \rangle. \tag{3.15}$$

Developing and using the continuity of b as well as the fact that

$$\langle \operatorname{grad} \phi_u \times n, \operatorname{grad} \phi_{g'} \rangle = 0, \tag{3.16}$$

we obtain

$$|\langle u \times n, g' \rangle| \le C \|u - \operatorname{grad} \phi_u\|_{-1/2} \|g' - \operatorname{grad} \phi_{g'}\|_{-1/2} \tag{3.17}$$

$$+ C \|u - \operatorname{grad} \phi_u\|_{-1/2} \|\operatorname{grad} \phi_{g'}\|_{-1/2} \tag{3.18}$$

$$+ C \|\operatorname{grad} \phi_u\|_{-1/2} \|g' - \operatorname{grad} \phi_{g'}\|_{-1/2}. \tag{3.19}$$

By the above estimates (3.12) and (3.13) it follows that

$$|\langle u \times n, g' \rangle| \le C h^{1/2} \|u\|_{-1/2} \|g'\|_{-1/2}. \tag{3.20}$$

Therefore

$$|\langle u \times n, g' \rangle| \le C h^{1/2} \|u\|_{-1/2} \|u'\|_{-1/2}. \tag{3.21}$$

Finally, for the last part, for any $\tilde{a}'$, $\langle u \times n, a' \rangle$ equals

$$\langle (u - \operatorname{grad} \phi_u) \times n, \tilde{a}' \rangle + \langle \operatorname{grad} \phi_u \times n, \tilde{a}' \rangle + \langle u \times n, (a' - \tilde{a}') \rangle. \tag{3.22}$$

Choosing $\tilde{a}' \in \mathbb{G}^1$ to be an approximation of a', one immediately obtains the proposition. □

Thus one sees that the reason for the degeneracy is that the subspace of X_h of elements which are in a sense discrete gradients does not have the same dimension as the subspace of rotationals.

3.2. Auxiliary spaces. Let $(\mathsf{S}_h^0, \mathsf{S}_h^1, \mathsf{S}_h^2)$, and $(\mathsf{S}_h'^0, \mathsf{S}_h'^1, \mathsf{S}_h'^2)$ be two triples of spaces of the type we discussed; more precisely, they should satisfy the null sequence condition, the discrete cohomology groups should have the "right" dimension, and S_h^1 and $\mathsf{S}_h'^1$ should satisfy the hypotheses (H0),...,(H3).

Two examples to keep in mind (the first one detailed and the second one mentioned in Christiansen and Nédélec [19]) are

- the case where S_h^1 and $\mathsf{S}_h'^1$ are equal and consist of lowest order Raviart–Thomas fields (then S_h^0 consists of continuous P^1 FE and S_h^2 of P^0 FE);
- the case where $(\mathsf{S}_h^0, \mathsf{S}_h^1, \mathsf{S}_h^2)$ corresponds to lowest order Raviart–Thomas fields, whereas $(\mathsf{S}_h'^0, \mathsf{S}_h'^1, \mathsf{S}_h'^2)$ corresponds to lowest order Brezzi–Douglas–Marini fields on the same mesh (then $\mathsf{S}_h'^0$ consists of continuous P^2 FE and $\mathsf{S}_h'^2$ of P^0 FE).

More generally (though it is not necessary) one might want to choose spaces such that the L_2-projections $\mathsf{S}_h^{2\bullet} \to \mathsf{S}_h'^{0\bullet}$, and $(\ker_{\operatorname{div}} \mathsf{S}_h^1) \times n \to \mathsf{S}_h'^1$ have kernels which are small in some sense (for instance have dimensions bounded by some small integer independently of h). Anticipating what follows, this would guarantee that the subspace $(\mathsf{S}_h^1)^\wedge$ of S_h^1, to be defined later, is almost all of S_h^1. If the first triple of

spaces is based on Raviart–Thomas fields of any order, one can take for this purpose Brezzi–Douglas–Marini fields *of the same order* on the same mesh in the second triple.

In fact, in addition to the hypothesis H0, ... ,H3 we will use some L^2 estimates for the spaces S_h^0 and S_h^2 (in the proofs of Lemmas 3.11 and 3.12) and an additional L^2 estimate for S_h^1 (proof of Proposition 4.2). Therefore we require in what follows that the spaces $(\mathsf{S}_h^0, \mathsf{S}_h^1, \mathsf{S}_h^2)$ and $(\mathsf{S}_h'^0, \mathsf{S}_h'^1, \mathsf{S}_h'^2)$ are the standard finite element spaces based on Raviart–Thomas or Brezzi–Douglas–Marini finite elements. The two triples can, however, have different orders m and m'. They can even be constructed on different meshes (with associated parameters h and h') as long as $(1/C)h \leq h' \leq Ch$. The most useful cases are $m' \geq m$ and $h' \leq h$.

Other Galerkin spaces are commonly used to solve boundary integral equations, and the method might work in the present state for such Galerkin spaces also. For instance, finite elements based on meshes with both triangular and quadrilateral elements pose no additional problem, once one has identified the appropriate null sequences called microlocal discretizations which are currently being developed.

3.3. Definition. Our starting point is to try to construct a preconditioner for the variational formulation of the EFIE on S_h^1. For this purpose we will use the auxiliary spaces $(\mathsf{S}_h'^0, \mathsf{S}_h'^1, \mathsf{S}_h'^2)$. As it turns out, with this preconditioner the EFIE is actually solved variationally on a subspace of S_h^1. However, we shall prove that this subspace (it will be denoted $(\mathsf{S}_h^1)^\wedge$) satisfies the hypotheses (H0),...,(H3), which ensures the well-posedness of the discrete problem.

Starting with a linear form $l \in \mathsf{S}_h^{1\star}$, determine the solution (u, q) of

$$\left\{ \begin{array}{l} u \in \mathsf{S}_h^1 \\ q \in \mathsf{S}_h^{2\bullet} \end{array} \right. \quad \left\{ \begin{array}{ll} \forall u' \in \mathsf{S}_h^1 & \langle u, u' \rangle + \langle q, \operatorname{div} u' \rangle = l(u'); \\ \forall q' \in \mathsf{S}_h^{2\bullet} & \langle q', \operatorname{div} u \rangle = 0. \end{array} \right. \tag{3.23}$$

Then to (u, q) associate the following element of $\mathsf{S}_h'^1$:

$$v = \mathcal{P}_{\mathsf{S}_h'^1}(u \times n) - \operatorname{rot} \mathcal{P}_{\mathsf{S}_h'^{0\bullet}}(q), \tag{3.24}$$

where for any space X_h, $\mathcal{P}_{\mathsf{X}_h}$ denotes the L_2-orthogonal projections onto X_h.

Let $\Theta_h : \mathsf{S}_h^{1\star} \to \mathsf{S}_h'^1, l \mapsto v$ be the composition of these two maps (defined by (3.23) and (3.24)), and let $\Theta_h^\star : \mathsf{S}_h'^{1\star} \to \mathsf{S}_h^{1\star\star} \approx \mathsf{S}_h^1$ be its adjoint. Then we put

$$\mathcal{Z}_h = \Theta_h^\star \mathcal{A}_h'(k) \Theta_h, \tag{3.25}$$

where $\mathcal{A}_h'(k) : \mathsf{S}_h'^1 \to \mathsf{S}_h'^{1\star}$ is the map induced by $\mathcal{A}(k)$.

Remark. In some cases it might be of interest to replace $\mathcal{A}_h'(k)$ by $\mathcal{A}_h'(k')$ for some different, possibly complex, wavenumber k'. In particular, a small perturbation $k' = k + i\epsilon$ guarantees invertibility even at resonant frequencies and is related to the limiting absorption principle. However, we will not discuss this possibility here.

3.4. Interpretation of the system. The invertibility in the sense of Babuska–Brezzi of the system (3.23) can be reinterpreted as the fact that the bilinear form b satisfies a uniform LBB inf-sup estimate on the spaces $\mathsf{S}_h^{1\#} \times \mathsf{S}_h^1$, where we have used the notation

$$\mathsf{S}_h^{1\#} = \{u \times n : u \in \mathsf{S}_h^1 \quad \text{and} \quad \operatorname{div} u = 0\} + \{\operatorname{rot} q : q \in \mathsf{S}_h^{0\bullet}\}. \tag{3.26}$$

To give a precise meaning to and prove this statement, notice that $\mathsf{S}_h^{1\#}$ is a subspace of $\mathrm{H}_{\mathrm{div}}^{-1}(\Gamma)$ (but contains vector-valued measures concentrated on the curved lines $\Xi_h^{-1}([S])$, where S is a segment in $\mathcal{T}_h$). We will also need the following lemma.

LEMMA 3.2. *Any $v \in \mathrm{H}^{-1+s}_{\mathrm{div}}(\Gamma)$, can be written in a unique way:*

$$v = u \times n - \mathrm{rot}\, q, \tag{3.27}$$

with $u \in \mathrm{H}^{s}_{\mathrm{div}}(\Gamma)$, $\mathrm{div}\, u = 0$, and $q \in \mathrm{H}^{s}(\Gamma)^{\bullet}$, and we have the equivalence of norms[2]

$$\|v\|^2_{-1+s} \approx |u|^2_s + |q|^2_s. \tag{3.28}$$

Proof. This can be proved using the HDRL (section 1.1). □

The lemma expresses that we have exhibited isomorphisms (for each s):

$$\{u \in \mathrm{H}^{s}_{\mathrm{div}}(\Gamma) \ : \ \mathrm{div}\, u = 0\} \times \mathrm{H}^{s}(\Gamma)^{\bullet} \to \mathrm{H}^{s-1}_{\mathrm{div}}(\Gamma). \tag{3.29}$$

In particular, the sum appearing in the definition of $\mathsf{S}^{1\#}_h$ is direct. Furthermore, we notice that

$$b(u \times n - \mathrm{rot}\, q, v') = \langle B(u \times n - \mathrm{rot}\, q), v' \rangle = -\langle u, v' \rangle - \langle q, \mathrm{div}\, v' \rangle, \tag{3.30}$$

where we have used the notation $\langle \cdot, \cdot \rangle$ for the three different standard dualities on

$$\mathrm{H}^{-1}_{\mathrm{rot}} \times \mathrm{H}^{0}_{\mathrm{div}} \ , \ \mathrm{H}^{0}_{\mathrm{T}} \times \mathrm{H}^{0}_{\mathrm{T}}, \quad \text{and} \quad \mathrm{H}^{0} \times \mathrm{H}^{0}. \tag{3.31}$$

Therefore, given $l \in \mathsf{S}^{1\star}_h$, if (u, q) solves system (3.23), then $v = u \times n - \mathrm{rot}\, q$ solves

$$v \in \mathsf{S}^{1\#}_h \quad \forall v' \in \mathsf{S}^{1}_h \quad -b(v, v') = l(v'), \tag{3.32}$$

and if v solves this equation, then, writing $v = u \times n - \mathrm{rot}\, q$ as in (3.26), (u, q) is also given by the *continuous* inverse of the map (3.29) and solves system (3.23).

Using the well-known properties of this system, one immediately obtains the following.

PROPOSITION 3.3. *There is $C > 0$ such that for all h*

$$\inf_{v \in \mathsf{S}^{1\#}_h} \sup_{v' \in \mathsf{S}^{1}_h} \frac{|b(v, v')|}{\|v\|_{-1} \|v'\|_0} \geq 1/C. \tag{3.33}$$

One also checks directly that these spaces have the same dimension.

3.5. Interpretation of the projections. According to Lemma 3.2 the projections defined by (3.24) correspond to a projection in the $\mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)$-norm. Lemma 3.16 and Proposition 3.18 further justify this interpretation.

3.6. A characterization of the kernel of $\boldsymbol{\Theta_h}$. In this paragraph, for any space X_h, $\mathcal{P}_{\mathsf{X}_h}$ is the orthogonal projection onto X_h for the usual L_2-inner product (on scalar or vector fields). The symbol $\perp$ is also relative to these inner products.

We introduce the following auxiliary spaces:

$$\mathsf{s}^0_h = \{ \ p \in \mathsf{S}^{0\bullet}_h \ : \ p \perp \mathsf{S}'^{2\bullet}_h \} \quad \text{and} \quad \mathsf{s}^2_h = \{ \ q \in \mathsf{S}^{2\bullet}_h \ : \ q \perp \mathsf{S}'^{0\bullet}_h \}. \tag{3.34}$$

Define also

$$(\mathsf{S}^1_h)^{\wedge} = \{v \in \mathsf{S}^1_h \ : \ v \perp \mathrm{rot}\, \mathsf{s}^0_h \quad \text{and} \quad \mathrm{div}\, v \perp \mathsf{s}^2_h\}. \tag{3.35}$$

[2]There is an obvious misprint in [19].

The introduction of s_h^0 and s_h^2 is justified by the two following lemmas, whereas that of $(\mathsf{S}_h^1)^\wedge$ is justified by Proposition 3.7. It shows that (for h small enough—we will not always repeat this condition—) $(\mathsf{S}_h^1)^\wedge$ is the range of $\Theta_h^\star$.

LEMMA 3.4. *There is h_0 such that for all $h < h_0$ and all divergence-free $u \in \mathsf{S}_h^1$*

$$\mathcal{P}_{\mathsf{S}_h'^1}(u \times n) = 0 \iff u \in \operatorname{rot} \mathsf{s}_h^0. \tag{3.36}$$

Proof. We use the fact that for all $\epsilon > 0$ there is $h_0 > 0$ such that for $h < h_0$

$$\forall a \in \mathsf{G}_h^1 \quad |a \times n - \mathcal{P}_{\mathsf{G}_h'^1}(a \times n)|_0 \leq \epsilon |a|_0. \tag{3.37}$$

Choosing a h_0 relative to a $\epsilon < 1$ we suppose from now on that $h < h_0$.

Pick $u \in \ker_{\mathsf{S}_h^1} \operatorname{div}$. Put $u = \operatorname{rot} p + a$, with $p \in \mathsf{S}_h^{0\bullet}$ and $a \in \mathsf{G}_h^1$.

(i). Suppose that the projection of $u \times n$ is zero. For all divergence-free $a' \in \mathsf{S}_h'^1$, we have

$$0 = \langle \mathcal{P}_{\mathsf{S}_h'^1}(-\operatorname{grad} p + a \times n), \overline{a'} \rangle = \langle (-\operatorname{grad} p + a \times n), \overline{a'} \rangle = \langle a \times n, \overline{a'} \rangle. \tag{3.38}$$

Put $a' = \mathcal{P}_{\mathsf{G}_h'^1}(a \times n)$. Then

$$|a|_0^2 = |\langle a \times n, \overline{a \times n} \rangle| = |\langle a \times n, (\overline{a \times n - a'}) \rangle| \leq \epsilon |a|_0^2. \tag{3.39}$$

Hence $a = 0$. Moreover, for all $v \in \mathsf{S}_h'^1$,

$$\langle \operatorname{rot} p \times n, v \rangle = \langle p, \operatorname{div} v \rangle. \tag{3.40}$$

It follows that $p \perp \mathsf{S}_h'^{2\bullet}$ and $u \in \operatorname{rot} \mathsf{s}_h^0$.

(ii). Conversely, if $u \in \operatorname{rot} \mathsf{s}_h^0$, then the above equality (3.40) shows that $u \times n \perp \mathsf{S}_h'^1$, and hence its projection is zero. ∎

LEMMA 3.5. *For all $q \in \mathsf{S}_h^{2\bullet}$ we have*

$$\operatorname{rot} \mathcal{P}_{\mathsf{S}_h'^{0\bullet}}(q) = 0 \iff \mathcal{P}_{\mathsf{S}_h'^{0\bullet}}(q) = 0 \iff q \in \mathsf{s}_h^2. \tag{3.41}$$

Proof. The proof is trivial. ∎

LEMMA 3.6. *The spaces $\mathcal{P}_{\mathsf{S}_h'^1}((\ker_{\mathrm{H}^0_{\operatorname{div}}(\Gamma)} \operatorname{div}) \times n)$ and $\operatorname{rot} \mathsf{S}_h'^0$ are L_2-orthogonal.*

Proof. Indeed, if $u \in \mathrm{H}^0_{\operatorname{div}}(\Gamma)$ is divergence-free and $p \in \mathsf{S}_h'^0$, then (since $\operatorname{rot} p \in \mathsf{S}_h'^1$)

$$\langle \mathcal{P}_{\mathsf{S}_h'^1}(u \times n), \operatorname{rot} p \rangle = \langle u \times n, \operatorname{rot} p \rangle = -\langle \operatorname{div} u, p \rangle = 0. \tag{3.42}$$ ∎

We now prove the following.

PROPOSITION 3.7. *For all $l \in \mathsf{S}_h^{1\star}$ we have*

$$\Theta_h(l) = 0 \iff \forall u' \in (\mathsf{S}_h^1)^\wedge \quad l(u') = 0. \tag{3.43}$$

Proof. Pick $l \in \mathsf{S}_h^{1\star}$. Let (u, q) be the solution of

$$\left\{ \begin{array}{l} u \in \mathsf{S}_h^1 \\ q \in \mathsf{S}_h^{2\bullet} \end{array} \right. \quad \left\{ \begin{array}{lll} \forall u' \in \mathsf{S}_h^1 & \langle u, u' \rangle + \langle q, \operatorname{div} u' \rangle & = \quad l(u'); \\ \forall q' \in \mathsf{S}_h^{2\bullet} & \langle q', \operatorname{div} u \rangle & = \quad 0. \end{array} \right. \tag{3.44}$$

With these definitions we have $\Theta_h(l) = 0$ if and only if

$$\mathcal{P}_{\mathsf{S}_h'^1}(u \times n) - \operatorname{rot} \mathcal{P}_{\mathsf{S}_h'^{0\bullet}}(q) = 0. \tag{3.45}$$

According to Lemma 3.6, this is in turn equivalent to

$$\mathcal{P}_{\mathsf{S}'^1_h}(u \times n) = 0 \quad \text{and} \quad \operatorname{rot} \mathcal{P}_{\mathsf{S}'^0_h \bullet}(q) = 0. \tag{3.46}$$

In Lemmas 3.4 and 3.5 we gave equivalent statements for these two conditions.

(i). If $\Theta_h(l) = 0$, then $u \in \operatorname{rot} \mathsf{s}^0_h$ and $q \in \mathsf{s}^2_h$; hence, for all $u' \in (\mathsf{S}^1_h)^\wedge$,

$$l(u') = \langle u, u' \rangle + \langle q, \operatorname{div} u' \rangle = 0. \tag{3.47}$$

That is to say, l vanishes on $(\mathsf{S}^1_h)^\wedge$.

(ii). If l vanishes on $(\mathsf{S}^1_h)^\wedge$, then

- for all $u' \in \ker_{\mathsf{S}^1_h} \operatorname{div}$ such that $u' \perp \operatorname{rot} \mathsf{s}^0_h$ we have (since $u' \in (\mathsf{S}^1_h)^\wedge$)

$$\langle u, u' \rangle = l(u') = 0, \tag{3.48}$$

  so $u \in \operatorname{rot} \mathsf{s}^0_h$;
- for all $q' \in \mathsf{S}^{2\bullet}_h$ such that $q' \perp \mathsf{s}^2_h$, picking $u' \in \mathsf{S}^1_h$ such that $\operatorname{div} u' = q'$ and $u' \perp \ker_{\mathsf{S}^1_h} \operatorname{div}$, we have (since $u' \in (\mathsf{S}^1_h)^\wedge$)

$$\langle q, q' \rangle = \langle q, \operatorname{div} u' \rangle = l(u') = 0, \tag{3.49}$$

  and hence $q \in \mathsf{s}^2_h$.

The proof is complete. ∎

3.7. Approximation properties of the range of $\Theta^\star_h$. We give yet another characterization of $(\mathsf{S}^1_h)^\wedge$, which will enable us to deduce its approximation properties.

LEMMA 3.8. *Pick $u \in \mathsf{S}^1_h$. Put $u = \operatorname{rot} p + a + g$ with $p \in \mathsf{S}^{0\bullet}_h$, $a \in \mathsf{G}^1_h$, and $g \in \mathsf{S}^1_h$ such that $g \perp \ker_{\mathsf{S}^1_h} \operatorname{div}$. Then we have*

$$u \in (\mathsf{S}^1_h)^\wedge \iff \operatorname{rot} p \perp \operatorname{rot} \mathsf{s}^0_h \quad \textit{and} \quad \operatorname{div} g \perp \mathsf{s}^2_h. \tag{3.50}$$

Proof. The proof is trivial. ∎

Now we give equivalent expressions for the above two conditions.

LEMMA 3.9. *Choose $q \in \mathsf{S}^{2\bullet}_h$. We have*

$$q \perp \mathsf{s}^2_h \iff \exists p \in \mathsf{S}'^{0\bullet}_h, \qquad q = \mathcal{P}_{\mathsf{S}^{2\bullet}_h}(p). \tag{3.51}$$

Proof. More generally, we have the following result: Let X be a Hilbert space. The orthogonal projection onto a closed subspace Y of X is written P_Y, and orthogonality is denoted by $\perp$. If Y and Z are two closed subspaces of X, we put

$$Y_{\perp Z} = \{x \in Y \ : \ x \perp Z\}. \tag{3.52}$$

Let A and B be two closed subspaces of X. Then

$$A_{\perp(A_{\perp B})} = P_A(B). \tag{3.53}$$

∎

LEMMA 3.10. *Choose $p \in \mathsf{S}^{0\bullet}_h$. We have*

$$\operatorname{rot} p \perp \operatorname{rot} \mathsf{s}^0_h \iff \exists q \in \mathsf{S}'^{2\bullet}_h, \qquad \operatorname{rot} p = \mathcal{P}_{\operatorname{rot} \mathsf{S}^{0\bullet}_h}(\operatorname{rot} \Delta^{-1} q). \tag{3.54}$$

Proof. We apply the same technique once again. For all $p \in \mathsf{S}^{0\bullet}_h$ and all $q \in \mathsf{S}'^{2\bullet}_h$ we have

$$\langle p, q \rangle = -\langle \operatorname{rot} p, \operatorname{rot} \Delta^{-1} q \rangle. \tag{3.55}$$

Hence

$$\operatorname{rot} p \in \operatorname{rot} \mathsf{s}_h^0 \iff p \in \mathsf{s}_h^0 \tag{3.56}$$

$$\iff p \perp \mathsf{S}_h'^{2\bullet} \tag{3.57}$$

$$\iff \operatorname{rot} p \perp \operatorname{rot} \Delta^{-1} \mathsf{S}_h'^{2\bullet} \tag{3.58}$$

$$\iff \operatorname{rot} p \perp \mathcal{P}_{\operatorname{rot} \mathsf{S}_h^{0\bullet}} (\operatorname{rot} \Delta^{-1} \mathsf{S}_h'^{2\bullet}). \tag{3.59}$$

So

$$\operatorname{rot} p \perp \operatorname{rot} \mathsf{s}_h^0 \iff \operatorname{rot} p \in \mathcal{P}_{\operatorname{rot} \mathsf{S}_h^{0\bullet}} (\operatorname{rot} \Delta^{-1} \mathsf{S}_h'^{2\bullet}). \tag{3.60}$$ ∎

We will also need the following two approximation results.

LEMMA 3.11. *There is $C > 0$ such that for all h and all $\phi \in \mathrm{H}^1(\Gamma)^\bullet$*

$$\inf\{|\phi - q|_0 \ : \ q \in \mathcal{P}_{\mathsf{S}_h^{2\bullet}} (\mathsf{S}_h'^{0\bullet})\} \leq Ch|\phi|_1. \tag{3.61}$$

Proof. Notice that

$$|\phi - \mathcal{P}_{\mathsf{S}_h^{2\bullet}} \mathcal{P}_{\mathsf{S}_h^{0\bullet}} \phi| \leq |\phi - \mathcal{P}_{\mathsf{S}_h^{0\bullet}} \phi| + |\mathcal{P}_{\mathsf{S}_h^{0\bullet}} \phi - \mathcal{P}_{\mathsf{S}_h^{2\bullet}} \mathcal{P}_{\mathsf{S}_h^{0\bullet}} \phi| \tag{3.62}$$

$$\leq Ch|\phi|_1 + Ch|\mathcal{P}_{\mathsf{S}_h^{0\bullet}} \phi|_1 \tag{3.63}$$

$$\leq Ch|\phi|_1. \tag{3.64}$$ ∎

LEMMA 3.12. *There is $C > 0$ such that for all h and all $\phi \in \mathrm{H}^2(\Gamma)^\bullet$*

$$\inf\{|\operatorname{rot} \phi - u|_0 \ : \ u \in \mathcal{P}_{\operatorname{rot} \mathsf{S}_h^{0\bullet}} (\operatorname{rot} \Delta^{-1} \mathsf{S}_h'^{2\bullet})\} \leq Ch|\operatorname{rot} \phi|_1. \tag{3.65}$$

Proof. We have

$$|\Delta^{-1} \mathcal{P}_{\mathsf{S}_h'^{2\bullet}} \Delta\phi - \phi|_1 \leq C|\mathcal{P}_{\mathsf{S}_h'^{2\bullet}} \Delta\phi - \Delta\phi|_{-1} \leq Ch|\Delta\phi|_0. \tag{3.66}$$

So, with $\psi = \Delta^{-1} \mathcal{P}_{\mathsf{S}_h^{2\bullet}} \Delta\phi$, we have

$$|\operatorname{rot} \psi - \operatorname{rot} \phi|_0 \leq Ch|\operatorname{rot} \phi|_1 \tag{3.67}$$

and

$$|\mathcal{P}_{\operatorname{rot} \mathsf{S}_h^{0\bullet}} \operatorname{rot} \psi - \operatorname{rot} \psi|_0 \leq Ch|\operatorname{rot} \psi|_1. \tag{3.68}$$

However,

$$|\operatorname{rot} \psi|_1 \leq C|\Delta^{-1} \mathcal{P}_{\mathsf{S}_h'^{2\bullet}} \Delta\phi|_2 \leq C|\mathcal{P}_{\mathsf{S}_h'^{2\bullet}} \Delta\phi|_0 \leq C|\Delta\phi|_0 \leq C|\operatorname{rot} \phi|_1. \tag{3.69}$$

Now combine inequalities (3.68) and (3.69) and conclude using (3.67). ∎

From the above results we deduce the following fundamental theorem.

THEOREM 3.13. *The spaces $X_h = (\mathsf{S}_h^1)^\wedge$ satisfy hypothesis* (H1).

Proof. Pick $u \in \mathrm{H}^1_{\mathrm{div}}(\Gamma)$. Consider its Helmholtz decomposition

$$u = \operatorname{rot} \phi + \alpha + \operatorname{grad} \psi. \tag{3.70}$$

(i) The field $\operatorname{rot} \phi$ is approximated using Lemma 3.12, which gives an element of X_h according to Lemmas 3.8 and 3.10.

(ii) The field α is approximated by $a = \Omega_h \alpha$, where Ω_h is defined by system (2.18).

(iii) The field $\Delta\psi$ is approximated by a $q \in \mathcal{P}_{\mathsf{S}_h^{2\bullet}} (\mathsf{S}_h'^{0\bullet})$ following Lemma 3.11. Then we consider $\Omega_h \operatorname{grad} \Delta^{-1} q$ which is in X_h according to Lemmas 3.8 and 3.9. ∎

3.8. Well-posedness.

PROPOSITION 3.14. *The spaces* $X_h = (\mathsf{S}^1_h)^\wedge$ *satisfy hypothesis* (H3).

Proof. Choose $u \in X_h$ such that

$$u \perp \ker_{X_h} \operatorname{div}. \tag{3.71}$$

Trivially we have $u \in \mathsf{S}^1_h$. Moreover, if u' is a divergence-free element of S^1_h it can be written $\operatorname{rot} p + u''$ with $p \in \mathsf{s}^0_h$ and a divergence-free $u'' \in X_h$. (To see this just remark that the L_2-orthogonal of $\operatorname{rot} \mathsf{s}^0_h$ in $\ker_{\mathsf{S}^1_h} \operatorname{div}$ is a subspace of X_h.) And since $u \perp \operatorname{rot} \mathsf{s}^0_h$ and $u \perp u''$ we therefore have

$$u \perp \ker_{\mathsf{S}^1_h} \operatorname{div}. \tag{3.72}$$

Then the proposition follows from the result known to hold for S^1_h. ∎

We have therefore reached the main goal of this section.

THEOREM 3.15. *The spaces* $(\mathsf{S}^1_h)^\wedge$ *satisfy the four hypothesis* (H0),...,(H3).

If $\mathcal{A}(k)$ were symmetric positive definite, then $\mathcal{Z}_h\mathcal{A}_h$ would determine an isomorphism $(\mathsf{S}^1_h)^\wedge \to (\mathsf{S}^1_h)^\wedge$, and, for a given h, the preconditioned conjugate gradients (PCG) algorithm would converge towards the variational solution on $(\mathsf{S}^1_h)^\wedge$, which by the above theorem is a good one.

In our indefinite case it might be that $\mathcal{Z}_h\mathcal{A}_h$ does not determine an isomorphism $(\mathsf{S}^1_h)^\wedge \to (\mathsf{S}^1_h)^\wedge$; this would be the case if the bilinear form induced by $\mathcal{A}$ on the range of Θ_h (as opposed to $\Theta^\star_h$) were degenerate, a question we have not settled. However, we are sure that the PCG algorithm constructs iterates that are in $(\mathsf{S}^1_h)^\wedge$, and later—in section 3.9—we will show how to construct stopping criteria that guarantee that the residual is small as a linear form on $(\mathsf{S}^1_h)^\wedge$. Thus one can *check* that the approximate solution given by the PCG algorithm is close to the variational solution on $(\mathsf{S}^1_h)^\wedge$. Theorem 3.15 ensures that this variational solution (exists, is unique, and) is close to the *best* (for any chosen norm) approximation on $(\mathsf{S}^1_h)^\wedge$ of the exact solution and that for small h this best approximation is a *good* approximation.

3.9. Stopping criterion. Proposition 3.7 shows that, for all $u \in (\mathsf{S}^1_h)^\wedge$, u solves the variational problem

$$u \in (\mathsf{S}^1_h)^\wedge \quad \text{and} \quad \forall v \in (\mathsf{S}^1_h)^\wedge \quad a(u,v) = l(v) \tag{3.73}$$

if and only if

$$\Theta_h(\mathcal{A}_h u - l_h) = 0, \tag{3.74}$$

which is in turn equivalent to

$$\|\Theta_h(\mathcal{A}_h u - l_h)\| = 0, \tag{3.75}$$

for any norm $\|\cdot\|$ on S'^1_h. We now set out to define norms on S'^1_h that are uniformly equivalent to the $\mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)$-norm but more readily computable.

The following lemma should be seen in relation to Lemma 3.2.

LEMMA 3.16. *There is* $C > 0$ *such that for all* $u \in \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)$*, if* $u = \operatorname{rot} p + v$ *with* $v \in \mathrm{H}^0_{\mathrm{T}}(\Gamma)$ *and* $p \in \mathrm{H}^0(\Gamma)$*, we have*

$$\|u\|^2_{-1} \le C\big(|p|^2_0 + |v|^2_0\big). \tag{3.76}$$

Proof. It holds

$$\|\operatorname{rot} p + v\|_{-1}^2 = |\operatorname{rot} p + v|_{-1}^2 + |\operatorname{div} v|_{-1}^2. \tag{3.77}$$

However,

$$|\operatorname{rot} p + v|_{-1}^2 \le 2\big(|\operatorname{rot} p|_{-1}^2 + |v|_{-1}^2\big) \le C\big(|p|_0^2 + |v|_{-1}^2\big). \tag{3.78}$$

Then one immediately concludes using $|\operatorname{div} v|_{-1} \le C|v|_0$. ∎

For convenience we state as a separate lemma the following fact which follows from (H3).

LEMMA 3.17. *There is $C > 0$ such that for all $u \in X_h$, if*

$$\forall u' \in X_h \quad \operatorname{div} u' = 0 \;\Rightarrow\; \langle u, u' \rangle = 0, \tag{3.79}$$

then

$$|u|_0 \le C|\operatorname{div} u|_{-1}. \tag{3.80}$$

Proof. Let ϕ be the solution of

$$\phi \in \mathrm{H}^1(\Gamma)^\bullet \quad \text{and} \quad \Delta\phi = \operatorname{div} u. \tag{3.81}$$

We have

$$|u|_0 \le |u - \operatorname{grad}\phi|_0 + |\operatorname{grad}\phi|_0 \le Ch|\operatorname{div} u|_0 + C|\operatorname{div} u|_{-1}. \tag{3.82}$$

The lemma then follows from an inverse inequality. ∎

It is a particular case of Lemma 3.2 that the converse inequality of Lemma 3.16 holds whenever v is such that $\operatorname{rot} v = 0$. This fact has the following discrete analogue.

PROPOSITION 3.18. *There is $C > 0$, such that for all h and all $u \in \mathsf{S}_h^1$, if $u = \operatorname{rot} p + v$ with $p \in \mathsf{S}_h^{0\bullet}$ and $v \perp \operatorname{rot} \mathsf{S}_h^{0\bullet}$, then*

$$|p|_0^2 + |v|_0^2 \le C\|u\|_{-1}^2. \tag{3.83}$$

Proof. Put $v = a + w$ with $a \in \mathsf{G}_h^1$ and $w \perp \ker_{\mathsf{S}_h^1} \operatorname{div}$. We have

$$|v|_0^2 = |a|_0^2 + |w|_0^2. \tag{3.84}$$

However, by Lemma 2.2 and Proposition 2.3 we have

$$|a|_0 \le C|a|_{-1} \le C\|u\|_{-1}, \tag{3.85}$$

and by Lemma 3.17 we have

$$|w|_0 \le C|\operatorname{div} w|_{-1} = |\operatorname{div} u|_{-1} \le \|u\|_{-1}. \tag{3.86}$$

So

$$|v|_0^2 \le C\|u\|_{-1}^2. \tag{3.87}$$

Moreover,

$$|p|_0^2 \le C|\operatorname{rot} p|_{-1}^2 \le C\big(|\operatorname{rot} p + v|_{-1}^2 + |v|_{-1}^2\big). \tag{3.88}$$

The proposition follows. ∎

Of course the same holds true for the spaces with a prime, which is in fact what will be of interest to us.

Let l be a linear form on S_h^1, and, as in the definition of the preconditioner, let (u, q) be the solution of system (3.23) so that

$$\Theta_h l = \mathcal{P}_{\mathsf{S}_h'^1}(u \times n) - \operatorname{rot} \mathcal{P}_{\mathsf{S}_h'^0 \bullet}(q). \tag{3.89}$$

Lemma 3.16 together with Proposition 3.18 now prove that we have the uniform (i.e., independent of h) equivalence of norms

$$\|\Theta_h l\|_{-1}^2 \approx |\mathcal{P}_{\mathsf{S}_h'^1}(u \times n)|_0^2 + |\mathcal{P}_{\mathsf{S}_h'^0 \bullet}(q)|_0^2. \tag{3.90}$$

A stopping criterion can therefore be a sufficient reduction of this norm. It is important to notice that to effectively compute these norms in the course of a conjugate gradients algorithm is a negligible task compared with the other ones, requiring only two sparse matrix products (at each iteration).

Another norm which is both natural and easily computable is $\|\Theta_h l\|_0$. Also, the quantity $(\|\Theta_h l\|_0 \|\Theta_h l\|_{-1})^{1/2}$, though not a norm, satisfies the interpolation inequality

$$\|\Theta_h l\|_{1/2} \leq C\big(\|\Theta_h l\|_0 \|\Theta_h l\|_{-1}\big)^{1/2} \tag{3.91}$$

and is therefore another good candidate for the construction of a stopping criterion. We have not determined to which extent the choice between these candidates really produces any significant differences on industrial problems.

4. Behavior of the iterates. We denote by Θ the continuous analogue of Θ_h, that is, the map $(\mathrm{H}^0_{\mathrm{div}}(\Gamma))^\star \to \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)$ which to l associates $u \times n - \operatorname{rot} q$, where u and q are the solutions of the continuous saddle-point problem. We also have $\mathcal{Z} = \Theta^\star \mathcal{A} \Theta$.

The Krylov subspaces are defined to be

$$\mathfrak{K}_h^n = \{P(\mathcal{Z}_h \mathcal{A}_h)\mathcal{Z}_h l|_{\mathsf{S}_h^1} \ : \ P \in \mathbb{C}[X], \ \deg P \leq n\}. \tag{4.1}$$

Their importance stems from the fact that—for fixed h—the PCG algorithm attempts to determine (by short recurrences) the sequence of solutions (u_h^n) of the problems

$$u \in \mathfrak{K}_h^n \quad \text{and} \quad \forall v \in \mathfrak{K}_h^n \quad a(u, v) = l(v). \tag{4.2}$$

For generalities about the PCG and related algorithms we refer to Barrett et al. [3] or Kelley [36]. Since we deal with complex-symmetric matrices, see also Freund [29]. In this section we investigate the convergence of the spaces $\mathfrak{K}_h^n$ towards their continuous analogues $\mathfrak{K}^n$, for fixed n as $h \to 0$, where naturally we have put

$$\mathfrak{K}^n = \{P(\mathcal{Z}\mathcal{A})\mathcal{Z} l \ : \ P \in \mathbb{C}[X], \ \deg P \leq n\}. \tag{4.3}$$

We will deduce from this results on the convergence as $h \to 0$ of the iterate u_h^n towards the solution u^n of

$$u \in \mathfrak{K}^n \quad \text{and} \quad \forall v \in \mathfrak{K}^n \quad a(u, v) = l(v). \tag{4.4}$$

We emphasize that for non-SPD problems the convergence or breakdown of the Lanczos process is as of today not completely understood. Here our point of view is to suppose that the continuous Lanczos process is well-defined up to iteration n, and then show that for small enough h, the discrete one is also well-defined, and yields

arbitrarily close iterates. If the ideal Lanczos process breaks down one should not expect the discrete one to behave well. We have not observed this pathology yet but should it occur one can consider in addition to various restart and look-ahead techniques perturbating the preconditioning operator.

Finally, we argue that the sequence (u^n) might converge superlinearly, as we show it to be the case for SPD operators, when the preconditioner is an inverse modulo a compact endomorphism.

4.1. Stability and convergence of Krylov subspaces.

PROPOSITION 4.1. *There is $C > 0$ such that for all h and all $l \in (\mathsf{S}_h^1)^\star$*

$$\|\Theta_h(l)\|_{-1} \leq C \sup_{v \in \mathsf{S}_h^1} \frac{|l(v)|}{\|v\|_0}. \tag{4.5}$$

Proof. Lemma 3.16 gives

$$\|\Theta_h(l)\|_{-1}^2 \leq C\big(|\mathcal{P}_{\mathsf{S}_h^{\prime 1}}(u \times n)|_0^2 + |\mathcal{P}_{\mathsf{S}_h^{\prime 0}\bullet}(q)|_0^2\big) \leq C\big(|u|_0^2 + |q|_0^2\big). \tag{4.6}$$

This gives the announced estimate. ∎

PROPOSITION 4.2. *There is $C > 0$ such that for all $l \in (\mathrm{H}_{\mathrm{div}}^{-1}(\Gamma))^\star$ and all h*

$$\|\Theta_h l|_{\mathsf{S}_h^1} - \Theta l\|_{-1} \leq Ch\|l\|_{-1\star}. \tag{4.7}$$

Proof. Let (u_h, q_h) be the solutions of the discrete saddle-point problem, and let (u_0, q_0) be the solutions of the continuous one. The well-known properties of this problem (in particular, Propositions 2.13 (p. 64) and 3.9 (p. 132) in Brezzi and Fortin [12]) yield

$$|u_h - u_0|_0^2 + |q_h - q_0|_0^2 \leq Ch^2(|u_0|_1^2 + |q_0|_1^2) \leq Ch^2\|l\|_{-1\star}^2. \tag{4.8}$$

Denoting for simplicity the L_2-orthogonal projections onto the appropriate spaces by $\mathcal{P}_h$, we have

$$|\mathcal{P}_h(u_h \times n) - (u_0 \times n)|_0 \tag{4.9}$$

$$\leq |\mathcal{P}_h(u_h \times n) - \mathcal{P}_h(u_0 \times n)|_0 + |\mathcal{P}_h(u_0 \times n) - u_0 \times n|_0 \tag{4.10}$$

$$\leq |(u_h \times n) - (u_0 \times n)|_0 + Ch|u_0 \times n|_1 \tag{4.11}$$

$$\leq Ch|u_0|_1. \tag{4.12}$$

Using the same technique, we also obtain

$$|\mathcal{P}_h(q_h) - q_0|_0 \leq Ch|q_0|_1. \tag{4.13}$$

This completes the proof, using Lemma 3.16. ∎

From these two propositions we deduce stability and convergence estimates for Θ_h in half-integer Sobolev norms.

COROLLARY 4.3. *There is $C > 0$ such that for all h and all $l \in \mathrm{H}_{\mathrm{div}}^{-1}(\Gamma)^\star$*

$$\|\Theta_h l|_{\mathsf{S}_h^1} - \Theta l\|_{-1/2} \leq Ch^{1/2}\|l\|_{-1\star}, \tag{4.14}$$

and there is $C > 0$ such that for all h and all $l \in (\mathsf{S}_h^1)^\star$

$$\|\Theta_h l\|_{-1/2} \leq C \sup_{v \in \mathsf{S}_h^1} \frac{|l(v)|}{\|v\|_{-1/2}}. \tag{4.15}$$

Proof. Let $\mathcal{Q}_h$ be the $\mathrm{H}^0_{\mathrm{div}}(\Gamma)$-orthogonal projection onto S'^1_h. The required properties of this projection were summarized in section 2.2. We have for all $l \in \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)^\star$

$$\|\Theta_h l|_{\mathsf{S}^1_h} - \Theta l\|_{-1/2} \le \|\Theta_h l|_{\mathsf{S}^1_h} - \mathcal{Q}_h\Theta l\|_{-1/2} + \|\mathcal{Q}_h\Theta l - \Theta l\|_{-1/2} \tag{4.16}$$

$$\le Ch^{-1/2}\|\Theta_h l|_{\mathsf{S}^1_h} - \mathcal{Q}_h\Theta l\|_{-1} + Ch^{1/2}\|\Theta l\|_0 \tag{4.17}$$

$$\le Ch^{-1/2}\big(\|\Theta_h l|_{\mathsf{S}^1_h} - \Theta l\|_{-1} + \|\Theta l - \mathcal{Q}_h\Theta l\|_{-1}\big) + \cdots \tag{4.18}$$

$$\le Ch^{1/2}\|l\|_{-1\star}, \tag{4.19}$$

and repeating the same sort of arguments, still supposing that $l \in \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)^\star$,

$$\|\Theta_h l|_{\mathsf{S}^1_h}\|_0 \le \|\Theta_h l - \mathcal{Q}_h\Theta l\|_0 + \|\mathcal{Q}_h\Theta l\|_0 \tag{4.20}$$

$$\le Ch^{-1}\|\Theta_h l - \mathcal{Q}_h\Theta l\|_{-1} + \|\Theta l\|_0 \tag{4.21}$$

$$\le Ch^{-1}\big(\|\Theta_h l - \Theta l\|_{-1} + \|\Theta l - \mathcal{Q}_h\Theta l\|_{-1}\big) + \cdots \tag{4.22}$$

$$\le C\|l\|_{-1\star}. \tag{4.23}$$

Combining this estimate with Proposition 4.1 by interpolation, we obtain, for $l \in \mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)^\star$,

$$\|\Theta_h l|_{\mathsf{S}^1_h}\|_{-1/2} \le C\|l\|_{-1/2\star}. \tag{4.24}$$

The apparently more refined version announced can then be deduced from the existence of an extension operator with norm one (the adjoint of the $\mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$-orthogonal projection onto S^1_h) or a Hahn–Banach theorem. ∎

We have similar estimates for $\Theta^\star_h$.

Corollary 4.4. *There is $C > 0$ such that for all h and all $l \in \mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)^\star$*

$$\|\Theta^\star_h l|_{\mathsf{S}'^1_h} - \Theta^\star l\|_{-1/2} \le Ch^{1/2}\|l\|_{-1\star}, \tag{4.25}$$

and there is $C > 0$ such that for all h and all $l \in (\mathsf{S}'^1_h)^\star$

$$\|\Theta^\star_h l\|_{-1/2} \le C \sup_{v\in\mathsf{S}'^1_h} \frac{|l(v)|}{\|v\|_{-1/2}}. \tag{4.26}$$

Proof. Using the fact that a (bounded) operator has the same norm as its adjoint we first obtain from Proposition 4.2 that

$$\|\Theta^\star_h l|_{\mathsf{S}'^1_h} - \Theta^\star l\|_{-1} \le Ch\|l\|_{-1\star}. \tag{4.27}$$

As in the proof of Corollary 4.3 this yields the first equation. The second one follows trivially from the second estimate of the same corollary. ∎

From this one deduces the following.

Corollary 4.5. *Let (l_h) be a sequence of linear forms on $\mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$ which converges to l (in the norm sense in the dual of $\mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$). Then $\Theta_h l_h|_{\mathsf{S}^1_h}$ converges to Θl in $\mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)$. Similarly, $\Theta^\star_h l_h|_{\mathsf{S}'^1_h}$ converges to $\Theta^\star l$.*

Proof. The technique of proof is very classical (see, for instance, Folland [28, Prop. 5.17, p. 169] for the just as easy case of constant l_h) and relies on Corollary 4.3 (and Corollary 4.4 for the second part) using the density of $\mathrm{H}^{-1}_{\mathrm{div}}(\Gamma)^\star$ in $\mathrm{H}^{-1/2}_{\mathrm{div}}(\Gamma)^\star$. ∎

Then we immediately obtain stability and approximation properties for the preconditioner $\mathcal{Z}_h$.

PROPOSITION 4.6. *There is $C > 0$ such that for all h and all $l \in \mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)^\star$*

$$\|\mathcal{Z}_h l|_{\mathsf{S}_h^1}\|_{-1/2} \leq C\|l\|_{-1/2\star}. \tag{4.28}$$

If a sequence of linear forms $l_h \in \mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)^\star$ converges to l in $\mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)^\star$, then $\mathcal{Z}_h l_h|_{\mathsf{S}_h^1}$ converges to $\mathcal{Z}l$.

We are now ready to prove the announced theorem.

THEOREM 4.7. *For all $l \in (\mathrm{H}_{\mathrm{div}}^{-1/2})^\star$ and all $n \in \mathbb{N}$, $(\mathcal{Z}_h\mathcal{A}_h)^n\mathcal{Z}_h l|_{\mathsf{S}_h^1}$ converges to $(\mathcal{Z}\mathcal{A})^n\mathcal{Z}l$ in $\mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)$.*

Proof. This follows from the above results using a simple recursion argument. ∎

Remark. We have not derived any optimal orders of convergence, for smoother than necessary data and perhaps higher order finite elements, though we do not expect this to yield any serious difficulties or require methods of proof different from the above ones. Nor have we tried to determine the minimum hypotheses on the regularity of the triangulations under which our conclusions hold; in particular, we have not determined to which extent the quasi-uniformity hypothesis (which is used for the inverse inequalities) can be relaxed. Of course also working on nonsmooth surfaces would put severe limitations on the range of Sobolev spaces we could use.

4.2. Convergence of the iterates. Let X be a Banach space and X_1, X_0 two closed subspaces. When these are nonzero, the *gap* from X_1 to X_0, denoted $\delta(X_1, X_0)$, is defined to be

$$\delta(X_1, X_0) = \sup_{u_1 \in X_1} \inf_{u_0 \in X_0} \|u_1 - u_0\|/\|u_1\|. \tag{4.29}$$

This definition is extended straightforwardly to the case when X_1 or X_0 is zero. For a thorough discussion of the gap we refer to Kato [35] but for us the definition is enough.

Suppose that X_0 *splits*, i.e., has a closed supplementary (for instance finite-dimensional spaces automatically split, as do closed subspaces of Hilbert spaces), so that we have a continuous projector $P : X \to X$ with range X_0. For all $u \in X$, one has

$$\forall u' \in X_0 \quad \|u - Pu\| = \|(u - u') - (Pu - u')\| = \|(u - u') - P(u - u')\|. \tag{4.30}$$

Hence

$$\|u - Pu\| \leq \|I - P\| \left(\inf_{u' \in X_0} \|u - u'\|/\|u\| \right) \|u\|. \tag{4.31}$$

In particular,

$$\forall u \in X_1 \quad \|u - Pu\| \leq \|I - P\|\delta(X_1, X_0)\|u\|. \tag{4.32}$$

Thus if (X_h) is a family of closed subspaces such that $\lim_h \delta(X_h, X_0) = 0$, then for sufficiently small h the spaces PX_h are closed in X_0 and P induces isomorphisms $X_h \to PX_h$ which are arbitrarily close in norm to the identity mapping on X_h.

PROPOSITION 4.8. *Let X be a reflexive Banach space and $\mathcal{A} : X \to X^\star$ a continuous linear map. Suppose X_0 is a closed subspace that splits yielding a projector*

P. Suppose X_h is another closed subspace, and that the induced maps $\mathcal{A}_0 : X_0 \to X_0^\star$, $\mathcal{A}_h : X_h \to X_h^\star$ satisfy the inf-sup conditions (1.20), (1.21), *with constants α_0 and α_h. Also put $\delta_h = \delta(X_h, X_0)$.*

Then $\mathcal{A}_0$ and $\mathcal{A}_h$ are invertible ; moreover, for any $l \in X^\star$, if we put $u_h = \mathcal{A}_h^{-1} l|_{X_h}$ and $u_0 = \mathcal{A}_0^{-1} l|_{X_0}$, we have for all $u' \in X_h$

$$\|u_h - u_0\| \leq \alpha_h^{-1}(1 + \alpha_0^{-1}\|\mathcal{A}\|)\|I - P\|\delta_h\|l\| + (1 + \alpha_h^{-1}\|\mathcal{A}\|)\|u_0 - u'\|. \tag{4.33}$$

Proof. That $\mathcal{A}_0$ and $\mathcal{A}_h$ are invertible is part of Theorem 1.1. Concerning the approximation property, we have (as usual we denote by a the bilinear form corresponding to $\mathcal{A}$)

$$\|u_h - u_0\| \leq \|u_h - u'\| + \|u' - u_0\| \leq \alpha_h^{-1} \sup_{v \in X_h} \frac{|a(u_h - u', v)|}{\|v\|} + \|u' - u_0\|. \tag{4.34}$$

Now (for $v \in X_h$) write

$$a(u_h - u', v) = a(u_h, v) + a(u_0 - u', v) - a(u_0, Pv) - a(u_0, v - Pv) \tag{4.35}$$

$$= l(v - Pv) + a(u_0 - u', v) - a(u_0, v - Pv). \tag{4.36}$$

Therefore,

$$|a(u_h - u', v)|/\|v\| \leq (1 + \alpha_0^{-1}\|a\|)(\|I - P\|)\delta_h\|l\| + \|a\|\,\|u_0 - u'\|. \tag{4.37}$$

This proves the proposition. □

Remark. Just as Theorem 1.1 this proposition can easily be extended to the more general setting of a continuous map $\mathcal{A} : X \to Y^\star$ and subspaces $X_0 \subset X$ and $Y_0 \subset Y$.

Suppose now (and this is the case for both the approximation of harmonic fields and the approximation of Krylov subspaces we were discussing) that we have a family (X_h) of subspaces of X and *surjective* linear maps $\Lambda_h : X_0 \to X_h$ such that $\lim_h \|\Lambda_h - I\| = 0$. When $\|\Lambda_h - I\| < 1$, Λ_h is invertible (so X_h is closed), and $\|\Lambda_h^{-1}\| \leq (1 - \|\Lambda_h - I\|)^{-1}$, and $\Lambda_h^{-1} - I$ considered as a map $X_h \to X$ has norm $\|\Lambda_h^{-1} - I\| \leq (1 - \|\Lambda_h - I\|)^{-1}\|\Lambda_h - I\|$. In particular,

$$\delta(X_h, X_0) \leq (1 - \|\Lambda_h - I\|)^{-1}\|\Lambda_h - I\|. \tag{4.38}$$

We also trivially have

$$\delta(X_0, X_h) \leq \|\Lambda_h - I\|. \tag{4.39}$$

Given some continuous bilinear form a on X, we define is_h and is_0 by

$$\mathrm{is}_h = \inf_{u \in X_h} \sup_{v \in X_h} \frac{|a(u,v)|}{\|u\|\,\|v\|} \quad \text{and} \quad \mathrm{is}_0 = \inf_{u \in X_0} \sup_{v \in X_0} \frac{|a(u,v)|}{\|u\|\,\|v\|}. \tag{4.40}$$

Some tedious elementary manipulations yield (independently of the existence of the map Λ_h) the inequality

$$\mathrm{is}_h \geq \mathrm{is}_0 \frac{\big(1 - \delta(X_h, X_0)\big)}{\big(1 + \delta(X_0, X_h)\big)} - \|a\| \left(\frac{\big(1 + \delta(X_h, X_0)\big)}{\big(1 - \delta(X_0, X_h)\big)} \delta(X_0, X_h) + \delta(X_h, X_0) \right). \tag{4.41}$$

Now if we plug estimates (4.38) and (4.39) into (4.41) we can conclude that as $h \to 0$, is_h becomes greater than $\mathrm{is}_0 - \epsilon$ for any $\epsilon > 0$. Combining this fact with Proposition 4.8, we obtain the following.

PROPOSITION 4.9. *Let X be a reflexive Banach space and $\mathcal{A} : X \to X^\star$ be a continuous linear map. Suppose that X_0 is a closed linear subspace that splits yielding a projector P and that $\mathcal{A}$ induces an isomorphism $X_0 \to X_0^\star$, with an inf-sup estimate α_0. Suppose we have a family (X_h) of subspaces of X, equipped with surjective linear continuous maps $\Lambda_h : X_0 \to X_h$, such that $\lim_h \|I - \Lambda_h\| = 0$.*

Then for any $0 < \alpha < \alpha_0$ there is $h_0 > 0$ such that, for all $h < h_0$, $\mathcal{A}$ induces isomorphisms $X_h \to X_h^\star$, and, for all $l \in X^\star$, with the notations of Proposition 4.8, *we have*

$$\|u_h - u_0\| \leq \alpha^{-1}(1 + \alpha^{-1}\|\mathcal{A}\|)(1 + \|I - P\|)\,\|\Lambda_h - I\|\,\|l\|. \tag{4.42}$$

Applying this proposition to the discrete and continuous Krylov spaces yields the following.

COROLLARY 4.10. *Fix an $n \in \mathbb{N}$. Suppose that $\dim \mathfrak{K}^n = n+1$ and that the map $\mathfrak{K}^n \to \mathfrak{K}^{n\star}$ induced by $\mathcal{A}$ is invertible. Then there is $h_n > 0$ such that for all $h < h_n$ the map $\mathfrak{K}_h^n \to \mathfrak{K}_h^{n\star}$ induced by $\mathcal{A}$ is invertible; moreover, given $l \in \mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)^\star$, the solutions u_h^n and u^n of* (4.2) *and* (4.4) *satisfy an estimate of the form*

$$\|u_h^n - u^n\|_{-1/2} \leq C\|\Lambda_h^n - I\|, \tag{4.43}$$

where $\Lambda_h^n : \mathfrak{K}^n \to \mathfrak{K}_h^n$ is the unique linear map that satisfies, for $0 \leq i \leq n$,

$$\Lambda_h^n : (\mathcal{Z}\mathcal{A})^i \mathcal{Z} l \mapsto (\mathcal{Z}_h \mathcal{A}_h)^i \mathcal{Z}_h l|_{\mathsf{S}_h^1}. \tag{4.44}$$

Of course Theorem 4.7 shows that (for fixed n) $\|\Lambda_h^n - I\| \to 0$, and the above mentioned question of regularity is whether for smooth l we have estimates of the form $\|\Lambda_h^n - I\| \leq Ch^s$ for $s > 0$.

It is also possible to give a slightly different and more algorithmic variant of this corollary. Namely, define an "abstract" conjugate gradients algorithm by skipping all the h indices in some implementation in terms of $\mathcal{Z}_h$ and $\mathcal{A}_h$ of the conjugate gradient algorithm on S_h^1. (It should be checked that this is actually possible.) Then if the abstract algorithm is well defined up to iteration n there is $h_n > 0$ such that for all $h < h_n$ the discrete algorithm is well defined up to iteration n. Convergence of the iterates is described by the above corollary. Notice that it covers the case of algorithms that can skip full rank Krylov subspaces on which $\mathcal{A}$ is degenerate, as long as the next Krylov subspace is also full rank and $\mathcal{A}$ is nondegenerate on it, so-called Look-ahead algorithms described in Parlett, Taylor, and Liu [46]. Abstract conjugate gradient algorithms are common folklore even for non-Hermitian operators and described for instance in Gutknecht [30].

Remark. We have not proved that the spectral condition number of the restriction of $\mathcal{Z}_h\mathcal{A}_h$ to $(\mathsf{S}_h^1)^\wedge$ is uniformly bounded, and we even suspect that this may not be true. More precisely, the spectral radius of the endomorphism induced by $\mathcal{Z}_h\mathcal{A}_h$ on $(\mathsf{S}_h^1)^\wedge$ is uniformly bounded but perhaps not that of its inverse. The lack of this property (which was the guide and main focus of Steinbach and Wendland [51] and Christiansen and Nédélec [17]) was our principal motivation for proving the convergence of the Krylov spaces and the corresponding approximate solutions.

4.3. Evidence of superlinear convergence. The above discussion (in particular Corollary 4.10) suggests that the behavior of the approximate solutions (u_h^n) is similar to the behavior of (u^n), at least for moderate n (compared with $\dim \mathsf{S}_h^1 \approx h^{-2}$). Concerning the behavior of (u^n), the convergence theory is more satisfactory in the SPD case than in the non-SPD case we are dealing with.

Here we remind the reader how, in the infinite-dimensional SPD case, the property of inversion up to a compact operator leads to *superlinear* convergence. In other words, with the proposed preconditioner for the first kind integral equation EFIE, one recovers the kind of convergence usually associated with second kind integral equations. Though preconditioning was not a focus at the time, such estimates seem to date back to Hayes [31]. The theory does not directly apply to the studied case, but (for smooth surfaces) it does give complementary convergence estimates on some related preconditioning techniques, in particular those described in Steinbach and Wendland [51], which were the starting point of the method we have described here. We also believe these developments to give good *indication* on the behavior we can expect for our present problem.

Suppose X is a real Hilbert space, $\mathcal{A} : X \to X^\star$ is linear continuous, and induces a symmetric positive definite bilinear form. Suppose $\mathcal{Z} : X^\star \to X$ is linear continuous and symmetric. Given $l = \mathcal{A}u^\star \in X^\star$ define the Krylov spaces as above with real polynomials only. We suppose that we do not provide an approximate solution—other than 0—to start the algorithm, though this can easily be accounted for.

Remark first that $\mathcal{A}$ induces a scalar product on X with associated norm $\|\cdot\|_{\mathcal{A}}$ and that u^n solves (4.4) if and only if $u \mapsto \|u - u^\star\|_{\mathcal{A}}$ is minimal on $\mathfrak{K}^n$ at u^n. Thus for all real polynomials P such that $\deg P \leq n$,

$$\|u^n - u^\star\|_{\mathcal{A}} \leq \|P(\mathcal{Z}\mathcal{A})\mathcal{Z}\mathcal{A}u^\star - u^\star\|_{\mathcal{A}}. \tag{4.45}$$

Hence for all P such that $\deg P \leq n+1$, and $P(0) = 1$,

$$\|u^n - u^\star\|_{\mathcal{A}} \leq \|P(\mathcal{Z}\mathcal{A})u^\star\|_{\mathcal{A}}. \tag{4.46}$$

Then remark that $\mathcal{Z}\mathcal{A}$ is continuous, and symmetric with respect to the bilinear form induced by $\mathcal{A}$, and therefore has a resolution of the identity E on the spectrum $\sigma = \sigma(\mathcal{Z}\mathcal{A}) \subset \mathbb{R}$ such that we can write (we refer to Rudin [50, Chap. 12] for definitions and notations)

$$\|P(\mathcal{Z}\mathcal{A})u^\star\|_{\mathcal{A}}^2 \leq \int_\sigma |P(\lambda)|^2 \mathrm{d}E_{u^\star,u^\star} \tag{4.47}$$

$$\leq \sup\{|P(\lambda)|^2 \ : \ \lambda \in \sigma \cap \operatorname{supp} \mathrm{d}E_{u^\star,u^\star}\} \, \|u^\star\|_{\mathcal{A}}^2. \tag{4.48}$$

Finally in the case were $\mathcal{Z}\mathcal{A} - I$ is compact we can put $\sigma = \{1\} \cup \{\lambda_i : i \in \mathbb{N}\}$, where $(|\lambda_i - 1|)$ is a decreasing sequence converging to 0. To be sure that the algorithm is well-defined for all n we suppose that $\mathcal{Z}$ is positive definite. Then $\lambda_i \neq 0$, and we can define polynomials P_n by

$$P_n \ : \ P_n(\lambda) = \Pi_{i=0}^n (1 - \lambda/\lambda_i). \tag{4.49}$$

Remark that for any i and any λ such that $|1 - \lambda| \leq |1 - \lambda_i|$

$$|1 - \lambda/\lambda_i| = |(\lambda_i - 1 + 1 - \lambda)/\lambda_i| \leq 2|1 - 1/\lambda_i|, \tag{4.50}$$

which gives

$$\sup_{\lambda \in \sigma} |P_n(\lambda)| \leq 2^{n+1} \Pi_{i=0}^n |1 - 1/\lambda_i|. \tag{4.51}$$

Since $|1 - 1/\lambda_i| \to 0$, this immediately implies superlinear convergence. More generally, if (λ_i) is a sequence of nonzero *complex* numbers, such that $(|\lambda_i - 1|)$ decreases

to 0, the estimates (4.50) (for complex λ) and (4.51) (with the same definition of P_n, and still with $\sigma = \{1\} \cup \{\lambda_i \ : \ i \in \mathbb{N}\}$) remain true, which might be of interest to other Krylov-subspace algorithms applied to non-SPD operators. Moreover, if the asymptotic behavior of the eigenvalues of the residual $\mathcal{Z}\mathcal{A} - I$ is known and we have an estimate of the form $|1-\lambda_i| \leq Ci^{-\alpha}$ for some $C > 0$, $\alpha > 0$, we get the convergence estimate (for a larger C and the same α)

$$\|u^n - u^\star\|_{\mathcal{A}} \leq C^n (n!)^{-\alpha} \|u^\star\|_{\mathcal{A}}. \tag{4.52}$$

More explicitly, returning to the case of operators on a smooth compact Riemannian manifold Γ, if the dimension of the manifold is N and the residual is an operator of order $-s$ for some $s > 0$ (i.e., in terms of Sobolev spaces, continuous $\mathrm{H}^{s'}(\Gamma) \to \mathrm{H}^{s'+s}(\Gamma)$ for all s') which commutes with the Laplacian on Γ, this holds with $\alpha = s/N$. This can be deduced from the eigenvalue asymptotics of the Laplacian, for which we refer to Taylor [52, Vol. II, Chap. 8]. An alternative and more general approach based on trace-class theory is exposed in Winther [56] and also gives a factor of the form $C^n(n!)^{-\alpha}$. When Γ has symmetries an even larger α might hold in the estimate (4.52) due to the degeneracy of eigenvalues. For the determination of the orders of different integral operators we refer to Nédélec [43]; in particular, the order of the residual in our preconditioning strategy for the EFIE is -2 (though we do not even claim to have proved that the PCG does not break down in this case).

5. Numerical results. In order to evaluate the performance of the preconditioner it is customary to show the convergence graphs. We use the notations of (1.29), and we denote by U_h^n the tuple of coordinates of the approximate solution u_h^n at iteration n in the chosen basis. The convergence graphs are then of the form

$$n \mapsto \log_{10}(\|A_h(k)U_h^n - V_h\|/\|V_h\|) \tag{5.1}$$

for a given choice of norms on the tuples. The standard norm is the ℓ_2 norm on tuples. From a functional point of view this is not very natural, but on the other hand functional norms are not readily computable—with a notable exception for those we defined in section 3.9.

5.1. Sphere. We start by showing convergence graphs for the canonical example of diffraction of a plane wave by the unit sphere for the wave lengths $\lambda = 8,\ 4,\ 2,\ 1$ ($k = 2\pi/\lambda$); see Figure 5.1. The discretization of the sphere has 2252 vertices and 4500 triangles, leading to 6750 degrees of freedom for lowest order Raviart–Thomas finite elements.

We consider here the case of the preconditioner using the same Galerkin space as the original and use a complex symmetric conjugate gradients algorithm.

Each graphic displays three curves. The thin line (upper graph) is obtained without preconditioning for the ℓ_2 norm on tuples; the dotted line (which stagnates) is obtained with the proposed preconditioner for the ℓ_2 norm on tuples; the third line (the bottom graph) is obtained with the proposed preconditioner for the natural norm defined in section 3.9.

We make the following comments:

- With the preconditioner, each iteration is a little more than twice as slow as without any preconditioner: we apply the Galerkin matrix once more, and do a considerable amount of sparse matrix manipulations.

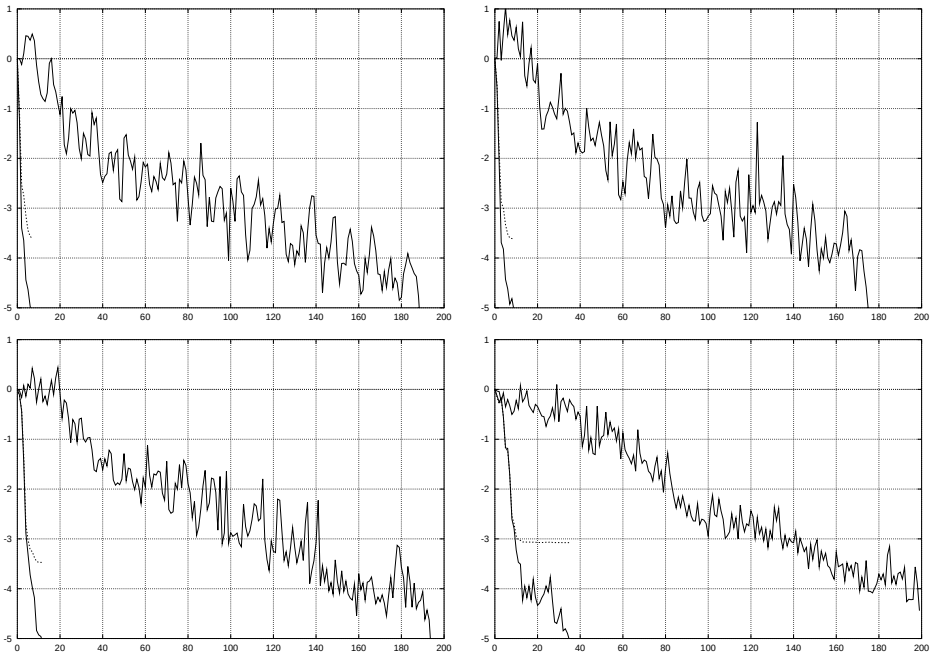


FIG. 5.1. *Convergence graphs for the unit sphere at* $\lambda = 8,\ 4,\ 2,\ 1$ *(top left to bottom right).*

- In the preconditioned case and in particular at $\lambda = 1$, the ℓ_2 norm (dotted line) of the residual stagnates, whereas the natural norm continues to decrease, confirming that the variational problem is indeed solved on a strict subspace.
- The preconditioner is particularly efficient at low frequencies—which is the only really ill-conditioned case on spheres.
- Usual stopping criteria vary from 10^{-2} to 10^{-5}, depending on the accuracy of the result required; for all these the preconditioned algorithm is several times faster than the algorithm without preconditioning.
- The auxiliary problems in the preconditioner were solved iteratively with a tolerance of 10^{-7} (saddle-point problem) and 10^{-8} (L^2 projections); accumulation of such errors and other round-off errors could also partly explain the instabilities observed at the last iterations at $\lambda = 1$.
- The far-field patterns deduced from the electric currents computed iteratively with and without preconditioner were not graphically distinguishable from those computed by standard factorization.
- Using the Brezzi–Douglas–Marini finite elements in the preconditioner (while still using Raviart–Thomas for the original problem) yields slightly better accuracy and requires slightly fewer iterations. Also the dotted line does not stagnate, confirming that the equation is then solved variationally on a much larger subspace. However, each iteration is much slower, since there are twice as many degrees of freedom in the preconditioner (at the lowest level).

5.2. Cavity. We consider now diffraction by a cavity. In cavities trapped rays create long range nontrivial interactions. This is a considerably more challenging problem than scattering by convex objects, since, without preconditioning, it happens that the iterates do not converge.

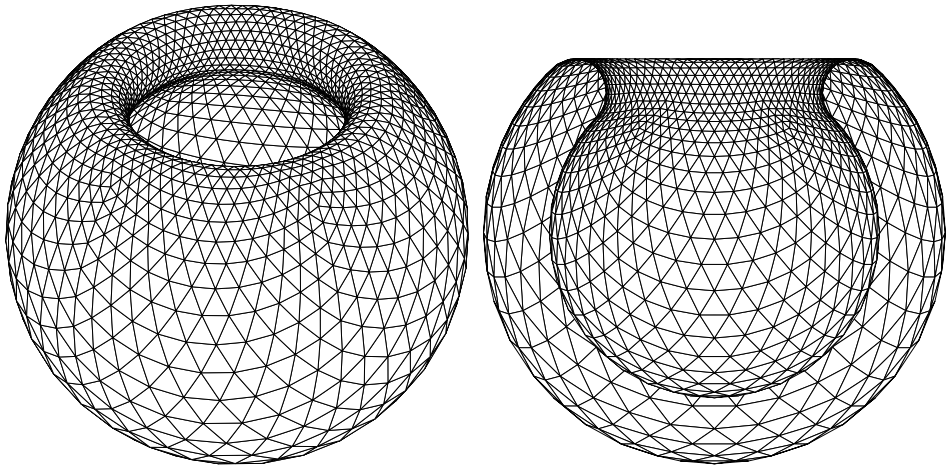


FIG. 5.2. *Cavity seen from outside and vertical section.*

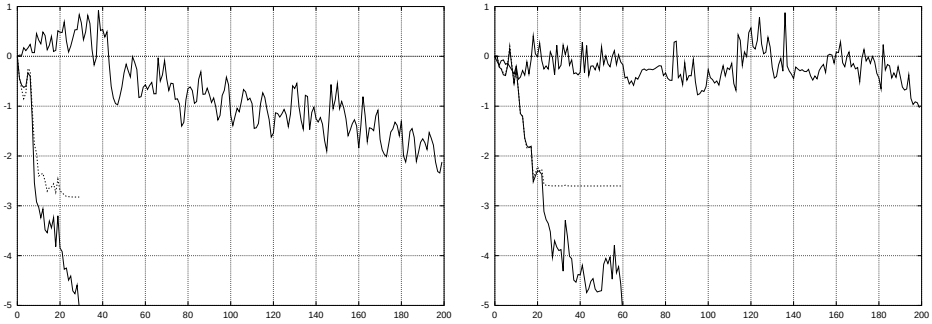


FIG. 5.3. *Convergence graphs for the cavity at* $\lambda = 8,\ 2$ *(left and right).*

The particular geometry we consider can be described as follows: take two concentric spheres, one with radius 5/6 and one with radius 7/6; excavate a cone with top at the origin and half-angle $\pi/4$ and join the interior surface with the exterior one with half a torus. See Figure 5.2. The mesh of the cavity was constructed from a mesh of the unit sphere (in fact the same one as in the preceding example) by successive deformations.

The cavity is lit by a horizontally polarized plane wave entering the cavity tangentially to its walls. (The wave vector makes an angle of $3\pi/4$ with the vertical direction.) The convergence graphs obtained for the cavity are displayed in Figure 5.3.

Notice that at $\lambda = 2$, the preconditioner not only speeds up convergence, it actually enables it.

In the preconditioned case, we observe a slow-down in the convergence, which we interpret as stemming from the fact that the Calderon formulas are less well represented on a discrete level. (The discrete iterates depart from the continuous ones.) It seems also that the slow-down always occurs slightly after the ℓ_2 norm of the residual stagnates. This stagnation could be indication that the current iterate is as close to the exact continuous solution as the exact Galerkin solution. Thus the

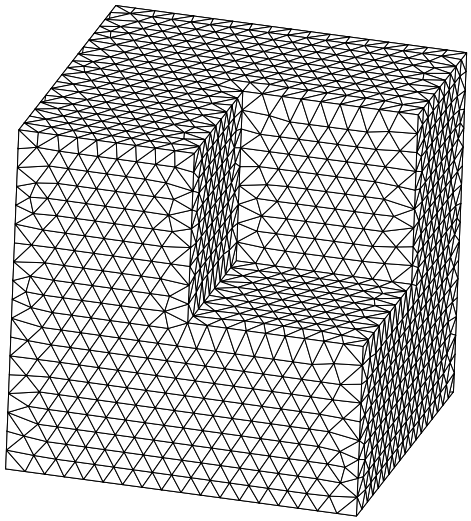


FIG. 5.4. *Indented cube seen from outside.*

stagnation in the ℓ_2 norm of the residual would be a good stopping criterion and then the slow-down would never be observed.

5.3. Indented cube. The Galerkin discretization of the EFIE is well known to perform well not only on smooth surfaces but also for the polyhedral ones that often occur in applications. Though we have justified the preconditioning technique only for smooth surfaces, it appears to perform well also on such nonsmooth ones. However, when the preconditioning operator $Z : X' \to X$ does not invert the operator $A : X \to X'$ up to a compact residual, but rather is such that ZA is an automorphism of X, one expects the ideal conjugate gradients algorithm to converge not superlinearly, but rather linearly, as is easily proven for SPD operators.

We show numerical results for the following geometry. The scattering object is the indented cube $[-1,1]^3 \backslash]0,1]^3$. The interior domain contains several types of singularities, both convex and nonconvex. Also when a plane wave with wave-vector $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, with $\sigma_i < 0$, hits the reentrant corner, geometrical optics predicts that it should be scattered mainly in the direction $-\sigma$, after three reflections. The mesh used for the numerical experiments is shown in Figure 5.4. It has 2164 vertices and 4324 triangles, leading to 6486 edges (and degrees of freedom for Raviart–Thomas finite elements).

In Figure 5.5 we show the convergence graphs obtained for an incident plane wave with wave-vector positively proportional to $(-1,-1,-1)$, with wavelength $\lambda = 8$ and $\lambda = 4$, and with horizontal polarization. Contrary to the case of smooth surfaces there might be significant loss of accuracy when solving the Galerkin problem on $(\mathsf{S}_h^1)^\wedge$ rather than S_h^1 when lowest degree Raviart–Thomas fields are used both in the problem formulation and the preconditioner. This problem would be remedied using Brezzi–Douglas–Marini fields in the preconditioner, since then $(\mathsf{S}_h^1)^\wedge$ has very low codimension in S_h^1. However, for the wavelengths used here the far-field patterns were not graphically distinguishable.

Perspectives. The study of the method on nonsmooth surfaces is still ongoing. In particular, the evaluation of the impact of singularities in the surface and corre-

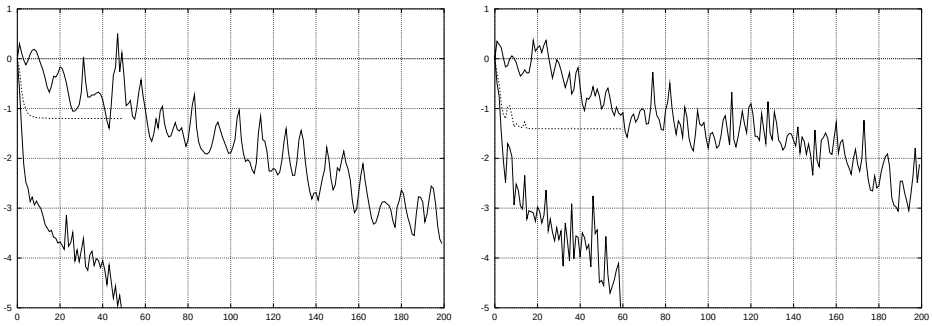


FIG. 5.5. *Convergence graphs for the indented cube at* $\lambda = 8,\ 4$ *(left and right).*

sponding mesh refinement strategies is important for many applications and could perhaps be achieved using recent results from Buffa, Costabel, and Schwab [13] and Hiptmair and Schwab [33].

As remarked on the numerical experiments, the preconditioning technique displays good stability at low frequencies. This stability can be enhanced by making the discrete Helmholtz decompositions still more explicit. For instance, focusing on the preconditioner, one applies separately the two terms of the operator in (1.18) to the two terms of the vector in (3.24). We shall come back elsewhere to this point, which is important for simulating semiconductor devices.

The method can also be extended to treat scattering by perfectly conducting simplical complexes (including open surfaces as well as branched ones, where more than two surfaces meet at an edge). In these cases one can no longer keep the same type of Galerkin spaces in the preconditioner as in the variational formulation of the EFIE, and the Calderon formulas, which require the surface to be orientable, need to be adapted. For some generalizations of the method described here we have obtained speed-ups comparable to the above ones, though the justifications are as of today at best intuitive.

Acknowledgments. We thank F. Béreux for his precious help with this project. We also thank both referees for many constructive remarks.

REFERENCES

[1] T. ABBOUD, *Etude mathématique et numérique de quelques problèmes de diffraction d'ondes électromagnétiques*, Ph.D. thesis, Ecole Polytechnique, Palaiseau, France, 1991.

[2] I. BABUSKA, *Error bounds for the finite element method*, Numer. Math., 16 (1971), pp. 322–333.

[3] R. BARRETT, M.W. BERRY, T.F. CHAN, J. DEMMEL, J. DONATO, J. DONGARRA, V. EIJKHOUT, R. POZO, C. ROMINE, AND H. VAN DER VORST, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1993.

[4] A. BENDALI, *Numerical analysis of the exterior boundary value problem for the time-harmonic Maxwell equations by a boundary finite element method, Part* 1: *The continuous problem*, Math. Comp., 43 (1984), pp. 29–46.

[5] A. BENDALI, *Numerical analysis of the exterior boundary value problem for the time-harmonic Maxwell equations by a boundary finite element method, Part* 2: *The discrete problem*, Math. Comp., 43 (1984), pp. 47–68.

[6] D. BOFFI, *Discrete compactness and Fortin operator for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[7] D. BOFFI, *A note on the de Rham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[8] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (1999), pp. 121–140.

[9] A. Bossavit, *Mixed finite elements and the complex of Whitney forms*, in The Mathematics of Finite Elements and Applications VI (Uxbridge, 1987), Academic Press, London, 1988, pp. 137–144.

[10] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 8 (1974), pp. 129–151.

[11] F. Brezzi, J. Douglas, Jr., and L.D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[12] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[13] A. Buffa, M. Costabel, and C. Schwab, *Boundary Element Methods for Maxwell's Equations on Non-Smooth Domains*, Technical report, SAM-ETH, Zürich, Switzerland, 2001.

[14] M. Cessenat, *Mathematical Methods in Electromagnetism, Linear Theory and Applications*, World Scientific, River Edge, NJ, 1996.

[15] S.H. Christiansen, *Discrete Fredholm properties and convergence estimates for the EFIE*, CMAP Technical report 454, Ecole Polytechnique, Palaiseau, France, 2001, submitted.

[16] S.H. Christiansen, F. Béreux, J.-C. Nédélec, and J.-P. Martinaud, *Algorithme de simulation électromagnétique, notamment des performances d'une antenne*, Thalès Airborne Systems, Patent in France, INPI Reg. No. 00 07456, 2000.

[17] S.H. Christiansen and J.-C. Nédélec, *Des préconditionneurs pour la résolution numérique des équations intégrales de frontière de l'acoustique*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 617–622.

[18] S.H. Christiansen and J.-C. Nédélec, *Preconditioners for the boundary element method in acoustics*, in Proceedings of the Fifth International Conference on Mathematical and Numerical Aspects of Wave Propagation, Santiago de Compostella, Spain, 2000, SIAM, Philadelphia, 2000, pp. 776–781.

[19] S.H. Christiansen and J.-C. Nédélec, *Des préconditionneurs pour la résolution numérique des équations intégrales de frontière de l'électromagnétisme*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 733–738.

[20] D.L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[21] H. Contopanagos, B. Dembart, M. Epton, J.J. Ottusch, V. Rokhlin, J. Visher, and S. Wandzura, *Well-Conditioned Boundary Integral Equations for Three-Dimensional Electromagnetic Scattering*, Research Report YALEU/DCS/RR-1198, Yale University, New Haven, CT, 2000.

[22] M. Costabel, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988) pp. 613–626.

[23] M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2 projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[24] A. DeLaBourdonnaye, *Décomposition de $\mathrm{H}_{\mathrm{div}}^{-1/2}(\Gamma)$ et nature de l'opérateur de Steklov-Poincaré du problème extérieur de l'électromagnétisme*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 369–372.

[25] L. Demkowicz, *Asymptotic convergence in finite and boundary element methods, Part* 1: *Theoretical results*, Comput. Math. Appl., 27 (1994), pp. 69–84.

[26] L. Demkowicz, *Asymptotic convergence in finite and boundary element methods, Part* 2: *The LBB constant for rigid and elastic scattering problems*, Comput. Math. Appl., 27 (1994), pp. 93–109.

[27] L. Demkowicz, P. Monk, C. Schwab, and L. Vardapetyan, *Maxwell eigenvalues and discrete compactness in two dimensions*, Comput. Math. Appl., 40 (2000), pp. 589–605.

[28] G.B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., John Wiley, New York, 1999.

[29] R.W. Freund, *Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices*, SIAM J. Sci. Statist. Comp., 13 (1992), pp. 425–448.

[30] M.H. Gutknecht, *Lanczos-Type Solvers for Non-Symmetric Linear Systems of Equations*, Acta. Numer. 6, Cambridge University Press, Cambridge, UK, 1997, pp. 271–397.

[31] R.M. Hayes, *Iterative Methods of Solving Linear Problems on Hilbert Space*, Nat. Bur. Standards Appl. Math. Ser. 39, U.S. Government Printing Office, Washington, D.C., 1954, pp. 71–103.

[32] R. Hiptmair, *Canonical construction of finite elements*, Math. Comp., 68 (1999), pp. 1325–1346.

[33] R. Hiptmair and C. Schwab, *Natural BEM for the Electric Field Integral Equation on Polyhedra*, Technical Report, SAM-ETH, Zürich, Switzerland, 2001.

[34] J.D. Jackson, *Classical Electrodynamics*, 2nd ed., John Wiley, New York, London, Sydney, 1975.
[35] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, New York, 1976.
[36] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.
[37] F. Kikuchi, *On a discrete compactness property for the Nédélec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. 1A Math., 36 (1989), pp. 479–490.
[38] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. I, Dunod, Paris, 1968.
[39] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. II, Dunod, Paris, 1968.
[40] J.-C. Nédélec, *Curved finite element methods for the solution of singular integral equations on surfaces in* $\mathbb{R}^3$, Comput. Methods Appl. Mech. Engrg., 8 (1976), pp. 61–80.
[41] J.-C. Nedelec, *Computation of eddy currents on a surface in* R^3 *by finite element methods*, SIAM J. Numer. Anal., 15 (1978), pp. 580–594.
[42] J.-C. Nédélec, *Notions sur les techniques d'éléments finis*, Publications SMAI, Ellipses, Paris, France, 1991.
[43] J.-C. Nédélec, *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, Springer-Verlag, Berlin, 2001.
[44] R.A. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems*, SIAM J. Numer. Anal., 19 (1982), pp. 349–357.
[45] L. Paquet, *Problèmes mixtes pour le système de Maxwell*, Ann. Fac. Sci. Toulouse Math. (5), 4 (1982), pp. 103–141.
[46] B.N. Parlett, D.R. Taylor, and Z.A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[47] S.S.M. Rao, D.R. Wilton, and A.W. Glisson, *Electromagnetic scattering by surfaces of arbitrary shape*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 409–418.
[48] P.A. Raviart and J.-M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, New York, 1977, pp. 292–315.
[49] J.E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II: Finite Element Methods (Part 1), P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 523–640.
[50] W. Rudin, *Functional Analysis*, 2nd ed., McGraw–Hill, New York, 1991.
[51] O. Steinbach and W.L. Wendland, *The construction of some efficient preconditioners in the boundary element method*, Adv. Comput. Math., 9 (1998), pp. 191–216.
[52] M. Taylor, *Partial Differential Equations, Vol.* I: *Basic Theory*, Springer-Verlag, New York, 1996.
[53] M. Taylor, *Partial Differential Equations, Vol.* II: *Qualitative Studies in Linear Equations*, Springer-Verlag, New York, 1996.
[54] I. Terrasse, *Résolution mathématique et numérique des équations de Maxwell instationnaires par une méthode de potentiels retardés*, Ph.D. thesis, Ecole Polytechnique, Palaiseau, France, 1993.
[55] W.L. Wendland, *Strongly elliptic boundary integral equations*, in The State of the Art in Numerical Analysis (Birmingham 1986), Inst. Math. Appl. Conf. Ser. New Ser. 9, A. Iserles and M. Powell, eds., Oxford University Press, New York, 1987, pp. 511–562.
[56] R. Winther, *Some superlinear convergence results for the conjugate gradient method*, SIAM J. Numer. Anal., 17 (1980), pp. 14–17.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1136–1183

SECOND ORDER NUMERICAL METHODS FOR FIRST ORDER HAMILTON–JACOBI EQUATIONS*

ADAM SZPIRO† AND PAUL DUPUIS‡

Abstract. We present practical numerical methods which produce provably second order approximations for a class of stationary first order Hamilton–Jacobi partial differential equations. Using probabilistic methods, we derive high order asymptotic expansions for a first order method and then use those results to design second order methods. We prove second order convergence for the solution and for its gradient on a subset of the domain where the solution is smooth. Although we limit our attention to second order schemes, in principle the techniques in this paper can be extended to arbitrarily high order methods. Examples illustrate the rate of convergence as well as global sharp resolution of discontinuities. The Hamilton–Jacobi equations we consider correspond to deterministic optimal control problems, and our rate of convergence results are valid for the value functions and for the optimal feedback controls.

Key words. Hamilton–Jacobi equations, numerical approximation, second order convergence, asymptotic expansion, Markov chain approximation, finite difference approximation

AMS subject classifications. 35B37, 49L20, 60F10, 60F05, 65N06, 65N12, 93E25

PII. S003614299935704X

1. Introduction.

This paper is concerned with the numerical solution of a class of first order stationary Hamilton–Jacobi partial differential equations (PDEs). We begin by specifying the PDEs of interest and giving a practical description of our proposed second order numerical methods. The PDEs that we study arise in connection with a variety of applications in computer vision, large deviations, and robust control [3], and they can be interpreted as dynamic programming equations for a class of deterministic optimal control problems on a finite domain. Under our assumptions, the value function $V^0(x)$ for a given control problem is the unique viscosity solution to the corresponding Hamilton–Jacobi PDE, and the optimal feedback control function $u^0(x)$ is obtained by taking the argument in the optimization part of the Hamilton–Jacobi equation on those parts of the domain where it is uniquely defined. We exploit this connection to develop the theoretical underpinnings of the proposed second order numerical methods. We give a detailed asymptotic analysis of a first order Markov chain-based numerical method and use that analysis to motivate our second order methods. Finally, we demonstrate under some assumptions that the proposed second order methods in fact yield second order convergent results and give several illustrative examples.

Much of the difficulty in solving equations of this type derives from the fact that the solutions need not be smooth on the entire domain. Under our assumptions, $V^0(x)$ is globally Lipschitz, but it need not be differentiable at all points. The control function $u^0(x)$ is closely related to the gradient of $V^0(x)$. It need not be uniquely

*Received by the editors May 28, 1999; accepted for publication (in revised form) January 7, 2002; published electronically September 12, 2002. This research was supported in part by the National Science Foundation grant NSF-DMS-9704426, the Army Research Office contract DAAH04-96-1-0075, and the Office of Naval Research contract ONR-N000014-96-1-0276.

http://www.siam.org/journals/sinum/40-3/35704.html

†Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02420 (adam@ll.mit.edu).

‡Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (dupuis@cfm.brown.edu).

defined at all points, and it is often discontinuous across those points where it is not uniquely defined. Typically, however, there are regions of strong regularity (RSRs) which are open and dense in the domain, and in those regions both $V^0(x)$ and $u^0(x)$ are as smooth as the problem data. Our asymptotic results hold on certain subsets of the RSRs, and we propose numerical methods which provide provably second order estimates for $V^0(x)$ and $u^0(x)$ on those regions, without sacrificing global convergence properties.

To our knowledge, there are no similar results which rigorously establish second order convergence of any fully implementable numerical method for nonlinear Hamilton–Jacobi PDEs in multiple dimensions (or for the closely related nonlinear conservation laws). Computational examples illustrate that the proposed methods also lead to well-behaved approximations which are second order convergent in the regions of interest. Given the difficulty of the problem, strong assumptions are to be expected for the theoretical analysis, and a major drawback to the current methodology is that one cannot determine the boundaries of the RSRs directly from the data. In addition, a key limiting assumption is contained in Assumption 5.1. That assumption requires that the optimal feedback controls be of fixed sign on the regions of interest, a condition which is typically only satisfied on a small subset of the domain. Nonetheless, the numerical methods appear to work well even where this assumption is violated. Perhaps more importantly, the analysis clarifies what sort of additional smoothness in the numerical methods would be required to obtain rigorous convergence results without such a restrictive assumption.

The numerical methods we consider are of finite difference type. We follow Kushner [24] and Kushner and Dupuis [25] and work in a probabilistic framework, beginning with the classical first order methods and proceeding to our proposed second order methods. The first order numerical methods can also be analyzed in a purely PDE framework using viscosity solution techniques [1, 6]. Several approaches to obtaining higher order approximations have been suggested in the literature, especially for the finite time case [12, 13, 27, 28, 23, 26, 22]. Many of these are based on numerical methods for conservation laws and have produced convincing numerical results. However, complete convergence proofs are not available, and adaptation to stationary problems has proven challenging [20]. We also note that the numerical method for conservation laws given in [21] is similar in philosophy to the first of our proposed second order methods.

We specify some notation. Let $\mathbb{R}^n$ be the n-dimensional Euclidian space, and let $\mathbb{Z}^n$ be the subset of $\mathbb{R}^n$ consisting of n-tuples of integers. For vectors $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ is the scalar product, $\|x\| = \sqrt{\langle x, x\rangle}$ is the Euclidean norm, $\|x\|_1 = \Sigma_{i=0}^n |x_i|$ is the l^1-vector norm, and $|x| = (|x_1|, \ldots, |x_n|)$ is the componentwise absolute value. For a process $X(\cdot)$ taking values in $\mathbb{R}^n$ and for $S < +\infty$, $\| X(\cdot) \|_S = \sup_{0\le t\le S} \|X(t)\|$ is the uniform L^2 norm. For any two subsets A and A' of $\mathbb{R}^n$, $d(A, A')$ denotes the minimum Euclidean distance between $\bar{A}$ and $\bar{A}'$. The positive part of a scalar is $a^+ = \max(a, 0)$, and its negative part is $a^- = -\min(a, 0)$. For a vector, the positive and negative parts are taken componentwise so that $x^\pm = (x_1^\pm, \ldots, x_n^\pm)$. In general, we use subscripts to denote the components of a vector, while superscripts index possibly vector valued quantities. Thus, x_i is always a scalar quantity, while x^i may denote a vector.

For a smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$ and for a positive integer r, put $D_i^r f(x) = \frac{\partial^r}{\partial^r x_i} f(x)$. The diagonal of the array of rth order partial derivatives is denoted by

$$D_\setminus^r f(x) = (D_1^r f(x), \ldots, D_n^r f(x)).$$

In the case $r = 1$, we use the standard notation $Df(x) = D^1_{\setminus} f(x)$ for the gradient of f. For $h > 0$, the operators $D^{h,\pm}$ are one-sided finite difference approximations to the gradient operator. The ith component of $D^{h,+}f(x)$ is given by

$$D_i^{h,+} f(x) = \frac{f(x + he_i) - f(x)}{h},$$

while the ith component of $D^{h,-}f(x)$ is given by

$$D_i^{h,-} f(x) = \frac{f(x) - f(x - he_i)}{h}.$$

In addition, $D^{h,c}f(x)$ is the centered difference approximation to the gradient vector $Df(x)$, while $D^{2,h}_{\setminus} f(x)$ is the centered difference approximation to the diagonal second derivative vector $D^2_{\setminus} f(x)$.

2. PDE and numerical methods. In this section, we describe the problem of interest and give a practical description of the proposed numerical methods. The remainder of the paper is concerned with a rigorous analysis of what is discussed here. Let $G \subset \mathbb{R}^n$ be open with compact closure, and assume that G satisfies uniform interior and exterior cone conditions (see [3] for definitions). Let b and c be C^∞ functions from $\mathbb{R}^n$ to $\mathbb{R}$, and let a be a C^∞ function from $\mathbb{R}^n$ to the space of symmetric positive definite $n \times n$ matrices. Notice that a is uniformly positive definite on G. Assume that $c(x) \geq c_0 > 0$ on G. Define

$$L(x,u) = \frac{1}{2}\langle (u - b(x)), a^{-1}(x)(u - b(x))\rangle + c(x), \tag{2.1}$$

and consider the PDE

$$\inf_u \left[\langle u, DV^0(x)\rangle + L(x,u)\right] = 0, \tag{2.2}$$

with the continuous boundary condition $V^0(x) = 0$ on ∂G. This class of PDEs is of Hamilton–Jacobi type with Hamiltonians convex in the gradient $DV^0(x)$ and includes several PDEs of general interest. For example, the Eikonal equation $\|DV^0(x)\|^2 = 2c(x)$ is obtained by letting $a(x)$ be the identity matrix and taking $b(x) = 0$. Equation (2.2) may not have a classical solution, but it is proved in [2, Theorem 6.1] to have a unique nonnegative viscosity solution $V^0(x)$. (See the discussion around Lemma 3.1.) The corresponding optimal control $u^0(x)$ is not uniquely defined everywhere, but it is given by $u^0(x) = -a(x)DV^0(x) + b(x)$ where $V^0(x)$ is smooth.

In general, there is no analytical way to identify the solution $V^0(x)$, so we are interested in numerical approximation methods. The starting place for our discussion is the first order method described in detail in section 3 and in a slightly more general setting in [3] . Those discussions are in terms of a Markov chain interpretation that is convenient for theoretical analysis. Here we restrict ourselves to a formal description of the numerical method. For $h > 0$ and for $u \in \mathbb{R}^n$ we define

$$\overline{\Delta t}^h(u) = \begin{cases} \frac{h}{\|u\|_1} & u \neq 0, \\ h & u = 0 \end{cases} \tag{2.3}$$

and

$$p^h(x, y|u) = \begin{cases} \frac{u_i^{\pm}}{\|u\|_1} & \text{if } y = x \pm he_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Then we take as a first order numerical approximation the unique solution $V^h(x)$ to

$$V^h(x) = \inf_u \left[\sum_{y \in \mathbb{R}^n} p^h(x, y|u) V^h(y) + \overline{\Delta t}^h(u) L(x, u) \right] \tag{2.5}$$

on $G^h = h\mathbb{Z}^n \cap G$ with zero boundary condition on the complement of G^h. Additionally, we define the approximate optimal control $u^h(x)$ to be the minimizing argument in (2.5). It is easy to see that (2.5) is equivalent to an upwind first order finite difference approximation to (2.2). As discussed in [3], this discrete equation can be solved efficiently by a Gauss–Seidel iteration, with the optimal value at each step computed analytically; the reader interested in implementation should consult [3] for further details. See also [30] for an interesting discussion of an alternative to Gauss–Seidel iteration for a subclass of problems. It can be shown that $V^h(x) \to V^0(x)$ on G. Additionally, it can be shown on the RSRs that $u^h(x) \to u^0(x)$ and that the convergence $V^h(x) \to V^0(x)$ is first order. See sections 3 and 4 and the reference cited there [3, 11].

Our primary interest in this paper is in second order numerical methods. Our second order methods are motivated by the formal asymptotic expansion

$$V^h(x) = V^0(x) + he^1(x) + O(h^2), \tag{2.6}$$

where formally $e^1(x)$ satisfies

$$\langle u^0(x), De^1(x) \rangle + \frac{1}{2} \langle |u^0(x)|, D^2_{\backslash} V^0(x) \rangle = 0 \tag{2.7}$$

with zero boundary condition. A rigorous analysis of this asymptotic expansion is given in sections 4 and 5. In this analysis, we draw liberally from ideas developed in [15, 10]. Note also that related results are obtained using viscosity solution methods in [17]. We observe that the asymptotic expansion motivates two distinct approaches to obtaining a second order approximation to $V^0(x)$.

Approximation Method I. The idea is to directly construct an approximation $e^{1,h}(x)$ to $e^1(x)$. Evidently, the unknowns $V^0(x)$ and $u^0(x)$ appear in relation (2.7), but it is possible to use the first order approximations $V^h(x)$ and $u^h(x)$ to construct the approximation by way of the linear equation

$$\langle u^{h,+}, D^{h,+} e^{1,h} \rangle - \langle u^{h,-}, D^{h,-} e^{1,h} \rangle + \frac{1}{2} \langle |u^h|, D^{2,h}_{\backslash} V^h \rangle = 0 \tag{2.8}$$

with zero boundary condition. This equation can be solved iteratively using an efficient Gauss–Seidel method. The second order approximations are defined to be

$$V^{h,*}(x) = V^h(x) - he^{1,h}(x)$$

and

$$u^{h,*}(x) = -a(x) D^{h,c} V^{h,*}(x) + b(x)$$

for all x in G^h.

Approximation Method II. Alternatively, (2.6) suggests Richardson extrapolation as a method to obtain a second order approximation. As such, we may define new approximations by

$$V^{h,*}(x) = 2V^h(x) - V^{2h}(x)$$

and

$$u^{h,*}(x) = 2u^h(x) - u^{2h}(x).$$

The form of $u^{h,*}(x)$ is motivated by an asymptotic expansion for $u^0(x)$ similar in form to (2.6). It should be noted that the $V^{h,*}(x)$ obtained here is not the same as in Method I. Both involve removing an approximation to the error term $e^1(x)$, but the approximations used in the two methods are in general not the same.

A detailed discussion of the derivation and convergence properties of these second order approximation methods is found in section 6, and computational examples are given in section 7. Both methods produce consistent numerical second order approximations to $V^0(x)$ and $u^0(x)$ on the RSRs when convergence is measured in the L^1 norm. Results are less consistent when measured in the L^∞ norm. No spurious oscillations are observed in the approximations to $V^0(x)$. Some oscillations are evident near discontinuities in the approximations to $u^0(x)$ (which is related to the gradient of $V^0(x)$), but the oscillations appear to be bounded as the grid is refined, indicating that the second order methods are numerically stable for both $V^0(x)$ and $u^0(x)$.

3. Connection with dynamic programming and first order numerical approximation. We describe a deterministic optimal control problem and its connection to the PDE discussed in the previous section. Then we describe in a Markov chain framework the first order numerical method used to approximate the solution, and we recall results which guarantee convergence of the numerical value function and of the numerical feedback control to their respective counterparts in the problem being approximated.

Let $G \subset \mathbb{R}^n$, $a(x)$, $b(x)$, $c(x)$, and $c_0 > 0$ be as in the previous section. For a control $\underline{u}^0(t)$ which is in $L^2([0,S];\mathbb{R}^n)$ for all $S < +\infty$ and for an initial condition $x \in G$, we define $\underline{X}^0(t)$ by the dynamics

$$\underline{X}^0(t) = x + \int_0^t \underline{u}^0(s)ds, \tag{3.1}$$

up to the time when it exits from the domain G. The corresponding generator $\mathcal{L}_u^0$ is given by

$$\mathcal{L}_u^0 f = \langle u, Df \rangle \tag{3.2}$$

for any smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$. We define the exit time $\underline{\tau}^0 = \inf[t : \underline{X}^0(t) \notin G]$. For the running cost

$$L(x,u) = \frac{1}{2}\langle (u - b(x)), a^{-1}(x)(u - b(x))\rangle + c(x),$$

we define the payoff functional

$$J^0(x,\underline{u}^0) = \int_0^{\underline{\tau}^0} L(\underline{X}^0(t), \underline{u}^0(t))dt.$$

The problem is to minimize the payoff by choosing a suitable control. Define the value function,

$$V^0(x) = \inf_{\underline{u}^0} J^0(x,\underline{u}^0),$$

where the infimum is over controls $\underline{u}^0$ which are in $L^2([0,S];\mathbb{R}^n)$ for all $S < +\infty$. We employ the underscore notation here to indicate trajectories which are obtained from an arbitrary control. The same notations, without the underscores, will be used later to refer to trajectories which are obtained through the application of an optimal control.

The dynamics in (3.1) involve an open loop control $\underline{u}^0(t)$ which is defined for all $t > 0$. It is generally desirable, from the point of view of robustness and for convenience of implementation, to consider controls which can be represented in the feedback form

$$\underline{X}^0(t) = x + \int_0^t \underline{u}^0(\underline{X}^0(s))ds. \tag{3.3}$$

A key feature of the RSRs is that the optimal open loop controls for all initial conditions in a region of strong regularity correspond to a unique smooth feedback function $u^0(x)$. The following lemma is proved by elementary arguments in [11, Lemma 2.1]. We will use the $T < +\infty$ from this lemma frequently in our analysis.

LEMMA 3.1. *$V^0(x)$ is bounded and uniformly Lipschitz on G, and there exists $T < +\infty$ such that every optimal trajectory exits from G by time $T-1$. Furthermore, there exists a compact set $U \subset \mathbb{R}^n$ such that every optimal open loop control is contained in the interior of U for each $0 \le t \le T-1$.*

The value function V^0 need not be differentiable on the entire domain G, but, given Lemma 3.1, it follows from [2, Theorem 6.1] that it is the unique nonnegative viscosity solution on G to the Hamilton–Jacobi dynamic programming equation (DPE)

$$\inf_u \left[\mathcal{L}_u^0 V^0(x) + L(x,u)\right] = 0, \tag{3.4}$$

with the continuous boundary condition $V^0(x) = 0$ on ∂G, where the generator $\mathcal{L}_u^0$ is defined in (3.2). See references [1] and [16] for a thorough account of the relationship between viscosity solutions of Hamilton–Jacobi PDEs and the value functions for various types of optimal control problems.

It turns out that V^0 is smooth on most of the domain G. Let Q be a subset of $\bar{G}$ which is open relative to G. We call Q an RSR if the following hold:

1. For each initial condition $x \in Q$, there is a unique optimal open loop control, and the corresponding trajectory $X_x^0(t)$ is contained in Q up to its exit time τ_x^0. The optimal trajectory meets ∂G nontangentially at a point z_x^0.
2. $V^0 \in C^\infty(Q)$.
3. There is a unique $u^0 \in C^\infty(Q)$ such that the optimal control can be represented in feedback form and is given by $u^0(x)$ for each $x \in Q$.
4. For $\xi \in Q \cap \partial G$, let T_ξ be the supremum of the exit times for optimal trajectories beginning in Q and exiting at ξ. The bijective map $\Xi(s,\xi)$ from

$$\{(s,\xi) : \xi \in Q \cap \partial G, 0 \le s < T_\xi\}$$

to Q given by $\Xi(\tau_x^0 - t, z_x^0) = X_x^0(t)$ is nonsingular in the sense that the quantities

$$\frac{\partial \Xi}{\partial s}(s,\xi), \frac{\partial \Xi}{\partial \nu_1}(s,\xi), \ldots, \frac{\partial \Xi}{\partial \nu_{n-1}}(s,\xi)$$

are linearly independent, where the ν_i are linearly independent tangent vectors to ∂G at the boundary point ξ.

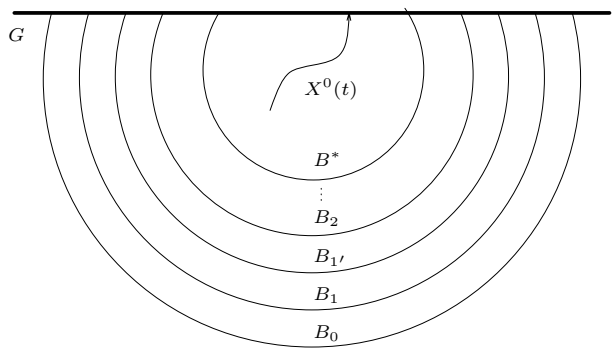


FIG. 1. *RSRs.*

The rather technical fourth condition in the above definition is the requirement that points in the RSRs not be conjugate. The uniqueness of optimal trajectories in the first condition is equivalent to the statement that the map $\Xi(s,\xi)$ is one-to-one. The nondegeneracy required in the fourth condition is a slightly stronger statement which precludes "almost" nonuniqueness in an infinitesimal sense. The classical method of characteristics and its application to proving the existence of RSRs for the present problem is discussed in the appendices of references [15, 17]. Detailed information on the structure of the regions of strong regularity for closely related problems can be found in references [4, 5, 14]. In general, the union of the RSRs is open and dense in the domain. We note that the C^∞ smoothness is a consequence of the assumed C^∞ smoothness of the data. Similar smoothness for finitely many derivatives would follow if the assumptions on the data were commensurately relaxed, and the asymptotic results in this paper would then be restricted to a finite order. For simplicity, we restrict ourselves to the case of C^∞ data.

Let B^* be a subset of $\bar{G}$ such that $\bar{B}^* \subset Q$. Our asymptotic results will hold on sets which include $\bar{B}^*$. Let $M \geq 1$ be a fixed integer. While our choice of M here is arbitrary, it corresponds to the order of the asymptotic expansion that we will obtain in section 5. Consider a region of strong regularity B_0, and, for $i = 1, \ldots, M$, regions of strong regularity B_i and $B_{i'}$ such that

$$\bar{B}^* \subset B_{M'} \subset \cdots \subset B_1 \subset \bar{B}_1 \subset B_0 \subset \bar{B}_0 \subset Q, \tag{3.5}$$

and such that

$$\bar{B}_{i+1} \subset B_{i'} \subset \bar{B}_{i'} \subset B_i \tag{3.6}$$

for each $i = 1, \ldots, M$, where B_{M+1} is taken to be B^* in (3.6). The existence of such nested RSRs follows from Theorem 2.4 in reference [17]. These relationships are illustrated in Figure 1.

Since V^0 is a classical solution to the DPE (3.4) on the RSR Q, the optimal feedback control can be explicitly evaluated there:

$$u^0(x) = -a(x)DV^0(x) + b(x). \tag{3.7}$$

We assume the following.

ASSUMPTION 3.2. *The boundary section $Q \cap \partial G$ is parallel to one of the coordinate hyperplanes. Furthermore, the minimum distance in the outward normal direction from $Q \cap \partial G$ to $\partial G/Q$ is equal to $\tilde{\delta} > 0$.*

We now present a first order Markov chain-based numerical approximation to the deterministic optimal control problem discussed above, along with known convergence results. The method of approximation by Markov chains was first described by Kushner [24], and an up-to-date treatment can be found in the book of Kushner and Dupuis [25]. Our approximation is essentially the one used in [3] and in [11]. In the sections which follow, we will present more detailed asymptotic results for this approximation and then describe a new numerical method which yields a qualitatively better rate of convergence.

Let $h > 0$ be a discretization parameter and define the discrete domain $G^h = h\mathbb{Z}^n \cap G$. For any $A \subset \mathbb{R}^n$, we define $A^h = h\mathbb{Z}^n \cap A^o$, where A^o is the interior of A. We consider limits as $h \to 0$, with the h chosen such that the hyperplane in which the boundary section $Q \cap \partial G$ lies lines up with the lattice $h\mathbb{Z}^n$. (See Assumption 3.2 and Figure 2.) We will construct a continuous time controlled jump Markov process on G^h which approximates the deterministic dynamics in (3.3).

Let $\underline{u}^h$ be any feedback control on G^h. Let $\underline{X}^h$ be the Markov process with controlled generator $\mathcal{L}_u^h$ given by

$$\mathcal{L}_u^h f = \langle u^+, D^{h,+} f \rangle - \langle u^-, D^{h,-} f \rangle$$

for any smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$. See section 1 for the notation in this definition. The corresponding stochastic dynamics will be called the h-dynamics. As in the description of the limit problem, we employ the underscore notation to indicate objects which are obtained from the application of an arbitrary possibly suboptimal feedback control.

Since we consider only feedback controls, it is straightforward to construct $\underline{X}^h$, as in section 4.3 of [25] and in [7]. We define a sequence of independently and identically distributed exponential random fields parameterized by u, with mean values specified as follows:

$$\overline{\Delta t}^h(u) = \begin{cases} \frac{h}{\|u\|_1} & u \neq 0, \\ h & u = 0. \end{cases} \tag{3.8}$$

Suppose that after $m-1$ jumps, $\underline{X}^h(s)$ is defined for $0 \leq s \leq t$ and that $\underline{X}^h(t) = x$. Then we take $\underline{X}^h(s) = x$ for all $t \leq s < t + \eta$, where the waiting time η is the exponential random variable obtained by evaluating the mth random field with the parameter value $u = \underline{u}^h(\underline{X}^h(s))$. If $u = 0$, then $\underline{X}^h(t + \eta) = x$, but otherwise it is conditionally distributed according to the jump probabilities

$$p^h(x, y|u) = \begin{cases} \frac{u_i^{\pm}}{\|u\|_1} & \text{if } y = x \pm he_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

It is easy to verify that the mean velocity of $\underline{X}^h$ at time t conditioned on $\underline{X}^h(t) = x$ is equal to $\underline{u}^h(x)$, so this is a consistent approximation to the limit dynamics in (3.3) if $\underline{u}^0(x)$ is replaced by $\underline{u}^h(x)$ there.

We now formulate the discrete approximation to the optimal control problem discussed above. Define the value function

$$V^h(x) = \inf_{\underline{u}^h} E_x \int_0^{\tau^h} L(\underline{X}^h(t), \underline{u}^h(\underline{X}^h(t)))dt,$$

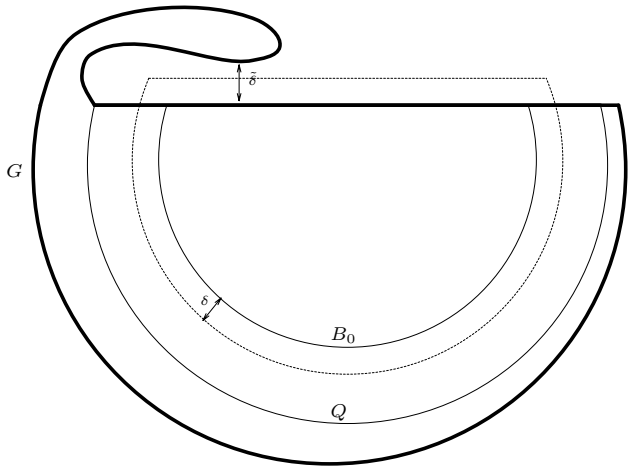


FIG. 2. *Region for smooth extension of* u^0.

where the exit time is $\underline{\tau}^h = \inf[t : \underline{X}^h(t) \notin G^h]$, and the infimum is over feedback controls $\underline{u}^h$. Using standard methods [25, section 4.3] it can be shown that V^h is the unique solution on G^h to the DPE

$$\inf_u \left[\mathcal{L}_u^h V^h(x) + L(x,u)\right] = 0, \tag{3.10}$$

with zero boundary condition on $h\mathbb{Z}^n/G^h$. It is straightforward to verify that (3.10) is equivalent to

$$V^h(x) = \inf_u \left[\sum_{y\in\mathbb{R}^n} p^h(x,y|u)V^h(y) + \overline{\Delta t}^h(u)L(x,u)\right] \tag{3.11}$$

and that the minimizing values of u are the same for these two equations. As suggested by the form of (3.11), the fixed point and an optimal feedback control can be found numerically using either Jacobi or Gauss–Seidel iteration schemes. We note that (3.11) is the DPE for a different approximating control problem, where, instead of the continuous time Markov chain described above, a discrete time Markov chain is used to approximate the deterministic dynamics. That is the approach taken in [3], where the time step $\overline{\Delta t}^h(u)$ is used to interpolate the Markov chain into continuous time. As discussed in [3], the choice of one-sided transition probabilities and of a control-dependent time step facilitates rapid convergence of the iterative schemes used to solve (3.11), and the required infima at each step can be evaluated analytically.

The DPE (3.10) gives rise to an approximate value function $V^h(x)$ on G^h as well as an approximate optimal feedback control $u^h(x)$ on G^h. Typical convergence results for numerical methods dealing with optimal control problems concern the convergence of the value functions $V^h(x)$ to the limit value $V^0(x)$. The following theorem is proved for the present problem in [3, Theorem 5.4].

THEOREM 3.3. *For any* $\varepsilon > 0$*, there exists* $h_0 > 0$ *such that*

$$|V^h(x) - V^0(x)| < \varepsilon$$

for all $0 < h \le h_0$ *and all* $x \in G^h$.

For applications it is often important to have a good numerical approximation to the optimal feedback control $u^0(x)$. Typically, this quantity is not uniquely defined on the entire domain, and until recently no general approximation theorems were available. The following theorem [11, Corollary 5.6] establishes convergence of the approximate optimal controls $u^h(x)$ to $u^0(x)$ on the RSRs for the present problem. Aside from its intrinsic value, this result plays a pivotal role in the detailed asymptotic analysis that we carry out in the this paper.

THEOREM 3.4. *For any $\varepsilon > 0$, there exists $h_0 > 0$ such that*

$$\|u^h(x) - u^0(x)\| < \varepsilon$$

for all $0 < h \le h_0$ and for all $x \in B_0^h$.

It is convenient to have $u^0(x)$ defined and Lipschitz on all of $\mathbb{R}^n$ and to have $u^h(x)$ defined on all of $h\mathbb{Z}^n$. We abuse notation by extending $u^0(x)$ to $\mathbb{R}^n$ and changing its values on the complement of $\bar{B}_0$. Let $\delta > 0$ be such that $\delta < d(B_0, \partial Q \cap G)$ and such that $\delta \le \bar{\delta}$, where $\bar{\delta}$ is as in Assumption 3.2; see Figure 2. We define a Lipschitz function $\tilde{u}^0(x)$ on $B_\delta(\bar{B}_0)$ by setting $\tilde{u}^0(x) = u^0(x)$ on $B_\delta(\bar{B}_0) \cap G$ and by extending it to $B_\delta(\bar{B}_0) \cap G^c$ as follows. For $x \in B_\delta(\bar{B}_0) \cap \partial G$ and for $0 \le \gamma \le \delta$, let $\tilde{u}^0(x + \gamma n) = u^0(x)$, where n is the outward normal vector at x; see Figure 2. Now let $\phi(x)$ be a C^∞ function on $\mathbb{R}^n$ taking values in $[0,1]$ such that $\phi(x) = 1$ on $B_{\delta/2}(\bar{B}_0)$ and $\phi(x) = 0$ outside of $B_\delta(\bar{B}_0)$. Such a function can be constructed by standard methods using a smooth convolution kernel [18, Theorem 0.17]. We can now redefine $u^0(x)$ to be equal to $\phi(x)\tilde{u}^0(x)$ on $B_\delta(\bar{B}_0)$ and zero everywhere else. This new $u^0(x)$ is Lipschitz on $\mathbb{R}^n$ and satisfies (3.7) on the region B_0. Finally, we put $u^h(x) = u^0(x)$ for all $x \in h\mathbb{Z}^n / B_0$. Notice that, by Lemma 3.1 and by Theorem 3.4, we have $u^0(x)$ and $u^h(x)$ contained in the compact set U for all x at which they are defined for $h > 0$ sufficiently small.

For most of what follows, we will be concerned only with initial conditions x in the RSR B_0. For $x \in B_0$, we define the optimal trajectory $X_x^0(t)$ for all $t \ge 0$ by applying the extended feedback control $u^0(x)$. Since B_0 is an RSR, $X_x^0(t)$ is optimally controlled until time τ_x^0, which is its first exit time from G and from the interior of B_0. We also define the exit location $z_x^0 = X_x^0(\tau_x^0)$. For $x \in B_0^h$, there is a unique process $X_x^h(t)$ defined for all $t \ge 0$ which is optimally controlled by $u^h(x)$ until it exits from B_0^h. We define the exit time $\tau_{x,B_0}^h = \inf[t : X_x^h(t) \notin B_0^h]$ and the exit location $z_{x,B_0}^h = X_x^h(\tau_{x,B_0}^h)$. We will often suppress the initial conditions in the subscripts of all of these notations.

The following lemma is proved just like [11, Lemma 2.3]. We use it to parlay information about convergence of trajectories into information about convergence of the corresponding exit times and locations.

LEMMA 3.5. *Let M_2,M_1 be RSRs such that $\bar{M}_1 \subset M_2 \subset B_0$. For each sufficiently small $\varepsilon > 0$, there exists $\eta > 0$ such that the following holds. Let X be a function of t that is continuous on the right with limits on the left and that has initial condition in M_2, and let τ_{M_2} and z_{M_2} be its exit time and location from the interior of M_2. If $x \in M_2$ is such that $\|\!| X - X_x^0 |\!\|_T \le \eta$ holds, then $\tau_{M_2} \le \tau_x^0 + \varepsilon$. If, in addition, $x \in M_1$, then it also follows that $|\tau_{M_2} - \tau_x^0| \le \varepsilon$ and $\|z_{M_2} - z_x^0\| \le \varepsilon$.*

Our proofs of the detailed asymptotic results in this paper depend upon having sharp estimates for the rate of convergence of the prelimit processes to the corresponding limit trajectories. We derive exponential rates of convergence in probability from the large deviations upper bound in Theorem B.1. For an RSR $M_2 \subset B_0$, consider a sequence of feedback control functions $\hat{u}^h(x)$ defined on $h\mathbb{Z}^n$ which satisfy

$\hat{u}^h(x) = u^0(x)$ for $x \notin M_2^h$. Suppose that, for each $\varepsilon > 0$, there exists $h_0 > 0$ such that

$$\|\hat{u}^h(x) - u^0(x)\| < \varepsilon$$

for all $0 < h \le h_0$ and for all $x \in M_2^h$. In particular, given Theorem 3.4, we can take for $\hat{u}^h(x)$ either $u^h(x)$ or $u^0(x)$. For $x \in M_2^h$, let $\hat{X}_x^h(t)$ be the process defined for $t \ge 0$ by applying the control $\hat{u}^h$ in the h-dynamics. Define $\hat{\tau}_{x,M_2}^h$ to be the exit time of $\hat{X}_x^h$ from M_2^h, and let $\hat{z}_{x,M_2}^h$ be its exit location.

LEMMA 3.6. *Let M_2, M_1 be RSRs such that $\bar{M}_1 \subset M_2 \subset B_0$, and let $\hat{u}^h(x)$ and its corresponding trajectories be as above. For any $\varepsilon > 0$, there exists $K > 0$ such that*

$$\text{(i)} \qquad P_x\left[\|\!| \hat{X}^h - X_x^0 |\!\|_T \ge \varepsilon\right] < \frac{1}{K} e^{-K/h},$$

$$\text{(ii)} \qquad P_x\left[\hat{\tau}_{M_2}^h > \tau_x^0 + \varepsilon\right] < \frac{1}{K} e^{-K/h}$$

holds for all $x \in M_2^h$ and for all sufficiently small $h > 0$. In addition,

$$\text{(iii)} \qquad P_x\left[|\hat{\tau}_{M_2}^h - \tau_x^0| \ge \varepsilon\right] < \frac{1}{K} e^{-K/h},$$

$$\text{(iv)} \qquad P_x\left[\|\hat{z}_{M_2}^h - z_x^0\| \ge \varepsilon\right] < \frac{1}{K} e^{-K/h}$$

holds for all $x \in M_1^h$ and for all sufficiently small $h > 0$.

Proof. Part (i) is obtained by applying the large deviations upper bound in Theorem B.1 with $n_1 = n, n_2 = 0$. In the lemma, we set $\bar{a}^h(x) = 0$, $\bar{a}(x) = 0$, $\bar{b}^h(x) = \hat{u}^h(x)$, and $\bar{b}(x) = u^0(x)$. We define the measure $\bar{\mu}^h(x)$ by

$$\bar{\mu}^h(x)(y) = \begin{cases} \hat{u}_i^{h,\pm}(x) & \text{if } y = \pm h e_i, \\ 0 & \text{otherwise,} \end{cases} \tag{3.12}$$

and we define the limit measure $\bar{\mu}(x)$ to put mass $\|u^0(x)\|_1$ at the origin. Finally, we take $F = \{\phi : \|\!| \phi - X_x^0 |\!\|_T \ge \varepsilon\}$, and we note that $I_x(\phi)$ is uniformly bounded away from zero for all $\phi \in F$. To get parts (ii)–(iv), pick $0 < \eta < \varepsilon$ to satisfy the conclusion of Lemma 3.5. Then pick $K > 0$ such that (i) holds with ε replaced by η, and parts (ii)–(iv) follow from Lemma 3.5. □

4. First term in the asymptotic expansion. In this section, we derive a first order asymptotic expansion for $V^h(x)$ around $V^0(x)$ in the RSR B_1. Our methodology is essentially that employed in [10], and Theorem 3.4 plays a key role in the proof. The first order expansion is of independent interest, as it establishes a rate of convergence for $V^h(x)$ as an approximation to $V^0(x)$ in the RSRs. Our primary interest in this paper, however, is in the fact that the first order expansion suggests that it is possible to obtain higher order convergence through a modification of the Markov chain approximation. In the sections which follow, we prove, under some additional assumptions, that such schemes can be implemented to yield second order convergence.

For each $x \in B_0$ and for each $u \in U$, we define the difference between the prelimit and the limit generators applied to $V^0(x)$ by

$$\begin{aligned} r^h(x,u) &= \mathcal{L}_u^h V^0(x) - \mathcal{L}_u^0 V^0(x), \\ &= \frac{1}{2}\, h\, \langle |u|, D_{\setminus}^2 V^0(x) \rangle + O(h^2), \end{aligned} \tag{4.1}$$

where the $O(h^2)$ term is uniform on $B_0 \times U$. We can now write the DPE (3.4) for $V^0(x)$ in the alternative form

$$\inf_u \left[\mathcal{L}_u^h V^0(x) + L(x,u) - r^h(x,u) \right] = 0. \tag{4.2}$$

Equation (4.2) is the DPE for an optimal control problem with Markov chain dynamics and with running cost equal to $L - r^h$. A comparison with the DPE for $V^h(x)$ in (3.10) suggests that the difference between $V^h(x)$ and $V^0(x)$ should be approximately equal to an integral of r^h along the optimal trajectory with initial condition x. The theorem below verifies that this intuition is correct, at least to a first order approximation. It is useful in what follows to define a compact notation for the running cost under a feedback control $u(x)$ by

$$L_u(x) = L(x, u(x)).$$

Similarly, we define

$$r_u^h(x) = r^h(x, u(x))$$

for all $x \in B_0$ and for all $h > 0$.

For $x \in B_0$, we define the error function

$$e^1(x) = \frac{1}{2} \int_0^{\tau_x^0} \langle \left| u^0(X_x^0(t)) \right|,\ D_{\setminus}^2 V^0(X_x^0(t)) \rangle dt, \tag{4.3}$$

and we note that the integrand is a scaled version of the first order approximation to r^h given in (4.1) above. Recall the region of strong regularity B_1 and the relationships specified in (3.5) and illustrated in Figure 1.

THEOREM 4.1. *The asymptotic expansion*

$$V^h(x) = V^0(x) + h e^1(x) + o(h) \tag{4.4}$$

holds as $h \to 0$, uniformly for $x \in B_1^h$.

Proof. We prove this theorem in two steps, first considering the upper bound on $V^h(x)$ and then the lower bound.

Upper bound. It is useful here to identify the suboptimal trajectories obtained by applying the limit optimal feedback control $u^0(x)$ in the h-dynamics. For an initial condition x in B_1^h, let $X^{h,0}$ be the process obtained by taking $\underline{u}^h = u^0$ in the Markov chain dynamics of section 3 with parameter h. Define the exit time $\tau_{B_0}^{h,0} = \inf[t : X^{h,0}(t) \notin B_0^h]$ and the exit location $z_{B_0}^{h,0} = X^{h,0}(\tau_{B_0}^{h,0})$. Since the infimum in (4.2) is achieved at $u^0(x)$, we can use a standard verification argument to establish for all $x \in B_1^h$ the representation

$$V^0(x) = E_x \left[\int_0^{\tau_{B_0}^{h,0}} L_{u^0}(X^{h,0}) - r_{u^0}^h(X^{h,0}) dt + V^0(z_{B_0}^{h,0}) \right]. \tag{4.5}$$

We use part (ii) of Lemma 3.6 and the strong Markov property to obtain the uniform integrability of the $\tau_{B_0}^{h,0}$ needed for the right-hand side of (4.5) to be finite. We obtain the following series of relations holding for x in B_1^h:

$$
\begin{aligned}
V^h(x) &\leq E_x\left[\int_0^{\tau_{B_0}^{h,0}} L_{u^0}(X^{h,0})dt + V^h(z_{B_0}^{h,0})\right] \\
&= V^0(x) + E_x\left[\int_0^{\tau_{B_0}^{h,0}} r_{u^0}^h(X^{h,0})dt\right] \\
&\quad + E_x\big[V^h(z_{B_0}^{h,0}) - V^0(z_{B_0}^{h,0})\big] \\
&= V^0(x) + \frac{1}{2}hE_x\left[\int_0^{\tau_{B_0}^{h,0}} \Big\langle |u^0(X^{h,0})|, D_{\setminus}^2 V^0(X^{h,0})\Big\rangle dt\right] \\
&\quad + E_x\big[V^h(z_{B_0}^{h,0}) - V^0(z_{B_0}^{h,0})\big] + O(h^2).
\end{aligned}
\tag{4.6}
$$

The first line follows from the definition of $V^h(x)$ and from the strong Markov property; the second line is a consequence of the representation in (4.5); and the third line is obtained from the estimate in (4.1), where the $O(h^2)$ term is uniform on B_1^h.

Using parts (i) and (iii) of Lemma 3.6, it is straightforward to see that the integral term in the last line of (4.6) is equal to $he^1(x) + o(h)$, uniformly for x in B_1^h. It remains to estimate the boundary term in the last line of (4.6). Let $\varepsilon > 0$ be equal to $d(B_1, \partial B_0 \cap G)$; see Figure 3 in the next section. Since $z_x^0 \in B_{q+1} \cap \partial G$, it follows from the fact that $h > 0$ is chosen so that the lattice $h\mathbb{Z}^n$ lines up with $Q \cap \partial G$ that if $\|z_{B_0}^{h,0} - z_x^0\| < \varepsilon$, then $z_{B_0}^{h,0} \in \partial G$, and in that case

$$V^h(z_{B_0}^{h,0}) = V^0(z_{B_0}^{h,0}) = 0.$$

Since $V^0(x)$ and $V^h(x)$ are uniformly bounded, part (iv) of Lemma 3.6 implies that the boundary term in the last line of (4.6) is equal to $O(e^{-K/h})$, uniformly for x in B_1^h. Combining these estimates, we have

$$V^h(x) \leq V^0(x) + he^1(x) + o(h),$$

holding uniformly for x in B_1^h.

Lower bound. Similarly to (4.6) in the proof of the upper bound, we obtain the

following series of relations holding for x in B_1^h:

$$
\begin{aligned}
V^0(x) &\leq E_x\left[\int_0^{\tau_{B_0}^h} L_{u^h}(X^h) - r_{u^h}^h(X^h)dt + V^0(z_{B_0}^h)\right] \\
&= V^h(x) - E_x\left[\int_0^{\tau_{B_0}^h} r_{u^h}^h(X^h)dt\right] \\
&\quad + E_x\left[V^0(z_{B_0}^h) - V^h(z_{B_0}^h)\right] \\
&= V^h(x) - \frac{1}{2}hE_x\left[\int_0^{\tau_{B_0}^h} \left\langle |u^h(X^h)|, D_{\setminus}^2 V^0(X^h)\right\rangle dt\right] \\
&\quad + E_x\left[V^0(z_{B_0}^h) - V^h(z_{B_0}^h)\right] + O(h^2).
\end{aligned} \tag{4.7}
$$

The first line is a consequence of the fact that $u^h(x)$ is suboptimal in the control problem corresponding to (4.2) and of the strong Markov property; the second line follows from the definition of $V^h(x)$; and the third line is obtained from the estimate in (4.1), where the $O(h^2)$ term is uniform on B_1^h.

As in the proof of the upper bound, we estimate the terms in the last line of (4.7) to obtain

$$V^0(x) \leq V^h(x) - he^1(x) + o(h),$$

holding uniformly for x in B_1^h. This time we need to use the convergence of u^h to u^0 from Theorem 3.4 in the application of Lemma 3.6 and then again to show that the integrand in the last line of (4.7) converges to the integrand in the definition of $e^1(x)$. □

Theorem 4.1 establishes the rate of convergence of $V^h(x)$ to $V^0(x)$ in the RSRs. It is not surprising that the convergence is first order, since the Markov chain approximation we consider gives rise to a discrete DPE (3.10) which could be obtained by replacing the derivatives in (3.4) with first order finite difference approximations. In fact, all numerical methods for problems of this type which have been proved to converge are intrinsically first order accurate.

What is promising about Theorem 4.1 is that (4.4) suggests that we can modify the present first order Markov chain method in order to obtain second order convergence in the RSRs. One approach is to approximate $e^1(x)$ and to subtract this approximation from $V^h(x)$. For this approach to be successful, two issues need to be addressed. First, we need to verify that the $o(h)$ error estimate in (4.4) can be sharpened to $O(h^2)$. Second, we need a first order approximation to $e^1(x)$. To get this we require first order estimates to the unknown quantities $u^0(x)$ and $D_{\setminus}^2 V^0(x)$. Both of these issues would be addressed if we could establish a higher order asymptotic expansion of the form

$$V^h(x) = V^0(x) + he^1(x) + h^2e^2(x) + O(h^3), \tag{4.8}$$

where both $e^1(x)$ and $e^2(x)$ are smooth functions of x. If (4.8) holds, then we can obtain first order approximations to the first and second derivatives of $V^0(x)$ by applying standard finite difference operators to $V^h(x)$. Since $u^0(x)$ is closely related

to $DV^0(x)$, this information can be used to obtain a first order approximation to $u^0(x)$. Another approach to obtaining second convergence is to apply Richardson extrapolation directly to the approximate value functions. If (4.8) is verified, then it is trivial that

$$2V^{h/2}(x) - V^h(x)$$

differs from $V^0(x)$ by an error which is $O(h^2)$, uniformly on the relevant RSR. In the sections which follow, we will specify conditions under which (4.8) can be verified, and then we will describe in detail how this information can be used to construct global numerical methods which provide second order approximation for both $V^0(x)$ and $u^0(x)$ in the RSRs.

5. Full asymptotic expansion. In this section, we establish an asymptotic expansion of $V^h(x)$ around $V^0(x)$ to order M, where M is an arbitrary positive integer. The results which follow require that we impose the following additional restriction on the region B_0.

ASSUMPTION 5.1. *There exists $\kappa > 0$ such that*

$$|u_i^0(x)| > \kappa$$

holds for each $i = 1, \ldots, n$ and for all $x \in B_0$. Since $u^0(x)$ is continuous on B_0, Assumption 5.1 implies that each component of the optimal control $u^0(x)$ has a fixed sign in the region B_0. Thus, without loss of generality, we can assume that there exists $\kappa > 0$ such that

$$u_i^0(x) > \kappa$$

for each $i = 1, \ldots, n$ and for all $x \in B_0$. Theorem 3.4 guarantees that, for $h > 0$ sufficiently small, we also have $u_i^h(x) > 0$ holding for each $i = 1, \ldots, n$ and for all $x \in B_0^h$. Thus, as long as we restrict the analysis which follows to sufficiently small $h > 0$, we can deal exclusively with forward differences in the numerical approximation. As such, we redefine the generator for the h-dynamics to be such that

$$\mathcal{L}_u^h f = \langle u, D^h f \rangle$$

for any smooth function f mapping $\mathbb{R}^n$ to $\mathbb{R}$. The operator D^h is taken to be $D^{h,+}$, the forward finite difference approximation to the gradient defined in section 1.

For convenience, we record here the DPEs for the limit problem and for the prelimit problem, with the generators fully written out. The limit DPE (3.4) takes the form

$$\begin{aligned} 0 &= \inf_u \left[\langle u, DV^0 \rangle + L \right] \\ &= -\frac{1}{2} \langle DV^0, aDV^0 \rangle + \langle b, DV^0 \rangle + c, \end{aligned} \tag{5.1}$$

and it holds for all $x \in B_0$, with the minimizer $u^0(x)$ given by (3.7). Similarly, the DPE (3.10) for the problem with h-dynamics can now be written in the form

$$\begin{aligned} 0 &= \inf_u \left[\langle u, D^h V^h \rangle + L \right] \\ &= -\frac{1}{2} \langle D^h V^h, aD^h V^h \rangle + \langle b, D^h V^h \rangle + c, \end{aligned} \tag{5.2}$$

where the minimizer $u^h(x)$ can be explicitly evaluated

$$u^h(x) = -a(x)D^hV^h(x) + b(x) \tag{5.3}$$

for all $x \in B_0^h$ and for all $h > 0$ sufficiently small.

In light of the discussion at the end of section 4, the condition in Assumption 5.1 is quite natural. In order for our approach to analyzing a second order numerical approximation to be successful, we require that both $e^1(x)$ and $e^2(x)$ in the asymptotic expansion (4.8) be smooth functions of x. In general, this fails to be true even for $e^1(x)$, owing to the $|u^0|$ term in the integrand of (4.3). Under the present assumptions, however, we can write $e^1(x)$ in the form

$$e^1(x) = \frac{1}{2}\int_0^{\tau_x^0} \langle u^0(X_x^0(t)),\ D_{\backslash}^2 V^0(X_x^0(t))\rangle dt, \tag{5.4}$$

which implies that $e^1(x)$ is a smooth function of x, and it satisfies the equation

$$\langle u^0, De^1\rangle + \frac{1}{2}\langle u^0, D_{\backslash}^2 V^0\rangle = 0, \tag{5.5}$$

for all x in B_0, with zero boundary condition on $\partial G \cap B_0$. The increased regularity can be accounted for by the fact that the Hamiltonian in the DPE (5.2) is smooth, whereas the Hamiltonian in the DPE (3.10) fails to be a continuously differentiable function of $D^{h,\pm}V^h(x)$. We remark that even in those regions where Assumption 5.1 fails to hold, our methods seem to work quite well; see the examples in section 7.

Before stating the main theorem of this section, we recall our standing assumptions. The domain $G \subset \mathbb{R}^n$ is open with compact closure and satisfies uniform interior and exterior cone conditions. We assume that b and c are C^∞ functions from $\mathbb{R}^n$ to $\mathbb{R}$ with $c(x) \geq c_0 > 0$ on G and that a is a C^∞ function from $\mathbb{R}^n$ to the space of symmetric positive definite $n \times n$ matrices. Additionally, we assume that we have RSRs as specified by (3.5)–(3.6) such that the flat boundary condition of Assumption 3.2 and the condition on the optimal controls given in Assumption 5.1 both hold.

It is useful for what follows to introduce some notation. For $h > 0$, we define a new approximate feedback control

$$\bar{u}^h(x) = \frac{1}{2}(u^0(x) + u^h(x)). \tag{5.6}$$

Combining (5.1) and (5.2), and noting that

$$\bar{u}^h(x) = -\frac{1}{2}a(x)D^hV^h(x) - \frac{1}{2}DV^0(x) + b(x) \tag{5.7}$$

holds for all x in B_0^h, we conclude that

$$\langle \bar{u}^h, D^h(V^h - V^0)\rangle + \langle \bar{u}^h, D^hV^0 - DV^0\rangle = 0 \tag{5.8}$$

holds on the region B_0^h.

THEOREM 5.2. *Recall the region B^* from* (3.5). *With $M \geq 1$ the arbitrary constant chosen in section* 3 *and with the $e^m(x)$ as given below, the asymptotic expansion*

$$V^h(x) = V^0(x) + \sum_{m=1}^{M} h^m e^m(x) + o(h^M) \tag{5.9}$$

holds as $h \to 0$, *uniformly for* $x \in (B^*)^h$.

The following corollary establishes an analogous asymptotic expansion for the optimal feedback control. It is a direct consequence of Theorem 5.2 and of the expressions for $u^0(x)$ and $u^h(x)$ given in (3.7) and (5.3).

COROLLARY 5.3. *With* $M \geq 1$ *the arbitrary constant chosen in section* 3 *and with the* $e^m(x)$ *as given below, the asymptotic expansion*

$$\begin{aligned} u^h(x) &= u^0(x) - a(x) \sum_{m=1}^{M-1} h^m D^h e^m(x) \\ &\quad - a(x) \sum_{m=1}^{M-1} h^m \frac{1}{(m+1)!} D_{\backslash}^{m+1} V^0(x) + o(h^{M-1}) \end{aligned} \tag{5.10}$$

holds as $h \to 0$, *uniformly for* $x \in (B^*)^h$.

In section 4, we were able to determine the value of $e^1(x)$ by considering the formal difference between the generators $\mathcal{L}_u^0$ and $\mathcal{L}_u^h$, and then through the new dynamic programming equation (4.2), considering $V^0(x)$ as the value function for a modified optimal control problem with the h-dynamics. We cannot use this approach to determine the form of $e^m(x)$ for $m \geq 2$ because, in addition to the higher order terms in the expansion of $r^h(x)$, there are approximation effects which result from the difference between the controls $u^h(x)$ and $u^0(x)$. Instead, we use a formal recursive procedure, illustrated below for the case $m = 2$.

Assume that the expansion in (5.10) holds to first order, so it follows that

$$\bar{u}^h(x) = u^0(x) - \frac{1}{2} h a(x) D e^1(x) - \frac{1}{4} h a(x) D_{\backslash}^2 V^0(x) + o(h). \tag{5.11}$$

We now combine (5.5) and (5.8) to obtain

$$\begin{aligned} &\langle \bar{u}^h, h^{-2} D^h [V^h - V^0 - h e^1] \rangle \\ &\quad + h^{-2} \langle \bar{u}^h, D^h V^0 - D V^0 \rangle - \frac{1}{2} h^{-1} \langle u^0, D_{\backslash}^2 V^0 \rangle \\ &\quad + h^{-1} \langle \bar{u}^h - u^0, D^h e^1 \rangle + h^{-1} \langle u^0, D^h e^1 - D e^1 \rangle = 0. \end{aligned}$$

Using the expression in (5.11), along with the fact that Taylor's theorem implies

$$D^h f(x) = \sum_{m=1}^{M} \frac{1}{m!} h^{m-1} D_{\backslash}^m f(x) + o(h^M) \tag{5.12}$$

for all smooth functions $f(x)$, we conclude that

$$\langle \bar{u}^h, h^{-2} D^h [V^h - V^0 - h e^1] \rangle + r^2 + o(1) = 0, \tag{5.13}$$

where we define $r^2(x)$ by

$$r^2(x) = \frac{1}{2}\langle u^0(x), D_\backslash^2 e^1(x)\rangle + \frac{1}{6}\langle u^0(x), D_\backslash^3 e^0(x)\rangle$$
$$- \frac{1}{2}\langle De^1(x), a(x)D_\backslash^2 e^0(x)\rangle - \frac{1}{2}\langle De^1(x), a(x)De^1(x)\rangle$$
$$- \frac{1}{8}\langle D_\backslash^2 e^0(x), a(x)D_\backslash^2 e^0(x)\rangle.$$

Heuristically, (5.13) suggests that we should set

$$e^2(x) = \frac{1}{2}\int_0^{\tau_x^0} r^2(X_x^0(t))dt.$$

The formal argument outlined above will be made rigorous when we prove Theorem 5.2 later in this section.

In order to record the general form of $e^m(x)$ for $m = 1, \ldots, M$, we adopt the convention $e^0(x) = V^0(x)$ and define

$$r^m(x) = \sum_{k=2}^{m+1} \frac{1}{k!}\langle u^0(x)\ ,\ D_\backslash^k e^{m+1-k}(x)\rangle$$
$$- \frac{1}{2}\sum_{l=2}^{m}\sum_{k=1}^{m+2-l} \frac{1}{k!l!}\langle D_\backslash^l e^0(x)\ ,\ a(x)D_\backslash^k e^{m+2-l-k}(x)\rangle$$
$$- \frac{1}{2}\sum_{j=1}^{m-1}\sum_{l=1}^{m-j}\sum_{k=1}^{m+2-j-l} \frac{1}{k!l!}\langle D_\backslash^l e^j(x)\ ,\ a(x)D_\backslash^k e^{m+2-j-l-k}(x)\rangle$$

for all x in B_0. Then recursively applying the heuristic outlined above suggests that we should define $e^m(x)$ by

$$e^m(x) = \frac{1}{2}\int_0^{\tau_x^0} r^m(X_x^0(t))dt \tag{5.14}$$

for each $m = 1, \ldots, M$ and for all x in B_0. By the method of characteristics, we then have that

$$\langle u^0, De^m\rangle + r^m = 0 \tag{5.15}$$

holds for each $m = 1, \ldots, M$ and for all x in B_0, with zero boundary conditions on $\partial G \cap B_0$. Since each $r^m(x)$ depends upon the $e^i(x)$ with $i < m$, we define the $e^m(x)$ recursively beginning with $m = 1$. We note that the conclusion of Lemma 5.7 serves to verify that we have properly defined the $e^m(x)$ for the problem at hand.

An alternative heuristic method to derive (5.15) is to assume that (5.9) holds and that the error term $o(h^M)$ is a smooth function of x. We apply the operator D^h to both sides of (5.9) and then use (5.12) to express $D^h V^h(x)$ in terms of derivatives of smooth functions. We substitute this expression into the DPE (5.2), formally differentiate the resulting equation m times with respect to the parameter h, and then set $h = 0$ to obtain the equation for $e^m(x)$.

The proof of Theorem 5.2 can be broken down into three lemmas to be applied recursively. To make the notation consistent, we define a new region $B_{0'}$ to be equal to B_0. See equations (3.5) and (3.6) along with Figure 1 for the relationships between the RSRs in the following lemmas. Given an asymptotic expansion for $D^h V^h(x)$ to q terms, the first lemma establishes the expansion for $V^h(x)$ to $q+1$ terms. The needed expansion for $D^h V^h(x)$ with $q = 0$ has been established in Theorem 3.4. Although it is not needed for the development, we also note that in section 4 we have already proved Lemma 5.4 for the case $q = 0$.

LEMMA 5.4. *Let $0 \le q \le M-1$ be an integer, and suppose that*

$$D^h V^h(x) = D^h V^0(x) + \sum_{m=1}^{q} h^m D^h e^m(x) + o(h^q) \tag{5.16}$$

holds as $h \to 0$, uniformly for x in $B^h_{q'}$. Then

$$V^h(x) = V^0(x) + \sum_{m=1}^{q+1} h^m e^m(x) + o(h^{q+1}) \tag{5.17}$$

holds as $h \to 0$, uniformly for x in B^h_{q+1}.

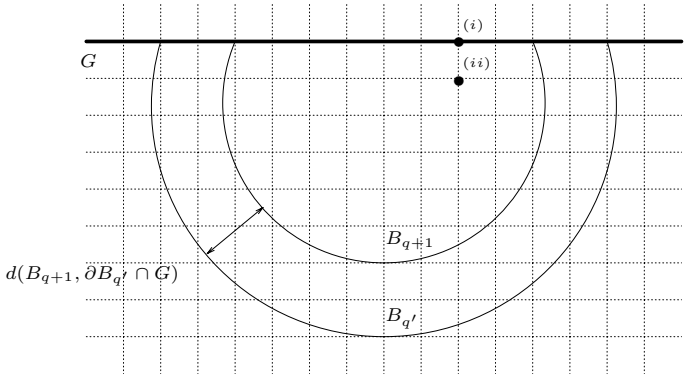


FIG. 3. *Boundary points.*

The next lemma establishes the expansion for $D^h V^h(x)$ to $q+1$ terms but only in a neighborhood of the boundary. The two cases in the lemma are illustrated in Figure 3.

LEMMA 5.5. *Let $0 \le q \le M-1$ be an integer, and suppose that* (5.16) *holds. Then, for each $p = 1, \ldots, n$,*

$$D^h_p V^h(x) = D^h_p V^0(x) + \sum_{m=1}^{q+1} h^m D^h_p e^m(x) + o(h^{q+1}) \tag{5.18}$$

holds as $h \to 0$, uniformly for $x \in \mathbb{R}^n$ such that either (i) *$x \in \partial G$ and $x + he_p \in B^h_{q+1}$ or* (ii) *$x \in B^h_{q+1}$ and $x + he_p \in \partial G$.*

Finally, given the conclusion of Lemma 5.5, the third lemma carries the asymptotic expansion for $D^h V^h(x)$ to $q+1$ terms on a smaller RSR.

LEMMA 5.6. *Let $0 \le q \le M-1$ be an integer, and suppose that* (5.16) *and* (5.18) *hold. Then*

$$D^h V^h(x) = D^h V^0(x) + \sum_{m=1}^{q+1} h^m D^h e^m(x) + o(h^{q+1}) \tag{5.19}$$

holds as $h \to 0$, uniformly for x in $B^h_{(q+1)'}$.

Given the nesting of the RSRs, these three lemmas complete one step of a recursive process. We can apply Lemma 5.4 with q replaced by $q+1$ and iterate until we have established the full asymptotic series to order M. In preparation for the proofs of these lemmas, we introduce some more notation. For $q = 1, \dots, M$ and for $x \in B_0$, we define

$$\Phi^{h,q}(x) = \frac{1}{h^q}\left[V^h(x) - V^0(x) - \sum_{m=1}^{q} h^m e^m(x)\right],$$

and, for each $p = 1, \dots, n$,

$$\begin{aligned}\Psi^{h,q}_p(x) &= D^h_p \Phi^{h,q}(x) \\ &= \frac{1}{h^q}\left[D^h_p V^h(x) - D^h_p V^0(x) - \sum_{m=1}^{q} h^m D^h_p e^m(x)\right],\end{aligned}$$

and then we put $\Psi^{h,q}(x) = (\Psi^{h,q}_1(x), \dots, \Psi^{h,q}_n(x))$. The conclusions in the above lemmas can now be formulated in terms of the convergence as $h \to 0$ of the $\Phi^{h,q+1}(x)$ and the $\Psi^{h,q+1}(x)$ to zero.

The proofs of Lemmas 5.4–5.6 involve some elementary but rather lengthy algebraic calculations. We summarize the results of these calculations in the following lemma, the proof of which is deferred to Appendix A. It should be emphasized that, while the proof of Lemma 5.7 is elementary, it is in fact the key step which confirms that we have properly defined the $e^m(x)$ for the particular numerical approximation that we are studying. We recall our convention that subscripts refer to the components of a vector, while superscripts act as indices for possibly vector valued quantities.

LEMMA 5.7. *Let $0 \le q \le M-1$ be an integer, and suppose that* (5.16) *holds. Then*

$$\langle \bar{u}^h, D^h \Phi^{h,q+1} \rangle + \phi^{h,q+1} = 0, \tag{5.20}$$

where $\phi^{h,q+1} = o(1)$ as $h \to 0$, uniformly for x in $B^h_{q'}$. Furthermore, for each $p = 1, \dots, n$

$$\langle \hat{u}^{h,q+1,p}, D^h \Psi^{h,q+1}_p \rangle + \langle \lambda^{h,q+1,p}, \Psi^{h,q+1} \rangle + \psi^{h,q+1}_p = 0, \tag{5.21}$$

where $\hat{u}^{h,q+1,p} = u^0 + o(1)$, $\lambda^{h,q+1,p} = D_p u^0 + o(1)$, and $\psi^{h,q+1}_p = o(1)$ as $h \to 0$, uniformly for x in $B^h_{q'}$.

Proof of Lemma 5.4. The conclusion in (5.17) is equivalent to the statement that $\Phi^{h,q+1}(x)$ converges to zero as $h \to 0$, uniformly for x in B^h_{q+1}. For such x, let $\bar{X}^h_x(t)$ be the process defined for $t \ge 0$ by applying the control $\bar{u}^h$ in the h-dynamics. Define $\bar{\tau}^h_{x,B_{q'}}$ to be the exit time of $\bar{X}^h_x$ from $B_{q'}$, and let $\bar{z}^h_{x,B_{q'}}$ be its exit location. In light of (5.20), a standard verification argument shows that $\Phi^{h,q+1}(x)$ satisfies

$$\Phi^{h,q+1}(x) = E_x\left[\int_0^{\bar{\tau}^h_{B_{q'}}} \phi^{h,q+1}(\bar{X}^h)dt + \Phi^{h,q+1}(\bar{z}^h_{B_{q'}})\right] \tag{5.22}$$

for each x in B^h_{q+1}. Part (ii) of Lemma 3.6 and the strong Markov property imply that the $\bar{\tau}^h_{B_{q'}}$ are uniformly integrable, which guarantees that the right-hand side in the above expression is finite. We use that $\phi^{h,q+1}(x)$ converges to zero uniformly on $B_{q'}$ and that the $E_x[\bar{\tau}^h_{B_{q'}}]$ are uniformly bounded, along with the fact that $\Phi^{h,q+1}(x)$ is equal to zero on ∂G, to conclude that

$$\Phi^{h,q+1}(x) = o(1) + P_x\big[\bar{z}^h_{B_{q'}} \notin \partial G\big] O(h^{-(q+1)})$$

holds uniformly for x in B^h_{q+1}. Let $\varepsilon > 0$ be equal to $d(B_{q+1}, \partial B_{q'} \cap G)$; see Figure 3. Since $z^0_x \in B_{q+1} \cap \partial G$, it follows from the fact that $h > 0$ is chosen so that the lattice $h\mathbb{Z}^n$ lines up with $Q \cap \partial G$ that if $\|\bar{z}^h_{x,B_{q'}} - z^0_x\| < \varepsilon$, then $\bar{z}^h_{x,B_{q'}} \in \partial G$. Thus, we can apply the exponential bound from part (iv) of Lemma 3.6 to conclude that $\Phi^{h,q+1}(x)$ converges to zero as $h \to 0$, uniformly for x in B^h_{q+1}. That completes the proof of the lemma. □

Proof of Lemma 5.5. For simplicity, we fix p and treat only case (ii), where $x \in B^h_{(q+1)'}$ and $x + he_p \in \partial G$. The proof for case (i) differs only in notation. Since V^h, V^0, and all of the e^m satisfy a zero boundary condition on ∂G, we have for each x satisfying condition (ii)

$$\begin{aligned}
&\frac{1}{h^{q+1}} \left| D^h_p V^h(x) - D^h_p V^0(x) - \sum_{m=1}^{q+1} h^m D^h_p e^m(x) \right| \\
&\qquad = \frac{1}{h^{q+1}} \left[\frac{1}{h} |V^h(x) - V^0(x) - \sum_{m=1}^{q+1} h^m e^m(x)| \right] \\
&\qquad = \frac{1}{h} |\Phi^{h,q+1}(x)|.
\end{aligned}$$

Thus, it suffices to prove that $\frac{1}{h}|\Phi^{h,q+1}(x)|$ converges to zero, uniformly for x satisfying condition (ii). Equation (5.22) implies that the relation

$$\frac{1}{h}|\Phi^{h,q+1}(x)| = \frac{1}{h} E_x\big[\bar{\tau}^h_{B_{q'}}\big] o(1) + P_x\big[\bar{z}^h_{B_{q'}} \notin \partial G\big] O(h^{-(q+2)}) \tag{5.23}$$

holds uniformly for x satisfying condition (ii). As in the proof of Lemma 5.4, the exponential bound from part (iv) of Lemma 3.6 implies that the second term in the right-hand side of (5.23) converges to zero as $h \to 0$, uniformly for x satisfying condition (ii). Thus, it remains to show that $\frac{1}{h} E_x[\bar{\tau}^h_{B_{q'}}]$ is uniformly bounded so that the first term in the right-hand side of (5.23) also converges to zero. We observe that

$$E_x\big[\bar{\tau}^h_{B_{q'}}\big] \le E_x\big[\bar{\tau}^h_{B_{q'}}; A_1\big] + E_x\big[\bar{\tau}^h_{B_{q'}}; A_2\big] + E_x\big[\bar{\tau}^h_{B_{q'}}; A_3\big],$$

where we define the events

$$A_1 = \big[\bar{X}^h(t) + he_p \in \partial G \ , \ 0 \le t \le \bar{\tau}^h_{B_{q'}}\big],$$

$$A_2 = \big[\bar{\tau}^h_{B_{q'}} > T\big],$$

$$A_3 = \big[\bar{\tau}^h_{B_{q'}} \le T \ , \ \exists\, t \in [0, \bar{\tau}^h_{B_{q'}}) \text{ such that } \bar{X}^h(t) + he_p \notin \partial G\big].$$

Given the nontangential exit property for the limit optimal trajectories and the convergence of $\bar{u}^h$ to u^0, there exists $\eta > 0$ such that the following holds. For sufficiently small $h > 0$, if $x \in B^h_{q+1}$ and if $x + he_p \in \partial G$, the probability of exiting from G in one jump is at least equal to η; see Figure 3. Furthermore, since $\|\bar{u}^h\|_1$ is bounded away from zero, the conditional expected value of each interjump waiting time is bounded by δh for some fixed $\delta > 0$. Thus,

$$E_x\left[\bar{\tau}^h_{B_{q'}}; A_1\right] \leq \delta h \eta \left(\sum_{i=0}^{\infty} i(1-\eta)^{i-1}\right) P_x\left[A_1\right] = O(h),$$

so $\frac{1}{h} E_x[\bar{\tau}^h_{B_{q'}}; A_1]$ is uniformly bounded for all x satisfying condition (ii). It follows from part (ii) of Lemma 3.6 and from the strong Markov property that $E_x[\bar{\tau}^h_{B_{q'}} | A_2]$ is uniformly bounded. Again applying part (ii) of Lemma 3.6, we obtain the uniform bound

$$P_x[A_2] = O(e^{-K/h}) \tag{5.24}$$

so that $\frac{1}{h} E_x[\bar{\tau}^h_{B_{q'}}; A_2]$ converges to zero uniformly for all x satisfying condition (ii). Since $E_x[\bar{\tau}^h_{B_{q'}} | A_3]$ is clearly bounded by T, we can complete the proof by obtaining a bound similar to the last display for the probability of the event A_3. On account of the fact that we use one-sided transition probabilities and given the nontangential exit property for the limit optimal trajectories, it follows from the fact that the boundary segment $B_{q'} \cap \partial G$ is parallel to one of the coordinate hyperplanes that $\bar{X}^h$ with initial condition x satisfying condition (ii) must move parallel to the boundary segment $B_{q'} \cap \partial G$ up to the time when it exits from $B_{q'}$; see Assumption 3.2 and Figure 3. Consequently, event A_2 can occur only if the exit location $\bar{z}^h_{B_{q'}}$ is not in ∂G. Thus, since $z^0_x \in B_{q+1} \cap G$, the fact that $d(B_{q+1}, \partial B_{q'} \cap G) > 0$ implies that we can obtain a bound for $P_x[A_3]$ analogous to (5.24) by applying part (iv) of Lemma 3.6. □

Proof of Lemma 5.6. The conclusion in (5.19) is equivalent to the statement that $\Psi^{h,q+1}(x)$ converges to zero as $h \to 0$, uniformly for x in $B^h_{(q+1)'}$. Our plan is to proceed as in the proof of Lemma 5.4, first obtaining a representation for $\Psi^{h,q+1}(x)$ analogous to the representation for $\Phi^{h,q+1}(x)$ given in (5.22) and then using that representation to show that $\Psi^{h,q+1}(x)$ converges to zero. In this case, however, we cannot use an entirely standard representation because $\Psi^{h,q+1}(x)$ is a vector quantity, and the equations for its components are coupled. We develop the representation fairly carefully in order to highlight what we believe to be a novel aspect of our approach that may prove useful in developing high order numerical methods based directly upon the use of higher order finite difference approximations; see the discussion in section 1. First, along the lines suggested in [19, section 5.4], we expand the state space so that $(x, p) \in \mathbb{R}^n \times \{1, \ldots, n\}$ is the state variable, rather than just x. To that end, we recall the quantities defined in Lemma 5.7 and abuse notation by defining

$$\Psi^{h,q+1}(x,p) = \Psi^{h,q+1}_p(x), \qquad \psi^{h,q+1}(x,p) = \psi^{h,q+1,p}(x),$$

$$\hat{u}^{h,q+1}(x,p) = \hat{u}^{h,q+1,p}(x), \qquad \lambda^{h,q+1}(x,p) = \lambda^{h,q+1,p}(x),$$

for each (x, p), and we note that $\Psi^{h,q+1}$ and $\psi^{h,q+1}$ are scalar valued, while $\hat{u}^{h,q+1}$ and $\lambda^{h,q+1}$ are vector valued. We can now regard p as an argument rather as an index

in (5.21) and obtain the scalar equation

$$\begin{aligned}
&\langle \hat{u}^{h,q+1}(x,p), D^h \Psi^{h,q+1}(x,p) \rangle \\
&+ \sum_{i=1}^{n} \lambda_i^{h,q+1}(x,p) \left[\Psi^{h,q+1}(x,i) - \Psi^{h,q+1}(x,p) \right] \\
&+ \sum_{i=1}^{n} \lambda_i^{h,q+1}(x,p) \Psi^{h,q+1}(x,p) + \psi^{h,q+1}(x,p) = 0,
\end{aligned} \tag{5.25}$$

holding for all $x \in B_{q+1}^h$ and for all $p = 1, \ldots, n$. If the $\lambda_i^{h,q+1}(x,p)$ were known to be nonnegative, it would be possible to regard them as probabilities of jumping from (x,p) to (x,i) and then to deduce a representation for $\Psi^{h,q+1}(x,p)$ in terms of a Markov chain directly from (5.25). Since the signs of the $\lambda_i^{h,q+1}(x,p)$ are indefinite, however, we must further expand the state space in a novel way to eliminate the possibility of negative transition probabilities. We consider state variable $(x,p,\sigma) \in \mathbb{R}^n \times \{1,\ldots,n\} \times \{-1,1\}$ and again abuse notation by defining

$$\Psi^{h,q+1}(x,p,\sigma) = \sigma \Psi^{h,q+1}(x,p), \qquad \psi^{h,q+1}(x,p,\sigma) = \sigma \psi^{h,q+1}(x,p),$$

$$\lambda^{h,q+1}(x,p,\sigma) = \lambda^{h,q+1}(x,p)$$

for each (x,p,σ). Notice that $\lambda^{h,q+1}(x,p,\sigma)$ is actually independent of σ; we introduce the dependence only for convenience. In light of this notation, the linear structure of (5.25) implies that we can write

$$\begin{aligned}
&\langle \hat{u}^{h,q+1}(x,p), D^h \Psi^{h,q+1}(x,p,\sigma) \rangle \\
&+ \sum_{i=1}^{n} \left(\lambda_i^{h,q+1}(x,p,\sigma) \right)^+ \left[\Psi^{h,q+1}(x,i,\sigma) - \Psi^{h,q+1}(x,p,\sigma) \right] \\
&+ \sum_{i=1}^{n} \left(\lambda_i^{h,q+1}(x,p,\sigma) \right)^- \left[\Psi^{h,q+1}(x,i,-\sigma) - \Psi^{h,q+1}(x,p,\sigma) \right] \\
&+ \| \lambda^{h,q+1}(x,p,\sigma) \|_1 \Psi^{h,q+1}(x,p,\sigma) + \psi^{h,q+1}(x,p,\sigma) = 0
\end{aligned} \tag{5.26}$$

for all $x \in B_{q+1}^h$, $p = 1, \ldots, n$, and $\sigma \in \{-1,1\}$. The left-hand side of (5.26) indicates the action on $\Psi^{h,q+1}(x)$ of a legitimate generator with positive transition probabilities, so we can use it to derive a representation for $\Psi^{h,q+1}(x)$ in terms of a Markov chain $\Xi^{h,q+1}(t)$ taking values in $\mathbb{R}^n \times \{1,\ldots,n\} \times \{-1,1\}$. We construct the Markov chain $\Xi^{h,q+1}$ with exponentially distributed waiting times with mean

$$\overline{\Delta t}^{h,q+1}(x,p,\sigma) = \frac{h}{\Gamma^{h,q+1}(x,p,\sigma)},$$

where

$$\Gamma^{h,q+1}(x,p,\sigma) = \| \hat{u}^{h,q+1}(x,p) \|_1 + h \| \lambda^{h,q+1}(x,p,\sigma) \|_1,$$

and with transition probabilities given by

$$p^{h,q+1}(x,p,\sigma;\xi) = \begin{cases} \frac{\hat{u}_i^{h,q+1}(x,p)}{\Gamma^{h,q+1}(x,p,\sigma)}, & \xi = (x+he_i,p,\sigma), \\ h\frac{\left(\lambda_i^{h,q+1}(x,p,\sigma)\right)^+}{\Gamma^{h,q+1}(x,p,\sigma)}, & \xi = (x,i,\sigma), \\ h\frac{\left(\lambda_i^{h,q+1}(x,p,\sigma)\right)^-}{\Gamma^{h,q+1}(x,p,\sigma)}, & \xi = (x,i,-\sigma), \\ 0 & \text{otherwise.} \end{cases}$$

For details on the construction of such a process, see the discussion in section 3. Let $X^{h,q+1}$ be the first component of $\Xi^{h,q+1}$. For initial conditions (x,p,σ) such that x is in $B_{(q+1)'}$, let $\tau^{h,q+1}_{B_{q+1}}$ be the first exit time of $X^{h,q+1}$ from B^h_{q+1}, and let $z^{h,q+1}_{B_{q+1}}$ be equal to its exit location. Although the full Markov chain $\Xi^{h,q+1}$ does not converge to a deterministic limit, Theorem B.1 implies that a large deviations upper bound holds for the component process $X^{h,q+1}$. Combined with Lemma 3.5, the large deviations upper bound implies the following analogue of Lemma 3.6. For any $\varepsilon > 0$, there exists $K > 0$ such that

$$\text{(i)} \qquad P_{x,p,\sigma}\left[\|\!| X^{h,q+1} - X^0_x |\!\|_T \geq \varepsilon\right] < \frac{1}{K}e^{-K/h},$$

$$\text{(ii)} \qquad P_{x,p,\sigma}\left[\tau^{h,q+1}_{B_{q+1}} > \tau^0_x + \varepsilon\right] < \frac{1}{K}e^{-K/h}$$

holds for all initial conditions such that x is in B_{q+1}, while

$$\text{(iii)} \qquad P_{x,p,\sigma}\left[|\tau^{h,q+1}_{B_{q+1}} - \tau^0_x| \geq \varepsilon\right] < \frac{1}{K}e^{-K/h},$$

$$\text{(iv)} \qquad P_{x,p,\sigma}\left[\|z^{h,q+1}_{B_{q+1}} - z^0_x\| \geq \varepsilon\right] < \frac{1}{K}e^{-K/h}$$

holds for all initial conditions such that x is in $B_{(q+1)'}$. The representation for $\Psi^{h,q+1}(x)$ now takes the form

$$\begin{aligned} &\Psi^{h,q+1}(x,p,\sigma) \\ &\quad = \left[E_{x,p,\sigma}\int_0^{\tau^{h,q+1}_{B_{q+1}}} e^{\int_0^t \|\lambda^{h,q+1}(\Xi^{h,q+1}(s))\|_1 ds}\psi^{h,q+1}(\Xi^{h,q+1}(t))dt\right. \\ &\qquad \left. + \ \Psi^{h,q+1}(\Xi^{h,q+1}(\tau^{h,q+1}_{B_{q+1}}))\right] \end{aligned} \tag{5.27}$$

for all (x,p,σ) such that x is in $B^h_{(q+1)'}$. Property (ii) above and the strong Markov property guarantee that

$$P_{x,p,\sigma}\left[\tau^{h,q+1}_{B_{q+1}} \geq kT\right] \leq \left(\frac{1}{K}\right)^k e^{-kK/h}$$

holds for all initial conditions such that x is in $B^h_{(q+1)'}$ where $h > 0$ is sufficiently small and for all integers $k \geq 1$. Since the $\lambda^{h,q+1}$ are uniformly bounded, this is sufficient to guarantee that the right-hand side of (5.27) is finite for sufficiently small $h > 0$ so that a standard verification argument can be used to establish that the indicated representation holds. In fact, since by Lemma 5.7 the $\psi^{h,q+1}(x,p,\sigma)$ converge uniformly to zero, this reasoning also implies that the first term on the right-hand side converges to zero as $h \to 0$. Thus, we can complete the proof of the lemma by showing that the second term also converges to zero. Since $\Psi^{h,q+1}(x,p,\sigma)$ is equal to zero whenever $x \in \partial G$, the second term in (5.27) is bounded by

$$P_{x,p,\sigma}\big[z^{h,q+1}_{B_{q+1}} \notin \partial G\big] O(h^{-(q+2)}).$$

Let $\varepsilon > 0$ be equal to $d(B_{(q+1)'}, \partial B_{q+1} \cap G)$. As in the proof of Lemma 5.4, Assumption 3.2 implies that if $\|z^{h,q+1}_{B_{q+1}} - z^0_x\| < \varepsilon$, then $z^{h,q+1}_{B_{q+1}} \in \partial G$. Thus, we can apply the exponential bound from property (iv) above to conclude that the last display converges to zero as $h \to 0$, uniformly for all (x,p,σ) such that x is in B_{q+1}. This completes the proof that $\Psi(x,p,\sigma)$ converges uniformly to zero and so establishes the lemma. □

6. Second order numerical approximations. The numerical method described in section 3 yields approximations to the value function and to the optimal control which are first order accurate as $h \to 0$ on the RSRs. This notion is made precise in Theorem 4.1 and in the stronger results in Theorem 5.2 and Corollary 5.3 which hold under Assumption 5.1. In this section, we exploit the detailed asymptotic information from section 5 to propose two different second order numerical methods. We prove in Theorem 6.1 that each of the modified numerical methods produces approximations to $V^0(x)$ and to $u^0(x)$ on the region B^* which are second order convergent as $h \to 0$. While these theoretical results apply only on RSRs where each component of the limit optimal control is bounded away from zero, in practice we do not know the locations of such regions. As such, our methods are of necessity applied on the entire domain. As the numerical examples in the next section illustrate, we obtain second order convergence even in large parts of the domain where the theoretical results do not apply.

Approximation Method I. For the purpose of defining a practical algorithm on the entire domain, we abuse notation by reverting to the optimal feedback control function $u^h(x)$ which is obtained on all of G^h by taking the maximizing argument in the DPE (3.10) before it is redefined on $h\mathbb{Z}^n/B_0$ in section 3. Additionally, we consider versions of X^h, τ^h, and z^h which are defined using that control up to the exit time from G^h. When we prove the convergence properties on B^*, we will again employ the modified control and trajectories which were defined at the end of section 3 and which were used in the analysis up to now.

The first step in the present algorithm is to obtain $V^h(x)$ and $u^h(x)$ by applying the first order method from section 3. By Theorem 5.2, we have the asymptotic result

$$V^h(x) = V^0(x) + he^1(x) + h^2e^2(x) + O(h^3), \tag{6.1}$$

holding for all $x \in (B^*)^h$, where the $e^m(x)$ are as defined in (5.14). For each x in G^h, we define an approximation to $e^1(x)$ by

$$e^{1,h}(x) = \frac{1}{2}E_x\left[\int_0^{\tau^h} \langle |u^h(X^h)|, D^{2,h}_{\setminus} V^h(X^h)\rangle dt\right],$$

and we note that $e^{1,h}(x)$ can be obtained by applying an iterative method to solve the linear equation

$$\langle u^{h,+}, D^{h,+}e^{1,h}\rangle - \langle u^{h,-}, D^{h,-}e^{1,h}\rangle + \frac{1}{2}\langle |u^h|, D^{2,h}_{\backslash} V^h\rangle = 0, \tag{6.2}$$

with zero boundary condition on the complement of G^h. To implement an iterative solver, we express (6.2) in the form

$$e^{1,h}(x) = \sum_{y\in\mathbb{R}^n} p^h(x,y|u^h(x))e^{1,h}(y) + \overline{\Delta t}^h(u^h(x))\frac{1}{2}\langle |u^h(x)|, D^{2,h}_{\backslash} V^h(x)\rangle, \tag{6.3}$$

which is analogous to the expression for the discrete DPE given in (3.11). Since the value function $V^h(x)$ is finite and since the running cost is bounded away from zero, every state in the optimally controlled Markov chain corresponding to (3.11) and (6.3) must communicate with the boundary. Thus, the right-hand side is a contraction, and the equation is guaranteed to have a unique fixed point. We now define a new approximation to $V^0(x)$ by subtracting a correction term from $V^h(x)$,

$$V^{h,*}(x) = V^h(x) - he^{1,h}(x), \tag{6.4}$$

and a new approximation to the optimal control by

$$u^{h,*}(x) = -a(x)D^{h,c}V^{h,*}(x) + b(x), \tag{6.5}$$

for all x in G^h, where $D^{h,c}$ is the centered difference approximation to the gradient operator. We will prove that $e^{1,h}(x)$ is a first order approximation to $e^1(x)$ on B^*, so it will follow directly from (6.1) that $V^{h,*}(x)$ is a second order approximation to $V^0(x)$. Additionally, a more detailed analysis of the asymptotic properties of $e^{1,h}(x)$ will be used to establish that $u^{h,*}(x)$ is a second order approximation to $u^0(x)$ on the region B^*.

We recall that the rate of convergence results that we will prove in Theorem 6.1 apply only on RSRs where each component of the limit optimal feedback control is bounded away from zero. A potential drawback of the method just described is that it may degrade the convergence properties of $V^h(x)$ and of $u^h(x)$ on those parts of the domain where second order convergence is not guaranteed. Our numerical experiments suggest that this is largely a theoretical concern, but nonetheless we mention here a possible approach to avoiding the problem. If we knew that the $e^{1,h}(x)$ were uniformly bounded, then it would follow from the definition of $V^{h,*}(x)$ that it has convergence properties at least as good as those proved for $V^h(x)$. That is, in addition to second order convergence on the RSRs which satisfy Assumption 5.1, the $V^{h,*}(x)$ would exhibit uniform convergence to $V^0(x)$ on the entire domain and first order convergence on the RSRs. In general, the $e^{1,h}(x)$ need not be uniformly bounded, but a sensible approach to guaranteeing this type of robustness would be to limit the norm of $D^{2,h}_{\backslash} V^h(x)$ in the definition of the $e^{1,h}(x)$. Similarly, if we could uniformly bound the $D^{h,c}e^{1,h}(x)$, then it would also follow that the $u^{h,*}(x)$ converge uniformly to the optimal feedback control on those parts of the RSRs where second order convergence is not assured. In any case, our numerical experiments have been done without any type of limiter, and we have not observed significant difficulties of this type.

Approximation Method II. Another approach to obtaining second order convergence on the RSRs is to apply Richardson extrapolation to the approximations

$V^h(x)$ and $u^h(x)$ obtained from the first order method in section 3. As in the description of Method I, we abuse notation by considering the optimal feedback control function $u^h(x)$ which is obtained on all of G^h by taking the maximizing argument in the DPE (3.10). For x in the grid $2h\mathbb{Z}^n$, we define a new approximation to the value function by

$$V^{h,*}(x) = 2V^h(x) - V^{2h}(x)$$

and a new approximate optimal feedback control by

$$u^{h,*}(x) = 2u^h(x) - u^{2h}(x).$$

It is a trivial consequence of Theorem 5.2 and of Corollary 5.3 that these are second order approximations to $V^0(x)$ and $u^0(x)$ on the region B^*.

An advantage of Richardson extrapolation over Method I is that there is no concern that the $V^{h,*}(x)$ or the $u^{h,*}(x)$ will exhibit qualitatively poorer convergence than their first order counterparts on those parts of the domain where Theorem 6.1 does not apply. This is a consequence of the fact that the new approximations are obtained by taking linear combinations of the first order approximations $V^h(x)$ and $u^h(x)$. This does not seem to be a practical problem even for Method I, and Method I has the advantages of producing somewhat smaller errors in our experiments and of being more straightforward to code because it involves only a single grid. Still, in circumstances where it is vital to maintain first order convergence outside of the regions where second order convergence is guaranteed, the present approach may be preferred.

THEOREM 6.1. *Let $V^{h,*}(x)$ and $u^{h,*}(x)$ be obtained by either Method* I *or* II *above. Then the estimates*

$$V^{h,*}(x) = V^0(x) + O(h^2)$$

and

$$u^{h,*}(x) = u^0(x) + O(h^2)$$

hold uniformly for all $x \in (B^*)^h$.

For the purposes of proving this theorem, we return to using the feedback control $u^h(x)$ and the corresponding trajectories which were defined at the end of section 3 and which were used for the analysis in the previous sections. Recall that $u^h(x)$ is defined for all x in $h\mathbb{Z}^n$ but is optimal only on the region B_0^h. Furthermore, Assumption 5.1 implies that each component of $u^h(x)$ is positive on the region B_0^h.

Proof of Theorem 6.1. As noted above, the conclusion of the theorem for the case of Method II is a trivial consequence of Theorem 5.2 and of Corollary 5.3. Thus, we proceed to prove the result for the case where $V^{h,*}(x)$ and $u^{h,*}(x)$ are obtained by Method I. We define the function

$$\Theta^h(x) = \frac{1}{h}\left[e^1(x) - e^{1,h}(x)\right]$$

for all x in $(B_0)^h$. Then it follows from (6.1) that

$$V^{h,*}(x) = V^0(x) + h^2\Theta^h(x) + h^2 e^2(x) + O(h^3), \tag{6.6}$$

so the estimate for $V^{h,*}(x)$ will be verified if we show that the $\Theta^h(x)$ are bounded uniformly on $(B^*)^h$. Define the function $\theta^h(x)$ by

$$\begin{aligned}\theta^h &= \left\langle \frac{1}{h}(u^0 - u^h), D^h e^1 \right\rangle + \left\langle u^0, \frac{1}{h}(De^1 - D^h e^1) \right\rangle \\ &\quad + \frac{1}{2}\left\langle \frac{1}{h}(u^0 - u^h), D^{2,h} V^h \right\rangle + \frac{1}{2}\left\langle u^0, \frac{1}{h}(D^2 V^0 - D^{2,h} V^h) \right\rangle.\end{aligned}$$

Given the fact that all of the components of $u^h(x)$ are assumed to be positive, (5.5) and (6.2) imply that $\Theta^h(x)$ satisfies

$$\langle u^h, D^h \Theta^h \rangle + \theta^h = 0, \tag{6.7}$$

for all x in $(B^*)^h$, with a zero boundary condition on $\partial G \cap (B^*)^h$. Thus, we have the representation

$$\Theta^h(x) = E_x \left[\int_0^{\tau^h_{B^*}} \theta^h(X^h) dt + \Theta^h(z^h_{B^*}) \right]. \tag{6.8}$$

The asymptotic expansions in Theorem 5.2 and in Corollary 5.3 imply that the $\theta^h(x)$ are uniformly bounded on $(B^*)^h$, so part (ii) of Lemma 3.6 implies that the integral term in (6.8) is bounded. Since the $\Theta^h(x)$ are uniformly bounded by $O(h^{-2})$ on B_0^h, we can apply the exponential estimate from part (iv) of Lemma 3.6 to bound the last term in (6.8). See the proof of Lemma 5.4, where details are given for a similar argument. We have shown that the $\Theta^h(x)$ are uniformly bounded on $(B^*)^h$, and this completes the proof of the first part of the theorem.

In light of the formula for $u^0(x)$ in (3.7) and the definition of $u^{h,*}(x)$, it is sufficient for the second part of the theorem to show that $D^{h,c} V^{h,*}(x)$ is a second order approximation to $D^{h,c} V^0(x)$, which is in turn a second order approximation to $DV^0(x)$. Since we do not know that $\Theta^h(x)$ satisfies a Lipschitz-type bound, we cannot derive such an estimate directly from (6.6). Instead, we will derive a first order asymptotic expansion of the form

$$\Theta^h(x) = \Theta(x) + O(h), \tag{6.9}$$

where $\Theta(x)$ is a smooth function. We can then substitute this expression into (6.6) and apply the second order centered difference operator $D^{h,c}$ to both sides of that equation in order to verify that $D^{h,c} V^{h,*}(x)$ is a second order approximation to $D^{h,c} V^0(x)$.

We define the function

$$\Theta(x) = \int_0^{\tau^0_x} \theta(X^0_x) dt,$$

for all x in B_0, where $\theta(x)$ is given by

$$\begin{aligned}\theta &= \left\langle \frac{1}{2} a D^2_{\backslash} V^0 + aDe^1, De^1 \right\rangle + \left\langle u^0, -\frac{1}{2} D^2_{\backslash} e^1 \right\rangle \\ &\quad + \frac{1}{2}\left\langle \frac{1}{2} a D^2_{\backslash} V^0 + aDe^1, D^2 V^0 \right\rangle + \frac{1}{2}\left\langle u^0, -\frac{1}{2} D^2 e^1 \right\rangle.\end{aligned}$$

Thus, $\Theta(x)$ satisfies the equation

$$\langle u^0, D\Theta \rangle + \theta = 0, \tag{6.10}$$

for all x in B^*, with a zero boundary condition on $\partial G \cap B^*$. In light of Theorem 5.2 and Corollary 5.3, a term by term comparison of the definitions of $\theta^h(x)$ and $\theta(x)$ reveals that

$$\theta^h(x) = \theta(x) + O(h),$$

uniformly for x in $(B^*)^h$. We can now combine (6.7) and (6.10) to obtain the relation

$$\langle u^h, D^h[\Theta^h - \Theta]\rangle + O(h) = 0,$$

holding for all x in $(B^*)^h$, with a zero boundary condition on $\partial G \cap (B^*)^h$. Just as in the proof of the first part of the theorem, we can derive a representation for the quantity $[\Theta^h(x) - \Theta(x)]$ and use the exponential estimates from Lemma 3.6 to establish that it is equal to $O(h)$. This implies that (6.9) holds uniformly on $(B^*)^h$ and completes the proof of the theorem. □

7. Examples. In this section, we present the results of experiments which illustrate the rates of convergence for the numerical methods described in this paper. For each example, we compute solutions using the first order method of section 3 and then using the two second order methods described in section 6. We solve the DPE (3.10), and for the second order Method I a linear version of that equation, by Gauss–Seidel iteration with a tolerance of 10^{-8}. Error values are indicated in the L^1 and L^∞ norms on the entire domain and on regions where the solution is known to be smooth. Approximate rates of convergence are determined by taking $\log_2(E_{m/2+1}/E_{m+1})$, where E_k is the error obtained using k gridpoints. In the case of the results obtained by Method II using Richardson extrapolation, the indicated number of gridpoints corresponds to the more refined of the two grids used in the calculation. In general, we observe second order or near second order convergence of the value functions in the L^1 norm on the entire domain and in the L^1 and L^∞ norms on the RSRs. In the case of the optimal controls, we generally observe second order convergence only in the L^1 norm on the RSRs. Results of this type are to be expected, since Assumption 5.1 is typically violated on some parts of the RSRs, so the theoretical rate of convergence results in Theorem 6.1 which would guarantee second order convergence in the L^∞ norm do not apply everywhere. In addition to the error values, we indicate the total number of iterative steps required to obtain each solution. As expected with Gauss–Seidel iteration, the number of iterations is essentially independent of the number of gridpoints. This last observation is very important in terms of the efficiency of practical calculations.

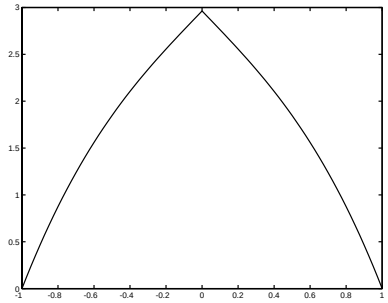
Example 1. One-dimensional problem. We begin with a one-dimensional example on the domain $[-1, 1]$. The running cost is taken to be

$$L(x, u) = (2 + 3x^2)^2 + \frac{1}{4}u^2,$$

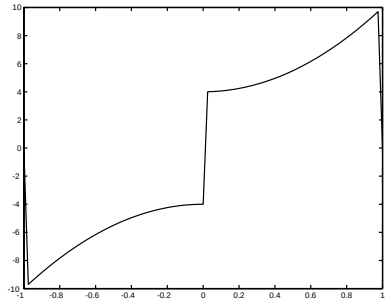
and we can analytically evaluate the solution

$$V(x) = \begin{cases} 2x + x^3 + 3, & x \le 0, \\ -2x - x^3 + 3, & x > 0. \end{cases}$$

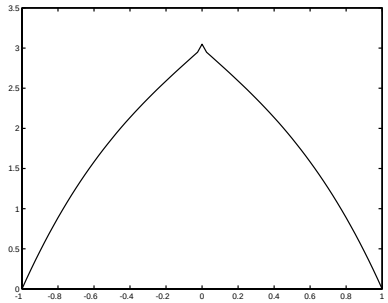
Approximations to the value function and to the optimal control are shown in Figure 4. The value function is approximated quite well by the first order method and by both second order methods, with only a slight overshoot appearing at the singularity when



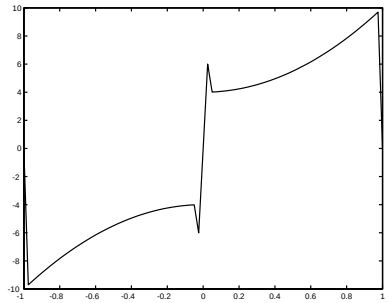
(a) Value Function
First Order



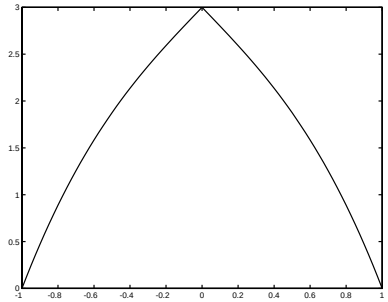
(b) Feedback Control
First Order



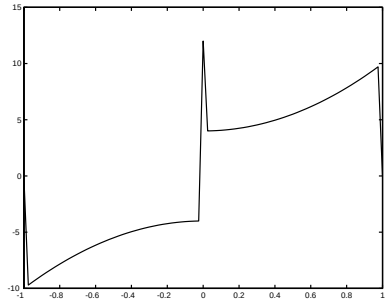
(c) Value Function
Method I



(d) Feedback Control
Method I



(e) Value Function
Method II



(f) Feedback Control
Method II

FIG. 4. *One-dimensional problem solutions.*

we use Method I. This is an apparent manifestation of our observation in section 6 that it is possible for this method to result in the loss of convergence at points of discontinuity. The discontinuity in the optimal control is resolved quite sharply, although sizable overshoots occur with both second order methods. We note, however, that these overshoots appear to remain bounded as the grid is refined, so they do not indicate a lack of numerical stability.

In Table 1, we indicate the errors in computing the value function using the three numerical schemes considered in this paper. We discuss the results for the optimal feedback control in the following paragraph. Errors are given for the entire domain and for those points which are at least a distance of 0.1 away from the singularity. Using the first order scheme, we obtain clear first order convergence of the value function in both the L^1 and the L^∞ norms on the entire domain and on the RSRs. Both second order schemes result in second order convergence of the value function in the L^1 norm on the entire domain as well as in the L^1 and L^∞ norms on the RSRs. Results of this quality are expected for a one-dimensional problem, since the optimal control cannot change sign within an RSR, and Theorem 6.1 is valid on every such region. Somewhat surprisingly, Method II also yields second order convergence of the value function in the L^∞ norm on the entire domain.

The situation with the optimal feedback control is unusual due to the fact that the problem is one-dimensional. Using the first order scheme, the approximate feedback controls on the RSRs are exactly (to machine accuracy) equal to the optimal feedback control for the limit problem. The following simple observation clarifies why this behavior is to be expected. Equations (5.1) and (5.2) imply for each x in the RSRs that

$$\frac{1}{2}\langle DV^0(x), a(x)DV^0(x)\rangle + \langle b(x), DV^0(x)\rangle + c(x) = 0$$

and

$$\frac{1}{2}\langle D^hV^h(x), a(x)DV^h(x)\rangle + \langle b(x), DV^h(x)\rangle + c(x) = 0.$$

In one dimension, these quadratic equations imply that there are only two possible values for $DV^0(x)$ and $D^hV^h(x)$. Since Theorem 3.4 guarantees that the $D^hV^h(x)$ converge to $DV^0(x)$, it must be that they are equal for sufficiently small $h > 0$. Given the relationships between the optimal feedback controls and $DV^0(x)$ and $D^hV^h(x)$, it follows that the approximate optimal feedback controls should be exactly correct. This observation applies only in one dimension, and our two-dimensional examples exhibit more typical convergence of the optimal feedback controls. For the present problem, Method II inherits the exact convergence of the optimal feedback controls, since it involves taking linear combinations of results obtained using the first order method, while Method I produces second order convergence in the L^1 and L^∞ norms in the RSRs.

Example 2. Perturbed escape time. Our next example is a two-dimensional problem which is obtained by perturbing the escape time problem on the unit square. The escape time problem has the simple running cost $\tilde{L}(x,u) = \frac{1}{2}\|u\|^2 + 1/2$, and its value function is known analytically to be given by the shortest distance to any of the four edges of the square. Thus, the complement of the diagonals $x_1 = x_2$ and $x_1 = -x_2$ is a maximal RSR, and the value function is linear in each connected component of that region. As in [11], we modify the data for this problem to obtain

TABLE 1
One-dimensional problem value function errors.

			L^1		L^∞		L^1 RSR		L^∞ RSR	
	Pts	Iter	Error	Ord	Error	Ord	Error	Ord	Error	Ord
1st order	21	4	1.95 e − 01	–	1.45 e − 01	–	1.66 e − 01	–	1.44 e − 01	–
	41	4	9.87 e − 02	1.0	7.38 e − 02	1.0	8.40 e − 02	1.0	7.31 e − 02	1.0
	81	4	4.97 e − 02	1.0	3.72 e − 02	1.0	4.23 e − 02	1.0	3.68 e − 02	1.0
	161	4	2.49 e − 02	1.0	1.87 e − 02	1.0	2.12 e − 02	1.0	1.85 e − 02	1.0
	321	4	1.25 e − 02	1.0	9.36 e − 03	1.0	1.06 e − 02	1.0	9.26 e − 03	1.0
	641	4	6.25 e − 03	1.0	4.68 e − 03	1.0	5.31 e − 03	1.0	4.64 e − 03	1.0
2nd order I	21	8	4.02 e − 02	–	1.77 e − 01	–	2.03 e − 02	–	2.25 e − 02	–
	41	8	1.06 e − 02	1.9	9.39 e − 02	0.9	5.06 e − 03	2.0	5.63 e − 03	2.0
	81	8	2.73 e − 03	2.0	4.85 e − 02	1.0	1.27 e − 03	2.0	1.41 e − 03	2.0
	161	8	6.93 e − 04	2.0	2.46 e − 02	1.0	3.16 e − 04	2.0	3.52 e − 04	2.0
	321	8	1.75 e − 04	2.0	1.24 e − 02	1.0	7.91 e − 05	2.0	8.79 e − 05	2.0
	641	8	4.38 e − 05	2.0	6.23 e − 03	1.0	1.98 e − 05	2.0	2.20 e − 05	2.0
2nd order II	21	8	1.00 e − 02	–	1.00 e − 02	–	8.00 e − 03	–	8.00 e − 03	–
	41	8	2.50 e − 03	2.0	2.50 e − 03	2.0	2.03 e − 03	2.0	2.25 e − 03	1.8
	81	8	6.25 e − 04	2.0	6.25 e − 04	2.0	5.06 e − 04	2.0	5.63 e − 04	2.0
	161	8	1.56 e − 04	2.0	1.56 e − 04	2.0	1.27 e − 04	2.0	1.40 e − 04	2.0
	321	8	3.91 e − 05	2.0	3.91 e − 05	2.0	3.16 e − 05	2.0	3.52 e − 05	2.0
	641	8	9.77 e − 06	2.0	9.77 e − 06	2.0	7.91 e − 06	2.0	8.79 e − 06	2.0

one with smooth data and a more interesting solution that can still be determined analytically. To that end, we introduce the C^∞ double bump function defined by

$$\chi(\xi) = \begin{cases} e^{-\lambda((\xi-m)^2-\sigma^2)^{-2}}, & \xi \in [m-\sigma, m+\sigma], \\ e^{-\lambda((-\xi-m)^2-\sigma^2)^{-2}}, & \xi \in [-m-\sigma, -m+\sigma], \\ 0 & \text{otherwise}, \end{cases}$$

where we use the parameter values $m = 0.7$, $\sigma = 0.5$, and $\lambda = 0.07$. Now we define a mollifier by

$$\Phi(x) = \chi(x_1 + x_2)\chi(x_1 - x_2)$$

for all $x = (x_1, x_2)$ in the unit square and then define the value function $V^0(x)$ by multiplying the value function for the escape time problem by $1+\Phi(x)$. The resulting function has the same RSRs as the escape time problem, and it maintains the simple structure in a neighborhood of the singularities. In a similar spirit, we define

$$a(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 3\sin(2\pi x_1)^2 \begin{bmatrix} 2 & 5 \\ 5 & 18 \end{bmatrix} \Phi(x)$$

and

$$b(x) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} x_1 \\ x_2 \sin((x_1^2 + x_2^2)^{1/2} - 1/2) \end{bmatrix} \Phi(x)$$

so that $a(x)$ is the identity and $b(x)$ is the zero vector in a neighborhood of the singularities. Now we define $c(x)$ on the RSRs by

$$c(x) = (1/2)\langle DV^0(x), a(x)DV^0(x)\rangle - \langle b(x), DV^0(x)\rangle.$$

Our use of a mollifier in defining all of the relevant functions ensures that the cost function $c(x)$ extends smoothly to $c(x) = 1/2$ at the singularities, and it turns out that $V^0(x)$ solves the limit control problem for the indicated cost structure.

In Table 2, we indicate the errors in computing the value function and the optimal controls using the three numerical schemes considered in this paper. Errors are given for the entire domain and for those regular points which are at least a distance of 0.1 away from the singularities. The first order method consistently produces first order convergence of the value function and of the optimal controls, except when the control errors are measured in the L^∞ norm on the entire domain. This is expected, since the convergence results for the optimal controls do not apply at the singular points. The qualitative behaviors of the two second order methods are essentially the same, although the absolute error values for the feedback control are somewhat smaller using Method I. We obtain second order convergence of the value functions in the L^1 norm on the entire domain, while first order convergence is maintained in the L^∞ norm. In the RSRs, convergence in the L^1 norm is second order, and convergence in the L^∞ norm is significantly better than first order but perhaps not second order. We note that the lack of clear second order convergence in the L^∞ norm is to be expected, since Assumption 5.1 does not hold everywhere in the RSRs for this problem.

In the case of the optimal controls, first order convergence is maintained in the L^1 norm on the entire domain, and we observe second order convergence in the L^1 norm in the RSRs. The errors on the RSRs measured in the L^∞ norm appear to decay only

TABLE 2
Perturbed escape time problem errors.

			L^1		L^∞		L^1 RSR		L^∞ RSR	
	Pts	Iter	Error	Ord	Error	Ord	Error	Ord	Error	Ord
Value Function, 1st order	21	12	2.14 e − 02	–	3.80 e − 02	–	6.34 e − 03	–	1.90 e − 02	–
	41	12	7.46 e − 03	1.5	1.95 e − 02	1.0	3.27 e − 03	1.0	1.02 e − 02	0.9
	81	13	2.90 e − 03	1.4	1.01 e − 02	0.9	1.69 e − 03	1.0	5.28 e − 03	1.0
	161	16	1.24 e − 03	1.2	5.19 e − 03	1.0	8.65 e − 04	1.0	2.72 e − 03	1.0
	321	16	5.69 e − 04	1.1	2.64 e − 03	1.0	4.40 e − 04	1.0	1.38 e − 03	1.0
	641	16	2.71 e − 04	1.1	1.33 e − 03	1.0	2.21 e − 04	1.0	6.91 e − 04	1.0
Value Function, 2nd order I	21	24	2.71 e − 02	–	6.58 e − 02	–	7.93 e − 03	–	1.18 e − 02	–
	41	28	7.28 e − 03	1.9	3.31 e − 02	1.0	1.55 e − 03	2.4	3.16 e − 03	1.9
	81	36	1.92 e − 03	1.9	1.68 e − 02	1.0	4.04 e − 04	1.9	9.31 e − 04	1.8
	161	39	4.96 e − 04	2.0	8.52 e − 03	1.0	1.05 e − 04	1.9	2.63 e − 04	1.8
	321	35	1.26 e − 04	2.0	4.27 e − 03	1.0	2.68 e − 05	2.0	9.93 e − 05	1.4
	641	35	3.18 e − 05	2.0	2.12 e − 03	1.0	6.77 e − 06	2.0	2.79 e − 05	1.8
Value Function, 2nd order II	21	20	1.66 e − 02	–	1.76 e − 02	–	1.40 e − 02	–	1.76 e − 02	–
	41	24	3.80 e − 03	2.1	5.23 e − 03	1.8	8.04 e − 04	4.1	2.85 e − 03	2.6
	81	25	1.17 e − 03	1.7	2.79 e − 03	0.9	2.99 e − 04	1.4	1.01 e − 03	1.5
	161	29	3.27 e − 04	1.8	1.40 e − 03	1.0	9.08 e − 05	1.7	3.51 e − 04	1.5
	321	32	8.67 e − 05	1.9	8.01 e − 04	0.8	2.48 e − 05	1.9	1.06 e − 04	1.7
	641	32	2.24 e − 05	2.0	4.07 e − 04	1.0	6.61 e − 06	1.9	3.42 e − 05	1.6

Table 2 (CONT.).

Feedback Control

			L^1		L^∞		L^1 RSR		L^∞ RSR	
	Pts	Iter	Error	Ord	Error	Ord	Error	Ord	Error	Ord
1st order	21	12	5.94 e − 01	–	1.86 e − 00	–	1.03 e − 01	–	2.66 e − 01	–
	41	12	3.24 e − 01	0.9	7.99 e − 01	0	5.39 e − 02	0.9	1.59 e − 01	0.7
	81	13	1.73 e − 01	0.9	1.86 e − 00	0	3.09 e − 02	0.8	8.94 e − 02	0.8
	161	16	8.94 e − 02	1.0	8.08 e − 01	0	1.67 e − 02	0.9	4.84 e − 02	0.9
	321	16	4.56 e − 02	1.0	8.13 e − 01	0	8.73 e − 03	0.9	2.54 e − 02	0.9
	641	16	2.30 e − 02	1.0	1.87 e − 00	0	4.45 e − 03	1.0	1.31 e − 02	1.0
2nd order I	21	24	4.89 e − 01	–	1.00 e − 00	–	1.73 e − 01	–	2.46 e − 01	–
	41	28	2.50 e − 01	1.0	1.00 e − 00	0	3.43 e − 02	2.3	9.58 e − 02	1.4
	81	36	1.24 e − 01	1.0	1.00 e − 00	0	8.38 e − 03	2.0	2.44 e − 02	2.0
	161	39	6.11 e − 02	1.0	1.00 e − 00	0	2.42 e − 03	1.8	1.22 e − 02	1.0
	321	35	3.03 e − 02	1.0	1.00 e − 00	0	6.86 e − 04	1.8	6.99 e − 03	0.8
	641	35	1.50 e − 02	1.0	1.00 e − 00	0	1.79 e − 04	1.9	3.85 e − 03	0.9
2nd order II	21	20	9.88 e − 01	–	3.79 e − 00	–	3.46 e − 01	–	3.38 e − 01	–
	41	24	5.07 e − 01	0.9	2.37 e − 00	0	5.72 e − 02	2.6	2.46 e − 02	3.8
	81	25	2.67 e − 01	1.0	3.80 e − 00	0	1.72 e − 02	1.7	1.20 e − 01	1.0
	161	29	1.29 e − 01	1.0	1.39 e − 00	0	5.12 e − 03	1.7	2.59 e − 02	2.2
	321	32	6.43 e − 02	1.0	8.09 e − 01	0	1.49 e − 03	1.8	3.03 e − 02	0
	641	32	3.21 e − 02	1.0	3.30 e − 00	0	4.09 e − 04	1.9	1.26 e − 02	1.3

at a first order rate. This can be accounted for by the fact that the present problem does not satisfy Assumption 5.1 on the maximal RSR, as the optimal controls for the limit problem are not of fixed sign. By considering subsets of the RSRs where the controls do not change signs, we can obtain more convincing second order convergence in the L^∞ norm for both the value function and the optimal control. The results in Table 3 are obtained by restricting our attention to those parts of the domain where $x > 0.5$ and $y > 0$ and where $x < -0.5$ and $y < 0$. This is a somewhat arbitrary choice, but it singles out a set of points where the sign changes do not seem to interfere with second order convergence.

In Figure 5, we display the exact values of $V^0(x)$ and of the first component of $u^0(x)$, as well as approximations obtained by using 41 points in the two second order numerical methods. The graphs of the feedback controls illustrate that our methods resolve discontinuities quite sharply. It is also worth noting the slight overshoot in the center of the value function obtained by Method I. We observed in section 6 that it is theoretically possible for convergence of the value function at singular points to be compromised when this second order method is employed. The overshoot is the only apparent manifestation of that possibility, and as predicted by the theory it does not appear when we compute the approximations using Method II.

Example 3. Quadratic running cost. Our final example is a two-dimensional problem on the unit square where the boundaries of the maximal RSR are curved. The running cost is defined to be $L(x,u) = \frac{1}{2}\|u\|^2 + 6x_1^2 + 1$. We do not have an analytic expression for the solution, but, by looking at approximate solutions, it is easy to visually identify the boundaries of the RSRs. In Figure 6(a), we indicate those boundaries, as well as the subset of the domain which we utilize for the purpose of calculating rates of convergence on the RSRs, and in Figure 6(b) we show approximations to the characteristics.

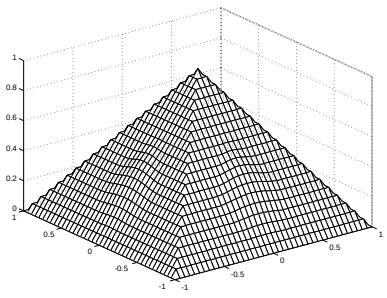
In Table 4, we give relative errors for the first order scheme and for the two second order schemes, both on the entire domain and on a subset of the RSRs. The relative errors are computed by comparing the solution computed using $m+1$ gridpoints to the solution computed using $m/2+1$ gridpoints. These values are not as reliable as absolute errors for determining rates of convergence, but they do give a reasonable indication, and the results are consistent with our expectations. As in the previous example, both second order methods produce second order convergence of the value functions when measured in the L^1 norm on the entire domain and on the RSRs. When these errors are measured in the L^∞ norm on the RSRs, Method I produces second order convergence, while the convergence with Method II is somewhat better than first order. In the case of the optimal controls, both methods apparently give rise to second order convergence in the L^1 norm on the RSRs, while preserving first order convergence in the L^1 norm on the entire domain and in the L^∞ norm on the RSRs. Again, since Assumption 5.1 does not appear to hold everywhere on the RSRs, results of this type are not surprising.

Figures 6(c)–6(f) show approximations with 41 points to the value function and to the first component of the optimal feedback control. There are no apparent qualitative differences in the two approximations to the value function. With Method I, the discontinuities in the control are resolved quite sharply. Method II seems to produce large overshoots, but we note that these are bounded as the grid is refined and do not indicate lack of numerical stability.

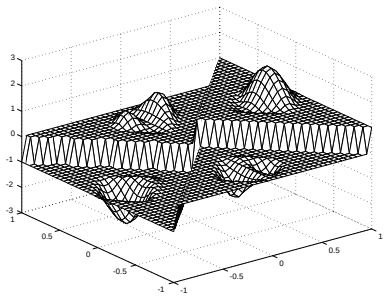
8. Conclusion. We have exhibited two global second order numerical methods for the solution of a class of Hamilton–Jacobi PDE which are related to deterministic

TABLE 3
Perturbed escape time problem errors on partial domain.

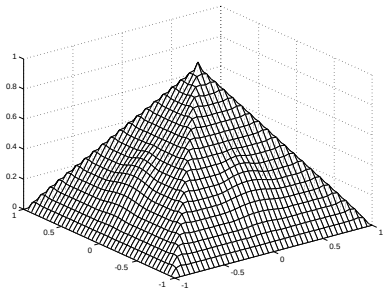
	Value function				Optimal control			
	Method I		Method II		Method I		Method II	
Pts	L^∞ Error	Ord	L^∞ Error	Ord	L^∞ Error	Ord	L^∞ Error	Ord
21	$7.87\,e-03$	–	$8.21\,e-03$	–	$1.49\,e-01$	–	$2.02\,e-01$	–
41	$2.27\,e-03$	1.8	$9.39\,e-04$	3.1	$3.27\,e-02$	2.2	$4.37\,e-02$	2.2
81	$6.45\,e-04$	1.8	$6.34\,e-04$	0.6	$1.24\,e-02$	1.4	$3.26\,e-02$	0.4
161	$1.74\,e-04$	1.9	$2.55\,e-04$	1.3	$3.49\,e-03$	1.8	$1.07\,e-02$	1.6
321	$4.55\,e-05$	1.9	$8.18\,e-05$	1.6	$9.97\,e-04$	1.8	$3.14\,e-03$	1.8
641	$1.16\,e-05$	2.0	$2.32\,e-05$	1.8	$2.65\,e-04$	1.9	$8.65\,e-04$	1.9



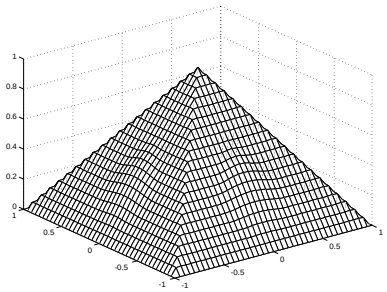
(a) Exact Value Function



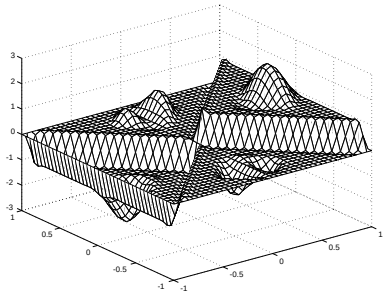
(b) Exact Feedback Control



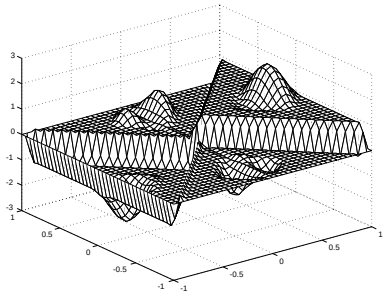
(c) Value Function
Method I



(d) Value Function
Method II



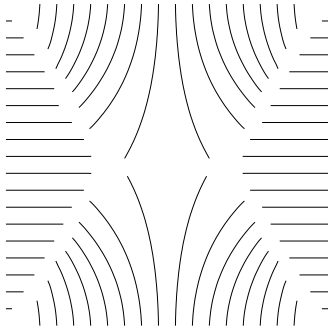
(e) Feedback Control
Method I



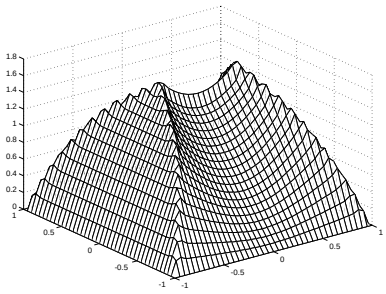
(f) Feedback Control
Method II

FIG. 5. *Perturbed escape time problem solutions.*

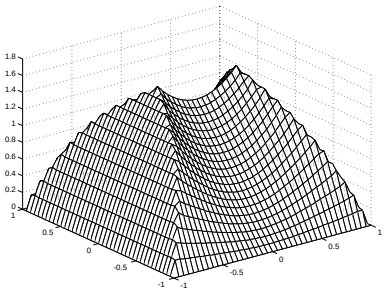
(a) Singularity Set and Smooth Region



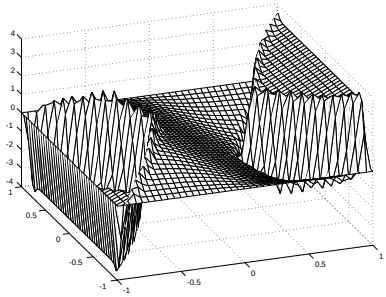
(b) Characteristics



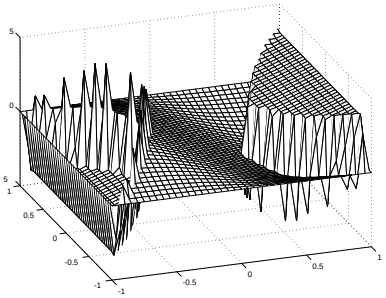
(c) Value Function Method I



(d) Value Function Method II



(e) Feedback Control Method I



(f) Feedback Control Method II

FIG. 6. *Quadratic running cost problem solutions.*

TABLE 4
Quadratic running cost problem relative errors.

Value Function				L^1		L^∞		L^1 RSR		L^∞ RSR	
		Pts	Iter	Rel error	Ord	Rel error	Ord	Rel error	Ord	Rel error	Ord
1st order		21	7	9.54 e − 02	–	9.73 e − 02	–	2.37 e − 02	–	6.05 e − 02	–
		41	7	4.14 e − 02	1.2	5.47 e − 02	0.8	1.47 e − 02	0.7	3.05 e − 02	1.0
		81	7	1.86 e − 02	1.2	2.90 e − 02	0.9	7.56 e − 03	1.0	1.70 e − 02	0.8
		161	7	8.85 e − 03	1.1	1.50 e − 02	1.0	3.81 e − 03	1.0	8.97 e − 03	0.9
		321	7	4.29 e − 03	1.0	7.67 e − 03	1.0	1.92 e − 03	1.0	4.49 e − 03	1.0
		641	7	2.11 e − 02	1.0	3.89 e − 03	1.0	9.67 e − 04	1.0	2.27 e − 03	1.0
2nd order I		21	14	1.51 e − 01	–	2.23 e − 01	–	2.61 e − 02	–	7.47 e − 02	–
		41	14	4.92 e − 02	1.6	1.37 e − 01	0.7	2.59 e − 03	3.3	1.08 e − 02	2.8
		81	14	1.44 e − 02	1.8	7.64 e − 02	0.8	5.17 e − 04	2.3	9.36 e − 04	3.5
		161	14	3.78 e − 03	1.9	4.15 e − 02	0.9	1.43 e − 04	1.9	2.62 e − 04	1.8
		321	14	9.99 e − 04	1.9	2.23 e − 02	0.9	3.78 e − 05	1.9	7.10 e − 05	1.9
		641	14	2.50 e − 04	2.0	1.11 e − 02	1.0	9.71 e − 06	2.0	1.84 e − 05	1.9
2nd order II		21	13	8.53 e − 02	–	6.14 e − 02	–	2.76 e − 02	–	4.29 e − 02	–
		41	14	2.63 e − 02	1.7	5.47 e − 02	0.2	1.01 e − 03	8.1	1.76 e − 03	7.9
		81	14	8.46 e − 03	1.6	4.09 e − 02	0.4	5.36 e − 04	0.9	1.26 e − 03	0.5
		161	14	1.79 e − 03	2.2	2.26 e − 02	0.9	1.76 e − 04	1.6	4.28 e − 04	1.6
		321	14	5.04 e − 04	1.8	9.51 e − 03	1.2	5.19 e − 05	1.8	1.24 e − 04	1.8
		641	14	1.27 e − 04	2.0	6.66 e − 03	0.5	1.41 e − 05	1.9	3.52 e − 05	1.8

Table 4 (CONT.).

Feedback Control

			L^1		L^∞		L^1 RSR		L^∞ RSR	
	Pts	Iter	Rel error	Ord	Rel error	Ord	Rel error	Ord	Rel error	Ord
1st order	21	7	1.20 e − 00	–	2.57 e − 00	–	1.35 e − 01	–	1.45 e − 01	–
	41	7	6.37 e − 01	0.9	3.29 e − 00	0	7.52 e − 02	0.8	8.69 e − 02	0.7
	81	7	3.28 e − 01	1.0	3.29 e − 00	0	3.98 e − 02	0.9	5.43 e − 02	0.7
	161	7	1.49 e − 01	1.1	3.45 e − 00	0	2.08 e − 02	0.9	3.03 e − 02	0.8
	321	7	8.31 e − 02	0.8	3.40 e − 00	0	1.06 e − 02	1.0	1.62 e − 02	0.9
	641	7	3.76 e − 02	1.1	3.57 e − 00	0	5.33 e − 03	1.0	8.41 e − 03	0.9
2nd order I	21	14	1.52 e − 00	–	1.33 e − 00	–	6.53 e − 01	–	1.07 e − 00	–
	41	14	9.49 e − 01	0.7	1.33 e − 00	0	2.01 e − 01	1.7	1.18 e − 00	0
	81	14	5.10 e − 01	0.9	1.36 e − 00	0	4.35 e − 03	5.5	6.70 e − 02	4.1
	161	14	2.68 e − 01	0.9	1.35 e − 00	0	9.65 e − 04	2.2	2.03 e − 03	5.0
	321	14	1.36 e − 01	1.0	1.40 e − 00	0	2.58 e − 04	1.9	7.48 e − 04	1.4
	641	14	6.87 e − 02	1.0	1.39 e − 00	0	6.71 e − 05	1.9	3.68 e − 04	1.0
2nd order II	21	13	3.79 e − 00	–	4.81 e − 00	–	6.31 e − 01	–	9.02 e − 01	–
	41	14	1.07 e − 00	1.8	2.48 e − 00	0	3.10 e − 02	1.0	1.04 e − 01	3.1
	81	14	7.54 e − 01	0.5	5.71 e − 00	0	6.03 e − 03	5.7	1.26 e − 02	3.0
	161	14	3.12 e − 01	1.3	3.27 e − 00	0	2.02 e − 03	1.6	5.53 e − 03	1.2
	321	14	2.14 e − 01	0.5	6.56 e − 00	0	6.11 e − 04	1.7	1.78 e − 03	1.6
	641	14	9.85 e − 02	1.1	6.95 e − 00	0	1.67 e − 04	1.9	8.33 e − 04	1.1

optimal control problems. Our analysis establishes that these methods produce provably second order approximations to the value function and to the optimal feedback control on certain subsets of the RSRs. The rate of convergence results apply only on those regions where each component of the limit optimal feedback control is bounded away from zero so that the sign remains constant. In practice, we observe second order convergence in the L^1 norm on the RSRs even where this assumption is violated, but the results are less favorable when the L^∞ norm is used to measure the errors.

The reason that our theory cannot accommodate changes in the signs of the optimal controls is that the transition probabilities given in (3.9) are singular as functions of the control u wherever any of the components of u are equal to zero. It is likely that we could eliminate the theoretical restrictions on the present method by applying our techniques to a first order method with smooth transition probabilities. In order for such an approach to be practical, however, it would be necessary that the new transition probabilities preserve the highly desirable qualitative properties of those considered here. The one-sided nature of the present transition probabilities is important in terms of achieving sharp resolution of the discontinuities, and their simple form allows us to evaluate the minima analytically at each step in the iterative process used to solve the discrete DPE. Additionally, the control-dependent mean jump times in (3.8) ensure that the number of Gauss–Seidel iterations needed to solve the DPE is small and essentially bounded as the grid is refined. We do not know of smooth transition probabilities which have these qualities, so it is not presently possible for us to improve upon the methods in this paper.

We conclude by suggesting other possible extensions to our work. In principle, the high order asymptotic analysis carried out in section 5 can be used to formulate numerical methods of arbitrarily high order, either by using a more refined correction term in Method I or by taking the linear combination of several approximate solutions in Method II. As a practical matter, since even the second order convergence is not as consistent as we might have hoped in the typical situation where Assumption 5.1 is violated, we do not pursue this avenue. Our analysis considers only homogeneous boundary conditions, but this should not be an essential restriction. Homogeneity at the boundary is used quite strongly in our proof of Lemma 5.5 and in an analogous step in the proof of Theorem 3.4 [11], so it would be necessary to find an alternative approach to estimating the gradient at points near the boundary. Next, the quadratic structure of the running cost as a function of u is probably not needed for the type of asymptotic analysis that we carry out. In practice, this structure is essential for the efficient solution of the discrete DPE as it enables us to evaluate the minima analytically, but it would also be interesting to construct higher order numerical methods for problems where the running cost does not have this form. Finally, we remark that it is worth considering the possibility of applying methods like ours to construct higher order numerical methods for the solution of the second order Hamilton–Jacobi PDEs which arise from stochastic control problems with dynamics given by controlled diffusions.

Appendix A. Proof of Lemma 5.7. The purpose of this appendix is to indicate the calculations which are used to prove Lemma 5.7. The argument essentially consists of elementary algebraic manipulations, but they are rather involved, so it is worthwhile to set out some of the key steps. We combine (5.8) and (5.15) with $m = 1, \dots, q+1$ to obtain the following relation holding on the region B_0^h:

$$
\begin{aligned}
&\langle \bar{u}^h, D^h \Phi^{h,q+1} \rangle \\
&\quad + \sum_{i=0}^{q+1} h^{-(q+1-i)} \, \langle u^0, D^h e^i - D e^i \rangle \\
&\quad - \sum_{m=1}^{q+1} \sum_{k=2}^{m+1} \frac{1}{k!} h^{-(q+1-m)} \langle u^0, D_\backslash^k e^{m+1-k} \rangle \qquad \Big\} \text{(a)} \\
&\quad + h^{-(q+1)} \langle \bar{u}^h - u^0, D^h e^0 - D e^0 \rangle \\
&\quad + \frac{1}{2} \sum_{m=1}^{q+1} \sum_{l=2}^{m} \sum_{k=1}^{m+2-l} \frac{1}{l!k!} h^{-(q+1-m)} \langle D_\backslash^l e^0, a D_\backslash^k e^{m+2-l-k} \rangle \qquad \Big\} \text{(b)} \\
&\quad + \sum_{j=1}^{q+1} h^{-(q+1-j)} \, \langle \bar{u}^h - u^0, D^h e^j \rangle \\
&\quad + \frac{1}{2} \sum_{m=1}^{q+1} \sum_{j=1}^{m-1} \sum_{l=1}^{m-j} \sum_{k=1}^{m+2-j-l} \frac{1}{l!k!} h^{-(q+1-m)} \langle D_\backslash^l e^j, a D_\backslash^k e^{m+2-j-l-k} \rangle \qquad \Big\} \text{(c)} \\
&= 0.
\end{aligned} \tag{A.1}
$$

The first part of Lemma 5.7 will be established if we show that each of the expressions (a)–(c) converges to zero, uniformly for x in $B_{q'}^h$. A key observation is that we can use (5.16) along with (3.7) and (5.7) to obtain the relation

$$
\bar{u}^h - u^0 = -\frac{1}{2} a \sum_{k=2}^{q+1} \frac{1}{k!} h^{k-1} D_\backslash^k e^0 - \frac{1}{2} a \sum_{i=1}^{q} \sum_{k=1}^{q+1} \frac{1}{k!} h^{i+k-1} D_\backslash^k e^i + o(h^q), \tag{A.2}
$$

holding uniformly on $B_{q'}^h$. We indicate the details of the manipulations for (b) and note that the calculations are similar for (a) and (c). Consider the following series of relations which is easily seen to imply that (b) converges to zero, uniformly on $B_{q'}$. Each line is explained after the display.

$$
\begin{aligned}
&h^{-(q+1)} \langle \bar{u}^h - u^0, D^h e^0 - D e^0 \rangle \\
&= -\frac{1}{2} \sum_{l=2}^{q+1} \sum_{k=2}^{q+3-l} \frac{1}{l!k!} h^{-(q+3-k-l)} \langle a D_\backslash^k e^0, D_\backslash^l e^0 \rangle \\
&\quad - \frac{1}{2} \sum_{i=1}^{q} \sum_{l=2}^{q+2-i} \sum_{k=1}^{q+3-l-i} \frac{1}{l!k!} h^{-(q+3-k-l-i)} \langle a D_\backslash^k e^i, D_\backslash^l e^0 \rangle + o(1)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\sum_{l=2}^{q+1}\sum_{k=2}^{q+3-l}\frac{1}{l!k!}h^{-(q+3-k-l)}\langle D_{\backslash}^{l}e^{0}, aD_{\backslash}^{k}e^{0}\rangle \\
&\quad -\frac{1}{2}\sum_{l=2}^{q+1}\sum_{k=1}^{q+2-l}\sum_{m=l+k-1}^{q+1}\frac{1}{l!k!}h^{-(q+1-m)}\langle D_{\backslash}^{l}e^{0}, aD_{\backslash}^{k}e^{m+2-l-k}\rangle + o(1) \\
&= -\frac{1}{2}\sum_{l=2}^{q+1}\sum_{k=2}^{q+3-l}\sum_{m=l+k-2}^{m=l+k-2}\frac{1}{l!k!}h^{-(q+1-m)}\langle D_{\backslash}^{l}e^{0}, aD_{\backslash}^{k}e^{m+2-l-k}\rangle \\
&\quad -\frac{1}{2}\sum_{l=2}^{q+1}\sum_{k=1}^{q+2-l}\sum_{m=l+k-1}^{q+1}\frac{1}{l!k!}h^{-(q+1-m)}\langle D_{\backslash}^{l}e^{0}, aD_{\backslash}^{k}e^{m+2-l-k}\rangle + o(1) \\
&= -\frac{1}{2}\sum_{m=2}^{q+1}\sum_{l=2}^{m}\sum_{k=1}^{m+2-l}\frac{1}{l!k!}h^{-(q+1-m)}\langle D_{\backslash}^{l}e^{0}, aD_{\backslash}^{k}e^{m+2-l-k}\rangle + o(1).
\end{aligned}
$$

The first equality is obtained by applying expression (A.2) and by using the Taylor expansion for $e^0(x)$, as in (5.12); to get the second equality, we rearrange the orders of summation and make the change of variables $m = i + l + k - 2$; we get the third equality by introducing m as a dummy variable in the first summation; the final equality is obtained by rearranging the orders of summation in each of the two terms in the previous line and then combining them into a single summation.

We now turn to verifying the second part of Lemma 5.7. It is easy to see that the following discrete product rule is valid for $p = 1, \ldots, n$:

$$
\begin{aligned}
D_p^h\langle f(x), g(x)h(x)\rangle &= \langle D_p^h f(x), g(x+he_p)h(x+he_p)\rangle \\
&\quad + \langle f(x), g(x+he_p)D_p^h h(x)\rangle \\
&\quad + \langle f(x), D_p^h g(x)h(x)\rangle.
\end{aligned}
$$

Our first application of this product rule to prove the second part of Lemma 5.7 actually involves the simpler case where $g(x)$ is the identity matrix. We state the rule for general $g(x)$ because it will be used later in dealing with the special case of proving the second part of Lemma 5.7 for $q = 0$. For each $p = 1, \ldots, n$, we apply the operator D_p^h to (A.1) and obtain

$$
\langle \bar{u}^h(x+he_p), D^h\Psi_p^{h,q+1}(x)\rangle + \langle D_p^h\bar{u}^h(x), \Psi^{h,q+1}(x)\rangle + \cdots = 0,
$$

where the unspecified quantities are those obtained by applying the product rule with $g(x)$ the identity matrix to the terms (a)–(c) in (A.1). This yields six new terms, and, as in the treatment of (b) above, each of these can be expanded and rearranged to obtain coefficients converging to zero. In the case $q \geq 1$, the condition in (5.16) implies that $D_p^h\bar{u}^h$ converges to $D_p u^0$, so these manipulations suffice to establish the lemma.

Unfortunately, this general calculation is not sufficiently detailed to prove the second part of Lemma 5.7 when $q = 0$, so we treat that as a special case. We do not generalize the manipulations which work for $q = 0$, as they are considerably more involved than the ones indicated above. By using the product rule to apply the

operator D_p^h directly to (5.1), (5.2), and (5.5) and then taking a linear combination of the resulting expressions, we obtain after some manipulations

$$\begin{aligned}
&\Big\langle -a(x+he_p)\left(\frac{1}{2}D^hV^h(x+he_p)+\frac{1}{2}D^hV^h(x)\right)\\
&\quad + b(x+he_p)-\frac{1}{2}ha(x+he_p)D_p^hDV^0(x)\ ,\ D^h\Psi_p^{h,1}(x)\Big\rangle\\
&+\Big\langle -a(x+he_p)D_p^hDV^0(x)-D_p^ha(x)\left(\frac{1}{2}D^hV^h(x)+\frac{1}{2}DV^0(x)\right)\\
&\quad + D_p^hb(x)\ ,\ \Psi^{h,1}(x)\Big\rangle\\
&+o(1)=0,
\end{aligned}$$

holding uniformly for x in B_1^h. Given the assumed convergence of the D^hV^h to DV^0 and the representation for u^0 in (3.7), it is easy to see that the above expression is consistent with (5.21), so it completes the proof of the lemma.

Appendix B. Large deviations upper bound. In this appendix, we state a general large deviations upper bound for a broad class of Markov processes with possibly discontinuous statistics. Our result is essentially an extension of [9, Theorem 1.1], and the reader interested in a proof should consult [9], as well as [29] for comments on the extension. We carefully define the needed notation and end this appendix with a statement of the theorem. Note that the notation in this appendix is independent of the notation in the preceding sections.

Let $\mathbb{R}^n = R^{n_1+n_2}$, and consider a sequence of Markov processes $X^h = (X_1^h, X_2^h)$ with trajectories in $\mathcal{D}([0,\infty):\mathbb{R}^n)$ and generators $\mathcal{L}^h$ such that, for any smooth function $f(x)$ mapping $\mathbb{R}^n$ to $\mathbb{R}$,

$$\begin{aligned}
\mathcal{L}^hf(x) &= \langle \bar{b}^h(x), Df(x)\rangle + \frac{h}{2}\mathrm{tr}[\bar{a}^h(x)D^2f(x)]\\
&\quad + \frac{1}{h}\int_{\mathbb{R}^n}[f(x+h\nu)-f(x)-h\langle \nu, Df(x)\rangle]\bar{\mu}^h(x)(d\nu).
\end{aligned}$$

For each $h>0$, $\bar{a}^h(x)$ and $\bar{b}^h(x)$ are uniformly bounded functions from $\mathbb{R}^n$ to the spaces of $n\times n$ matrices and n-vectors, respectively, and $\bar{\mu}^h(x)$ is a function from $\mathbb{R}^n$ to the space of nonnegative measures on $\mathbb{R}^n$ such that $\bar{\mu}^h(x)$ is uniformly bounded and has uniformly compact support for $x\in\mathbb{R}^n$. We consider block decompositions

$$\bar{a}^h(x)=\begin{bmatrix}\bar{a}_{11}^h(x) & \bar{a}_{12}^h(x)\\ \bar{a}_{21}^h(x) & \bar{a}_{22}^h(x)\end{bmatrix},\qquad \bar{b}^h(x)=\begin{bmatrix}\bar{b}_1^h(x)\\ \bar{b}_2^h(x)\end{bmatrix},$$

where $\bar{a}_{11}^h(x)$ is an $n_1\times n_1$ matrix valued function, $\bar{b}_1^h(x)$ is an n_1-vector valued function, and the other blocks are of appropriate sizes to complete the decompositions. Also, we let $\bar{\mu}_1^h(x)$ be the marginal measure of $\bar{\mu}^h(x)$ on $\mathbb{R}^{n_1}$. In general, we will employ without comment the notation $x=(x_1,x_2)$, where $x_i\in\mathbb{R}^{n_i}$ for $i=1,2$.

For the space of finite measures on $\mathbb{R}^{n_1}$, we define a metric $d(\cdot,\cdot)$ as follows. For two such measures η_1 and η_2, the distance $d(\eta_1,\eta_2)$ is defined to be the supremum of

$$\left|\int_{\mathbb{R}^{n_1}} f(\xi)\eta_1(d\xi)-\int_{\mathbb{R}^{n_1}} f(\xi)\eta_2(d\xi)\right|$$

over all functions $f(\xi)$ for which $\|f\|_\infty \leq 1$ and which satisfy a Lipschitz condition with constant no greater than 1. We note that $d(\cdot,\cdot)$ metrizes weak convergence of probability measures [8, Proposition 11.3.2]. Assume the following:

1. The $\bar{a}_{11}^h(x)$ and $\bar{b}_1^h(x)$ are uniformly bounded for all $x \in \mathbb{R}^n$ and for all $h > 0$.
2. There exists a compact set $K \subset \mathbb{R}^{n_1}$ and a constant $M < +\infty$ such that $\bar{\mu}_1^h(x)(K^c) = 0$ and $\bar{\mu}_1^h(x)(K) \leq M$ for all $x \in \mathbb{R}^n$ and for all $h > 0$.
3. There exist functions $\bar{a}_{11}(x)$ and $\bar{b}_1(x)$ such that

$$\bar{a}_{11}^h(x) \to \bar{a}_{11}(x), \qquad \bar{b}_1^h(x) \to \bar{b}_1(x)$$

   hold as $h \to 0$, uniformly for x in $\mathbb{R}^n$ with x_1 in compact subsets of $\mathbb{R}^{n_1}$.
4. There exists a measure valued function $\bar{\mu}_1(x)$ such that

$$d\Big(\bar{\mu}_1^h(x), \bar{\mu}_1(x)\Big) \to 0$$

   holds as $h \to 0$, uniformly for x in $\mathbb{R}^n$ with x_1 in compact subsets of $\mathbb{R}^{n_1}$.

Notice that conditions 1 and 2 are uniform for all $h > 0$, and hence are more restrictive than what is assumed earlier in the definition of $\mathcal{L}^h$. We now proceed to define the rate function for our large deviations upper bound. For each $x \in \mathbb{R}^n$ and for each vector $\alpha \in \mathbb{R}^{n_1}$, define the convex function

$$H(x,\alpha) = \langle \bar{b}_1(x), \alpha\rangle + \frac{1}{2}\mathrm{tr}[\bar{a}_{11}(x)\alpha\alpha^t] + \int_{\mathbb{R}^{n_1}} [e^{\langle \nu,\alpha\rangle} - 1 - \langle \nu,\alpha\rangle]\mu_1(x)(d\nu),$$

and then for each $x_1 \in \mathbb{R}^{n_1}$

$$H_1(x_1,\alpha) = \sup_{x_2 \in \mathbb{R}^{n_2}} H((x_1,x_2),\alpha).$$

We further define the upper semicontinuous regularization,

$$h_1(x_1,\alpha) = \lim_{\delta \to 0} h_{1,\delta}(x_1,\alpha),$$

where

$$h_{1,\delta}(x_1,\alpha) = \sup_{\|y_1 - x_1\| \leq \delta} H(y_1,\alpha).$$

Consider the Legendre–Fenchel transform given by

$$l(x_1,\beta) = \sup_{\alpha \in \mathbb{R}^{n_1}} [\langle \beta,\alpha\rangle - h_1(x_1,\alpha)]$$

for each $\beta \in \mathbb{R}^{n_1}$. Given $T < +\infty$ and for each $x \in \mathbb{R}^n$, we define the rate function $I_{x_1}(\phi_1)$ by

$$I_{x_1}(\phi_1) = \int_0^T l(\phi_1(s), \dot{\phi}_1(s))ds$$

for absolutely continuous functions ϕ_1 taking values in $\mathbb{R}^{n_1}$ which satisfy $\phi_1(0) = x_1$, and we set $I_{x_1}(\phi_1)$ to be $+\infty$ otherwise.

We can now state the main large deviations theorem.

THEOREM B.1. *Assume conditions* 1–4 *above. Given a compact set* $C \subset \mathbb{R}^{n_1}$, *the following hold:*

(i) *Given $L < +\infty$ and for each $x_1 \in \mathbb{R}^{n_1}$, define*

$$\Phi_{x_1}(L) = \{\phi_1 \in \mathcal{D}([0,T] : \mathbb{R}^{n_1}) : I_{x_1}(\phi_1) \leq L\}.$$

Then the set $\bigcup_{x_1 \in C} \Phi_{x_1}(L)$ is compact.

(ii) *For each closed set $F \subset \mathcal{D}([0,T] : \mathbb{R}^{n_1})$,*

$$\limsup_{h \to 0} h \log P_x\{X_1^h \in F\} \leq - \inf_{\phi_1 \in F} I_{x_1}(\phi_1)$$

holds uniformly for $x \in \mathbb{R}^n$ such that $x_1 \in C$.

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhauser, Boston, 1997.

[2] M. Bardi and P. Soravia, *Hamilton-Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.

[3] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[4] P. Cannarsa, A. Mennucci, and C. Sinestrari, *Regularity results for solutions of a class of Hamilton-Jacobi equations*, Arch. Rational Mech. Anal., 140 (1997), pp. 197–223.

[5] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 273–298.

[6] M. G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[7] M. H. A. Davis, *Piecewise deterministic markov processes: A general class of non-diffusion stochastic models*, J. Roy. Statist. Soc. Ser. B, 46 (1984), pp. 353–388.

[8] R. M. Dudley, *Real Analysis and Probability*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

[9] P. Dupuis, R. S. Ellis, and A. Weiss, *Large deviations for Markov processes with discontinuous statistics* I: *General upper bounds*, Ann. Probab., 19 (1991), pp. 1280–1297.

[10] P. Dupuis and M. James, *Rates of convergence for approximation schemes in optimal control*, SIAM J. Control Optim., 36 (1998), pp. 719–741.

[11] P. Dupuis and A. Szpiro, *Convergence of the optimal feedback policies in a numerical method for a class of deterministic optimal control problems*, SIAM J. Control Optim., 40 (2001), pp. 393–420.

[12] M. Falcone and R. Ferretti, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.

[13] M. Falcone and R. Ferretti, *Convergence analysis for a class of high-order semi-Lagrangian advection schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 909–940.

[14] W. Fleming, *The Cauchy problem for a nonlinear first order partial differential equation*, J. Differential Equation, 5 (1969), pp. 515–530.

[15] W. H. Fleming, *Stochastic control for small noise intensities*, SIAM J. Control, 9 (1971), pp. 473–517.

[16] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[17] W. H. Fleming and P. E. Souganidis, *Asymptotic series and the method of vanishing viscosity*, Indiana Univ. Math. J., 35 (1986), pp. 425–447.

[18] G. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[19] M. Freidlin, *Functional Integration and Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1985.

[20] S. Gottlieb, *Convergence to Steady-State of Weighted ENO Schemes, Norm Preserving Runge-Kutta Methods and a Modified Conjugate Gradient Method*, Ph.D. thesis, Brown University, Providence, RI, 1998.

[21] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[22] C. Hu and C.-W. Shu, *A discontinuous Galerkin finite element method for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (1999), pp. 666–690.
[23] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.
[24] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[25] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1992.
[26] C.-T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.
[27] P.-L. Lions and P. Souganidis, *Convergence of MUSCL and filtered schemes for scalar conservation law and Hamilton-Jacobi equations*, Numer. Math., 69 (1995), pp. 441–470.
[28] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
[29] A. Szpiro, *Asymptotic Analysis and High Order Markov Chain Based Numerical Methods for Optimal Control Problems and Related Hamilton-Jacobi Partial Differential Equations*, Ph.D. thesis, Brown University, Providence, RI, 1999.
[30] J. N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 1184–1205

A STABLE TIME DISCRETIZATION OF THE STEFAN PROBLEM WITH SURFACE TENSION*

BEN SCHWEIZER†

Abstract. We present a time discretization for the single phase Stefan problem with Gibbs–Thomson law. The method resembles an operator splitting scheme with an evolution step for the temperature distribution and a transport step for the dynamics of the free boundary. The evolution step involves only the solution of a linear equation that is posed on the old domain. We prove that the proposed scheme is stable in function spaces of high regularity. In the limit $\Delta t \to 0$ we find strong solutions of the continuous problem. This proves consistency of the scheme, and additionally it yields a new short-time existence result for the continuous problem.

Key words. free boundary problem, time discretization, operator splitting

AMS subject classifications. 35R35, 65M12, 80A22

PII. S003614290037232X

1. Introduction.

The Stefan problem is a set of equations that describe the melting of ice or the growth of ice cristals. At time t the ice (or the water) occupies a region Ω_t, and the second phase occupies the complement of Ω_t. The position of the interface $\partial\Omega_t$ is not known a priori but must be determined together with the temperature distribution $\Theta(t)$. Several sets of evolution equations can be found in the extensive literature (see [11] for equations and further references). Commonly used is the heat equation (2.1) in the domain Ω_t (in the two phase problem another heat equation is posed in the complement of Ω_t). The latent heat relates the normal heat flux on the free boundary (or its jump across the boundary) with the speed of the free boundary as in (2.2). In order to determine the evolution we need one more boundary condition. Various possibilities are studied for that: (a) fixed temperature $\Theta = 0$, (b) the Gibbs–Thomson relation $\Theta \sim \kappa$ with κ being the mean curvature of the boundary, and (c) kinetic undercooling: temperature plus a multiple of the speed is proportional to the mean curvature. In the paper at hand we are interested in case (b), the Gibbs–Thomson relation (2.3).

The aim of this paper is to introduce a stable time discretization of the two-dimensional free boundary value problem. We consider the single phase problem for simplicity; the two phase problem can be treated with the same method. Since the domain changes with time, it is not clear what equations we should pose at every time step, how to define a new domain, and how to define a temperature distribution on the new domain. Thinking of the numerical use of the scheme it is desirable that at each time step only a linear equation must be solved. This linear equation should be posed on the old domain. Our scheme will provide exactly this. As a by-product of our stability result in Theorem 2.2 we find a short-time existence result for (2.1)–(2.3) in Corollary 2.3. Such a result (in different function spaces) was proved earlier by Radkevich in [8]. Our approach is more elementary in the sense that it involves less functional-analytic machinery.

*Received by the editors May 11, 2000; accepted for publication (in revised form) March 12, 2002; published electronically September 12, 2002.
http://www.siam.org/journals/sinum/40-3/37232.html
†Institut für Angewandte Mathematik, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany (ben.schweizer@iwr.uni-heidelberg.de).

Another time discretization for the Gibbs–Thomson law was introduced by Luckhaus in [6]. His approach assumes only very low regularity such that solutions can be defined past geometric singularities. For that it is necessary to use explicitly the new domain in the definition of the time step. In this context we wish to mention the work of Bänsch [3] dealing with a time discretization for the Navier–Stokes equations with a free boundary. Also in these more complicated equations the new geometry is needed in the definition of the new iterate.

Let us compare the Gibbs-Thomson law (b) with kinetic undercooling (c). The term introduced in case (c) is regularizing; mathematically it has the effect that one can regard the equations as a coupled system of a heat equation and an equation for the motion of the free boundary. The regularity properties of the two evolution equations allow us to iterate the two solution operators. The fixed point is a solution of the original problem. Such an iteration is used by Chen and Reitich in [4] and by Abergel et al. in [1] in order to derive an existence result in case (c). A spatial semidiscretization was studied by Veeser in [10]. In contrast to case (c), it seems impossible to decouple the equations in case (b).

This paper is organized as follows. In section 2 we present the operator splitting scheme (OS) for a time discretization. Each time step consists of (1) defining an auxiliary velocity field v, (2) solving a linear equation with transport term $v \cdot \nabla$, and (3) defining the new domain and a temperature field on the new domain by advection. In Theorem 2.2 and Corollary 2.3 we state our main result: the proposed scheme (OS) is stable and consistent.

Within this paper we introduce three different schemes. Scheme (OS) is the numerically applicable scheme in physical variables. The analysis of (OS) is the goal of this paper, and the results are collected in section 2. In order to prove our results we introduce a linear Crank–Nicolson-type scheme (CN) for unknowns (u, h). (CN) is defined on a fixed domain and considers a given right-hand side f; detailed a priori estimates are derived in section 3. The next step is to consider scheme (CN) with a right-hand side of the form $f = f(u, h)$. Note that this is in general not a practical numerical scheme, since f may depend on the values of the solution at later times. The special choice of $f(u, h)$ in section 4 is motivated by the original equations and their transformation to a fixed domain. We prove the existence of solutions and a priori estimates. In section 5 we conclude that the original scheme (OS) inherits these properties.

As already mentioned, our analysis is based on the study of a linear problem. This linear problem is obtained by transforming the equations onto a rectangle and linearizing them. This defines an operator in the unknown quantities temperature distribution u and height function of the free boundary h. This linear operator has a compact inverse with regularizing properties. It allows us to solve instationary problems with a time discretization (CN). The discretization can be proven to be stable by a testing procedure. Since the nonlinearity requires regular solutions, we apply the results also to discrete time derivatives and to second spatial derivatives of the time-discrete solutions. This yields estimates in function spaces of high regularity. In section 3 we collect estimates for (CN), the semidiscrete equations on a fixed domain. Some care must be taken of compatibility conditions of the initial values.

Note that similar facts of the corresponding linearized problem were used in [9] in order to treat the Navier–Stokes equations with a free boundary. Let us again compare cases (b) and (c): in case (c) the properties of the linear operator can be shown with an iteration that solves successively for u and h. In the case at hand one actually has to study the coupled system.

In section 4 we consider a time discretization of transformed equations and apply the results of section 3. It turns out to be of importance in which point we linearize the equations. Concerning the mean curvature operator of the Gibbs–Thomson law it is sufficient to linearize it about the initial values. This is different for the nonlinearity introduced by the domain transformation: it has a different structure and cannot be treated by introducing error terms on the right-hand side (see Lemma 3.2 and remarks thereafter). We have to use in every time step the linearization of the equations on the current "old" domain. This introduces time-dependent coefficient matrices in the equations, but this way the transformation respects the variational character of the problem. As it turns out, the scheme (OS) mimics this strategy of linearization.

We encounter the fact that the solution of the discrete equations does not satisfy maximal regularity estimates. Therefore we have to be careful in the discretization of the nonlinearity.

In section 5 we prove Theorem 2.2 for scheme (OS). The idea is to transform the operator splitting scheme onto a reference domain and to apply the results of section 3. It will turn out that the transformation of scheme (OS) is actually identical to the scheme of section 4. The results of section 4 imply the stability of the transformed scheme and therefore the stability of the original scheme. Since (OS) is consistent with the continuous equations we can conclude that weak limits of the discrete solutions define strong solutions of the original problem.

2. The free boundary problem and the time discretization. We denote the domain that is covered with ice (or water) at time t by Ω_t. For notational convenience we assume that the free boundary is given as the graph of a single function. We study the two-dimensional case and write $S := [0,1]_{per}$ for the unit interval with identified endpoints. A function defined on S is automatically periodic; in particular, all derivatives (if defined) coincide in the endpoints. We write the domain as

$$\Omega_t = \{(x,y) | x \in S, 0 < y < h(t,x)\}.$$

The height function h will be close to 1, and we can always parametrize Ω_t over the standard rectangle $S \times (0,1)$. Again, all functions on the rectangle are automatically periodic on the lateral boundaries. We introduce the time-dependent function $H(t,x) = (x, h(t,x))$ to parametrize the upper boundary of Ω_t. In the following we will often omit the argument t. By a rescaling argument we can assume that the physical constants latent heat, surface tension, and thermal diffusion are all equal to 1. The physical equations then read

$$\partial_t \Theta = \Delta \Theta \qquad \text{in } \bigcup_{t>0} \{t\} \times \Omega_t, \tag{2.1}$$

$$\partial_t h = -(n \cdot \nabla \Theta) \circ H \sqrt{1 + |\partial_x h|^2} \qquad \text{on } \{(t,x) | t > 0, x \in S\}, \tag{2.2}$$

$$\Theta \circ H = \kappa \qquad \text{on } \{(t,x) | t > 0, x \in S\}. \tag{2.3}$$

Here

$$\kappa := -\partial_x \left(\frac{\partial_x h}{\sqrt{1 + |\partial_x h|^2}} \right)$$

is the mean curvature of the free boundary, n is the exterior normal of Ω, and $n_2 = (1 + |\partial_x h|^2)^{-1/2}$ the second component of n. The above equations are complemented with a boundary condition for Θ on the lower boundary, say,

$$\Theta(t,x,0) = \psi(t,x).$$

For notational convenience we will use $\psi \equiv 0$ in the following. All results remain valid for smooth ψ. Additionally, initial values $\Theta(t=0) = \Theta_0$ and $h(t=0) = h_0 > 0$ are imposed.

Later on we will use the linearization of the mean curvature operator about $h = h_0$,

$$\underline{\Delta} h := \; -D\kappa(h_0) \cdot h = \partial_x \left(\frac{\partial_x h}{\sqrt{1 + |\partial_x h_0|^2}^3} \right).$$

For smooth and small h_0 the properties of $\underline{\Delta}$ are similar to those of $\Delta_x = \partial_x^2$, therefore the notation.

We now introduce a uniform discretization of the time interval $(0, T)$ by $t_k := k \cdot \Delta t$. Note that nonuniform time partitions can also be treated with our method. The pair (Θ^k, h^k) is meant to approximate $(\Theta(t_k), h(t_k))$. We set $(\Theta^0, h^0) := (\Theta_0, h_0)$. The function h^k defines the domain $\Omega^k := \{(x, y) | x \in S, 0 < y < h^k(x)\}$ and the normal vector n^k. We use $H^k(x) := (x, h^k(x)) \in \mathbb{R}^2$. In the following definition we need functions $\Theta^{(-1)} := \tilde{\Theta}^{(-1)} := \Theta^0$ for the first execution of Step 1. We define $H^{(-1)}$ and $n^{(-1)}$ via $h^{(-1)} := h^0$.

Let us motivate in advance (2.5): let Θ solve $\partial_t \Theta = \Delta \Theta$ on the time-dependent domain Ω_t. We consider $\tilde{\Theta}(t, .) := \Theta(t, \Phi(t, .))$, where $\Phi(t, .)$ parametrizes Ω_t over the fixed domain Ω_{t_0}: $\Phi(t, .) : \Omega_{t_0} \to \Omega_t$. Then $\tilde{\Theta}$ satisfies

$$\partial_t \tilde{\Theta} = (\partial_t \Theta) \circ \Phi + (\nabla \Theta) \circ \Phi \cdot \partial_t \Phi = \Delta \Theta|_\Phi + \partial_t \Phi \cdot \nabla \Theta|_\Phi.$$

If we want to calculate on a given domain (the "old" domain Ω_{t_0}), then we have to include a convective term in the heat equation. In the numerical scheme it remains to choose a guess for the corresponding velocity field.

DEFINITION 2.1. We assume that an initial domain Ω^0 is given by h^0 and an initial temperature by $\Theta^0 : \Omega^0 \to \mathbb{R}$. Let $X^0 : R \to \Omega^0$ be a parametrization of Ω^0.

The operator splitting scheme (OS) for a time discretization of (2.1)–(2.3) is defined by the following three steps; they are executed beginning with $k = 0$.

Step 1. We use the temperature data of the last time step in order to define a vertical velocity field $v^k = (v_1, v_2) = (0, v_2) : \Omega^k \to \mathbb{R}^2$ with boundary values

$$n^{k-1} \circ H^{k-1} \cdot v^k \circ H^k = \left(n^{k-1} \cdot \nabla \frac{\Theta^{k-1} + \tilde{\Theta}^{k-1}}{2} \right) \circ H^{k-1}$$

by the linear interpolation

$$v^k(x, y) = \frac{y}{h^k(x)} v^k(x, h^k(x)). \tag{2.4}$$

Step 2. Find $\tilde{\Theta}^k : \Omega^k \to \mathbb{R}$ and $h^{k+1} : [0, 1] \to \mathbb{R}$ with

$$\frac{\tilde{\Theta}^k - \Theta^k}{\Delta t} = \Delta \left(\frac{\Theta^k + \tilde{\Theta}^k}{2} \right) + v^k \cdot \nabla \frac{\Theta^k + \tilde{\Theta}^k}{2} \text{ in } \Omega^k, \tag{2.5}$$

$$\frac{h^{k+1} - h^k}{\Delta t} = -\sqrt{1 + |\partial_x h^k|^2} \left(n^k \cdot \nabla \frac{\Theta^k + \tilde{\Theta}^k}{2} \right) \circ H^k \quad \text{in } [0, 1], \tag{2.6}$$

$$\tilde{\Theta}^k \circ H^k + \underline{\Delta}(h^{k+1} - h^k) = \kappa(h^k) \quad \text{in } [0, 1]. \tag{2.7}$$

On the lower boundary we impose $\tilde{\Theta}^k(x,0) = \psi(x,t_k)$. We slightly change the definition in the first time step $k=0$. There we use Θ_0 instead of $\frac{1}{2}(\Theta^0 + \tilde{\Theta}^0)$ in the convective term of (2.5).

Step 3. The function h^{k+1} defines the new domain Ω^{k+1}. We now want to define a temperature field Θ^{k+1} on the new domain. We set

$$X^{k+1}(x,y) := X^k(x,y) + \left(0, \frac{X_2^k(x,y)}{h^k(x)}\right) \cdot (h^{k+1} - h^k)(x), \tag{2.8}$$

$$\Theta^{k+1} \circ X^{k+1} := \tilde{\Theta}^k \circ X^k \qquad \text{in } R. \tag{2.9}$$

We will show that the above scheme can be used to define uniquely $(\Theta^k, X^k)_{k=0,\dots,K}$. The functions Θ^k are defined on domains that depend on time (on k). The domains are always parametrized by $X^k = (X_1^k, X_2^k)$. In order to formulate estimates we introduce the pairs $(u^k, h^k) := (\Theta^k \circ X^k, h^k)$. The functions u^k are then defined on the time-independent domain R.

The main result of this paper is the following theorem. It is proved together with its corollary in section 5.

THEOREM 2.2. *Let the initial values (u_0, h_0) satisfy the regularity and compatibility assumption, Assumption* 5.1, *and let $h_0 - 1$ be small in $C^{0,1}(S)$. Let the initial domain be parametrized over the rectangle $R = S \times (0,1)$ with a diffeomorphism $X^0 \in H^{4+1/2}(R)$ with $X^0 - id$ small in $C^{0,1}(R)$, $X_1^0(x,y) = x$ and $\partial_2 X_2^0(.,1) = 1$.*

Then, on a small time interval $I = (0,T)$ the scheme (OS) has a unique solution for $k = 1, \dots, K$ with $t_K < T$. The scheme is stable: the linear interpolant (u,h) of $(u^k, h^k)_k$ satisfies the estimate

$$\begin{aligned} &\|h\|_{L^\infty(I;H^{4+1/2}(S))} + \|h\|_{W^{1,\infty}(I;H^{2+1/2}(S))} \\ &\quad + \|u\|_{L^\infty(I;H^3(R))} + \|u\|_{W^{1,\infty}(I;H^1(R))} \le C. \end{aligned}$$

The number C and the time interval I depend only on the initial values (Θ_0, h_0). They are independent of the time-step size Δt.

COROLLARY 2.3. *Consider solutions $(u,X)_{\Delta t}$ as in Theorem* 2.2. *For a subsequence $\Delta t \to 0$ there holds*

$$(u,X)_{\Delta t} \longrightarrow (\tilde{u}, \tilde{X}) \qquad \text{for} \quad \Delta t \to 0 \tag{2.10}$$

in the norms of $L^2(I;H^2(R)) \cap H^1(I;L^2(R))$ and of $H^1(I;H^3(R))$. The limit function $(\Theta, X) := (\tilde{u} \circ \tilde{X}^{-1}, \tilde{X})$ is a strong solution of the physical problem (2.1)–(2.3).

Note that in the above results no smallness assumption is made on Θ_0; the velocity of the boundary can be large, and convective effects must be included in the scheme. On the other hand, we assume smallness of X^0. This is not a severe restriction, since one could parametrize all domains Ω^k over a reference domain that is close to Ω^0. Then smallness of X^0 is guaranteed.

A remark on implementations of the scheme. In the stability result we use the assumption that initially the height function is almost constant. This is done in order to simplify the proofs. It would be sufficient to have the initial domain close to a smooth reference domain (which is no restriction if the initial values are smooth).

Running the scheme is possible only for small times. This is because one of the following may happen: (1) The domain transformation onto the reference domain introduces large errors. (2) Using the linearization of the mean curvature operator about the initial values is no longer appropriate. (3) A geometric singularity makes a

smooth parametrization impossible. Note that this is possible also for the continuous equations.

The best we can expect of the discretization is to work well as long as there exist continuous solutions of the system, that is, until problem (3) appears. In general, our method will fail to work before that time, due to problem (1) or (2). In this case one may continue with a restart: choose a new smooth reference domain, calculate the new linearized mean curvature operator, and restart the scheme.

3. A Crank–Nicolson scheme on the reference domain. After a transformation of (2.1)–(2.3) onto the reference domain $R = S \times (0,1)$ the equations have a linearization of the form (3.1)–(3.3). This section is devoted to the study of these linear equations on the rectangle.

$$\partial_t u = \nabla \cdot A(t) \nabla u + f_0 \qquad \text{in } R, \tag{3.1}$$

$$\partial_t h = -a(t) \cdot \nabla u(.,1) + f_1 \qquad \text{on } S, \tag{3.2}$$

$$u(.,1) = -\underline{\Delta} h + f_2 \qquad \text{on } S. \tag{3.3}$$

We assume $a(t) = e_2 \cdot A(t)$ and $A(t) : R \to \mathbb{R}^{2\times 2}$. In the following we always impose without further mentioning the condition $u = \psi = 0$ (and $u^k = 0$) on the lower boundary $\{(x,y)|y = 0\}$. This also enables us to make use of the Poincaré inequality in what follows. A natural time discretization of (3.1)–(3.3) is the following Crank–Nicolson scheme.

DEFINITION 3.1. *We denote the following scheme by (CN). In every time step we define* $u^{k+1} : R \to \mathbb{R}$, $h^{k+1} : S \to \mathbb{R}$ *as the solution of*

$$\frac{u^{k+1} - u^k}{\Delta t} = \nabla \cdot A^k \nabla \frac{u^k + u^{k+1}}{2} + f_0^k \qquad \textit{in } R, \tag{3.4}$$

$$\frac{h^{k+1} - h^k}{\Delta t} = -a^k \cdot \nabla \left(\frac{u^k + u^{k+1}}{2} \right)(.,1) + f_1^k \qquad \textit{on } S, \tag{3.5}$$

$$u^{k+1}(.,1) = -\underline{\Delta} h^{k+1} + f_2^{k+1} \qquad \textit{on } S. \tag{3.6}$$

Notation. In the following we will denote the averages of solutions at intermediate points as $u^{k+1/2} := \frac{u^k + u^{k+1}}{2}$. The linear interpolant of the values (u^k, h^k) will always be denoted by (u,h), and linear interpolants of $f^k = (f_0^k, f_1^k, f_2^k)$ are denoted by $f = (f_0, f_1, f_2)$. We will once also use the linear interpolant of the values $u^{k+1/2}$; it will be denoted by $\bar{u}$.

In the scheme (CN) the matrices A^k will be uniformly close to the identity $I_2 \in \mathbb{R}^{2\times 2}$. Nevertheless, it will be of importance to use the coefficient matrices in (3.4) and the corresponding oblique derivatives in (3.5). Loosely speaking, we must avoid any error term f_1^k in (3.5). This statement is made precise in the subsequent lemma. The lemma gives a result on the resolvent problem corresponding to (3.1)–(3.3). It introduces function spaces that are natural for the problem.

LEMMA 3.2 (the resolvent problem in energy spaces). *Let* $A : R \to \mathbb{R}^{2\times 2}$ *be a field of uniformly elliptic and symmetric matrices. Then for* $\lambda > 0$ *the equations*

$$\lambda u - \nabla \cdot A \nabla u = g_0 \qquad \textit{in } R, \tag{3.7}$$

$$\lambda h + e_2 \cdot A \cdot \nabla u(.,1) = g_1 \qquad \textit{on } S, \tag{3.8}$$

$$u(.,1) + \Delta_x h = g_2 \qquad \textit{on } S, \tag{3.9}$$

together with $u(.,0) = 0$, have a unique solution (u,h). It satisfies the resolvent estimate

$$\begin{aligned}\lambda^2\|u\|_0^2 + \|\nabla\cdot A\nabla u\|_0^2 + \lambda^2\int_S |\partial_x h|^2 + \lambda\|h\|_{2+1/2}^2 \\ \leq C\,(\|g_0\|_0^2 + \|g_1\|_1^2 + \lambda^2\|g_2\|_{-1}^2 + \lambda\|g_2\|_{1/2}^2),\end{aligned} \tag{3.10}$$

with C independent of λ. Here $\|.\|_s$ denotes the norm of H^s.

Proof. To prove existence we assume $g_2 = 0$; this can be achieved by defining the new unknown to be $h - \Delta_x^{-1} g_2$. We find u as the minimizer of

$$\begin{aligned}E(u) := \lambda\int_R u^2 + \int_R A\nabla u\cdot\nabla u - \lambda\int_S \Delta_x^{-1}u(.,1)\cdot u(.,1) \\ - 2\int_R g_0\cdot u + 2\int_S g_1\cdot u(.,1)\end{aligned}$$

in $\{u \in H^1(R) | u(.,0) = 0, \int_S u(.,1) = 0\}$. Here the operator Δ_x^{-1} is defined by prescribing vanishing averages. With the function $\tilde{h} := \Delta_x^{-1}u(.,1)$ the pair $(u,\tilde{h})$ solves (3.7), (3.9) exactly and (3.8) up to a constant function. Defining $h(x) := \bar{h} + \tilde{h}(x)$ with an appropriate constant $\bar{h}$ we obtain a solution to (3.7)–(3.9).

To find the a priori estimate we multiply (3.7) with $\lambda u - \nabla\cdot A\nabla u$ and integrate over R. This yields

$$\lambda^2\int_R |u|^2 + \int_R |\nabla\cdot A\nabla u|^2 - 2\lambda\int_R u\,\nabla\cdot A\nabla u = \int_R (\lambda u - \nabla\cdot A\nabla u)g_0.$$

With another integration by parts we find

$$\begin{aligned}\lambda^2\int_R |u|^2 + \int_R |\nabla\cdot A\nabla u|^2 + 2\lambda\int_R \nabla u\cdot A\nabla u \\ + 2\lambda\int_S (g_2 - \Delta_x h)(\lambda h - g_1) = \int_R (\lambda u - \nabla\cdot A\nabla u)g_0.\end{aligned}$$

The third term is positive, and in the fourth term we perform an integration by parts over S. We find an estimate for the first three terms on the left-hand side of (3.10). The estimate for $\sqrt{\lambda}h \in H^{2+1/2}(S)$ then follows from regularity for (3.9). ∎

We read the above lemma as follows: the linearized problem has a good resolvent operator, and we can expect high regularity of solutions of the coupled problem. There are two restrictive points. In (3.10) an estimate of λg_2 is needed on the right-hand side. This means that in the time-dependent problem the time derivative of f_2 must be controlled. The second difficulty is the regularity property that is assumed for g_1. In particular, we cannot insert an error of the form "trace of a first derivative of u." This is the reason why we use the oblique derivatives in (3.5).

DEFINITION 3.3. *For a solution (u,h) we define the Banach space $Y := Y_u \times Y_h$ with*

$$\begin{aligned}Y_u &:= L^\infty(0,T;H^1(R)) \cap H^1(0,T;L^2(R)),\\ Y_h &:= L^\infty(0,T;H^{2+1/2}(S)) \cap H^1(0,T;H^1(S)).\end{aligned}$$

To control the right-hand side we define the Banach spaces

$$\begin{aligned}X_0 &:= L^2(0,T;L^2(R)),\\ X_1 &:= L^2(0,T;H^1(S)),\\ X_2 &:= L^\infty(0,T;H^{1/2}(S)) \cap H^1(0,T;H^{-1}(S)).\end{aligned}$$

Observe that the above are not the maximal regularity spaces of the continuous equations. For that we would expect additional estimates for $u \in L^2(I;H^2)$ and $h \in L^2(I;H^{3+1/2})$. However, the above Crank–Nicolson scheme will not provide such an estimate. It can provide it at best for the interpolant of the midpoints $\frac{1}{2}(u^k+u^{k+1})$.

LEMMA 3.4 (the scheme (CN) in energy spaces). *Assume that the coefficient matrices A^k in Definition* 3.1 *are symmetric and satisfy*

$$\sup_k \|A^k - I_2\|_{L^\infty(R)} + \sum_k \left\| \frac{A^{k+1}-A^k}{\Delta t} \right\|_{L^\infty(R)} \Delta t < \delta. \tag{3.11}$$

We consider initial values $u_0 \in H^1(R)$, $h_0 \in H^{2+1/2}(S)$. Let (3.6) *be satisfied for the initial values $(u^0,h^0) := (u_0,h_0)$; that is,* (3.6) *holds for $k=-1$. Given a right-hand side $(f^k)_k$ we will write estimates in terms of the linear interpolant $f : I \to L^2(R)^2 \times L^2(S) \times L^2(S)$. Let the time interval $I=(0,T)$ and $\delta > 0$ be small enough.*

Then for every $K \in \mathbb{N}$ with $K \cdot \Delta t \le T$ the linear scheme (CN) of Definition 3.1 *has a unique solution $(u^k,h^k)_{k=0,\dots,K}$. The linear interpolant (u,h) of $(u^k,h^k)_k$ satisfies the estimate*

$$\|(u,h)\|_Y \le C_1 \|u_0\|_{H^1(R)} + C_2 \ (\|f_0\|_{X_0} + \|f_1\|_{X_1} + \|f_2\|_{X_2}) \tag{3.12}$$

with C_1 and C_2 independent of Δt.

The estimate (3.12) *can be improved: on the right-hand side we can replace $\|f_2\|_{X_2}$ by $\|f_2\|_{H^1(0,T;H^{-1}(S))} + C_\star$, where $C_\star$ has the property that for some $C > 0$ every solution of* (3.6) *satisfies*

$$\|h^{k+1}\|_{H^{2+1/2}(S)} \le C_\star + C\|u^{k+1}\|_{H^1(R)}.$$

Proof. The proof of this lemma relies on a testing procedure; it is analogous to the proof of the resolvent estimate of Lemma 3.2. We multiply (3.4) with $-\nabla \cdot A^k \nabla(\frac{u^k+u^{k+1}}{2})$. An integration over R yields

$$\begin{aligned} &\int_R \nabla \frac{u^{k+1}-u^k}{\Delta t} \cdot A^k \nabla \frac{u^k+u^{k+1}}{2} + \left\| \nabla \cdot A^k \nabla \frac{u^k+u^{k+1}}{2} \right\|_{L^2(R)}^2 \\ &- \int_S a^k \cdot \nabla \left(\frac{u^k+u^{k+1}}{2} \right) \frac{u^{k+1}-u^k}{\Delta t} = - \int_R f_0^k \cdot \nabla \cdot A^k \nabla \frac{u^k+u^{k+1}}{2}. \end{aligned} \tag{3.13}$$

We use the symmetry of A^k to calculate for the first term

$$\begin{aligned} &\int_R \nabla \frac{u^{k+1}-u^k}{\Delta t} \cdot A^k \nabla \frac{u^k+u^{k+1}}{2} \\ &\quad = \frac{1}{2\Delta t} \int_R A^k \nabla u^{k+1} \cdot \nabla u^{k+1} - \frac{1}{2\Delta t} \int_R A^k \nabla u^k \cdot \nabla u^k. \end{aligned}$$

To evaluate the boundary integral we use (3.5) with index k and (3.6) with the indices k and $k+1$:

$$\begin{aligned} &\int_S a^k \cdot \nabla \left(\frac{u^k+u^{k+1}}{2} \right) \left(\frac{u^{k+1}-u^k}{\Delta t} \right) \\ &\quad = \int_S \left(\frac{h^{k+1}-h^k}{\Delta t} - f_1^k \right) \cdot \left(\underline{\Delta} \frac{h^{k+1}-h^k}{\Delta t} - \frac{f_2^{k+1}-f_2^k}{\Delta t} \right). \end{aligned}$$

Inserting this into (3.13) we find

$$\begin{aligned}
&\frac{1}{2\Delta t}\int_R A^{k+1}\nabla u^{k+1}\cdot\nabla u^{k+1} - \frac{1}{2\Delta t}\int_R A^k\nabla u^k\cdot\nabla u^k \\
&+ \left\|\nabla\cdot A^k\nabla\frac{u^k+u^{k+1}}{2}\right\|^2_{L^2(R)} - \int_S \frac{h^{k+1}-h^k}{\Delta t}\cdot\underline{\Delta}\frac{h^{k+1}-h^k}{\Delta t} \\
&= -\int_R f_0^k\cdot\nabla\cdot A^k\nabla\frac{u^k+u^{k+1}}{2} + \int_R \frac{A^{k+1}-A^k}{2\Delta t}\nabla u^{k+1}\cdot\nabla u^{k+1} \\
&- \int_S\left(\frac{h^{k+1}-h^k}{\Delta t} - f_1^k\right)\cdot\frac{f_2^{k+1}-f_2^k}{\Delta t} - \int_S f_1^k\cdot\underline{\Delta}\frac{h^{k+1}-h^k}{\Delta t}.
\end{aligned}$$

Multiplication with Δt and summing up over $k = 0,\dots,K-1$ we find

$$\begin{aligned}
&\int_R A^K\nabla u^K\cdot\nabla u^K + \sum_k\left\|\nabla\cdot A^k\nabla\frac{u^k+u^{k+1}}{2}\right\|^2_{L^2(R)}\Delta t \\
&+\sum_k\int_S\left|\partial_x\left(\frac{h^{k+1}-h^k}{\Delta t}\right)\right|^2\Delta t \;\le\; 2\|\nabla u^0\|^2_{L^2(R)} \\
&+2\sum_k\left\|\frac{A^{k+1}-A^k}{\Delta t}\right\|_{L^\infty(R)}\cdot\left\|\nabla u^{k+1}\right\|^2_{L^2(R)}\;\Delta t \\
&+C\sum_k\left\{\|f_0^k\|_0^2 + \int_S[|\partial_x f_1^k|^2 + |f_1^k|^2] + \left\|\frac{f_2^{k+1}-f_2^k}{\Delta t}\right\|^2_{-1}\right\}\Delta t.
\end{aligned} \tag{3.14}$$

For the linear interpolant (u,h) of the sequence (u^k,h^k) we find with (3.4) the estimate

$$\|u\|_{L^\infty(I;H^1(R))} + \|\partial_t u\|_{L^2(I;L^2(R))} + \|\partial_x\partial_t h\|_{L^2(I;L^2(S))} \le C. \tag{3.15}$$

It remains to prove spatial regularity properties of h. Since traces of u^k are bounded in the space $l^\infty(\{0,\dots,K\};H^{1/2}(S))$, (3.6) implies the regularity of h. The improved version of the estimate mimics this argument. □

The nonlinearity of the original problem requires the control of the domain in regular norms. Estimates of higher order can be derived by considering derivatives of solutions. They satisfy again equations of the type (3.4)–(3.6), and we can apply Lemma 3.4.

We introduce a notation. As before we write g for the linear interpolant of a set of functions $(g^k)_k$. We will write $\bar\partial_t g$ for the linear interpolant of the discrete time derivatives $\frac{g^k-g^{k-1}}{\Delta t}$. In this way we can use also time derivatives of $\bar\partial_t g$; they are piecewise constant functions with values $\frac{g^{k+1}-2g^k+g^{k-1}}{(\Delta t)^2}$. The function $\bar\partial_t g$ is defined on the time interval $(\Delta t, T)$, and all norms are calculated on that interval.

PROPOSITION 3.5 (the scheme (CN) with higher regularity). *Let the compatibility assumption, Assumption* 3.6, *on the initial values be satisfied. Assume that the coefficient matrices are symmetric and satisfy*

$$\begin{aligned} \sup_k \|A^k - I_2\|_{C^0(\bar{R})} &< \delta, \\ \sum_k \left\{ \|\nabla A^k\|^2_{H^2(R)} + \|\nabla A^k(.,1)\|^2_{H^2(S)} \right\} \Delta t &< \delta^2, \\ \sum_k \left\{ \left\| \frac{A^{k+1} - A^k}{\Delta t} \right\|^2_{H^1(R)\cap L^\infty(R)} + \left\| \frac{A^{k+1} - A^k}{\Delta t}(.,1) \right\|^2_{H^1(S)} \right\} \Delta t &< \delta^2. \end{aligned} \tag{3.16}$$

On the initial values we assume $\|h_0 - 1\|_{C^{0,1}(S)} < \delta$. *Let* $T > 0$ *and* $\delta > 0$ *be small enough and* $(u^k, h^k)_k$ *be a solution of scheme (CN). Then the linear interpolant* (u, h) *satisfies*

$$\begin{aligned} &\|\bar{\partial}_t(u,h)\|_Y + \|\partial_x^2(u,h)\|_Y \\ &\quad \le C_1 \left[\|\nabla \cdot A(0)\nabla u_0 + f_0(0)\|_{H^1(R)} + \|\partial_x^2 u_0\|_{H^1(R)} \right] \\ &\quad\quad + C_2 \left[\|\bar{\partial}_t f_0\|_{X_0} + \|\bar{\partial}_t f_1\|_{X_1} + \|\bar{\partial}_t f_2\|_{X_2} \right] \\ &\quad\quad + C_3 \left[\|\partial_x^2 f_0\|_{X_0} + \|\partial_x^2 f_1\|_{X_1} + \|\partial_x^2 f_2\|_{X_2} \right] \\ &\quad\quad + C_4\delta \left[\|f_0\|_{L^\infty(I;H^1(R))} + \|A\|_{L^\infty(I;H^2(R))} \right]. \end{aligned} \tag{3.17}$$

The linear interpolant $\bar{u}$ *of the midpoint values* $u^{k+1/2}$ *satisfies additionally the regularity estimate*

$$\|\bar{u}\|_{L^\infty(I;H^3(R))} \le C_5(c_0 + \|f_0\|_{L^\infty(I;H^1(R))} + \sup_k \|A^k\|_{H^2(R)}), \tag{3.18}$$

where c_0 *denotes the right-hand side of* (3.17).

Proof. The assumptions on A are stronger than those in Lemma 3.4. In particular, we know that a unique discrete solution exists on a small time interval and that it satisfies the estimate (3.12).

Part I. Time derivatives. We introduce discrete derivatives

$$\tilde{u}^k := \frac{u^k - u^{k-1}}{\Delta t}, \qquad \tilde{h}^k := \frac{h^k - h^{k-1}}{\Delta t} \tag{3.19}$$

for all $k = 1, \dots, K$. We now use the definition of (u^k, h^k) in (3.4)–(3.6). Taking the equations with index k and subtracting the equations with index $k-1$ yields for the new functions the following set of equations:

$$\begin{aligned} \frac{\tilde{u}^{k+1} - \tilde{u}^k}{\Delta t} &= \nabla \cdot A^k \nabla \frac{\tilde{u}^k + \tilde{u}^{k+1}}{2} + \frac{f_0^k - f_0^{k-1}}{\Delta t} \\ &\quad + \nabla \cdot \left(\frac{A^k - A^{k-1}}{\Delta t} \nabla \frac{u^{k-1} + u^k}{2} \right), \end{aligned} \tag{3.20}$$

$$\begin{aligned} \frac{\tilde{h}^{k+1} - \tilde{h}^k}{\Delta t} &= -a^k \cdot \nabla \left(\frac{\tilde{u}^k + \tilde{u}^{k+1}}{2} \right)(.,1) + \frac{f_1^k - f_1^{k-1}}{\Delta t} \\ &\quad - \frac{a^k - a^{k-1}}{\Delta t} \cdot \nabla \left(\frac{u^{k-1} + u^k}{2} \right)(.,1), \end{aligned} \tag{3.21}$$

$$\tilde{u}^{k+1}(.,1) = -\underline{\Delta}\tilde{h}^{k+1} + \frac{f_2^{k+1} - f_2^k}{\Delta t}. \tag{3.22}$$

We read these equations as follows: $(\tilde u^k, \tilde h^k)_{k=1,\dots,K}$ is a solution of the scheme (CN) of Definition 3.1 with initial values $(\tilde u^1, \tilde h^1)$. The right-hand side is

$$\begin{aligned}
\tilde f_0^k &:= \frac{f_0^k - f_0^{k-1}}{\Delta t} + \nabla \cdot \left(\frac{A^k - A^{k-1}}{\Delta t} \nabla \frac{u^k + u^{k-1}}{2} \right), \\
\tilde f_1^k &:= \frac{f_1^k - f_1^{k-1}}{\Delta t} - \frac{a^k - a^{k-1}}{\Delta t} \cdot \nabla \left(\frac{u^k + u^{k-1}}{2} \right) (.,1), \\
\tilde f_2^k &:= \frac{f_2^k - f_2^{k-1}}{\Delta t}
\end{aligned}$$

for $k = 1, \dots, K$. We next apply Lemma 3.4 on the sequence $(\tilde u^k, \tilde h^k)_k$. We recall the notation $(\tilde u, \tilde h)$ for the linear interpolant of $(\tilde u^k, \tilde h^k)_k$ and introduce $\tilde f_i$ for the linear interpolant of $(\tilde f_i^k)_k$. Note that the domain of definition is $(\Delta t, T)$; on this time interval we have by Lemma 3.4

$$\|(\tilde u, \tilde h)\|_Y \le c_0 \|\tilde u^1\|_{H^1(R)} + c_1 \left[\|\tilde f_0\|_{X_0} + \|\tilde f_1\|_{X_1} + \|\tilde f_2\|_{X_2} \right]. \tag{3.23}$$

The discrete time derivatives $\bar\partial_t f_i$ of f_i enter the bound (3.17) explicitly. It remains to estimate the contributions

$$\begin{aligned}
&\nabla \cdot \left(\frac{A^k - A^{k-1}}{\Delta t} \nabla \frac{u^k + u^{k-1}}{2} \right) \in X_0, \\
&\frac{a^k - a^{k-1}}{\Delta t} \cdot \nabla \left(\frac{u^k + u^{k-1}}{2} \right) (.,1) \in X_1.
\end{aligned}$$

We find $c > 0$ such that

$$\begin{aligned}
&\sum_k \left\| \nabla \cdot \left(\frac{A^k - A^{k-1}}{\Delta t} \nabla \frac{u^k + u^{k-1}}{2} \right) \right\|_{L^2(R)}^2 \Delta t \\
&\quad \le c \sum_k \left\| \frac{A^k - A^{k-1}}{\Delta t} \right\|_{H^1(R)}^2 \Delta t \cdot \sup_k \left\| \nabla \frac{u^k + u^{k-1}}{2} \right\|_{H^2(R)}^2, \\
&\sum_k \left\| \frac{a^k - a^{k-1}}{\Delta t} \cdot \nabla \left(\frac{u^k + u^{k-1}}{2} \right) (.,1) \right\|_{H^1(S)}^2 \Delta t \\
&\quad \le c \sum_k \left\| \frac{a^k - a^{k-1}}{\Delta t} (.,1) \right\|_{H^1(S)}^2 \Delta t \cdot \sup_k \left\| \nabla \frac{u^k + u^{k-1}}{2} (.,1) \right\|_{H^1(S)}^2.
\end{aligned}$$

With the assumptions on A and f_i Lemma 3.4 yields for $\bar\partial_t (u,h) = (\tilde u, \tilde h)$ the estimate

$$\|\bar\partial_t (u,h)\|_Y \le c_0 \, \|\tilde u^1\|_{H^1(R)} + c_1 \delta \, \sup_k \left\| \frac{u^k + u^{k-1}}{2} \right\|_{H^3(R)} + c_2, \tag{3.24}$$

where c_2 depends only on the norms of $\bar\partial_t f_i$.

In order to treat the second term on the right-hand side we now show estimate (3.18). This is done with the help of the original equation (3.4). The elliptic equation with the boundary condition (3.5) yields the estimate

$$\begin{aligned}
\left\| u^{k+1/2} \right\|_{H^3(R)} \le C \big(&\|f_0^k\|_{H^1(R)} + \|f_1^k\|_{H^{3/2}(S)} \\
&+ \|\bar\partial_t (u,h)\|_Y + \|A^k\|_{H^2(R)} \big).
\end{aligned}$$

Here the norm of f_1 is controlled by the right-hand side of (3.17). We have $\bar\partial_t f_1$ bounded in $L^2(0,T;H^1(S))$ and f_1 bounded in $L^2(0,T;H^3(S))$. An interpolation yields an estimate in $L^\infty(0,T;H^{3/2}(S))$ for f_1,

$$\sup_k \|f_1^k\|_{H^{3/2}(S)} \le C\left\{\|\partial_t f_1\|_{X_1} + \|\partial_x^2 f_1\|_{X_1}\right\}.$$

Equation (3.18) is shown.

We now insert (3.18) into estimate (3.24) and find with new constants c_0, c_1, and c_2,

$$\begin{aligned}\|\bar\partial_t(u,h)\|_Y \le c_0\ &\|\tilde u^1\|_{H^1(R)}\\ &+ c_1\delta\left(\|f_0\|_{L^\infty(I;H^1(R))} + \|A\|_{L^\infty(I;H^2(R))}\right) + c_2,\end{aligned} \tag{3.25}$$

where c_2 depends only on the norms of $\bar\partial_t f_i$.

Part II. Spatial derivatives. Estimate (3.25) does not suffice for the analysis of the nonlinear problem. Note that the best spatial estimate for the boundary so far is $h \in C^\alpha(I;H^{2+1/2})$. We next want to derive an estimate for $h \in L^\infty(I;H^{4+1/2}(S))$ to have good control of the regularity of the boundary. This estimate could be derived from an estimate for $u \in L^\infty(I;H^3(R))$. A similar estimate does appear in (3.18) but only for interpolants of $\frac12(u^k+u^{k+1})$ and not for interpolants of u^k. In order to derive the regularity estimate on h we perform an analysis of second spatial derivatives of the semidiscrete solution. While we used discrete derivatives in Part I we can now use classical derivatives. We introduce

$$\hat u^k := \partial_x^2 u^k, \qquad \hat h^k := \partial_x^2 h^k. \tag{3.26}$$

As in Part I we will use the fact that $(\hat u^k, \hat h^k)_k$ is a solution of scheme (CN) for an appropriate right-hand side. To be precise, $(\hat u^k, \hat h^k)_k$ satisfies (3.4)–(3.6) with $(f_0^k, f_1^k, f_2^k)_k$ replaced by $(\hat f_0^k, \hat f_1^k, \hat f_2^k)_k$, defined by

$$\begin{aligned}\hat f_0^k &:= \partial_x^2 f_0^k + \nabla\cdot\left(\left[\partial_x^2, A^k\right]\nabla\frac{u^k+u^{k+1}}{2}\right),\\ &\left[\partial_x^2, A^k\right]w = (\partial_x^2 A^k)w + 2(\partial_x A^k)\partial_x w \quad \forall w,\end{aligned} \tag{3.27}$$

$$\begin{aligned}\hat f_1^k &:= \partial_x^2 f_1^k - \left[\partial_x^2, a^k\right]\cdot\nabla\frac{u^k+u^{k+1}}{2},\\ &\left[\partial_x^2, a^k\right]\cdot w = (\partial_x^2 a^k)w + 2(\partial_x a^k)\partial_x w \quad \forall w,\end{aligned} \tag{3.28}$$

$$\begin{aligned}\hat f_2^k &:= \partial_x^2 f_2^k - \partial_x\left(\left[\partial_x^2, \gamma_0\right]\partial_x h^k\right),\\ &\left[\partial_x^2, \gamma_0\right]w = \left(\partial_x^2\gamma_0\right)w + 2\partial_x\gamma_0\partial_x w \quad \forall w,\end{aligned} \tag{3.29}$$

where we introduced the abbreviation

$$\gamma_0 = \frac{1}{\sqrt{1+|h_0'|^2}^3}.$$

We now use Lemma 3.4. With the notation $I_K = \{0,\dots,K\}$ and $\bar u^k := \frac{u^k+u^{k+1}}{2}$ we have to show estimates for

$$\begin{aligned}&\nabla\partial_x^2 A^k\cdot\nabla\bar u^k,\ \nabla\partial_x A^k\cdot\nabla\partial_x\bar u^k,\\ &\qquad\partial_x^2 A^k\cdot\Delta\bar u^k,\ \partial_x A^k\cdot\Delta\partial_x\bar u^k \in l^2(I_K;L^2(R)),\\ &\partial_x^2 a^k\cdot\nabla\bar u^k(.,1),\ \partial_x a^k\cdot\nabla\partial_x\bar u^k(.,1) \in l^2(I_K;H^1(S)),\end{aligned}$$

and additionally estimates in $l^2(I_K; H^{-1/2}(S))$ for the discrete time derivatives of the functions

$$\partial_x^4 h_0 \cdot \partial_x h^k,\ \partial_x^3 h_0 \cdot \partial_x^2 h^k,\ \partial_x^2 h_0 \cdot \partial_x^3 h^k.$$

On the functions of the last line we additionally have to give an estimate in $l^\infty(I_K; H^{1/2}(S))$ or we use the improved version of estimate (3.12). We use the latter and see from the original equation (3.6) for h^{k+1} that we can use $C_\star = C\|f_2^k\|_{H^{2+1/2}(S)} \leq C(\|\partial_x^2 f_2\|_{X_2} + \|\partial_t f_2\|_{X_2})$, where C depends only on $\|h_0\|_{H^{4+1/2}(S)}$. All the above error terms can be estimated by a small multiple of the solution norm in (3.17). While the other terms can be estimated directly, the most intricate term is the one containing second derivatives of the trace of derivatives of $\bar{u}^k$. It suffices to estimate for the interpolation

$$\partial_x^2 \nabla \bar{u}(.,1) \in L^2(I; L^2(S))$$

by the norms of u and f in (3.17). This estimate can be derived from (3.4) if we differentiate that equation twice with respect to x. We use $\partial_t \partial_x^2 u, \partial_x^2 f_0 \in L^2(I; L^2(R))$, and, for the boundary condition, $\partial_t \partial_x^2 h \in L^2(I; H^1(S))$. Elliptic theory yields $\partial_x^2 \bar{u} \in L^2(I; H^2(R))$ and therefore the result.

We can now apply Lemma 3.4 which yields the Y-estimates for $\partial_x^2(u,h)$. The compatibility condition ((3.6) is satisfied for $k=-1$) holds, since we took only second derivatives on both sides. Note that without the estimates of the time derivative we could not have derived the spatial estimates on $\bar{u}$ but only estimates on higher x-derivatives of u.

Part III. The first time step. It remains to control the first discrete time derivative $\tilde{u}^1 \in H^1(R)$ of (3.25) by the first term on the right-hand side of (3.17). This is done in the subsequent lemma which concludes the proof of the proposition. ∎

ASSUMPTION 3.6. *We assume that $A = A(0)$ is a Sym$(\mathbb{R}^2)$-valued function of class $H^3(R)$, sufficiently close to the identity in $L^\infty(R)$.*

The compatibility conditions for the discrete scheme read

$$u_0(.,1) = -\underline{\Delta} h_0 + f_2^0, \tag{3.30}$$

$$(\nabla \cdot A\nabla u_0 + f_0^0)(.,1) = \underline{\Delta}(a \cdot \nabla u_0(.,1) - f_1^0) + \frac{f_2^1 - f_2^0}{\Delta t}. \tag{3.31}$$

LEMMA 3.7. *Let Assumption 3.6 be satisfied. Then the solution (u^1, h^1) for the first time step in scheme (CN) satisfies*

$$\left\| \frac{u^1 - u^0}{\Delta t} \right\|_{H^1(R)} \leq C\ \|\nabla \cdot A\nabla u_0 + f_0\|_{H^1(R)} \tag{3.32}$$

with C independent of Δt.

Proof. We write $\underline{\Delta} = \partial_x(\gamma_0 \partial_x)$ with γ_0 close to 1 in $L^\infty(S)$. We use $a = e_2 \cdot A$ and study the operator

$$B : \begin{pmatrix} u \\ h \end{pmatrix} \mapsto \begin{pmatrix} \nabla \cdot A\nabla u \\ -a \cdot \nabla u(.,1) \end{pmatrix}$$

defined on

$$D(B) := \left\{ (u,h) \in X_0 | u \in H^2(R), u(.,1) = -\underline{\Delta} h, u(.,0) = 0 \right\},$$

a subset of the space

$$X_0 := \left\{ (u,h) | \int_0^1 h = 0 \right\} \subset X := L^2(R) \times H^1(S).$$

On X_0 we use the scalar product

$$\left\langle \begin{pmatrix} u \\ h \end{pmatrix}, \begin{pmatrix} \hat{u} \\ \hat{h} \end{pmatrix} \right\rangle := \int_R u \cdot \hat{u} + \int_S \gamma_0 \, \partial_x h \cdot \partial_x \hat{h}.$$

Then the operator B is densely defined in X_0, it has a compact inverse by Lemma 3.2, and it is symmetric. By the spectral theorem we find a complete set of eigenfunctions (σ^j, η^j) of B; that is,

$$\lambda_j \sigma^j - \nabla \cdot A \nabla \sigma^j = 0, \tag{3.33}$$

$$\lambda_j \eta^j + a \cdot \nabla \sigma^j(.,1) = 0, \tag{3.34}$$

$$\sigma^j(.,1) + \underline{\Delta} \eta^j = 0. \tag{3.35}$$

In order to have a basis (σ^j, η^j) of X (and not only on X_0) we extend the basis by eigenfunctions of the form $\sigma(x,y) = U(y)$, $h(x) = 1$.

The functions (σ^j, η^j) can be normalized such that

$$\int_R \sigma^j \cdot \sigma^l + \int_S \gamma_0 \, \partial_x \eta^j \partial_x \eta^l = \delta_{jl}. \tag{3.36}$$

Furthermore, one verifies that all eigenvalues are negative, and orthogonality also holds with the scalar product

$$\int_R A \nabla \sigma^j \cdot \nabla \sigma^l = -\lambda_j \delta_{jl}. \tag{3.37}$$

This scalar product defines a norm equivalent to the H^1-norm by the Poincaré inequality. We denote the Hilbert space corresponding to the product $(v,w) \mapsto \int_R Av \cdot w$ in the following by L^2_A. We next consider pairs $(u,h) = \sum_{j=1}^{\infty} c_j (\sigma^j, \eta^j)$. For $(u,h) \in D(B)$ we can conclude with $u_N := \sum_{j=1}^{N} c_j \sigma^j$ that $\|\nabla u_N\|_{L^2_A} \le \|\nabla u\|_{L^2_A}$ and

$$\left\| \nabla \sum_j^{\infty} c_j \sigma^j \right\|^2_{L^2_A(R)} = \sum_j^{\infty} |c_j|^2 \, |\lambda_j|. \tag{3.38}$$

In particular, if one side in this equality is finite, then the other is also finite.

We now expand the initial values and the right-hand side in terms of eigenfunctions and write

$$(u^0, h^0 - \underline{\Delta}^{-1} f_2^0) = \sum_j a_j (\sigma^j, \eta^j), \quad (u^1, h^1 - \underline{\Delta}^{-1} f_2^1) = \sum_j b_j (\sigma^j, \eta^j),$$

$$\left(f_0^0, f_1^0 - \underline{\Delta}^{-1} \frac{f_2^1 - f_2^0}{\Delta t} \right) = \sum_j d_j (\sigma^j, \eta^j).$$

Here $\underline{\Delta}^{-1}$ denotes any right inverse of $\underline{\Delta}$. Equations (3.4), (3.5) for the first time step translate into

$$\frac{b_j - a_j}{\Delta t} = \lambda_j \frac{b_j + a_j}{2} + d_j \qquad \forall j \in \mathbb{N}. \tag{3.39}$$

We find

$$b_j = \frac{1}{\frac{1}{\Delta t} - \frac{1}{2}\lambda_j} \left(a_j \left[\frac{1}{\Delta t} + \frac{1}{2}\lambda_j \right] + d_j \right).$$

Therefore

$$\frac{b_j - a_j}{\Delta t} = \frac{\lambda_j a_j + d_j}{1 - \frac{1}{2}\lambda_j \Delta t}. \tag{3.40}$$

We have to estimate the H^1-norm of the function $\sum_j \frac{b_j - a_j}{\Delta t}\sigma_j$ by the H^1-norm of the function $\sum_j (\lambda_j a_j + d_j)\sigma^j$. We use (3.38) for the following two pairs that are both in $D(B)$ by the compatibility assumption:

$$\left(u^1 - u^0, h^1 - h^0 - \underline{\Delta}^{-1}(f_2^1 - f_2^0)\right),$$

$$\left(\nabla \cdot A\nabla u_0 + f_0^0, -a \cdot \nabla u_0(., 1) + f_1^0 - \underline{\Delta}^{-1}\frac{f_2^1 - f_2^0}{\Delta t}\right).$$

We can calculate

$$\begin{aligned}
\left\| \nabla \sum_j \frac{b_j - a_j}{\Delta t}\sigma_j \right\|^2_{L^2_A(R)} &= \sum_j |\lambda_j| \left| \frac{b_j - a_j}{\Delta t} \right|^2 \\
&\leq \sum_j |\lambda_j| \, |\lambda_j a_j + d_j|^2 = \left\| \nabla \sum_j (\lambda_j a_j + d_j)\sigma_j \right\|^2_{L^2_A(R)} \\
&\leq C \|\nabla \cdot A\nabla u_0 + f_0^0\|^2_{H^1}.
\end{aligned}$$

This concludes the proof. ∎

4. A discretization of the transformed equations. We perform some elementary calculations for the transformation of (2.1)–(2.3) onto a reference domain. Our aim is to replace the temperature $\Theta(t) : \Omega_t \to \mathbb{R}$ by the new unknown $u(t) : R \to \mathbb{R}$. We denote the upper boundary of Ω by Γ and the upper boundary of R by $\Gamma_R = \{(x, 1) : x \in S\}$. Given a domain transformation $\Psi : \Omega \to R$ we use $u \circ \Psi = \Theta$ and, in the calculation below, also $v \circ \Psi = \varphi$. We define

$$B_{ij} := \nabla_j \Psi_i, \quad J := \det(B)^{-1}, \quad A := J \cdot B \cdot B^t. \tag{4.1}$$

We see that the equation

$$\int_\Omega \nabla\Theta \cdot \nabla\varphi + \int_\Omega f \circ \Psi\varphi - \int_\Gamma g \circ \Psi\varphi = 0 \qquad \forall \varphi \in C^1(\Omega)$$

transforms into

$$\begin{aligned}
&\int_R (B^t \cdot \nabla u) \cdot (B^t \cdot \nabla v) \; J + \int_R fv \; J \\
&\qquad - \int_{\Gamma_R} gv\sqrt{1 + |\partial_x h|^2} = 0 \qquad \forall v \in C^1(R).
\end{aligned}$$

We conclude that the equation

$$\Delta\Theta = f \circ \Psi \quad \text{in } \Omega, \quad n \cdot \nabla\Theta = g \circ \Psi \quad \text{on } \Gamma$$

transforms into

$$\nabla \cdot A\nabla u = J\ f \quad \text{in } R, \quad e_2 \cdot A\nabla u = g\sqrt{1+|h'|^2} \quad \text{on } \Gamma_R.$$

Therefore the physical equations (2.1)–(2.3) transform into

$$J\partial_t u + J\partial_t \Psi \cdot \nabla u = \nabla \cdot A\nabla u, \tag{4.2}$$

$$\partial_t h = -e_2 \cdot A\nabla u, \tag{4.3}$$

$$u|_h + \underline{\Delta} h = \underline{\Delta} h + \kappa(h). \tag{4.4}$$

The equations formally coincide with (3.1)–(3.3) if we set

$$f_0 := (1-J)\partial_t u - J\ (\partial_t \Psi) \cdot \nabla u, \tag{4.5}$$

$$f_1 := 0, \qquad f_2 := \underline{\Delta} h + \kappa(h). \tag{4.6}$$

We now want to choose a discretization of (4.2)–(4.4). The idea is to define matrices A^k as in (4.1) and to define f_i^k as in (4.5), (4.6). In order to proceed we have to define domain transformations $\Psi^k : \Omega^k \to R$ that we can insert in (4.1). We define Ψ^k as the inverse of functions $X^k : R \to \Omega^k$ with

$$X^{k+1}(x,y) - X^k(x,y) = \frac{X_2^k(x,y)}{h^k(x)}(h^{k+1}(x) - h^k(x))e_2.$$

We choose an initial parametrization X^0 as in Theorem 2.2.

To discretize formula (4.5) we have to discretize $\partial_t \Psi$. Since the definition of X is consistent with the continuous equation

$$\partial_t \Psi^{-1}(t,x,y) = \frac{(\Psi^{-1})_2(x,y)}{h(x)}\ \partial_t h(t,x)\ e_2,$$

we find from $\partial_t(\Psi \circ \Psi^{-1}) = 0$ the continuous equation

$$\partial_t \Psi(t,\xi,\zeta) = -\partial_\zeta \Psi \cdot \frac{\zeta}{h}\partial_t h(t,\xi).$$

Because of $J = (\partial_\zeta \Psi_2)^{-1}$ the right-hand side of the discrete scheme can be defined consistently by

$$\begin{aligned} f_0^k &:= (1-J^k)\frac{u^{k+1}-u^k}{\Delta t} + \frac{X_2^k}{h^k}\ \frac{h^k - h^{k-1}}{\Delta t}\partial_y \frac{u^k + u^{k+1}}{2}, \\ f_1^k &:= 0, \qquad f_2^{k+1} := \underline{\Delta} h^k + \kappa(h^k). \end{aligned} \tag{4.7}$$

In the definition of f_0^0, the first time step, we insert the formal time derivative of h instead of $\frac{h^0 - h^{-1}}{\Delta t}$, and we use u_0 instead of $\frac{u^0+u^1}{2}$. To have f_2 defined on the whole time interval we set $f_2^0 = \underline{\Delta} h_0 + \kappa(h_0) \equiv f_2^1$. This defines a discrete scheme that is consistent with (4.2)–(4.4). Note that in the above definition f_0^k depends on u^{k+1}.

An assumption concerning the compatibility of the initial values will be needed. This is not an artifact of the discretization—the same is true for the continuous

equations. Let (u,h) be a classical solution such that $\partial_t(u,h)$ is continuous in $t=0$. We conclude that the formal time derivative $\tilde{\partial}_t(u,h)$ defined by (4.2) and (4.3) must satisfy on the boundary the time derivative of (4.4). We will therefore use later on the following assumption.

ASSUMPTION 4.1. *We say that the compatibility conditions for the continuous equations are satisfied if for $u_0 \in H^3(R)$ the formal time derivative $\tilde{\partial}_t u_0$ is in $H^1(R)$ and*

$$u_0(.,1) - \kappa(h_0) = 0, \tag{4.8}$$

$$\tilde{\partial}_t u_0(.,1) + \underline{\Delta}\tilde{\partial}_t h_0 = 0. \tag{4.9}$$

THEOREM 4.2. *Let the initial values (u_0,h_0) satisfy the compatibility condition of Assumption 4.1 and let $h_0 - 1$ be small in $C^{0,1}(S)$. We consider scheme (CN) with f_i^k as in (4.7) and A^k defined by (4.1).*

Then there exists $T > 0$ such that the scheme (CN) has a unique solution $(u^k,h^k)_k$. The linear interpolants (u,h) of $(u^k,h^k)_k$ and $\bar{u}$ of $\frac{1}{2}(u^k+u^{k+1})$ satisfy the estimate

$$\|\bar{\partial}_t(u,h)\|_Y + \|\partial_x^2(u,h)\|_Y + \|\bar{u}\|_{L^\infty(I;H^3(R))} \le C, \tag{4.10}$$

where C and T depend only on the norm of the initial values and are independent of Δt.

Proof. The proof is given in three parts (A)–(C). Part (A) is concerned with the initial values and their compatibility. In part (B) the crucial estimates on solutions are derived with the help of Proposition 3.5 on the scheme (CN) with a fixed right-hand side. In part (C) we show the existence of a bounded solution.

(A) Compatibility of initial values. We want to use Proposition 3.5. In order to do so, we have to guarantee that the compatibility assumption, Assumption 3.6, is satisfied. By definition of f_2^0, (3.30) holds. Concerning (3.31) we observe that $f_2^1 - f_2^0 = 0$. In the above scheme the time derivative $\frac{u^1-u^0}{\Delta t}$ appears in f_0^0. This in general changes the compatibility condition for the scheme. However, our construction imposed $J^0 = 1$ on the upper boundary and therefore

$$f_0^0(.,1) = -\tilde{\partial}_t h(0)\ \partial_y u^0(.,1).$$

Then the discrete compatibility assumption (3.31) coincides with the continuous version (4.9).

(B) Improvement of a priori bounds. This part of the proof is based on estimate (3.17). We use the constant C_1 and the first term of the right-hand side of that estimate and define

$$C_0 := 2C_1\left[\|\nabla\cdot A(0)\nabla u_0 + f_0(0)\|_{H^1(R)} + \|\partial_x^2 u_0\|_{H^1(R)}\right].$$

We will show that given $\delta > 0$ we can choose a small $T > 0$ and a small a priori bound for $\|h_0 - 1\|_{C^{0,1}(S)}$ such that for every solution (u,h)

$$\begin{aligned} &\|\bar{\partial}_t(u,h)\|_Y + \|\partial_x^2(u,h)\|_Y \le 2C_0 \\ \Rightarrow &\|\bar{\partial}_t(u,h)\|_Y + \|\partial_x^2(u,h)\|_Y \le C_0. \end{aligned} \tag{4.11}$$

This is shown in four steps. With a constant C independent of δ and T there holds the following:

1. $\bar{u} \in L^\infty(I;H^2(R))$ is bounded by C.
2. The coefficients A^k defined by (4.1) satisfy $(3.16)_\delta$.

3. The norms of $\bar{\partial}_t f$ and of $\partial_x^2 f$ on the right-hand side of (3.17) are bounded by $C\delta$.
4. The norms of $f_0 \in L^\infty(I; H^1(R))$ and of $A \in L^\infty(I; H^2(R))$ are bounded by C.

Once we have shown 1–4, we can choose a new $\delta > 0$ and $T > 0$ and use Proposition 3.5 to obtain the implication (4.11).

Now consider a solution (u, h) with the bound $2C_0$ as in (4.11).

1. Regularity of $\bar{u}$. The function f_0 is bounded in $L^\infty(I; L^2(R))$ (see below). We use the elliptic equation (3.4) for $u^{k+1/2}$:

$$\nabla \cdot A^k \nabla u^{k+1/2} = \frac{u^{k+1} - u^k}{\Delta t} - f_0^k \in L^2(R).$$

The boundary condition (3.5) is smooth enough to imply the desired estimate for $\sup_k \|u^{k+1/2}\|_{H^2(R)}$.

2. Estimates for A. By an interpolation we see that for some $\alpha > 0$ the function h is also bounded as

$$h \in C^\alpha(I; H^4(S)).$$

Then the matrix $B = \nabla\Psi$ satisfies

$$B \in C^\alpha(I; H^3(R)), \qquad B(.,1) \in C^\alpha(I; H^3(S)).$$

Since $H^3(R)$ is an algebra (see, e.g., [2]), the matrix A satisfies estimates in the same spaces. Choosing T small we immediately infer the first two lines in (3.16).

In order to verify the third line we again use an interpolation: with $p > 2$ we find an estimate for

$$\partial_t h \in L^p(I; H^3(S)).$$

This implies an estimate for

$$\partial_t B \in L^p(I; H^2(R)).$$

Again, $\partial_t A$ satisfies estimates in the same space. If necessary we choose a smaller T in order to infer the third line in (3.16).

We turn to the estimates for f_i. The function f_1 vanishes identically, and all estimates are trivial.

3. and 4. Estimates for f_0. We first consider the term $(1 - J)\partial_t u$. The factor $(1 - J)$ is small in $L^\infty(I; L^\infty(R))$ by smallness of h_0 in $C^{0,1}(S)$. We use

$$\begin{aligned}
&\partial_t u \in Y_u \Rightarrow \partial_t^2 u \in L^2L^2 \Rightarrow \partial_t[(1 - J)\partial_t u] \in L^2L^2,\\
&\partial_x^2 u \in Y_u \Rightarrow \partial_t\partial_x^2 u \in L^2L^2 \Rightarrow \partial_x^2[(1 - J)\partial_t u] \in L^2L^2,\\
&\partial_t u \in Y_u \Rightarrow \partial_t u \in L^\infty H^1 \Rightarrow [(1 - J)\partial_t u] \in L^\infty H^1.
\end{aligned}$$

These implications together with their corresponding estimates give the desired estimate for the first term in f_0. Note that the smallness of, e.g., $\partial_t[(1 - J)\partial_t u] = (1 - J)\partial_t^2 u - \partial_t J\partial_t u$ follows for the first term by smallness of $1 - J$, for the second term by a compactness argument: $\partial_t J \in L^\infty H^{3/2}$ and $\partial_t u \in L^\infty H^1$ imply (for small T) smallness of the second term in L^2L^2. The estimate of $(1 - J)\partial_t^2 u$ is the only place where we use the smallness of h_0.

The other term of f_0 has the regularity properties of $\tilde\partial_t\Psi\cdot\nabla\bar u$. We use step 1 with the estimate for $\bar u\in L^\infty H^2$. It yields

$$\tilde\partial_t\Psi\in L^\infty(I;H^{2+1/2}(R)),\quad \nabla\bar u\in L^\infty(I;H^1(R)),$$

and we find the estimate for $f_0\in L^\infty(I;H^1(R))$. The estimates for $\partial_t f_0$ and $\partial_x^2 f_0$ are direct. Smallness of the $L^2(I)$-norms follows by the compactness argument.

Estimates for f_2. Concerning f_2 we have to take special care of the first time step. However, let us first consider f_2 as defined by $f_2^1,\dots,f_2^K$: the functions f_2^k are composed from first and second derivatives of h. Remember that the operator $-\underline\Delta h$ is the linearization of the mean curvature $\kappa(h)$ in $h=h_0$. By the $2C_0$-bound of (4.11) we can estimate the differences $\partial_x h-\partial_x h_0$ pointwise by a small number (depending on T). Then f_2 has the form $f_2=-\kappa(h)-\underline\Delta h=G(\partial_x h,\partial_x h_0)\cdot(1,\partial_x^2 h)$ with $G(0,0)=0$ and G differentiable. We find the estimate

$$\|f_2\|\le C\,\varepsilon\,\|h\|,$$

where the norms are those of (3.17) and of (4.11), and ε is arbitrarily small for T small.

Let us now consider the first time step. $f_2^0\in H^{2+1/2}(S)$ by Assumption 4.1. There holds $f_2^1-f_2^0=0$, and we find the estimate for the first discrete time derivative of f_2. The second discrete time derivative is

$$\bar\partial_t^2 f_2(0):=\frac{f_2^2-2f_2^1+f_2^0}{(\Delta t)^2}=\frac{f_2^2-f_2^1}{(\Delta t)^2}=\frac{\kappa(h^1)+\underline\Delta h^1-\kappa(h^0)-\underline\Delta h^0}{(\Delta t)^2}.$$

We introduce $T[\partial_x h]:=\frac{\partial_x h}{\sqrt{1+|\partial_x h|^2}}$ to write

$$\bar\partial_t^2 f_2(0)=-\frac{1}{(\Delta t)^2}\partial_x\left(T[\partial_x h^1]-T[\partial_x h^0]-T'[\partial_x h^0]\cdot\partial_x(h^1-h^0)\right).$$

We find

$$\|\bar\partial_t^2 f_2(0)\|_{H^{-1}(S)}\le C\left\|\frac{\partial_x(h^1-h^0)}{\Delta t}\right\|^2_{L^\infty(S)}\le C\left\|\frac{u^1-u^0}{\Delta t}\right\|^2_{H^1(R)}.$$

(C) Existence of a solution—the continuity argument. Note that a time step of scheme (CN) with f as in (4.7) is still a linear equation for (u^{k+1},h^{k+1}). We see that the single time step can always be solved as long as $1-J^k$ and $\frac{X_2^k}{h^k}$ $\frac{h^k-h^{k-1}}{\Delta t}$ are small in L^∞. Still, it could happen that on the time interval $(0,t^k)$ the solution has norm less than C_0 and on the time interval $(0,t^{k+1})$ the solution has a norm larger than $2C_0$. We will show that this cannot happen.

We connect the initial values (u_0,h_0) with a continuous path $(u_\lambda,h_\lambda)_{\lambda\in[0,1]}$ with the trivial initial values $(u_1,h_1)=0$. This can be done in such a way that (u_λ,h_λ) satisfies the compatibility condition for all $\lambda\in[0,1]$. If scheme (CN) with f as in (4.7) and with initial values (u_λ,h_λ) has a solution on $I=(0,T)$ we denote this solution by (u^λ,h^λ). This family of solutions has the following two properties.

1. Every weak limit $\lim_{\lambda\to\lambda_0}(u^\lambda,h^\lambda)$ in the topology of (4.11) of bounded solutions is again a bounded solution. This follows immediately, since we can take the limit in all equations.

2. If $(u^{\lambda_0}, h^{\lambda_0})$ is a solution, bounded by C_0, then also in a neighborhood $(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon)$ of λ_0 there exist solutions that are bounded by C_0. This follows because we deal with a fixed (finite) number of time steps. The norm of the solution depends continuously on λ. In general the norm might exceed the value C_0, but we can achieve that it does not exceed $2C_0$. Now property (4.11) ensures that the norm remains bounded by C_0.

We combine the above facts 1 and 2 to conclude. The set

$$\{\lambda \in [0,1] | \text{ a solution } (u,h)_\lambda \text{ exists and } \|(u,h)_\lambda\| \le C_0\}$$

is a nonempty ($\lambda = 1$ is in the set), closed (by property 1), and open (by property 2) subset of $[0,1]$. Therefore $\lambda = 0$ is in the above set, and therefore a solution (u^0, h^0) to initial values (u_0, h_0) exists and satisfies the estimate. This concludes the proof of the theorem. ∎

COROLLARY 4.3. *Let h_0, u_0, and $T > 0$ as in the last theorem. Then, for a subsequence $\Delta t \to 0$, the solutions (u^k, h^k) converge to solutions of (4.2)–(4.4). In particular, (2.1)–(2.3) with compatible initial conditions possess a solution on a short-time interval.*

Proof. By the above theorem the solutions $(u,h)_{\Delta t}$ of the discrete problems are uniformly bounded. Therefore there exists a subsequence with a weak limit (u,h). The convergence is strong for $u \in L^2(I; H^2(R)) \cap H^1(I; L^2(R))$ and for $h \in L^2(I; H^5(R)) \cap H^1(I; H^3(R))$. Because of consistency in the definition of A and f we can conclude that (u,h) is a strong solution to the transformed equations (4.2)–(4.4). The transformed solution (Θ, h) is a solution of the original problem. ∎

5. Proof of Theorem 2.2. Theorem 4.2 yields a stable discretization of the original equations. The drawback for a use as a numerical scheme is the need to transform all equations onto a fixed domain. It is more natural to use the operator splitting scheme (OS). We will prove in this section the stability of scheme (OS) as it was stated in Theorem 2.2. The proof uses a transformation of the discrete scheme onto a fixed domain. It will turn out that scheme (OS) is in fact identical to the scheme (CN) of section 4.

ASSUMPTION 5.1. *Let n be the normal vector of the initial domain given by h_0. We introduce the formal time derivatives in $t = 0$ by*

$$\begin{aligned} \tilde{\partial}_t \Theta|_{t=0} &:= \Delta \Theta_0, \\ \tilde{\partial}_t h|_{t=0} &:= -n_2^{-1} \, (n \cdot \nabla \Theta_0) \circ H_0. \end{aligned}$$

We impose on the initial values the regularity $\tilde{\partial}_t \Theta|_{t=0} \in H^1(\Omega_0)$ and the compatibility conditions

$$\begin{aligned} \Theta \circ H_0 &= \kappa(h_0), \\ \tilde{\partial}_t \Theta|_{t=0}(x, h_0(x)) + \partial_2 \Theta_0(x, h_0(x)) \cdot \tilde{\partial}_t h|_{t=0}(x) &= D\kappa(h_0) \tilde{\partial}_t h|_{t=0}(x). \end{aligned}$$

Proof of Theorem 2.2. We introduce the following functions:

$$\begin{aligned} u^k &:= \Theta^k \circ X^k : R \to \mathbb{R}, \\ \bar{v}^k &:= v^k \circ X^k : R \to \{0\} \times \mathbb{R} \subset \mathbb{R}^2, \\ \tilde{u}^k &:= \tilde{\Theta}^k \circ X^k. \end{aligned}$$

We now interpret scheme (OS) as a scheme for (u^k, h^k). Step 3 of (OS) reads in the new notation

$$\tilde{u}^k = u^{k+1}.$$

We use this identity to write the equations of Step 2 in terms of u^{k+1}. We use the transformation of section 4 with corresponding B^k, A^k, J^k.

$$J^k \ \frac{u^{k+1}-u^k}{\Delta t} = \nabla \cdot \left(A^k \nabla \frac{u^k+u^{k+1}}{2} \right) + J^k \ \bar{v}^k \cdot (B^k)^t \cdot \nabla \frac{u^k+u^{k+1}}{2} \qquad \text{in } R, \tag{5.1}$$

$$\frac{h^{k+1}-h^k}{\Delta t} = -e_2 \cdot A^k \cdot \nabla \frac{u^k+u^{k+1}}{2}(.,1) \qquad \text{in } [0,1], \tag{5.2}$$

$$u^{k+1}(.,1) + \underline{\Delta} h^{k+1} = \kappa(h^k) + \underline{\Delta} h^k \qquad \text{in } [0,1]. \tag{5.3}$$

This is nothing but scheme (CN) with the right-hand side

$$f_0^k := (1-J^k) \ \frac{u^{k+1}-u^k}{\Delta t} - J^k \ \bar{v}^k \cdot (B^k)^t \cdot \nabla \frac{u^k+u^{k+1}}{2},$$
$$f_1^k := 0, \qquad f_2^{k+1} := \kappa(h^k) + \underline{\Delta} h^k,$$

where in the definition of f_0^0 the convective term is calculated explicitly. The scheme is identical to that of section 4, since

$$J^k \ e_2 \cdot (B^k)^t = e_2 \quad \text{and} \quad \bar{v}^k(x,y) = \frac{X_2^k(x,y)}{h^k(x)} \ \frac{h^k(x)-h^{k-1}(x)}{\Delta t} e_2.$$

Theorem 2.2 is a consequence of Theorem 4.2. □

Corollary 2.3 follows from the theorem just as Corollary 4.3 followed from Theorem 4.2. Let us demonstrate without referring to section 4 that the scheme is consistent. From (2.9) and (2.5) we conclude

$$\Theta^{k+1} \circ X^{k+1} - \Theta^k \circ X^k = \left(\Delta \frac{\Theta^k + \tilde{\Theta}^k}{2} \right) \circ X^k + \left(v \cdot \nabla \frac{\Theta^k + \tilde{\Theta}^k}{2} \right) \circ X^k.$$

In the limit $\Delta t \to 0$ we infer

$$\partial_t (\Theta \circ X) = (\Delta \Theta) \circ X + (v \cdot \nabla \Theta) \circ X.$$

This yields the original equation (2.1), since by definition of v in (2.4)

$$\nabla \Theta \cdot \partial_t X(x,y) = \nabla \Theta \cdot \left(\frac{X_2(x,y)}{h(x)} \partial_t h \right) \ e_2 = v \cdot \nabla \Theta.$$

REFERENCES

[1] F. Abergel, D. Hilhorst, F. Issard-Roch, and J. Scheid, *Local existence and uniqueness of a Stefan problem with surface tension*, Appl. Anal., 60 (1996), pp. 219–240.

[2] R.A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[3] E. Bänsch, *Numerical Methods for the Instationary Navier–Stokes Equations with a Free Capillary Surface*, Habilitationsschrift, Universität Freiburg, 1998.

[4] X. Chen and F. Reitich, *Local existence and uniqueness of solutions of the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[5] J. Crank and P. Nicolson, *A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type*, Proc. Cambridge Philos. Soc., 43 (1947), pp. 50–67.

[6] St. Luckhaus, *Solutions for the two-phase Stefan problem with the Gibbs-Thomson law for the melting temperature*, European J. Appl. Math., 1 (1990), pp. 101–111.

[7] A.M. Meirmanov, *The Stefan problem with surface tension in the three dimensional case with spherical symmetry: Nonexistence of the classical solution*, European J. Appl. Math., 5 (1994), pp. 1–19.

[8] E.V. Radkevich, *On conditions for the existence of a classical solution of the modified Stefan problem (the Gibbs-Thompson law)*, Math. USSR-Sb., 75 (1993), pp. 221–246.

[9] B. Schweizer, *Free boundary fluid systems in a semigroup approach and oscillatory behavior*, SIAM J. Math. Anal., 28 (1997), pp. 1135–1157.

[10] A. Veeser, *Error estimates for semi-discrete dendritic growth*, Interfaces Free Bound., 1 (1999), pp. 227–255.

[11] A. Visintin, *Models of Phase Transitions*, Progr. Nonlinear Differential Equations Appl. 28, Birkhäuser, Boston, 1996.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1207–1229

AN ADAPTIVE UZAWA FEM FOR THE STOKES PROBLEM: CONVERGENCE WITHOUT THE INF-SUP CONDITION*

EBERHARD BÄNSCH†, PEDRO MORIN‡, AND RICARDO H. NOCHETTO§

Abstract. We introduce and study an adaptive finite element method (FEM) for the Stokes system based on an Uzawa outer iteration to update the pressure and an elliptic adaptive inner iteration for velocity. We show linear convergence in terms of the outer iteration counter for the pairs of spaces consisting of continuous finite elements of degree k for velocity, whereas for pressure the elements can be either discontinuous of degree $k-1$ or continuous of degree $k-1$ and k. The popular Taylor–Hood family is the sole example of stable elements included in the theory, which in turn relies on the stability of the continuous problem and thus makes no use of the discrete inf-sup condition. We discuss the realization and complexity of the elliptic adaptive inner solver and provide consistent computational evidence that the resulting meshes are quasi-optimal.

Key words. a posteriori error estimators, adaptive mesh refinement, convergence, data oscillation, performance, quasi-optimal meshes

AMS subject classifications. 65N12, 65N15, 65N30, 65N50, 65Y20

PII. S0036142901392134

1. Introduction.

Adaptive finite element methods (FEM) have become essential tools in science and engineering for the numerical solution of multiscale phenomena governed by partial differential equations (PDE). We refer to [1, 20] for references on adaptivity and restrict the list of papers to those strictly related to our work.

Computational experience strongly suggests that, starting from a coarse mesh, adaptive algorithms converge within any prescribed tolerance in a finite number of steps, but their convergence for general—even linear—problems is largely an open question. This issue has been recently tackled for *elliptic* problems, in the multidimensional setting, by Morin, Nochetto, and Siebert [15, 16], exploiting an idea of Dörfler [11]. In [11, 15, 16], the fact that the elliptic operator is *positive definite* (or *coercive*) plays a fundamental role.

In this article we devise an adaptive finite element algorithm for the Stokes problem and prove its convergence. The essential difference with elliptic problems is that the Stokes operator is *not* positive definite but rather leads to a saddle-point problem. The role of coercivity is thus played by the weaker condition of sole invertibility given by the inf-sup condition (1.2) below.

*Received by the editors July 11, 2001; accepted for publication (in revised form) February 11, 2002; published electronically September 27, 2002. This research was partially supported by NSF-DAAD grant INT-9910086.

http://www.siam.org/journals/sinum/40-4/39213.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, 10117 Berlin, Germany and Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany (baensch@ wias-berlin.de).

‡Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, 3000 Sante Fe, Argentina and Instituto de Matemática Aplicada del Litoral (IMAL), Güemes 3450, 3000 Santa Fe, Argentina (pmorin@math.unl.edu.ar). The research of this author was partially supported by Programa FOMEC de la Universidad Nacional del Litoral and CONICET of Argentina and NSF grant DMS-9971450. This work was partly developed while this author was visiting the University of Maryland.

§Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu). The research of this author was partially supported by NSF grant DMS-9971450.

To be more specific, let Ω be a polygonal (polyhedral) domain in $\mathbb{R}^d$ for $d \geq 2$, and let $\mathbb{V} = \left(\mathring{H}^1(\Omega)\right)^d$ be the usual Sobolev space of vector-valued square integrable functions, having also square integrable first derivatives whose trace vanishes on $\partial\Omega$. Let $\mathbb{P} := \mathring{L}^2(\Omega)$ be the space of square integrable functions with mean value zero. Then, the weak form of the Stokes problem in its primitive variables reads as follows: Find a pair $(\mathbf{u}, p) \in \mathbb{V} \times \mathbb{P}$ such that for all $(\mathbf{v}, q) \in \mathbb{V} \times \mathbb{P}$,

$$\int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} - \int_\Omega p \, \operatorname{div} \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \quad \text{and} \quad \int_\Omega q \, \operatorname{div} \mathbf{u} = 0, \tag{1.1}$$

where throughout this paper we assume $\mathbf{f} \in \left(L^2(\Omega)\right)^d$. To avoid confusion, vector-valued functions will always be denoted with boldface characters.

The existence and uniqueness of solutions to (1.1) are equivalent to the so-called inf-sup condition,

$$\inf_{q \in \mathbb{P}} \sup_{\mathbf{v} \in \mathbb{V}} \frac{\int_\Omega q \, \operatorname{div} \mathbf{v}}{\|q\| \, \|\nabla \mathbf{v}\|} > 0, \tag{1.2}$$

which holds for the pair $(\mathbb{V}, \mathbb{P})$ as defined above [3]. Hereafter, $\|\cdot\| = \|\cdot\|_\Omega$, and for any domain G, $\|w\|_G = (\int_G |w|^2)^{1/2}$ denotes the usual $L^2(G)$-norm for scalar- as well as vector- and matrix-valued functions on G.

The classical approach to solving the Stokes equations with finite elements is the following. Let $\mathcal{T}_h$ be a triangulation of Ω, and let $\mathbb{V}_h \subset \mathbb{V}$, $\mathbb{P}_h \subset \mathbb{P}$ be finite element spaces defined on $\mathcal{T}_h$. Find a pair $(\mathbf{U}_h, P_h) \in \mathbb{V}_h \times \mathbb{P}_h$ such that

$$\int_\Omega \nabla \mathbf{U}_h : \nabla \mathbf{V}_h - \int_\Omega P_h \, \operatorname{div} \mathbf{V}_h = \int_\Omega \mathbf{f} \cdot \mathbf{V}_h \quad \forall \, \mathbf{V}_h \in \mathbb{V}_h, \tag{1.3}$$

$$\int_\Omega Q_h \, \operatorname{div} \mathbf{U}_h = 0 \quad \forall \, Q_h \in \mathbb{P}_h; \tag{1.4}$$

discrete functions will always be written in capitals. Again, this discrete problem admits a unique solution if and only if the discrete inf-sup condition

$$\inf_{Q_h \in \mathbb{P}_h} \sup_{\mathbf{V}_h \in \mathbb{V}_h} \frac{\int_\Omega Q_h \, \operatorname{div} \mathbf{V}_h}{\|Q_h\| \, \|\nabla \mathbf{V}_h\|} \geq \kappa > 0 \tag{1.5}$$

holds. Moreover, the following *optimal a priori bound* holds:

$$\|\nabla(\mathbf{u} - \mathbf{U}_h)\| + \|p - P_h\| \leq C_\kappa \left(\inf_{\mathbf{V}_h \in \mathbb{V}_h} \|\nabla(\mathbf{u} - \mathbf{V}_h)\| + \inf_{Q_h \in \mathbb{P}_h} \|p - Q_h\| \right), \tag{1.6}$$

where C_κ is a positive constant depending only on κ [3]. When a pair of finite element spaces $(\mathbb{V}_h, \mathbb{P}_h)$ satisfies (1.5), with κ independent of h, the method is called *stable*.

In this article, exploiting an idea introduced in [8] in the context of wavelet approximations to the Stokes problem, we propose and analyze an adaptive FEM for the solution of the Stokes problem. This algorithm consists of an inexact Uzawa iteration at an infinite-dimensional level, and the inner solve is based upon a convergent adaptive FEM for elliptic problems. Amazingly, the *convergence of our adaptive Uzawa algorithm (AUA) does not need the discrete* inf-sup *condition* (1.5) *but rather the continuous* inf-sup *condition* (1.2). This allows for unstable pairs $(\mathbb{V}_h, \mathbb{P}_h)$.

In section 2 we will precisely state the algorithm and prove its convergence for the pairs of spaces consisting of continuous finite elements of degree k for velocity,

whereas for pressure the elements can be either discontinuous of degree $k-1$ or continuous of degree $k-1$ and k. These elements are all *unstable*, except for the Taylor–Hood elements, which consist of continuous elements of degree k for velocity and degree $k-1$ for pressure. We stress that adaptivity is an inherently *nonlinear* process, which appears to detect and exploit the stability of the underlying PDE, namely (1.2), regardless of the finite element spaces. This is perhaps the most salient consequence of our work, which reproduces in the finite element setting the crucial observation made in [7, 8] for wavelets.

This may seem to contradict the celebrated theory of mixed methods [3]. However, it is important to realize that the jth iterate $(\mathbf{U}_j, P_j)$ of our algorithm is not necessarily a solution of the discrete Stokes problem (1.3)–(1.4); it is just an *approximate* solution. Therefore our notion of convergence is fundamentally different from the customary one arising from a priori error analysis in which (1.5) plays a central role and asymptotics is understood in the sense that the meshsize h_j of partition $\mathcal{T}_j$ satisfies $h_j \to 0$: we think of $j \to \infty$ rather than $h_j \to 0$. Depending on the flatness of $\mathbf{u}$ and p, our algorithm may yield convergence even for h_j not tending to zero globally.

Although our theory covers only the class of elements mentioned above and described more specifically in (2.5) and (2.6), extensive computations show convergence for other combinations of elements. Moreover, the computational rate of convergence in terms of degrees of freedom is always optimal. This will be discussed in detail in section 3.

The rest of the article is organized as follows. In section 2 we introduce the AUA and prove its convergence. In section 3 we present numerical evidence showing that the meshes obtained through the AUA are quasi-optimal for any pair of finite element spaces. In section 4 we discuss a posteriori error estimates specially designed for the inexact Uzawa iteration, which are used to stop the outer iterations. Finally, we investigate the properties and complexity of the elliptic inner solver ELLIPTIC in section 5.

In what follows, unless specified otherwise, C will represent a positive constant, possibly depending on mesh-regularity, and the refinements will be done using bisection [2, 19], thus ensuring mesh-regularity.

2. The AUA. We start this section by describing the *exact* Uzawa algorithm in infinite dimensions as an iteration to solve (1.1). Given $p_0 \in \mathbb{P}$, we seek, for $j \geq 1$,

$$
\begin{aligned}
&\mathbf{u}_j \in \mathbb{V}: \qquad \int_\Omega \nabla \mathbf{u}_j : \nabla \mathbf{v} = \int_\Omega \mathbf{f}\cdot\mathbf{v} + \int_\Omega p_{j-1}\ \mathrm{div}\,\mathbf{v} \quad \forall \mathbf{v} \in \mathbb{V},\\
&p_j \in \mathbb{P}: \qquad \int_\Omega p_j q = \int_\Omega p_{j-1} q - \alpha \int_\Omega q\ \mathrm{div}\,\mathbf{u}_j \quad \forall q \in \mathbb{P}.
\end{aligned}
\tag{2.1}
$$

Recall that $\mathbb{V} = \big(\mathring{H}^1(\Omega)\big)^d$, $\mathbb{V}^* = \big(H^{-1}(\Omega)\big)^d$, and $\mathbb{P} = \mathring{L}^2(\Omega)$ and let us denote with $\langle\cdot,\cdot\rangle$ the pairing between $\mathbb{V}$ and $\mathbb{V}^*$ as well as the inner product in $\mathbb{P}$. Let us define the operators $-\Delta$, ∇, and div as follows:

$$
\begin{aligned}
-\Delta : \mathbb{V} \to \mathbb{V}^* \quad \langle -\Delta \mathbf{v},\, \mathbf{w}\rangle &:= \int_\Omega \nabla\mathbf{v} : \nabla\mathbf{w} \quad \forall \mathbf{w} \in \mathbb{V},\\
\nabla : \mathbb{P} \to \mathbb{V}^* \quad \langle \nabla q,\, \mathbf{w}\rangle &:= -\int_\Omega q\ \mathrm{div}\,\mathbf{w} \quad \forall \mathbf{w} \in \mathbb{V},\\
\mathrm{div} : \mathbb{V} \to \mathbb{P} = \mathbb{P}^* \quad \langle \mathrm{div}\,\mathbf{v},\, q\rangle &:= \int_\Omega q\,\mathrm{div}\,\mathbf{v} \quad \forall q \in \mathbb{P}.
\end{aligned}
$$

The *Schur complement operator* $\mathcal{S} : \mathbb{P} \to \mathbb{P}$ is defined by

$$\mathcal{S} := -\operatorname{div}(-\Delta)^{-1}\nabla \tag{2.2}$$

and turns out to be positive definite, self-adjoint, and bounded [6]. Moreover, the Uzawa iteration (2.1) can be written in terms of $\mathcal{S}$ as

$$p_j = (I - \alpha\mathcal{S})p_{j-1} + \alpha F, \tag{2.3}$$

where $F := -\operatorname{div}(-\Delta)^{-1}\mathbf{f}$. Therefore, if $0 < \alpha < 2/\|\mathcal{S}\|_{\mathcal{L}(\mathbb{P},\mathbb{P})}$, then

$$\beta := \big\|I - \alpha\mathcal{S}\big\|_{\mathcal{L}(\mathbb{P},\mathbb{P})} < 1, \tag{2.4}$$

where $\|\cdot\|_{\mathcal{L}(\mathbb{P},\mathbb{P})}$ denotes the norm in the space of bounded linear operators from the Hilbert space $\mathbb{P}$ into itself. Since $\|\mathcal{S}\|_{\mathcal{L}(\mathbb{P},\mathbb{P})} \le 1$ (see [17]), we could take $0 < \alpha < 2$; we chose $\alpha = 1$ in the numerical experiments of section 3.

From now on, $j \ge 0$ will always denote the Uzawa iteration counter, and $\mathcal{T}_j$ will be the jth shape-regular partition of Ω. If k is the polynomial degree for velocity, and l is that for pressure, then we study the pairs of *continuous* finite element spaces

$$\mathbb{V}_j = \mathcal{P}^k(\mathcal{T}_j) \cap \mathbb{V}, \quad \mathbb{P}_j = \mathcal{P}^l(\mathcal{T}_j) \cap \mathbb{P}, \quad l = k, k-1 \ge 1, \tag{2.5}$$

as well as the *discontinuous* finite element spaces

$$\mathbb{V}_j = \mathcal{P}^k(\mathcal{T}_j) \cap \mathbb{V}, \quad \mathbb{P}_j = \mathcal{P}_d^{k-1}(\mathcal{T}_j) \cap \mathbb{P}, \quad k \ge 1. \tag{2.6}$$

Hereafter, $\mathcal{P}_d^k(\mathcal{T}_j)$ denotes the space of—scalar-valued as well as vector-valued—(possibly *discontinuous*) functions that, restricted to an element T, are polynomials of degree $\le k$ for all $T \in \mathcal{T}_j$, and $\mathcal{P}^k(\mathcal{T}_j)$ denotes the subspace of *continuous* functions of $\mathcal{P}_d^k(\mathcal{T}_j)$. We observe that $l = k-1$ in (2.5) corresponds to the popular Taylor–Hood family of finite elements. Any other choice in either (2.5) or (2.6) yields an *unstable* pair of spaces.

Our AUA builds upon a convergent adaptive algorithm for elliptic problems, the procedure ELLIPTIC of section 5, which replaces the first equation in (2.1) by an *approximation*. To introduce such a procedure, we first consider the following auxiliary elliptic problem: Given $\mathbf{f} \in (L^2(\Omega))^d$ and a pressure function $P_{j-1} \in \mathbb{P}_{j-1}$ for $j \ge 1$, solve

$$\mathbf{u}_j \in \mathbb{V} : \int_\Omega \nabla\mathbf{u}_j : \nabla\mathbf{v} = \int_\Omega \mathbf{f}\cdot\mathbf{v} + \int_\Omega P_{j-1}\operatorname{div}\mathbf{v} \quad \forall\mathbf{v} \in \mathbb{V}. \tag{2.7}$$

In contrast with (2.1), we observe that P_{j-1} is discrete in (2.7). If ε_j stands for an adjustable error tolerance, then the procedure ELLIPTIC,

$$(\mathcal{T}_j, \mathbf{U}_j) \leftarrow \text{ELLIPTIC}(\mathcal{T}_{j-1}, P_{j-1}, \varepsilon_j, \mathbf{f}),$$

finds adaptively a refined mesh $\mathcal{T}_j$ of $\mathcal{T}_{j-1}$ and solves the discrete *elliptic* problem

$$\mathbf{U}_j \in \mathbb{V}_j : \quad \int_\Omega \nabla\mathbf{U}_j : \nabla\mathbf{V} = \int_\Omega \mathbf{f}\cdot\mathbf{V} + \int_\Omega P_{j-1}\operatorname{div}\mathbf{V} \quad \forall\ \mathbf{V} \in \mathbb{V}_j, \tag{2.8}$$

within the prescribed error bound

$$\big\|\nabla(\mathbf{u}_j - \mathbf{U}_j)\big\| \le C\varepsilon_j, \tag{2.9}$$

with $C > 0$ independent of j. We point out that this estimate is not standard in that the right-hand side $\mathbf{f} - \nabla P_{j-1}$ of (2.7) may not be in $L^2(\Omega)$ when dealing with discontinuous pressures. This issue is assessed in section 5.

In addition, let $\Pi_j : L^2(\Omega) \to \mathbb{P}_j$ denote the orthogonal L^2-projection into $\mathbb{P}_j$. In section 5, we will show the existence of a constant C independent of j such that the output function $\mathbf{U}_j$ of ELLIPTIC satisfies

$$\| \operatorname{div} \mathbf{U}_j - \Pi_j \operatorname{div} \mathbf{U}_j \| \leq C\varepsilon_j. \tag{2.10}$$

The pressure update is performed by the procedure

$$P_j \leftarrow \text{UPDATE}(\mathcal{T}_j, P_{j-1}, \mathbf{U}_j, \alpha)$$

which computes, according to (2.1) with $\mathbb{P}$ replaced by $\mathbb{P}_j$,

$$P_j \in \mathbb{P}_j : \quad \int_\Omega P_j Q = \int_\Omega P_{j-1} Q - \alpha \int_\Omega Q \operatorname{div} \mathbf{U}_j \quad \forall Q \in \mathbb{P}_j,$$

or equivalently,

$$P_j = P_{j-1} - \alpha\, \Pi_j \operatorname{div} \mathbf{U}_j. \tag{2.11}$$

We are now in a position to introduce the AUA. This algorithm consists of an inexact inner solve using ELLIPTIC in place of (2.1), followed by an update of pressure given by UPDATE. A similar algorithm was first considered by Dahlke, Hochmuth, and Urban [8, 9] in the context of adaptive wavelet methods, which builds upon Elman and Golub [12].

Adaptive Uzawa Algorithm (AUA)

Choose parameters $0 < \alpha < 2$, $0 < \gamma < 1$, $\varepsilon_0 > 0$; set $j = 1$.
1. Select any initial mesh $\mathcal{T}_0$ and any function $P_0 \in \mathbb{P}_0$.
2. Update $\varepsilon_j \leftarrow \gamma \varepsilon_{j-1}$.
3. Compute $(\mathcal{T}_j, \mathbf{U}_j) \leftarrow \text{ELLIPTIC}(\mathcal{T}_{j-1}, \mathbf{f}, P_{j-1}, \varepsilon_j)$.
4. Compute $P_j \leftarrow \text{UPDATE}(\mathcal{T}_j, P_{j-1}, \mathbf{U}_j, \alpha)$.
5. Update $j \leftarrow j + 1$.
6. Go to step 2.

We observe that the AUA makes sense for any pair of spaces $(\mathbb{V}_h, \mathbb{P}_h)$, even *unstable* pairs; this freedom is further investigated in section 3.

THEOREM 2.1. *Let $\alpha > 0$ satisfy* (2.4)*, and let* ELLIPTIC *fulfill* (2.9) *and* (2.10)*. Then, there exist positive constants C_1 and $\delta < 1$ such that the iterates $(\mathbf{U}_j, P_j)$ produced by the* AUA *satisfy*

$$\big\|\nabla(\mathbf{u} - \mathbf{U}_j)\big\| + \big\|p - P_j\big\| \leq C_1 \delta^j.$$

Proof. Let us first observe that (2.7) implies $\mathbf{u}_j = (-\Delta)^{-1}(\mathbf{f} - \nabla P_{j-1})$ for any $j \geq 1$. Hence

$$\begin{aligned} P_j &= P_{j-1} - \alpha\, \Pi_j \operatorname{div} \mathbf{U}_j \\ &= P_{j-1} - \alpha \operatorname{div} \mathbf{u}_j + \alpha \operatorname{div}(\mathbf{u}_j - \mathbf{U}_j) + \alpha(I - \Pi_j) \operatorname{div} \mathbf{U}_j \\ &= (I - \alpha \mathcal{S}) P_{j-1} - \alpha \operatorname{div}(-\Delta)^{-1} \mathbf{f} + \alpha \operatorname{div}(\mathbf{u}_j - \mathbf{U}_j) + \alpha(I - \Pi_j) \operatorname{div} \mathbf{U}_j, \end{aligned}$$

where $\mathcal{S}$ stands for the Schur operator (2.2). Analogously, the exact p satisfies

$$p = (I - \alpha\mathcal{S})p - \alpha \operatorname{div}(-\Delta)^{-1}\mathbf{f},$$

which implies

$$p - P_j = (I - \alpha\mathcal{S})(p - P_{j-1}) - \alpha \operatorname{div}(\mathbf{u}_j - \mathbf{U}_j) - \alpha(I - \Pi_j)\operatorname{div}\mathbf{U}_j.$$

Therefore, in view of (2.4), (2.9), and (2.10), together with property $\|\operatorname{div}\mathbf{v}\| \le \|\nabla\mathbf{v}\|$ [17], we get

$$\begin{aligned}\|p - P_j\| &\le \beta\|p - P_{j-1}\| + \alpha\|\nabla(\mathbf{u}_j - \mathbf{U}_j)\| + \alpha\|(I - \Pi_j)\operatorname{div}\mathbf{U}_j\| \\ &\le \beta\|p - P_{j-1}\| + C\alpha\varepsilon_j = \beta\|p - P_{j-1}\| + C\alpha\varepsilon_0\gamma^j,\end{aligned}$$

where γ is the reduction factor used in step 2 of the AUA. By induction we obtain

$$\|p - P_j\| \le \beta^j\|p - P_0\| + C\alpha\varepsilon_0 \sum_{\ell=0}^{j-1} \beta^\ell\gamma^{j-\ell}, \tag{2.12}$$

and setting $\eta := \max\{\beta, \gamma\}$, we thus have

$$\|p - P_j\| \le \|p - P_0\|\eta^j + \alpha\varepsilon_0 j\eta^j \le C\delta^j \tag{2.13}$$

for some positive constants C and $\eta < \delta < 1$.

To obtain a similar bound for $\|\nabla(\mathbf{u} - \mathbf{U}_j)\|$, we first observe that

$$\int_\Omega \nabla(\mathbf{u} - \mathbf{u}_j) : \nabla\mathbf{v} = \int_\Omega (p - P_{j-1})\operatorname{div}\mathbf{v} \le \|p - P_{j-1}\|\,\|\nabla\mathbf{v}\| \quad \forall\mathbf{v} \in \mathbb{V},$$

whence $\|\nabla(\mathbf{u} - \mathbf{u}_j)\| \le \|p - P_{j-1}\|$. Since

$$\|\nabla(\mathbf{u} - \mathbf{U}_j)\| \le \|\nabla(\mathbf{u} - \mathbf{u}_j)\| + \|\nabla(\mathbf{u}_j - \mathbf{U}_j)\| \le \|p - P_{j-1}\| + \varepsilon_j,$$

(2.13) yields the desired assertion. ∎

Several comments about the AUA and its convergence properties are now in order.

Remark 2.1. For discontinuous pressure spaces $\mathcal{P}_d^l(\mathcal{T}_j)$, $l \ge k-1$, the procedure UPDATE of the AUA hinges upon a pressure correction within the subspace $\operatorname{div}\mathbb{V}_j \subset \mathcal{P}_d^{k-1}(\mathcal{T}_j)$. Consequently, for $l \ge k-1$, the output of the AUA is insensitive to l because the *effective* pressure space is

$$\mathbb{P}_j = \operatorname{div}\mathbb{V}_j;$$

this justifies the restriction $l = k-1$ in (2.6). In contrast, if we enforce continuity of pressure, as in (2.5), then UPDATE works within the subspace $\Pi_j \operatorname{div}\mathbb{V}_j$ of $\mathbb{P}_j = \mathcal{P}^l(\mathcal{T}_j)$ for any $l \ge k-1$, and the output of UPDATE does depend on l.

Remark 2.2. It is remarkable that the discrete inf-sup condition (1.5) plays no role in our analysis. In fact, the above proof hinges solely on the continuous inf-sup condition (1.2) or, equivalently, on the stability of the infinite-dimensional problem (property $\beta < 1$ of $\mathcal{S}$). This observation was first made by Dahlke, Hochmuth, and Urban [8, 9], and very recently exploited by Dahlke, Dahmen, and Urban [7] and Cohen, Dahmen, and DeVore [5], in the context of wavelet approximations of the Stokes system.

Remark 2.3. Unstable elements such as (2.6) are known to yield checkerboard patterns in pressure [3]. One may thus wonder whether any adaptive procedure, which extracts discrete regularity via a posteriori error estimation, may be misled by pressure oscillations and thus fail to produce *selective local* mesh refinement. A possible cure for pressure oscillations within the classical mixed finite element context consists of having a uniformly refined mesh for velocity [3]. In view of (2.8), it turns out that $(\mathbf{U}_j, P_{j-1})$ is a solution to (1.3) with P_{j-1} defined on a grid $\mathcal{T}_{j-1}$ coarser than $\mathcal{T}_j$. This may be regarded as a built-in *stabilization*, but different from the usual one because $\mathcal{T}_j$ is never a global refinement of $\mathcal{T}_{j-1}$ and (1.4) is never fulfilled. This is confirmed by the numerical experiments of section 3, which show optimal meshes for these elements. It thus seems that the nonlinear process associated with adaptivity selects the least amount of refinement necessary to stabilize the method.

Remark 2.4. The procedure ELLIPTIC of the AUA entails an inner loop of the form SOLVE → ESTIMATE → REFINE for the symmetric and coercive elliptic problem (2.7). To achieve the error reduction of (2.9), two ingredients are necessary. First, we need upper and local lower a posteriori error bounds for (2.8). Second, we need a marking strategy and associated error reduction result (2.9). These issues are discussed in section 5.

Remark 2.5. Parameters α, γ, and ε_0 control the behavior of the AUA. The convergence of the AUA, but not its rate, is independent of γ and ε_0 but *not* of α because it dictates the size of the reduction factor β in (2.4). Even though the AUA converges for any choice of γ and ε_0, provided $0 < \gamma < 1$ and $\varepsilon_0 > 0$, its *performance* is greatly influenced by them, especially for unstable elements. In particular, if $\beta < \gamma < 1$, then the *complexity* of ELLIPTIC is independent of j, as will be shown in section 5.2.

Remark 2.6. To stop the AUA it is necessary to have a posteriori error estimators especially designed for the pair $(\mathbf{U}_j, P_{j-1})$, which is *not* a solution of the discrete Stokes problem over $\mathcal{T}_j$. This issue is further investigated in section 4.

3. Experiments and mesh optimality. In this section we focus on the computational performance of the algorithm. We analyze it not only for the elements of Theorem 2.1 but also for cases beyond this. All numerical experiments were carried out using the finite element toolbox ALBERT [18, 19], which provides a flexible programming environment for adaptive finite element computations. Some pictures (Figures 3.2, 3.3, 3.5, 3.7) were produced with the graphics package GRAPE [13].

In order to have an appropriate test bed for the algorithm, we consider two examples in two dimensions and one in three dimensions and run simulations with the AUA for several pairs of elements. They can be divided into three groups, all containing unstable elements: elements of type (2.6), elements of type (2.5)—which include the Taylor–Hood elements—and the continuous unstable pair $\mathcal{P}^1$-$\mathcal{P}^2$ which is not covered by our theory. We always use the following parameters and initial guess:

$$\alpha = 1.0, \quad \gamma = 0.95, \quad \varepsilon_0 = 2.0, \quad P_0 = 0. \tag{3.1}$$

In order to compare our method with the classical adaptive approach to solving the Stokes equations, we also run experiments with a conventional adaptive strategy of the form SOLVE → ESTIMATE → REFINE. For these experiments we use the Taylor–Hood elements $\mathcal{P}^2$-$\mathcal{P}^1$ and $\mathcal{P}^3$-$\mathcal{P}^2$ and the usual residual-type error estimators. An important difference with the AUA is that in SOLVE we solve the saddle-point problem (1.3), (1.4).

The comparative results for the three groups of elements as well as for the conventional approach are reported below in Tables 3.1, 3.2, and 3.3. To describe the information they contain, let us assume that we expect a relation of the form

$$\mathrm{ERR}_j := \|\nabla(u - U_j)\| + \|p - P_j\| \approx C\, N_j^{-r/d},$$

where N_j denotes the number of degrees of freedom (DOFs) at the step j of the outer loop in the AUA, $r = \min\{k, l+1\}$ is the order of the FEM, and d is the dimension. We then define the experimental orders of convergence EOC_j to be

$$\mathrm{EOC}_j := -d\,\frac{\log(\mathrm{ERR}_j/\mathrm{ERR}_{j-1})}{\log(N_j/N_{j-1})},$$

and EOC to be the asymptotic value of EOC_j for large values of j. We also introduce the average error decay (AED) in the energy norm for consecutive outer iterations of the AUA, and the number of DOFs for which the relative energy error

$$\frac{\|\nabla(u - U_j)\| + \|p - P_j\|}{\|\nabla u\| + \|p\|}$$

is less than or equal to prescribed tolerances of 10%, 5%, 1%, and 0.1%, respectively. To compute the errors, we integrated elementwise using a quadrature rule exact for polynomials up to degree 17 in two dimensions and 7 in three dimensions.

We show pictures of pressure (Figures 3.2, 3.3, 3.5), the variable most sensitive to instabilities, and corresponding meshes for several elements at 5% relative accuracy. The velocity never exhibits oscillations, is always well approximated, and is thus not depicted.

We also report curves depicting the relative energy error decay in terms of DOFs and compare them with the optimal slope $-r/d$.

Finally, we draw some conclusions common to all the experiments.

3.1. Example: Smooth solution in two dimensions. Let $\Omega := (-1,1) \times (-1,1)$ and let the velocity $\mathbf{u}$ and pressure p be given by

$$\mathbf{u}(x,y) := \begin{bmatrix} 2y\cos(x^2+y^2) \\ -2x\cos(x^2+y^2) \end{bmatrix}, \qquad p = e^{-10(x^2+y^2)} - p_m,$$

where p_m is such that $\int_\Omega p = 0$ and the forcing $\mathbf{f}$ is computed as $\mathbf{f} = -\Delta u + \nabla p$.

We report the computational results in Table 3.1 and the error decays in Figure 3.1. The behavior of the pressure is illustrated for several pairs of elements in Figures 3.2 and 3.3.

3.2. Example: Singular solution in two dimensions. We consider the L-shaped domain

$$\Omega := \big((-1,1) \times (-1,1)\big) \setminus \big([0,1] \times [-1,0]\big)$$

with reentrant angle $\omega = 3\pi/2$ at the origin. Let $\alpha \approx 0.544$ be an approximation of the smallest root of the nonlinear equation [10]:

$$\frac{\sin^2(\alpha\omega) - \alpha^2 \sin^2\omega}{\alpha^2} = 0.$$

TABLE 3.1

Example 3.1: *EOC, AED per outer iteration, and DOFs to reach tolerances of* 10%, 5%, 1%, 0.1%. *The first* 7 *rows correspond to the AUA and the last* 2 *to the saddle-point problem (compare them with rows* 5 *and* 6 *of the AUA).*

Spaces	EOC	AED	DOFs for relative error of			
			10%	5%	1%	0.1%
$\mathcal{P}^1$-$\mathcal{P}^0_d$	1.075	0.948	6570	24826	448786	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1_d$	2.029	0.951	834	1538	6930	70578
$\mathcal{P}^3$-$\mathcal{P}^2_d$	2.997	0.950	266	1010	1754	8570
$\mathcal{P}^1$-$\mathcal{P}^1$	1.044	0.948	2715	9867	227991	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1$	1.994	0.950	295	403	3403	22791
$\mathcal{P}^3$-$\mathcal{P}^2$	2.878	0.952	211	211	947	4331
$\mathcal{P}^1$-$\mathcal{P}^2$	0.905	0.950	21931	109279	$> 10^6$	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1$	2.057	/	295	403	3403	21351
$\mathcal{P}^3$-$\mathcal{P}^2$	2.321	/	211	211	947	4331

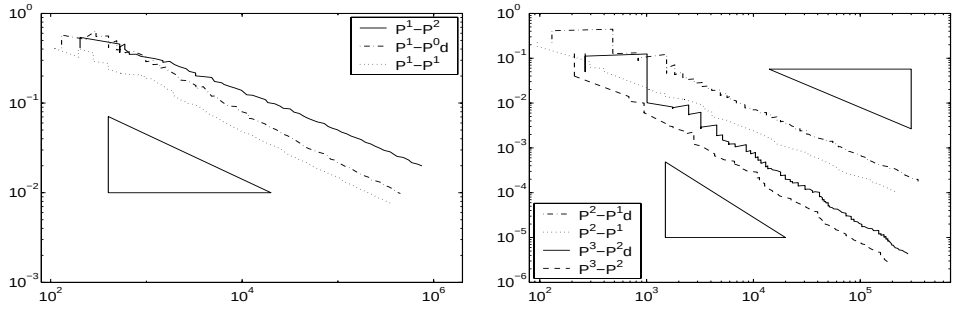


FIG. 3.1. *Example* 3.1: *Relative energy error versus DOFs. Triangles showing optimal decay have slopes* $-1/d = -1/2$ *(left) and* $-2/d = -1$, $-3/d = -3/2$ *(right), respectively. Quasi-optimality of the resulting meshes is thus evident.*

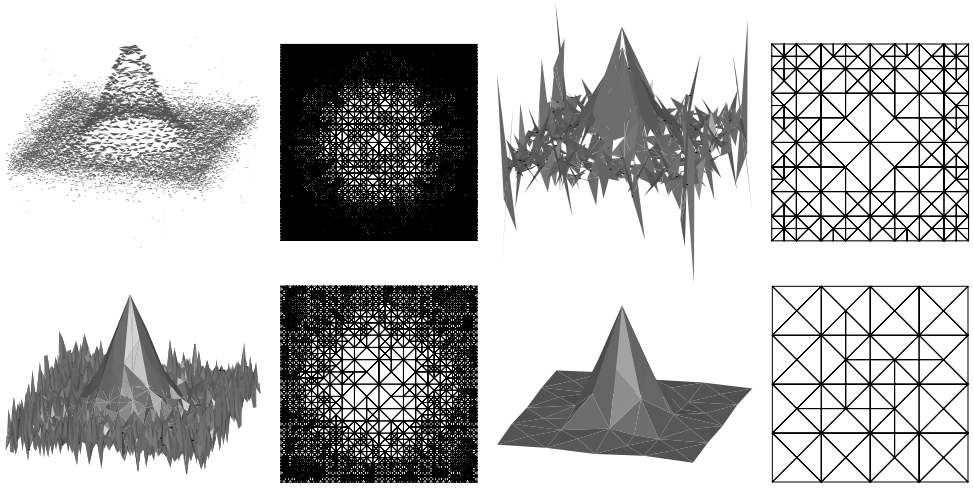


FIG. 3.2. *Example* 3.1: *Pressures and meshes for tolerance of* 5% *and finite element pairs (respectively, outer iteration number/DOFs);* $\mathcal{P}^1$-$\mathcal{P}^0_d$ (60/24826), $\mathcal{P}^2$-$\mathcal{P}^1_d$ (50/1538), $\mathcal{P}^1$-$\mathcal{P}^1$ (50/9867), $\mathcal{P}^2$-$\mathcal{P}^1$ (45/403). *The oscillations for unstable pairs do not persist under further selective refinement.*

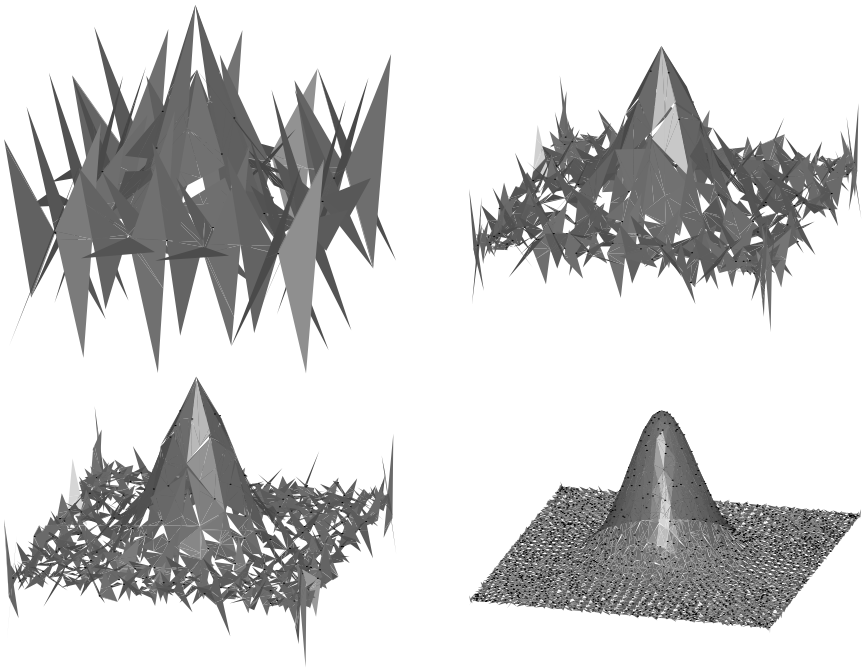


FIG. 3.3. *Example* 3.1: *Sequence of pressures for the unstable pair* $\mathcal{P}^2$-$\mathcal{P}^1_d$ *and outer iterations (respectively, DOFs) j* = 20 (*DOFs* = 482), 60 (2802), 70 (4066), 120 (40986). *Oscillations are present in the early stages of adaptivity but are cured later by selective refinement.*

The exact velocity $\mathbf{u}$ and pressure p are given in polar coordinates by [10, 21]

$$\mathbf{u}(r,\varphi) = r^\alpha \begin{bmatrix} \cos(\varphi)\psi'(\varphi) + (1+\alpha)\sin(\varphi)\psi(\varphi) \\ \sin(\varphi)\psi'(\varphi) - (1+\alpha)\cos(\varphi)\psi(\varphi) \end{bmatrix} = r^\alpha \big(\psi'(\varphi)\mathbf{e}_r - (1+\alpha)\psi(\varphi)\mathbf{e}_\varphi\big)$$

and

$$p(r,\varphi) = -r^{\alpha-1}\frac{(1+\alpha)^2\psi'(\varphi) + \psi'''(\varphi)}{1-\alpha},$$

where $\psi(\varphi)$ is the function

$$\psi(\varphi) = \frac{\sin\left((1+\alpha)\varphi\right)\cos(\alpha\omega)}{1+\alpha} - \cos\left((1+\alpha)\varphi\right) + \frac{\sin\left((\alpha-1)\varphi\right)\cos(\alpha\omega)}{1-\alpha} + \cos((\alpha-1)\varphi).$$

The forcing term is $\mathbf{f} = \mathbf{0}$.

We report the computational results in Table 3.2 and the error decays in Figure 3.4. The behavior of the pressure is illustrated for several pairs of elements in Figure 3.5. In contrast to Example 3.1, the singular nature of p makes selective refinement apparently more effective in this example, which is less prone to oscillations.

3.3. Example: Smooth solution in three dimensions. We consider the cube $\Omega = (-1,1)^3$, and the exact velocity $\mathbf{u}$ and pressure p,

$$\mathbf{u}(x,y,z) = \begin{bmatrix} 2y\cos(x^2+y^2) \\ -2x\cos(x^2+y^2) \\ 0 \end{bmatrix}, \qquad p = \mu e^{-\lambda(x^2+y^2+z^2)} - p_m,$$

TABLE 3.2

Example 3.2: *EOC, AED per outer iteration, and DOFs to reach tolerances of* 10%, 5%, 1%, 0.1%. *The first* 7 *rows correspond to the AUA and the last* 2 *to the saddle-point problem (compare them with rows* 5 *and* 6 *of the AUA).*

Spaces	EOC	AED	DOFs for relative error of			
			10 %	5%	1%	0.1%
$\mathcal{P}^1$-$\mathcal{P}^0_d$	1.116	0.946	3288	9680	164398	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1_d$	1.992	0.950	1058	1940	9314	85686
$\mathcal{P}^3$-$\mathcal{P}^2_d$	2.984	0.950	986	1598	5054	20882
$\mathcal{P}^1$-$\mathcal{P}^1$	1.250	0.943	1434	4971	62979	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1$	2.043	0.948	802	1200	3913	27387
$\mathcal{P}^3$-$\mathcal{P}^2$	3.182	0.948	1125	1757	3153	9749
$\mathcal{P}^1$-$\mathcal{P}^2$	0.907	0.948	7751	41603	$> 10^6$	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1$	2.087	/	668	1012	3273	26708
$\mathcal{P}^3$-$\mathcal{P}^2$	3.425	/	1125	1757	3153	9985

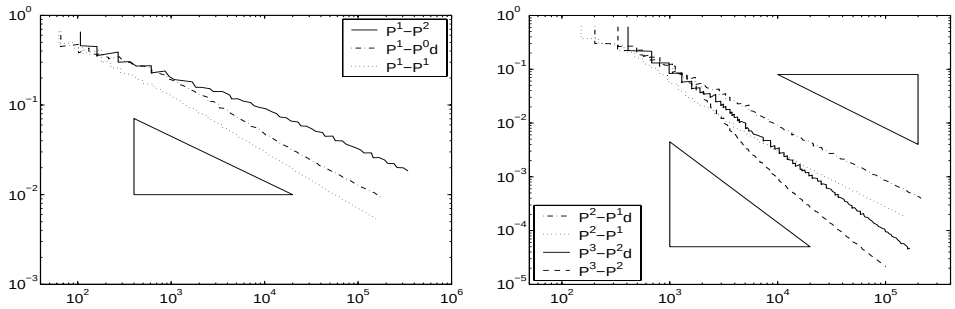


FIG. 3.4. *Example* 3.2: *Relative energy error versus DOFs. Triangles showing optimal decay have slopes* $-1/d = -1/2$ *(left) and* $-2/d = -1$, $-3/d = -3/2$ *(right), respectively. Quasi-optimality of the resulting meshes is thus evident.*

where p_m is such that $\int_\Omega p = 0$. The forcing term $\mathbf{f}$ is computed as $\mathbf{f} = -\Delta u + \nabla p$.

We report the computational results in Table 3.3 and the error decays in Figure 3.6, both for $\mu = 1, \lambda = 10$. Meshes for two finite element pairs are shown in Figure 3.7 for $\mu = 10, \lambda = 300$.

3.4. Conclusions. We now collect and comment on the consistent information about the AUA extracted from the experiments of sections 3.1–3.3.

- Tables 3.1, 3.2, and 3.3 show that error decay in each outer iteration of the AUA is about 0.95, regardless of example and pair of elements. This is a consequence of the choice of $\gamma = 0.95$ of (3.1) and is further discussed in Remark 5.6. Tables 3.1 and 3.2 also reveal convergence for the unstable pair $\mathcal{P}^1$-$\mathcal{P}^2$ in two dimensions, which is not covered by our theory.
- Tables 3.1, 3.2, and 3.3 show that the EOC is optimal for all element pairs and examples and obeys the formula $r = \min\{k, l+1\}$.
- Figures 3.1, 3.4, and 3.6 demonstrate that the relation between error and number of DOFs is optimal for all element pairs and examples: the slopes of the curves match those of the triangles, namely, $-r/d$. The resulting meshes are thus *quasi-optimal* in all cases.

TABLE 3.3

Example 3.3 ($\mu = 1, \lambda = 10$): *EOC, AED per outer iteration, and DOFs to reach tolerances of* 10%, 5%, 1%, 0.1%. *The first* 5 *rows correspond to the AUA and the last* 2 *to the saddle-point problem (compare them with rows* 4 *and* 5 *of the AUA).*

Spaces	EOC	AED	DOFs for relative error of			
			10 %	5%	1%	0.1%
$\mathcal{P}^1$-$\mathcal{P}^0_d$	1.059	0.948	$> 10^6$	$> 10^6$	$> 10^6$	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1_d$	1.872	0.950	23799	128903	$> 10^6$	$> 10^6$
$\mathcal{P}^3$-$\mathcal{P}^2_d$	2.415	0.949	1509	57159	320815	$> 10^6$
$\mathcal{P}^2$-$\mathcal{P}^1$	2.149	0.951	3112	10472	86316	$> 10^6$
$\mathcal{P}^3$-$\mathcal{P}^2$	3.117	0.952	1154	6736	25696	136208
$\mathcal{P}^2$-$\mathcal{P}^1$	2.062	/	3112	10728	71564	$> 10^6$
$\mathcal{P}^3$-$\mathcal{P}^2$	3.239	/	1154	6736	25696	136208

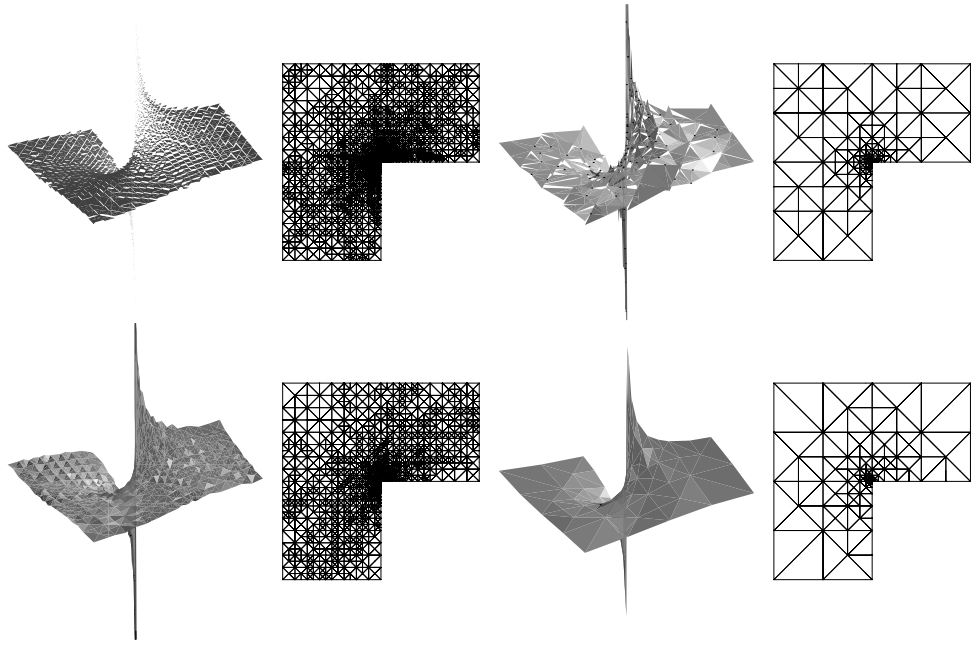


FIG. 3.5. *Example* 3.2: *Pressures and meshes for tolerance of* 5% *and finite element pairs (respectively, outer iteration number/DOFs);* $\mathcal{P}^1$-$\mathcal{P}^0_d$ (50/9680), $\mathcal{P}^2$-$\mathcal{P}^1_d$ (35/1940), $\mathcal{P}^1$-$\mathcal{P}^1$ (50/4971), $\mathcal{P}^2$-$\mathcal{P}^1$ (50/1200). *The oscillations for unstable elements do not persist under further selective refinement.*

- Tables 3.1, 3.2, and 3.3 corroborate the fact that higher order elements are superior to lower order elements for piecewise analytic solutions such as those in Examples 3.1–3.3. For a given tolerance, they need many fewer DOFs than lower order elements.
- Tables 3.1, 3.2, and 3.3 display very similar performance between the AUA, with the element pairs of (2.5) and $l = k - 1$, and the saddle-point approach with the Taylor–Hood families $\mathcal{P}^2$-$\mathcal{P}^1$ and $\mathcal{P}^3$-$\mathcal{P}^2$ (last 2 rows of these tables).
- The stable element pairs (2.5) with $l = k - 1$ exhibit a slightly better performance than the corresponding unstable pairs (2.6). This is documented in Tables 3.1,

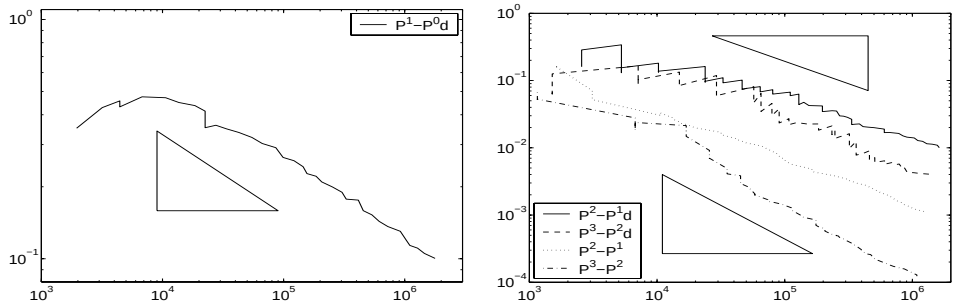


FIG. 3.6. *Example* 3.3 ($\mu = 1, \lambda = 10$): *Relative energy error versus DOFs. Triangles have slopes* $-1/d = -1/3$ *(left) and* $-2/d = -2/3$, $-3/d = -1$ *(right), respectively. Quasi-optimality of the resulting meshes is thus evident.*

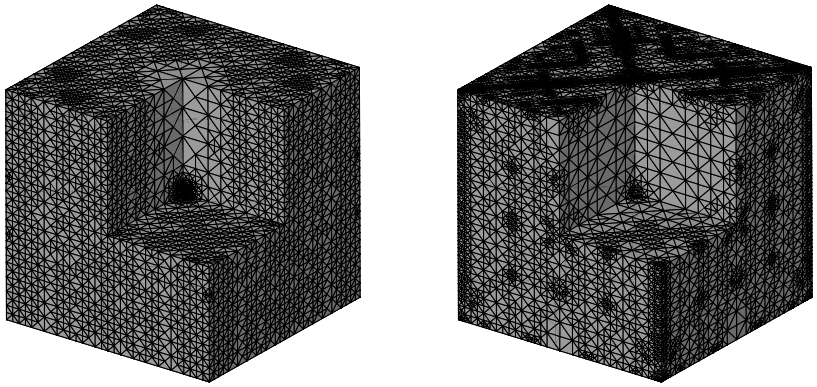


FIG. 3.7. *Example* 3.3 ($\mu = 10, \lambda = 300$): *Mesh for finite element pair* $\mathcal{P}^2$-$\mathcal{P}^1$ *at outer iteration number* $j = 105$, *DOFs* $= 1063176$ *(left), and mesh for finite element pair* $\mathcal{P}^2$-$\mathcal{P}^1_d$ *at* $j = 70$, *DOFs* $= 2007799$ *(right). The first octant has been removed for visualization purposes.*

3.2, and 3.3 in terms of DOFs for a given tolerance, and in Figures 3.2, 3.3, and 3.5 in terms of oscillations.

- It is important to note that oscillations tend to zero in L^2, thereby giving rise to convergence of pressure in L^2. However, as suggested by Figures 3.2 and 3.3, this might be a rather weak concept of convergence in practice, which is in contrast to common belief.

4. A posteriori error estimators. In this section we derive a posteriori error estimators for the pair $(\mathbf{U}_j, P_{j-1})$, which are instrumental to stopping the outer loop in the AUA. We start by defining the bilinear form $\mathcal{L} : (\mathbb{V} \times \mathbb{P}) \times (\mathbb{V} \times \mathbb{P}) \to \mathbb{R}$,

$$\mathcal{L}\left[(\mathbf{v}, q), (\mathbf{w}, r)\right] := \int_\Omega \nabla \mathbf{v} : \nabla \mathbf{w} - \int_\Omega q \ \mathrm{div}\, \mathbf{w} + \int_\Omega r \ \mathrm{div}\, \mathbf{v},$$

and noting that (1.1) is equivalent to finding a pair $(\mathbf{u}, p) \in \mathbb{V} \times \mathbb{P}$ such that

$$\mathcal{L}\left[(\mathbf{u}, p), (\mathbf{v}, q)\right] = \int_\Omega \mathbf{f} \cdot \mathbf{v} \quad \forall (\mathbf{v}, q) \in \mathbb{V} \times \mathbb{P}.$$

Moreover, the *continuous* inf-sup condition (1.2) is equivalent to the existence of a constant $\Lambda > 0$ such that

$$\|\nabla \mathbf{v}\| + \|q\| \leq \Lambda \sup_{(\mathbf{w},r)\in \mathbb{V}\times\mathbb{P}} \frac{\mathcal{L}\,[(\mathbf{v},q),(\mathbf{w},r)]}{\|\nabla \mathbf{w}\| + \|r\|}. \tag{4.1}$$

This property alone, or equivalently, the *stability* of the continuous problem, is responsible for a posteriori error estimates for the Stokes system, even for unstable elements. Therefore, the mere derivation of a posteriori error estimators is no guarantee of convergence of any adaptive algorithm based on them.

In what follows we derive both global upper and local lower a posteriori error bounds for the pair $(\mathbf{U}_j, P_{j-1})$. This pair is a solution of (2.8), but not of the discrete Stokes problem (1.3)–(1.4) for the following two reasons:

- pressure P_{j-1} is piecewise polynomial in the mesh $\mathcal{T}_{j-1}$, which is coarser than $\mathcal{T}_j$;
- equation (1.4) is not fulfilled.

Altogether, this makes our error analysis a bit unusual. However, since the same techniques reported in [1, 20] apply in our context, we only sketch the proofs for completeness. We first set $\mathbf{e}_u := \mathbf{u} - \mathbf{U}_j$ and $e_p := p - P_{j-1}$ and observe that, from (1.1) and (2.8), we have

$$\mathcal{L}\,[(\mathbf{e}_u, e_p),(\mathbf{w},r)] = \sum_{T\in\mathcal{T}_j} \left(\int_T \mathbf{f}\cdot\mathbf{w} - (\nabla \mathbf{U}_j - P_{j-1}\mathbf{I}) : \nabla \mathbf{w} - r\ \mathrm{div}\,\mathbf{U}_j \right)$$

for any $\mathbf{w} \in \mathbb{V}$, $r \in \mathbb{P}$; here $\mathbf{I} \in \mathbb{R}^{d\times d}$ stands for the identity matrix. Since the matrix $\mathbf{T}_j := \nabla \mathbf{U}_j - P_{j-1}\mathbf{I}$ plays a crucial role, we introduce the *jump residual*,

$$\mathbf{J}_j := [\![\mathbf{T}_j \cdot \mathbf{n}]\!] = [\![\nabla \mathbf{U}_j \cdot \mathbf{n} - P_{j-1}\mathbf{n}]\!], \tag{4.2}$$

which indicates the jump of the vector-valued function $\mathbf{T}_j \cdot \mathbf{n}$ across interelement sides S. Such a jump is independent of the choice of the normal $\mathbf{n}$ to S and is defined as zero for boundary sides. We also introduce the *interior residual*,

$$\mathbf{R}_j := \mathbf{f} + \Delta \mathbf{U}_j - \nabla P_{j-1}, \tag{4.3}$$

which is computed elementwise.

LEMMA 4.1 (upper bound). *Let $\{(\mathbf{U}_j, P_j)\}_{j=1}^{\infty}$ be the sequence of solutions produced by the* AUA. *Then there exists a constant C^* depending only on mesh shape-regularity such that the following a posteriori upper bound for the error of the pair $(\mathbf{U}_j, P_{j-1})$ holds:*

$$\|\nabla(\mathbf{u} - \mathbf{U}_j)\| + \|p - P_{j-1}\| \leq C^* \left(\sum_{T\in\mathcal{T}_j} \zeta_j(T)^2 \right)^{1/2},$$

where the local error indicators $\zeta_j(T)$ are given by

$$\zeta_j(T)^2 = h_T^2 \|\mathbf{R}_j\|_T^2 + h_T \|\mathbf{J}_j\|_{\partial T}^2 + \|\,\mathrm{div}\,\mathbf{U}_j\|_T^2 \quad \forall\ T \in \mathcal{T}_j,$$

and the quantity h_T represents the diameter of the element $T \in \mathcal{T}_j$.

Proof. Exploiting (2.8), we deduce Galerkin orthogonality $\mathcal{L}[(\mathbf{e}_u, e_p), (\mathbf{W}, 0)] = 0$ for all $\mathbf{W} \in \mathbb{V}_j$. We then have

$$\mathcal{L}\left[(\mathbf{e}_u, e_p), (\mathbf{w}, r)\right] = \mathcal{L}\left[(\mathbf{e}_u, e_p), (\mathbf{w} - \mathbf{W}, r)\right] \quad \forall\ \mathbf{W} \in \mathbb{V}_j,$$

and integrating by parts, with $\mathbf{z} = \mathbf{w} - \mathbf{W}$ we obtain

$$\mathcal{L}\left[(\mathbf{e}_u, e_p), (\mathbf{w}, r)\right] = \sum_{T \in \mathcal{T}_j} \left(\int_T \mathbf{R}_j \mathbf{z} + \frac{1}{2} \int_{\partial T} \mathbf{J}_j \mathbf{z} - \int_T r \ \mathrm{div}\, \mathbf{U}_j \right).$$

Finally, taking $\mathbf{W} \in \mathbb{V}_j$ as the Clément interpolant of $\mathbf{w}$, we arrive at

$$\|\mathbf{z}\|_T \leq C h_T \left\|\nabla \mathbf{w}\right\|_{\mathcal{N}_j(T)}, \qquad \|\mathbf{z}\|_{\partial T} \leq C h_T^{1/2} \left\|\nabla \mathbf{w}\right\|_{\mathcal{N}_j(T)},$$

where $\mathcal{N}_j(T)$ is the union of all elements of $\mathcal{T}_j$ sharing at least a vertex with $T \in \mathcal{T}_j$. This, together with (4.1), leads to the assertion. ∎

Remark 4.1. The a posteriori error analysis for the Stokes system is based exclusively on satisfaction of the momentum equation (1.3), or (2.8), but not of the incompressibility equation (1.4). In fact, (1.4) is not valid in either our setting or when stabilizing terms are added [14].

Before stating the local lower error bound, we need to introduce the concept of *data oscillation*, which accounts for information missing due to the averaging process associated with the FEM. Given a subset of elements $\mathcal{F}$ of $\mathcal{T}_j$, we set

$$\mathrm{osc}(\mathbf{f}, \mathcal{F}) := \left(\sum_{T \in \mathcal{F}} h_T^2 \|\mathbf{f} - \mathbf{f}_T\|_T^2 \right)^{1/2}, \tag{4.4}$$

where $\mathbf{f}_T$ is the (local) L^2-projection of $\mathbf{f}$ into the polynomial space $\mathcal{P}^{k-1}(T)$, and k is the polynomial degree of the velocity space $\mathbb{V}_j$. Given an element $T \in \mathcal{T}_j$ we designate with $\mathcal{F}_j(T)$ either the set of elements of $\mathcal{T}_j$ sharing a face with T or their union. This abuse of notation will not lead to confusion.

LEMMA 4.2 (lower bound). *Let $\{(\mathbf{U}_j, P_j)\}_{j=1}^{\infty}$ be the sequence of solutions produced by the* AUA. *Then there exists a constant C_*, depending only on mesh shape-regularity, such that the following local a posteriori lower bound for the error of the pair $(\mathbf{U}_j, P_{j-1})$ holds:*

$$\zeta_j(T) \leq C_* \Big(\left\|\nabla(\mathbf{u} - \mathbf{U}_j)\right\|_{\mathcal{F}_j(T)} + \left\|p - P_{j-1}\right\|_{\mathcal{F}_j(T)} + \mathrm{osc}(\mathbf{f}, \mathcal{F}_j(T)) \Big).$$

We omit the proof because it is the same as that in [1, 21, 20]. This result shows that the upper bound is sharp (global efficiency), and implies that *local efficiency* refining where the local indicators $\zeta_j(T)$ are large is always necessary to reduce the error. This property seems to be distinctive of finite elements and in fact is not valid for wavelets.

5. ELLIPTIC: Realization and complexity. In this section we first define the procedure ELLIPTIC and prove the key properties (2.9) and (2.10) for the finite element families (2.5) and (2.6). Second, we analyze the complexity of ELLIPTIC in terms of the number of iterations necessary to achieve (2.9).

5.1. Realization. The study of convergence of adaptive FEM for elliptic problems in the multidimensional setting started with the seminal work by Dörfler [11] and was further developed by Morin, Nochetto, and Siebert in [15, 16]. In this section we will follow the approach in [15, 16] to state the algorithm and prove its convergence for a special class of H^{-1} right-hand sides. This is the class of L^2 vector-valued functions plus gradients of functions in the pressure space $\mathbb{P}_h$ which might have discontinuities across interelement boundaries.

In this section we drop the outer counter j and relabel the input arguments $\mathcal{T}_{j-1}$, P_{j-1}, ε_j of ELLIPTIC as $\mathcal{T}^0$, P, ε and relabel the output $\mathcal{T}_j, \mathbf{U}_j$ as $\mathcal{T}, \mathbf{U}$. Hence

$$(\mathcal{T}, \mathbf{U}) \leftarrow \text{ELLIPTIC}(\mathcal{T}^0, P, \varepsilon, \mathbf{f}). \tag{5.1}$$

To avoid confusion we always use *superscripts*, instead of subscripts, whenever inner iterates of ELLIPTIC are involved. Consequently, for $i \geq 1$ we denote by $\mathcal{T}^i$ a refinement of $\mathcal{T}^{i-1}$, by $\mathbb{V}^i$ the corresponding finite element space for velocities, and by $\mathbf{U}^i$ the solution to the following discrete elliptic problem:

$$\mathbf{U}^i \in \mathbb{V}^i : \quad \int_\Omega \nabla \mathbf{U}^i : \nabla \mathbf{V} = \int_\Omega \mathbf{f} \cdot \mathbf{V} + \int_\Omega P \operatorname{div} \mathbf{V} \quad \forall\, \mathbf{V} \in \mathbb{V}^i. \tag{5.2}$$

This is the discretization of (2.7) or, equivalently, of the elliptic PDE $-\Delta \mathbf{u} = \mathbf{f} - \nabla P$ with $\mathbf{u} = \mathbf{u}_j$; since $\mathbf{f} - \nabla P \in \mathbb{V}^*$, there exists a unique solution to (2.7). We notice that P does *not* change with i and that, when ELLIPTIC stops, (5.2) becomes (2.8). According to (4.2) and (4.3), the residuals of (5.2) are

$$\mathbf{J} := [\![\nabla \mathbf{U}^i \cdot \mathbf{n} - P\mathbf{n}]\!], \qquad \mathbf{R} := \mathbf{f} + \Delta \mathbf{U}^i - \nabla P. \tag{5.3}$$

LEMMA 5.1. *Let $\mathbf{u}$ be the solution of (2.7). For $i \geq 1$, let $\mathcal{T}^i$ be a refinement of $\mathcal{T}^{i-1}$ and let $\mathbf{U}^i$ be the solution to (5.2). Let the local error indicators $\eta^i(T)$ be*

$$\eta^i(T)^2 := h_T^2 \, \|\mathbf{R}\|_T^2 + h_T \, \|\mathbf{J}\|_{\partial T}^2 \quad \forall\, T \in \mathcal{T}^i.$$

Then there exist two constants K^*, K_* depending only on mesh shape-regularity, but otherwise independent of $\mathbf{u}$, $\mathbf{f}$, P, and $\mathcal{T}^i$, such that

$$\left\| \nabla(\mathbf{u} - \mathbf{U}^i) \right\|^2 \leq K^* \sum_{T \in \mathcal{T}^i} \eta^i(T)^2, \tag{5.4}$$

$$\eta^i(T)^2 \leq K_* \big(\left\| \nabla(\mathbf{u} - \mathbf{U}^i) \right\|^2_{\mathcal{F}^i(T)} + \operatorname{osc}(\mathbf{f}, \mathcal{F}^i(T))^2 \big) \quad \forall\, T \in \mathcal{T}^i. \tag{5.5}$$

Proof. We note that the error equation can be written equivalently as

$$\int_\Omega \nabla(\mathbf{u} - \mathbf{U}^i) : \nabla \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v} - \int_\Omega (\nabla \mathbf{U}^i - P\mathbf{I}) : \nabla \mathbf{v} \quad \forall\, \mathbf{v} \in \mathbb{V}.$$

The argument now proceeds as in Lemma 4.1 with $r = 0$; see also [20]. ∎

Remark 5.1. Regardless of the adaptive algorithm used to reduce the error, the estimate (5.4) allows us to measure the error $\mathbf{u} - \mathbf{U}^i$ up to a factor K^*. Stopping the inner iterations of ELLIPTIC when $\sum_{T \in \mathcal{T}^i} \eta^i(T)^2 < \varepsilon^2$ thus guarantees (2.9).

To motivate the subsequent discussion about the convergence of adaptive FEM for elliptic problems, we observe that consecutive spaces are nested $\mathbb{V}^i \subset \mathbb{V}^{i+1}$, whence $\mathbf{U}^i - \mathbf{U}^{i+1} \in \mathbb{V}^{i+1}$. Consequently, using the orthogonal decomposition $\mathbf{u} - \mathbf{U}^i = (\mathbf{u} - \mathbf{U}^{i+1}) + (\mathbf{U}^{i+1} - \mathbf{U}^i)$, the Pythagoras theorem yields

$$\left\| \nabla(\mathbf{u} - \mathbf{U}^{i+1}) \right\|^2 = \left\| \nabla(\mathbf{u} - \mathbf{U}^i) \right\|^2 - \left\| \nabla(\mathbf{U}^i - \mathbf{U}^{i+1}) \right\|^2. \tag{5.6}$$

The error reduction is thus exactly $\|\nabla(\mathbf{U}^i - \mathbf{U}^{i+1})\|^2$. In order to guarantee that the error decreases a fixed proportion of the current error $\|\nabla(\mathbf{u}-\mathbf{U}^i)\|$, we have to bound $\|\nabla(\mathbf{U}^i - \mathbf{U}^{i+1})\|$ from below by $\|\nabla(\mathbf{u} - \mathbf{U}^i)\|$; this is the key idea from [11]. In view of Lemma 5.1, this reduces to showing a local lower bound for $\|\nabla(\mathbf{U}^i - \mathbf{U}^{i+1})\|_{\mathcal{F}^i(T)}$ in terms of $\eta^i(T)$.

The following lemma states such a lower bound and is crucial for the error reduction property that leads to convergence. Its proof is different from that in [16] due to the presence of the singular term ∇P in (5.2) and is postponed until the end of this section. We say that an element $T \in \mathcal{T}^i$ satisfies the *interior node* property if

(5.7) *the element $T \in \mathcal{T}^i$, as well as each of its faces, contains a node of the finer mesh $\mathcal{T}^{i+1}$ in its interior.*

LEMMA 5.2. *Let $\mathcal{T}^{i+1}$ be a refinement of $\mathcal{T}^i$, and let $T \in \mathcal{T}^i$ be an element for which every $T' \in \mathcal{F}^i(T)$ satisfies* (5.7)*. Then, there exists a constant K_*, depending only on mesh shape-regularity, such that*

$$\eta^i(T)^2 \le K_*\big(\|\nabla(\mathbf{U}^{i+1} - \mathbf{U}^i)\|^2_{\mathcal{F}^i(T)} + \operatorname{osc}(\mathbf{f}, \mathcal{F}^i(T))^2\big).$$

Remark 5.2. We refer to [15] for a thorough discussion about the requirement (5.7) and its importance for convergence.

We now present the following algorithm, which was first proposed in [15], based on a marking strategy due to Dörfler [11].

ELLIPTIC$(\mathcal{T}^0, P, \varepsilon, \mathbf{f})$

Choose parameters $0 < \theta,\ \theta_f < 1$, and set $i = 0$.
1. Compute the discrete solution $\mathbf{U}^i \in \mathbb{V}^i$ of (5.2) over $\mathcal{T}^i$.
2. Compute the local indicators $\eta^i(T)$.
3. If $\big(\sum_{T\in\mathcal{T}^i} \eta^i(T)^2\big)^{1/2} \le \varepsilon$, return the pair $(\mathcal{T}^i, \mathbf{U}^i)$ to AUA.
4. Mark a subset $\hat{\mathcal{T}}^i \subset \mathcal{T}^i$ such that

$$\textstyle\sum_{T\in\hat{\mathcal{T}}^i} \eta^i(T)^2 \ge \theta \sum_{T\in\mathcal{T}^i} \eta^i(T)^2.$$

5. Define $\tilde{\mathcal{T}}^i$ to be the set of all elements $T' \in \mathcal{F}^i(T)$ for $T \in \hat{\mathcal{T}}^i$.
6. Enlarge (if necessary) $\tilde{\mathcal{T}}^i$ to satisfy

$$\textstyle\sum_{T\in\tilde{\mathcal{T}}^i} h_T^2 \|f - f_T\|_T^2 \ge \theta_f \sum_{T\in\mathcal{T}^i} h_T^2 \|f - f_T\|_T^2.$$

7. Refine $\mathcal{T}^i$ so that every element $T' \in \tilde{\mathcal{T}}^i$ satisfies (5.7).
8. Update $i \leftarrow i + 1$ and go to step 1.

Remark 5.3. Let us note that step 6 implies the existence of a constant $\rho_f \in (0, 1)$, depending only on θ_f and mesh shape-regularity, such that

$$\operatorname{osc}(f, \mathcal{T}^{i+1}) \le \rho_f \operatorname{osc}(f, \mathcal{T}^i) \quad \forall\ i \ge 0. \tag{5.8}$$

This assertion has been proved in [15] for linear finite elements, but that proof remains valid for any polynomial degree and is thus omitted.

Remark 5.4. Since (5.2) is not the ultimate goal of the AUA, but rather an intermediate problem, we used $\theta = \theta_f = 0.1$ in the experiments of section 3 for the refinement decisions of ELLIPTIC to be rather conservative. This yields a suitable

work balance between ELLIPTIC and UPDATE. Moreover, step 6 of ELLIPTIC plays no significant role in practice for smooth $\mathbf{f}$ [15].

For ELLIPTIC we have the following result, which guarantees that (2.9) can be achieved with finite iterations. We show in Proposition 5.6 that this number of iterations is indeed independent of the outer counter j.

PROPOSITION 5.3. *Let $\mathbf{U}^i$ be a sequence of finite element solutions produced by* ELLIPTIC. *Then, there exist two constants C_0 and $\rho < 1$, depending only on the parameters θ, θ_f of* ELLIPTIC*, such that*

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\| \le \rho^i \max\left\{ \left\|\nabla(\mathbf{u}-\mathbf{U}^0)\right\|, C_0\,\mathrm{osc}(f,\mathcal{T}^0)\right\}. \tag{5.9}$$

Proof. By virtue of Lemmas 5.1, 5.2, and step 4 of ELLIPTIC, we have that

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\|^2 \le \frac{K^*}{\theta}\sum_{T\in\hat{\mathcal{T}}^i}\eta^i(T)^2 \le \frac{1}{\lambda}\left(\left\|\nabla(\mathbf{U}^i-\mathbf{U}^{i+1})\right\|^2 + \mathrm{osc}(f,\mathcal{T}^i)^2\right),$$

with $\lambda = \frac{\theta}{(d+2)\,K^*\,K_*}$. Combining this with (5.6) we arrive at

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^{i+1})\right\|^2 \le (1-\lambda)\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\|^2 + \mathrm{osc}(f,\mathcal{T}^i)^2.$$

If $\mu > 0$ is sufficiently small so that $\rho_e^2 := 1-\lambda+\mu^2 < 1$, as well as

$$\mathrm{osc}(f,\mathcal{T}^i) \le \mu\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\|, \tag{5.10}$$

we then get the error reduction formula

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^{i+1})\right\| \le \rho_e\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\|. \tag{5.11}$$

To prove (5.9) we set $\rho := \max\{\rho_e,\rho_f\}$, $C_0 := (\mu\,\rho)^{-1}$, with ρ_f as in Remark 5.3, and argue by induction. Since the claim holds trivially for $i=0$, we assume that it holds for $i \ge 0$. Then, we have the two alternatives

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\| > \rho^{i+1}C_0\,\mathrm{osc}(f,\mathcal{T}^0), \tag{5.12}$$

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\| \le \rho^{i+1}C_0\,\mathrm{osc}(f,\mathcal{T}^0). \tag{5.13}$$

In case (5.12), we see from Remark 5.3 that $\mathrm{osc}(f,\mathcal{T}^i) \le \rho_f^i\,\mathrm{osc}(f,\mathcal{T}^0)$, whence

$$\mathrm{osc}(f,\mathcal{T}^i) \le \mu\rho^{i+1}\frac{\mathrm{osc}(f,\mathcal{T}^0)}{\rho\mu} = \mu\rho^{i+1}C_0\,\mathrm{osc}(f,\mathcal{T}^0) < \mu\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\|.$$

Consequently, (5.10) holds and, by (5.11) and the induction assumption, we deduce

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^{i+1})\right\| \le \rho\left\|\nabla(\mathbf{u}-\mathbf{U}^i)\right\| \le \rho^{i+1}\max\left\{\left\|\nabla(\mathbf{u}_j-\mathbf{U}^0)\right\|, C_0\,\mathrm{osc}(f,\mathcal{T}^0)\right\}.$$

On the other hand, exploiting that $\mathcal{T}^{i+1}$ is a refinement of $\mathcal{T}^i$, and thus that the error cannot increase $\|\nabla(\mathbf{u}-\mathbf{U}^{i+1})\| \le \|\nabla(\mathbf{u}-\mathbf{U}^i)\|$, we handle (5.13) as follows:

$$\left\|\nabla(\mathbf{u}-\mathbf{U}^{i+1})\right\| \le \rho^{i+1}C_0\,\mathrm{osc}(f,\mathcal{T}_0) \le \rho^{i+1}\max\left\{\left\|\nabla(\mathbf{u}-\mathbf{U}^0)\right\|, C_0\,\mathrm{osc}(f,\mathcal{T}^0)\right\}.$$

The proof is thus complete. ∎

We now verify property (2.10) for the families (2.5) and (2.6). We note that $\mathrm{div}\,\mathcal{P}^k(\mathcal{T}_j) \subset \mathcal{P}_d^{k-1}(\mathcal{T}_j)$, whence Π_j reduces to the identity for (2.6) and thus (2.10) is trivially satisfied. The case (2.5) is more delicate and is the subject of our next result.

PROPOSITION 5.4. *The following interpolation estimate is valid:*

$$\|g - \Pi_j g\| \le C \left(\sum_{T \in \mathcal{T}_j} h_T \, \|[\![g]\!]\|^2_{\partial T} \right)^{1/2} \qquad \forall\ g \in \mathcal{P}_d^{k-1}(\mathcal{T}_j). \tag{5.14}$$

Proof. We recall that $\Pi_j g \in \mathbb{P}_j$ is the best L^2-approximation in $\mathbb{P}_j = \mathcal{P}^l(\mathcal{T}_j)$ of g and $l \ge k-1$. To prove the assertion we could simply replace $\Pi_j g$ by any interpolant of g into $\mathbb{P}_j$. We now construct an interpolation operator I_j closely related to the Clément operator [4]. Let ω_i be the star of $\mathcal{T}_j$ corresponding to the node x_i, and let $g_i \in P^l(\omega_i)$ be the L^2-projection of g into the space of continuous *piecewise* polynomials $P^l(\omega_i)$:

$$g_i \in P^l(\omega_i): \quad \int_{\omega_i} (g - g_i) q = 0 \quad \forall q \in P^l(\omega_i).$$

We then set $I_j g(x_i) := g_i(x_i)$ and recall that to estimate the error $g - I_j g$ it suffices to bound $g - g_i$ over ω_i for all i [4]. To this end, we first scale ω_i to a reference situation of unit size and then realize that, since $g - g_i$ is piecewise polynomial, all its norms are equivalent. In particular, we claim that the seminorm

$$|g - g_i|_{\omega_i} := \left(\sum_{S \subset \omega_i} \|[\![g]\!]\|^2_S \right)^{1/2}$$

is a norm. In fact, if $|g - g_i|_{\omega_i} = 0$ then $g - g_i$ is continuous in ω_i, $g - g_i \in \mathcal{P}^l(\omega_i)$, and thus $g - g_i$ is orthogonal to itself, whence $g - g_i = 0$. A scaling back to ω_i yields the power of meshsize asserted in (5.14) and concludes the proof. ∎

We point out that Proposition 5.4 remains true if Π_j is an L^2-projection into any space of continuous piecewise polynomials containing $\mathcal{P}^{k-1}(\mathcal{T}_j)$.

COROLLARY 5.5. *There exists a constant $C > 0$, depending only on mesh shape-regularity and k, such that*

$$\|\nabla \mathbf{V} - \Pi_j \nabla \mathbf{V}\| \le C \left(\sum_{T \in \mathcal{T}_j} h_T \, \|[\![\nabla \mathbf{V} \cdot \mathbf{n}]\!]\|^2_{\partial T} \right)^{1/2} \qquad \forall\ \mathbf{V} \in \mathcal{P}^k(\mathcal{T}_j).$$

Proof. We take $g \in \mathcal{P}_d^{k-1}(\mathcal{T}_j)$ to be any partial derivative of $\mathbf{V} \in \mathcal{P}^k(\mathcal{T}_j)$ and observe that $|[\![g]\!]| \le |[\![\nabla \mathbf{V} \cdot \mathbf{n}]\!]|$ because V being continuous makes the jump $[\![\nabla \mathbf{V}]\!]$ vanish in any tangential direction. We now apply Proposition 5.4. ∎

To derive (2.10) from Corollary 5.5 in case (2.5), we further note that if $\mathbb{P}_{j-1}$ is a space of continuous finite elements, then the jump residual of (5.3) reduces to the jumps of $\nabla \mathbf{U}_j$, which are bounded by ε_j when ELLIPTIC stops.

Proof of Lemma 5.2. We first prove the following estimate for the interior residual $\mathbf{R}$ of (5.3), provided $T \in \mathcal{T}^i$ has a node of the finer mesh $\mathcal{T}^{i+1}$ in its interior:

$$h_T^2 \, \|\mathbf{R}\|_T^2 \le C \big(\big\|\nabla(\mathbf{U}^{i+1} - \mathbf{U}^i)\big\|_T^2 + \operatorname{osc}(\mathbf{f}, T)^2 \big). \tag{5.15}$$

We recall that $\mathbf{f}_T$ denotes the orthogonal L^2-projection of any vector-valued function $\mathbf{f}$ into $\mathcal{P}^{k-1}(T)$. Then, since the degree of the pressure space is $\ell \le k$, $(\Delta \mathbf{U}^i - \nabla P)_T = \Delta \mathbf{U}^i - \nabla P$, whence $\mathbf{R} - \mathbf{R}_T = \mathbf{f} - \mathbf{f}_T$.

Now let φ_T be the canonical continuous piecewise linear basis function of the triangulation $\mathcal{T}^{i+1}$ corresponding to the node inside T; thus $\operatorname{supp}\varphi_T \subset T$. Since $\mathbf{R}_T$ is a polynomial, both $\|\mathbf{R}_T\|_T^2$ and $\int_T |R_T|^2\varphi_T$ are equivalent up to a constant depending on mesh-regularity. Therefore, integrating by parts and using the fact that $\mathbf{R}_T\varphi_T \in \mathbb{V}^{i+1}$, we get

$$\begin{aligned}\|\mathbf{R}_T\|_T^2 &\le C\int_T |\mathbf{R}_T|^2\varphi_T = \int_T \mathbf{R}\cdot(\mathbf{R}_T\varphi_T) + \int_T (\mathbf{R}_T-\mathbf{R})\cdot(\mathbf{R}_T\varphi_T)\\ &= \int_T \nabla(\mathbf{U}^{i+1}-\mathbf{U}^i):\nabla(\mathbf{R}_T\varphi_T) + \int_T (\mathbf{f}_T-\mathbf{f})\cdot(\mathbf{R}_T\varphi_T)\\ &\le C\big(\big\|\nabla(\mathbf{U}^{i+1}-\mathbf{U}^i)\big\|_T \|\nabla(\mathbf{R}_T\varphi_T)\|_T + \|\mathbf{f}_T-\mathbf{f}\|_T\,\|\mathbf{R}_T\varphi_T\|_T\big).\end{aligned}$$

Since $\mathbf{R}_T\varphi_T \in \mathbb{V}^{i+1}$, applying the inverse inequality $\|\nabla(\mathbf{R}_T\varphi_T)\|_T \le Ch_T^{-1}\|\mathbf{R}_T\varphi_T\|_T$, together with the triangle inequality $\|\mathbf{R}\|_T \le \|\mathbf{R}_T\|_T + \|\mathbf{f}-\mathbf{f}_T\|$, results in (5.15).

We next consider a side S of $\mathcal{T}^i$ having a node of $\mathcal{T}^{i+1}$ in its interior and prove the following estimate for the residual $\mathbf{J}$ in (5.3):

$$h_T\|\mathbf{J}\|_S^2 \le C\big(h_T\|\mathbf{R}\|_{\mathcal{F}^i(S)}^2 + \big\|\nabla(\mathbf{U}^{i+1}-\mathbf{U}^i)\big\|_{\mathcal{F}^i(S)}^2\big). \tag{5.16}$$

Let us first observe that $\mathbf{J}$ is a polynomial of degree at most $k-1$ on S. In fact, if $P \in \mathbb{P} = \mathcal{P}_d^{k-1}(\mathcal{T}_j)$, then this is obvious (case (2.6)), and if $P \in \mathbb{P} = \mathcal{P}^l(\mathcal{T}_j)$, then P does not jump and $\mathbf{J} = [\![\nabla\mathbf{U}_j\cdot\mathbf{n}]\!]$ (case (2.5)). Therefore, $\mathbf{J}$ admits a piecewise polynomial extension to $\mathcal{F}^i(S)$ of degree at most $k-1$, which is still denoted by $\mathbf{J}$ (simply scale to the master element and extend $\mathbf{J}$ as a constant in the direction normal to S).

Now let φ_S be the continuous piecewise linear basis function of the triangulation $\mathcal{T}^{i+1}$ corresponding to the node inside S; thus $\operatorname{supp}\varphi_S \subset \mathcal{F}^i(S)$. Hence $\mathbf{J}\varphi_S \in \mathbb{V}^{i+1}$ and $\operatorname{supp}(\mathbf{J}\varphi_S) \subset \mathcal{F}^i(S)$. Since $\|\mathbf{J}\|_S^2$ is equivalent to $\int_S |\mathbf{J}|^2\varphi_S$, integrating by parts and using the fact that $\mathbf{J}\varphi_S \in \mathbb{V}^{i+1}$, we obtain

$$\begin{aligned}\|\mathbf{J}\|_S^2 &\le C\int_S |\mathbf{J}|^2\varphi_S = \int_S [\![\nabla\mathbf{U}^i\cdot\mathbf{n} - P\mathbf{n}]\!]\cdot\mathbf{J}\varphi_S\\ &= -\int_{\mathcal{F}^i(S)} \mathbf{R}\cdot\mathbf{J}\varphi_S + \int_{\mathcal{F}^i(S)} \nabla(\mathbf{U}^{i+1}-\mathbf{U}^i):\nabla(\mathbf{J}\varphi_S)\\ &\le \|\mathbf{R}\|_{\mathcal{F}^i(S)}\|\mathbf{J}\varphi_S\|_{\mathcal{F}^i(S)} + \big\|\nabla(\mathbf{U}^{i+1}-\mathbf{U}^i)\big\|_{\mathcal{F}^i(S)}\|\nabla(\mathbf{J}\varphi_S)\|_{\mathcal{F}^i(S)}\\ &\le \left(\|\mathbf{R}\|_{\mathcal{F}^i(S)} + \frac{1}{h_S}\big\|\nabla(\mathbf{U}^{i+1}-\mathbf{U}^i)\big\|_{\mathcal{F}^i(S)}\right)\|\mathbf{J}\varphi_S\|_{\mathcal{F}^i(S)}.\end{aligned}$$

Multiplying by h_S and using the equivalence of $\|\mathbf{J}\|_S$ and $\frac{1}{h_T}\|\mathbf{J}\|_{\mathcal{F}^i(S)}^2$, we arrive at the desired estimate (5.16).

To complete the proof, we let $T \in \mathcal{T}^i$ satisfy the assumption that all elements $T' \in \mathcal{F}^i(T)$ possess the interior node property (5.7). We realize that for all those T' we can apply (5.15) and next insert the bound for $\|\mathbf{R}_{T'}\|_{T'}$ into (5.16). ∎

Remark 5.5. The above proof uncovers the need for the relation $k \ge l$ between the polynomial degrees k for velocity and l for pressure. If this were not true, then ∇P would differ from $(\nabla P)_T$ and we should then account for the oscillation $\nabla P - (\nabla P)_T$, which is not given data. Since procedure UPDATE reveals no accuracy gain for $l \ge k$, our assumption $l \le k$ in (2.5) and (2.6) is not a serious restriction.

5.2. Complexity. We now turn to the analysis of the complexity of ELLIPTIC. We first observe that the lower bound (5.5), together with the rates of convergence (5.9) and of oscillation reduction (5.8), implies

$$\begin{aligned}\sum_{T\in\mathcal{T}^i}\eta^i(T)^2 &\le K_*(d+2)\big(\left\|\nabla(\mathbf{u}_j-\mathbf{U}^i)\right\|^2+\operatorname{osc}(\mathbf{f},\mathcal{T}^i)^2\big)\\ &\le K_1\rho^{2i}\left\|\nabla(\mathbf{u}_j-\mathbf{U}^0)\right\|^2+K_2\rho^{2i}\operatorname{osc}(\mathbf{f},\mathcal{T}^0)^2,\end{aligned}\tag{5.17}$$

where the constants K_1, K_2 depend only on mesh-regularity and the parameters θ, θ_f of ELLIPTIC. Therefore, the stopping criterion in step 3 of ELLIPTIC can be fulfilled in a finite number of iterations.

A fundamental question that remains open is whether this number can be *bounded uniformly* with respect to the outer iteration counter j. The answer is affirmative and is the subject of the following statement.

PROPOSITION 5.6. *Let the tolerance reduction factor γ of the AUA satisfy $\gamma>\beta$, where $\beta=\left\|I-\alpha\mathcal{S}\right\|_{\mathcal{L}(\mathbb{P},\mathbb{P})}<1$ is defined in* (2.4)*. Then, the number of iterations in the inner loop of* ELLIPTIC *is bounded by a constant which depends only on* $\mathbf{f}$*, the initial pressure guess P_0, the initial triangulation $\mathcal{T}_0$ of the AUA, the ratio β/γ, and the parameters θ and θ_f of* ELLIPTIC*, but not on the outer index j.*

Proof. By the preceding comment, the number of iterations of ELLIPTIC is bounded for all outer counters j. It is thus sufficient to consider the case $j>1$.

We need to prove the existence of a constant C such that for some $i<C$,

$$\sum_{T\in\mathcal{T}^i}\eta^i(T)^2\le\varepsilon^2.$$

We recall that the initial mesh $\mathcal{T}^0$ of the inner loop is always taken to be the mesh $\mathcal{T}_{j-1}$ of the previous outer loop. Since the term $\Delta\mathbf{U}_j-\nabla P_{j-1}$ of the interior residual $\mathbf{R}_{j-1}$ does not oscillate, using definition (4.4), we get

$$\begin{aligned}\operatorname{osc}(\mathbf{f},\mathcal{T}^0)=\operatorname{osc}(\mathbf{R}_{j-1},\mathcal{T}_{j-1}) &\le\left(\sum_{T\in\mathcal{T}_{j-1}}h_T^2\|\mathbf{R}_{j-1}\|_T^2\right)^{1/2}\\ &\le\left(\sum_{T\in\mathcal{T}_{j-1}}h_T^2\|\mathbf{R}_{j-1}\|_T^2+h_T\|\mathbf{J}_{j-1}\|_{\partial T}^2\right)^{1/2}\\ &=\left(\sum_{T\in\mathcal{T}_{j-1}}\eta_{j-1}(T)^2\right)^{1/2}\le\varepsilon_{j-1}=\varepsilon_0\gamma^{j-1},\end{aligned}$$

where the last inequality is guaranteed by the stopping criterion (step 3) of ELLIPTIC. This accounts for the second term in (5.17).

To estimate $\left\|\nabla(\mathbf{u}_j-\mathbf{U}^0)\right\|$ in (5.17), we first split it into three parts:

$$\left\|\nabla(\mathbf{u}_j-\mathbf{U}^0)\right\|\le\|\nabla(\mathbf{u}_j-\mathbf{u}_{j-1})\|+\|\nabla(\mathbf{u}_{j-1}-\mathbf{U}_{j-1})\|+\left\|\nabla(\mathbf{U}_{j-1}-\mathbf{U}^0)\right\|.$$

By virtue of (2.9), we have

$$\|\nabla(\mathbf{u}_{j-1}-\mathbf{U}_{j-1})\|\le C\varepsilon_{j-1}=C\varepsilon_0\gamma^{j-1}.$$

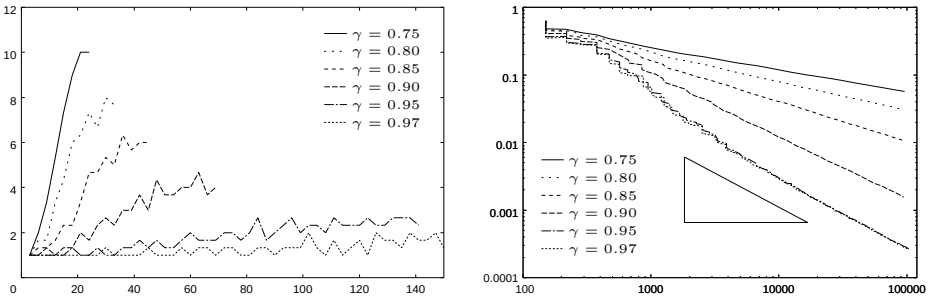


FIG. 5.1. *Example* 3.2: *Number of inner iterations versus outer interations (left) and error decay versus DOFs (right) for different values of* γ. *A triangle shows the optimal decay.*

Since $\mathbb{V}^0 = \mathbb{V}_{j-1}$, both $\mathbf{U}^0$ and $\mathbf{U}_{j-1}$ belong to $\mathbb{V}_{j-1}$ and, by (2.8), they satisfy

$$\int_\Omega \nabla(\mathbf{U}^0 - \mathbf{U}_{j-1}) : \nabla \mathbf{V} = \int (P_{j-1} - P_{j-2}) \operatorname{div} \mathbf{V} \quad \forall\, \mathbf{V} \in \mathbb{V}_{j-1}.$$

Hence, taking $\mathbf{V} = \mathbf{U}^0 - \mathbf{U}_{j-1}$ implies

$$\left\|\nabla(\mathbf{U}_{j-1} - \mathbf{U}^0)\right\| \le \|P_{j-1} - P_{j-2}\| \le \|P_{j-1} - p\| + \|p - P_{j-2}\|.$$

A similar energy argument, based on the fact that $\mathbf{u}_j$ is the solution to (2.7), yields the same estimate for $\|\nabla(\mathbf{u}_j - \mathbf{u}_{j-1})\|$. To derive a suitable estimate for $\|p - P_j\|$, we now improve (2.12) as follows:

$$\|p - P_j\| \le \beta^j \|p - P_0\| + C\alpha\varepsilon_0 \sum_{\ell=0}^{j-1} \beta^\ell \gamma^{j-\ell} \le \beta^j \|p - P_0\| + C\alpha\varepsilon_0 \gamma^j \frac{1 - (\beta/\gamma)^j}{1 - (\beta/\gamma)}.$$

Inserting the previous estimates back into (5.17), we find a constant K, depending on f, the initial pressure guess P_0, the initial triangulation $\mathcal{T}_0$ of the AUA, the parameters θ, θ_f of ELLIPTIC, and the ratio $\beta/\gamma < 1$, such that

$$\left(\sum_{T \in \mathcal{T}^i} \eta^i(T)^2 \right)^{1/2} \le K \rho^i \gamma^j. \tag{5.18}$$

Therefore, $K\rho^i\gamma^j \le \varepsilon_j = \varepsilon_0\gamma^j$ whenever $\rho^i \le \varepsilon_0/K$, and the assertion is proved. ∎

Remark 5.6. It might seem at first sight that $\gamma > \beta$ is an artificial requirement of the proof and thus that the result should still hold for any $\gamma < 1$. If $\gamma < \beta$, then the above proof would also give (5.18) with β instead of γ, whence

$$i \le C_1 j + C_2$$

for appropriate constants $C_1, C_2 > 0$. This linear growth is corroborated by the simulations leading to Figure 5.1, which depicts the number of inner loops i versus the outer loop counter j for Example 3.2 with the Taylor–Hood element $\mathcal{P}^2$-$\mathcal{P}^1$.

Remark 5.7. Since β is not known in general, the requirement $\beta < \gamma < 1$ may seem restrictive in practice. On the other hand, a value of γ too close to 1 results in a large number of outer iterations. We found a practical compromise $\gamma = 0.95$ for all simulations of section 3 that leads to a number of inner iterations between 3 and 5.

Acknowledgment. The authors would like to thank Kunibert G. Siebert for his assistance with the implementation of discontinuous finite elements within ALBERT and for many valuable suggestions and discussions about this work.

REFERENCES

[1] M. Ainsworth and J. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley, New York, 2000.

[2] E. Bänsch, *Local mesh refinement in* 2 *and* 3 *dimensions*, Impact Comput. Sci. Engrg., 3 (1991), pp. 181–191.

[3] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[4] P. Clément, *Approximation by finite element functions using local regularizations*, Rev. Française Automat. Informat. Recherche Opérationalle Sér. Rouge Anal. Numér., 9 (1975), pp. 77–84.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive Wavelet Methods* II—*Beyond the Elliptic Case*, IGPM Report, Rheinisch-Westfälische Technische Hochschule (RWTH), Aachen, Germany, 2000.

[6] M. Crouzeix, *Etude d'une méthode de linéarisation. Résolution numérique des équations de Stokes stationnaires. Application aux équations de Navier-Stokes stationnaires*, in Approximation et Méthodes Itératives de Résolution d'Inéquations Variationnelles et de Problèmes Non Linéaires, IRIA Cahier 12, Inst. Recherche Informat. Automat., Rocquencourt, France, 1974, p. 139.

[7] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems—optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[8] S. Dahlke, R. Hochmuth, and K. Urban, *Adaptive wavelet methods for saddle point problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1003–1022.

[9] S. Dahlke, R. Hochmuth, and K. Urban, *Convergent adaptive wavelet methods for the Stokes problem*, in Multigrid Methods VI, E. Dick, K. Riemslagh, and J. Vierendeels, eds., Springer-Verlag, Berlin, 2000, pp. 66–72.

[10] M. Dauge, *Stationary Stokes and Navier-Stokes systems on two- or three-dimensional domains with corners. Part* I: *Linearized equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97.

[11] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[12] H.C. Elman and G.H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle-point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[13] GRAPE—*GRAphics Programming Environment*, Manual, Version 5.0 SFB 256, University of Bonn, Bonn, Germany, 1995.

[14] D. Kay and D. Silvester, *A posteriori error estimation for stabilized mixed approximations of the Stokes equations*, SIAM J. Sci. Comput., 21 (1999), pp. 1321–1336.

[15] P. Morin, R.H. Nochetto, and K.G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[16] P. Morin, R.H. Nochetto, and K.G. Siebert, *Basic Principles for Convergence of Adaptive Higher-Order FEM*, manuscript.

[17] J-H. Pyo, *The Gauge-Uzawa and Related Projection Finite Element Methods for the Evolution Navier-Stokes Equations*, Ph.D. Dissertation, University of Maryland, College Park, MD, 2002.

[18] A. Schmidt and K.G. Siebert, *ALBERT—software for scientific computations and applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2000), pp. 105–122.

[19] A. Schmidt and K.G. Siebert, *Concepts of the finite element toolbox ALBERT*, Notes Numer. Fluid Mech., to appear.

[20] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley, New York, 1996.

[21] R. Verfürth, *A posteriori error estimators or the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1230–1262

ADAPTIVE WAVELET METHODS FOR SADDLE POINT PROBLEMS—OPTIMAL CONVERGENCE RATES*

STEPHAN DAHLKE†, WOLFGANG DAHMEN‡, AND KARSTEN URBAN§

Abstract. In this paper an adaptive wavelet scheme for saddle point problems is developed and analyzed. Under the assumption that the underlying continuous problem satisfies the inf-sup condition, it is shown in the first part under which circumstances the scheme exhibits asymptotically optimal complexity. This means that within a certain range the convergence rate which relates the achieved accuracy to the number of involved degrees of freedom is asymptotically the same as the error of the best wavelet N-term approximation of the solution with respect to the relevant norms. Moreover, the computational work needed to compute the approximate solution stays proportional to the number of degrees of freedom. It is remarkable that compatibility constraints on the trial spaces such as the Ladyzhenskaya–Babuška–Brezzi (LBB) condition do not arise. In the second part the general results are applied to the Stokes problem. Aside from the verification of those requirements on the algorithmic ingredients the theoretical analysis had been based upon, the regularity of the solutions in certain Besov scales is analyzed. These results reveal under which circumstances the work/accuracy balance of the adaptive scheme is even asymptotically better than that resulting from preassigned uniform refinements. This in turn is used to select and interpret some first numerical experiments that are to quantitatively complement the theoretical results for the Stokes problem.

Key words. saddle point problems, wavelet bases, norm equivalences, adaptive refinements, fast approximate operator application, Uzawa iteration

AMS subject classifications. 35B65, 41A25, 41A46, 42C15, 46E35, 65F99, 65N12, 65N55

PII. S003614290139233X

1. Introduction.

This paper draws on two major sources of motivation. First, it has recently been shown in [8] that certain adaptive wavelet schemes are asymptotically optimal for a wide class of self-adjoint elliptic operator equations. This means that the achieved accuracy in the energy norm expressed in terms of the numbers of involved degrees of freedom is asymptotically the same as the rate of the *best N-term approximation*, i.e., the minimal number of basis functions needed to approximate the solution within the given accuracy tolerance. Moreover, (up to additional log-factors in sorting operations; see also Remark 4.10 below) it was shown that the computational work needed to compute the approximate solution stays proportional to the number of degrees of freedom. While the class of operator equations covers boundary value problems for partial differential equations as well as singular integral equations, symmetry did play a crucial role in the analysis and design of the scheme. These techniques have meanwhile been extended to noncoercive problems through *wavelet least squares formulations* [9].

Second, in [14] the results of a predecessor [13] of [8] also for the symmetric elliptic case have been extended to saddle point problems. The key idea there was

*Received by the editors July 12, 2001; accepted for publication (in revised form) April 8, 2002; published electronically, September 27, 2002. This work has been supported in part by the Deutsche Forschungsgemeinschaft grants Da 117/8–2, SFB 401 and the TMR Network "Wavelets in Numerical Simulation" funded by the European Commission.

http://www.siam.org/journals/sinum/40-4/39233.html

†Department of Mathematics and Computer Science, Philipps-University Marburg, Hans Meerwein Str., Lahnberge, D-35032 Marburg, Germany (dahlke@mathematik.uni-marburg.de).

‡RWTH Aachen, Institut für Geometrie und Praktische Mathematik, Templergraben 55, 52056 Aachen, Germany (dahmen@igpm.rwth-aachen.de).

§Department of Numerical Analysis, University of Ulm, D-89069 Ulm, Germany (kurban@mathematik.uni-ulm.de).

to use an outer Uzawa iteration and to solve the interior symmetric positive definite problems by a scheme of the type considered in [13]. However, no statements about the efficiency of such schemes in terms of convergence rates and work count was made in [14].

In this paper we also consider saddle point problems actually under slightly weaker assumptions than in [14] and propose an adaptive wavelet scheme for their numerical solution. In order to avoid (among other things) the squaring of condition numbers, it is based as in [14] on an outer Uzawa iteration although it differs from the scheme in [14] in several essential ways. It draws on detailed algorithmic ingredients from [8] which allow one to quantify concrete computational steps and estimate their complexity, which results in a somewhat different balance of accuracies. It also applies when the symmetric bilinear form is only elliptic on the kernel of the constraint operator.

On a more fundamental level, in the same spirit as in [8, 9], there are two essential features that distinguish the present approach from [13, 14] and more so from classical discretization. The first one is that through appropriate wavelet bases the original continuous problem is transformed right from the beginning into an *equivalent* problem which is *well-posed in the Euclidean metric*. All essential computational steps refer then to approximation in ℓ_2 and therefore bear a great potential of being portable to other problem classes. In fact, many of the basic routines developed in [2, 8] in the context of elliptic problems can be used here as well. The second important point is that the wavelet representation allows us to think of performing, up to a controlled perturbation, an iteration on the full infinite dimensional problem realized through the *adaptive approximate application* of the full infinite dimensional operators. The tolerances have to be chosen so that the convergence speed of the perturbed realizable iteration is indeed governed by the properties of the ideal infinite dimensional iteration.

This offers, in particular, a first intuitive explanation for the following fact which at first glance strikes one as a paradox; namely, compatibility constraints on the choice of trial functions such as the *Ladyzhenskaya–Babuška–Brezzi (LBB) condition* do *not* arise. In fact, recall that even when the infinite dimensional saddle point problem is well-posed and hence satisfies an inf-sup condition, inappropriate choices of finite dimensional trial spaces could lead to discrete problems with poor stability properties; that is, the inverses of the corresponding system matrices may have arbitrarily large norms. This fact is relevant whenever linear systems are to be solved for any such given pair of trial spaces. In the present context this situation will never arise. Instead an iterative process is conceptually applied to the *full infinite dimensional problem*, where each iteration involves an adaptive application of the underlying infinite dimensional operators within a certain stage dependent dynamic accuracy tolerance. This process is inherently nonlinear. Roughly speaking, proper adaptation in the above sense inherits the stability of the infinite dimensional problem. In this sense adaptation not only reduces complexity but also *stabilizes* the computation *automatically*.

The paper is organized as follows. After formulating the problem in section 2 we describe and analyze an adaptive method in section 3. It will be shown in section 4 under which conditions on the algorithmic ingredients it exhibits an *asymptotically optimal accuracy/work balance* in the following sense. Whenever the exact solution has, within a certain range of exponents s, an error of *best N-term approximation* with respect to an underlying wavelet basis decaying like N^{-s}, then the error achieved by the adaptive scheme also decays like N^{-s}, where N is the number of used degrees of freedom. Moreover, the computational work stays proportional to N. A key role

in this context is played by the *compressibility range* of the involved operators in wavelet coordinates. Given this property, one can employ a certain adaptive scheme for applying the operator to any finitely supported vector with optimal accuracy/work balance [8].

In section 5 the general results are applied to the *Stokes problem*. Specifically, we investigate in section 5.3 the compressibility range of the wavelet representation of the Stokes operator for a certain family of wavelet bases and derive sharp estimates for this range. This identifies the range of decay rates for which the general results from the preceding sections apply.

It should be stressed that the scheme works without *any* a priori assumptions on the solution, while its complexity is analyzed under the assumption that the solution has a certain order of best N-term approximations and the involved operators in wavelet coordinates have a certain compressibility range (see section 4). Certain rates of the decay of best N-term approximation, in turn, are (almost) equivalent to a certain regularity of the solution in a *Besov scale*. Roughly speaking, when the Sobolev regularity of the solution is lower than its Besov regularity, the adaptive scheme is expected to offer even an *asymptotically better* accuracy/work balance than linear schemes. To see whether or under which circumstances the adaptive scheme can be rigorously proven to offer even an asymptotically better accuracy/work balance than schemes based on uniform preassigned mesh refinements, we investigate in section 5.4 the Besov regularity of singularity solutions for the Stokes problem. The results show that in two spatial dimensions sufficiently high order wavelet bases would give rise to adaptive schemes with *arbitrarily* high convergence rates.

Finally, in section 6 we present some numerical experiments essentially guided by the above-mentioned theoretical considerations. Here we make use of the software developed in [2] as well as in [25]. The results confirm that the adaptive scheme performs essentially independently of the pairing of trial functions for velocities and pressure. For instance, the rate of decay of the best N-term approximation is met within a factor two when both velocities and pressure are approximated by piecewise linear trial functions.

After completion of this work we became aware of related investigations in [4] pursuing similar ideas in a finite element context. There, convergence in the sense of [14] is proven for a similar Uzawa technique without establishing, however, rigorous estimates for the corresponding work/accuracy balance.

2. Saddle point problems.

2.1. The setting. Let X, M denote Hilbert spaces with norms $\|\cdot\|_X, \|\cdot\|_M$, respectively. Dual pairings on $X \times X'$ and $M \times M'$ (X', M' denoting the duals of X, M, respectively) will always be denoted by $\langle\cdot,\cdot\rangle$. It will be clear from the context which spaces are referred to. Suppose that $a(\cdot,\cdot)$ is a continuous symmetric bilinear form on $X \times X$ and that $b(\cdot,\cdot)$ is a continuous bilinear form on $X \times M$; i.e.,

$$|a(v,w)| \lesssim \|v\|_X \|w\|_X, \quad |b(q,v)| \lesssim \|v\|_X \|q\|_M.$$

We shall often write $A \lesssim B$ to indicate the existence of an absolute constant $c > 0$ such that $A \le cB$. In addition, $A \sim B$ means $A \lesssim B \lesssim A$.

Moreover, denoting by $B : X \to M'$ the operator induced by $b(p,v) = \langle p, Bv\rangle$ and setting $V := \ker B$, assume that $a(\cdot,\cdot)$ is elliptic on V and $b(\cdot,\cdot)$ satisfies the *inf-sup*

condition

$$a(v,v) \geq \alpha \|v\|_X^2, \quad v \in V, \qquad \inf_{q \in M} \sup_{v \in X} \frac{b(q,v)}{\|v\|_X \|q\|_M} > \beta, \qquad \alpha, \beta > 0. \tag{2.1.1}$$

It is well known that then the variational problem

$$\begin{array}{rcl} a(u,v) + b(p,v) & = & \langle f, v\rangle, \quad v \in X, \\ b(q,u) & = & \langle q, g\rangle, \quad q \in M, \end{array} \tag{2.1.2}$$

has a unique solution $U = (u,p) \in X \times M$ for any $f \in X'$, $g \in M'$; see, e.g., [5]. Defining $A : X \to X'$ by $a(v,w) = \langle v, Aw\rangle$, $v \in X$, (2.1.2) is equivalent to the 2×2 block operator equation

$$\mathcal{L}U := \begin{pmatrix} A & B' \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} =: F, \tag{2.1.3}$$

where $\mathcal{L}$ is an isomorphism from $X \times M$ into its dual $X' \times M'$; i.e., there exist positive constants $c_{\mathcal{L}}, C_{\mathcal{L}}$ such that

$$c_{\mathcal{L}} \left(\|v\|_X^2 + \|q\|_M^2 \right)^{1/2} \leq \left\| \mathcal{L} \begin{pmatrix} v \\ q \end{pmatrix} \right\|_{X' \times M'} \leq C_{\mathcal{L}} \left(\|v\|_X^2 + \|q\|_M^2 \right)^{1/2}. \tag{2.1.4}$$

Classical examples are mixed formulations of second order elliptic boundary value problems, the Stokes problem, or the system obtained when appending essential boundary conditions by Lagrange multipliers.

2.2. Wavelet coordinates. Now suppose that we have wavelet bases $\Psi_X = \{\psi_{X,\lambda} : \lambda \in \mathcal{J}_X\}$, $\Psi_M = \{\psi_{M,\lambda} : \lambda \in \mathcal{J}_M\}$ for X and M at our disposal ($\mathcal{J}_X$ and $\mathcal{J}_M$ being the corresponding index sets) such that for suitable diagonal matrices $\mathbf{D}_X$, $\mathbf{D}_M$ and constants c_X, C_X, c_M, C_M one has

$$c_X \|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \leq \|\mathbf{v}^T \mathbf{D}_X^{-1} \Psi_X\|_X \leq C_X \|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)}, \tag{2.2.1}$$

and likewise

$$c_M \|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)} \leq \|\mathbf{q}^T \mathbf{D}_M^{-1} \Psi_M\|_M \leq C_M \|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)}, \tag{2.2.2}$$

where $\mathbf{v}^T \mathbf{D}_X^{-1} \Psi_X := \sum_{\lambda \in \mathcal{J}_X} d_{X,\lambda}^{-1} v_\lambda \psi_{X,\lambda}$. The validity of such norm equivalences will be crucial in what follows. Note that often M is a closed subspace of finite codimension in a larger Hilbert space $\hat{M}$ for which (2.2.2) holds. For instance, in the case of the Stokes problem, M is the space of all L_2 functions with zero mean. Thus the arrays of wavelet coefficients of elements in M will, in general, form a closed subspace $\ell_{2,0}(\mathcal{J}_M)$ of finite codimension in $\ell_2(\mathcal{J}_M)$.

At this point we dispense with any additional technical details about the precise nature of the basis functions but refer to [7, 15] for surveys and further references; see also the comments in connection with numerical realizations below. A further important property is the *cancellation property* which entails near sparseness of wavelet representations for many operators. This also will be detailed when necessity arises.

Defining now for any two countable arrays Θ, Φ and some inner product $c(\cdot,\cdot)$ the matrix $c(\Theta,\Phi) := (c(\theta,\phi))_{\theta \in \Theta, \phi \in \Phi}$, consider as usual the scaled wavelet representations

$$\mathbf{A} := a(\mathbf{D}_X^{-1} \Psi_X, \mathbf{D}_X^{-1} \Psi_X), \quad \mathbf{B} := b(\mathbf{D}_M^{-1} \Psi_M, \mathbf{D}_X^{-1} \Psi_X), \tag{2.2.3}$$

as well as the arrays $\mathbf{f} := \mathbf{D}_X^{-1}\langle \Psi_X, f\rangle$, $\mathbf{g} := \mathbf{D}_M^{-1}\langle \Psi_M, g\rangle$, and $\mathbf{F} := (\mathbf{f}^T, \mathbf{g}^T)^T$. Then (2.1.2) or (2.1.3) is equivalent to the following two by two block matrix system:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{2.2.4}$$

It will make things much more transparent when working from now on exclusively in the ℓ_2 setting.

2.3. Well-posedness in ℓ_2. It follows from (2.2.1) and (2.2.2) together with (2.1.4) that the operator

$$\mathbf{L} := \begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} : \ell_2(\mathcal{J}) := \ell_2(\mathcal{J}_X) \times \ell_{2,0}(\mathcal{J}_M) \to \ell_2(\mathcal{J}), \quad \mathcal{J} := \mathcal{J}_X \times \mathcal{J}_M,$$

is an isomorphism; i.e., there exist positive constants c_L, C_L such that for $\mathbf{V} := (\mathbf{v}^T, \mathbf{q}^T)^T \in \ell_2(\mathcal{J})$, $\|\mathbf{V}\|^2_{\ell_2(\mathcal{J})} = \|\mathbf{v}\|^2_{\ell_2(\mathcal{J}_X)} + \|\mathbf{q}\|^2_{\ell_2(\mathcal{J}_M)}$

$$c_L\|\mathbf{V}\|_{\ell_2(\mathcal{J})} \le \|\mathbf{L}\mathbf{V}\|_{\ell_2(\mathcal{J})} \le C_L\|\mathbf{V}\|_{\ell_2(\mathcal{J})}, \quad \mathbf{V} \in \ell_2(\mathcal{J}); \tag{2.3.1}$$

see, e.g., [15, 21] for further details. Clearly, c_L, C_L can be expressed in terms of the constants $c_{\mathcal{L}}, C_{\mathcal{L}}$, c_Y, C_Y for $Y \in \{X, M\}$. Furthermore, there exist constants C_B, C'_A such that

$$\|\mathbf{B}\mathbf{v}\|_{\ell_2(\mathcal{J}_M)} \le C_B\|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)}, \quad \|\mathbf{B}^T\mathbf{q}\|_{\ell_2(\mathcal{J}_X)} \le C_B\|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)}, \tag{2.3.2}$$

and

$$\|\mathbf{A}\mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \le C'_A\|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)}. \tag{2.3.3}$$

2.4. The Schur complement. In many cases a somewhat stronger property than the first relation in (2.1.1) is valid; namely, that

$$a(v, v) \sim \|v\|^2_X, \qquad v \in X, \tag{2.4.1}$$

which, of course means that $\mathbf{A}$ is invertible on all of $\ell_2(\mathcal{J}_X)$. In this case, block elimination reduces (2.2.4) to the so-called reduced system

$$\mathbf{S}\mathbf{p} = \mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{g}, \tag{2.4.2}$$

involving the (infinite dimensional) *Schur complement*

$$\mathbf{S} := \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T : \ell_{2,0}(\mathcal{J}_M) \to \ell_{2,0}(\mathcal{J}_M), \tag{2.4.3}$$

which is symmetric positive definite and, under the above assumptions, in fact an automorphism on $\ell_{2,0}(\mathcal{J}_M)$; i.e., there exist positive constants c_S, C_S such that

$$c_S\|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)} \le \|\mathbf{S}\mathbf{q}\|_{\ell_2(\mathcal{J}_M)} \le C_S\|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)}, \quad \mathbf{q} \in \ell_{2,0}(\mathcal{J}_M). \tag{2.4.4}$$

Once $\mathbf{p}$ has been determined from (2.4.2) it remains to solve the positive definite problem

$$\mathbf{A}\mathbf{u} = \mathbf{f} - \mathbf{B}^T\mathbf{p}. \tag{2.4.5}$$

However, under the weaker assumption (2.1.1) on the bilinear form $a(\cdot,\cdot)$ one first has to take a precaution whose variational counterpart is sometimes referred to as *augmented Lagrangian method*. In the present setting it boils down to considering the matrix

$$\hat{\mathbf{A}} := \mathbf{A} + c\mathbf{B}^T\mathbf{B}, \tag{2.4.6}$$

where c is some sufficiently large but fixed positive constant.

REMARK 2.1. *Assume that* (2.1.1) *hold. Let* c *in* (2.4.6) *be any fixed positive constant, when the operator* A *in* (2.1.3) *is positive semidefinite on* X*, and otherwise satisfy* $c > C_B^2 C'_A/c_L^4$*, where* c_L, C_B, C'_A *are the constants from* (2.4.3), (2.3.2), *and* (2.3.3)*, respectively. Then the matrix* $\hat{\mathbf{A}}$ *is an automorphism on* $\ell_2(\mathcal{J}_X)$*; i.e., there exist positive constants* c_A, C_A *such that*

$$c_A\|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \le \|\hat{\mathbf{A}}\mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \le C_A\|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)}, \quad \mathbf{v} \in \ell_2(\mathcal{J}_X). \tag{2.4.7}$$

Proof. In order to identify concrete conditions on c for later purposes, we include the proof although it is in principle standard. It follows from (2.3.2) and (2.3.3) that $\hat{\mathbf{A}}$ is bounded on $\ell_2(\mathcal{J}_X)$. Moreover, by (2.3.1) the matrix $\mathbf{L}^T\mathbf{L} = \mathbf{L}^2$ is positive definite on $\ell_2(\mathcal{J})$. Since $\mathbf{B}\mathbf{B}^T$ is a principal block of $\mathbf{L}^2$, it is positive definite on $\ell_{2,0}(\mathcal{J}_M)$. This entails that $\hat{\mathbf{A}}$ is also injective on $\ell_2(\mathcal{J}_X)$. To see this, note that by the first relation in (2.1.1), $\mathbf{v}^T\hat{\mathbf{A}} \neq \mathbf{0}$ for $\mathbf{v} \in \ker\mathbf{B}$. On the other hand, when $\mathbf{v}$ is in the range of $\mathbf{B}^T$, i.e., $\mathbf{v} = \mathbf{B}^T\mathbf{q}$ for some $\mathbf{q} \in \ell_{2,0}(\mathcal{J}_M)$, then one has

$$\mathbf{v}^T\hat{\mathbf{A}}\mathbf{v} = \mathbf{q}^T\mathbf{B}\mathbf{A}\mathbf{B}^T\mathbf{q} + c\|\mathbf{B}\mathbf{B}^T\mathbf{q}\|^2_{\ell_2(\mathcal{J}_M)}. \tag{2.4.8}$$

Noting that, by (2.3.1), $\|\mathbf{B}\mathbf{B}^T\mathbf{q}\|^2_{\ell_2(\mathcal{J}_M)} \ge c_L^4\|\mathbf{q}\|^2_{\ell_2(\mathcal{J}_M)}$, (2.4.8) is readily seen to be strictly positive under the above assumptions on c whenever $\mathbf{q} \neq \mathbf{0}$. This confirms the injectivity of $\hat{\mathbf{A}}$ on $\ell_2(\mathcal{J}_X)$. By symmetry, (2.4.8) also implies surjectivity. Due to the boundedness of $\hat{\mathbf{A}}$, the claim follows now from the inverse mapping theorem. ∎

One easily verifies that (2.2.4) is equivalent to the system

$$\begin{pmatrix} \hat{\mathbf{A}} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{f}} \\ \mathbf{g} \end{pmatrix}, \qquad \hat{\mathbf{f}} := \mathbf{f} + c\mathbf{B}^T\mathbf{g}, \tag{2.4.9}$$

where $\hat{\mathbf{A}}$ is given by (2.4.6). By Remark 2.1, for suitable $c > 0$, block elimination can be applied to this new system (2.4.9), which then reduces to the coupled systems (2.4.2), (2.4.5) with $\mathbf{A}$ and $\mathbf{f}$ replaced by $\hat{\mathbf{A}}$ and $\hat{\mathbf{f}}$, respectively.

To simplify notation we will use the following convention throughout the remainder of the paper. We will always set

$$\mathbf{A} := \mathbf{D}_X^{-1}a(\Psi_X,\Psi_X)\mathbf{D}_X^{-1} + c\mathbf{B}^T\mathbf{B}, \quad \mathbf{f} := \mathbf{D}_X^{-1}\langle\Psi_X, f\rangle + c\mathbf{B}^T\mathbf{g}, \tag{2.4.10}$$

with $\mathbf{B} := \mathbf{D}_M^{-1}b(\Psi_M,\Psi_X)\mathbf{D}_X^{-1}$ as in (2.2.3) and $\mathbf{g} := \mathbf{D}_M^{-1}\langle\Psi_M, g\rangle$. When the bilinear form $a(\cdot,\cdot)$ satisfies the stronger assumption (2.4.1), the constant c in (2.4.10) can be chosen to be zero. Otherwise, c will always be assumed in the following to be a fixed positive constant as specified in Remark 2.1. Thus, without loss of generality, we can always make use of the reduced systems (2.4.3), (2.4.5) with a proper interpretation of the matrix $\mathbf{A}$ according to the above convention. Consequently, $\mathbf{A}$ satisfies in this sense (2.4.7).

A standard way of formulating finite dimensional problems is to take Galerkin discretizations for (2.1.2). As soon as one *fixes* a pair of finite dimensional trial spaces in X and M, for instance, spanned by collections of wavelets, the corresponding Galerkin discretization gives rise to a finite dimensional linear system, e.g., in terms of a principal finite submatrix of (2.2.4). However, it is well known that stability of the infinite dimensional problem does not guarantee the finite dimensional problems to be uniformly stable as well. Compatibility constraints in terms of the LBB condition come into play. It will be seen that this will *not* be the case in the following adaptive framework.

3. An adaptive Uzawa strategy.

3.1. Infinite dimensional Uzawa iteration. The idea is to use a stationary iterative scheme for the solution of the reduced system (2.4.2), which is essentially the Uzawa strategy proposed in [14]. In contrast, we formulate it here directly for the discrete infinite dimensional ℓ_2-problem (2.2.4). To this end, we first have to address an issue which is somewhat hidden in the ℓ_2-setting. The spaces X, M are always function spaces on some domain Ω. As will be explained in more detail later, the wavelet bases Ψ_X and Ψ_M are then typically constructed as *Riesz bases* for the corresponding spaces $L_2(\Omega)$; i.e., in addition to the norm equivalences (2.2.1), (2.2.2), one also has

$$\|\mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \sim \|\mathbf{v}^T\Psi_X\|_{L_2(\Omega)}, \quad \|\mathbf{q}\|_{\ell_2(\mathcal{J}_M)} \sim \|\mathbf{q}^T\Psi_M\|_{L_2(\Omega)}. \tag{3.1.1}$$

This means that there exist *dual* bases $\tilde\Psi_X$, $\tilde\Psi_M$ in $L_2(\Omega)$ which are also Riesz bases and satisfy

$$(\Psi_X, \tilde\Psi_X) = \mathbf{id}, \quad (\Psi_M, \tilde\Psi_M) = \mathbf{id}, \tag{3.1.2}$$

where $(\cdot,\cdot)$ denotes the standard inner product in $L_2(\Omega)$. In full agreement with the fact that the operator B maps X onto M', one observes that for $v = \mathbf{v}^T\mathbf{D}_X^{-1}\Psi_X$ the array $\mathbf{Bv}$ represents expansion coefficients of Bv with respect to the *dual basis* $\tilde\Psi_M$. In fact,

$$\begin{aligned}(\mathbf{Bv})^T\mathbf{D}_M\tilde\Psi_M &= \mathbf{v}^T\langle B\mathbf{D}_X^{-1}\Psi_X, \Psi_M\rangle\mathbf{D}_M^{-1}\mathbf{D}_M\tilde\Psi_M = \mathbf{v}^T\langle B\mathbf{D}_X^{-1}\Psi_X, \Psi_M\rangle\tilde\Psi_M \\ &= B(\mathbf{v}^T\mathbf{D}_X^{-1}\Psi_X) = \langle Bv, \Psi_M\rangle\tilde\Psi_M = Bv.\end{aligned}$$

Likewise, the array $\mathbf{g}$ consists by definition of the wavelet coefficients with respect to the dual basis $\tilde\Psi_M$. On the other hand, the unknown array $\mathbf{p}$ in the reduced system (2.4.2) contains coefficients with respect to the primal basis Ψ_M. Now, as mentioned before, in some cases the space M is actually a closed subspace of a somewhat larger Hilbert space characterized by Ψ_M. Therefore the wavelet coefficients of elements of M with respect to Ψ_M (or $\mathbf{D}_M^{-1}\Psi_M$) satisfy certain constraints which generally depend on the particular wavelet basis. To change representations if necessary, observe that, in view of (3.1.2), $\tilde\Psi_M = (\tilde\Psi_M, \tilde\Psi_M)\Psi_M$, so that such a change of bases is realized by the matrix

$$\mathbf{R} := (\tilde\Psi_M, \tilde\Psi_M) \tag{3.1.3}$$

because $\tilde{\mathbf{p}}^T\tilde\Psi_M = \tilde{\mathbf{p}}^T\mathbf{R}\Psi_M = (\mathbf{R}\tilde{\mathbf{p}})^T\Psi_M$. It immediately follows from (3.1.1) that both $\mathbf{R}$ and $\mathbf{R}^{-1} = (\Psi_M, \Psi_M)$ are bounded on $\ell_2(\mathcal{J}_M)$:

$$\|\mathbf{R}\|_{\ell_2(\mathcal{J}_M)\to\ell_2(\mathcal{J}_M)} \le C_R. \tag{3.1.4}$$

Since $\mathbf{S}$ is positive definite and satisfies (2.4.4), there exists therefore some positive ω (e.g., $\omega < 2/(C_S C_R)$) such that

$$\rho := \|\mathbf{id} - \omega \mathbf{RS}\|_{\ell_2(\mathcal{J}_M)\to\ell_2(\mathcal{J}_M)} < 1. \tag{3.1.5}$$

Then the infinite dimensional version of the Uzawa scheme reads as follows.

UZAWA: *Given any* $\mathbf{p}_0 \in \ell_{2,0}(\mathcal{J}_M)$*, compute for* $i = 1, 2, \ldots$

$$\mathbf{A}\mathbf{u}_i = \mathbf{f} - \mathbf{B}^T \mathbf{p}_{i-1}, \tag{3.1.6}$$

$$\mathbf{p}_i = \mathbf{p}_{i-1} + \omega \mathbf{R}(\mathbf{B}\mathbf{u}_i - \mathbf{g}). \tag{3.1.7}$$

This is known to converge when $\rho < 1$. In fact, since $\mathbf{u} = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\mathbf{p})$, it is easy to see that

$$\mathbf{p} - \mathbf{p}_i = (\mathbf{id} - \omega \mathbf{RS})(\mathbf{p} - \mathbf{p}_{i-1}),$$

so that

$$\|\mathbf{p} - \mathbf{p}_i\|_{\ell_2(\mathcal{J}_M)} \le \rho^i \|\mathbf{p} - \mathbf{p}_0\|_{\ell_2(\mathcal{J}_M)}. \tag{3.1.8}$$

Moreover, it has been shown in [14] that for $\mathbf{p}_0 = \mathbf{0}$ one has

$$\|\mathbf{p} - \mathbf{p}_i\|_{\ell_2(\mathcal{J}_M)} \le \|\mathbf{A}^{-1}\mathbf{f}\|_{\ell_2(\mathcal{J}_X)} \|\omega \mathbf{RSB}\|_{\ell_2(\mathcal{J}_X)\to\ell_2(\mathcal{J}_M)} \frac{\rho^i}{1-\rho}. \tag{3.1.9}$$

3.2. The adaptive scheme. As in [9] the key idea is to apply the above Uzawa iteration to the infinite dimensional problem. In view of (3.1.6) and (3.1.7), this involves three tasks, namely, adding sequences with generally infinite support such as the data $\mathbf{f}$ and $\mathbf{g}$, the application of *infinite* matrices like $\mathbf{B}$ or $\mathbf{B}^T$ to *finitely supported* vectors, as well as the solution of elliptic problems involving the infinite matrix $\mathbf{A}$. Of course, in practice neither one of these tasks can be performed *exactly*. Therefore one has to employ suitable approximations whose accuracy will depend on the current stage of the algorithm and which will be described next.

To this end, we shall not distinguish formally between finitely supported vectors and infinite sequences in $\ell_2(\mathcal{J}')$, where in what follows $\mathcal{J}' \in \{\mathcal{J}_X, \mathcal{J}_M\}$, but rather we will view both quantities as sequences (expanded by zero entries if necessary).

The first basic ingredient is the routine **NCOARSE** from [8] which solves the following task.

NCOARSE $[\eta, \mathbf{v}] \to (\bar{\mathbf{v}}, \Lambda)$ determines for a given *finitely supported* vector $\mathbf{v}$ a vector $\bar{\mathbf{v}}$ with smallest possible support Λ such that

$$\|\mathbf{v} - \bar{\mathbf{v}}\|_{\ell_2(\mathcal{J})} \le \eta. \tag{3.2.1}$$

For a detailed description of this routine and the analysis of its computational complexity, see [8, Properties 6.1, 6.3]. In particular, **NCOARSE** will be used to approximate the arrays $\mathbf{f}^\circ := \mathbf{D}_X^{-1}\langle \Psi_X, f\rangle_X$ and $\mathbf{g}$ of given data by finitely supported vectors. The way to think about **NCOARSE** in this context can be formulated as follows.

ASSUMPTION 3.1. *In a preprocessing step for a given target accuracy, sufficiently many (wavelet) coefficients in the arrays* $\mathbf{f}^\circ := \mathbf{D}_X^{-1}\langle \Psi_X, f\rangle$ *and* $\mathbf{g}$ *are made available and ordered by size.*

In many applications f and g are simple and, as model data given by the user, are considered here as completely accessible. Coarser approximations of the data are then obtained by applying **NCOARSE** to these preprocessed finite arrays (see [8, section 6.1] for a more detailed discussion).

The second basic ingredient taken from [8] is an approximate application of an infinite matrix to a finitely supported vector. Given an infinite matrix $\mathbf{C}$ (as a mapping from $\ell_2(\mathcal{J}'')$ to $\ell_2(\mathcal{J}')$ for any pair $(\mathcal{J}', \mathcal{J}'') \in \{\mathcal{J}_X, \mathcal{J}_M\}^2$), we use the scheme **APPLY** from [8] for serving the following purpose.

APPLY $[\eta, \mathbf{C}, \mathbf{v}] \to (\mathbf{w}, \Lambda)$ produces for any finitely supported input vector $\mathbf{v}$ a vector $\mathbf{w}$ with finite support $\Lambda \subset \mathcal{J}'$ such that

$$\|\mathbf{C}\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J}')} \le \eta. \tag{3.2.2}$$

A scheme with this property has been developed in [8, section 6.4]; see [2] for implementation issues and numerical experiments. We postpone a quick description of the relevant features along with estimates for its computational cost to a later section.

Note that, in particular, the routines **APPLY** and **NCOARSE** allow us to approximately evaluate the right-hand sides of (3.1.6) and (3.1.7).

So the remaining task in an approximate Uzawa iteration of the form (3.1.6), (3.1.7) is to solve the operator equation (3.1.6) with system matrix $\mathbf{A}$. This is an elliptic problem in the sense of [8], and we will make heavy use of the results obtained there; also see [2] for implementations and numerical tests. The scheme from [8] is also built solely on the above routines **NCOARSE** and **APPLY**. There are, however, two minor points that need to be briefly addressed.

First, in [8] the matrix $\mathbf{A}$ is just the wavelet representation of the underlying elliptic operator, while in the present situation, $\mathbf{A}$ has the form (2.4.10) for some positive constant c, when $a(\cdot,\cdot)$ is not elliptic on all of X. Nevertheless, once a scheme **APPLY** for wavelet representations is available, a scheme for applying matrices of the form (2.4.10) with $c \neq 0$ is easily obtained from such a building block as follows. To simplify notation we set $\mathbf{A}^\circ := \mathbf{D}_X^{-1} a(\Psi_X, \Psi_X) \mathbf{D}_X^{-1}$.

APPLY$^*[\eta, \mathbf{A}, \mathbf{v}] \to (\mathbf{w}, \Lambda)$:

(i) **APPLY** $[\eta/2, \mathbf{A}^\circ, \mathbf{v}] \to (\mathbf{w}_1, \Lambda_1)$;

(ii) **APPLY** $[\eta/4cC_B, \mathbf{B}, \mathbf{v}] \to (\mathbf{w}_2, \Lambda_2)$;

(iii) **APPLY** $[\eta/4, c\mathbf{B}^T, \mathbf{w}_2] \to (\mathbf{w}_3, \Lambda_3)$ and set $\mathbf{w} := \mathbf{w}_1 + \mathbf{w}_3, \quad \Lambda := \Lambda_1 \cup \Lambda_3$.

REMARK 3.2. *One easily derives from* (3.2.2) *that the output* $\mathbf{w}$ *produced by* **APPLY**$^*[\eta, \mathbf{A}, \mathbf{v}]$ *satisfies for* $\mathbf{A}$ *given by* (2.4.10)

$$\|\mathbf{A}\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J}_X)} \le \eta. \tag{3.2.3}$$

Moreover, it is also clear that up to a uniform constant, the work/accuracy balance for **APPLY*** *is the same as that for* **APPLY**. *Note that the matrix* $\mathbf{B}^T\mathbf{B}$ *is, of course, never computed.*

According to (3.1.6), **UZAWA** requires the solution of elliptic problems. For this purpose we shall employ here an adaptive scheme referred to as **ELLSOLVE**, developed and analyzed in [8]; also see [9] for the identification of those requirements on the routine **APPLY** in this context that ensure asymptotically optimal complexity. This will allow us, in particular, to employ the variant **APPLY*** in place of the original scheme without changing the asymptotical work/accuracy rate.

To explain the features of the scheme **ELLSOLVE** from [8], consider for $\mathbf{A}$ as above the elliptic problem

$$\mathbf{A}\mathbf{u} = \mathbf{h} \tag{3.2.4}$$

for some $\mathbf{h} \in \ell_2(\mathcal{J}_X)$, and denote its exact solution by $\hat{\mathbf{u}}$. The scheme **ELLSOLVE** solves the following task (see Algorithm III in [8, Theorem 7.6]).

ELLSOLVE $[\varepsilon, \mathbf{A}, \mathbf{v}, \mathbf{h}] \to (\bar{\mathbf{u}}, \Lambda)$: Given $\varepsilon > 0$ and an approximate solution $\mathbf{v}$ to (3.2.4), then the output $\bar{\mathbf{u}}$ with finite support Λ satisfies

$$\|\hat{\mathbf{u}} - \bar{\mathbf{u}}\|_{\ell_2(\mathcal{J}_X)} \le \varepsilon. \tag{3.2.5}$$

The second point is that in [8] the right-hand side data are assumed to be a given array of wavelet coefficients as explained above that can be preprocessed. In the present situation the right-hand side data are composed of such preprocessable data like $\mathbf{f}$ and an additional matrix/vector product involving dynamically updated entities. We therefore have to approximate these data by finitely supported vectors that can then be processed as in [8, sections 7.2, 7.3]. The corresponding perturbations can be estimated as follows.

REMARK 3.3. *Again consider* (3.2.4) *and suppose that approximate finitely supported right-hand side data* $\mathbf{h}_\eta \in \ell_2(\mathcal{J}_X)$ *are given such that*

$$\|\mathbf{h} - \mathbf{h}_\eta\|_{\ell_2(\mathcal{J}_X)} \le \eta. \tag{3.2.6}$$

Then the output $\bar{\mathbf{u}}$ *of* **ELLSOLVE** $[\eta, \mathbf{A}, \mathbf{v}, \mathbf{h}_\eta]$ *satisfies*

$$\|\hat{\mathbf{u}} - \bar{\mathbf{u}}\|_{\ell_2(\mathcal{J}_X)} \le \varepsilon + c_A^{-1}\eta. \tag{3.2.7}$$

Proof The claim follows from (3.2.5) combined with (2.4.7) to estimate the perturbation effect. □

Next, we will describe the computation of a finitely supported $\mathbf{h}_\eta$ when $\mathbf{h} = \mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}$; see (3.1.6). Recall from (2.4.10) that

$$\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1} = \mathbf{f}^\circ - \mathbf{B}^T(\bar{\mathbf{p}}_{i-1} - c\mathbf{g}), \qquad \mathbf{f}^\circ = \mathbf{D}_X^{-1}\langle \Psi_X, f\rangle,$$

which thus involves coarsening the given (preprocessed) data $\mathbf{f}^\circ, \mathbf{g}$ and a multiplication by $\mathbf{B}^T$. The respective concrete accuracy tolerances are given in the following routine.

RHS $[\bar{\mathbf{p}}, \eta] \to (\mathbf{h}_\eta, \Lambda^h)$:
Given a finitely supported $\bar{\mathbf{p}}$, the routine **RHS** computes a vector $\mathbf{h}_\eta$ with finite support Λ^h satisfying

$$\|\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}} - \mathbf{h}_\eta\|_{\ell_2(\mathcal{J}_X)} \le \eta \tag{3.2.8}$$

as follows:

(i) **NCOARSE** $[\eta/3, \mathbf{f}^\circ] \to (\bar{\mathbf{f}}, \Lambda^f)$,
NCOARSE $[\eta/3cC_B, \mathbf{g}] \to (\bar{\mathbf{g}}, \Lambda^g)$, and set $\mathbf{r} := \bar{\mathbf{p}} - c\bar{\mathbf{g}}$.
(ii) **APPLY** $[\eta/3, \mathbf{B}^T, \mathbf{r}] \to (\mathbf{w}, \Lambda^w)$ and set $\mathbf{h}_\eta := \bar{\mathbf{f}} - \mathbf{w}$, $\Lambda^h := \Lambda^f \cup \Lambda^w$.

Since by (3.2.1), $\|(\bar{\mathbf{p}} - c\mathbf{g}) - \mathbf{r}\|_{\ell_2(\mathcal{J}_X)} \le \eta/3C_B$, the estimate (3.2.8) indeed readily follows from (3.2.2).

Our numerical realization of the *ideal* (infinite dimensional) Uzawa scheme (3.1.6), (3.1.7) has the following structure. A fixed uniformly bounded number K, depending only on the constants associated with the wavelet bases and the mapping properties

of the involved operators, of approximate applications of (3.1.6), (3.1.7) are applied, which is then followed by a coarsening step before the iteration is further resumed. Such an *iteration block* will be arranged to *advance* the current approximate solutions so as to reduce the current error bounds by a fixed factor. Before giving a precise description, we would like to stress that the Uzawa scheme as a gradient method for the reduced system (2.4.2) treats in some sense $\mathbf{p} \in \ell_{2,0}(\mathcal{J}_M)$ as the "preferred" variable. In fact, the accuracy of the approximate solution to the elliptic problem (3.1.6) need not be too high relative to the current accuracy of the approximation of $\mathbf{p}$. In order to formulate now the basic iteration block as a concrete routine, we will use the following choice for the number K of perturbed iterations before the next coarsening step. Let γ_i denote any positive summable numbers, e.g., $\gamma_i = (1+i)^{-2}$. It will be convenient to assume always that, as in this example, $\gamma_i \le 1$, $i \in \mathbb{N}$. Moreover, we need some control parameters. Set

$$C_1 := \omega(C_R C_B + 2)\gamma + 1, \tag{3.2.9}$$

where $\gamma := \sum_{i=0}^{\infty} \gamma_i$, and let K denote the smallest integer such that

$$\rho^K \max\left\{C_1, (\rho c_A)^{-1} C_B C_1 + 1\right\} \le 1/10. \tag{3.2.10}$$

Note that, since c_A, C_B appear in lower, respectively, upper, bounds, $(\rho c_A)^{-1} C_B$ will typically be larger than one, so that the maximum will usually be attained by the second term in the curly brackets in the left-hand side of (3.2.10).

Now we have collected all the ingredients for composing the core of a computational version of **UZAWA**.

ADV $[\bar{\mathbf{u}}, \bar{\mathbf{p}}, \delta] \to (\tilde{\mathbf{u}}, \tilde{\mathbf{p}}, \Lambda_u, \Lambda_q)$:
Given current approximations $\bar{\mathbf{u}}, \bar{\mathbf{p}}$ of the solution to (2.2.4) such that

$$\|\bar{\mathbf{u}} - \mathbf{u}\|_{\ell_2(\mathcal{J}_X)} \le \delta, \quad \|\bar{\mathbf{p}} - \mathbf{p}\|_{\ell_2(\mathcal{J}_M)} \le \delta, \tag{3.2.11}$$

ADV $[\bar{\mathbf{u}}, \bar{\mathbf{p}}, \delta]$ produces new approximations $\tilde{\mathbf{u}}, \tilde{\mathbf{p}}$ as follows:
(i) Set $i = 1$, $\bar{\mathbf{p}}_0 := \bar{\mathbf{p}}$, $\bar{\mathbf{u}}_0 := \bar{\mathbf{u}}$.
(ii) If $i \le K$, go to (iii); else
  NCOARSE $[2\delta/5, \bar{\mathbf{p}}_{i-1}] \to (\tilde{\mathbf{p}}, \Lambda_p)$;
  NCOARSE $[2\delta/5, \bar{\mathbf{u}}_{i-1}] \to (\tilde{\mathbf{u}}, \Lambda_u)$; STOP;
(iii) **RHS** $[\bar{\mathbf{p}}_{i-1}, c_A \gamma_i \rho^i \delta/2] \to (\mathbf{h}_i, \Lambda_i^h)$;
(iv) **ELLSOLVE** $[\gamma_i \rho^i \delta/2, \mathbf{A}, \bar{\mathbf{u}}_{i-1}, \mathbf{h}_i] \to (\bar{\mathbf{u}}_i, \Lambda_i^X)$.
(v) **NCOARSE** $[\gamma_i \rho^i \delta/2C_R, \mathbf{g}] \to (\hat{\mathbf{g}}_i, \hat{\Lambda}_i)$;
  APPLY $[\gamma_i \rho^i \delta/2, \mathbf{R}, \hat{\mathbf{g}}_i] \to (\mathbf{g}_i, \Lambda_i^q)$;
  APPLY $[\gamma_i \rho^i \delta/2C_R, \mathbf{B}, \bar{\mathbf{u}}_i] \to (\hat{\mathbf{p}}_i, \hat{\Lambda}_i)$;
  APPLY $[\gamma_i \rho^i \delta/2, \mathbf{R}, \hat{\mathbf{p}}_i] \to (\mathbf{p}_i', \Lambda_i^p)$;
  set $\bar{\mathbf{p}}_i = \bar{\mathbf{p}}_{i-1} + \omega(\mathbf{p}_i' - \mathbf{g}_i)$; and $i + 1 \to i$ and go to (ii).

The routine **ADV** *advances* given finitely supported approximate solutions $(\bar{\mathbf{u}}, \bar{\mathbf{p}})$ to a new pair $(\tilde{\mathbf{u}}, \tilde{\mathbf{p}})$ by a finite number of perturbed Uzawa steps, followed by a coarsening step. It will be shown later that the tolerances in **ADV** are chosen so as to reduce the error bounds of the new approximations by at least a factor of two. The final application of **NCOARSE** in step (ii) of **ADV**, following the iteration block, will be seen later to play an important role with regard to asymptotically optimal complexity. Roughly speaking, it ensures that only sufficiently significant coefficients are propagated.

Of course, when the characterization of the space M does not entail any constraints on the wavelet coefficients, $\mathbf{R}$ can be replaced by the identity in (3.1.7), in which case step (v) of **ADV** simplifies in an obvious manner.

To formulate the main algorithm, recall that by (2.3.1)

$$\|\mathbf{u}\|^2_{\ell_2(\mathcal{J}_X)} + \|\mathbf{p}\|^2_{\ell_2(\mathcal{J}_M)} \le c_L^{-1}\left(\|\mathbf{f}^\circ\|^2_{\ell_2(\mathcal{J}_X)} + \|\mathbf{g}\|^2_{\ell_2(\mathcal{J}_M)}\right).$$

Therefore the right-hand side gives a bound for the initial error when using $\mathbf{0}$ as the initial guess for $\mathbf{u}, \mathbf{p}$, respectively. The complete adaptive Uzawa iteration can be described now as follows.

UZAWA^c $[\mathbf{A}, \mathbf{B}, \mathbf{f}, \mathbf{g}, \varepsilon] \to (\mathbf{u}(\varepsilon), \mathbf{p}(\varepsilon))$:
Set $\mathbf{\Lambda}_0 := (\Lambda_0^X, \Lambda_0^M) \subset \mathcal{J} =: \mathcal{J}_X \times \mathcal{J}_M$ to be empty, $\Lambda_0^M = \Lambda_0^X = \emptyset$, $\mathbf{p}_0 = \bar{\mathbf{p}}_0 = \mathbf{0}$, $\bar{\mathbf{u}} = \mathbf{0}$, $\delta_0 := c_L^{-1/2}(\|\mathbf{f}^\circ\|^2_{\ell_2(\mathcal{J}_X)} + \|\mathbf{g}\|^2_{\ell_2(\mathcal{J}_M)})^{1/2}$, $J = 0$, choose a target accuracy ε.
(i) **ADV** $[\bar{\mathbf{u}}, \bar{\mathbf{p}}, \delta_J] \to (\tilde{\mathbf{u}}, \tilde{\mathbf{p}}, \Lambda_u, \Lambda_q)$;
(ii) Set $\delta_{J+1} := \delta_J/2$;
If $\delta_{J+1} \le \varepsilon$, set $\mathbf{u}(\varepsilon) := \tilde{\mathbf{u}}, \mathbf{p}(\varepsilon) := \tilde{\mathbf{p}}$; STOP;
Else, set $\bar{\mathbf{u}} = \tilde{\mathbf{u}}$, $\bar{\mathbf{p}} = \tilde{\mathbf{p}}$, $J + 1 \to J$ and go to (i).

3.3. Convergence. The convergence of **UZAWA**^{c} relies on the error reduction caused by **ADV**.

PROPOSITION 3.4. *Given a scheme* **APPLY** *such that* (3.2.2) *holds, under the above Assumption* 3.1 *concerning* **NCOARSE** *on the data* $\mathbf{f}^\circ, \mathbf{g}$, *the vectors* $\tilde{\mathbf{u}}, \tilde{\mathbf{p}}$ *produced by* **ADV** $[\bar{\mathbf{u}}, \bar{\mathbf{p}}, \delta]$ *above satisfy*

$$\|\tilde{\mathbf{u}} - \mathbf{u}\|_{\ell_2(\mathcal{J}_X)} \le \delta/2, \quad \|\tilde{\mathbf{p}} - \mathbf{p}\|_{\ell_2(\mathcal{J}_M)} \le \delta/2. \tag{3.3.1}$$

Hence, after finitely many steps the scheme **UZAWA**^{c} *produces finitely supported solutions* $(\mathbf{u}(\varepsilon), \mathbf{p}(\varepsilon))$ *satisfying*

$$\|\mathbf{u} - \mathbf{u}(\varepsilon)\|_{\ell_2(\mathcal{J}_X)} \le \varepsilon, \quad \|\mathbf{p} - \mathbf{p}(\varepsilon)\|_{\ell_2(\mathcal{J}_M)} \le \varepsilon. \tag{3.3.2}$$

Proof. Set $\mathbf{p}_0 := \bar{\mathbf{p}}_0 = \bar{\mathbf{p}}$, $\bar{\mathbf{u}}_0 := \bar{\mathbf{u}}$ and observe that

$$\begin{aligned}
\mathbf{p}_i - \bar{\mathbf{p}}_i &= \mathbf{p}_{i-1} + \omega\mathbf{R}(\mathbf{B}\mathbf{u}_i - \mathbf{g}) - \bar{\mathbf{p}}_{i-1} - \omega(\mathbf{p}_i' - \mathbf{g}_i) \\
&= \mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1} + \omega(\mathbf{R}\mathbf{B}\mathbf{u}_i - \mathbf{p}_i' - \mathbf{R}\mathbf{g} + \mathbf{g}_i) \\
&= (\mathbf{id} - \omega\mathbf{R}\mathbf{S})(\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}) + \omega(\mathbf{R}(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)(\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}) \\
&\qquad + \mathbf{R}\mathbf{B}\mathbf{u}_i - \mathbf{p}_i' + \mathbf{g}_i - \mathbf{R}\mathbf{g}).
\end{aligned} \tag{3.3.3}$$

Since $\mathbf{A}\mathbf{u}_i = \mathbf{f} - \mathbf{B}^T\mathbf{p}_{i-1}$, we can replace $\mathbf{B}^T\mathbf{p}_{i-1}$ by $\mathbf{f} - \mathbf{A}\mathbf{u}_i$ to obtain

$$\begin{aligned}
&\omega\left(\mathbf{R}(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)(\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}) + \mathbf{R}\mathbf{B}\mathbf{u}_i - \mathbf{p}_i' + \mathbf{g}_i - \mathbf{R}\mathbf{g}\right) \\
&= \omega\left(\mathbf{R}(\mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{B}\mathbf{u}_i + \mathbf{B}\mathbf{u}_i - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \mathbf{p}_i' + \mathbf{g}_i - \mathbf{R}\mathbf{g}\right) \\
&= \omega\left(\mathbf{R}\mathbf{B}\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \mathbf{p}_i' + (\mathbf{g}_i - \mathbf{R}\mathbf{g})\right).
\end{aligned} \tag{3.3.4}$$

Thus

$$\mathbf{R}\mathbf{B}\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \mathbf{p}_i' = \mathbf{R}\mathbf{B}\left(\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \bar{\mathbf{u}}_i\right) + (\mathbf{R}\mathbf{B}\bar{\mathbf{u}}_i - \mathbf{p}_i'). \tag{3.3.5}$$

Hence combining (3.3.3), (3.3.4), and (3.3.5) and recalling (3.1.5) yields

$$\begin{aligned}
\|\mathbf{p}_i - \bar{\mathbf{p}}_i\|_{\ell_2(\mathcal{J}_M)} &\le \rho\|\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}\|_{\ell_2(\mathcal{J}_M)} + \omega\left(\|\mathbf{R}\mathbf{B}\left(\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \bar{\mathbf{u}}_i\right)\|_{\ell_2(\mathcal{J}_M)}\right. \\
&\quad \left. + \|\mathbf{R}\mathbf{B}\bar{\mathbf{u}}_i - \mathbf{p}_i'\|_{\ell_2(\mathcal{J}_M)} + \|\mathbf{g}_i - \mathbf{R}\mathbf{g}\|_{\ell_2(\mathcal{J}_M)}\right) \\
&\le \rho\|\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}\|_{\ell_2(\mathcal{J}_M)} + \omega C_R C_B\|\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \bar{\mathbf{u}}_i\|_{\ell_2(\mathcal{J}_X)} \\
&\quad + 2\omega\gamma_i\rho^i\delta,
\end{aligned} \tag{3.3.6}$$

where we have used the tolerances in step (v) of **ADV**. By (3.2.8) we have for the output $\mathbf{h}_i$ of step (iii) in **ADV** that $\|\mathbf{h}_i - (\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1})\|_{\ell_2(\mathcal{J}_X)} \le c_A\gamma_i\rho^i\delta/2$, which, in view of the tolerances in step (iv) of **ADV** and (3.2.7), implies that

$$\|\mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}) - \bar{\mathbf{u}}_i\|_{\ell_2(\mathcal{J}_X)} \le \gamma_i\rho^i\delta. \tag{3.3.7}$$

Therefore, we deduce from (3.3.6) that

$$\|\mathbf{p}_i - \bar{\mathbf{p}}_i\|_{\ell_2(\mathcal{J}_M)} \le \rho\|\mathbf{p}_{i-1} - \bar{\mathbf{p}}_{i-1}\|_{\ell_2(\mathcal{J}_M)} + \omega(C_RC_B + 2)\gamma_i\rho^i\delta. \tag{3.3.8}$$

Iterating this estimate, and bearing in mind that $\mathbf{p}_0 = \bar{\mathbf{p}}_0$, provides

$$\|\mathbf{p}_i - \bar{\mathbf{p}}_i\|_{\ell_2(\mathcal{J}_M)} \le \omega(C_RC_B + 2)\rho^i\delta\sum_{l=1}^{i}\gamma_l. \tag{3.3.9}$$

Since by (3.1.8) and the assumption $\|\mathbf{p} - \mathbf{p}_i\|_{\ell_2(\mathcal{J}_M)} \le \rho^i\|\mathbf{p} - \mathbf{p}_0\|_{\ell_2(\mathcal{J}_M)} = \rho^i\|\mathbf{p} - \bar{\mathbf{p}}\|_{\ell_2(\mathcal{J}_M)} \le \rho^i\delta$, we conclude that

$$\|\mathbf{p} - \bar{\mathbf{p}}_i\|_{\ell_2(\mathcal{J}_M)} \le \left(\omega(C_RC_B + 2)\sum_{l=1}^{i}\gamma_l + 1\right)\rho^i\delta \le C_1\rho^i\delta, \tag{3.3.10}$$

where C_1 is the constant from (3.2.9). In view of (3.2.10), this gives

$$\|\mathbf{p} - \bar{\mathbf{p}}_K\|_{\ell_2(\mathcal{J}_M)} \le \delta/10. \tag{3.3.11}$$

Now recall that by step (ii) of **ADV**, the final approximation $\tilde{\mathbf{p}}$ is obtained by coarsening $\bar{\mathbf{p}}_K$. Thus

$$\|\mathbf{p} - \tilde{\mathbf{p}}\|_{\ell_2(\mathcal{J}_M)} \le \|\mathbf{p} - \bar{\mathbf{p}}_K\|_{\ell_2(\mathcal{J}_M)} + \|\bar{\mathbf{p}}_K - \tilde{\mathbf{p}}\|_{\ell_2(\mathcal{J}_M)} \le \left(\frac{2}{5} + \frac{1}{10}\right)\delta = \frac{\delta}{2}, \tag{3.3.12}$$

as claimed.

It remains to estimate the accuracy of $\bar{\mathbf{u}}_K$. Denoting by $\hat{\mathbf{u}}_i$ the exact solution of $\mathbf{A}\hat{\mathbf{u}}_i = \mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1}$, (3.3.7) says that $\|\hat{\mathbf{u}}_i - \bar{\mathbf{u}}_i\|_{\ell_2(\mathcal{J}_X)} \le \gamma_i\rho^i\delta$. Writing

$$\mathbf{u} - \bar{\mathbf{u}}_i = \mathbf{u} - \hat{\mathbf{u}}_i + \hat{\mathbf{u}}_i - \bar{\mathbf{u}}_i = \mathbf{A}^{-1}\mathbf{B}^T(\bar{\mathbf{p}}_{i-1} - \mathbf{p}) + \hat{\mathbf{u}}_i - \bar{\mathbf{u}}_i, \tag{3.3.13}$$

one infers from (3.3.10) that

$$\|\mathbf{u} - \bar{\mathbf{u}}_i\|_{\ell_2(\mathcal{J}_X)} \le (c_A\rho)^{-1}C_BC_1\rho^i\delta + \gamma_i\rho^i\delta = \left((c_A\rho)^{-1}C_BC_1 + \gamma_i\right)\rho^i\delta.$$

Again, since $\gamma_i \le 1$, we conclude from (3.2.10) that

$$\|\mathbf{u} - \bar{\mathbf{u}}_K\|_{\ell_2(\mathcal{J}_X)} \le \delta/10, \tag{3.3.14}$$

so that by the same reasoning as in (3.3.12), $\tilde{\mathbf{u}}$ produced by $\mathbf{NCOARSE}\,[2\delta/5, \bar{\mathbf{u}}_K]$ satisfies $\|\mathbf{u} - \tilde{\mathbf{u}}\|_{\ell_2(\mathcal{J}_X)} \le \delta/2$. This completes the proof. □

As an immediate consequence of the norm equivalences (2.2.1), (2.2.2) one has the following fact.

COROLLARY 3.5. *Let $u = \mathbf{u}^T\mathbf{D}_X^{-1}\Psi_X$, $p = \mathbf{p}^T\mathbf{D}_M^{-1}\Psi_M$ be the exact solution of (2.1.2). Then the finite expansions $u(\varepsilon) := \mathbf{u}^T(\varepsilon)\mathbf{D}_X^{-1}\Psi_X$, $p(\varepsilon) = \mathbf{p}^T(\varepsilon)\mathbf{D}_M^{-1}\Psi_M$ with terms from the finite index sets $\Lambda_{u(\varepsilon)} \subset \mathcal{J}_X$, $\Lambda_{q(\varepsilon)} \subset \mathcal{J}_M$ satisfy*

$$\|u - u(\varepsilon)\|_X \le C_X\varepsilon, \quad \|p - p(\varepsilon)\|_M \le C_M\varepsilon \tag{3.3.15}$$

uniformly in ε, where C_X, C_M are the constants in (2.2.1)*, respectively,* (2.2.2)*.*

To keep things transparent we have based the above considerations on the simplest version (3.1.6), (3.1.7) of an Uzawa iteration. It will be seen below that already this version gives rise to asymptotically optimal convergence properties. Of course, similar results would be obtained for different accuracy tolerances as long as they differ by constants leading possibly to different values of K. Nevertheless, several more important possibilities suggest themselves for realizing quantitative improvements, e.g., replacing the Richardson iteration by a gradient or conjugate gradient iteration. This avoids the need for estimating step size parameters and should speed error reduction. Note that these variants still involve only the same algorithmic tasks, namely, approximate application of operators in the above sense. Furthermore, the number K of subiterations is likely to be too pessimistic. Therefore it would be preferable to monitor the error decay as follows. Note that $\mathbf{p}_i - \mathbf{g}_i$ in step (v) of **ADV** approximates $\mathbf{R}(\mathbf{B}\mathbf{u}_i - \mathbf{p}_i)$ and, in view of (3.1.6), (3.1.7), the residual $\mathbf{R}(\mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{g} - \mathbf{S}\mathbf{p}_{i-1})$. By (2.4.4) and the bounded invertibility of $\mathbf{R}$, this residual can be bounded from below and above by fixed constant multiples of the current error of the approximate solution of the reduced system (2.4.2). Thus monitoring $\|\mathbf{p}_i' - \mathbf{g}_i\|_{\ell_2(\mathcal{J}_M)}$ can be used as a stopping criterion. This is expected to result in frequent early termination of step (ii) in **ADV**. These points will be taken up in more detail elsewhere.

4. Complexity analysis. Of course, the central questions now are how do we come up with an **APPLY** scheme with the desired properties and what is the computational cost of $\mathbf{UZAWA}^c$ for a given target accuracy ε. In the present generality, cost will be measured by storage requirements and the number of flops required by the scheme (being well aware of the fact that this is not the full story).

4.1. Best *N*-term approximation. As in [8] we will relate the performance of the adaptive scheme to what could be achieved at best, namely, the approximation of the solution in terms of *possibly few degrees of freedom* within the given discretization context—here determined by the underlying wavelet bases. Note that, in view of (3.3.15), it suffices to deal with the conceptually much simpler approximation in $\ell_2(\mathcal{J})$. To explain this, it is useful to recall first the following notion of *best N-term approximation* in ℓ_2:

$$\sigma_{N,\ell_2(\mathcal{J}')}(\mathbf{v}) := \inf_{\mathbf{w},\#\operatorname{supp}\mathbf{w}\le N} \|\mathbf{v}-\mathbf{w}\|_{\ell_2(\mathcal{J}')}, \tag{4.1.1}$$

where $\ell_2(\mathcal{J}')$ stands again for $\ell_2(\mathcal{J}_X)$ or $\ell_2(\mathcal{J}_M)$. Thus $\sigma_{N,\ell_2(\mathcal{J}')}(\mathbf{v})$ describes the error as a function of the number of degrees of freedom when the (possibly infinitely supported) vector is approximated by a vector with at most N nonzero entries whose value and position can be freely chosen. Thus the approximant is not taken from any fixed linear space but from the nonlinear manifold of all vectors with at most N nonzero entries. This notion is well understood for ℓ_2; see, e.g., [19]. Obviously, $\sigma_{N,\ell_2(\mathcal{J}')}(\mathbf{v})$ is realized by retaining the N *largest* coefficients in $\mathbf{v}$ which are, of course, unknown when $\mathbf{v}$ is a solution of a system of equations. To understand how this error behaves, denote for any $\mathbf{v} \in \ell_2(\mathcal{J}')$ by $\mathbf{v}^* = \{v_{\lambda_l}\}_{l\in\mathbb{N}} =: \{v_l^*\}_{l\in\mathbb{N}}$ its *decreasing rearrangement* in the sense that $|v_{\lambda_l}| \ge |v_{\lambda_{l+1}}|$ and let

$$\Lambda(\mathbf{v},N) := \{\lambda_l : l = 1,\dots,N\}, \quad \mathbf{v}_N := \mathbf{v}|_{\Lambda(\mathbf{v},N)}. \tag{4.1.2}$$

It is clear that $\mathbf{v}_N$ is a best N-term approximation of $\mathbf{v}$.

In particular, it will be important to characterize the sequences in $\ell_2(\mathcal{J}')$ whose best N-term approximation behaves like N^{-s} for some $s > 0$. The following facts are well known [8, 19]. Let for $0 < \tau < 2$

$$|\mathbf{v}|_{\ell_\tau^w(\mathcal{J}')} := \sup_{n\in\mathbb{N}} n^{1/\tau}|v_n^*|, \quad \|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')} := \|\mathbf{v}\|_{\ell_2(\mathcal{J}')} + |\mathbf{v}|_{\ell_\tau^w(\mathcal{J}')}. \tag{4.1.3}$$

It is easy to see that for any $\tau < \tau' \le 2$

$$\|\mathbf{v}\|_{\ell_{\tau'}(\mathcal{J}')} \lesssim \|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')} \le 2\|\mathbf{v}\|_{\ell_\tau(\mathcal{J}')}, \tag{4.1.4}$$

so that by Jensen's inequality, in particular, $\ell_\tau^w(\mathcal{J}') \subset \ell_2(\mathcal{J}')$.

We shall use the following characterization of decay rates of best N-term approximation in $\ell_2(\mathcal{J}')$; see [8, Proposition 3.2].

PROPOSITION 4.1. *Let*

$$\frac{1}{\tau} = s + \frac{1}{2}. \tag{4.1.5}$$

Then $\mathbf{v}$ *belongs to* $\ell_\tau^w(\mathcal{J}')$ *if and only if* $\sigma_{N,\ell_2(\mathcal{J}')}(\mathbf{v}) \lesssim N^{-s}$, *and one has the error estimate*

$$\|\mathbf{v} - \mathbf{v}_N\|_{\ell_2(\mathcal{J}')} \lesssim N^{-s}\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')}. \tag{4.1.6}$$

In complete analogy one can define $\|\cdot\|_{\ell_\tau^w(\mathcal{J})}$ for $\ell_\tau^w(\mathcal{J}) := \ell_\tau^w(\mathcal{J}_X \times \mathcal{J}_M)$ by forming the rearrangements from both component vectors $\mathbf{v} \in \ell_2(\mathcal{J}_X)$, $\mathbf{p} \in \ell_{2,0}(\mathcal{J}_M)$ and regrouping the entries to both component vectors.

We will make use of the following result from [8] which interrelates best N-term approximation in ℓ_2 with the routine **NCOARSE**; see [8, Property 6.3].

PROPOSITION 4.2. *Suppose that* $\mathbf{v} \in \ell_2(\mathcal{J}')$ *and a finitely supported* $\mathbf{w}$ *satisfies for some tolerance* $\eta > 0$

$$\|\mathbf{v} - \mathbf{w}\|_{\ell_2(\mathcal{J}')} \le \eta/5.$$

Then (as has been used before), the output $\bar{\mathbf{w}}$ *of* $\mathbf{NCOARSE}\,[\mathbf{w}, 4\eta/5]$ *satisfies* $\|\mathbf{v} - \bar{\mathbf{w}}\|_{\ell_2(\mathcal{J}')} \le \eta$. *Moreover, when* $\mathbf{v} \in \ell_\tau^w(\mathcal{J}')$ *and* $\frac{1}{\tau} = s + \frac{1}{2}$ *for some* $s > 0$, *then there exists a constant* C, *depending only on* s *when* s *tends to infinity, such that*

$$\|\mathbf{v} - \bar{\mathbf{w}}\|_{\ell_2(\mathcal{J}')} \le C\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')}(\#\operatorname{supp}\bar{\mathbf{w}})^{-s}, \tag{4.1.7}$$

and

$$\|\bar{\mathbf{w}}\|_{\ell_\tau^w(\mathcal{J}')} \le C\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')}, \quad \#\operatorname{supp}\bar{\mathbf{w}} \le C\|\mathbf{v}\|^{1/s}_{\ell_\tau^w(\mathcal{J}')}\eta^{-1/s}. \tag{4.1.8}$$

Best N-term approximation will be one important ingredient in the realization of the approximate application of infinite matrices represented by **APPLY**. The other one is the (a priori known) quasi sparseness of wavelet representations which can be formalized as follows; see [8].

DEFINITION 4.3. *A matrix* $\mathbf{C}$ *belongs to the class* $\mathcal{C}_{s^*}$ *if for every* $s < s^*$ *there exists a positive summable sequence* $(\alpha_j)_{j\ge0}$ *and for every* $j \ge 0$ *there exists a matrix* $\mathbf{C}_j$ *with at most* $2^j\alpha_j$ *nonzero entries per row and column such that*

$$\|\mathbf{C}_j - \mathbf{C}\| \lesssim \alpha_j 2^{-sj}. \tag{4.1.9}$$

A matrix in $\mathcal{C}_{s^*}$ is called compressible *or sometimes s^*-compressible.*

Compressibility of a wavelet representation of certain operators follows from the above-mentioned cancellation properties of the wavelets; see [8] as well as section 5.3 for concretizations.

Now suppose that the (possibly infinite) matrix $\mathbf{C}$ (defined on $\ell_2(\mathcal{J}')$, say) is known to be compressible in the sense of (4.1.9) for some range of $s > 0$. For any given finitely supported $\mathbf{v} \in \ell_2(\mathcal{J}')$, let $\mathbf{v}_{[j]} := \mathbf{v}_{2^j}$ denote its best 2^j-term approximation in $\ell_2(\mathcal{J}')$. We shall numerically approximate $\mathbf{Cv}$ by using the vector

$$\mathbf{w}_k := \mathbf{C}_k\mathbf{v}_{[0]} + \mathbf{C}_{k-1}(\mathbf{v}_{[1]} - \mathbf{v}_{[0]}) + \cdots + \mathbf{C}_0(\mathbf{v}_{[k]} - \mathbf{v}_{[k-1]}) \tag{4.1.10}$$

for a certain value of k determined by the desired numerical accuracy. This leads to a practical scheme $\mathbf{APPLY}\ [\eta, \mathbf{C}, \mathbf{v}] \to (\mathbf{w}, \Lambda)$, whose detailed description is given in [8], section 6.4; see also [2]. For later use we recall its properties; see Property 6.4 in [8].

PROPOSITION 4.4. *Assume that $\mathbf{C} \in \mathcal{C}_{s^*}$. Given a tolerance $\eta > 0$ and a vector $\mathbf{v}$ with finite support, the algorithm* **APPLY** *produces a vector $\mathbf{w} = \mathbf{w}(\mathbf{v}, \eta)$ which satisfies* (3.2.2).

Moreover, if $\mathbf{v} \in \ell_\tau^w(\mathcal{J}')$, with $\tau = (s + 1/2)^{-1}$ and $0 < s < s^*$, then the following properties hold:

(i) *The size of the output Λ is bounded by*

$$\#(\Lambda) \le C\|\mathbf{v}\|^{1/s}_{\ell_\tau^w(\mathcal{J}')}\eta^{-1/s}, \tag{4.1.11}$$

and the number of entries of $\mathbf{C}$ that need to be computed is $\le C\|\mathbf{v}\|^{1/s}_{\ell_\tau^w(\mathcal{J}')}\eta^{-1/s}$.

(ii) *The number of* arithmetic operations *needed to compute $\mathbf{w}(\mathbf{v}, \eta)$ does not exceed $C\eta^{-1/s}\|\mathbf{v}\|^{1/s}_{\ell_\tau^w(\mathcal{J}')} + 2N$ with $N := \#\operatorname{supp}\mathbf{v}$.*

(iii) *The number of operations for* sorting *needed to assemble the slices $\mathbf{v}_{[j]}$ of $\mathbf{w}(\mathbf{v}, \eta)$, $j = 0, 1, \ldots, \lfloor \log N \rfloor$, does not exceed $CN \log N$.*

(iv) *The output vector $\mathbf{w}$ satisfies*

$$\|\mathbf{w}\|_{\ell_\tau^w(\mathcal{J}')} \le C\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{J}')}. \tag{4.1.12}$$

As for the log terms for sorting, see Remark 4.10 at the end of this section. We shall make use of the following fact; see [8].

REMARK 4.5. *It follows from Proposition* 4.1 *and Proposition* 4.4(i) *that any matrix $\mathbf{C} \in \mathcal{C}_{s^*}$ is bounded on ℓ_τ^w when τ is related to $s < s^*$ by* (4.1.5).

As mentioned above, wavelet representations of differential operators are compressible. Therefore the following observation is useful.

REMARK 4.6. *When $\mathbf{A}^\circ = \mathbf{D}_X^{-1}a(\Psi_X, \Psi_X)\mathbf{D}_X^{-1}$ and $\mathbf{B}$ belong to $\mathcal{C}_{s^*}$ for some $s^* > 0$, then one easily shows that the scheme* $\mathbf{APPLY}^*$ *inherits all the properties described in Proposition* 4.4 *above; see* [8, Property 6.4].

The complexity estimates in (ii) and (iii) of Proposition 4.4 hold under the assumption that the entries of $\mathbf{C}$ are accessible during the calculation. In fact, the subsequent developments will always be based on the following assumption.

ASSUMPTION 4.7. *The entries of the matrices $\mathbf{A}^\circ$ and $\mathbf{B}$ are accessible at unit cost.*

Using piecewise polynomial wavelets, this assumption can be realized for constant coefficient operators in a relatively straightforward manner. This task becomes much more delicate under more general circumstances, e.g., when isoparametric mappings

are involved in the construction of the wavelets; see section 5.2 below. In [3] a fast evaluation scheme is developed that computes sufficiently accurate approximations to the summands on the right-hand side of (4.1.10) at a computational cost that still satisfies the bounds in (ii), (iii) of Proposition 4.4 above. Thus Assumption 4.7 is justified for a wide range of practically relevant situations.

With Remark 4.6 at hand, we are now in the position for estimating the complexity analysis of **ELLSOLVE** based on the results in [8, 9] with the **APPLY** scheme for compressible matrices replaced, if necessary, by the extended version **APPLY*** introduced above. The fact that in the present context **ELLSOLVE** applies to varying auxiliary problems with little a priori information on the corresponding intermediate solutions prevents us from applying the results from [8] directly. Nevertheless, we can extract from the analysis in [8, 9] some facts that will apply in the present situation as well. This is most transparent when considering the simplified scheme in [9] which (in the very spirit of the current approach) for the special case of an elliptic (coercive) problem is based on a simple iteration for (3.2.4) of the form

$$\hat{\mathbf{u}}^{n+1} = \hat{\mathbf{u}}^n + \bar{\omega}(\mathbf{h} - \mathbf{A}\hat{\mathbf{u}}^n). \tag{4.1.13}$$

In particular, when the right-hand sides are already finitely supported as in the present situation, the scheme consists of at most $\bar{K}$ perturbed iterations of the form (4.1.13), employing **APPLY*** and **NCOARSE** with judiciously chosen accuracy tolerances, followed by a coarsening step so as to reduce a current error bound by a factor of two, say (see the algorithm **SOLVE** in section 4.2 of [9]). This implies the following fact.

PROPOSITION 4.8. *Consider the problem* (3.2.4) *and suppose that the initial approximation* $\mathbf{v}$ *used as input for* **ELLSOLVE** *satisfies*

$$\|\hat{\mathbf{u}} - \mathbf{v}\|_{\ell_2(\mathcal{J}_X)} \le \bar{\varepsilon} \tag{4.1.14}$$

for some $\bar{\varepsilon} > \varepsilon$. *Moreover, assume that* s *and* τ *are related by* (4.1.5) *and that*

$$\varepsilon \le \bar{C}\bar{\varepsilon} \tag{4.1.15}$$

for some positive constant $\bar{C}$. *Then the output* $\bar{\mathbf{u}}$ *and* $\Lambda := \operatorname{supp} \bar{\mathbf{u}}$ *of* **ELLSOLVE** $[\varepsilon, \mathbf{A}, \mathbf{v}, \mathbf{h}]$ *satisfies*

$$\#(\Lambda) \le \hat{C}\left(\#(\operatorname{supp}\mathbf{v}) + \left(\|\mathbf{v}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)} + \|\mathbf{h}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)}\right)\varepsilon^{-1/s}\right),$$

$$\|\bar{\mathbf{u}}\|_{\ell^w_\tau(\mathcal{J}_X)} \le \hat{C}\left(\|\mathbf{v}\|_{\ell^w_\tau(\mathcal{J}_X)} + \|\mathbf{h}\|_{\ell^w_\tau(\mathcal{J}_X)}\right). \tag{4.1.16}$$

Moreover, the number of arithmetic operations required for the computation of $\bar{\mathbf{u}}$ *remains bounded by*

$$\hat{C}\left\{\#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\left(\|\mathbf{v}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)} + \|\mathbf{h}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)}\right)\right\}. \tag{4.1.17}$$

An additional factor $\hat{C}\log\varepsilon^{-1}$ *is allowed for operations spent on sorting arrays (see Remark* 4.10*). The constant* $\hat{C}$ *depends in all cases only on the constants in* (2.4.7), (2.2.1), *on* s *when* s *tends to infinity, and on the constant* $\bar{C}$ *in* (4.1.15).

Proof. In view of (4.1.15), only a uniformly bounded number of blocks of perturbed iterations (4.1.13) separated by coarsening steps is needed to reduce the current error bound from $\bar{\varepsilon}$ to ε; see Proposition 4.2 in [9]. This number depends clearly on the bound $\bar{C}$ for the ratio $\bar{\varepsilon}/\varepsilon$. Each block, in turn, involves a uniformly bounded

number $\bar{K}$ of perturbed applications of (4.1.13), where $\bar{K}$ depends only on the constants in (2.4.7) and (2.2.1). The claim follows now immediately from Propositions 4.2 and 4.4 (see also the proof of Theorem 5.7 in [9]). ∎

The main result can now be formulated as follows.

THEOREM 4.9. *Assume that the scaled wavelet representations* $\mathbf{A}^\circ$, $\mathbf{B}$ *in* (2.2.4), *and* $\mathbf{R}$ *from* (3.1.3) *belong to* $\mathcal{C}_{s^*}$ *for some* $s^* > 0$, *where the underlying wavelet bases* Ψ_X, Ψ_M *satisfy* (2.2.1), (2.2.2). *Moreover, assume that* (2.1.2) *is well-posed in the sense of* (2.1.4). *If the exact solution* (u, p) *of* (2.1.2) *satisfies for some* $s < s^*$

$$\inf_{\#\mathrm{supp}\mathbf{v}\le N} \|u-\mathbf{v}^T\mathbf{D}_X^{-1}\Psi_X\|_X \lesssim N^{-s}, \quad \inf_{\#\mathrm{supp}\mathbf{q}\le N} \|p-\mathbf{q}^T\mathbf{D}_M^{-1}\Psi_M\|_M \lesssim N^{-s}, \ \ N\to\infty, \tag{4.1.18}$$

then the approximations $(\mathbf{u}(\varepsilon), \mathbf{p}(\varepsilon))$ *produced by* **UZAWA**^{c} *satisfy*

$$\|u-\mathbf{u}(\varepsilon)^T\mathbf{D}_X^{-1}\Psi_X\|_X \lesssim (\#\mathrm{supp}\,\mathbf{u}(\varepsilon))^{-s}, \quad \|p-\mathbf{p}(\varepsilon)^T\mathbf{D}_M^{-1}\Psi_M\|_M \lesssim (\#\mathrm{supp}\,\mathbf{p}(\varepsilon))^{-s}. \tag{4.1.19}$$

Moreover, under the Assumptions 3.1 *and* 4.7, *the computational work needed to compute* $\mathbf{u}(\varepsilon), \mathbf{p}(\varepsilon)$ *is also of the order* $\varepsilon^{-1/s}$ *(except for additional* log *terms for sorting).*

Proof. First note that, by (2.2.1) and (2.2.2), $\sigma_{N,\ell_2(\mathcal{J}_X)}(\mathbf{u}) \lesssim N^{-s}$ and $\sigma_{N,\ell_2(\mathcal{J}_M)}(\mathbf{p}) \lesssim N^{-s}$. Proposition 4.1 says that then $\mathbf{u} \in \ell_\tau^w(\mathcal{J}_X)$ and $\mathbf{p} \in \ell_\tau^w(\mathcal{J}_M)$. It follows now from (2.2.4) and Remark 4.5 that $\mathbf{g} \in \ell_\tau^w(\mathcal{J}_M)$. Since by the same argument $\mathbf{B}^T\mathbf{p}, \mathbf{A}\mathbf{u} \in \ell_\tau^w(\mathcal{J}_X)$, (2.4.5) says that also $\mathbf{f} \in \ell_\tau^w(\mathcal{J}_X)$, i.e.,

$$\|\mathbf{g}\|_{\ell_\tau^w(\mathcal{J}_M)} \lesssim \|\mathbf{u}\|_{\ell_\tau^w(\mathcal{J}_X)}, \quad \|\mathbf{f}\|_{\ell_\tau^w(\mathcal{J}_X)} \lesssim \|\mathbf{u}\|_{\ell_\tau^w(\mathcal{J}_X)} + \|\mathbf{p}\|_{\ell_\tau^w(\mathcal{J}_M)}. \tag{4.1.20}$$

We proceed now estimating the computational cost of one call of **ADV** adhering to the notation used in this context before. We will make frequent use of the fact that all accuracy tolerances appearing in **ADV** remain, in view of the uniform boundedness of K, proportional to the current accuracy $\delta = \delta_J$ in the Jth call of **ADV** in **UZAWA**^{c}. We begin with step (ii) of **ADV**, since the effect of **NCOARSE** determines also the properties of the input of **ADV**. To this end, recall that the perturbed iterate $\bar{\mathbf{p}}_K$, which forms one of the inputs of the coarsening step, is a finitely supported approximation of $\mathbf{p}$, satisfying the error estimate (3.3.11). It is important to note that, regardless of the sizes of its support, Proposition 4.2 implies that then

$$\#(\mathrm{supp}\,\tilde{\mathbf{p}}) \le C\delta^{-1/s}\|\mathbf{p}\|^{1/s}_{\ell_{\tau'}^w(\mathcal{J}_M)}, \qquad \|\tilde{\mathbf{p}}\|_{\ell_{\tau'}^w(\mathcal{J}_M)} \le C\|\mathbf{p}\|_{\ell_{\tau'}^w(\mathcal{J}_M)}, \tag{4.1.21}$$

where C depends only on s when s tends to infinity. Likewise, in view of the error bound (3.3.14) for $\bar{\mathbf{u}}_K$, Proposition 4.2 ensures that

$$\|\tilde{\mathbf{u}}\|_{\ell_\tau^w(\mathcal{J}_X)} \le C\|\mathbf{u}\|_{\ell_\tau^w(\mathcal{J}_X)}, \quad \#\mathrm{supp}\,\tilde{\mathbf{u}} \le C\delta^{-1/s}\|\mathbf{u}\|^{1/s}_{\ell_\tau^w(\mathcal{J}_X)}, \tag{4.1.22}$$

where, as before, $\delta = \delta_J$ is the current accuracy level in the Jth call of **ADV** in **UZAWA**^{c}. As before in (4.1.21), the constant C is independent of $\bar{\mathbf{u}}_K$.

We still have to control the computational cost of the intermediate steps (iii)–(iv) of **ADV**, leading to the final perturbed iterates $\bar{\mathbf{p}}_K, \bar{\mathbf{u}}_K$, which are then subjected to the coarsening step that led to the above estimates. To this end, we infer from Remark 4.5, Propositions 4.2 and 4.4 that

$$\#(\mathrm{supp}\,\bar{\mathbf{p}}_i) \le C\delta^{-1/s}\left(\|\bar{\mathbf{u}}_i\|^{1/s}_{\ell_\tau^w(\mathcal{J}_X)} + \|\mathbf{g}\|^{1/s}_{\ell_\tau^w(\mathcal{J}_M)}\right) + \#(\mathrm{supp}\,\bar{\mathbf{p}}_{i-1}), \quad i = 1,\dots,K, \tag{4.1.23}$$

and

$$\|\bar{\mathbf{p}}_i\|_{\ell^w_{\tau'}(\mathcal{J}_M)} \leq C\left(\|\bar{\mathbf{p}}_{i-1}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\bar{\mathbf{u}}_i\|_{\ell^w_\tau(\mathcal{J}_X)}\right), \quad 1,\dots,K. \tag{4.1.24}$$

Thus we have to estimate next the quantities $\|\bar{\mathbf{u}}_i\|_{\ell^w_\tau(\mathcal{J}_X)}$, $\#\operatorname{supp}\bar{\mathbf{u}}_i$, $i = 1,\dots,K$. To this end, we first have to determine the accuracy of $\bar{\mathbf{u}}_{i-1}$ as an initial guess for $\mathbf{ELLSOLVE}\,[\gamma_i\rho^i\delta/2, \mathbf{A}, \bar{\mathbf{u}}_{i-1}, \mathbf{h}_i]$. In fact, a little care is needed because the right-hand sides $\mathbf{h}_i$ change. To this end, let $\check{\mathbf{u}}_i$ denote the exact solution of $\mathbf{A}\check{\mathbf{u}}_i = \mathbf{h}_i$; see (iii) in **ADV**. Then, by (3.3.13), for $\delta = \delta_J$ in the Jth call of **ADV** in $\mathbf{UZAWA}^c$ one obtains for some constant C

$$\begin{aligned}
\|\check{\mathbf{u}}_i - \bar{\mathbf{u}}_{i-1}\|_{\ell_2(\mathcal{J}_X)} &\leq \|\check{\mathbf{u}}_i - \mathbf{u}\|_{\ell_2(\mathcal{J}_X)} + \|\mathbf{u} - \bar{\mathbf{u}}_{i-1}\|_{\ell_2(\mathcal{J}_X)} \\
&\leq c_A^{-1}\|\mathbf{f} - \mathbf{B}^T\mathbf{p} - \mathbf{h}_i\|_{\ell_2(\mathcal{J}_X)} + C\delta \\
&\leq c_A^{-1}\left(\|\mathbf{f} - \mathbf{B}^T\mathbf{p} - (\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1})\|_{\ell_2(\mathcal{J}_X)}\right. \\
&\qquad \left. + \|\mathbf{f} - \mathbf{B}^T\bar{\mathbf{p}}_{i-1} - \mathbf{h}_i\|_{\ell_2(\mathcal{J}_X)}\right) + C\delta \\
&\leq c_A^{-1}C_B\|\mathbf{p} - \bar{\mathbf{p}}_{i-1}\|_{\ell_2(\mathcal{J}_M)} + \gamma_i\rho^i\delta/2 + C\delta \\
&\leq C'\delta_J,
\end{aligned}$$

where we have used (3.3.10) and (3.2.8). Thus, the ratio of initial and target accuracies in each call of **ELLSOLVE** remains uniformly bounded by a constant $\bar{C}$ depending on the number K in **ADV**, so that Proposition 4.8 applies. To this end, consider first $i = 1$ in step (iv) of **ADV**. By the above bound (4.1.21) on $\bar{\mathbf{p}}_0 = \tilde{\mathbf{p}}^{J-1}$, Remark 4.5, Propositions 4.2, 4.4, and steps (i), (ii) in **RHS**, we conclude that

$$\|\mathbf{h}_1\|_{\ell^w_\tau(\mathcal{J}_X)} \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{f}\|_{\ell^w_\tau(\mathcal{J}_X)}) \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J}_X)}), \tag{4.1.25}$$

where we have used (4.1.20) in the last step. Here and in what follows, unless stated otherwise, C will be a constant (that may vary from place to place) which is independent of $\mathbf{u}, \mathbf{p}$ and at most dependent on the problem constants as before. Proposition 4.8 combined with (4.1.22) now implies

$$\begin{aligned}
&\|\bar{\mathbf{u}}_1\|_{\ell^w_\tau(\mathcal{J}_X)} \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J}_X)}), \\
&\#(\operatorname{supp}\bar{\mathbf{u}}_1) \leq C\left(\#(\operatorname{supp}\bar{\mathbf{u}}_0) + \delta^{-1/s}(\|\mathbf{p}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)})\right).
\end{aligned} \tag{4.1.26}$$

Again keeping (4.1.21) in mind and substituting (4.1.26) in (4.1.24) for $i = 1$, we obtain

$$\|\bar{\mathbf{p}}_1\|_{\ell^w_\tau(\mathcal{J}_M)} \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J}_X)}). \tag{4.1.27}$$

We can now repeat this argument K times and obtain that for all $i \leq K$,

$$\begin{aligned}
&\|\bar{\mathbf{u}}_i\|_{\ell^w_\tau(\mathcal{J}_X)} \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J}_X)}), \\
&\#(\operatorname{supp}\bar{\mathbf{u}}_i) \leq C\left(\#(\operatorname{supp}\bar{\mathbf{u}}_0) + \delta^{-1/s}(\|\mathbf{p}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)})\right), \\
&\|\bar{\mathbf{p}}_i\|_{\ell^w_\tau(\mathcal{J}_M)} \leq C(\|\mathbf{p}\|_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|_{\ell^w_\tau(\mathcal{J}_X)}), \\
&\#(\operatorname{supp}\bar{\mathbf{p}}_i) \leq C\left(\#(\operatorname{supp}\bar{\mathbf{p}}_{i-1}) + \delta^{-1/s}(\|\mathbf{p}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_M)} + \|\mathbf{u}\|^{1/s}_{\ell^w_\tau(\mathcal{J}_X)})\right).
\end{aligned} \tag{4.1.28}$$

In view of the operations count given in Propositions 4.4, 4.8, estimate (4.1.28) says that the computational cost in steps (iii)–(v) remains proportional to $\delta^{-1/s}$. Of course,

the constant C of proportionality depends on the number of steps K and may build up if the perturbed iterations were simply continued. However, the thresholding in step (ii) of **ADV** produces a new constant that no longer depends on K and in this sense sets the proportionality constant back. Thus, we conclude that, under the given assumptions on the exact solutions $\mathbf{u}, \mathbf{p}$, the convergence rate N^{-s} is indeed preserved by $\mathbf{UZAWA}^c$ within the claimed bounds for the corresponding computational work. The assertion follows now directly from Corollary 3.5, (3.3.15). □

REMARK 4.10. *One should note that a strict ordering of the wavelet coefficients by size is actually not essential. What matters is to group the coefficients in binary bins, i.e., to collect all those coefficients whose modulus falls into* $[a2^{-j}, a2^{-j+1})$*, say. In this way one can avoid the logarithmic terms appearing in the work counts for sorting; see* [1].

5. Applications to the Stokes problem. In this section the above developments will be applied to a classical example, namely the *Stokes problem*. In particular, we shall identify suitable wavelet bases and determine the compressibility range s^* for which Theorem 4.9 ensures asymptotically optimal performance.

5.1. The continuous problem. We consider a Lipschitz domain $\Omega \subset \mathbb{R}^d$ and assume for simplicity homogeneous boundary conditions; i.e.,

$$-\Delta u + \nabla p = f, \quad \nabla \cdot u = 0 \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad u|_{\partial\Omega} = 0. \tag{5.1.1}$$

The standard L_2 inner product on a domain G will be denoted by $\langle v, w\rangle_G := \int_G v(x)\, w(x)\, dx$, where we will drop the subscript whenever the inner product refers to Ω. The mixed formulation takes the form (2.1.2) with

$$X = H_0^1(\Omega)^d, \quad M = L_{2,0}(\Omega) := \left\{ v \in L_2(\Omega) : \int_\Omega v(x)\, dx = 0 \right\}, \tag{5.1.2}$$

and

$$a(u, v) := (\nabla u, \nabla v), \qquad b(q, v) := -\langle \nabla \cdot v, q\rangle. \tag{5.1.3}$$

It is well known that (2.1.1) holds in this case even with the stronger relation (2.4.1), so that (2.1.4) is true for (5.1.2). In view of the preceding discussion we have to address the following issues. First, we identify a class of suitable wavelet bases which will be employed later in numerical experiments. Then we determine the compressibility range of the corresponding wavelet representations. Next, we discuss the regularity of the solution to (5.1.1) in a certain scale of Besov spaces. Although this information has no effect on the algorithmic realization, it will allow us to determine under which principal circumstances the adaptive scheme offers even an asymptotically better work/accuracy balance than discretizations based on uniform mesh refinements. These results will guide the selection of our test examples.

5.2. Wavelet representation. When Ω can be partitioned into regular parametric images $\Omega_l = \kappa_l(\square)$ of the unit d-cube $\square := (0, 1)^d$, one can use the constructions from [6, 17] yielding conforming trial spaces for the velocities and pressure. We proceed now collecting the relevant properties of these bases in the present context.

We will reserve the notation Ψ_X for the wavelet basis for $X = H_0^1(\Omega)^d$; i.e., each wavelet $\psi_{X,\lambda}$ is a vector valued function with components $\psi_{\lambda,i}, \lambda \in \mathcal{J}_X$, $i = 1, \ldots, d$. A wavelet $\psi_{\lambda,i}$ which is supported in a single patch Ω_l is then constructed as a linear combination of tensor product B-splines of (coordinatewise) order m_X (which is for

simplicity taken to be the same for each component i) composed with κ_l^{-1}. Wavelets whose support intersects several domains are obtained by suitably patching together such functions across interfaces; see [6, 17] for details. At this point a word on the nature of the indices λ is in order. Without going into details, λ encodes the spatial location of the wavelet $\psi_{X,\lambda}$ as well as its scale denoted by $|\lambda|$. We will employ only compactly supported wavelets whose supports then scale like $\operatorname{diam}(\operatorname{supp}\psi_\lambda) \sim 2^{-|\lambda|}$. The coarsest scale $|\lambda| = 0$ corresponds to finitely many functions, which roughly speaking span the polynomial part in an expansion. Thus for each component i the corresponding *multiresolution spaces* $S_{i,J} := \operatorname{span}\{\psi_{\lambda,i} : |\lambda| < J\}$ can be viewed as trial spaces on meshes of size 2^{-J}. To have a conforming discretization the $S_{i,J}$ are arranged to be contained in $H_0^1(\Omega)$. Being generated by m_Xth order B-splines, they realize approximation order m_X in $H^{m_X}(\Omega) \cap H_0^1(\Omega)$. Such a basis can be realized for any order $m_X \in \mathbb{N}$. We will later vary this order, keeping in mind that the restrictions to a patch Ω_l satisfy

$$\Psi_X|_{\Omega_l} \subset H^{m_X - 1/2}(\Omega_l)^d; \tag{5.2.1}$$

see [24]. Moreover, recall that a wavelet basis consists of two disjoint collections of functions Ψ_X^+ and Ψ_X^- (and analogously for Ψ_M). As indicated above, Ψ_X^+ is comprised of finitely many *scaling functions* of level $|\lambda| = 0$ whose preimages under the parametric mappings span all polynomials of order m_X on $\square$ (up to boundary conditions). The infinite collection Ψ_X^- contains the "true wavelets" in the following sense. In fact, the construction of Ψ_X involves a second important parameter $\tilde{m}_X$. Given any m_X, one can take any $\tilde{m}_X \in \mathbb{N}$, $\tilde{m}_X \geq m_X$ such that $m_X + \tilde{m}_X$ is even and arrange Ψ_X so that for any $\psi_{\lambda,i}$ supported in Ω_l the following m_Xth order *moment conditions* hold:

$$(P, \psi_{\lambda,i})_{\Omega_l} = 0 \qquad \text{for all } P \in \Pi_{\tilde{m}_X,\kappa_l},\quad \psi_{\lambda,i} \in \Psi_i^-, \tag{5.2.2}$$

where $(\cdot,\cdot)_{\Omega_l}$ denotes the standard inner product on the subdomain Ω_l. Here $\Pi_{\tilde{m},\kappa_l} := \{P : P = g_l Q \circ \kappa_l^{-1}, Q \in \Pi_{\tilde{m}}\}$, where $g_l := |\det \partial\kappa_l^{-1}|$ and $\Pi_{\tilde{m}}$ denotes the space of all polynomials of degree less than $\tilde{m}$. With a slight abuse of terminology we will refer to the elements of Π_{m,κ_l} simply as *polynomials*. In fact, since by assumption the g_l are smooth and bounded away from zero, the local approximation properties of $\Pi_{\tilde{m},\kappa_l}$ are the same as those of $\Pi_{\tilde{m}}$, which is what matters for the compression properties.

The pressure functions will be expanded in a basis $\Psi_M = \{\psi_{M,\lambda} : \lambda \in \mathcal{J}_M\}$, which is also generated by B-splines of order m_M in the above sense. Likewise the order of moment conditions will be denoted by $\tilde{m}_M$, i.e.,

$$(P, \psi_{M,\lambda})_{\Omega_l} = 0 \quad \text{for all } P \in \Pi_{\tilde{m}_M,\kappa_l},\ \psi_{M,\lambda} \in \Psi_M^-. \tag{5.2.3}$$

REMARK 5.1. *There are some important distinctions between Ψ_X and Ψ_M though (aside from the fact that Ψ_X is vector and Ψ_M is scalar valued). First, $\psi_{M,\lambda}$ do not satisfy any boundary conditions. Moreover, the moment conditions hold everywhere in Ω since all wavelets are always fully supported in a single patch Ω_l; i.e., the wavelets need not be continuous across patch interfaces.*

Since by (5.2.3) the wavelets in Ψ_M^- have zero mean, an ab initio wavelet basis for $L_2(\Omega)$ can easily be transformed into one for the constrained space $L_{2,0}(\Omega)$ by modifying only the finitely many elements in Ψ_M^+, a fact that will be important later in the numerical realization.

It has been shown in [6, 17] that bases Ψ_X and Ψ_M satisfy the norm equivalences (2.2.1) and (2.2.2) with scaling weights

$$(D_X)_\lambda := 2^{|\lambda|}, \quad (D_M)_\lambda := 1. \tag{5.2.4}$$

In fact, the alternative choice $(D_X)_\lambda := a(\psi_{X,\lambda}, \psi_{X,\lambda})^{1/2}$ typically gives rise to quantitatively better results, but we will stick with (5.2.4) for simplicity.

Hence, the resulting wavelet representations $\mathbf{A}$ and $\mathbf{B}^T$ are of the following form:

$$\mathbf{A} = (a_{\lambda,\lambda'})_{\lambda,\lambda'\in\mathcal{J}_X}, \quad a_{\lambda,\lambda'} = \sum_{i,l=1}^{d} 2^{-(|\lambda|+|\lambda'|)} \int_\Omega \frac{\partial\psi_{\lambda,i}}{\partial x_l}\frac{\partial\psi_{\lambda',i}}{\partial x_l}dx, \tag{5.2.5}$$

$$\mathbf{B}^T = (b_{\lambda,\lambda'})_{\lambda\in\mathcal{J}_X,\lambda'\in\mathcal{J}_M}, \quad b_{\lambda,\lambda'} = -\sum_{i=1}^{d} 2^{-|\lambda|} \int_\Omega \psi_{M,\lambda'}(x)\frac{\partial\psi_{\lambda,i}}{\partial x_i}(x)dx. \tag{5.2.6}$$

5.3. Compression properties. The matrices $\mathbf{A}$, $\mathbf{B}$, defined by (5.2.5) and (5.2.6), are known to be compressible in a range that depends on the regularity of the wavelets; see [8]. However, the special piecewise polynomial nature of the above bases allows us to establish a somewhat larger range of compressibility compared with the general estimates in [8]. In this subsection, we briefly discuss the compression properties of the matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{B}^T$. The analysis is based on the following version of the *Schur lemma* which follows from interpolation between ℓ_∞ and ℓ_1.

LEMMA 5.2. *Let $\mathbf{T} = (T_{l,l'})_{l\in\mathcal{I},l'\in\mathcal{I}'}$ be a matrix and let $\mathcal{I},\mathcal{I}'$ be countable index sets. Suppose that there exist sequences $(\varpi_l)_{l\in\mathcal{I}}$ and $(\tilde{\varpi}_{l'})_{l'\in\mathcal{I}'}$ such that*

$$\sum_{l'\in\mathcal{I}'} |T_{l,l'}|\tilde{\varpi}_{l'} \le c\varpi_l \qquad \text{and} \qquad \sum_{l\in\mathcal{I}} |T_{l,l'}|\varpi_l \le c\tilde{\varpi}_{l'}, \qquad l\in\mathcal{I},\ l'\in\mathcal{I}'; \tag{5.3.1}$$

then $\|\mathbf{T}\| \le c$.

Our numerical examples refer to the L-shaped domain $\Omega = (-1,1)^2 \setminus (-1,0]^2$. Thus Ω can be decomposed, e.g., into three subpatches Ω_l, $l = 1,2,3$, each being a simple translate of the unit square $(0,1)^2$. The spaces Π_{m,κ_l} consist then of polynomials in the classical sense. The moment conditions (5.2.2) hold then on all of Ω also for those wavelets whose support overlaps more than one subdomain. In this case the truncation rule that produces the compressed matrices $\mathbf{A}_j$ from (4.1.9) reads as follows; see [8, 2]. In order to indicate the role of the spatial dimension we keep the general notation although the example refers to $d = 2$. Given j, set

$$\tilde{a}_{\lambda,\nu} := \begin{cases} a_{\lambda,\nu}, & \big||\lambda| - |\nu|\big| \le j/d, \\ 0, & \text{else.} \end{cases} \tag{5.3.2}$$

Unless otherwise stated, we shall henceforth use the abbreviation $m = m_X$, $\tilde{m} = \tilde{m}_X$.

THEOREM 5.3. *For the matrix $\mathbf{A}$ defined by* (5.2.5) *and any $\epsilon > 0$, the following compression estimate holds:*

$$\|\mathbf{A} - \mathbf{A}_J\| \lesssim 2^{-J(m-3/2-\epsilon)/d}, \quad \textit{i.e.,} \quad \mathbf{A}\in\mathcal{C}_{s^*}, \quad s^* = (m-3/2)/d. \tag{5.3.3}$$

Proof. The estimate (5.3.3) can be established by using Lemma 5.2 with $\mathcal{I} = \mathcal{I}' = \mathcal{J}_X$ and $\varpi_\lambda = \tilde{\varpi}_\lambda = 2^{|\lambda|(1-d)}$ for all $\lambda\in\mathcal{J}_X$. To this end, let $\Omega_{\lambda,i}$, $\mathbb{P}_m$ denote the support of the i-component of ψ_λ and the space of polynomials of degree at most m. We recall that derivatives of wavelets are again wavelets with the order of vanishing

moments increased by one [22]. Exploiting this fact, and recalling that $\tilde{m} \geq m$ and $|\lambda'| \geq |\lambda|$, we obtain

$$|(\nabla\psi_{X,\lambda}, \nabla\psi_{X,\lambda'})| \leq \sum_{i,l=1}^{d} \inf_{P\in\mathbb{P}_m} \left| \left(\frac{\partial\psi_{\lambda,i}}{\partial x_l} - P, \frac{\partial\psi_{\lambda',i}}{\partial x_l} \right) \right|$$
$$\lesssim 2^{|\lambda'|} \sum_{i,l=1}^{d} \inf_{P\in\mathbb{P}_m} \left\| \frac{\partial\psi_{\lambda,i}}{\partial x_l} - P \right\|_{L_2(\Omega_{\lambda',i})},$$

where we have applied (2.2.1) with the weights from (5.2.4) to estimate the term $\|\frac{\partial\psi_{\lambda',i}}{\partial x_l}\|_{L_2}$ by $2^{|\lambda'|}$. Setting $j := |\lambda|$, $j' := |\lambda'|$, since $\frac{\partial\psi_{\lambda,i}}{\partial x_l} \in H^s$, $s < m - 3/2$, a classical Whitney-type estimate therefore yields

$$|\,(\nabla\psi_\lambda, \nabla\psi_{\lambda'})\,| \lesssim \sum_{i,l=1}^{d} 2^{j'} 2^{-j'(m-3/2-\epsilon)} \left| \frac{\partial\psi_{\lambda,i}}{\partial x_l} \right|_{H^{m-3/2-\epsilon}}$$
$$\lesssim \sum_{i=1}^{d} 2^{j'} 2^{-j'(m-3/2-\epsilon)} |\psi_{\lambda,i}|_{H^{m-1/2-\epsilon}}$$
$$\lesssim 2^{j'} 2^{-j'(m-3/2-\epsilon)} 2^{j(m-1/2-\epsilon)} \lesssim 2^{(j-j')(m-3/2-\epsilon)} 2^{j+j'}.$$

Thus, taking the scaling matrix $\mathbf{D}_X$ into account and treating the case $j' \leq j$ in an analogous fashion, we derive

$$|a_{\lambda,\lambda'}| \lesssim 2^{-||\lambda|-|\lambda'||(m-3/2-\epsilon)}. \tag{5.3.4}$$

In view of (5.3.3) and (5.3.1), we have to estimate $\sum_{|j-j'|>J/d} \sum_{|\lambda'|=j'} |a_{\lambda,\lambda'}| 2^{j'(1-d)}$. Again consider the case $j' > j$ first and observe that (5.3.4) can be refined for certain entries because in the present case the wavelets are piecewise polynomial. In fact, the nonvanishing entries correspond only to the wavelets $\psi_{\lambda'}$ for which the support of one component $\psi_{\lambda',i}$ intersects the corresponding singular support $\mathcal{S}_{\lambda',i}$ of $\psi_{\lambda,i}$. The set $\mathcal{S}_{\lambda',i}$ can be viewed as a submanifold of dimension $d-1$ with measure of the order $2^{-j(d-1)}$. Consequently, for $j' > j$, there are at most a fixed constant multiple of $2^{(j'-j)(d-1)}$ many wavelets possessing a nontrivial intersection with $\mathcal{S}_{\lambda',i}$. Therefore we obtain

$$\sum_{|\lambda'|=j'} |a_{\lambda,\lambda'}| \lesssim 2^{(j-j')(m-3/2-\epsilon)} 2^{(j'-j)(d-1)} \lesssim 2^{(j-j')(m-3/2-\epsilon+1-d)} \tag{5.3.5}$$

and hence finally

$$\sum_{j'-j>J/d} \sum_{|\lambda'|=j'} |a_{\lambda,\lambda'}| 2^{j'(1-d)} \lesssim \sum_{j'=j+J/d}^{\infty} 2^{(j-j')(m-3/2-\epsilon+1-d)} 2^{j'(1-d)} \tag{5.3.6}$$
$$\lesssim 2^{j(1-d)} 2^{-J(m-3/2-\epsilon)/d}.$$

The case $j' \leq j$ can be treated analogously, and the second condition in (5.3.1) follows in this case by symmetry, which confirms (5.3.3). ∎

REMARK 5.4. *By combining the results in* [8] *with the analysis in* [13]*, one derives the following bound for the range of compressibility of the wavelet representation of an elliptic differential operator of order* $2t$*:*

$$s^* := \max\left\{0, \min\left\{\frac{\sigma}{d} - \frac{1}{2}, \frac{2t+2\tilde{m}}{d}\right\}\right\}.$$

Here the parameter σ *must satisfy* $t + \sigma < \gamma$*, where* γ *bounds the Sobolev regularity of the wavelets. In the present case one has* $t = 1,\ \gamma = m - 1/2$*, i.e.,* $\sigma = m - 3/2$*, and hence* $s^* = (m-3/2)/d - 1/2$*. Therefore* (5.3.3) *ensures in any spatial dimension a gain in the compression range by* $1/2$ *when compared with the usual estimate in* [8, Proposition 3.4].

For more general domains, when the κ_l are no longer affine, some constructions of wavelet bases guarantee the full order of vanishing moments (5.2.2) only for those wavelets that are supported in a single patch Ω_l. Those wavelets overlapping several subdomains still have at least first order moments, and hence their gradients have second order moments. Of course, this occurs only along a $(d-1)$ dimensional manifold and can be compensated by modifying the compression rule (5.3.2). Moreover, those entries $a(\psi_\lambda, \psi_{\lambda'})$, for which the supports overlap each other but their singular supports (cut regions of tensor product B-splines) do not intersect, are no longer zero. However, since one of the wavelets is arbitrarily smooth throughout the integration domain, the order of vanishing moments increases to $\tilde{m}_X + 1$ so that these entries are much smaller than the remaining ones, which suffices as well. Alternatively, one can employ the construction from [18], where vanishing moments are not constrained through patch interfaces.

A similar result can also be established for the matrix $\mathbf{B}^T$ defined in (5.2.6).

THEOREM 5.5. *Suppose that the order* m_X *of the multiresolution spaces for the velocity space* X *and the order* $\tilde{m}_M$ *of the vanishing moments of the pressure wavelets defined in* (5.2.3) *satisfy* $\tilde{m}_M \geq m_X - 1$*. Then for the matrix* $\mathbf{B}^T$ *defined in* (5.2.6) *and any* $\epsilon > 0$*, the following compression estimate holds:*

$$\|\mathbf{B}^T - \mathbf{B}_J^T\| \lesssim 2^{-J(m-3/2-\epsilon)/d}, \quad \textit{i.e.,} \quad \mathbf{B}^T \in \mathcal{C}_s,\ s < s^* = (m-3/2)/d. \tag{5.3.7}$$

The proof of Theorem 5.5 follows the lines of the proof of Theorem 5.3 with $\mathcal{I} = \mathcal{J}_X,\ \mathcal{I}' = \mathcal{J}_M,\ \varpi_\lambda = 2^{|\lambda|(1-d)},\ \lambda \in \mathcal{J}_X$, and $\tilde{\varpi}_{\lambda'} = 2^{|\lambda'|(1-d)}, \lambda' \in \mathcal{J}_M$.

To determine finally the compressibility of the matrix $\mathbf{R}$ from (3.1.3), we can apply the same reasoning for $\partial\psi_{\lambda,i}/\partial x_l$ and $\psi_{M,\lambda'}$ replaced by $\tilde{\psi}_{M,\lambda}$. Since in this case no derivatives are involved and $\tilde{\Psi}_M$ is patchwise defined just as Ψ_M is, the compressibility range is again determined by the order m_M of the primal basis Ψ_M (which limits the order of the polynomials that can be subtracted in the inner products) and the Sobolev regularity $\tilde{\gamma}_M$ of the dual basis $\tilde{\Psi}$ inside each patch Ω_l. Combining tensor products of the wavelet bases on $[0,1]$ from [16] with parametric mappings allows one to realize therefore any desired order s_R^* of compressibility for $\mathbf{R}$, provided that m_M and $\tilde{\gamma}_M$ are chosen accordingly.

Theorems 5.3 and 5.5 tell us now in which range for a given choice of wavelet bases the general results Theorem 4.9 and Corollary 3.5 assert asymptotically optimal accuracy/work balance for the adaptive solution of the Stokes problem.

5.4. Regularity theory for the Stokes problem. So far we have presented some numerical tools to serve as input for an adaptive scheme that realizes asymptotically optimal convergence rates in (essentially) linear time within a certain range

of error decay orders determined by the compressibility of the involved wavelet representations. A natural question is whether at all or under which circumstances the corresponding accuracy/work balance is better than for technically much simpler schemes based, e.g., on uniformly refined meshes—in brief, when does adaptivity pay? It turns out that this question is inherently related to the *regularity* of the approximated solution. More precisely, while a given order of best approximation from trial spaces for preassigned uniform meshes (referred to as *linear schemes*) is characterized by the *Sobolev* regularity of the approximant, the order of *nonlinear* or *best N-term approximation* is (almost) characterized by the regularity in a certain *Besov scale* to be specified in a moment; see also [19]. To explain this let H^t denote a (closed subspace of a) Sobolev space such as $H_0^1(\Omega)$, respectively, $H_0^1(\Omega)^d$ or $L_{2,0}(\Omega)$ for $t = 0$ and let Υ denote a wavelet basis in H^t satisfying a norm equivalence of the form (2.2.1) with suitable scaling matrix $\mathbf{D}^t$. In analogy to (4.1.1), let

$$\sigma_{N,H^t}(v) := \inf_{\mathbf{w},\#\mathbf{w}\le N} \|v - \mathbf{w}^T(\mathbf{D}^t)^{-1}\Upsilon\|_{H^t} \tag{5.4.1}$$

denote the error of best wavelet N-term approximation in H^t. The following fact has been shown in [12].

PROPOSITION 5.6. *Whenever $t \le r$ for some $r \in \mathbb{R}_+$ depending on the regularity of the wavelet basis, let*

$$\frac{1}{\alpha} = \frac{r-t}{d} + \frac{1}{2} \tag{5.4.2}$$

for some $r \in \mathbb{R}$. Then (for a sufficiently regular basis Υ) one has

$$\sum_{N=1}^{\infty} \Big(N^{(r-t)/d}\sigma_{N,H^t}(v)\Big)^{\alpha} < \infty \quad \text{if and only if} \quad v \in B_\alpha^r(L_\alpha(\Omega)). \tag{5.4.3}$$

Note that $B_\alpha^r(L_\alpha(\Omega))$ is the largest space of smoothness r in L_α which is still embedded in H^t, since (5.4.2) marks the Sobolev embedding line; see the "DeVore diagram" in [19]. Clearly, (2.2.1) says that for $v = \mathbf{v}^T(\mathbf{D}^t)^{-1}\Upsilon$ one has

$$\sigma_{N,H^t}(v) \sim \sigma_{N,\ell_2}(\mathbf{v}). \tag{5.4.4}$$

Moreover, (5.4.3), (5.4.4) mean that when $v \in B_\alpha^r(L_\alpha(\Omega))$ the error of the best N-term approximation of its wavelet coefficients $\mathbf{v}$ decays at least like $\sigma_{N,\ell_2}(\mathbf{v}) \lesssim N^{-(r-t)/d}$. This is sharp in the sense that the exponent $s = (r-t)/d$ is best possible. This subtle gap in the characterization of the Besov spaces is due to the small difference between the classical spaces ℓ_τ (characterizing wavelet coefficients for elements in the Besov space) and the weak-type space ℓ_τ^w characterizing best N-term approximation of the wavelet coefficient sequences in ℓ_2 [19].

These facts suggest asking for the regularity of the solution (u, p) of the Stokes problem (5.1.1) in the relevant Besov scales.

We shall briefly review now some results from [10, 20, 23] concerning the regularity of the solution to the Stokes problem (5.1.1) for the L-shaped domain, which are relevant for the subsequent selection of numerical examples. In our case, one can identify the singular part u_S of the velocity which is independent of smooth right-hand sides f and describes the influence of the domain.

In fact, by specializing the results in [10, 11], one can identify solutions u_S of the Stokes problem (referred to as *singular solution*) exhibiting the strongest singularity

induced by the reentrant corner for smooth right-hand sides. The following result can then be established.

THEOREM 5.7. *Any singular solution u_S of the Stokes problem* (5.1.1) *on the L-shaped domain satisfies*

$$u_S \in B^r_\tau(L_\tau(\Omega))^2 \qquad \text{for all} \quad r > 0, \quad \frac{1}{\tau} = \frac{r-1}{2} + \frac{1}{2}. \tag{5.4.5}$$

Noting that $\nabla p = f + \Delta u$, one concludes (by the shift properties of the gradient and the Laplacian in Besov spaces) that also the pressure has arbitrarily high Besov regularity along the critical embedding line through L_2; see [10]. Furthermore, one can likewise determine (for smooth right-hand sides f) the singular parts p_S of the pressure; see also [20, 23]. The relevant conclusions for the present context can be formulated as follows.

REMARK 5.8. *One can verify that*

$$\begin{aligned} &u \in H^r(\Omega)^2, \quad r < r^*_X \approx 1.54448373678246 \qquad \text{and} \\ &p \in H^r(\Omega), \quad r < r^*_M \approx 0.54448373678246 \end{aligned}$$

(i.e., $u \notin H^{r^*_X}(\Omega)^2$, $p \notin H^{r^*_M}(\Omega)$), which limits the convergence rate of uniform refinements. On the other hand, u and p both have arbitrary *high Besov regularity. Hence, in principle, wavelet bases with high order regularity would give rise to correspondingly high order adaptive approximation rates.*

6. Numerical results. In this section, we present some numerical experiments for the Stokes problem on the planar L-shaped domain $\Omega = (-1,1)^2 \setminus (-1,0]^2$. We employ different versions from the family of wavelet bases Ψ_X and Ψ_M from section 5.2 for velocities and pressure, respectively.

Our objective is not to present a fully matured code but to gain additional quantitative insight that complements the preceding theoretical results of primarily asymptotic nature. This concerns the quantitative effect of "violating" the LBB condition and the tradeoff between larger supports and better compressibility when using higher order wavelets as well as suggestions for further algorithmic variants and developments. For instance, the theoretical estimates, e.g., on the number K of iterations in **ADV**, are presumably overly conservative. So it would be interesting to see experimentally whether typically smaller numbers suffice or whether monitoring residuals pays to realize significantly earlier terminations. Furthermore, we wish to see how the scheme copes with highly singular cases suggested by the discussion in section 5.4 compared with more regular solutions. More extensive tests of variants derived from first experiences will be presented elsewhere.

6.1. Discretization of the pressure. Recall from (5.1.2) that $L_{2,0}(\Omega)$ is the appropriate pressure space. Hence the zero mean constraint requires special care. Here we exploit the fact that all wavelets in Ψ^-_M have, according to (5.2.3), vanishing moments of order $\tilde{m}_M \geq m_M \geq 1$, so that

$$\int_\Omega \psi_{M,\lambda}(x)\,dx = 0, \qquad \lambda \in \mathcal{J}_M, |\lambda| > j_0.$$

Hence for any $q = \mathbf{q}^T \Psi_M$ one has

$$\int_\Omega q(x)\,dx = \sum_{|\lambda|=j_0} q_\lambda \int_\Omega \psi_{M,\lambda}(x)\,dx =: \sum_{|\lambda|=j_0} q_\lambda \alpha_\lambda =: I_\Omega(q).$$

On the other hand, the scaling functions form a partition of unity, i.e.,

$$1 \equiv \sum_{|\lambda|=j_0} \tilde{\alpha}_\lambda \, \psi_{M,\lambda}(x), \qquad x \in \Omega, \qquad \tilde{\alpha}_\lambda := \int_\Omega \tilde{\psi}_{M,\lambda}(x)\, dx = (1, \tilde{\psi}_{M,\lambda}),$$

where $\{\tilde{\psi}_{M,\lambda} : |\lambda| = j_0\}$ is the (explicitly known) *dual basis* for the scaling functions in Ψ_M, i.e., $(\psi_{M,\lambda}, \tilde{\psi}_{M,\lambda'}) = \delta_{\lambda,\lambda'}$; see [6, 17]. Thus, denoting by $\mu(\Omega)$ the Lebesgue measure of Ω, we obtain a projection $P_0 : L_2(\Omega) \to L_{2,0}(\Omega)$ by

$$P_0(q) := \sum_{|\lambda|=j_0} \Big(q_\lambda - \frac{I_\Omega(q)}{\mu(\Omega)} \tilde{\alpha}_\lambda \Big) \psi_{M,\lambda} + \sum_{|\lambda|>j_0} q_\lambda \psi_{M,\lambda}$$

that factors out constants. Hence, realizing the zero mean constraint requires modifications only on the *coarsest* level, whereas the wavelet coefficients remain unchanged. Since operators are applied only approximately, corresponding corrections are needed after applying $\mathbf{B}$ and also after coarsening. Since the projection P_0 depends on the particular primal wavelet basis for $L_2(\Omega)$, all arrays have to refer to the same basis so that the Riesz map $\mathbf{R} = (\tilde{\Psi}_M, \tilde{\Psi}_M)$ is needed in the second step (3.1.7) of the Uzawa iteration.

Note that the present way of factoring out constants is only a first convenient option. A drawback reflected by the experiments below is that due to the nature of P_0 always *all* coarse scale functions will be involved in the pressure approximations. In particular, for higher order trial functions this number grows, so that at least for the first few refinement steps the work/accuracy balance of the scheme is less favorable for the pressure component. Local coarse scale basis functions would remedy this effect.

A detailed description of the routines **APPLY** and **NCOARSE** can be found in [2, 8] combined with the above provisions with respect to the matrix $\mathbf{B}$. As mentioned before, the routine **ELLSOLVE** is essentially the adaptive Poisson solver from [2]. This indicates the principal potential of recycling these basic routines for the treatment of problems with increasing complexity.

6.2. Description of the test cases. We wish to report below on two different test cases. Example (I) corresponds to *the most singular* solution described in section 5.4. As can be seen in Figure 6.1, the pressure exhibits a strong singularity at the reentrant corner. In order to keep the effort for computing an exact reference solution as moderate as possible, we have computed an approximation of the exact solution by truncating p. Of course, this limits the number of iterations of the adaptive algorithm for which meaningful comparisons can be made.

Example (II) involves a pressure which is *localized* around the reentrant corner, has strong gradients, but is smooth. More precisely, we have chosen an exact solution for the velocity which is very similar to the one above and a pressure solution which is constant around the reentrant corner and multiplied by a smooth cutoff function. These functions are displayed in Figure 6.2.

6.3. Choice of the parameters. We expect that some of the constants resulting from the analysis are actually too pessimistic. For instance, deriving estimates for the constants in the norm equivalences, we have estimated K to be in the range of 15, which turned out to entail unnecessarily high accuracy in the treatment of the inner Poisson problems while the pressure approximation and hence the right-hand side for the Laplace problem are still poor. Several numerical experiments with different trial functions and for different test cases indicate that $K = 3$ already seems to suffice

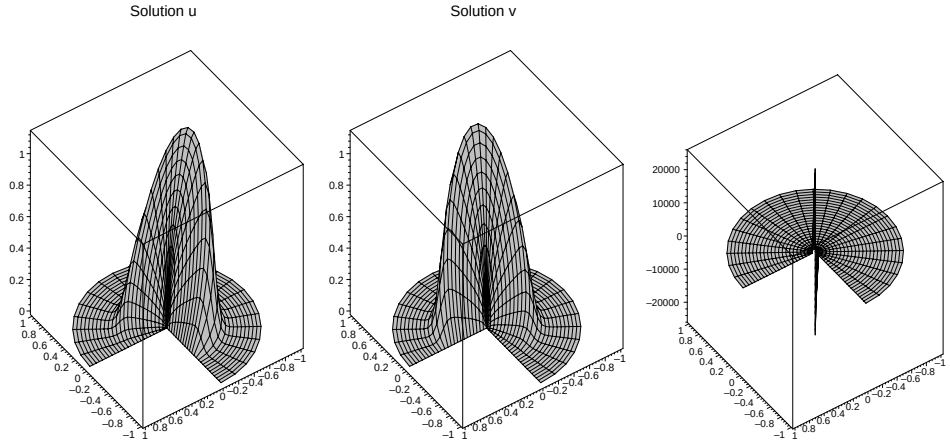


FIGURE 6.1. *Exact solution for the first example. Velocity components (left and middle) and pressure (right). The pressure function exhibits a strong singularity and is only shown up to* $r = 0.001$ *in polar coordinates.*

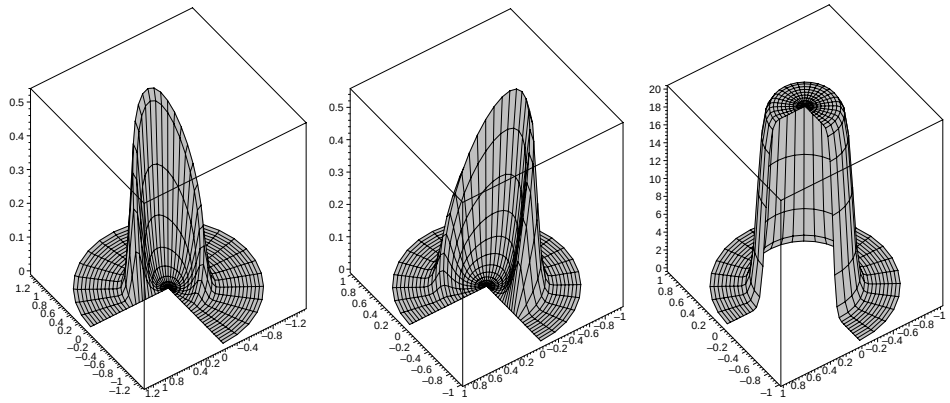


FIGURE 6.2. *Exact solution for the second example. Velocity components (left and middle) and pressure (right).*

and that the alternatives discussed in section 3 are in these cases not necessary. All subsequent results are therefore based on this choice. Moreover, we have used $\rho = 0.6$ and $\omega = 1.3$ in all experiments.

6.4. Rate of convergence. Table 6.1 displays the results for Example (I), employing piecewise linear trial functions for the velocity and piecewise constant functions for the pressure. We are interested in the relation between the error produced for a given number of degrees of freedom by the adaptive scheme and the error of best N-term approximation with respect to the underlying wavelet basis. To describe the results we denote by $\mathbf{u}^1, \mathbf{u}^2$ the wavelet coefficient arrays of the first and second velocity component and for $\mathbf{x} \in \{\mathbf{u}^1, \mathbf{u}^2, \mathbf{p}\}$ by

$$\rho_{\mathbf{x}} := \frac{\|\mathbf{x} - \mathbf{x}_{\mathbf{\Lambda}}\|_{\ell_2}}{\|\mathbf{x} - \mathbf{x}_{\#\mathbf{\Lambda}}\|_{\ell_2}}, \qquad r_{\mathbf{x}} := \frac{\|\mathbf{x} - \mathbf{x}_{\mathbf{\Lambda}}\|_{\ell_2}}{\|\mathbf{x}\|_{\ell_2}},$$

the ratio of the error of the adaptive approximation and the corresponding best N-term approximation and the relative errors of the solution components x_Λ, respec-

TABLE 6.1

Results for Example (I). *Numbers of adaptively generated degrees of freedom, ratio to best N-term approximation, and relative errors.*

It	δ	$\#\Lambda_{\mathbf{u}^1}$	$\rho_{\mathbf{u}^1}$	$r_{\mathbf{u}^1}$	$\#\Lambda_{\mathbf{u}^2}$	$\rho_{\mathbf{u}^2}$	$r_{\mathbf{u}^2}$	$\#\Lambda_{\mathbf{p}}$	$\rho_{\mathbf{p}}$	$r_{\mathbf{p}}$
1	11.730947	33	1.04	0.6838	34	1.04	0.6744	768	130.35	1.0024
2	5.865474	84	1.26	0.3427	83	1.24	0.3447	768	130.40	1.0028
3	2.932737	193	1.32	0.1530	184	1.31	0.1541	768	15.37	0.5234
4	1.466368	446	1.29	0.0821	450	1.29	0.0897	929	4.15	0.2218
5	0.733184	1070	1.27	0.0434	1065	1.27	0.0456	1211	2.58	0.1034

TABLE 6.2

Results for Example (II). *Numbers of adaptively generated degrees of freedom, ratio to best N-term approximation, and relative error.*

It	δ	$\#\Lambda_{\mathbf{u}^1}$	$\rho_{\mathbf{u}^1}$	$r_{\mathbf{u}^1}$	$\#\Lambda_{\mathbf{u}^2}$	$\rho_{\mathbf{u}^2}$	$r_{\mathbf{u}^2}$	$\#\Lambda_{\mathbf{p}}$	$\rho_{\mathbf{p}}$	$r_{\mathbf{p}}$
1	15.636636	278	28.20	1.2936	364	60.31	2.1867	768	6.96	0.3329
2	7.818318	261	8.30	0.4028	295	16.10	0.7003	768	3.76	0.1800
3	3.909159	234	3.72	0.1995	274	5.63	0.2617	768	1.80	0.0863
4	1.954580	180	1.25	0.0886	249	2.08	0.1056	810	1.22	0.0452
5	0.977290	233	1.14	0.0615	267	1.29	0.0615	980	1.07	0.0231
6	0.488645	298	1.11	0.0480	321	1.17	0.0470	1276	1.05	0.0117
7	0.244322	456	1.35	0.0398	505	1.43	0.0265	1551	1.09	0.0061
8	0.122161	704	1.36	0.0250	724	1.39	0.0177	1842	1.24	0.0035

tively. Recall from Corollary 3.5 that these quantities also reflect the error in the energy norms. We see that the velocity approximation is from the beginning very close to its best N-term approximation. For the reasons indicated above this is different for the pressure. The application of P_0 fills up the coarsest level, which in this example has 768 degrees of freedom. To explain this in more detail assume that the adaptive method picks exactly one scaling function, so that the degree of freedom for the pressure would be 1. Since the integral of a scaling function is not zero, the pressure projection P_0 produces a nonzero constant whose expansion involves *all* scaling function coefficients. This is the reason why at the early stage of the refinement process the work accuracy balance for the pressure is less favorable. However, the last two iterates shown in the table indicate that the scheme catches up with the optimal rate. Local coarse scale bases would of course yield better results already from the beginning of the adaptive refinements.

The results for Example (II) are shown in Table 6.2, and plots of the approximations are displayed in Figure 6.3. We see that the computed approximations differ only by a very moderate factor from the best N-term approximation. The results suggest the following directions for more systematic implementations. The simple Richardson iteration should be replaced (possibly after a few initial steps) by gradient or conjugate gradient steps. This should speed up convergence and avoid a necessarily pessimistic estimation of step size parameters. Since all algorithmic ingredients still require the same type of (approximate) matrix/vector multiplications, one can employ the same routines. One should then include, however, monitoring residuals which, due to (2.3.1), should detect rapid convergence for a possible early termination of the iterations in **ADV** (ii). Moreover, higher order wavelets should be tested to exploit larger compressibility ranges.

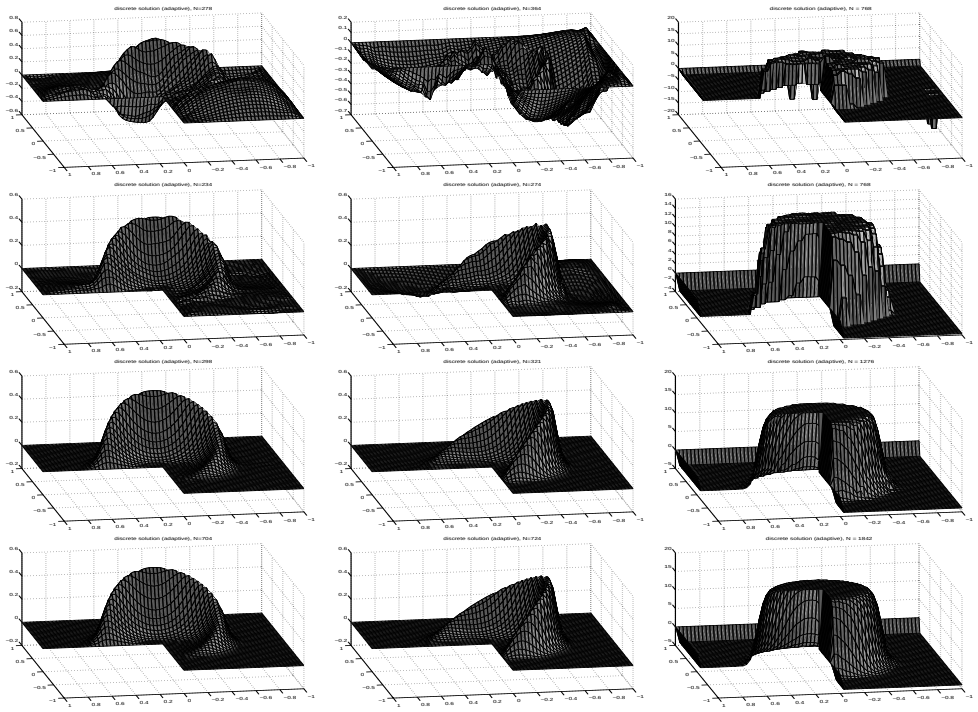


FIGURE 6.3. *First, third, sixth, and eighth approximations for Example* (II). *First and second velocity component (left and middle columns) and pressure (right column).*

6.5. High order discretizations. Recall from section 5.3 that the compressibility range of the wavelet representations grows with increasing regularity and hence order of the wavelet bases; see Theorems 5.3, 5.5. Moreover, the regularity results from Theorem 5.7 and Remark 5.8 indicate that the larger the compressibility range of the wavelet representations, the more an adaptive scheme would gain at least asymptotically over uniform refinements. This suggests investigating the quantitative effect of employing higher order spline wavelets.

We now compare discretizations of various orders for the pressure in the second example. In Figure 6.4, we have shown the relative error versus the number of unknowns in a logarithmic scale. Comparing the slopes of the best N-term approximation, we obtain the expected asymptotic gain for increasing orders, again at the end with moderate values for the ratios $\rho_{\mathbf{x}}$. However, we also see that the fast decay of the rate of the best N-term approximation is delayed more and more for an increasing order of trial functions. For instance, for piecewise cubic wavelets, we obtain an almost horizontal line until $N \approx 2000$. This is, on one hand, due to some technical restrictions of the particular patchwise tensor product wavelet bases used here that require a certain coarsest level j_0 on each patch. The values for j_0 are shown in Table 6.3 for different orders. We see that j_0 increases with m (the case $m = 2$ is somewhat special due to the very local character of primal and dual functions). We display also the number of unknowns for the coarsest level, i.e., the number of scaling functions on level $j = j_0$. On the other hand, as pointed out before, the nature of P_0 keeps all coarse scale basis functions active. This explains why the slope of the best N-term approximation is almost horizontal until all scaling functions are used up. There are

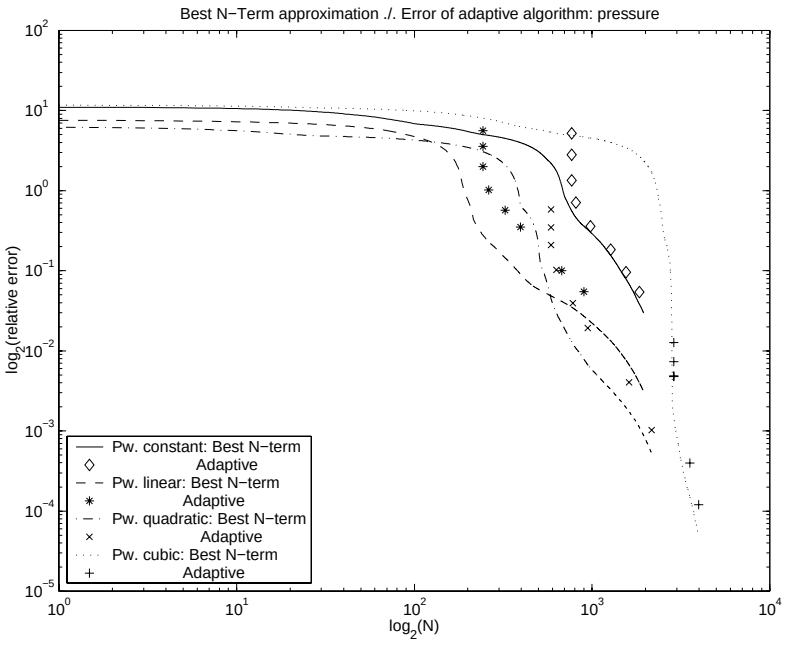


FIGURE 6.4. *Relative error versus number of unknowns for spline wavelets of different order for the discretization of the pressure in the second example.*

TABLE 6.3

Minimal level j_0 and number of scaling functions N_Φ on the minimal level for different order discretizations.

$m, \tilde{m}$	1,3	2,2	3,3	4,4
j_0	4	3	4	5
N_Φ	705	242	587	2882

several ways to alleviate this problem also for higher order discretizations. Aside from using local coarse scale basis functions with zero mean, one can take a fictitious domain approach and append the boundary conditions by Lagrange multipliers. This allows one to use periodic wavelet bases on the fictitious domain where the minimal level can be always chosen as $j_0 = 0$ for all values of m and $\tilde{m}$. This issue will be addressed elsewhere.

6.6. The LBB condition. At first glance it is somewhat puzzling that in the analysis of the adaptive Uzawa method the LBB condition did not play any role. Roughly speaking, this is due to the fact that conceptually at every stage of the algorithm the full infinite dimensional operator is applied within a certain tolerance that has to be chosen tight enough to inherit the stability properties of the original infinite dimensional problem. This effect of adaptive schemes in connection with saddle point problems and also with more complex variational problems has been observed first in [9]; see also [14] for saddle point problems. Hence it is interesting to study the quantitative influence of the choice of bases. Therefore, we have included a combination of bases for which pairs of fixed finite dimensional subspaces would violate the LBB condition, namely, piecewise linear trial functions for both velocity and pressure. The results are displayed in Table 6.4. We see that the rate of the best

TABLE 6.4
Results for the second example with piecewise linear trial functions for velocity and pressure. Note that in this case the number of degrees of freedom for the coarsest level is 243.

It	δ	$\#\Lambda_{\mathbf{u}^1}$	$\rho_{\mathbf{u}^1}$	$r_{\mathbf{u}^1}$	$\#\Lambda_{\mathbf{u}^2}$	$\rho_{\mathbf{u}^2}$	$r_{\mathbf{u}^2}$	$\#\Lambda_{\mathbf{p}}$	$\rho_{\mathbf{p}}$	$r_{\mathbf{p}}$
1	16.743449	1	1.00	0.9293	1	1.00	0.9300	243	6.27552	0.3354
2	8.371724	1	1.00	0.9304	1	1.00	0.9292	243	3.98811	0.2131
3	4.185862	5	1.00	0.7586	5	1.00	0.7588	243	2.23810	0.1196
4	2.092931	20	1.13	0.4064	24	1.45	0.3979	262	2.08107	0.0612
5	1.046466	61	1.47	0.2107	77	1.79	0.2107	324	2.72102	0.0339
6	0.523233	178	1.33	0.1060	198	1.52	0.1306	396	2.81079	0.0209
7	0.261617	294	1.19	0.0533	286	1.46	0.0744	674	2.21371	0.0108
8	0.130808	478	1.25	0.0271	531	1.46	0.0362	899	1.83271	0.0071

N-term approximation is still matched fairly well with ratios that are only slightly larger than in Table 6.2 for the piecewise linear/piecewise constant discretization. Note that the oscillations in the pressure approximation for unstable elements shown by the experiments in [4] are not observed in the present context; see Figure 6.3. This seems to result from the different pressure update.

REFERENCES

[1] A. BARINKA, *Fast Evaluation Tools for Adaptive Wavelet Schemes*, Ph.D. thesis, RWTH Aachen, Aachen, Germany, in preparation.

[2] A. BARINKA, T. BARSCH, P. CHARTON, A. COHEN, S. DAHLKE, W. DAHMEN, AND K. URBAN, *Adaptive wavelet schemes for elliptic problems—implementation and numerical experiments*, SIAM J. Sci. Comput., 23 (2001), pp. 910–939.

[3] A. BARINKA, S. DAHLKE, AND W. DAHMEN, *Adaptive Application of Operators in Standard Wavelet Representation*, manuscript, 2002

[4] E. BÄNSCH, P. MORIN, AND R.H. NOCHETTO, *An adaptive Uzawa FEM for Stokes: Convergence without the inf-sup*, preprint, WIAS, Berlin, 2001.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[6] A. CANUTO, A. TABACCO, AND K. URBAN, *The wavelet element method, part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[7] A. COHEN, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis #7, P.G. Ciarlet and J.L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 417–711.

[8] A. COHEN, W. DAHMEN, AND R.A. DEVORE, *Adaptive wavelet methods for elliptic operator equations—Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[9] A. COHEN, W. DAHMEN, AND R.A. DEVORE, *Adaptive wavelet methods* II - *Beyond the elliptic case*, IGPM report # 199, RWTH Aachen, Aachen, Germany, 2000; Found. Comp. Math., to appear.

[10] S. DAHLKE, *Besov regularity for the Stokes problem*, in Advances in Multivariate Approximation, W. Haußmann, K. Jetter, and M. Reimer, eds., Mathematical Research 107, Wiley VCH, Berlin, 1999, pp. 129–138.

[11] S. DAHLKE, *Besov regularity for elliptic boundary value problems on polygonal domains*, Appl. Math. Lett., 12 (1999), pp. 31–36.

[12] S. DAHLKE, W. DAHMEN, AND R.A. DEVORE, *Nonlinear approximation and adaptive techniques for solving elliptic operator equations*, in Multiscale Wavelet Methods for PDEs, W. Dahmen, A. Kurdila, and P. Oswald, eds., Academic Press, San Diego, 1997, pp. 237–284.

[13] S. DAHLKE, W. DAHMEN, R. HOCHMUTH, AND R. SCHNEIDER, *Stable multiscale bases and local error estimation for elliptic problems*, Appl. Numer. Math., 23 (1997), pp. 21–48.

[14] S. DAHLKE, R. HOCHMUTH, AND K. URBAN, *Adaptive wavelet methods for saddle point problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1003–1022.

[15] W. DAHMEN, *Wavelet methods for PDEs—some recent developments*, J. Comput. Appl. Math., 128 (2001), pp. 133–185.

[16] W. DAHMEN, A. KUNOTH, AND K. URBAN, *Biorthogonal spline-wavelets on the interval—*

stability and moment conditions, Appl. Comput. Harmon. Anal., 6 (1999), pp. 132–196.
[17] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.
[18] W. Dahmen and R. Schneider, *Wavelets on manifolds* I. *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.
[19] R.A. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.
[20] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Mason, Paris, Springer-Verlag, Berlin, 1992.
[21] A. Kunoth, *Wavelet Methods—Elliptic Boundary Value Problems and Control Problems*, Teubner-Verlag, Stuttgart, 2001
[22] P.G. Lemarié-Rieusset, *Analyses multi-résolutions non orthogonales, Commutation entre Projecteurs et Derivation et Ondelettes Vecteurs à divergence nulle*, Rev. Mat. Iberoamericana, 8 (1992), pp. 221–236 (in French).
[23] J. Osborn, *Regularity of solutions to the Stokes problem in a polygonal domain*, in Symposium on Numerical Solutions of Partial Differential Equations III, B. Hubbart, ed., Academic Press, New York, 1976, pp. 393–411.
[24] P. Oswald, *On function spaces related to finite element approximation theory*, Z. Anal. Anwendungen, 9 (1990), pp. 43–64.
[25] J. Vorloeper, *Multiskalenverfahren und Gebietszerlegungsmethoden*, Masters thesis, RWTH Aachen, Aachen, Germany, 1999 (in German).

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1263–1280

SUPERCONVERGENCE FOR THE GRADIENT OF FINITE ELEMENT APPROXIMATIONS BY L^2 PROJECTIONS*

BJØRN-OVE HEIMSUND†, XUE-CHENG TAI†, AND JUNPING WANG‡

Abstract. A gradient recovery technique is proposed and analyzed for finite element solutions which provides new gradient approximations with high order of accuracy. The recovery technique is based on the method of least-squares surface fitting in a finite-dimensional space corresponding to a coarse mesh. It is proved that the recovered gradient has a high order of superconvergence for appropriately chosen surface fitting spaces. The recovery technique is robust, efficient, and applicable to a wide class of problems such as the Stokes and elasticity equations.

Key words. finite element methods, superconvergence, error estimates, adaptive refinement

AMS subject classifications. 65N30, 65N15, 65F10

PII. S003614290037410X

1. Introduction.

It has been known for a long time that finite element solutions of partial differential equations can have superconvergence in some subregions of the domain [26, 23, 8, 1]. *Superconvergence* is a phenomenon that the numerical solution converges to the exact solution at a rate higher than the optimal order error estimate. To exploit superconvergence in the finite element method, several methods have been proposed in the literature in the last 30 years. The method of local averaging has turned out to be a common and useful technique in the study of superconvergence in most of the existing results; see, for example, [23, 8, 1, 7, 35, 19, 18, 20, 17, 21, 25, 26, 10, 13] and the references therein. In theory, all the existing results require the underlying finite element mesh to have some special properties such as uniformity [23, 7, 21], local point symmetry [25, 26], local translation invariance [1, 26], or orthogonality (e.g., rectangular partition) [8, 10, 13, 19, 18, 20, 28, 34].

The Zienkiewicz and Zhu (ZZ) method [32, 33] is a procedure which postprocesses the gradient of the finite element solution by using a discrete least-squares fitting on a local patch with high order polynomials. Due to its high efficiency and robustness, the ZZ postprocessing has been widely used for mesh adaptivity and error control in finite element methods [32, 33, 5, 6]. For appropriately chosen discrete norms, this procedure has been computationally justified to yield some superconvergence for the gradient. If the underlying finite element partition is uniform or rectangular, one can provide a theoretical proof for the ZZ method [31, 34, 30] by using some existing superconvergent estimates [8, 23, 17, 35, 18].

Our objective of this paper is twofold. First, we modify the ZZ method by applying a global least-squares fitting to the gradient of the finite element approximation. The surface fitting space consists of continuous or discontinuous piecewise polynomials of high order on a coarse partition. Second, we provide a theoretical analysis for

*Received by the editors June 21, 2000; accepted for publication (in revised form) February 25, 2002; published electronically September 27, 2002.

http://www.siam.org/journals/sinum/40-4/37410.html

†Department of Mathematics, University of Bergen, Johannes Brunsgt. 12, 5007 Bergen, Norway (bjornoh@math.uib.no, Xue-Cheng.Tai@math.uib.no). The research of the second author was partially supported by the Research Council of Norway (NFR) under grant 128224/431.

‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (jwang@mines.edu). The research of this author was supported in part by the NSF grant DMS-9706985.

the modified ZZ (MZZ) method by establishing a superconvergence estimate for the recovered gradient/flux on general quasi-uniform meshes. To the authors' knowledge, our result is the first that gives a theoretical proof for the superconvergence of the ZZ method with some modifications under general assumptions for the finite element partition. The essential idea behind the approach is the use of a coarser mesh and a higher order of polynomials which can be translated to the method of "long" and "accurate" finite difference quotients. The same idea has been applied in [15] to yield asymptotically exact a posteriori estimators for the pointwise gradient error.

Our presentation follows a framework established in Wang [27] (see also [29]), where the least-squares surface fitting (the projection method) was applied to the finite element solution u_h in order to produce a new and better approximation for the original unknown function $u = u(x)$ and its gradient ∇u. The approach of this paper is different in that the projection will be applied directly to the numerical gradient ∇u_h in order to provide a superconvergent numerical solution for ∇u. Like all the existing results in superconvergence, our results are based on a certain regularity assumption for the exact solution of the underlying model problem.

For simplicity of discussion, our superconvergence result will be presented only for Dirichlet boundary value problems. The results can be extended to Neumann and Robin boundary conditions without any difficulty.

The paper is organized as follows. In section 2, we introduce a model problem for which the required regularity condition is satisfied. In section 3, we present an extension of the ZZ method by using a global least-squares fitting in a high order finite element space corresponding to a coarse mesh. Some error estimates for the new gradient approximation will be derived in section 4. In sections 5 and 6 we apply the error estimate to show that the projected gradient is superconvergent if the fitting space is properly chosen. Section 7 applies the gradient recovery scheme to mesh adaption, and section 8 gives numerical results and comparison for various adaptive schemes.

2. A model problem. To illustrate the idea, we consider boundary value problems for the second order elliptic equation. Let Ω be an open bounded domain in $\mathbb{R}^d, d = 2, 3$. Denote by $x = (x_1, \ldots, x_d)$ the points in Ω. Let $\partial_i = \frac{\partial}{\partial x_i}$ be the partial derivative operator in the direction of $x_i, i = 1, \ldots, d$. The Dirichlet boundary value problem seeks a function $u = u(x)$ such that $u(x) = g(x)$ for any $x \in \partial\Omega$ and

$$\sum_{i,j=1}^{d} \partial_j(a_{ij}\partial_i u) + \sum_{i=1}^{d} b_i \partial_i u + cu = f \qquad \text{in } \Omega, \tag{1}$$

where $\boldsymbol{a} = (a_{ij})_{i,j=1}^d$ is the coefficient tensor which is symmetric, bounded, and uniformly positive definite in the domain Ω with measurable entries $a_{ij} = a_{ij}(x)$. The other coefficients $\boldsymbol{b} = (b_i(x))_{i=1}^d$ and $c = c(x)$ are assumed to ensure a uniqueness of solutions for (1).

Standard notations for Sobolev spaces and norms are adopted in this paper. For an $s \geq 0$, which may not be an integer, and a given domain Ω, $H^s(\Omega)$ denotes the Sobolev space with norm $\|\cdot\|_s$ as defined in [14]. The space $H_0^s(\Omega)$ is a closed subspace of $H^s(\Omega)$ that is the closure of $C_0^s(\Omega)$ (the set of compact-supported C^s functions) in the norm of $H^s(\Omega)$. For $s < 0$, $H^s(\Omega)$ is defined to be the dual space of $H^{-s}(\Omega)$; see [14] for details. The Sobolev space $H^0(\Omega)$ coincides with $L^2(\Omega)$, in which case the norm and inner product are denoted by $\|\cdot\|$ and $(\cdot, \cdot)$, respectively.

Let

$$a(u,v) = \sum_{i,j=1}^{d} \int_\Omega a_{ij}\partial_i u \partial_j v dx + \sum_{i=1}^{d} \int_\Omega b_i \partial_i uv dx + \int_\Omega cuv dx$$

be a bilinear form defined in $H^1(\Omega) \times H^1(\Omega)$. A weak form for the problem (1) seeks a function $u \in H^1(\Omega)$ such that $u = g$ on $\partial\Omega$ and

$$a(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega). \tag{2}$$

Here we have assumed that the boundary data $g \in H^{1/2}(\partial\Omega)$.

Let $s \geq 1$ be a positive real number. Assume that the dual problem of (2) has H^s regularity in the sense that, for any given $f \in H^{s-2}(\Omega)$, the problem

$$a(v,w) = (f,v) \quad \forall v \in H_0^1(\Omega)$$

has a unique solution $w \in H_0^1(\Omega) \cap H^s(\Omega)$ such that

$$\|w\|_s + \left\| \frac{\partial w}{\partial \boldsymbol{n_a}} \right\|_{s-\frac{3}{2},\partial\Omega} \leq C\|f\|_{s-2}, \tag{3}$$

where $\boldsymbol{n}$ is the unit outward normal vector of $\partial\Omega$ and $\frac{\partial w}{\partial \boldsymbol{n_a}} = (\boldsymbol{a}\nabla w)\cdot \boldsymbol{n}$ denotes the normal component of the flux variable on the boundary $\partial\Omega$ for the dual solution w. It is well known that the bilinear form $a(\cdot,\cdot)$ is bounded in $H^1(\Omega)$. In other words, there exists a constant C such that

$$|a(u,v)| \leq C\|u\|_1\|v\|_1 \quad \forall\ u,v \in H^1(\Omega).$$

The finite element solution of (2) is a function $u_h = u_h(x)$ from a finite element space $S_h \subset H^1(\Omega)$ associated with a prescribed finite element partition Ω_h such that $u_h(x) = g_h(x)$ for all $x \in \partial\Omega$ and

$$a(u_h,v) = (f,v) \quad \forall v \in S_h^0. \tag{4}$$

Here $S_h^0 = H_0^1(\Omega) \cap S_h$ and g_h is a certain approximation of the Dirichlet boundary data g. Let Λ_h be the restriction of the finite element space S_h on the boundary of Ω. For simplicity, we shall deal with polygonal or polyhedral domain Ω so that the boundary $\partial\Omega$ is exactly represented by the finite element partition Ω_h. Among many possibilities, we are particularly interested in two cases for the approximate boundary data:

- *g_h is the standard nodal interpolation of g in Λ_h for sufficiently smooth g.*
- *g_h is the L^2 projection of g in Λ_h.*

We recall that the L^2 projection of g in Λ_h is given by solving the following system of linear equations:

$$\langle g_h, v\rangle = \langle g, v\rangle \quad \forall v \in \Lambda_h, \tag{5}$$

where $\langle\cdot,\cdot\rangle$ is the standard L^2-inner product on $\partial\Omega$.

Assume that S_h consists of continuous piecewise polynomials of order $k \geq 1$. Let h be the mesh parameter for the finite element partition Ω_h. The finite element space S_h is assumed to have the following approximation property:

$$\inf_{v\in S_h} (\|w-v\| + h\|w-v\|_1) \leq Ch^m\|w\|_m \quad \forall w \in H^m(\Omega)$$

for any $0 \leq m \leq k+1$.

3. Gradient recovery by projections. To obtain an approximate gradient and flux with superconvergence, we consider a new finite-dimensional space $\mathcal{L}_\tau$ with parameter $\tau \gg h$ and a higher order approximation property than S_h [15]. The functions in $\mathcal{L}_\tau$ are vector valued and will be employed to approximate the exact gradient/flux variable $\boldsymbol{q} = \boldsymbol{a}\nabla u$. In practice, the mesh parameter τ is proportional to h^α for some $\alpha \in (0,1)$ in order to obtain a superconvergent approximation from the projection space $\mathcal{L}_\tau$. Details can be found from Wang [27].

For simplicity, assume that $\mathcal{L}_\tau$ is a finite element space associated with another finite element partition Ω_τ and consists of piecewise polynomials of order $r \geq 0$. The finite element space $\mathcal{L}_\tau$ is required to satisfy the following properties:

- Inverse property.

$$\|\boldsymbol{v}_\tau\|_{[H^m(K)]^d} \leq C\tau^{-m}\|\boldsymbol{v}_\tau\|_{[L^2(K)]^d} \quad \forall \boldsymbol{v}_\tau \in \mathcal{L}_\tau, \quad \forall K \in \Omega_\tau \tag{6}$$

  for all nonnegative integer $m \geq 0$.

- Approximation property.

$$\inf_{\boldsymbol{v}_\tau \in \mathcal{L}_\tau} \|\boldsymbol{v} - \boldsymbol{v}_\tau\|_0 \leq C\tau^m \|\boldsymbol{v}\|_m \quad \forall \boldsymbol{v} \in [H^m(\Omega)]^d, \quad 0 \leq m \leq r+1. \tag{7}$$

- Smoothness property.

$$\mathcal{L}_\tau \subset [H^{s-1}(\Omega)]^d. \tag{8}$$

  Here $s \geq 1$ is associated with the regularity of the dual problem as indicated in (3).

We emphasize that the space $\mathcal{L}_\tau$ can be replaced by other finite-dimensional spaces as trigonometric functions, B-splines, and any special functions if the domain is of special type. In such cases the approximation property and the inverse inequality will be different, and the forthcoming analysis must be modified accordingly.

3.1. Recovery based on a mixed formulation. Our objective here is to provide a very accurate approximation for the flux variable $\boldsymbol{q}$ by using the finite element solution u_h. The relation between the flux $\boldsymbol{q} = \boldsymbol{q}(x)$ and the original function $u = u(x)$ can be rewritten as follows:

$$\boldsymbol{a}^{-1}\boldsymbol{q} = \nabla u. \tag{9}$$

Let

$$H(\mathrm{div};\Omega) = \{\boldsymbol{v} : \ \boldsymbol{v} \in [L^2(\Omega)]^d, \ \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\}$$

be equipped with the norm

$$\|\boldsymbol{v}\|_H = \left(\|\boldsymbol{v}\|^2 + \|\nabla \cdot \boldsymbol{v}\|^2\right)^{1/2}.$$

By testing (9) against any $\boldsymbol{v} \in H(\mathrm{div};\Omega)$ we arrive at

$$(\boldsymbol{a}^{-1}\boldsymbol{q}, \boldsymbol{v}) = -(u, \nabla \cdot \boldsymbol{v}) + \langle g, \boldsymbol{v} \cdot \boldsymbol{n}\rangle, \tag{10}$$

where we have employed the integration by parts to the term on the right-hand side.

Equation (10) can be employed to provide a new flux/gradient recovery $\tilde{\boldsymbol{q}}_\tau$ defined as follows:

$$(\boldsymbol{a}^{-1}\tilde{\boldsymbol{q}}_\tau, \boldsymbol{v}) = -(u_h, \nabla \cdot \boldsymbol{v}) + \langle g, \boldsymbol{v} \cdot \boldsymbol{n}\rangle \quad \forall \boldsymbol{v} \in \mathcal{L}_\tau. \tag{11}$$

The new flux approximation $\tilde{\boldsymbol{q}}_\tau$ will be denoted by

$$\tilde{\boldsymbol{q}}_\tau = \tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h.$$

It is clear that $\tilde{\boldsymbol{Q}}_\tau$ can be regarded as a linear operator onto the fitting space $\mathcal{L}_\tau$.

Since (11) was obtained by using test functions in the space $H(\mathrm{div};\Omega)$, the fitting space $\mathcal{L}_\tau$ has to be constructed as a finite element subspace of $H(\mathrm{div};\Omega)$. In practical computation, the standard mixed finite element spaces of Raviart and Thomas [24], Brezzi et al. [2, 3], Brezzi, Douglas, and Marini [4], and Douglas and Wang [11] can be employed to accomplish the goal. Of course, one can also use continuous finite element spaces in the place of $\mathcal{L}_\tau$. The well-known *inf-sup* condition is no longer an issue in this procedure because the flux computed is based on a Galerkin approximation of the scalar variable.

3.2. Recovery based on L^2 projections. Let $\boldsymbol{Q}_\tau$ be the weighted L^2 projection onto the fitting space $\mathcal{L}_\tau$ with respect to the weighted inner product $(\boldsymbol{a}^{-1}\cdot,\cdot)$. More precisely, for any $\boldsymbol{v} \in \left[L^2(\Omega)\right]^d$, the projection $\boldsymbol{Q}_\tau \boldsymbol{v}$ is a function in $\mathcal{L}_\tau$ such that

$$(\boldsymbol{a}^{-1}\boldsymbol{Q}_\tau \boldsymbol{v}, \boldsymbol{\phi}) = (\boldsymbol{a}^{-1}\boldsymbol{v}, \boldsymbol{\phi}) \quad \forall \boldsymbol{\phi} \in \mathcal{L}_\tau. \tag{12}$$

It follows from the definition of the Galerkin approximation u_h that $\boldsymbol{q}_h = \boldsymbol{a}\nabla u_h$ is an approximate solution of the exact flux variable $\boldsymbol{q}$. In addition, it is not hard to derive the following error estimate:

$$\|\boldsymbol{q} - \boldsymbol{q}_h\| \le C\|u - u_h\|_1$$

for some constant C.

With the L^2-projection operator $\boldsymbol{Q}_\tau$, we can provide a new flux approximation given as follows:

$$\boldsymbol{q} = \boldsymbol{a}\nabla u \approx \boldsymbol{Q}_\tau \boldsymbol{q}_h. \tag{13}$$

From the definition of $\boldsymbol{Q}_\tau$, we see that the new flux approximation $\boldsymbol{Q}_\tau \boldsymbol{q}_h$ satisfies the following system of equations:

$$(\boldsymbol{a}^{-1}\boldsymbol{Q}_\tau \boldsymbol{q}_h, \boldsymbol{\phi}) = (\nabla u_h, \boldsymbol{\phi}) \quad \forall \boldsymbol{\phi} \in \mathcal{L}_\tau. \tag{14}$$

When the fitting space $\mathcal{L}_\tau$ consists of discontinuous piecewise polynomials of order k on each element of Ω_τ, our flux recovery method is closely related to the ZZ [32, 33] patch recovery technique. The difference lies on the selection of the fitting space $\mathcal{L}_\tau$ and the way that the projection was defined. The ZZ method uses a discrete version of the L^2-inner product, and the fitting space is based on a patch of elements from the original finite element partition Ω_h.

In practical computation, the recovery space contains polynomials of higher order than the original finite element space. In other words, the value of the parameter r is normally larger than k.

4. Error estimates. The objective of this section is to analyze the approximation formulas (13) and (11). The accuracy of the approximations is given in Theorem 4.1 for the case when the boundary data g is approximated by its L^2 projection in Λ_h. In case that g_h is the nodal point interpolation or other approximations satisfying (29), the corresponding superconvergence estimate is given in Theorem 4.2.

For simplicity of notation, we use the following element-wise Sobolev norms:

$$|||v|||_{m,h} = \left(\sum_{K \in \Omega_h} |v|^2_{H^m(K)} \right)^{1/2}, \qquad |||v|||_{m,\tau} = \left(\sum_{K \in \Omega_\tau} |v|^2_{H^m(K)} \right)^{1/2},$$

where

$$|v|_{H^m(K)} = \left(\sum_{K \in \Omega_h} \sum_{|\alpha|=m} \int_K |D^\alpha v|^2 dx \right)^{1/2}$$

is the seminorm of $v \in H^m(K)$. Here $\alpha = (\alpha_1, \ldots, \alpha_d), \alpha_i \geq 0$ is a multi-index and $D^\alpha v = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} v$. Thus, our function v above needs only to be in the Sobolev space $H^m(K)$ over each element K from Ω_h or Ω_τ in order to guarantee that the norms exist. If (6) and (7) are satisfied, then it is not hard to prove the following estimate:

$$\|\boldsymbol{v} - \boldsymbol{Q}_\tau \boldsymbol{v}\|_0 \leq C\tau^m |||\boldsymbol{v}|||_{m,\tau}, \quad 0 \leq m \leq r+1. \tag{15}$$

4.1. Projection space $\mathcal{L}_\tau$ of class $H(\mathrm{div};\Omega)$. Let us analyze the approximation schemes (11) and (13) by assuming that the projection space $\mathcal{L}_\tau$ is of class $H(\mathrm{div};\Omega)$. The following theorem is concerned with the case when the Dirichlet boundary data is approximated by L^2 projections.

THEOREM 4.1. *Let u be the exact solution of* (2) *and u_h be its finite element approximation given by* (4). *Let $\boldsymbol{q} = \boldsymbol{a}\nabla u$ be the flux/gradient with the obvious approximation $\boldsymbol{q}_h = \boldsymbol{a}\nabla u_h$ and let $\mathcal{G}_\tau$ be a postprocessing operator given by either $\boldsymbol{Q}_\tau$ or $\tilde{\boldsymbol{Q}}_\tau$ as in the previous section. Assume that the approximate boundary value g_h is taken to be the $L^2(\partial\Omega)$ projection of g and* (3) *and* (8) *are valid for an $s \in [1, k+1]$. Assume that the fitting (projection) space $\mathcal{L}_\tau$ is constructed so that $\mathcal{L}_\tau \subset H(\mathrm{div};\Omega)$. Then*

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C\tau^{r+1} |||\boldsymbol{q}|||_{r+1,\tau} + C(h\tau^{-1})^{s-1} \|u - u_h\|_1. \tag{16}$$

Proof. First, we provide a proof for $\mathcal{G}_\tau = \tilde{\boldsymbol{Q}}_\tau$. To this end, we observe that

$$\|\tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h - \boldsymbol{Q}_\tau \boldsymbol{q}\|_0 \leq C \sup_{\boldsymbol{\phi} \in \mathcal{L}_\tau, \|\boldsymbol{\phi}\|_{L^2}=1} (\boldsymbol{a}^{-1}\tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h - \boldsymbol{a}^{-1}\boldsymbol{Q}_\tau \boldsymbol{q}, \boldsymbol{\phi}). \tag{17}$$

For a given $\boldsymbol{\phi} \in \mathcal{L}_\tau$, it is true that

$$\begin{aligned}(\boldsymbol{a}^{-1}\tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h - \boldsymbol{a}^{-1}\boldsymbol{Q}_\tau \boldsymbol{q}, \boldsymbol{\phi}) &= (\boldsymbol{a}^{-1}\tilde{\boldsymbol{Q}}_\tau(\boldsymbol{a}\nabla u_h) - \boldsymbol{a}^{-1}\boldsymbol{Q}_\tau(\boldsymbol{a}\nabla u), \boldsymbol{\phi}) \\ &= -(u_h, \nabla \cdot \boldsymbol{\phi}) + \langle g, \boldsymbol{\phi} \cdot \boldsymbol{n} \rangle - (\nabla u, \boldsymbol{\phi}) = (u - u_h, \nabla \cdot \boldsymbol{\phi}).\end{aligned}$$

Define $w \in H_0^1(\Omega)$ to be the solution of

$$a(v, w) = (v, \nabla \cdot \boldsymbol{\phi}) \quad \forall v \in H_0^1(\Omega). \tag{18}$$

Applying the theory of distributions, it can be proved that

$$a(v, w) - \left\langle \frac{\partial w}{\partial \boldsymbol{n}_{\boldsymbol{a}}}, v \right\rangle = (\nabla \cdot \boldsymbol{\phi}, v) \quad \forall v \in H^1(\Omega).$$

Using (2), (4), (5), and the above equation, it is easy to see that, for any $v \in S_h^0$ and $\xi \in \Lambda_h$,

$$(u-u_h, \nabla\cdot\phi) = a(u-u_h, w) - \left\langle \frac{\partial w}{\partial \boldsymbol{n_a}}, u-u_h \right\rangle$$
$$= a(u-u_h, w-v) - \left\langle \frac{\partial w}{\partial \boldsymbol{n_a}} - \xi, g-g_h \right\rangle.$$

It follows that

$$|(u-u_h, \nabla\cdot\phi)| \le C\|u-u_h\|_1\|w-v\|_1 + \|\frac{\partial w}{\partial \boldsymbol{n_a}} - \xi\|_{-\frac{1}{2},\partial\Omega}\|u-u_h\|_{\frac{1}{2},\partial\Omega}.$$

Using the trace inequality

$$\|v\|_{\frac{1}{2},\partial\Omega} \le C\|v\|_1 \quad \forall v \in H^1(\Omega)$$

and the interpolation estimates

$$\inf_{v\in S_h} \|w-v\|_1 \le Ch^{s-1}\|w\|_s, \tag{19}$$

$$\inf_{\xi\in\Lambda_h} \left\| \frac{\partial w}{\partial \boldsymbol{n_a}} - \xi \right\|_{-\frac{1}{2},\partial\Omega} \le Ch^{s-1}\|w\|_s, \tag{20}$$

we obtain

$$|(u-u_h, \nabla\cdot\boldsymbol{\phi})| \le Ch^{s-1}\|u-u_h\|_1\|w\|_s \tag{21}$$

for an $s \in [1, k+1]$. Next, we use the H^s regularity assumption (3) to obtain

$$|(u-u_h, \nabla\cdot\boldsymbol{\phi})| \le Ch^{s-1}\|u-u_h\|_1\|\nabla\cdot\boldsymbol{\phi}\|_{s-2} \le Ch^{s-1}\tau^{1-s}\|u_h-u\|_1\|\boldsymbol{\phi}\|_0,$$

where we have also used the inverse property (6) in the last inequality. Collecting all the estimates we obtain

$$\|\tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h - \boldsymbol{Q}_\tau \boldsymbol{q}\|_0 \le Ch^{s-1}\tau^{1-s}\|u_h-u\|_1, \tag{22}$$

which, together with (15), gives the desired error estimate for $\mathcal{G}_\tau = \tilde{\boldsymbol{Q}}_\tau$.

To analyze the case $\mathcal{G}_\tau = \boldsymbol{Q}_\tau$, it suffices to estimate $\|\boldsymbol{Q}_\tau(\boldsymbol{q}-\boldsymbol{q}_h)\|_0$. Since

$$\|\boldsymbol{Q}_\tau(\boldsymbol{q}-\boldsymbol{q}_h)\|_0 \le C \sup_{\boldsymbol{\phi}\in\mathcal{L}_\tau, \|\boldsymbol{\phi}\|_0=1} (\boldsymbol{a}^{-1}\boldsymbol{Q}_\tau(\boldsymbol{q}-\boldsymbol{q}_h), \boldsymbol{\phi}) \tag{23}$$

and

$$(\boldsymbol{a}^{-1}\boldsymbol{Q}_\tau(\boldsymbol{q}-\boldsymbol{q}_h), \boldsymbol{\phi}) = (\nabla(u-u_h), \boldsymbol{\phi}), \tag{24}$$

then it is sufficient to estimate $|(\nabla(u-u_h), \boldsymbol{\phi})|$. Recall that, by assumption, we have $\mathcal{L}_\tau \subset H(\mathrm{div};\Omega)$. Thus, it follows from the integration by parts that

$$(\nabla(u_h-u), \boldsymbol{\phi}) = (u-u_h, \nabla\cdot\boldsymbol{\phi}) + \langle u_h-u, \boldsymbol{\phi}\cdot\boldsymbol{n}\rangle. \tag{25}$$

Let $w \in H_0^1(\Omega)$ be defined as the solution of the following problem:

$$a(v,w) = -(\nabla\cdot\boldsymbol{\phi}, v) \qquad \forall v \in H_0^1(\Omega). \tag{26}$$

It follows from (2), (4), and (5) that, for any $v \in S_h^0$ and $\xi \in \Lambda_h$,

$$
\begin{aligned}
&(\nabla(u_h - u), \boldsymbol{\phi}) = (u - u_h, \nabla \cdot \boldsymbol{\phi}) + \langle u_h - u, \boldsymbol{\phi} \cdot \boldsymbol{n} \rangle \\
&= a(u - u_h, w) + \left\langle \tfrac{\partial w}{\partial \boldsymbol{n_a}}, u_h - u \right\rangle + \langle \boldsymbol{\phi} \cdot \boldsymbol{n}, u_h - u \rangle \\
&= a(u - u_h, w - v) + \left\langle \tfrac{\partial w}{\partial \boldsymbol{n_a}} + \boldsymbol{\phi} \cdot \boldsymbol{n} - \xi, u_h - u \right\rangle .
\end{aligned} \tag{27}
$$

Using the standard approximation property of Λ_h and the trace inequality in Sobolev spaces as in (20)–(21), we obtain

$$
\inf_{\xi \in S_h} \left\| \frac{\partial w}{\partial \boldsymbol{n_a}} + \boldsymbol{\phi} \cdot \boldsymbol{n} - \xi \right\|_{-\frac{1}{2}, \partial\Omega} \leq C h^{s-1} \tau^{1-s} \|\boldsymbol{\phi}\|_0 .
$$

It is also not hard to see that

$$
\inf_{v \in S_h^0} |a(u - u_h, w - v)| \leq C h^{s-1} \tau^{1-s} \|u - u_h\|_1 \|\boldsymbol{\phi}\|_0 .
$$

Substituting the above two estimates into (27), we obtain

$$
|(\nabla(u_h - u), \boldsymbol{\phi})| \leq C h^{s-1} \tau^{1-s} \|u - u_h\|_1 \|\boldsymbol{\phi}\|_0 ,
$$

which implies that

$$
\|\boldsymbol{Q}_\tau(\boldsymbol{q} - \boldsymbol{q}_h)\|_0 \leq C h^{s-1} \tau^{1-s} \|u - u_h\|_1 . \tag{28}
$$

This completes the proof of the theorem. □

If the exact solution is sufficiently smooth, then we have from the estimate (16) that

$$
\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C(u, \boldsymbol{q}) \left(\tau^{r+1} + \tau^{1-s} h^{k+s-1} \right) .
$$

Assume that the model problem has the H^{k+1} regularity (i.e., $s = k + 1$). Thus,

$$
\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C(u, \boldsymbol{q}) \left(\tau^{r+1} + \tau^{-k} h^{2k} \right) .
$$

By choosing $\tau = h^\alpha$, the above estimate becomes

$$
\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C(u, \boldsymbol{q}) \left(h^{\alpha(r+1)} + h^{(2-\alpha)k} \right) ,
$$

which is optimized when

$$
\alpha(r+1) = (2 - \alpha)k \Longleftrightarrow \alpha = \frac{2k}{r + k + 1} .
$$

The corresponding error estimate is given by

$$
\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C(u, \boldsymbol{q}) h^{\frac{2k(r+1)}{r+k+1}} .
$$

With $k = 3$ and $r = 3$, the above estimate implies an accuracy of order $h^{\frac{8}{3}}$ which is much better than the optimal order h^2.

In practical computation, the Dirichlet boundary data is often approximated by a scheme different from the L^2 projection. Thus, the estimate in Theorem 4.1 is

no longer valid for such problems. Our next goal of this section is to derive some superconvergence for general approximation schemes of the Dirichlet boundary data.

THEOREM 4.2. *Assume that* (3) *and* (8) *are valid for an* $s \in [3/2, k+1]$ *and that* g_h *is an approximation of the Dirichlet data* g *on the boundary such that*

$$\|g - g_h\|_{0,\partial\Omega} \le Ch^{k+1}\|g\|_{k+1,\partial\Omega}. \tag{29}$$

Assume that the projection space $\mathcal{L}_\tau \subset H(\mathrm{div};\Omega)$. *Then there exists a constant* C *such that*

$$\|\boldsymbol{q} - \tilde{\boldsymbol{Q}}_\tau \boldsymbol{q}_h\|_0 \le C\tau^{r+1}|\!|\!|\boldsymbol{q}|\!|\!|_{r+1,\tau} + Ch^{s-1}\tau^{1-s}\|u - u_h\|_1 + Ch^{k+1}\tau^{-\frac{1}{2}}\|g\|_{k+1,\partial\Omega}.$$

Proof. The proof is similar to that of Theorem 4.1. The only modification is on the treatment of $(u - u_h, \nabla\cdot\boldsymbol{\phi})$. To this end, we observe that

$$(u - u_h, \nabla\cdot\boldsymbol{\phi}) = a(u - u_h, w - v) - \left\langle \frac{\partial w}{\partial \boldsymbol{n_a}}, g - g_h \right\rangle.$$

Thus,

$$\begin{aligned}
|(u - u_h, \nabla\cdot\boldsymbol{\phi})| &\le Ch^{s-1}\|u - u_h\|_1\|w\|_s + \left\|\frac{\partial w}{\partial \boldsymbol{n_a}}\right\|_{0,\partial\Omega}\|g - g_h\|_{0,\partial\Omega} \\
&\le Ch^{s-1}\|u - u_h\|_1\|\nabla\cdot\boldsymbol{\phi}\|_{s-2} + Ch^{k+1}\|w\|_{\frac{3}{2}}\|g\|_{k+1,\partial\Omega} \\
&\le Ch^{s-1}\tau^{1-s}\|u_h - u\|_1\|\boldsymbol{\phi}\|_0 + Ch^{k+1}\tau^{-1/2}\|g\|_{k+1,\partial\Omega}\|\boldsymbol{\phi}\|_0.
\end{aligned}$$

The rest of the proof is similar to that of Theorem 4.1 and is omitted. ∎

Theorem 4.2 shows that if the Dirichlet boundary data is not approximated by the L^2 projection, then the superconvergence estimate for the recovered flux/gradient approximation will suffer. In fact, our estimate of Theorem 4.2 ensures only a superconvergence of order $O(h^{k+1})$ for sufficiently smooth solution u and the projection space $\mathcal{L}_\tau$.

4.2. Discontinuous projection space $\mathcal{L}_\tau$. The flux approximation scheme (13) or (14) is well defined for discontinuous projection space $\mathcal{L}_\tau$. When discontinuous finite elements are employed in the projection method, the computation of the recovered flux/gradient can be implemented locally on each element $K \in \Omega_\tau$, which results in a great saving of computer time and efficiency. However, due to the use of integration by parts in (25), the superconvergence established in Theorems 4.1 and 4.2 is no longer applicable to discontinuous projection space. Our objective of this section is to provide a superconvergent theory for the approximation scheme (13) when $\mathcal{L}_\tau$ contains discontinuous finite element functions.

Let K be any element from the partition Ω_τ. It is not hard to show that there exists a constant C independent of K and v such that

$$\int_{\partial K} v^2 ds \le C\left((\tau^{-1} + \epsilon^{-1})\int_K v^2 dx + \epsilon\int_K |\nabla v|^2 dx\right), \tag{30}$$

where $\epsilon > 0$ is any real number.

THEOREM 4.3. *Let* u *be the exact solution of* (2) *and let* u_h *be its finite element approximation given by* (4). *Let* $\boldsymbol{q} = \boldsymbol{a}\nabla u$ *be the flux/gradient with the obvious approximation* $\boldsymbol{q}_h = \boldsymbol{a}\nabla u_h$ *and let* $\mathcal{G}_\tau$ *be a postprocessing operator given by* $\boldsymbol{Q}_\tau$. *Then,*

for any projection space $\mathcal{L}_\tau$ which is a piecewise polynomial of order r, we have for any $\epsilon > 0$

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \le C\tau^{r+1} |\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau} + C\tau^{-\frac{1}{2}}((\tau^{-\frac{1}{2}} + \epsilon^{-\frac{1}{2}})\|u - u_h\|_0 + \epsilon^{\frac{1}{2}}\|u - u_h\|_1). \tag{31}$$

Proof. Since $\mathcal{G}_\tau = \boldsymbol{Q}_\tau$, then

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \le \|\boldsymbol{q} - \boldsymbol{Q}_\tau \boldsymbol{q}\|_0 + \|\boldsymbol{Q}_\tau \boldsymbol{q} - \boldsymbol{Q}_\tau \boldsymbol{q}_h\|_0.$$

The error $\|\boldsymbol{q} - \boldsymbol{Q}_\tau \boldsymbol{q}\|_0$ can be estimated by using (15). To estimate $\|\boldsymbol{Q}_\tau(\boldsymbol{q} - \boldsymbol{q}_h)\|_0$, we see from (23) and (24) that it suffices to deal with $(\nabla(u - u_h), \boldsymbol{\phi})$ for any $\boldsymbol{\phi} \in \mathcal{L}_\tau$ such that $\|\boldsymbol{\phi}\| = 1$. To this end, using the integration by parts we obtain

$$(\nabla(u - u_h), \boldsymbol{\phi}) = \sum_{K \in \Omega_\tau} \int_K (u_h - u) \nabla \cdot \boldsymbol{\phi} dx + \sum_{K \in \Omega_\tau} \int_{\partial K} (u - u_h) \boldsymbol{\phi} \cdot \boldsymbol{n}_K ds. \tag{32}$$

The first term on the right-hand side of (32) can be bounded as follows:

$$\left| \sum_{K \in \Omega_\tau} \int_K (u_h - u) \nabla \cdot \boldsymbol{\phi} dx \right| \le \|u - u_h\|_0 \, \|\nabla \cdot \boldsymbol{\phi}\|_0 \le C\tau^{-1} \|u - u_h\|_0, \tag{33}$$

where we have used the standard inverse estimate for $\|\nabla \cdot \boldsymbol{\phi}\|$. To estimate the second term on the right-hand side of (32), we use the Schwarz inequality to obtain

$$\begin{aligned} \left| \sum_{K \in \Omega_\tau} \int_{\partial K} (u - u_h) \boldsymbol{\phi} \cdot \boldsymbol{n}_K ds \right| &\le \sum_{K \in \Omega_\tau} \int_{\partial K} |u - u_h| |\boldsymbol{\phi} \cdot \boldsymbol{n}_K| ds \\ &\le \sum_{K \in \Omega_\tau} \|u - u_h\|_{0,\partial K} \|\boldsymbol{\phi}\|_{0,\partial K} \\ &\le \left(\sum_{K \in \Omega_\tau} \|u - u_h\|^2_{0,\partial K} \right)^{\frac{1}{2}} \left(\sum_{K \in \Omega_\tau} \|\boldsymbol{\phi}\|^2_{0,\partial K} \right)^{\frac{1}{2}}. \end{aligned} \tag{34}$$

It follows from (30) that

$$\|u - u_h\|^2_{0,\partial K} \le C \left((\tau^{-1} + \epsilon^{-1}) \|u - u_h\|^2_{0,K} + \epsilon \|\nabla(u - u_h)\|^2_{0,K} \right).$$

Similarly, from (30) with $\epsilon = \tau$ we have

$$\|\boldsymbol{\phi}\|^2_{0,\partial K} \le C(\tau^{-1} \|\boldsymbol{\phi}\|^2_{0,K} + \tau \|\nabla \boldsymbol{\phi}\|^2_{0,K}).$$

Substituting the above two estimates into (34) yields

$$\begin{aligned} &\left| \sum_{K \in \Omega_\tau} \int_{\partial K} (u - u_h) \boldsymbol{\phi} \cdot \boldsymbol{n}_K ds \right| \\ &\le C \left((\tau^{-1} + \epsilon^{-1}) \|u - u_h\|^2_0 + \epsilon \|\nabla(u - u_h)\|^2_0 \right)^{\frac{1}{2}} \left(\tau^{-1} \|\boldsymbol{\phi}\|^2_0 + \tau \|\nabla \boldsymbol{\phi}\|^2_0 \right)^{\frac{1}{2}}. \end{aligned}$$

Now using the standard inverse inequality and the fact that $\|\boldsymbol{\phi}\|_0 = 1$, we obtain

$$\tau^{-1} \|\boldsymbol{\phi}\|^2_0 + \tau \|\nabla \boldsymbol{\phi}\|^2_0 \le C\tau^{-1}.$$

Thus,

$$\left|\sum_{K\in\Omega_\tau}\int_{\partial K}(u-u_h)\boldsymbol{\phi}\cdot\boldsymbol{n}_K ds\right| \tag{35}$$
$$\le C\tau^{-\frac{1}{2}}((\tau^{-\frac{1}{2}}+\epsilon^{-\frac{1}{2}})\|u-u_h\|_0+\epsilon^{\frac{1}{2}}\|\nabla(u-u_h)\|_0).$$

The combination of (32) with (33) and (35) gives

$$|(\nabla(u-u_h),\boldsymbol{\phi})|\le C\tau^{-\frac{1}{2}}((\tau^{-\frac{1}{2}}+\epsilon^{-\frac{1}{2}})\|u-u_h\|_0+\epsilon^{\frac{1}{2}}\|\nabla(u-u_h)\|_0),$$

which completes the proof. ∎

The discontinuous projection space $\mathcal{L}_\tau$ has many distinguished features in theory and application. In practical implementation, it allows a local and parallel computation of the projected flux $\boldsymbol{Q}_\tau\boldsymbol{q}_h$. In addition, one does not need to worry about any special treatment of the boundary condition $u=g$. From the analysis of Theorem 4.3, we see that the estimate (31) does not require the regularity/smoothness assumptions (3) and (8). However, in order to get a superconvergence from the estimate (31), the L^2 norm of the error must have a higher order of convergence than the H^1 norm. This is often accomplished via a duality argument which requires a certain regularity for the dual problem.

For illustration, we consider a model problem where the flux $\boldsymbol{q}$ and the solution u satisfy

$$C(u)=|\!|\!|u|\!|\!|_{k+1,h}<\infty,\quad C(\boldsymbol{q})=|\!|\!|\boldsymbol{q}|\!|\!|_{r+1,\tau}<\infty. \tag{36}$$

Assume that the H^2 regularity is satisfied for the dual problem. Then the following error estimate is well known:

$$\|u-u_h\|_0+h\|\nabla(u-u_h)\|_0\le C(u)h^{k+1}.$$

Substituting the above with $\tau=h^\alpha$ and $\epsilon=h$ into (31) yields

$$\|\boldsymbol{q}-\mathcal{G}_\tau\boldsymbol{q}_h\|_0\le C(u,\boldsymbol{q})(h^{\alpha(r+1)}+h^{k+0.5-0.5\alpha}).$$

The above estimate is optimized when

$$\alpha(r+1)=k+0.5-0.5\alpha\Longleftrightarrow\alpha=\frac{k+0.5}{r+1.5},$$

which gives the following error estimate:

$$\|\boldsymbol{q}-\mathcal{G}_\tau\boldsymbol{q}_h\|_0\le C(u,\boldsymbol{q})h^{\frac{(k+0.5)(r+1)}{r+1.5}}. \tag{37}$$

With $k=2$ and $r=3$ we obtain

$$\|\boldsymbol{q}-\mathcal{G}_\tau\boldsymbol{q}_h\|_0\le C(u,\boldsymbol{q})h^{\frac{20}{9}},$$

which is much better than the optimal order error estimate $\mathcal{O}(h^2)$ for the straightforward gradient approximation $\boldsymbol{q}_h=\boldsymbol{a}\nabla u_h$.

For problems with reentrant corners in the domain or discontinuous data in the coefficient tensor $\{a_{ij}\}$, the H^2 regularity for the dual problem is not satisfied. The dual problem, however, has the $H^{1+\sigma}(\Omega)$ regularity for some $\sigma\in(0,1)$. For sufficiently smooth solution u and the flux $\boldsymbol{q}$, it is possible to show that

$$\|u-u_h\|_0+h^\sigma\|\nabla(u-u_h)\|_0\le C(u)h^{k+\sigma}.$$

By choosing ϵ properly in Theorem 4.3, one is able to determine a value of α in $\tau=h^\alpha$ which gives a superconvergence for the gradient. Details of this analysis are omitted.

5. A relation with ZZ patch recovery. For a given element K of S_h, the ZZ method projects ∇u_h to a covering patch $\tilde{K}$ consisting of K and several neighboring elements [32, 33]. The projection space in the ZZ method is the restriction of S_h on each patch $\tilde{K}$. The superconvergence of the original ZZ patch recovery has been observed only through numerical experiments with specially defined discrete L^2-inner products.

We shall modify the ZZ patch recovery method as follows. First, we replace the ZZ projection space by the space of polynomials of order $r \geq 0$ on each patch. Second, we assume that each patch is of size τ which is larger than the original mesh size h. By adjusting the size τ and the fitting polynomial order r, we are able to obtain a superconvergence for the MZZ patch recovery method.

We now present a detailed discussion on the MZZ method. Based on the finite element partition for S_h, we shall first divide the mesh domain Ω into many nonoverlapping and simply connected subdomains Ω_i. Each subdomain is a union of finite elements of Ω_h. Assume that the partition $\{\Omega_i\}$ is regular in the sense that each Ω_i is of diameter proportional to τ and contains a ball of diameter also proportional to τ. The finite element space $\mathcal{L}_\tau$ is defined as

$$\mathcal{L}_\tau = \{\boldsymbol{v} = (v_1, \ldots, v_d) : \quad v_i|_{\Omega_i} \in P_r \ \forall i\}.$$

In other words, for $\boldsymbol{v} \in \mathcal{L}_\tau$, each component of $\boldsymbol{v}$ in Ω_i is the restriction of a polynomial of order r. Notice that $\boldsymbol{v}$ can be discontinuous on the interface between the subdomains. Here are some points of why discontinuous fitting functions are preferable:

- Each subdomain (element) Ω_i is constructed from the elements of Ω_h by regrouping. Thus, the implementation of the projection operator $\boldsymbol{Q}_\tau$ is computationally feasible, since the corresponding numerical integration for the matrix problem of $\boldsymbol{Q}_\tau$ is easy to compute.
- As the functions $\boldsymbol{v}$ can be totally discontinuous on the interfaces, the boundary of the element Ω_i does not need to be straight lines.
- The projection operator $\boldsymbol{Q}_\tau$ can be computed on each subdomain in parallel, and the projections over the subdomains do not interact with each other.

An error estimate can be established for the MZZ scheme by using Theorem 4.3. In fact, from the estimate (31) with $\epsilon = h$ we have

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C\tau^{r+1} |\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau} + \tau^{-\frac{1}{2}} (h^{-\frac{1}{2}} \|u - u_h\|_0 + h^{\frac{1}{2}} \|u - u_h\|_1),$$

where $\mathcal{G}_\tau$ is given by $\boldsymbol{Q}_\tau$. If the exact solution u is sufficiently smooth, then

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C(\tau^{r+1} |\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau} + \tau^{-\frac{1}{2}} h^{k+0.5} |\!|\!| u |\!|\!|_{k+1,h}). \tag{38}$$

For a given τ, we can choose the order of polynomials of $\mathcal{L}_\tau$ properly such that $\tau^{r+1} \leq \sqrt{\frac{h}{\tau}}\, h^k$. Hence, the gain on the convergence for the flux/gradient is of a factor $\sqrt{\frac{h}{\tau}}$.

In case that u has only a limited regularity, we need to choose the mesh size τ properly to obtain the best possible superconvergence. For example, if $\boldsymbol{q} \in [H^2(\Omega)]^d$ and $u \in H^2(\Omega)$ and S_h contains continuous piecewise linear functions, we take the projection space $\mathcal{L}_\tau$ to be the restriction of linear functions on each subdomain (or patch) Ω_i. For any $\boldsymbol{v} \in \mathcal{L}_\tau$, we have

$$\boldsymbol{v}|_{\Omega_i} = a_0^i + \sum_{j=1}^{d} a_j^i x_j.$$

In computing the projection of the flux over each Ω_i, we need only to compute the coefficients $a_0^i, a_1^i, \ldots, a_d^i$ on each Ω_i by using the standard least-squares method. As $k = r = 1$, we have that

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C\tau^2 \|\!|\boldsymbol{q}|\!\|_{2,\tau} + C(h/\tau)^{0.5} h \|\!|u|\!\|_{2,h}.$$

By choosing $\tau = h^{\frac{3}{5}}$, we arrive at the following superconvergence:

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq Ch^{\frac{6}{5}} (\|\!|\boldsymbol{q}|\!\|_{2,\tau} + \|\!|u|\!\|_{2,h}).$$

The gain for the convergence order for the flux or gradient is then 1/5. If we use quadratic functions for $\mathcal{L}_\tau$, then we need to take $\tau = h^{\frac{3}{7}}$, and the gain for the convergence order is 2/7. In case that $r = 3$, the gain of the convergence order can be 1/3.

For the original ZZ method, it is typical that $r = k$ and $\tau = Lh$ for some fixed value $L \geq 1$. Correspondingly, our estimate implies that

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq CL^{k+1} h^{k+1} \|\!|\boldsymbol{q}|\!\|_{k+1,\tau} + C\sqrt{L^{-1}}\, h^k \|\!|u|\!\|_{k+1,h},$$

which does not claim any superconvergence for the recovered flux or gradient approximation.

6. A remark on continuous least-squares surface fitting. Locality and parallelization are the main features in using discontinuous finite elements to fit the approximate flux. However, as indicated by (38), the maximum gain on the order of convergence with discontinuous projection space is $h^{\frac{1}{2}}$ over the optimal order error estimate. In fact, our convergence analysis in previous sections suggests that continuous finite element fitting spaces should be used in order to achieve a high order of superconvergence for the recovered flux/gradient approximation.

Let $\mathcal{L}_\tau$ be a finite element space of class C^0 consisting of continuous piecewise polynomials of order r over each element. Recall that the finite element partition for Ω_h does not need to be a refinement of the elements of Ω_τ. It is well known that, for any $\epsilon \in (0, \frac{1}{2})$, $\mathcal{L}_\tau \subset H^{\frac{3}{2}-\epsilon}(\Omega)$. In other words, assumption (8) is satisfied with $s = 2.5 - \epsilon$. Assume that (3) is also valid with $s = 2.5 - \epsilon$. An application of Theorem 4.1 shows that the convergence for the recovered flux approximations is given by

$$\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \leq C\tau^{r+1} \|\!|\boldsymbol{q}|\!\|_{r+1,\tau} + C(h/\tau)^{1.5-\epsilon} \|u - u_h\|_1,$$

where $\mathcal{G}_\tau$ is either $\tilde{\boldsymbol{Q}}_\tau$ or $\boldsymbol{Q}_\tau$. If u is sufficiently smooth, we can choose the order of polynomials of $\mathcal{L}_\tau$ properly such that $\tau^{r+1} \leq C(h/\tau)^{1.5-\epsilon}$. In such a case, the gain of the convergence for the flux and gradient is of a factor $(h/\tau)^{1.5-\epsilon}$. For simplicity of discussion, we shall assume $\epsilon = 0$ in the rest of this section. In case that u has only a limited regularity, we need to choose r according to the regularity of u and choose τ such that $O(\tau^{r+1}) = (h/\tau)^{1.5}\, h^k$. In Table 1, we show some theoretical gain of the convergence order for the flux/gradient with different values of r and k. In theory, the computational result can only be better than this.

In a similar manner, the improvement on the convergence of the flux/gradient would be of a factor $O(h/\tau)^{(\ell+1.5)}$ if $\mathcal{L}_\tau$ is a finite element space of class C^ℓ for $0 \leq \ell \leq k - 1.5$. In case that u has only a limited regularity, we need to choose the mesh size τ properly to get the best possible superconvergence.

TABLE 1
The value of β in $\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}_h\|_0 \le Ch^{k+\beta}$.

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
$k=1$	0.4286	0.6667	0.8182	0.9231	1.0000
$k=2$		0.3333	0.5455	0.6923	0.8000
$k=3$			0.2727	0.4615	0.6000
$k=4$				0.2308	0.4000

7. An application to mesh adaptivity. The superconvergence estimates can be used to refine the finite element mesh adaptively. Let us note that

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \le \|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}\|_0 + \|\mathcal{G}_\tau(\boldsymbol{q} - \boldsymbol{q}_h)\|_0 + \|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_0.$$

Assume that the finite element solution u_h is nontrivial in the sense that

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \cong h^k |\!|\!| \boldsymbol{q} |\!|\!|_{k,h}. \tag{39}$$

From (15), (22), and (28), and assuming that $\boldsymbol{q}$ has the needed regularity, we obtain

$$\begin{aligned}
&\|\boldsymbol{q} - \mathcal{G}_\tau \boldsymbol{q}\|_0 + \|\mathcal{G}_\tau(\boldsymbol{q} - \boldsymbol{q}_h)\|_0 \\
&\le C\tau^{r+1} |\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau} + \|\mathcal{G}_\tau(\boldsymbol{q} - \boldsymbol{q}_h)\|_0 \\
&\le C\frac{\tau^{r+1}}{h^k} \frac{|\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau}}{|\!|\!| \boldsymbol{q} |\!|\!|_{k,h}} \|\boldsymbol{q} - \boldsymbol{q}_h\|_0 + C\left(\frac{h}{\tau}\right)^{s-1} \|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \\
&= \alpha \|\boldsymbol{q} - \boldsymbol{q}_h\|_0,
\end{aligned}$$

where

$$\alpha = C\frac{\tau^{r+1}}{h^k} \frac{|\!|\!| \boldsymbol{q} |\!|\!|_{r+1,\tau}}{|\!|\!| \boldsymbol{q} |\!|\!|_{k,h}} + C\left(\frac{h}{\tau}\right)^{s-1}.$$

The mesh parameters τ and r can be chosen properly to ensure that $\alpha \to 0$ when $\tau \to 0, h \to 0$. For simplicity, let us take $\tau = \kappa h$. Thus,

$$\alpha = C(\boldsymbol{q})\kappa^{r+1} h^{r+1-k} + C\kappa^{-(s-1)}.$$

By letting $\kappa \to \infty$ and $\kappa \le o(h^{k/(r+1)-1})$, we have $\alpha \to 0$. In fact, the choices we have discussed for k, r, h, τ in sections 5 and 6 will all guarantee that $\alpha \to 0$. Thus,

$$(1 - \alpha)\|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \le \|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_0.$$

We emphasize that the right-hand side of the above estimate is computable. In practical computations, the value of α is small but not known exactly. We can produce a mesh to guarantee that

$$\|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_0 \le \varepsilon, \tag{40}$$

where ε stands for a prescribed tolerance. To this end, for a given coarse mesh $\mathcal{L}_\tau$, we compute the maximum value of the error indicator over all the coarse mesh elements:

$$\eta_\tau = \max_{K \in \Omega_\tau} \|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_{0,K}.$$

We choose a parameter $\theta \in (0,1)$. For a given coarse mesh element $K \in \Omega_\tau$, we refine K if

$$\|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_{0,K} \geq \theta \eta_\tau.$$

The refinement process is stopped if either (40) is satisfied or the memory limit has been reached, or the change of the computed solution in the energy norm is less than a given tolerance.

Under assumption (39), the error indicator $\|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_0$ is in fact equivalent to the true error due to the fact that

$$\begin{aligned}\|\mathcal{G}_\tau \boldsymbol{q}_h - \boldsymbol{q}_h\|_0 &\leq \|\mathcal{G}_\tau(\boldsymbol{q}_h - \boldsymbol{q})\|_0 + \|\mathcal{G}_\tau \boldsymbol{q} - \boldsymbol{q}\|_0 + \|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \\ &\leq (1+\alpha)\|\boldsymbol{q} - \boldsymbol{q}_h\|_0.\end{aligned}$$

Note that we refine the coarse mesh Ω_τ instead of the fine mesh Ω_h. The fine mesh Ω_h is always produced from Ω_τ by refining each coarse mesh element into several smaller elements. See [5] for some results about using averaging-type error estimators and the ZZ method for mesh refinement for general unstructured meshes.

8. Numerical experiments. Two meshes Ω_τ and Ω_h are needed in the computation. The coarse mesh Ω_τ is produced by the adaptive strategy of section 7. The fine mesh Ω_h is always produced from Ω_τ. To produce Ω_h, each coarse mesh element is refined into 4 elements by connecting the edge middle points or refined uniformly twice to produce 16 elements for two-dimensional problems. Continuous piecewise linear finite element functions over Ω_τ and Ω_h are used for the projection $\mathcal{G}_\tau$ and for the solution of the finite element approximation u_h, respectively.

The proposed algorithms are tested for

$$-\nabla \cdot (a \nabla u) = f \quad \text{on} \quad \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega \tag{41}$$

with $\Omega = (0,1) \times (0,1), f = 2\pi^2 \sin(\pi x)\sin(\pi y), g = 0, a = 1$. The exact solution is easily seen to be $u = \sin(\pi x)\sin(\pi y)$.

The global coarse mesh recovery is as described earlier, and the element-wise coarse mesh recovery works by projecting the gradient in S_h to the fitting space $\mathcal{L}_\tau$ consisting of piecewise linear functions over Ω_τ. Equation (14) is thus solved for each element in Ω_τ independently, and these local solutions are combined to a global solution by an averaging procedure. As we shall see, this gives a worse convergence rate than the global projection, but it is still superconvergent.

We use Figures 1 and 2 to show the computational results. In the plot, the x-axis represents the degree of freedom of the mesh. The y-axis represents the L^2 error of the gradient. Note that both axes are scaled using $\log_{10}$. Figure 2 compares the mesh quality produced by the superconvergence error estimator and the error estimator of Johnson [16] and Eriksson and Johnson [12]. The error for the finite element solution over S_h, i.e., the mesh produced by the superconvergence error estimator, is slightly better than the error for the finite element solution for the mesh produced by the error estimator of [16, 12]. To reach the same accuracy, we need a much smaller degree of freedom in our new method; see also Figure 1. The projected gradient over Ω_τ has a better convergence order, as can be seen from Figure 1. The convergence rate for the different errors are plotted in Figure 1. The convergence rate and the accuracy of the MZZ and the ZZ methods are nearly the same.

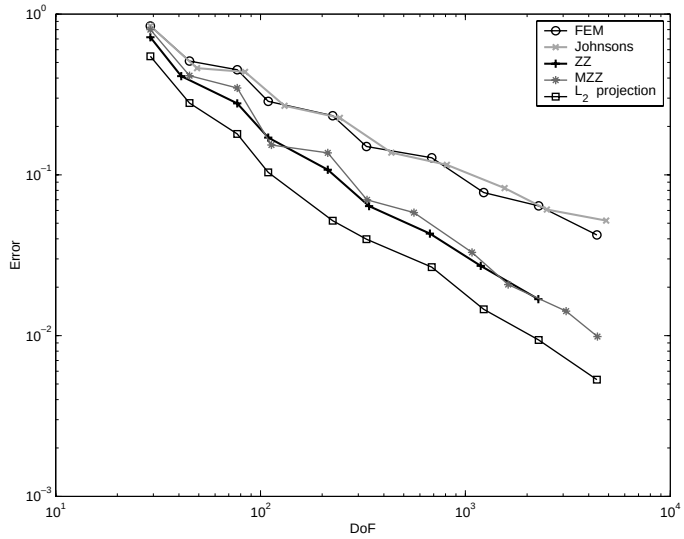


FIG. 1. *Comparison of the mesh quality produced by different recovery methods compared to the error estimator of Johnson* [16] *and Eriksson and Johnson* [12]. *All the errors are measured in* L^2 *for the gradients. FEM refers to the error* $\|\nabla u_h - \nabla u\|_0$, *where* u_h *is the finite element solution over* S_h.

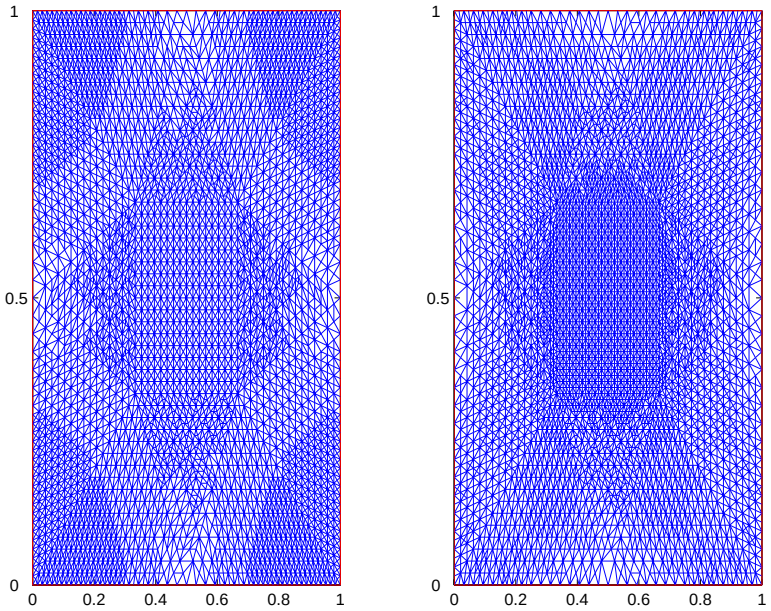


FIG. 2. *The left is the mesh produced by the superconvergence estimator of this paper, and the right is the mesh of the residual error estimator of Johnson* [16] *and Eriksson and Johnson* [12]. *The* L^2 *error for the left mesh is* 0.0094 *with* 2279 *nodes, and the right mesh has an* L^2 *error of* 0.061 *with* 2497 *nodes.*

Our computational experiment reveals that the convergence rate of standard Galerkin finite element solution over S_h is 1.2. For ZZ recovery the order of convergence is 1.72. The MZZ method proposed earlier has a convergence order of 1.74.

Johnson's method gives a rate of 1.08. Finally, the L^2 projection has a convergence of order 1.84. We clearly see that the L^2 projection has superior performance. The convergence order is evaluated by using the formula $\log_{10} \|\nabla u_h - \nabla u\|_0 / \log_{10}(\sqrt{DoF})$ as in [6, 22], where DoF stands for the total number of nodal points in the finite element partition. The continuous least-squares surface fitting is easy to implement and has the best accuracy.

REFERENCES

[1] J. H. Bramble and A. H. Schatz, *Higher order local accuracy by averaging in the finite element method*, Math. Comp., 31 (1977), pp. 94–111.

[2] F. Brezzi, J. Douglas, R. Durán, and L. Marini, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 52 (1987), pp. 237–250.

[3] F. Brezzi, J. Douglas, M. Fortin, and L. Marini, *Efficient rectangular mixed finite elements in two and three spaces variables*, RAIRO Modél Math. Anal. Numér., 21 (1987), pp. 581–604.

[4] F. Brezzi, J. Douglas, and L. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[5] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids Part* I: *Low order conforming, nonconforming and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[6] Z. Chen and S. Dai, *On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients*, SIAM J. Sci. Comput., to appear.

[7] C. Chen and Y. Huang, *High Accuracy Theory for Finite Element Methods*, Hunan Science and Technology Publishing House, Hunan, People's Republic of China, 1995 (in Chinese).

[8] J. Douglas and T. Dupont, *Superconvergence for Galerkin methods for the two-point boundary problem via local projections*, Numer. Math, 21 (1973), pp. 270–278.

[9] J. Douglas and J. Roberts, *Global estimate for mixed finite elements methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[10] J. Douglas and J. Wang, *Superconvergence of mixed finite element spaces on rectangular domains*, Calcolo, 26 (1989), pp. 121–134.

[11] J. Douglas and J. Wang, *A new family of spaces in mixed finite element methods for rectangular elements*, Comput. Appl. Math., 12 (1993), pp. 183–197.

[12] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* I: *A linear model problem*, SIAM J. Numer. Anal., 28, (1991), pp. 43–77.

[13] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[14] P. Grisvard, *Elliptic Problems on Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, MA, 1985.

[15] W. Hoffmann, A. H. Schatz, L. B. Wahlbin, and G. Wittum, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes. Part* 1*: A smooth problem and globally quasi-uniform meshes*, Math. Comp., 70 (2001), pp. 897–909.

[16] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Studentlitteratur, Lund, Sweden, Cambridge University Press, Cambridge, UK, 1987.

[17] M. Krizek and P. Neittaanmaki, *Superconvergence phenomenon in the finite element method arising from averaging gradients*, Numer. Math., 45 (1984), pp. 105–116.

[18] R. Lazarov, A. B. Andreev, and M. Hatri, *Superconvergence of the gradients in the finite element method for some elliptic and parabolic problems,* in Variational-Difference Methods in Mathematical Physics, Part II, Moscow, 1984, pp. 13–25.

[19] Q. Lin, *An integral identity and interpolated postprocess in superconvergence*, Research Report 90-7, Institute of Systems Science, Academia Sinica, Beijing, China, 1990, pp. 1–6.

[20] Q. Lin, *Global error expansion and superconvergence for higher order interpolation of finite elements*, J. Comput. Math., supplementary issue, 1992, pp. 286–289.

[21] Q. Lin and Q. Zhu, *The Preprocessing and Postprocessing for the Finite Element Method*, Hunan Scientific & Technical Publishers, Hunan, People's Republic of China, 1994.

[22] P. Morin, R. H. Nochetto, and K. Siebert, *D*ata oscillation and convergence of adaptive mesh, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[23] L. A. Oganesyan and L. A. Rukhovetz, *Study of the rate of convergence of variational difference scheme for second order elliptic equations in two-dimensional field with a smooth boundary,* U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 158–183.

[24] P. Raviart and J. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in the Mathematics Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977, pp. 292–315.

[25] A. H. Schatz, I. H. Sloan, and L. B. Wahlbin, *Superconvergence in finite element methods and meshes that are locally symmetric with respect to a point*, SIAM J. Numer. Anal., 33 (1996), pp. 505–521.

[26] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, Berlin, 1995.

[27] J. Wang, *A superconvergence analysis for finite element solutions by the least-squares surface fitting on irregular meshes for smooth problems*, J. Math. Study, 33 (2000), pp. 229–243.

[28] J. Wang, *Superconvergence and extrapolation for mixed finite element methods on rectangular domains*, Math. Comp., 56 (1991), pp. 477–503.

[29] J. Wang and X. Ye, *Superconvergence of finite element approximations for the Stokes problem by projection methods*, SIAM J. Numer. Anal., 39 (2001), pp. 1001–1013.

[30] Z. M. Zhang, Jr., *Ultraconvergence of the patch recovery techniques* II, Math. Comp., 69 (1999), pp. 141–158.

[31] Z. M. Zhang and H. D. Victory, Jr., *Mathematical analysis of Zienkiewicz-Zhu's derivative patch recovery technique*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 507–524.

[32] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates,* Part 1*,* Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

[33] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates,* Part 2, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

[34] Z. M. Zhang and J. Z. Zhu, *Superconvergence of the derivative path recovery technique and a posteriori error estimation*, in The Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations, Springer, New York, 1995, pp. 431–450.

[35] M. Zlámal, *Superconvergence and reduced integration in the finite element method*, Math. Comp., 32 (1978), pp. 663–685.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1281–1310

A WIGNER-MEASURE ANALYSIS OF THE DUFORT–FRANKEL SCHEME FOR THE SCHRÖDINGER EQUATION*

PETER A. MARKOWICH†, PAOLA PIETRA‡, CARSTEN POHL§, AND HANS PETER STIMMING†

Abstract. We apply Wigner transform techniques to the analysis of the Dufort–Frankel difference scheme for the Schrödinger equation and to the continuous analogue of the scheme in the case of a small (scaled) Planck constant (semiclassical regime). In this way we are able to obtain sharp conditions on the spatial-temporal grid which guarantee convergence for average values of observables as the Planck constant tends to zero. The theory developed in this paper is *not* based on local and global error estimates and does *not* depend on whether or not caustics develop. Numerical test examples are presented to help interpret the theory and to compare the Dufort–Frankel scheme to other difference schemes for the Schrödinger equation.

Key words. Schrödinger equation, finite difference scheme, semiclassical limit, Wigner measure

AMS subject classifications. 65M06, 65M12, 35B40, 81Q05

PII. S0036142900381734

1. Introduction.

We shall analyze the Dufort–Frankel discretization scheme for the linear Schrödinger equation

$$\varepsilon u_t^\varepsilon = i\frac{\varepsilon^2}{2}\Delta u^\varepsilon - iV(x)u^\varepsilon, \quad x \in \mathbb{R}^d, \quad t \in \mathbb{R} \tag{1.1}$$

$$u^\varepsilon(x, t=0) = u_I^\varepsilon(x), \qquad x \in \mathbb{R}^d. \tag{1.2}$$

Here $0 < \varepsilon \leq \varepsilon_0 < 0$ is the (scaled) Planck constant (generally assumed to be small in what follows), $V = V(x)$ is a given electrostatic potential, and $u^\varepsilon = u^\varepsilon(x,t)$ is the (complex valued) wave function. In classical quantum physics, this function is used to compute observables (including the primary physical quantities), which are quadratic functions (or functionals) of $u^\varepsilon(t)$ [LL]. In this paper, we want to study the behavior of the discretization scheme in the limit case $\varepsilon \to 0$, the so-called semiclassical limit. As the Schrödinger equation propagates oscillations of wavelength ε, the wave function u^ε does not converge strongly, for example, in $L_t^\infty(L_x^2)$, and weak convergence is not sufficient for passing to the limit in the observables. By introducing the Wigner measure (see [G2], [LP], [MMP], [GMMP], [Wi] or other tools of microlocal analysis [G1], [T]) this passage to the limit in the macroscopic densities becomes possible.

The highly oscillatory nature of the solutions of (1.1) poses a problem if this equation is solved numerically in the case of a small scaled Planck constant ε. The

*Received by the editors November 27, 2000; accepted for publication (in revised form) February 15, 2002; published electronically September 27, 2002. This research was supported by the European TMR network "Asymptotic Methods in Kinetic Equations," grant ERB-FMBX-CT-0157.
http://www.siam.org/journals/sinum/40-4/38173.html

†Institut für Mathematik, Universität Wien, Boltzmanngasse 9, A-1090 Wien, Austria (peter.markowich@univie.ac.at). The first author was supported by the "Wittgenstein 2000 Award," founded by the Austrian Research Fund FWF. The fourth author was supported by the Austrian government START prize project "Nonlinear Schrödinger and Quantum Boltzmann equations," grant FWF Y-137-TEC.

‡Istituto di Analisi Numerica del C.N.R., Via Ferrata 1, I-27100 Pavia, Italy (pietra@dragon.ian.pv.cnr.it).

§SAP AG, Neurottstrasse 16, D-69190 Walldorf, Germany (carsten.pohl@sap.com).

oscillations may disturb the numerical solution in such a way that the physical observables come out completely wrong. If the microlocal techniques used to analyze the semiclassical limit for the IVP (1.1)–(1.2) are adapted to the analysis of finite difference discretizations, sharp conditions can be found which guarantee the convergence of observables. This has been done in [MPP] for other numerical schemes, in particular for the Crank–Nicolson and Leap-Frog schemes.

For the sake of transparency we shall now set up the Dufort–Frankel scheme for the case of one space dimension ($d = 1$). Also, the subsequent analysis will mainly focus on the case $d = 1$. Generalizations to $d > 1$ are immediate for tensor product grids, and the results remain valid without modifications.

We choose a spatial mesh size $\Delta x > 0$ and a temporal mesh size $\Delta t > 0$ and denote the grid points

$$x_j := j\,\Delta x, \quad t_n := n\,\Delta t, \quad n, j \in \mathbb{Z}.$$

The Dufort–Frankel scheme for (the one-dimensional version of) (1.1) reads

$$\varepsilon \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = i\frac{\varepsilon^2}{2} \frac{u_{j+1}^n - (u_j^{n+1} + u_j^{n-1}) + u_{j-1}^n}{(\Delta x)^2} - iV(x_j)\frac{u_j^{n+1} + u_j^{n-1}}{2}, \quad n \in \mathbb{Z}, \quad j \in \mathbb{Z} \tag{1.3}$$

$$u_j^0 = u_I^\varepsilon(x_j), \qquad j \in \mathbb{Z}, \tag{1.4}$$

$$u_j^1 = \tilde{u}_{I,j}^\varepsilon, \qquad j \in \mathbb{Z}. \tag{1.5}$$

Obviously, the solution u_j^n is considered an approximation of $u^\varepsilon(x_j, t_n)$. Observe that if in (1.3) the time average $(u_j^{n+1} + u_j^{n-1})/2$ is replaced by u_j^n, we obtain the Leap-Frog scheme.

Since the Dufort–Frankel scheme is a two-step scheme (in time), additional initial data at $t = \Delta t$ (or $t = -\Delta t$) have to be prescribed. Later on we shall discuss the choice of the values $\tilde{u}_{I,j}^\varepsilon$ in detail.

Although, formally, the scheme (1.3) is implicit, the $(n+1)$st time level contributions on the right-hand side of (1.3) occur only on the diagonal, thus the computational effort involved in solving (1.3) is exactly the same as for an explicit scheme (e.g., the Leap-Frog scheme). The scheme also is time reversible and has some favorable stability properties, which make it suitable for the numerical solution of the Schrödinger equation. This was done in [IR] and [W] for both linear and nonlinear Schrödinger equations for the case of $\varepsilon > 0$ fixed, i.e., *not* close to the semiclassical regime. There the authors use the classical stability-consistency method described below but do not address the question of the behavior of the discretization in the case of the limit $\varepsilon \to 0$.

The classical analysis of finite difference schemes relies on the stability-consistency concept. The local discretization error of a difference scheme is calculated by inserting the "exact" solution of the differential equation into the difference scheme and calculating the residual by, say, Taylor expansion. Replacing u_j^n in (1.3) by $u^\varepsilon(x_j, t_n)$ gives

$$\begin{aligned} \varepsilon u_t^\varepsilon = i\frac{\varepsilon^2}{2} u_{xx}^\varepsilon - iVu^\varepsilon - i\frac{\varepsilon^2}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 u_{tt}^\varepsilon \\ + O\left(\varepsilon^2(\Delta x)^2 u_{xxxx}^\varepsilon\right) + O\left(\varepsilon(\Delta t)^2 u_{ttt}^\varepsilon\right) \\ + O\left((\Delta t)^2 u_{tt}^\varepsilon\right) + O\left(\varepsilon^2\left(\frac{\Delta t}{\Delta x}\right)^2 (\Delta t)^2 u_{tttt}^\varepsilon\right). \end{aligned} \tag{1.6}$$

It is then clear that the scheme, for fixed ε, is consistent only if $\frac{\Delta t}{\Delta x}$ goes to zero with $\Delta t \to 0$ and $\Delta x \to 0$. Actually, since typically

$$\left\| \frac{\partial^{k_1+k_2}}{\partial x^{k_1} \partial t^{k_2}} u^\varepsilon(t) \right\|_{L^2(\mathbb{R})} = O\left(\frac{1}{\varepsilon^{k_1+k_2}} \right), \qquad k_1, k_2 \in \mathbb{N} \cup \{0\}, \tag{1.7}$$

the local discretization error is

$$l^\varepsilon = O\left(\left(\frac{\Delta t}{\Delta x} \right)^2 \left(1 + \left(\frac{\Delta t}{\varepsilon} \right)^2 \right) + \left(\frac{\Delta t}{\varepsilon} \right)^2 + \left(\frac{\Delta x}{\varepsilon} \right)^2 \right). \tag{1.8}$$

Even when $\frac{\Delta t}{\Delta x}$ is $o(1)$, l^ε tends to zero as ε, Δt, and Δx tend to zero only when temporal and spatial $O(\varepsilon)$-wavelength oscillations are accurately resolved by the temporal-spatial grid.

Note that the above computation of the local discretization error is highlighted by rewriting (1.3) as

$$\begin{aligned} \varepsilon \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} &= i \frac{\varepsilon^2}{2} \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} - iV(x_j) u_j^n \\ &\quad - i \frac{\varepsilon^2}{2} \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} \left(\left(\frac{\Delta t}{\Delta x} \right)^2 + \left(\frac{\Delta t}{\varepsilon} \right)^2 V(x_j) \right). \end{aligned} \tag{1.9}$$

Stability then implies that the local discretization error is not amplified (enough to ruin convergence).

We shall obtain convergence results by both the classical method and by the Wigner measure, finding that the classical method result requires much more restrictions on the choice of discretization steps than the Wigner measure method. We identify the semiclassical Wigner measure (on the scale ε) for combinations of ε and the space and time mesh sizes and conclude convergence of all (smooth) observables in exactly those cases for which the Wigner measure of the numerical scheme is identical to the Wigner measure of the Schrödinger equation itself.

We remark that the analysis of the Dufort–Frankel scheme (which is frequently used in applications) is technically much more challenging than the analysis for the other schemes presented in [MPP]. This is mainly due to the mixing of spatial and temporal grids in this two-step scheme.

The paper is organized as follows. In section 2, we shall give a brief presentation of the analytical tools and calculate the Wigner measure of the modified equation related to the form (1.9) of the scheme (in order to get a "feeling" for the analysis). In section 3, we shall derive a conservation property of the Dufort–Frankel scheme and from this the stability, as well as the convergence by the classical method, keeping track explicitly of the parameter ε in the convergence estimate. In section 4, we shall calculate the Wigner measure of the scheme, and in section 5 we present a sample of numerical computations illustrating the theory.

2. The modified Schrödinger equation. Let μ be a positive parameter (not necessarily small) and consider

$$\varepsilon v_t^{\varepsilon,\mu} = i \frac{\varepsilon^2}{2} \Delta v^{\varepsilon,\mu} - iV(x) v^{\varepsilon,\mu} - i \frac{\varepsilon^2}{2} \mu^2 v_{tt}^{\varepsilon,\mu}, \quad x \in \mathbb{R}^d, \quad t \in \mathbb{R}, \tag{2.1}$$

$$v^{\varepsilon,\mu}(x, t=0) = v_I^{\varepsilon,\mu}(x), \tag{2.2}$$

$$v_t^{\varepsilon,\mu}(x, t=0) = p_I^{\varepsilon,\mu}(x), \qquad x \in \mathbb{R}^d. \tag{2.3}$$

Notice that (2.1) is a "continuous" version of (1.9). We expect to get important asymptotic information on (1.9) by first analyzing (2.1). The superscript ε for the solution $u = u^\varepsilon$ of the Schrödinger equation (1.1) and ε, μ for the solution $v = v^{\varepsilon,\mu}$ of the modified Schrödinger equation (2.1) were introduced to emphasize the dependence of the solution on the parameters. Where convenient, these superscripts will be skipped in what follows to simplify the notation.

At first we remark that the transformation $z^{\varepsilon,\mu} := \exp(-i\frac{t}{\varepsilon\mu^2})v^{\varepsilon,\mu}$ gives the Klein–Gordon equation with potential $V(x) + \frac{1}{2\mu^2}$:

$$\frac{\varepsilon^2\mu^2}{2} z_{tt}^{\varepsilon,\mu} = \frac{\varepsilon^2}{2}\Delta z^{\varepsilon,\mu} - \left(V(x) + \frac{1}{2\mu^2}\right) z^{\varepsilon,\mu}, \quad x \in \mathbb{R}^d, \quad t \in \mathbb{R} \tag{2.4}$$

$$z^{\varepsilon,\mu}(x, t = 0) = v_I^{\varepsilon,\mu}(x), \tag{2.5}$$

$$z_t^{\varepsilon,\mu}(x, t = 0) = p_I^{\varepsilon,\mu}(x) - \frac{i}{\varepsilon\mu^2} v_I^{\varepsilon,\mu}(x), \qquad x \in \mathbb{R}^d. \tag{2.6}$$

The IVP (2.4) does not conserve total charge (as opposed to the Schrödinger equation).

However, energy conservation is easily proven by multiplying (2.1) by $\overline{v}_t = \overline{v}_t^{\varepsilon,\mu}$ ("–" stands for complex conjugation) and integrating over $\mathbb{R}^d$. We obtain

$$\begin{aligned} E^{\varepsilon,\mu}(t) &:= \int_{\mathbb{R}^d} \left(\frac{\varepsilon^2}{2} |\nabla v^{\varepsilon,\mu}|^2 + \frac{\varepsilon^2}{2}\mu^2 |v_t^{\varepsilon,\mu}|^2 + V(x)|v^{\varepsilon,\mu}|^2 \right) dx \\ &= \int_{\mathbb{R}^d} \left(\frac{\varepsilon^2}{2} |\nabla v_I^{\varepsilon,\mu}|^2 + \frac{\varepsilon^2}{2}\mu^2 |p_I^{\varepsilon,\mu}|^2 + V(x)|v_I^{\varepsilon,\mu}|^2 \right) dx \\ &=: E_I^{\varepsilon,\mu} \quad \forall t \in \mathbb{R}. \end{aligned} \tag{2.7}$$

For the sake of simplicity we shall assume from now on that

$$\begin{aligned} &V \in C^\infty(\mathbb{R}^d) \quad \forall l_1, \ldots, l_d \in \mathbb{N} : \quad \frac{\partial^{l_1+\cdots+l_d}}{\partial x_1^{l_1} \cdots \partial x_d^{l_d}} V \in L^\infty(\mathbb{R}^d), \\ &V(x) \geq 0 \text{ on } \mathbb{R}^d. \end{aligned} \tag{2.8}$$

We conclude L^2-stability in the following lemma.

LEMMA 2.1. *The estimate*

$$\|v^{\varepsilon,\mu}(t)\|_{L^2(\mathbb{R}^d)} \leq \mu\sqrt{2E_I^{\varepsilon,\mu}} + \|v_I^{\varepsilon,\mu}\|_{L^2(\mathbb{R}^d)} \tag{2.9}$$

holds for all $t \in \mathbb{R}$.

Proof. Multiplication of (2.1) by the conjugate $\overline{v}$, integration over $\mathbb{R}^d \times (0, T)$, and taking real parts gives

$$\int_{\mathbb{R}^d} |v(T)|^2 dx = \int_{\mathbb{R}^d} |v(0)|^2 dx + \varepsilon\mu^2 \int_{\mathbb{R}^d} \mathrm{Im}\,(v_t(0)\overline{v}(0) - v_t(T)\overline{v}(T)) dx$$

and, using (2.7),

$$\int_{\mathbb{R}^d} |v(T)|^2 dx \leq \int_{\mathbb{R}^d} |v(0)|^2 dx + \mu\sqrt{2E_I^{\varepsilon,\mu}} \left(\|v(T)\|_{L^2(\mathbb{R}^d)} + \|v(0)\|_{L^2(\mathbb{R}^d)} \right).$$

Then (2.9) is immediate. ∎

The L^2-stability is uniform in ε and μ if the initial energy and the initial local charge are bounded uniformly in ε and μ and if $\mu \leq \mu_0 < \infty$.

The following theory will be largely based on Wigner transforms. For functions $v, u \in L^2(\mathbb{R}^d)$ we define their Wigner transform on the scale $\varepsilon > 0$ as the phase space function

$$w^\varepsilon(u,v)(x,\xi) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} u\left(x - \frac{\varepsilon}{2}\eta\right) \overline{v}\left(x + \frac{\varepsilon}{2}\eta\right)\ e^{i\xi\cdot\eta}\ d\eta. \tag{2.10}$$

Obviously, $w^\varepsilon(u,v)$ is bilinear and $w^\varepsilon(u,u) =: w^\varepsilon[u]$ is a real-valued function.

The mathematical (and physical) literature on Wigner transforms is substantial, and we are not able to give a detailed review here. As basic references we recommend [G1], [LP], [GMMP], [MPP].

Here we remark only that for $a \in S(\mathbb{R}^{2d}_{x,\xi})$ we have

$$E^\varepsilon_a[u] := \int_{\mathbb{R}^d} (a(x,\varepsilon D)^W u)\ \overline{u}\ dx = \int_{\mathbb{R}^d_x \times \mathbb{R}^d_\xi} a(x,\xi)\ w^\varepsilon[u](x,\xi)\ dxd\xi, \tag{2.11}$$

where $a(\,.\,,\varepsilon D)^W$ denotes the Weyl operator associated with the symbol $a(x,\xi)$, defined as

$$a(\,.\,,\varepsilon D)^W\varphi(x) := \frac{1}{(2\pi)^m} \int_{\mathbb{R}^d_y} \int_{\mathbb{R}^d_\xi} a\left(\frac{x+y}{2}, \varepsilon\xi\right) \varphi(y) e^{i(x-y)\cdot\xi} d\xi dy \tag{2.12}$$

(see [HM3], [GMMP]). Physically, the left-hand side of (2.11) denotes the average value of the observable a in the state u (cf. [LL]), which by (2.11) can be obtained by testing the symbol $a(x,\xi)$ against the Wigner transform of the function u.

Note that the average values of observables (2.11) are the physically important quantities which can be obtained as a "postprocessing" from the wave function $u = u^\varepsilon$ after having solved the Schrödinger equation.

Also, we remark that a can be chosen independent of x in (2.11) such that

$$\rho(x) := |u(x)|^2 = \int_{\mathbb{R}^d_\xi} w^\varepsilon[u](x,\xi)\ d\xi. \tag{2.13}$$

Similarly,

$$J(x) := \varepsilon \mathrm{Im}\ (\overline{u}(x)\nabla u(x)) = \int_{\mathbb{R}^d_\xi} \xi\ w^\varepsilon[u](x,\xi)\ d\xi \tag{2.14}$$

and, formally,

$$e(x) := \frac{\varepsilon^2}{2}|\nabla u(x)|^2 - \frac{\varepsilon^2}{2}\mathrm{Re}\ (\overline{u}(x)\Delta u(x)) = \int_{\mathbb{R}^d_\xi} |\xi|^2 w^\varepsilon[u](x,\xi)\ d\xi. \tag{2.15}$$

Here ρ, J, e denote the position density, current density, and energy density of the state u, respectively.

Also, let $u^\varepsilon = u^\varepsilon(x)$ be a sequence, uniformly bounded in $L^2(\mathbb{R}^d)$ as $\varepsilon \to 0$. Then there exists a subsequence $(\varepsilon_k) \rightharpoonup 0$ such that

$$w^{\varepsilon_k}[u^{\varepsilon_k}] \overset{\varepsilon_k \rightharpoonup 0}{\rightharpoonup} w^0 = w^0[u^{\varepsilon_k}] \quad \text{in } S'(\mathbb{R}^{2d}_{x,\xi}), \tag{2.16}$$

where w^0 is a positive phase space measure, called the Wigner measure of u^ε on the scale ε_k.

It is now our goal to compute the Wigner measure(s) of the solution $v = v^{\varepsilon,\mu}$ of (2.1) on the ε-scale with, say, $\mu = \mu(\varepsilon) > 0$. We set

$$w^{\varepsilon,\mu} := w^{\varepsilon}[v^{\varepsilon,\mu}] = w^{\varepsilon}(v^{\varepsilon,\mu}, v^{\varepsilon,\mu}).$$

Differentiation of the bilinear Wigner transform with respect to t gives

$$\frac{\partial}{\partial t} w^{\varepsilon,\mu} = \frac{2}{\varepsilon}\, \mathrm{Im}\, w^{\varepsilon}\left(\frac{\varepsilon^2}{2}\Delta v - Vv,\ v\right) - \varepsilon\, \mathrm{Im}\, w^{\varepsilon}\left(\mu^2 v_{tt},\ v\right), \tag{2.17}$$

using $w^{\varepsilon}(f,g) = \overline{w^{\varepsilon}(g,f)}$, $\partial_t w^{\varepsilon}(v,v) = 2\ \mathrm{Re}\, w(v_t, v)$, and (2.1).

To proceed we shall use the following lemma.

LEMMA 2.2. *Let $P \in C^{\infty}(\mathbb{R}^m_x \times \mathbb{R}^m_{\xi})$ satisfy for some $M \geq 0, C_{\alpha} \geq 0$*

$$|\partial^{\alpha}_{x,\xi} P(x,\xi)| \leq C_{\alpha}(1 + |\xi|)^M \quad \forall \alpha \in \mathbb{N}^m_0 \times \mathbb{N}^m_0. \tag{2.18}$$

Then, if f, g lie in a bounded subset of $L^2(\mathbb{R}^m)$, the expansion

$$w^{\varepsilon}\left(P(\,.\,, \varepsilon D)^W f, g\right) = P w^{\varepsilon}(f,g) + \frac{\varepsilon}{2i}\{P, w^{\varepsilon}(f,g)\} + O(\varepsilon^2)$$

holds in $S'(\mathbb{R}^m \times \mathbb{R}^m)$ uniformly for all symbols $P = P(x,\xi)$ satisfying (2.18).

Here $\{\ .\ ,\ .\ \}$ denotes the Poisson bracket

$$\{f,g\} = \nabla_{\xi} f \nabla_x g - \nabla_x f \nabla_{\xi} g. \tag{2.19}$$

A proof of Lemma 2.2 can be found in [GMMP]. We can now pass to the limit in the first term of the right-hand side setting $P(x,\xi) = \frac{1}{2}|\xi|^2 + V(x)$:

$$\frac{2}{\varepsilon}\, \mathrm{Im}\, w^{\varepsilon}\left(\frac{\varepsilon^2}{2}\Delta v - Vv,\ v\right) \rightharpoonup \left\{\ \frac{1}{2}|\xi|^2 + V(x),\ w^0\right\},$$

where

$$w^0 := \ \mathrm{w}\text{-}\lim_{\varepsilon \to 0}\ w^{\varepsilon}(v^{\varepsilon,\mu}, v^{\varepsilon,\mu})$$

(after extraction of a subsequence). To pass to the limit in the second term we multiply by a real-valued smooth test function φ with sufficient decay at $|x| = \infty$, $|\xi| = \infty$ and integrate by parts:

$$\begin{aligned} &\varepsilon\ \mathrm{Im} \int_{\mathbb{R}^{2d}_{x,\xi}} \int_{\mathbb{R}_t} \varphi w^{\varepsilon}(\mu^2 v_{tt}, v)\ dtdxd\xi \\ &= -\ \varepsilon \int_{\mathbb{R}^{2d}_{x,\xi}} \int_{\mathbb{R}_t} \varphi\ \mathrm{Im}\, w^{\varepsilon}(\mu v_t, \mu v_t)\ dtdxd\xi \\ &\quad -\ \mu \int_{\mathbb{R}^{2d}_{x,\xi}} \int_{\mathbb{R}_t} \varphi_t\ \mathrm{Im}\, w^{\varepsilon}(\varepsilon\mu v_t, v)\ dtdxd\xi. \end{aligned} \tag{2.20}$$

The first term is zero, since $w^{\varepsilon}(f,f)$ is real-valued, and the second term can be estimated using

$$\|w^{\varepsilon}(f,g)\|_{\mathcal{A}^*} \leq \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}, \tag{2.21}$$

where $\mathcal{A}$ is a suitable Banach space of test functions (cf. [LP]) containing $\mathcal{D}(\mathbb{R}^{2d}_{x,v})$. Choosing $\varphi \in C_0^\infty(\mathbb{R}_t; \mathcal{A})$ we obtain

$$\begin{aligned}&\varepsilon \left| \operatorname{Im} \int_{\mathbb{R}^{2d}_{x,v}} \varphi w^\varepsilon(\mu^2 v_{tt}, v)\, dtdxdv \right| \\ &\le \mu K(\varphi) \sup_{t\in\mathbb{R}} \left\| \varepsilon\mu v_t(t) \right\|_{L^2(\mathbb{R}^d)} \left\| v(t) \right\|_{L^2(\mathbb{R}^d)}.\end{aligned}$$

We have thus proven the following lemma.

LEMMA 2.3. *Let $E_I^{\varepsilon,\mu} \le L$ for $\varepsilon \to 0$ and let $\mu \to 0$. Then the Wigner measures satisfy*

$$w_t^0 + \left\{ \frac{1}{2}|\xi|^2 + V(x),\ w^0 \right\} = 0, \tag{2.22}$$

$$w^0(t=0) = w^0[v_I^{\varepsilon,\mu}]. \tag{2.23}$$

Note that the possible nonuniqueness of the Wigner measures stem from the possible nonuniqueness of the initial Wigner measures $w^0[v_I^{\varepsilon,\mu}]$. (Different subsequences might give different limits.)

If μ does not tend to zero we have to proceed differently. Therefore we use (2.1) to compute the second term on the right-hand side of (2.20). (The first term is zero!) For $\mu = \text{const}$ we obtain

$$\begin{aligned}w^\varepsilon(\mu^2 v_t, v) &= -\frac{i}{\varepsilon} w^\varepsilon \left(\mu^2 \left(\frac{\varepsilon^2}{2}\Delta v - Vv \right), v \right) + i\frac{\varepsilon}{2} w^\varepsilon(\mu^4 v_{tt}, v) \\ &= -\frac{i}{\varepsilon}\mu^2 \left(\frac{|\xi|^2}{2} + V(x) \right) w^\varepsilon \\ &\quad + \frac{1}{2}\mu^2 \left\{ \frac{|\xi|^2}{2} + V(x), w^\varepsilon \right\} + O(\varepsilon\mu^2) \\ &\quad + i\frac{\varepsilon}{2}\mu^4 w^\varepsilon(v_{tt}, v)\end{aligned}$$

such that

$$\begin{aligned}\varepsilon \operatorname{Im} w^\varepsilon(\mu^2 v_t, v) = &-\mu^2 \left(\frac{|\xi|^2}{2} + V(x) \right) w^\varepsilon + O(\varepsilon^2\mu^2) \\ &+ \frac{\varepsilon^2\mu^4}{2} \operatorname{Re} w^\varepsilon(v_{tt}, v).\end{aligned}$$

Since, for $r, s \in \mathbb{R}^d$,

$$\big(\overline{v}(r,t)v(s,t)\big)_{tt} = \overline{v}(r,t)_{tt} v(s,t) + 2\overline{v}_t(r,t)v_t(s,t) + \overline{v}(r,t)v_{tt}(s,t)$$

we conclude that

$$\operatorname{Re} w^\varepsilon(v_{tt}, v) = \frac{1}{2} w^\varepsilon(v,v)_{tt} - w^\varepsilon(v_t, v_t),$$

and thus

$$\begin{aligned}\frac{\varepsilon^2\mu^4}{2} \operatorname{Re} w^\varepsilon(v_{tt}, v) = &\frac{\varepsilon^2\mu^4}{4} w^\varepsilon(v,v)_{tt} \\ &- \mu^2 w^\varepsilon \left(\frac{\varepsilon\mu}{\sqrt{2}} v_t, \frac{\varepsilon\mu}{\sqrt{2}} v_t \right).\end{aligned}$$

We obtain as $\varepsilon \to 0$

$$\varepsilon \mathrm{Im}\, w^\varepsilon(\mu^2 v_t, v) \overset{\varepsilon\to 0}{\to} -\mu^2 \left(\frac{|\xi|^2}{2} + V(x)\right) w^0 - \mu^2 w^{(1)},$$

where

$$w^{(1)} := \text{w–}\lim_{\varepsilon\to 0} w^\varepsilon \left(\frac{\varepsilon\mu}{\sqrt{2}} v_t^{\varepsilon,\mu}, \frac{\varepsilon\mu}{\sqrt{2}} v_t^{\varepsilon,\mu}\right). \tag{2.24}$$

Clearly, $w^{(1)}$ exists as a positive measure if the initial energy $E_I^{\varepsilon,\mu}$ is uniformly bounded as $\varepsilon \to 0$.

LEMMA 2.4. *Let $E_I^{\varepsilon,\mu} \le L$ for $\varepsilon \to 0$ and let $\mu > 0$ be constant. Then*

$$\left(1 + \mu^2 \left(\frac{|\xi|^2}{2} + V(x)\right)\right) w_t^0 + \left\{\frac{1}{2}|\xi|^2 + V(x),\ w^0\right\} + \mu^2 w_t^{(1)} = 0,$$

$$w^0(t=0) = w^0[v_I^{\varepsilon,\mu}].$$

We are thus left with calculating $w^{(1)}$.

Therefore we shall apply the homogenization theory for systems developed in [GMMP]. We start by defining the new variables

$$q := i\sqrt{V(x)}v, \quad r := \frac{\varepsilon}{\sqrt{2}}\mu v_t, \quad s := \frac{\varepsilon}{\sqrt{2}}\nabla v \tag{2.25}$$

and rewrite (2.1) as the system

$$\varepsilon \begin{pmatrix} q \\ r \\ s \end{pmatrix}_t + i \begin{pmatrix} 0 & -\frac{\sqrt{2}}{\mu}\sqrt{V} & 0 \\ -\frac{\sqrt{2}}{\mu}\sqrt{V} & -\frac{2}{\mu^2} & i\frac{\varepsilon}{\mu}\,\mathrm{div}_x \\ 0 & i\frac{\varepsilon}{\mu}\nabla_x & 0 \end{pmatrix} \begin{pmatrix} q \\ r \\ s \end{pmatrix} = 0, \tag{2.26}$$

$$q(t=0) = i\sqrt{V(x)}v_I^{\varepsilon,\mu}, \quad r(t=0) = \frac{\varepsilon}{\sqrt{2}}\mu p_I^{\varepsilon,\mu}, \quad s(t=0) = \frac{\varepsilon}{\sqrt{2}}\nabla v_I^{\varepsilon,\mu}, \tag{2.27}$$

with the symbol matrix

$$P^\mu(x,\xi) := \begin{pmatrix} 0 & -\frac{\sqrt{2V(x)}}{\mu} & 0 \\ -\frac{\sqrt{2V(x)}}{\mu} & -\frac{2}{\mu^2} & -\frac{1}{\mu}\xi^T \\ 0 & -\frac{1}{\mu}\xi & 0 \end{pmatrix}. \tag{2.28}$$

Then (2.26) reads

$$\varepsilon\psi_t + iP^\mu(x, \varepsilon D)^W \psi = 0, \qquad \psi(t=0) = \psi_I^{\varepsilon,\mu}, \tag{2.29}$$

where we set $\psi = (q,\ r,\ s)$ with $\psi_I^{\varepsilon,\mu}$ defined by (2.27), $D = -i\nabla_x$, and the superscript "W" denotes the Weyl operator which in (2.29) agrees with the operator generated by P^μ as left symbol matrix. Here we wrote (2.29) with the Weyl operator in accordance with [GMMP].

Since $P^\mu(x,\xi)$ is pointwise symmetric, $P^\mu(x,\varepsilon D)^W$ is formally self-adjoint (self-adjointness is easily concluded from the assumed regularity of $V(x)$) and, then, $e^{\varepsilon,\mu} := |\psi|^2$ is an $L^1(\mathbb{R}^d)$-conserved quantity; in fact,

$$E^{\varepsilon,\mu} = \int_{\mathbb{R}^d} e^{\varepsilon,\mu}(x,t)dx, \tag{2.30}$$

where $E^{\varepsilon,\mu}$ is the total energy defined in (2.7).

We can now define the Wigner matrix

$$W^\varepsilon(t) := \begin{pmatrix} w^\varepsilon(q,q) & w^\varepsilon(q,r) & w^\varepsilon(q,s_1) & \cdots & w^\varepsilon(q,s_d) \\ w^\varepsilon(r,q) & w^\varepsilon(r,r) & w^\varepsilon(r,s_1) & \cdots & w^\varepsilon(r,s_d) \\ w^\varepsilon(s_1,q) & w^\varepsilon(s_1,r) & w^\varepsilon(s_1,s_1) & \cdots & w^\varepsilon(s_1,s_d) \\ \vdots & \vdots & \vdots & & \vdots \\ w^\varepsilon(s_d,q) & w^\varepsilon(s_d,r) & w^\varepsilon(s_d,s_1) & \cdots & w^\varepsilon(s_d,s_d) \end{pmatrix} \tag{2.31}$$

and conclude, as for the scalar case,

$$W^\varepsilon \rightharpoonup W^0 \quad \text{in } L^\infty\left(\mathbb{R}_t;\ S'(\mathbb{R}^d_x \times \mathbb{R}^d_\xi)^{d+2}\right), \tag{2.32}$$

where the hermitian matrix W^0 is positive definite in the sense of measures; i.e.,

$$\xi^T W^0 \xi \in L^\infty\left(\mathbb{R}_t;\ \mathcal{M}^+(\mathbb{R}^d_x \times \mathbb{R}^d_\xi)\ \right) \quad \forall \xi \in \mathbb{R}^d.$$

($\mathcal{M}^+$ denotes the space of positive measures.)

The result from the theory in [GMMP], which we shall restate here, requires the following assumptions on the matrix symbol $P^\mu(x,\xi)$:

(A1) (i) There exists a closed subset E of $\mathbb{R}^d_x \times \mathbb{R}^d_\xi$ such that, for every $(x,\xi) \notin E$, the eigenvalues of $P^\mu(x,\xi)$ can be ordered as follows:

$$\lambda_1(x,\xi) < \cdots < \lambda_l(x,\xi),$$

where, for $1 \le k \le l$, the multiplicity of λ_k does not depend on $(x,\xi) \notin E$.

(ii) For $1 \le k \le l$, the Hamiltonian flow of λ_k leaves invariant the set

$$\Omega = (\mathbb{R}^d_x \times \mathbb{R}^d_\xi) \setminus E.$$

(iii) E is a null set of the measure $w^0_I = \operatorname{tr}(W^0_I)$.

For $1 \le k \le l$, we denote by $\Pi_q(x,\xi)$ the orthogonal projection of $\mathbb{C}^d$ on the eigenspace associated with $\lambda_k(x,\xi)$.

(A2) $\exists \sigma \in \mathbb{R}$: $P^\mu \in S^\sigma(\mathbb{R}^d)^{d+2\times d+2}$ uniformly in ε, which means that for all $\alpha,\beta \in \mathbb{N}_0$ there exists $C_{\alpha,\beta}$ such that for all $l,k \in \{1,\dots,d\}$ and for all $\varepsilon \in (0,\varepsilon_0]$ we have

$$\left| \frac{\partial^{\alpha+\beta}}{\partial x_k^\alpha \partial \xi_l^\beta} P^\mu(x,\xi) \right| \le C_{\alpha,\beta}(1+|\xi|)^{\sigma-\beta}$$

for all $(x,\xi) \in \mathbb{R}^d_x \times \mathbb{R}^d_\xi$.

For the convergence of the energy density $e^{\varepsilon,\mu}$, we also need more assumptions on the initial data $\psi_I^{\varepsilon,\mu}$:

(A3) $\psi_I^{\varepsilon,\mu}$ is bounded in $L^2(\mathbb{R}^d_x)^{d+2}$, ε-oscillatory, and compact at infinity, which means that, for every continuous compactly supported function φ on $\mathbb{R}^d$,

$$\overline{\lim_{\varepsilon\to 0}} \int_{|\xi|\geq R/\varepsilon} \left|\widehat{\varphi\psi}_I^{\varepsilon,\mu}(\xi)\right|^2 \to 0 \quad \text{as } R\to+\infty, \tag{2.33}$$

respectively,

$$\overline{\lim_{\varepsilon\to 0}} \int_{|x|\geq R} |\psi_I^{\varepsilon,\mu}(x)|^2 dx \to 0 \quad \text{as } R\to+\infty. \tag{2.34}$$

We state the main result of [GMMP].

LEMMA 2.5. *Let $P^\mu(x,\xi)$ be essentially self-adjoint on $L^2(\mathbb{R}^d)^{d+2}$ and satisfy* (A1), (A2), *and let $\psi_I^{\varepsilon,\mu}$ satisfy* (A3). *Let $\{.,.\}$ denote the Poisson bracket* (2.19).

(i) *For $1\leq k\leq l$, we denote by $w_k^0(t)$ the continuously t-dependent positive scalar measure on $\mathbb{R}^d_x\times\mathbb{R}^d_\xi$ defined by*

$$\begin{aligned} \frac{\partial}{\partial t} w_k^0 + \{\lambda_k, w_k^0\} &= 0 \quad \text{on } \mathbb{R}_t\times\Omega,\\ w_k^0(t=0) &= tr\,(\Pi_k W_I^0) \quad \text{on } \Omega,\\ w_k^0(t,E) &= 0, \qquad t\in\mathbb{R}. \end{aligned} \tag{2.35}$$

Then the scalar Wigner transform of $\psi^\varepsilon(t)$, defined by $w^\varepsilon(x,\xi,t) := tr\, W^\varepsilon(t)$ (which is defined in (2.31)), *converges locally uniformly in t to*

$$w^0(x,\xi,t) = \sum_{k=1}^{l} w_k^0(x,\xi,t), \tag{2.36}$$

and $n^\varepsilon(t,x) := |\psi^{\varepsilon,\mu}|^2$ converges locally uniform in t to

$$n^0(t,x) = \int_{\mathbb{R}^d_\xi} w^0(t,x,d\xi). \tag{2.37}$$

(ii) *For $1\leq k\leq l$, we set $F_k = [\Pi_k,\{\lambda_k,\Pi_k\}] + \frac{1}{2}\sum_{j=1}^l(\lambda_k-\lambda_j)\Pi_k\{\Pi_j,\Pi_j\}\Pi_k$; denote by W_k^0 the continuously t-dependent positive matrix-valued measure on $\mathbb{R}^d_x\times\mathbb{R}^d_\xi$ defined by*

$$\begin{aligned} \frac{\partial}{\partial t} W_k^0 + \{\lambda_k, W_k^0\} &= [W_k^0, F_k] \quad \text{on } \mathbb{R}_t\times\Omega,\\ W_k^0(t=0) &= \Pi_k W_I^0 \Pi_k \quad \text{on } \Omega,\\ W_k^0(t,E) &= 0, \qquad t\in\mathbb{R}. \end{aligned} \tag{2.38}$$

Then the Wigner matrix (2.31) *converges in $L^\infty(\mathbb{R}_t, S')$ weak - $*$ to*

$$W^0 = \sum_{k=1}^{l} W_k^0. \tag{2.39}$$

Denoting $W^0 = (W^0_{ij})_{i,j=1}^{d+2}$, we obtain $w^{(1)}$ of Lemma 2.4 as $W^0_{2,2} = w^0(r,r) = w^{(1)}$, according to the definition in (2.31).

As already mentioned (cf. (2.30)), the energy density $n^\varepsilon = |\psi^\varepsilon|^2$ is equal to the total energy of (2.1). Consequently, by (2.37) we also have convergence of the energy of (2.1), provided that $\psi_I^{\varepsilon,\mu}$ fulfills (A3).

It remains to check the above assumptions on P^μ. The eigenvalues of $P^\mu(x,\xi)$ are

$$\begin{aligned}
\lambda_1 &= -\frac{1}{\mu^2}\left(1+\sqrt{1+\mu^2(|\xi|^2+2V(x))}\right),\\
\lambda_2 &= 0,\\
\lambda_3 &= -\frac{1}{\mu^2}\left(1-\sqrt{1+\mu^2(|\xi|^2+2V(x))}\right),
\end{aligned}$$

where λ_2 is of multiplicity d. If we now set $E = \{(x,\xi)|\xi = 0, V(x) = 0\}$, we have

$$\lambda_1(x,\xi) < \lambda_2(x,\xi) < \lambda_3(x,\xi) \quad \forall (x,\xi) \in \mathbb{R}^d_x \times \mathbb{R}^d_\xi \setminus E.$$

We now assume $w^0_I(\{(x,\xi)|\xi = 0, V(x) = 0\}) = 0$. Then (A1) is satisfied, since the Hamiltonian flows of λ_1, λ_2, and λ_3 map E into E bijectively. The corresponding projector matrices are

$$\Pi_1(x,\xi) = \frac{\mu^2}{2(\theta^2+\theta)}\begin{pmatrix} 2V & \frac{\sqrt{2V}}{\mu}(1+\theta) & \sqrt{2V}\xi^T \\ \frac{\sqrt{2V}}{\mu}(1+\theta) & \frac{1}{\mu^2}(1+\theta)^2 & \frac{1}{\mu}(1+\theta)\xi^T \\ \sqrt{2V}\xi & \frac{1}{\mu}(1+\theta)\xi & \xi\otimes\xi \end{pmatrix},$$

$$\Pi_2(x,\xi) = \begin{pmatrix} \sum_{k=1}^d \frac{\xi_k^2}{\xi_k^2+2V} & 0 & -\frac{\sqrt{2V}\xi_1}{\xi_1^2+2V} & \cdots & -\frac{\sqrt{2V}\xi_d}{\xi_d^2+2V} \\ 0 & 0 & 0 & \cdots & 0 \\ -\frac{\sqrt{2V}\xi_1}{\xi_1^2+2V} & 0 & \frac{2V}{\xi_1^2+2V} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ -\frac{\sqrt{2V}\xi_d}{\xi_d^2+2V} & 0 & 0 & & \frac{2V}{\xi_d^2+2V} \end{pmatrix},$$

$$\Pi_3(x,\xi) = \frac{\mu^2}{2(\theta^2-\theta)}\begin{pmatrix} 2V & \frac{\sqrt{2V}}{\mu}(1-\theta) & \sqrt{2V}\xi^T \\ \frac{\sqrt{2V}}{\mu}(1-\theta) & \frac{1}{\mu^2}(1-\theta)^2 & \frac{1}{\mu}(1-\theta)\xi^T \\ \sqrt{2V}\xi & \frac{1}{\mu}(1-\theta)\xi & \xi\otimes\xi \end{pmatrix}.$$

Here we denoted $\theta = \theta^\mu(x,\xi) := \sqrt{1+\mu^2(|\xi|^2+2V(x))}$. (A2) holds by the smoothness assumption on $V(x)$. We leave the (very tedious) explicit calculation of the right-hand sides of the transport equations (2.38) to the interested reader.

In the case $\mu \to \infty$, the transport equation (2.35) reads, for $1 \le k \le 3$,

$$\frac{\partial}{\partial t} w_k^0 = 0, \qquad w_k^0(t=0) = \operatorname{tr}\,(\Pi_k W_I^0),$$

since for the derivatives of the eigenvalues of $P^\mu(x,\xi)$

$$\begin{aligned}
\nabla_x \lambda_{1,3} &= \frac{\mp \nabla_x V}{\sqrt{1+\mu^2(|\xi|^2+2V)}},\\
\nabla_\xi \lambda_{1,3} &= \frac{\pm \xi}{\sqrt{1+\mu^2(|\xi|^2+2V)}},
\end{aligned}$$

$$\nabla_x \lambda_2 = \nabla_\xi \lambda_2 = 0$$

we have as $\mu \to \infty$

$$\lim_{\mu\to\infty} \nabla_x \lambda_k(x,\xi) = \lim_{\mu\to\infty} \nabla_\xi \lambda_k(x,\xi) = 0, \qquad k = 1,2,3.$$

Since also $\lim_{\mu\to\infty} \lambda_k(x,\xi) = 0$ holds for $k = 1,2,3$, (2.38) in this case is

$$\frac{\partial}{\partial t} W_k^0 = 0, \qquad W_k^0(t=0) = \Pi_k W_I^0 \Pi_k,$$

and the Wigner measure matrix remains constant in time.

3. Energy estimates and the consistency-stability method. We now consider the inhomogeneous version of the scheme (1.3):

$$\varepsilon \frac{u^{n+1}(x) - u^{n-1}(x)}{2\Delta t} = i\frac{\varepsilon^2}{2} \frac{u^n(x+\Delta x) - (u^{n+1}(x) + u^{n-1}(x)) + u^n(x-\Delta x)}{(\Delta x)^2} \tag{3.1}$$

$$- iV(x)\frac{u^{n+1}(x) + u^{n-1}(x)}{2} + if^n(x), \quad x \in \mathbb{R}, \quad n \in \mathbb{Z}$$

$$u^0(x) = u_I^\varepsilon(x), \qquad x \in \mathbb{R} \tag{3.2}$$

$$u^1(x) = \tilde{u}_I^\varepsilon(x), \qquad x \in \mathbb{R}. \tag{3.3}$$

For the sake of convenience in the subsequent computation we extended (1.3) from the grid $\{j\Delta x | j \in \mathbb{Z}\}$ to $\mathbb{R}$; i.e., $u^n(j\Delta x) = u_j^n$. $f^n(x)$ denotes an inhomogeneity on the time level $t = n\Delta t$.

We set for the following

$$\delta = \frac{\Delta t}{\Delta x}, \qquad \gamma = \frac{\varepsilon \Delta t}{(\Delta x)^2} \tag{3.4}$$

and define the functional

$$\begin{aligned} E(f,g) := {} & (1+\gamma^2)\left(\|f\|_{L^2}^2 + \|g\|_{L^2}^2\right) \\ & + \delta^2 \int_{\mathbb{R}} V(x)\left(|f|^2 + |g|^2\right) dx \\ & - \operatorname{Re}\left(i\gamma(1-i\gamma)\int_{\mathbb{R}} (f(x-\Delta x) + f(x+\Delta x))\overline{g(x+\Delta x)}dx\right) \end{aligned} \tag{3.5}$$

for $f, g \in L^2(\mathbb{R};\mathbb{C})$. At first we prove the following lemma.

LEMMA 3.1.

$$\frac{1}{2}\left(\|f\|_{L^2}^2 + \|g\|_{L^2}^2\right) \le E(f,g).$$

Proof. The third term in $E(f,g)$ is clearly bounded above by $\gamma\sqrt{1+\gamma^2}(\|f\|_{L^2}^2 + \|g\|_{L^2}^2)$. Therefore

$$E(f,g) \ge \left(1 + \gamma^2 - \gamma\sqrt{1+\gamma^2}\right)\left(\|f\|_{L^2}^2 + \|g\|_{L^2}^2\right),$$

and the assertion follows since

$$1+\gamma^2-\gamma\sqrt{1+\gamma^2}=\frac{\sqrt{1+\gamma^2}}{\gamma+\sqrt{1+\gamma^2}}\geq\frac{1}{2}. \qquad \square$$

We now show that $E(u^n,u^{n+1})$ is a "kind of energy" for (3.1).

LEMMA 3.2. *Let $u^n, n=1,2,\ldots$ satisfy* (3.1). *Then*

$$\begin{aligned}E(u^n,u^{n+1})=E(u^{n-1},u^n)&+\frac{\Delta t}{\varepsilon}\mathrm{Re}\left(\int_{\mathbb{R}}f^n(x)(\overline{u^{n+1}(x)}+\overline{u^{n-1}(x)})\,dx\right)\\&+\gamma\frac{\Delta t}{\varepsilon}\mathrm{Im}\left(\int_{\mathbb{R}}f^n(x)(\overline{u^{n+1}(x)}-\overline{u^{n-1}(x)})\,dx\right).\end{aligned}\tag{3.6}$$

Proof. The assertion of the lemma follows immediately by multiplying (3.1) by the complex conjugate of $(1+i\gamma)u^{n+1}+(1-i\gamma)u^{n-1}$ and integration over $\mathbb{R}$. $\square$

In particular, for the homogeneous problem (3.1) with $f^n=0$ for all $n\in\mathbb{N}$, we conclude "energy" conservation

$$E(u^n,u^{n+1})=E(u^{n-1},u^n),\qquad n=1,2,\ldots \tag{3.7}$$

(cf. [W]) and an $L^2(\mathbb{R})$-estimate from Lemma 3.1:

$$\|u^n\|_{L^2}^2+\|u^{n+1}\|_{L^2}^2\leq E(u_I^\varepsilon,\tilde{u}_I^\varepsilon),\qquad n\in\mathbb{N}. \tag{3.8}$$

We shall now use (3.8) to derive an appropriate choice for the function $\tilde{u}_I^\varepsilon$. Therefore we use the following lemma.

LEMMA 3.3. *$E(f,g)$ can be rewritten as*

$$\begin{aligned}E(f,g)=\|f\|_{L^2}^2&+\|g\|_{L^2}^2\\&+\gamma^2\int_{\mathbb{R}}\left|g(x)-\frac{f(x-\Delta x)+f(x+\Delta x)}{2}\right|^2dx\\&+\frac{\gamma^2}{4}\int_{\mathbb{R}}|f(x+\Delta x)-f(x-\Delta x)|^2\,dx\\&-\gamma\mathrm{Im}\left(\int_{\mathbb{R}}\overline{(f(x-\Delta x)+f(x+\Delta x))}g(x)\,dx\right).\end{aligned}\tag{3.9}$$

The form (3.9) of $E(f,g)$ is easily verified by a direct computation.

We now choose

$$\tilde{u}_I^\varepsilon(x):=\frac{u_I^\varepsilon(x+\Delta x)+u_I^\varepsilon(x-\Delta x)}{2}. \tag{3.10}$$

Then (3.9) simplifies to

$$\begin{aligned}E(u_I^\varepsilon,\tilde{u}_I^\varepsilon)=\|u_I^\varepsilon\|_{L^2}^2&+\left\|\frac{u_I^\varepsilon(\,.\,+\Delta x)+u_I^\varepsilon(\,.\,-\Delta x)}{2}\right\|_{L^2}^2\\&+\gamma^2\Delta x^2\int_{\mathbb{R}}\left|\frac{u_I^\varepsilon(x+\Delta x)-u_I^\varepsilon(x-\Delta x)}{2\Delta x}\right|^2dx.\end{aligned}$$

Assuming that

$$\left\|\frac{d^l}{dx^l}u_I^\varepsilon\right\|_{L^2}=O\left(\frac{1}{\varepsilon^l}\right),\qquad l=0,1,\ldots, \tag{3.11}$$

we conclude that

$$E(u_I^\varepsilon(x), \tilde{u}_I^\varepsilon(x)) = O\left(1 + \left(\frac{\Delta t}{\Delta x}\right)^2\right) \tag{3.12}$$

and L^2-stability follows for the homogeneous problem (3.1) with (3.10) by (3.8) if $\frac{\Delta t}{\Delta x}$ is bounded, which is also required by consistency of the Dufort–Frankel scheme. We summarize in the following proposition.

PROPOSITION 3.1. *Let* (3.11) *hold (for* $l = 0$ *and* $l = 1$*). Then the solution* $u^n = u^n(x)$ *of the homogeneous version (i.e.,* $f^n(x) = 0$*,* $n = 1, 2, \ldots$*) of* (3.1) *with* $\tilde{u}_I^\varepsilon$ *given by* (3.10) *satisfies*

$$\|u^n\|_{L^2} = O\left(1 + \frac{\Delta t}{\Delta x}\right) \tag{3.13}$$

uniformly as $\varepsilon \to 0^+$*. In particular, the scheme is* L^2*-stable if* $\frac{\Delta t}{\Delta x}$ *is bounded as* $\Delta t \to 0$*,* $\Delta x \to 0$*.*

Now we expand the recursion (3.6) in the form

$$E(u^n, u^{n+1}) = E(u^0, u^1) + \frac{\Delta t}{\varepsilon} \sum_{k=1}^{n} A_k,$$

collecting the two last terms on the right-hand side of (3.6) as A_n. Rewriting the terms corresponding to the last term in (3.6) as

$$\begin{aligned}
&\sum_{k=1}^{n} f^k(\overline{u^{k+1}} - \overline{u^{k-1}}) \\
&= \sum_{k=2}^{n-1} (f^{k-1} - f^{k+1})\overline{u^k} + f^{n-1}\overline{u^n} + f^n\overline{u^{n+1}} - f^1\overline{u^0} - f^2\overline{u^1}
\end{aligned}$$

we estimate $E(u^n, u^{n+1})$ as

$$\begin{aligned}
E(u^n, u^{n+1}) \le E(u^0, u^1) &+ \Delta t \sum_{k=1}^{n} \left(\frac{1}{2}\|u^{k+1}\|_{L^2}^2 + \frac{1}{2}\|u^{k-1}\|_{L^2}^2 + \left\| \frac{f^k}{\varepsilon} \right\|_{L^2}^2 \right) \\
&+ \gamma \frac{\Delta t}{2} \left(\|u^n\|_{L^2}^2 + \|u^{n+1}\|_{L^2}^2 + \left\| \frac{f^{n-1}}{\varepsilon} \right\|_{L^2}^2 + \left\| \frac{f^n}{\varepsilon} \right\|_{L^2}^2 \right. \\
&\qquad \left. + \|u^0\|_{L^2}^2 + \|u^1\|_{L^2}^2 + \left\| \frac{f^1}{\varepsilon} \right\|_{L^2}^2 + \left\| \frac{f^2}{\varepsilon} \right\|_{L^2}^2 \right) \\
&+ \gamma \frac{2\Delta t}{\varepsilon} \Delta t \sum_{k=1}^{n-2} \frac{1}{2} \left(\left\| \frac{f^k - f^{k+2}}{2\Delta t} \right\|_{L^2}^2 + \|u^k\|_{L^2}^2 \right).
\end{aligned}$$

By Lemma 3.1, and as $\gamma \Delta t = \varepsilon \delta^2$, we obtain for $n\Delta t \le T$

$$\begin{aligned}
E(u^n, u^{n+1}) \le \left(1 + \frac{\varepsilon\delta^2}{2}\right) E(u^0, u^1) &+ \frac{\varepsilon\delta^2}{2} E(u^n, u^{n+1}) + \Delta t \sum_{k=1}^{n} E(u^{k-1}, u^k) \\
&+ \left(T + \frac{\varepsilon\delta^2}{2}\right) \max_{k=1,\ldots,n} \left\| \frac{f^k}{\varepsilon} \right\|_{L^2}^2 + T\,\delta^2 \max_{k=1,\ldots,n-2} \left\| \frac{f^k - f^{k+2}}{2\Delta t} \right\|_{L^2}^2.
\end{aligned}$$

Assuming that Δt is small enough so $\delta = \Delta t/\Delta x$ remains bounded, we have by the discrete Gronwall lemma

$$E(u^n, u^{n+1}) \le CE(u^0, u^1) + C \max_{k=1,\dots,n-1} \left\| \frac{f^k}{\varepsilon} \right\|_{L^2}^2 + C \max_{k=1,\dots,n-3} \left\| \frac{f^k - f^{k+2}}{2\Delta t} \right\|_{L^2}^2, \tag{3.14}$$

with $C = C(T)$.

THEOREM 3.1. *Let the exact solution u^ε of* (1.1) *satisfy* (1.7) *and let $\tilde{u}_I^\varepsilon$ be chosen as in* (3.10). *Then*

$$\|u^n - u^\varepsilon(n\Delta t)\|_{L^2(\mathbb{R})}^2 = O\left(\frac{\Delta t^2}{\varepsilon^3}\right) + O\left(\frac{\Delta x^2}{\varepsilon^3}\right) + O\left(\frac{\delta^2}{\varepsilon}\right). \tag{3.15}$$

Proof. Inserting the local discretization error obtained in (1.8) as inhomogeneity into (3.6), we obtain from (3.14):

$$\begin{aligned} &E(u^n - u^\varepsilon(n\Delta t), u^{n+1} - u^\varepsilon((n+1)\Delta t)) \\ &\le\ C\ E(u^0 - u_I^\varepsilon, u^1 - u^\varepsilon(\Delta t)) \\ &+ C\left(\frac{\Delta t^2}{\varepsilon^3} + \frac{\Delta x^2}{\varepsilon^3} + \frac{1}{\varepsilon}\left(\frac{\Delta t}{\Delta x}\right)^2 + \frac{\Delta t^2}{\varepsilon^3}\left(\frac{\Delta t}{\Delta x}\right)^2\right), \end{aligned} \tag{3.16}$$

and the statement then follows from Lemma 3.1 and (3.12). ∎

The estimate (3.15) implies the same bound for the errors of all observables $|E_a^\varepsilon[u^\varepsilon(n\Delta t)] - E_a^\varepsilon[u^n]|$, $a \in S$. We see that in order to obtain $L^2(\mathbb{R})$-convergence of the ("discrete") solution of the Dufort–Frankel scheme to the ("continuous") solution of the Schrödinger equation by Theorem 3.1 we need

$$\frac{\Delta t^2}{\varepsilon^3} \to 0, \quad \frac{\Delta x^2}{\varepsilon^3} \to 0, \quad \frac{\delta^2}{\varepsilon} \to 0. \tag{3.17}$$

4. Calculation of the Wigner measure of the Dufort–Frankel scheme. In this section we take a different approach to analyzing the convergence behavior of the Dufort–Frankel scheme and of the associated observables. We shall calculate its Wigner measure, as we already did with the modified equation (2.1) in section 2, and determine conditions on the mesh such that the Wigner measure of the Dufort–Frankel scheme is identical to the Wigner measure of the Schrödinger equation. By this method we will see that weaker conditions on the grid than those determined in section 3 suffice to obtain accurate observables.

To apply the Wigner measure approach to the solution of the scheme, we need uniform (in ε) L^2-stability of the scheme. We assume from now on that $\frac{\Delta t}{\Delta x} \to \delta_0 < \infty$ as $\Delta t \to 0$ and $\Delta x \to 0$ and choose (3.10) as second initial layer $\tilde{u}_I^\varepsilon$. Under these conditions the (uniform) L^2-stability of the scheme is provided by Proposition 3.1.

We consider the scheme in the form (1.9). To be able to apply Lemma 2.2, we need to rewrite the scheme in a pseudodifferential way. We formulate the x-direction part of the finite difference operator as a discrete Weyl operator, which is, according to (2.12), generated by the symbol

$$Q_{\Delta x,\varepsilon}(x,\xi) := -\frac{1}{2}\left(\frac{\varepsilon}{\Delta x}\right)^2 \left(e^{i\frac{\Delta x}{\varepsilon}\xi} - 2 + e^{-i\frac{\Delta x}{\varepsilon}\xi}\right) + V(x).$$

With this definition (and extending u_j^n to $\mathbb{R}$ as in the previous chapter), the scheme (1.9) reads

$$\begin{aligned}\varepsilon\frac{u_\sigma^{n+1}(x)-u_\sigma^{n-1}(x)}{2\Delta t} =& -iQ_{\Delta x,\varepsilon}\left(x,\varepsilon D\right)^W u_\sigma^n(x)\\ &-\frac{i}{2}\left(\varepsilon^2\delta^2+V\Delta t^2\right)\frac{u_\sigma^{n+1}(x)-2u_\sigma^n(x)+u_\sigma^{n-1}(x)}{\Delta t^2},\end{aligned}$$

$$x\in\mathbb{R},\qquad n\in\mathbb{Z}, \tag{4.1}$$

$$u_\sigma^0(x)=u_\sigma^I(x),\qquad x\in\mathbb{R}, \tag{4.2}$$

$$u_\sigma^1(x)=\tilde{u}_\sigma^I(x),\qquad x\in\mathbb{R}. \tag{4.3}$$

Here (and in what follows) we denote by σ the vector of the small parameters on which the solution depends: $\sigma:=(\varepsilon,\Delta x,\Delta t)$. The Weyl operator associated with the symbol $Q_{\Delta x,\varepsilon}(x,\xi)$ is

$$Q_{\Delta x,\varepsilon}(x,\varepsilon D)^W\varphi=-\frac{\varepsilon^2}{2}\frac{\varphi(x+\Delta x)-2\varphi(x)+\varphi(x-\Delta x)}{\Delta x^2}+V(x)\varphi(x).$$

It was shown in [MPP] that

$$Q_{\Delta x,\varepsilon}\psi\overset{\varepsilon,\Delta x\to 0}{\longrightarrow}Q\psi\quad\forall\,\psi\in S(\mathbb{R}_x^d\times\mathbb{R}_\xi^d) \tag{4.4}$$

holds if and only if $\Delta x/\varepsilon\to 0$, where $Q(x,\xi)=\frac{1}{2}|\xi|^2+V$ is the generating symbol of the Weyl operator of the continuous equation (1.1). If $\Delta x/\varepsilon\to 1/\rho$, for some $\rho>0$, then

$$Q_{\Delta x,\varepsilon}\to Q_\rho:=-\frac{1}{2}\rho^2(e^{-i\xi/\rho}-2+e^{i\xi/\rho})+V(x),$$

in the sense of (4.4), independently of ε and Δx. In the case $\Delta x/\varepsilon\to\infty$, the limit of $Q_{\Delta x,\varepsilon}$ does not approximate Q in any way and therefore no reasonable numerical results can be expected in this case, which will not be investigated further. For a detailed review on discrete pseudodifferential operators we refer to [MP], [M1], [M2], [M3].

We use the Wigner transform (2.10) and define, for $n,m\in\mathbb{Z}$, $w^{n,m}:=w^\varepsilon(u_\sigma^n,u_\sigma^m)$ and $w^n:=w^\varepsilon(u_\sigma^n,u_\sigma^n)$. In order to obtain the evolution equation for the Wigner transform, we observe that

$$\begin{aligned}&\varepsilon\frac{w^{n+1}-w^{n-1}}{2\Delta t}\\ &=w^\varepsilon\left(\varepsilon\frac{u_\sigma^{n+1}-u_\sigma^{n-1}}{2\Delta t},u_\sigma^{n+1}\right)+w^\varepsilon\left(u_\sigma^{n-1},\varepsilon\frac{u_\sigma^{n+1}-u_\sigma^{n-1}}{2\Delta t}\right).\end{aligned} \tag{4.5}$$

Using the identity (4.1), and applying Lemma 2.2, we obtain

$$\begin{aligned}&\frac{w^{n+1}-w^{n-1}}{2\Delta t}\\ &=-i\left(\frac{\Delta t}{\varepsilon}Q_{\Delta x,\varepsilon}-\eta\right)\frac{w^{n,n+1}-w^{n-1,n}}{\Delta t}-\frac{1}{2}\{Q_{\Delta x,\varepsilon},w^{n,n+1}+w^{n-1,n}\}\\ &\quad-i\eta\frac{w^{n+1}-w^{n-1}}{2\Delta t}\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{4}\{V, w^{n+1} + w^{n-1} - 2(w^{n,n+1} + w^{n-1,n})\} + O(\varepsilon) \\
&- \frac{1}{2}\Bigg\{V, \frac{1-\eta^2+2i\eta}{1+\eta^2} w^{n-1} \\
&\quad + \frac{2i-2\eta}{1+\eta^2}\Big(\Big(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\Big) w^{n-1,n} - i\frac{\Delta t}{2}\{Q_{\Delta x,\varepsilon} - V, w^{n-1,n}\}\Big) \\
&\quad - \Delta t \frac{1-\eta^2+2i\eta}{(1+\eta^2)^2}\Bigg(\{V, w^{n-1}\} \\
&\quad + i\left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) \partial_x V \left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) \partial_\xi w^{n-1,n}\Bigg)\Bigg\},
\end{aligned}
$$

where we denoted

$$
\eta = \eta(x) := \gamma + \frac{\Delta t}{\varepsilon} V(x) = \frac{\varepsilon}{\Delta t}\delta^2 + \frac{\Delta t}{\varepsilon} V(x)
$$

and used (4.1) to express $w^{n-1,n+1}$. Now we set $a^n := \operatorname{Im} w^{n,n+1}$ and $b^n := \operatorname{Re} w^{n,n+1}$, and obtain as the real part of the above equation

$$
\begin{aligned}
\frac{w^{n+1} - w^{n-1}}{2\Delta t} &= \left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) \frac{a^n - a^{n-1}}{\Delta t} \\
&\quad - \frac{1}{2}\{Q_{\Delta x,\varepsilon} - V, b^n + b^{n-1}\} - \frac{1}{4}\left\{V, w^{n+1} + w^{n-1}\right\} \\
&\quad - \frac{1}{2}\Bigg\{V, \frac{2}{1+\eta^2}\Bigg(\frac{1-\eta^2}{2} w^{n-1} \\
&\qquad - \eta\left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) b^{n-1} - \left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) a^{n-1} \\
&\qquad - \frac{\Delta t}{2}\left(\eta\{Q_{\Delta x,\varepsilon} - V, a^{n-1}\} + \{Q_{\Delta x,\varepsilon} - V, b^{n-1}\}\right)\Bigg) \\
&\qquad - \frac{\Delta t}{(1+\eta^2)^2}\Bigg((1-\eta^2)\{V, w^{n-1}\} \\
&\qquad - (1-\eta^2)\partial_x V \left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) \partial_\xi a^{n-1} \\
&\qquad - 2\eta\, \partial_x V \left(\frac{\Delta t}{\varepsilon} Q_{\Delta x,\varepsilon} - \eta\right) \partial_\xi b^{n-1}\Bigg)\Bigg\} + O(\varepsilon).
\end{aligned} \tag{4.6}
$$

We denote

$$
W^0(t) := \text{w–}\lim_{\sigma\to 0}\ w^\varepsilon(u^n_\sigma, u^n_\sigma)
$$

for $t = t_n$ fixed (where the limit is understood to hold for appropriate subsequences) and denote by $w^0(t)$ the Wigner measure of the Schrödinger equation (1.1), (1.2) (in one dimension):

$$
w^0(t) := \text{w–}\lim_{\sigma\to 0}\ w^\varepsilon(u^\varepsilon(t), u^\varepsilon(t)).
$$

THEOREM 4.1. *For the scheme* (4.1) *let* $\sigma \to 0$, $\frac{\Delta x}{\varepsilon} \to 1/\rho_0 < \infty$, *and* $\frac{\Delta t}{\Delta x} \to \delta_0 < \infty$. *Furthermore, let* $\gamma = \frac{\varepsilon \Delta t}{\Delta x^2} \to \gamma_0 < \infty$ *and set* $\tilde{u}^I_\sigma = \tilde{u}^\varepsilon_I$ *with definition* (3.10).

Case 1. $\Delta t/\varepsilon \to 0$.

Also let $\Delta x/\varepsilon \to 0$. *Then any Wigner measure* W^0 *of scheme* (4.1)–(4.3) *satisfies*

$$W^0_t + \{Q, W^0\} = 0, \qquad t \in \mathbb{R}, \tag{4.7}$$

$$W^0(t=0) = w^0[u^\varepsilon_I].$$

So the Wigner measure W^0 *of scheme* (4.1) *is the same as the Wigner measure* w^0 *of the Schrödinger equation* (1.1), (1.2).

If, on the other hand, $\Delta x/\varepsilon \to 1/\rho_0$ *for some* $0 < \rho_0 < \infty$, *then* W^0 *satisfies the above equation with* Q *replaced by* Q_{ρ_0}. *Thus the Wigner measure of the scheme is different from the Wigner measure of the continuous equation.*

Case 2. $\Delta t/\varepsilon \to \omega_0 \in (0, \infty)$.

Then there are initial data $u^\varepsilon_I \in L^2(\mathbb{R})$, *uniformly as* $\varepsilon \to 0$, *such that the Wigner measure* W^0 *of the scheme is different from the Wigner measure* w^0 *of* (1.1), (1.2).

For the proof of Theorem 4.1 we shall use the following lemma.

LEMMA 4.1. *If* $\gamma \to \gamma_0 \in [0, \infty)$ *and* $\frac{\Delta t}{\varepsilon} \to 0$, *then* $\varepsilon \frac{u^{n+1}_\sigma - u^n_\sigma}{\Delta t}$ *is bounded in* $L^2(\mathbb{R})$, *uniformly in* n, *as* $\sigma \to 0$.

Proof of Lemma 4.1. Define $v^n := \varepsilon \frac{u^{n+1}_\sigma - u^n_\sigma}{\Delta t}$. Since v^n satisfies (1.3), according to Lemma 3.1, (3.7), and (3.8)

$$\|v^n\|_{L^2} \le E(v^n, v^{n+1}) = E(v^0, v^1) \le K(\gamma)(\|v^0\|_{L^2} + \|v^1\|_{L^2}),$$

where K is bounded on compact subsets of $[0, \infty)$. We have

$$v^0 = \varepsilon \frac{u^1_\sigma - u^0_\sigma}{\Delta t} = \varepsilon \frac{\frac{u^\varepsilon_I(x+\Delta x) + u^\varepsilon_I(x-\Delta x)}{2} - u^\varepsilon_I(x)}{\Delta t}$$

by (3.10), so $\|v^0\|^2_{L^2} \le C \frac{1}{2} \frac{\varepsilon}{\Delta t} \frac{(\Delta x)^2}{\varepsilon^2} = C/\gamma$ by assumption (3.11) on u^ε_I. For v^1 we find

$$v^1 = \varepsilon \frac{u^2_\sigma - u^1_\sigma}{\Delta t} = 2\varepsilon \frac{u^2_\sigma - u^0_\sigma}{2\Delta t} + 2\varepsilon \frac{u^0_\sigma - u^1_\sigma}{2\Delta t}. \tag{4.8}$$

The second term on the right-hand side is bounded (in L^2) by C/γ, by the argument above. For the first term, using the scheme gives

$$\begin{aligned} \varepsilon \frac{u^2_\sigma - u^0_\sigma}{2\Delta t} &= i \frac{\varepsilon^2}{2} \frac{u^1_\sigma(x+\Delta x) - (u^2_\sigma(x) + u^0_\sigma(x)) + u^1_\sigma(x-\Delta x)}{(\Delta x)^2} \\ &\quad - iV(x) \frac{u^2_\sigma(x) + u^0_\sigma(x)}{2} \\ &= A(x) - i \frac{\varepsilon^2}{2} \frac{u^2_\sigma - u^0_\sigma}{\Delta x^2} - iV(x) \frac{u^2_\sigma - u^0_\sigma}{2}, \end{aligned}$$

with $A(x) := i \frac{\varepsilon^2}{4} \frac{u^0_\sigma(x+2\Delta x) - 2u^0_\sigma(x) + u^0_\sigma(x-2\Delta x)}{(\Delta x)^2} - iV(x) u^0_\sigma(x)$ bounded in L^2. The above equation is equivalent to

$$\varepsilon \frac{u^2_\sigma - u^0_\sigma}{2\Delta t} \left(1 + i\varepsilon \frac{\Delta t}{\Delta x^2} + i \frac{\Delta t}{\varepsilon} V(x)\right) = A(x).$$

So, by (4.8), $\|v^1\|_{L^2}$ is bounded uniformly and the result follows. ∎

Proof of Theorem 4.1. Since the scheme is uniformly (in ε) L^2-stable under the given conditions, the measures

$$\begin{aligned} W^0(t) &:= \text{w–}\lim_{\sigma\to 0}\ w^\varepsilon(u_\sigma^n, u_\sigma^n),\\ A^0(t) &:= \text{w–}\lim_{\sigma\to 0}\ \operatorname{Im}\{w^\varepsilon(u_\sigma^n, u_\sigma^{n+1})\},\\ B^0(t) &:= \text{w–}\lim_{\sigma\to 0}\ \operatorname{Re}\{w^\varepsilon(u_\sigma^n, u_\sigma^{n+1})\} \end{aligned}$$

(for $t = t_n$ fixed) exist, after selection of a subsequence. Assume $\Delta x/\varepsilon \to 0$. By (4.4), we then have $Q_{\Delta x,\varepsilon}(x,\xi) \to Q(x,\xi)$.

Case 1. Letting $\sigma \to 0$ in (4.6) we obtain

$$W_t^0 = -\gamma_0 A_t^0 - \{Q, B^0\} + \frac{1}{1+\gamma_0^2}\{V, B^0 - W^0 - \gamma_0 A^0\} \tag{4.9}$$

for some $\gamma_0 \in [0,\infty)$. Now observe that

$$w^\varepsilon(u_\sigma^n, u_\sigma^{n+1}) = w^\varepsilon(u_\sigma^n, u_\sigma^n) + \frac{\Delta t}{\varepsilon} w^\varepsilon\left(u_\sigma^n, \varepsilon\frac{u_\sigma^{n+1} - u_\sigma^n}{\Delta t}\right). \tag{4.10}$$

By Lemma 4.1 and (2.21), we conclude that

$$w^\varepsilon(u_\sigma^n, u_\sigma^{n+1}) - w^\varepsilon(u_\sigma^n, u_\sigma^n) \rightharpoonup 0 \tag{4.11}$$

in the case $\frac{\Delta t}{\varepsilon} \to 0$, and thus

$$B^0 \equiv W^0, \qquad A^0 \equiv 0.$$

Then from (4.9) we conclude (4.7). Also, by the choice (3.10) for $\tilde u_\sigma^I$, we have $\lim_{\sigma\to 0} w^0(u_I^\varepsilon, \tilde u_\sigma^I) = w^0[u_I^\varepsilon]$, which is real-valued, and so $B_I^0 = w^0[u_I^\varepsilon]$ and $A_I^0 = 0$.

In the case $\Delta x/\varepsilon \to 1/\rho_0$, we have $Q_{\Delta x,\varepsilon}(x,\xi) \to Q_{\rho_0}(x,\xi)$, and the same arguments hold true if Q is replaced by Q_{ρ_0} in (4.9).

Case 2. Since $\gamma = \frac{\Delta t}{\varepsilon}(\frac{\varepsilon}{\Delta x})^2$, the conditions on γ and on $\Delta x/\varepsilon$ imply that $\Delta x/\varepsilon \to 1/\rho_0$ for some $\rho_0 < \infty$ such that $Q_{\Delta x,\varepsilon}(x,\xi) \to Q_{\rho_0}(x,\xi)$. Assume now that the Wigner measure W^0 of the scheme would be the same as the one of (1.1). Then it has to satisfy $B^0 \equiv W^0 \equiv w^0$, $A^0 \equiv 0$, as we saw in the previous case. Inserting this into (4.6), after letting $\sigma \to 0$ we find

$$W_t^0 + \{Q_{\rho_0}, W^0\} = \left\{V, \omega_0 \frac{\eta_0}{1+\eta_0^2} Q_{\rho_0} W^0\right\}, \qquad t \in \mathbb{R}, \tag{4.12}$$

with $\eta_0 = \gamma_0 + \omega_0 V(x)$ for $\gamma \to \gamma_0 < \infty$, which is not the correct equation. Obviously, there are data w_I^0 such that W^0 is different from w^0, which is a contradiction. ∎

Note that the case $\Delta t/\varepsilon \to \infty$ is excluded by the assumptions, since both $\Delta t/\Delta x$ and $\Delta x/\varepsilon$ have to be bounded.

W^0 is the unique Wigner measure of scheme (4.1) if the initial data u_I^ε is chosen such that the Wigner measure $w^0[u_I^\varepsilon]$ is unique.

In Case 1, for $\Delta x/\varepsilon \to 0$, the transport equation for the Wigner measure of the scheme is identical to the one for the Wigner measure of the continuous equation, which was derived in Lemma 2.3, and the initial conditions coincide. Therefore, the Wigner measure of the scheme coincides with the Wigner measure of the original

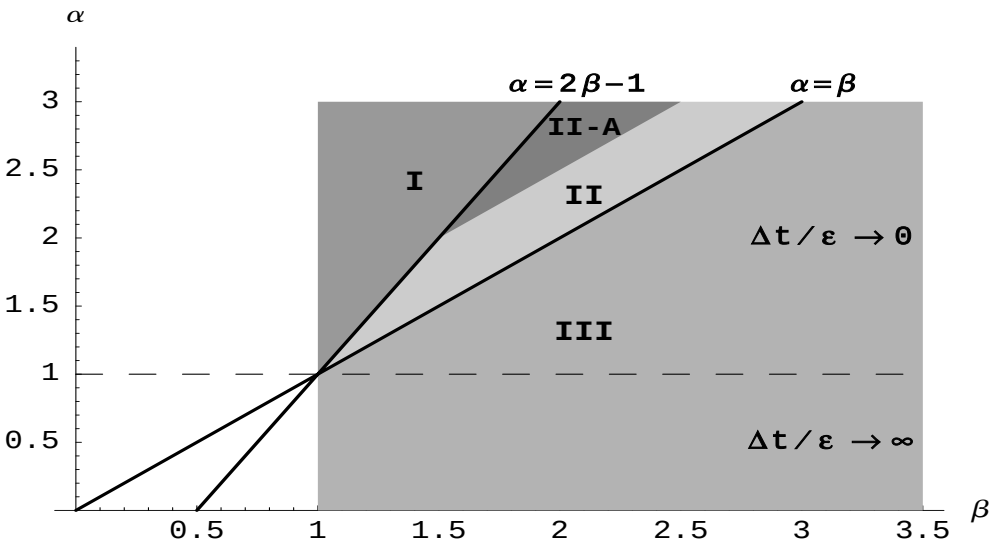


FIG. 1. *γ and δ for different exponents.*

equation. So, as a consequence of Theorem 4.1, we have convergence of the scheme under the condition that $\Delta t/\varepsilon \to 0$, $\Delta x/\varepsilon \to 0$, and $\gamma = \frac{\Delta t}{\Delta x^2}\varepsilon \to \gamma_0 < \infty$ (which also imply the stability condition $\delta < \infty$). We also have convergence of all smooth observables, since the limit of average values of observables with symbol in S can be obtained from the Wigner measure by using (2.11). These convergence conditions are much weaker than the requirements (3.17) of the "classical" consistency-stability method; however, no error estimate for the observables is provided. In all other cases, the Wigner measure is, for general initial data, different from the Wigner measure of the Schrödinger equation, and observables of the scheme are not guaranteed to converge to the exact observables.

The case $\gamma \to \infty$ is not covered by Theorem 4.1. Numerical evidence indicates that there are observables which do not converge correctly (see section 5, Figures 13–15).

To clarify the meaning of the conditions of Theorem 4.1 in terms of restrictions on the discretization steps, we consider a choice of Δt and Δx as powers of ε. So let $\Delta t = \omega_0 \varepsilon^\alpha$ and $\Delta x = \frac{1}{\rho_0}\varepsilon^\beta$, which means

$$\delta = \frac{\Delta t}{\Delta x} = \omega_0 \rho_0 \ \varepsilon^{\alpha-\beta},$$

$$\gamma = \frac{\varepsilon \Delta t}{\Delta x^2} = \omega_0 \rho_0^2 \ \varepsilon^{1+\alpha-2\beta}.$$

Figure 1 shows α versus β. We consider only the cases $\beta \geq 1$, where the "spatial convergence" (4.4) is given. The corresponding cases of Theorem 4.1 are the following:

- Region I. $\delta \to 0$ and $\gamma \to \gamma_0 < \infty$. Case 1.
- Region II. $\delta \to 0$ and $\gamma \to \infty$. No assertion in Theorem 4.1.
- Region II-A. $\delta \to 0$, $\gamma \to \infty$, and $\alpha, \beta > 1.5$, $\alpha > \beta + 0.5$. No assertion in Theorem 4.1 but convergence according to (3.17).
- Region III. $\delta \to \infty$ and $\gamma \to \infty$. No uniform L^2-stability.

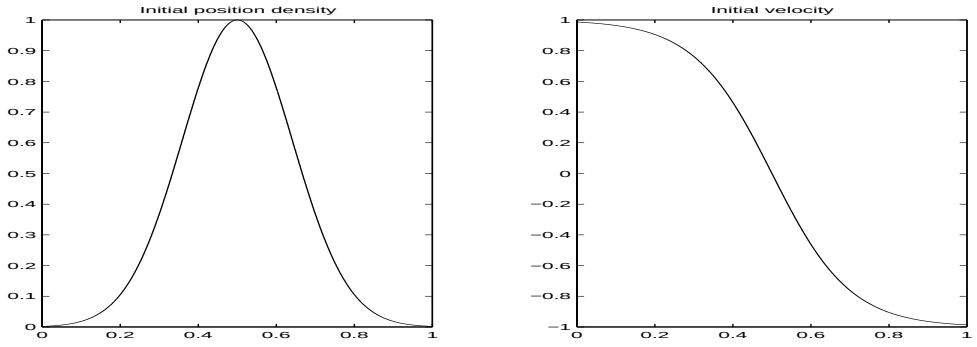


FIG. 2. *Initial condition* (5.2).

Case 2 corresponds to the point $\alpha = \beta = 1$. As already pointed out, there are choices of the spatial-temporal mesh (corresponding to Region II) which are not included in Theorem 4.1. However, among them there are cases for which the result (3.17) shows convergence (corresponding to Region II-A) and, consequently, the observables do converge correctly. Note that the mesh restrictions of (3.17) require $\alpha > 2$; that is, Δt is $O(\varepsilon^{2+r})$ for some positive r, which is computationally unaffordable.

5. Examples and numerical results. In [MPP] different numerical schemes for the linear Schrödinger equation were analyzed by the same approach as in this paper. Those are the forward and backward Euler, Crank–Nicolson, and Leap-Frog schemes. In order to be able to put our results in line with the results obtained there, we shall use the same numerical examples here as in [MPP]. We consider WKB-type initial data in one space dimension:

$$u_I^\varepsilon(x) = \sqrt{n_I(x)}\exp\Big(\frac{i}{\varepsilon}S_I(x)\Big), \qquad x \in \mathbb{R}^m, \tag{5.1}$$

with n_I and S_I real-valued and independent of ε.

We choose the following data, shown in Figure 2,

$$n_I = \big(\exp(-25(x-0.5)^2)\big)^2, \tag{5.2}$$

$$\frac{d}{dx}S_I(x) = -\tanh\left(5(x-0.5)\right) \tag{5.3}$$

on the interval $[0,1]$ (imposing periodic boundary conditions). Equation (3.10) is chosen for u_σ^1. Here the characteristics of the free Burgers equation (which is the classical limit of the velocity equation before singularities occur; cf. [GM]), given by $x(t) = v_I(s)t + s$ (where $x(0) = s \in \mathbb{R}$), intersect in finite time because the initial velocity $v_I(x) = \frac{d}{dx}S_I(x)$ is compressive. The curves which separate the areas without intersection of characteristics from the area where intersection occurs are called caustics and are given as follows (obtained from a simple calculation):

$$x_{1,2}(t) = 0.5 \pm \ln\left(\frac{\sqrt{t}+\sqrt{t-0.2}}{\sqrt{0.2}} \mp \frac{t-0.2+\sqrt{t(t-0.2)}}{t+\sqrt{t(t-0.2)}}\right),$$

emanating from the focus $t = 0.2,\ x = 0.5$.

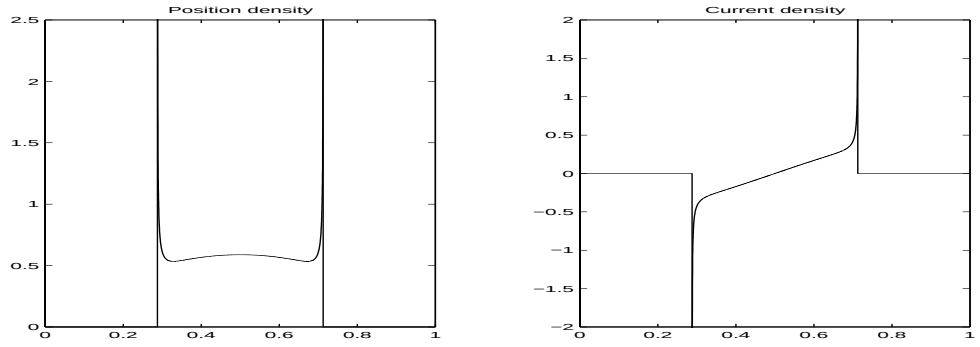


FIG. 3. n^0 *and* J^0 *at* $t = 0.54$.

Once the solution u_j^n of (1.3) has been computed, we compute its discrete position and current densities $n_\sigma,\ J_\sigma$, which are (in the one-dimensional case)

$$n_\sigma(x_j, t_n) = |u_j^n|^2, \qquad n, j \in \mathbb{Z},$$

$$J_\sigma(x_j, t_n) = \varepsilon\ \mathrm{Im}\left(\overline{u_j^n}\frac{u_{j+1}^n - u_j^n}{\Delta x}\right), \qquad n, j \in \mathbb{Z}.$$

In the continuous case, the limits of the position density (2.13) and the current density (2.14) are recovered as moments of the Wigner measure if u^ε is ε-oscillatory. (They cannot be obtained by means of (2.11), since the symbols are not in S.) Under this condition (which is defined in (2.33)), we have

$$n^\varepsilon \to n^0 := \int_{\mathbb{R}^d_\xi} w^0(x, d\xi, t)$$

and

$$J^\varepsilon \to J^0 := \int_{\mathbb{R}^d_\xi} \xi\ w^0(x, d\xi, t).$$

Again we refer to [GMMP] for more details. The ε-oscillatory condition is satisfied by the initial data (5.2)–(5.3). Although in general this property is not preserved by the Dufort–Frankel scheme, it is preserved in the constant coefficient case (which is the case in the examples presented). Thus, in this case we have $n_\sigma \overset{\sigma\to 0}{\longrightarrow} n^0$ and $J_\sigma \overset{\sigma\to 0}{\longrightarrow} J^0$, provided that the Wigner measures are identical.

n^0 and J^0 are L^1_{loc} functions, assuming infinite value on the caustics. They are shown in Figure 3. For $\varepsilon > 0$, the continuous observables n^ε and J^ε are oscillating with wavelength $\mathrm{O}(\varepsilon)$ in those areas where two or more characteristics intersect.

The following pictures show the computed densities n_σ and J_σ at time $t = 0.54$, i.e., after the caustics develop. For reference purposes the weak limits n^0 and J^0 are also depicted, using dashed lines.

Figures 4–6 refer to Case 1, $\Delta t/\varepsilon \to 0$, with $\Delta x/\varepsilon \to 0$ and $\gamma \to 0$, for $\varepsilon = 2\cdot 10^{-3}$, $1\cdot 10^{-3}$, and $0.5\cdot 10^{-3}$, respectively. The discretization parameters are $\Delta x = \varepsilon^{1.2}, \Delta t = \varepsilon^{1.5}$. We set $V = 0$. In this case the transport equations for the Wigner measure of the difference scheme and of the continuous problem coincide. The obtained solutions

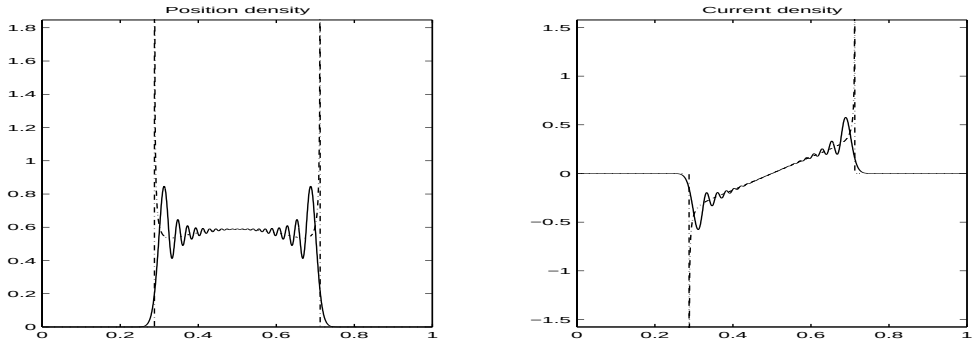


FIG. 4. $\varepsilon = 2 \cdot 10^{-3}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.155$, $\gamma = 0.54$, $V = 0$.

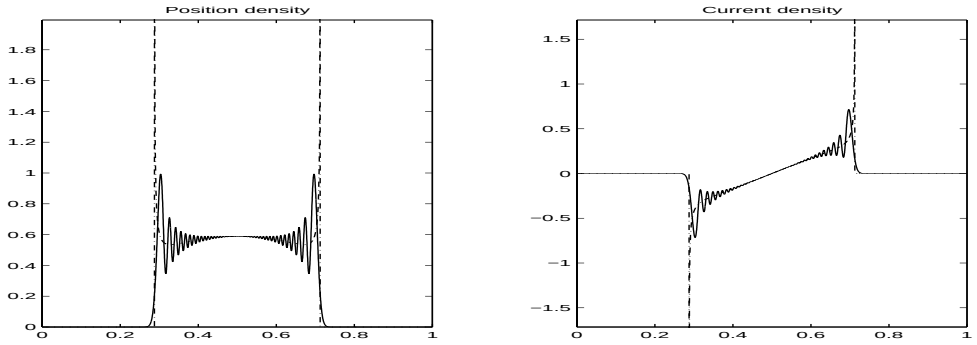


FIG. 5. $\varepsilon = 10^{-3}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.126$, $\gamma = 0.50$, $V = 0$.

oscillate around n^0 and J^0 with wavelength ε. It can be observed that the number of oscillations doubles when ε is halved. The amplitude of the oscillations does not grow (except for the first one). There is evidence that the sequences $\{n_\sigma\}_\sigma$ and $\{J_\sigma\}_\sigma$ weakly converge to n^0 and J^0. One can say that these solutions are good approximations of n^ε and J^ε for the selected ε. We remark that the constraints of the consistency-stability analysis are not satisfied, and this choice of parameters would be discarded, according to (3.17).

Figures 7–9 correspond to Case 1 with a constant γ. The parameters are $\varepsilon = 4 \cdot 10^{-4}$, $2 \cdot 10^{-4}$, and 10^{-4}, with $\Delta x = \varepsilon^{1.2}$, $\Delta t = \varepsilon^{1.4}$, and $V = 0$. So we have $\gamma = \gamma_0 = 1$, and the Wigner measure of the scheme is still correct. The figures indicate that, as in the previous example, the sequences $\{n_\sigma\}_\sigma$ and $\{J_\sigma\}_\sigma$ weakly converge to n^0 and J^0.

Figures 10–12 show Case 1 in the situation $\Delta x/\varepsilon \to 1/\rho_0 > 0$. Here the Wigner measure of the scheme is different from the continuous one. n_σ and J_σ converge to functions with smaller support than n^0 and J^0. At fixed ε the amplitude of oscillations is larger than in the example of Figures 4–6. In this case, there is no convergence to n^0 and J^0, and n_σ, J_σ are poor approximations of n^ε and J^ε for the corresponding fixed ε.

Figures 13–15 explore the case not covered in Theorem 4.1. The consistency condition $\delta \to 0$ is satisfied, but the condition $\gamma \to \gamma_0 < \infty$ is violated. The parameters are $\varepsilon = 8 \cdot 10^{-4}$, $4 \cdot 10^{-4}$, and $2 \cdot 10^{-4}$ respectively, and $\Delta x = \varepsilon^{1.35}$, $\Delta t = \varepsilon^{1.4}$ ($\gamma = \varepsilon^{-0.3}$). The results show a smaller support than the continuous limits, similarly

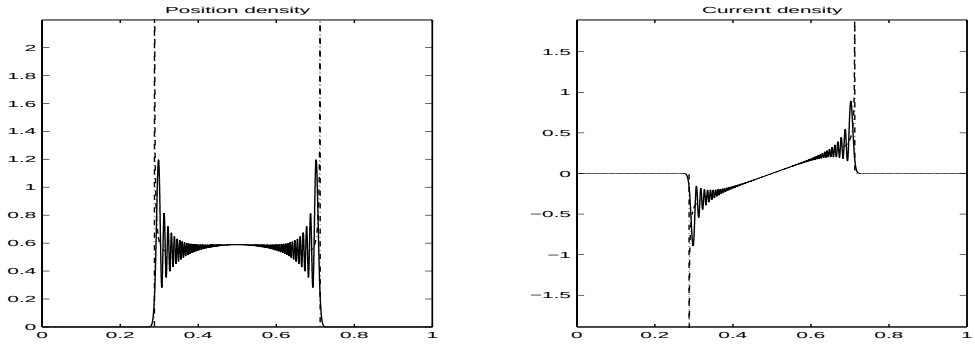


FIG. 6. $\varepsilon = 0.5 \cdot 10^{-3}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.102$, $\gamma = 0.47$, $V = 0$.

to the previous example, and there is no convergence of n_σ and J_σ to n^0 and J^0.

The results of Theorem 4.1, being asymptotic statements, are valid only for "sufficiently small" ε. Statements like "$\Delta t/\varepsilon \to 0$" must be interpreted as "$\Delta t/\varepsilon$ small" when applied to a fixed ε. Since, in contrast to (1.1), the numerical scheme is not gauge invariant, the choice of those Δx and Δt which guarantee a good approximation depends on V.

Figures 16 and 17 show a situation of Case 1 with a positive constant potential V. We show the results for $\varepsilon = 10^{-3}$ and $\varepsilon = 5 \cdot 10^{-4}$ with the discretization parameters $\Delta t = \varepsilon^{1.3}$ and $\Delta x = \varepsilon^{1.1}$, and $V = 3$. According to the relations of the discretization parameters, the conditions for the correct Wigner measure $\Delta t/\varepsilon,\ \Delta x/\varepsilon \to 0$, $\gamma < \infty$ as $\varepsilon \to 0$ are satisfied. The computed densities are clearly poor approximations of n^ε, respectively, J^ε. Moreover, they also exhibit fast oscillations in time as well as in space, even before caustics develop, as shown in Figure 18, where the first time layers of n^σ for $0 < t < 4.78 \cdot 10^{-3}$ are plotted for $\varepsilon = 10^{-3}$. The oscillations are of almost the same wavelength as the discretization steps.

Figures 19–21 show results for increasing values of V. The obtained results deviate more and more with growing V from n^0 and J^0.

Compared to the Crank–Nicolson scheme, in Case 1 the results are of the same quality as in the case of convergence there. The choice of discretization parameters is less restrictive for the Crank–Nicolson scheme, since the conditions $\Delta t/\Delta x \to 0$ and $\gamma < \infty$ are not needed. That scheme also has the advantage of better conservation properties; it conserves total charge as well as a discrete version of the energy. However, these advantages have to be traded in for an implicit computation complexity, which is a serious disadvantage in performance critical problems.

In comparison to the Leap-Frog scheme, we also have the same quality of results in the convergent case. The Leap-Frog scheme has similar conservation properties, as it conserves some discrete analogon of the energy. The choice of the discretization steps is slightly more restrictive there, since there is the stability condition

$$\frac{\Delta t}{\Delta x^2}\varepsilon + \frac{\Delta t}{2\varepsilon} V_{\max} < \frac{1}{2}$$

to be satisfied, in comparison with the convergence condition $\gamma = \frac{\Delta t}{\Delta x^2}\varepsilon \to \gamma_0 < \infty$ of the Dufort–Frankel scheme.

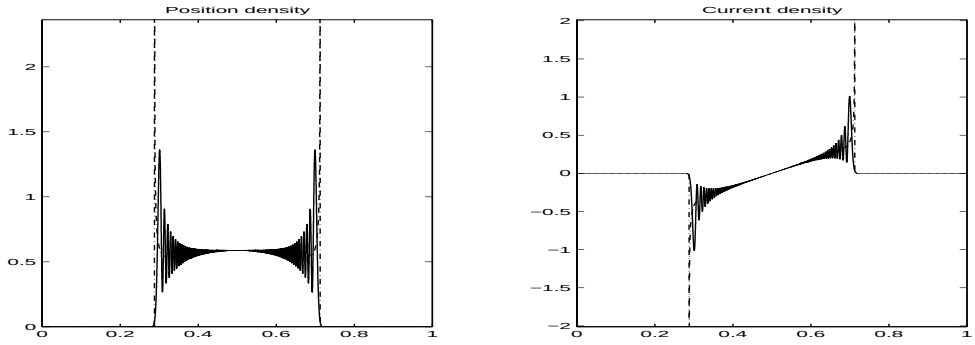


FIG. 7. $\varepsilon = 4 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.209$, $\gamma = 1$, $V = 0$.

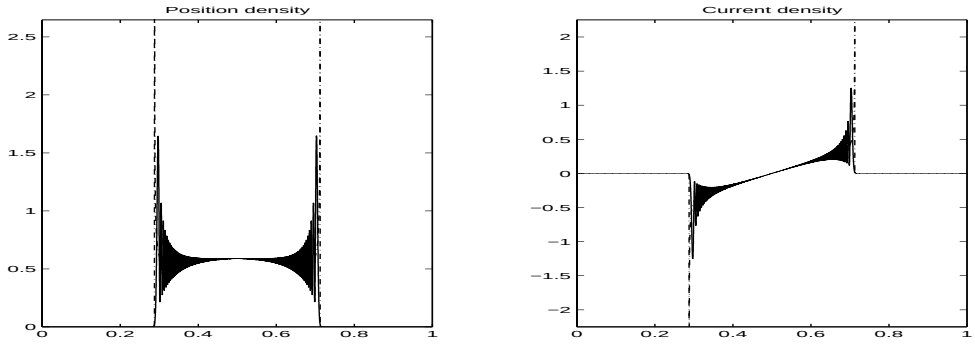


FIG. 8. $\varepsilon = 2 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.182$, $\gamma = 1$, $V = 0$.

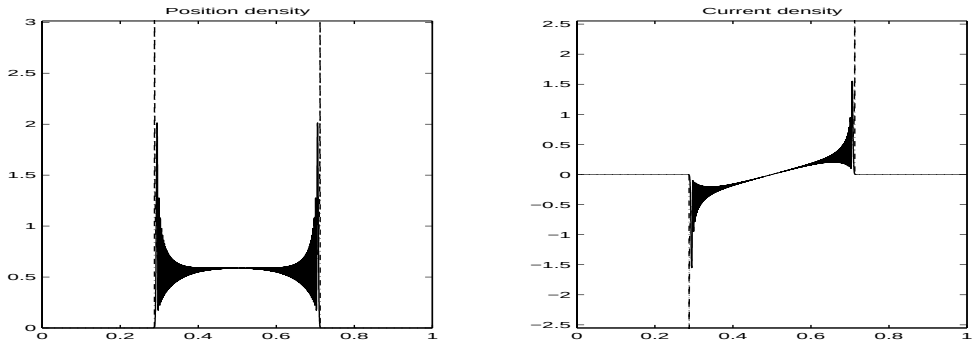


FIG. 9. $\varepsilon = 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.2}$, $\delta = 0.159$, $\gamma = 1$, $V = 0$.

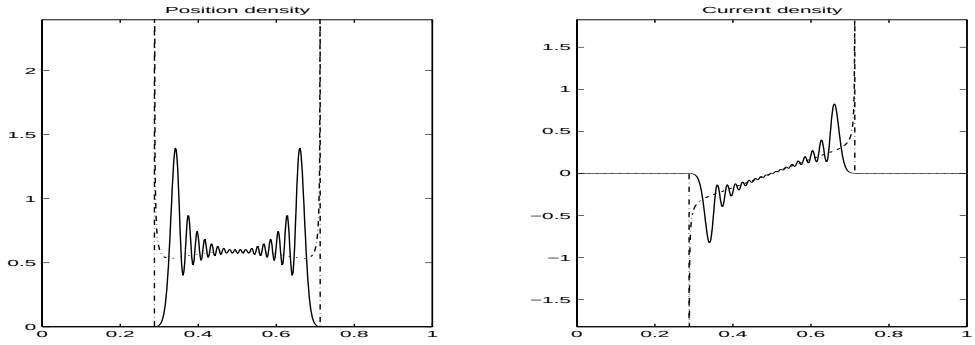


FIG. 10. $\varepsilon = 2 \cdot 10^{-3}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon$, $\delta = \gamma = 0.045$, $V = 0$.

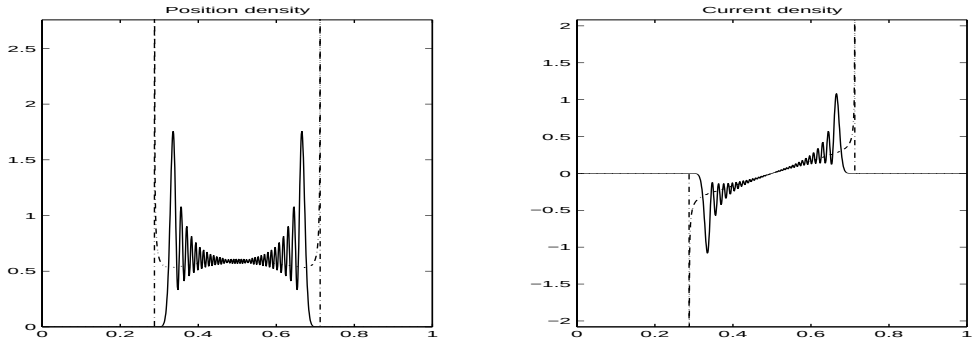


FIG. 11. $\varepsilon = 10^{-3}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon$, $\delta = \gamma = 0.032$, $V = 0$.

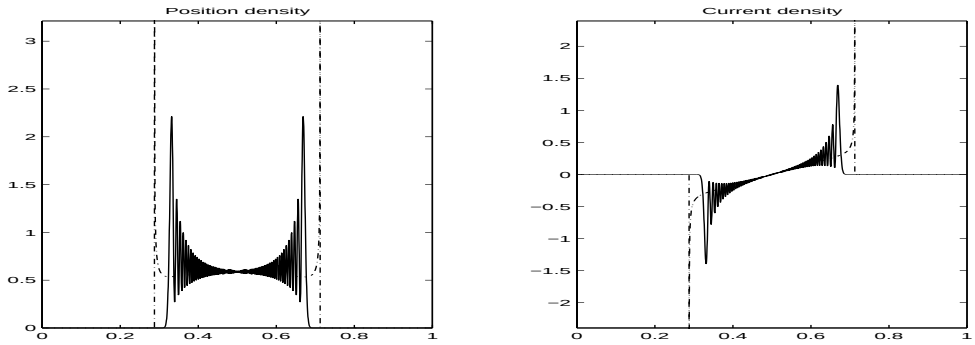


FIG. 12. $\varepsilon = 5 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.5}$, $\Delta x = \varepsilon$, $\delta = \gamma = 0.022$, $V = 0$.

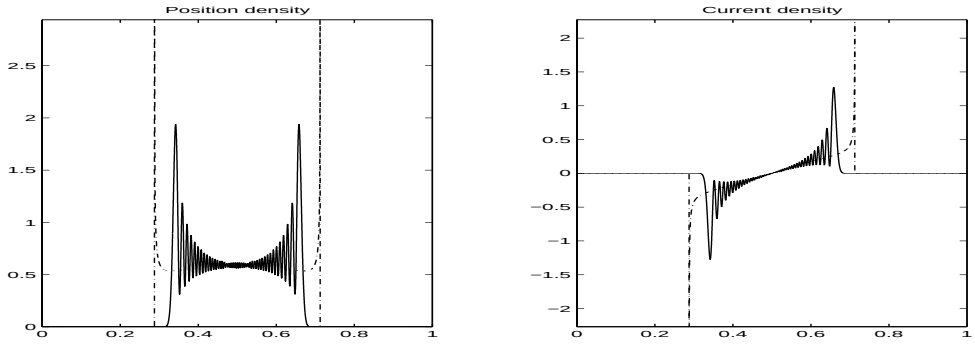


FIG. 13. $\varepsilon = 8 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.35}$, $\delta = 0.70$, $\gamma = 8.49$, $V = 0$.

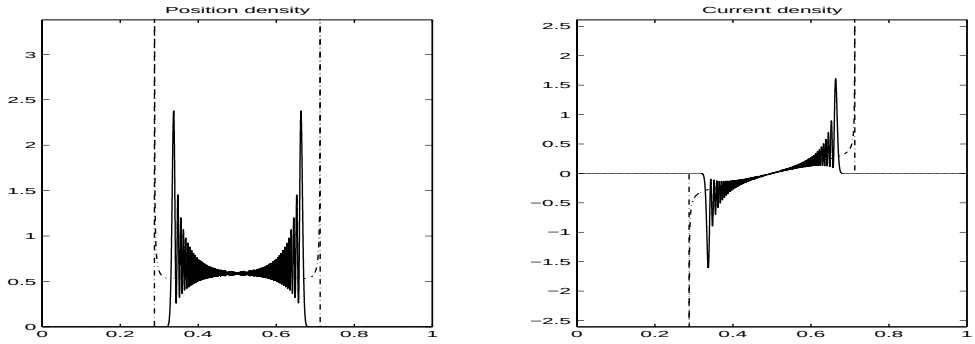


FIG. 14. $\varepsilon = 4 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.35}$, $\delta = 0.676$, $\gamma = 10.46$, $V = 0$.

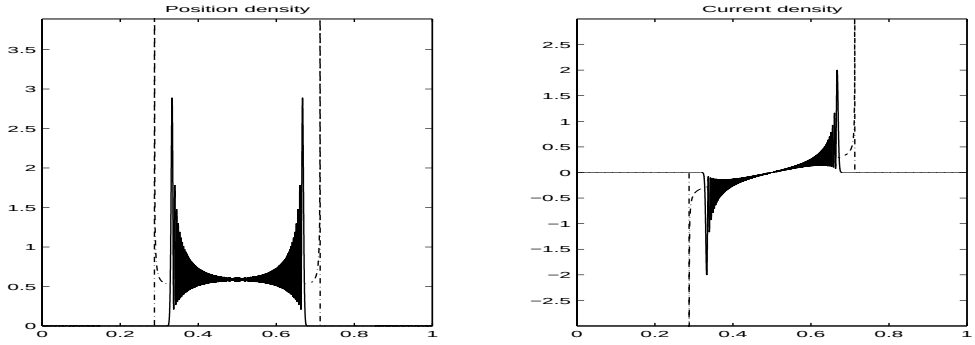


FIG. 15. $\varepsilon = 2 \cdot 10^{-4}$, $\Delta t = \varepsilon^{1.4}$, $\Delta x = \varepsilon^{1.35}$, $\delta = 0.653$, $\gamma = 12.87$, $V = 0$.

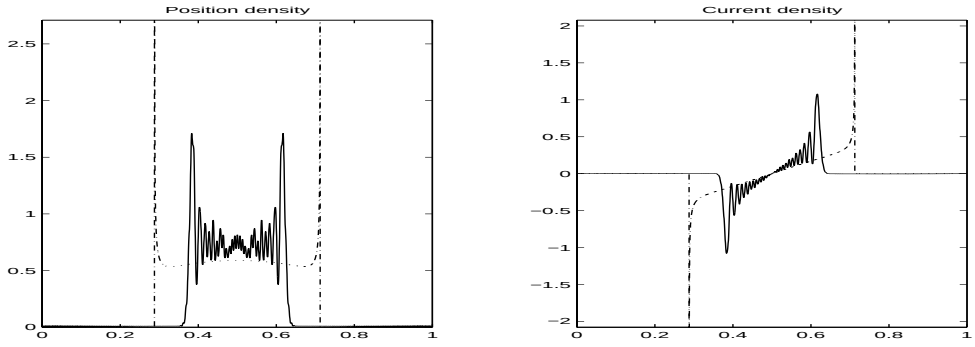


FIG. 16. $\varepsilon = 10^{-3}$, $\Delta x = \varepsilon^{1.1}$, $\Delta t = \varepsilon^{1.3}$, $\delta = 0.251$, $\gamma = 0.501$, $V = 3$.

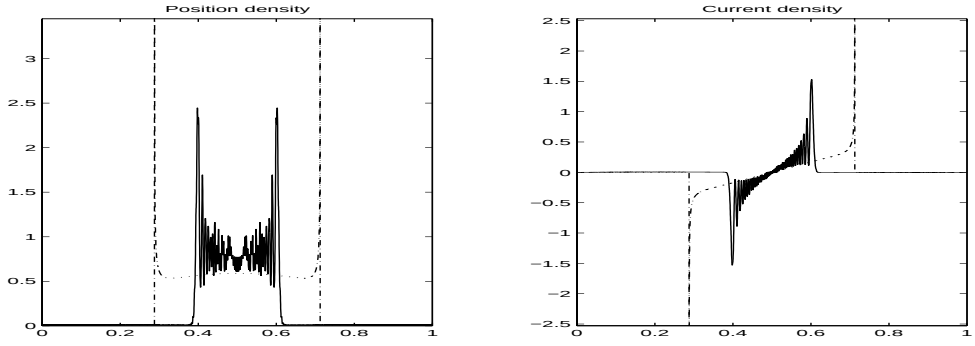


FIG. 17. $\varepsilon = 5 \cdot 10^{-4}$, $\Delta x = \varepsilon^{1.1}$, $\Delta t = \varepsilon^{1.3}$, $\delta = 0.219$, $\gamma = 0.468$, $V = 3$.

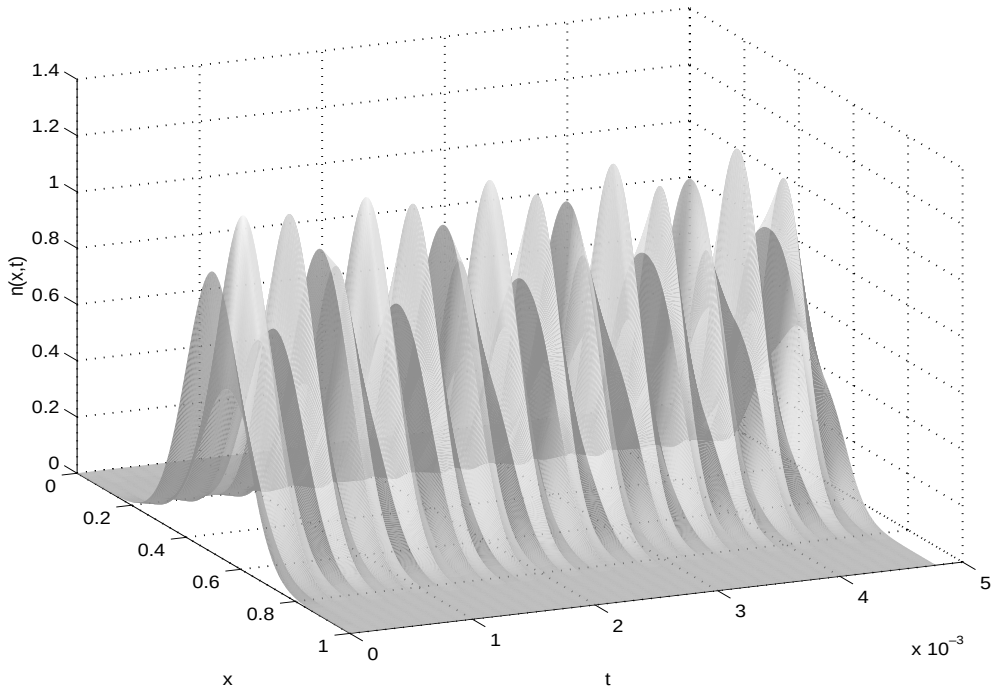


FIG. 18. $\varepsilon = 10^{-3}$, $\Delta t = \varepsilon^{1.3}$, $\Delta x = \varepsilon^{1.1}$, $V \equiv 3$, $0 < t < 4.78 \cdot 10^{-3}$.

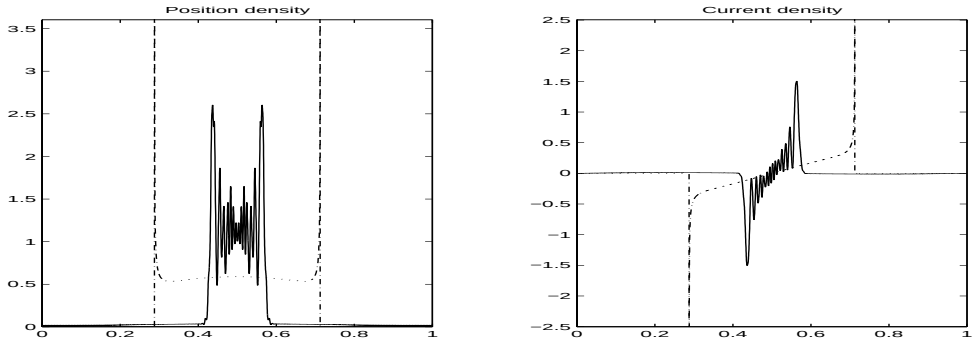


FIG. 19. $\varepsilon = 10^{-3}, \Delta x = \varepsilon^{1.2}$, $\Delta t = \varepsilon^{1.5}$, $V = 25$.

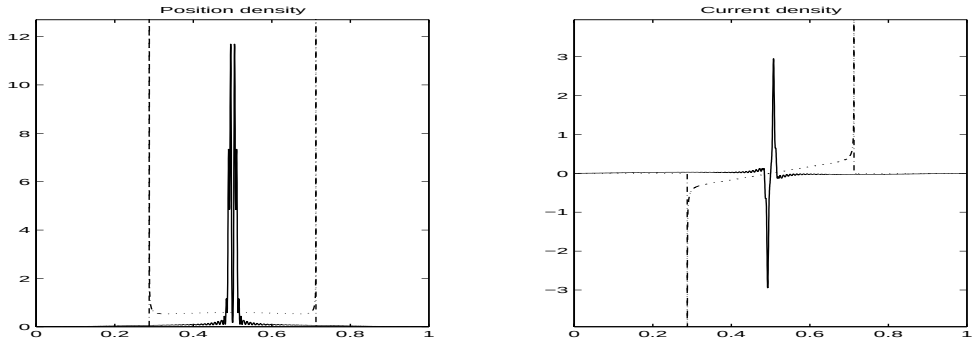


FIG. 20. $\varepsilon = 10^{-3}$, $\Delta x = \varepsilon^{1.2}$, $\Delta t = \varepsilon^{1.5}$, $V = 50$.

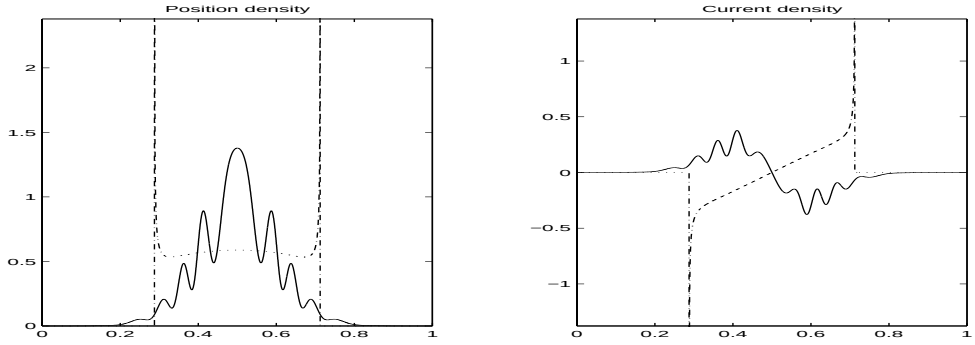


FIG. 21. $\varepsilon = 10^{-3}$, $\Delta x = \varepsilon^{1.2}$, $\Delta t = \varepsilon^{1.5}$, $V = 500$.

Acknowledgment. We want to thank the anonymous referee for providing some very helpful remarks which significantly improved the paper.

REFERENCES

[G1] P. Gérard, *Mesures semi-classiques et ondes de Bloch*, Séminaire E.D.P. École Polytechnique Palaiseau, XVI (1990/1991), pp. 1–19.

[G2] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.

[GM] I. Gasser and P A. Markowich, *Quantum hydrodynamics, Wigner transforms and the classical limit*, Asymptot. Anal., 14 (1997), pp. 97–116.

[GMMP] P. Gérard, P A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 321–377.

[GMMP2] P. Gérard, P A. Markowich, N. J. Mauser, and F. Poupaud, *Erratum: Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 53 (2000), pp. 280–281.

[HM3] L. Hörmander, *The Analysis of Linear Partial Differential Operators* III, Springer-Verlag, Berlin, 1985

[IR] F. Ivanauskas and M. Radžiūnas, *On convergence and stability of the explicit difference method for solution of nonlinear Schrödinger equations*, SIAM J. Numer. Anal., 36 (1999), pp. 1466–1481.

[JLL] S. Jin, C. D. Levermore, and D. W. McLaughlin, *The behavior of solutions of the NLS equation in the semiclassical limit*, in Singular Limits of Dispersive Waves, Plenum Press, New York, London, 1994, pp. 235–255.

[LL] L. D. Landau and E. M. Lifschitz, *Lehrbuch der Theoretischen Physik* III: *Quantenmechanik*, Akademie-Verlag, Berlin, 1985.

[LP] P. L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.

[M1] V. Maslov, *The theory of bicharacteristics for difference schemes*, Uspekhi Mat. Nauk, 23 (1968), p. 243.

[M2] V. Maslov, *Méthodes Opératorielles*, Editions Mir, Moscow, 1972.

[M3] V. Maslov, *The characteristics of pseudo-differential operators and difference schemes*, Actes Cong. Intern. Math., 2 (1970), pp. 755–769.

[MMP] P. A. Markowich, N. J. Mauser, and F. Poupaud, *A Wigner function approach to semiclassical limits: Electrons in a periodic potential*, J. Math. Phys., 35 (1994), pp. 1066–1094.

[MP] P. A. Markowich and F. Poupaud, *The pseudo-differential approach to finite differences revisited*, Calcolo, 36 (1999), pp. 161–186.

[MPP] P. A. Markowich, P. Pietra, and C. Pohl, *Numerical approximation of quadratic observables of Schrödinger-type equations in the semi-classical limit*, Numer. Math., 81 (1999), pp. 595–630.

[T] L. Tartar, *H–measures: A new approach for studying homogenization, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.

[W] L. Wu, *Dufort–Frankel-type methods for linear and nonlinear Schrödinger equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1526–1533.

[Wi] E. Wigner, *On the quantum correction for the thermodynamic equilibrium*, Phys. Rev. II, 40 (1932), pp. 749–759.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1311–1329

TOPOLOGICAL AND ε-ENTROPY FOR LARGE VOLUME LIMITS OF DISCRETIZED PARABOLIC EQUATIONS*

GABRIEL J. LORD† AND JACQUES ROUGEMONT†

Abstract. We consider semidiscrete and fully discrete approximations of nonlinear parabolic equations in the limit of unbounded domains, which by a scaling argument is equivalent to the limit of vanishing viscosity. We define the spatial density of ε-entropy, topological entropy, and dimension for the attractors and show that these quantities are bounded. We also provide practical means of computing lower bounds on them. The proof uses the property that solutions lie in Gevrey classes of analyticity, which we define in a way that does not depend on the size of the spatial domain. As a specific example we discuss the complex Ginzburg–Landau equation.

Key words. parabolic partial differential equations, large volume, entropy, dimension

AMS subject classifications. 35K55, 35B41, 37M25, 65M60

PII. S0036142901392328

1. Introduction.

We consider the following general parabolic equation:

$$\partial_t u = \nu \Delta u + \gamma u + F(u) , \qquad x \in [-L\pi, L\pi]^d , \qquad t \geq 0 , \tag{1.1}$$

for a complex valued function $u = u(x,t)$ and bounded continuous initial condition $u(x,0) = u_0(x)$. We restrict ourselves to $L \in \mathbb{N}$ for convenience. The coefficients of (1.1) satisfy

$$\nu \in \mathbb{C} , \qquad \mathrm{Re}(\nu) > 0 , \quad \gamma \in \mathbb{R} ,$$

and we assume that $\mathrm{Re}(F)$ and $\mathrm{Im}(F)$ are real analytic functions of $\mathrm{Re}(u)$ and $\mathrm{Im}(u)$.

We are interested in the large volume limit ($L \to \infty$) of the long time dynamics (in particular the attractor) of (1.1) and its approximation by numerical schemes. In the latter case we are interested in the limit when the mesh size of our discretization is kept constant while taking the limit $L \to \infty$, thereby obtaining an infinite-dimensional but still discrete system. (See section 6 for results of upper semicontinuity of the attractors in terms of the different parameters of the problem.)

We remark that, by a scaling transformation, the large volume limit can be interpreted as a small viscosity limit. The rescaled function $v(y,t) = u(Ly,t)$ with $y \in [-\pi,\pi]^d$ satisfies the following equation:

$$\partial_t v = \frac{\nu}{L^2} \Delta v + \gamma v + F(v) ,$$

with periodic boundary conditions on $[-\pi,\pi]^d$. It is, however, easier to work with (1.1) (with periodic boundary conditions) and take $L \to \infty$. Indeed, since the problem on the full space $\mathbb{R}^d$ is well posed, we have a priori bounds for all $L < \infty$. In fact, we view the periodic boundary conditions on $[-L\pi, L\pi]^d$ for large L as an approximation of the infinite volume.

*Received by the editors July 12, 2001; accepted for publication (in revised form) March 14, 2002; published electronically September 27, 2002.
http://www.siam.org/journals/sinum/40-4/39232.html

†Department of Mathematics, Heriot–Watt University, Edinburgh EH14 4AS, United Kingdom (G.J.Lord@ma.hw.ac.uk, J.Rougemont@ma.hw.ac.uk). The work of the second author was supported by the Fonds National Suisse de la Recherche Scientifique and the EPSRC GR/R29949/01.

For each fixed $L < \infty$, (1.1) generates a semiflow Φ_L^t. We discretize this time evolution spatially by truncating to a finite number of (Fourier) modes. We make this truncation by multiplying by a smooth function in Fourier space (rather than a sharp indicator function) to have better control as $L \to \infty$ (when the spectrum becomes dense). We then discretize in time using an explicit scheme inspired by [26]. This scheme is amenable to analysis and also proves to be an efficient numerical scheme for smooth initial conditions.

It is not the purpose of this paper to prove the existence of global attractors for (1.1) or for the discretizations; this has been considered in different setups in a large number of publications (see, for example, [27, 24, 3, 1, 29]). Instead, we assume the existence of a semiflow and of a family of global attractors, $\widehat{\mathcal{A}}(L)$, for the continuous and discrete problems (see Definition 3.2).

We compute bounds on statistical quantities that are valid for both the discrete and continuous systems. The first of these statistical quantities is the (Kolmogorov) ε-entropy

$$H_\varepsilon := \limsup_{L\to\infty} \frac{\log \mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L))}{(2L\pi)^d} ,$$

where $\mathcal{N}$ is the minimum number of balls of radius ε in the topology of L^∞ that are needed to cover the attractor $\widehat{\mathcal{A}}(L)$ (see Definition 3.3). We prove that H_ε is a finite number in Theorem 4.3. We thereby get a bound on the upper density of dimension

$$d_{\text{up}} = \limsup_{\varepsilon\to 0} \frac{H_\varepsilon}{\log \varepsilon^{-1}} .$$

This is to be compared with the results of Kolmogorov and Tikhomirov [15], where they obtain a bound of the same type for the set of all entire analytic functions of exponential type. This result follows from a sampling result for such functions (Proposition D.3); namely, any of these analytic functions can be reconstructed by interpolation of a discrete set of values. Although the functions on $\widehat{\mathcal{A}}$ are not entire functions, they are still determined by a discrete sampling.

Remark that it is appropriate to take the L^∞ topology, since the diameter of $\widehat{\mathcal{A}}(L)$ does not depend on L in this topology, unlike the topology of Sobolev spaces of nonzero order. We remark that the L^∞ topology is stronger than the L^2 topology, and hence our results do not follow from [9, 8, 29].

We also wish to emphasize here that the order of the limits in our definition of d_{up} is important. A more "naive" definition would be

$$\widehat{d}_{\text{up}} = \limsup_{L\to\infty} \limsup_{\varepsilon\to 0} \frac{\log \mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L))}{(2L\pi)^d \log \varepsilon^{-1}} .$$

The two limits do not commute in general; see [5]. We believe our approach is more natural from an experimental/numerical point of view, in the sense that L is a parameter that can be varied in a series of measurements/simulations made at a fixed accuracy ε.

We also consider the density of topological entropy in section 5. We show that the spatial densities satisfy the analogue of the following well-known inequalities [14, 22]:

$$\mathcal{V} \le h_{\text{top}} \le \lambda d_{\text{up}} ,$$

where $\mathcal{V}$ is the volume expansion rate, λ is the largest Lyapunov exponent, h_{top} is the topological entropy, and d_{up} is the upper Hausdorff dimension.

The paper is organized as follows. In the remainder of this section we introduce the notation for the paper. In section 2 the semidiscrete and fully discrete approximations to (1.1) are presented. In section 3, we define the density of ε-entropy, topological entropy, of upper dimension and the volume growth rate and state our assumptions on the equation and its approximations. A key result of the paper is Lemma 4.2 (proved in Appendix A), which states that the evolution has a fast local smoothing effect, a property which allows us to establish upper bounds on the ε-entropy (section 4). This is then applied in section 5 to show that the topological entropy is finite. We also show that it is bounded below by the volume expansion rate (section 5.2). We discuss the upper semicontinuity of the attractors in section 6. Technical proofs are given at the end of the paper: Appendix B contains a proof of analyticity for the fully discrete scheme, Appendix C contains a lemma on analytic functions, and Appendix D recalls some results on Gevrey and Bernstein classes.

1.1. Notation. We use the following conventions: $\overline{z}$ is the complex conjugate of z and $|z| = \sqrt{z\overline{z}}$, its modulus. A function $f = f_1 + if_2$ with both f_1 and f_2 real analytic is identified with the vector-valued function $f = (f_1, f_2)$. Its analytic extension to the complex plane has the form $(f_1 + ig_1, f_2 + ig_2)$, and we write $|f| = (|f_1|^2 + |f_2|^2 + |g_1|^2 + |g_2|^2)^{1/2}$ which, on the real axis, is equal to the modulus of the complex function f. The convolution of two functions f, g is denoted $f \star g(x) := \int f(x-y)g(y)dy$.

If u is a function of t (time) and x (space), then we consider it either as a function of two variables with values in $\mathbb{C}$, written $u(x,t) \in \mathbb{C}$, or as a function of time with values in the functions of x, written $u(t) \in \mathcal{C}_{\mathrm{b}}(\mathbb{R}^d)$ (the set of bounded continuous functions). A function in the set $\mathcal{C}_{\mathrm{per}}([-L\pi, L\pi]^d)$ of $2L\pi$-periodic continuous functions will often be identified with its lift (by periodic extension) to $\mathcal{C}_{\mathrm{b}}(\mathbb{R}^d)$.

The spaces $\mathcal{C}_{\mathrm{b}}(\mathbb{R}^d)$ and $\mathcal{C}_{\mathrm{per}}([-L\pi, L\pi]^d)$ are Banach spaces with the sup norm $\|\cdot\|_\infty$ and may be viewed as subspaces of $\big(L^\infty(\mathbb{R}^d), \|\cdot\|_\infty\big)$ and $\big(L^\infty([-L\pi, L\pi]^d), \|\cdot\|_\infty\big)$, respectively. We also make extensive use of the Gevrey class $\mathcal{G}_\alpha(C)$ and the Bernstein class $\mathcal{B}_\sigma(C)$. These are both discussed in Appendix D. If $\mathrm{Re}\big(f\big)$ and $\mathrm{Im}\big(f\big)$ belong to the Gevrey class $\mathcal{G}_\alpha(C)$, we use the notation $f \in [\mathcal{G}_\alpha(R)]^2$ (similarly for $\mathcal{B}_\sigma(C)$).

We denote by $\mathcal{T}$ the standard Fourier transform operator

$$\big(\mathcal{T} f\big)(k) \;:=\; \frac{1}{(2\pi)^d} \int e^{ik\cdot x} f(x)\, dx \;, \qquad \big(\mathcal{T}^{-1} f\big)(x) \;:=\; \int e^{-ik\cdot x} f(x)\, dk \;.$$

The Fourier series operator for $2L\pi$-periodic functions is denoted with the same symbol:

$$\big(\mathcal{T} f\big)_n \;:=\; \frac{1}{(2L\pi)^d} \int_{|x|\le L\pi} e^{in\cdot x/L} f(x)\, dx \;, \qquad \big(\mathcal{T}^{-1} f\big)(x) \;:=\; \sum_{n\in\mathbb{Z}^d} e^{-in\cdot x/L} f_n \;. \tag{1.2}$$

We introduce two different smooth cutoff functions (see Figure 1). The first of these, φ, acts in real space and serves as a weight in L^p norms in order to get bounds that do not depend on L.

DEFINITION 1.1. *Let φ be a real-space cutoff function satisfying*

$$\varphi(x) > 0 \quad \forall x \in \mathbb{R}^d, \quad \varphi(-x) \,=\, \varphi(x) \;, \quad \int \varphi(x)\, dx \,=\, 1 \;, \quad \left\| \frac{\nabla\varphi}{\varphi} \right\|_\infty \,<\, \infty \;,$$

and, moreover, φ^{-1} is a tempered distribution, i.e., $\int \varphi^{-1} f < \infty$ for any Schwartz function f.

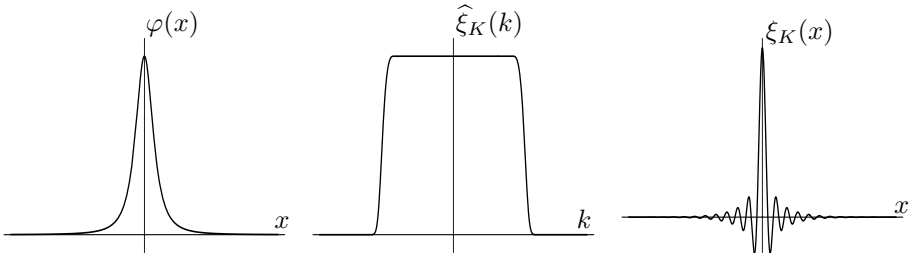


FIG. 1. *The cutoff functions of Definitions* 1.1–1.2.

Examples. The function

$$\varphi(x) \;=\; \frac{1}{(1+|C_\varphi x|^2)^{d/2+1}}$$

satisfies all of our requirements. (Here, C_φ is a normalization constant determined by the equation $\int \varphi = 1$, similarly for C_ψ below.) However, the function

$$\psi(x) \;=\; \frac{1}{\cosh(C_\psi x_1)\cdots\cosh(C_\psi x_d)}\;,$$

which has a sharper decay at infinity, cannot be used because it fails the last property; namely, $\cosh(x)$ is not a tempered distribution. The importance of this may be seen in Lemma 4.2.

Note that for (1.1) the function ψ could be used and would provide sharper bounds in our proofs. This does not work, however, with the truncation to a finite number of modes (such as given by the semidiscrete system (2.3) or fully discrete system (2.5)).

Our second cutoff function, ξ_K, is defined in terms of its Fourier transform. It smoothly truncates to a finite set of Fourier modes and hence produces a finite-dimensional problem.

DEFINITION 1.2. *Let $K > 1$ and let $\widehat{\xi}_K$ be a $\mathcal{C}^\infty$ function taking the following values:*

$$\widehat{\xi}_K(k) \;=\; \begin{cases} 1 & \text{if } |k| \le K-1\;, \\ 0 & \text{if } |k| \ge K\;. \end{cases}$$

Its inverse Fourier transform $\xi_K = \mathcal{T}^{-1}(\widehat{\xi}_K)$ is an (entire) Schwartz function.

Note that if f is a Schwartz function, then $\xi_K \star f$ is a Schwartz function whose Fourier transform has support in $[-K,K]^d$, and hence it belongs to $\mathcal{B}_K(C)$ for some C; see [23].

2. Semidiscrete and fully discrete approximations. In this section, we propose a spatial discretization of (1.1) and a fully discrete scheme.

2.1. Galerkin scheme. The semidiscretization we describe here is a spectral method. Let $N \in \mathbb{N}$; then we use the Fourier cutoff ξ_K of Definition 1.2 with $K = N$ to define the operators P^N and $\mathrm{Q}^N f := f - \mathrm{P}^N f$, where

$$\mathrm{P}^N f \;:=\; \xi_N \star f \;=\; \mathcal{T}^{-1}\left(\widehat{\xi}_N \mathcal{T}(f)\right)\;, \tag{2.1}$$

i.e.,

$$\mathrm{P}^N \sum_{n\in\mathbb{Z}^d} f_n e^{-in\cdot x/L} \;:=\; \sum_{|n|\le NL} \widehat{\xi}_N(n/L) f_n e^{-in\cdot x/L}\;.$$

Notice that P^N truncates to $(2NL)^d$ modes, not $(2N)^d$. In this way, $\mathcal{T}\big(\mathrm{P}^N f\big)$ has support contained in $[-N,N]^d$ for all L. The operator P^N is not a projector, since $\mathrm{P}^N\mathrm{P}^N \neq \mathrm{P}^N$.

The Galerkin approximation is defined as follows: the solution $u(x,t)$ to (1.1) is replaced by a finite Fourier series

$$u^N(x,t) := \sum_{|n|\leq NL} u_n(t) e^{-in\cdot x/L} \ . \tag{2.2}$$

The evolution equation is obtained by applying P^N to the nonlinear term of (1.1) and to the initial condition u_0:

$$\partial_t u^N = \big(\gamma + \nu\Delta\big)u^N + \mathrm{P}^N F(u^N) \ , \qquad u^N(x,0) = \big(\mathrm{P}^N u_0\big)(x) \ . \tag{2.3}$$

2.2. Fully discrete scheme. The time discretization is an exact exponential integrator for the linear part and a simple (order 1) quadrature for the nonlinear term appearing in the variation of constants formula. It is similar to that considered in [26], although they need a different definition of discrete Gevrey space, which depends on the time step. The full discretization is obtained by applying this time discretization to the Galerkin scheme (2.3). We use this particular scheme because it makes it straightforward to prove that solutions are (Gevrey) analytic functions (uniformly in the parameters of the scheme; see Appendix B), a fact that we rely on heavily in the next sections.

Let $\mathcal{L} = \gamma + \nu\Delta$ and $\mathcal{K}(x,t)$ be the convolution kernel associated with the operator $\exp(t\mathcal{L})$:

$$\mathcal{K}(x,t) = \frac{1}{(2\pi)^d}\int e^{-ik\cdot x + (\gamma - \nu|k|^2)t}\, dk \ . \tag{2.4}$$

Note that the operator P^N commutes with $\mathcal{K}\star\cdot$, since both are convolution operators. Let $h > 0$ denote the time step. Then the fully discrete approximation to $u(x,t)$ is defined iteratively by

$$u^N\big((n+1)h\big) = \mathcal{K}(h) \star \big(u^N(nh) + h\mathrm{P}^N F\big(u^N(nh)\big)\big) \ . \tag{2.5}$$

In terms of the Fourier coefficients, (2.2), we get

$$\begin{aligned} u^N_m\big((n+1)h\big) &= e^{h\lambda_m}\left(u^N_m(nh) + h\mathrm{P}^N\mathcal{T}F\big(\mathcal{T}^{-1}u^N(nh)\big)_m\right) \\ &= e^{h\lambda_m}(u^N_m(nh) + h\widehat{\xi}_N(m/L)\mathcal{T}F\big(\mathcal{T}^{-1}u^N(nh)\big)_m) \ , \end{aligned}$$

where $\{\lambda_m\}_{m\in\mathbb{Z}^d}$ are the eigenvalues of $\mathcal{L}$; namely, $\lambda_m = \gamma - \nu|m|^2/L^2$, n is the time index, m is the Fourier index, and $\mathcal{T}$ is the Fourier transform (1.2).

For the purposes of analysis, it is useful to consider this scheme in terms of piecewise solutions of a linear differential equation. Indeed, $u^N(x,(n+1)h)$ is the solution at time $t = h$ of

$$\partial_t u(x,t) = \nu\Delta u(x,t) + \gamma u(x,t) \tag{2.6}$$

with initial condition $u^N(x,nh) + h\mathrm{P}^N F\big(u^N(x,nh)\big)$ at $t = 0$.

Remark. We could apply our techniques to other numerical schemes. We require only the numerical approximation to belong to the Gevrey class $\mathcal{G}_\alpha(C)$ of bounded real analytic functions for some $\alpha > 0$, $C > 0$ (see Appendix D). There exists many wavelet and finite element schemes satisfying this requirement; see [7, 18]. In particular, Propositions D.2 and D.3 provide a natural example of a different basis of analytic functions on which our problem can be decomposed and then a truncation applied: this basis consists of the functions

$$\Psi_{j,k}(x) \;=\; \frac{3e^{ik\cdot x}\sin\bigl(2x-\frac{1}{3}j\pi\bigr)\sin\bigl(6x-j\pi\bigr)}{(6x-j\pi)^2}$$

for $j,k\in\mathbb{Z}^d$. These functions have the advantage of being localized both in real space, and in Fourier space, although the numerical implementation is more involved.

3. Definitions and assumptions. Since we are interested in the large volume limit we specify this dependence in the definitions below.

ASSUMPTION 3.1. *For initial data $u_0\in\mathcal{C}_{\rm per}([-L\pi,L\pi]^d)$, we assume that*

- (1.1) *is the generator of a semiflow $\Phi_L^t : u_0\mapsto u(t)$;*
- *for all $N>N_0$ the semidiscrete* (2.3) *is the generator of a semiflow $\Phi_{L,N}^t : u_0\mapsto u(t)$;*
- *for all $N>N_0$ and $h<h_0$ the fully discrete equation* (2.5) *is the generator of a semiflow $\Phi_{L,N,h}^t : u_0\mapsto u(t)$ with $t=nh$, $n\in\mathbb{N}$.*

Furthermore, we assume for each of the semiflows above that there exists constants $\alpha>0$ and $R>0$, independent of L and t, such that $\mathrm{Re}\bigl(u(t)\bigr)$ and $\mathrm{Im}\bigl(u(t)\bigr)$ belong to the Gevrey class $\mathcal{G}_\alpha(R)$ for all $t>T(u)$, and so $u(t)\in[\mathcal{G}_\alpha(R)]^2$. In other words, the following sets are absorbing balls for their corresponding semiflows:

$$\begin{aligned}
B(L) &:= \mathcal{C}_{\rm per}([-L\pi,L\pi]^d)\bigcap[\mathcal{G}_\alpha(R)]^2\ ,\\
B_N(L) &:= \mathrm{P}^N\mathcal{C}_{\rm per}([-L\pi,L\pi]^d)\bigcap[\mathcal{G}_\alpha(R)]^2\ ,\\
B_{N,h}(L) &:= \mathrm{P}^N\mathcal{C}_{\rm per}([-L\pi,L\pi]^d)\bigcap[\mathcal{G}_\alpha(R)]^2\ .
\end{aligned}$$

Throughout the paper we use $\widehat{\Phi}^t$ to denote any of the semiflows (with t taken appropriately) defined above and $\widehat{B}(L)$ to denote the corresponding absorbing balls.

We next define the attractors of the different evolutions introduced above.

DEFINITION 3.2. *We define the following invariant attracting sets for the flows defined in Assumption* 3.1:

$$\begin{aligned}
\mathcal{A}(L) &:= \bigcap_{t>0}\Phi_L^t\bigl(B(L)\bigr)\ ,\\
\mathcal{A}_N(L) &:= \bigcap_{t>0}\Phi_{L,N}^t\bigl(B_N(L)\bigr)\ ,\\
\mathcal{A}_{N,h}(L) &:= \bigcap_{n\in\mathbb{N}}\Phi_{L,N,h}^{nh}\bigl(B_{N,h}(L)\bigr)\ .
\end{aligned}$$

Throughout the paper we use $\widehat{\mathcal{A}}(L)$ to denote any of the above attracting sets.

Clearly, finite trigonometric sums like (2.2) are entire functions. However, the assumption that there exists a strip around the real axis where u^N is bounded by the same constant for all N is not trivial. Results of this type are known for a number of parabolic partial differential equations of the form (1.1), under the assumption that

F is dissipative in an appropriate sense (see, for example, [1, 25]). For numerical approximations, existence of semiflows and global attractors is a well-considered problem (see, for example, [24]). Gevrey regularity of solutions for numerical schemes has not been so widely considered; two different approaches are in [17, 26]. Appendix B contains a sketch of how to obtain this result for the fully discrete scheme given by (2.5). The proof relies only on an a priori L^∞ bound on the solutions and the assumption that the nonlinearity F is analytic.

We next introduce the notion of ε-entropy. The proof that this is a finite quantity will be given in section 4. From this we define the upper density of dimension.

DEFINITION 3.3. *Let Y be a subset of a metric space X. A set $\mathcal{U} = \{U_1, \dots, U_N\}$ of open sets in X is called a cover of Y if $\bigcup_{n=1}^N U_n \supset Y$. It is called an ε-cover if* $\max_{n=1,\dots,N} \operatorname{diam}(U_n) \le \varepsilon$.

Let $\widehat{\mathcal{A}}(L)$ be endowed with the metric defined by the norm $\|\cdot\|_\infty$. Let

$$\mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L)) := \inf\bigl\{\operatorname{card}(\mathcal{U}) : \mathcal{U} \text{ is an } \varepsilon\text{-cover of } \widehat{\mathcal{A}}(L)\bigr\} .$$

We define the ε-entropy H_ε as the limit

$$H_\varepsilon := \limsup_{L\to\infty} \frac{\log \mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L))}{(2L\pi)^d} .$$

The upper density of dimension d_{up} is defined by

$$d_{\mathrm{up}} := \limsup_{\varepsilon\to 0} \frac{H_\varepsilon}{\log \varepsilon^{-1}} .$$

Remark. In [4, 5, 6], H_ε was defined with a limit instead of a limit superior. The existence of the limit followed from a subadditivity argument which cannot be used here because of the boundary conditions. That is, the set $\widehat{\mathcal{A}}(L)$ we are considering here changes with L, whereas, in the papers [4, 5, 6], only the topology on $\mathcal{A}$ depended on L, not the set itself. See also [31, 32] for similar results.

Another more classical notion of entropy is the topological entropy. It serves to measure to complexity of a dynamical system. Similarly to the previous definition, we consider here the spatial density of topological entropy. See section 5 for results on the topological entropy.

DEFINITION 3.4. *For $\tau > 0$, we define a pseudometric $d_{m,\tau}$ on $\mathcal{C}_{\mathrm{per}}([-L\pi, L\pi]^d)$ by*

$$d_{m,\tau}(u, v) := \max_{k=0,\dots,m-1} \|\widehat{\Phi}^{k\tau}(u) - \widehat{\Phi}^{k\tau}(v)\|_\infty .$$

An (m, ε)-cover of $\widehat{\mathcal{A}}(L)$ is a collection of open sets whose diameter in the metric $d_{m,\tau}$ is at most ε and whose union contains $\widehat{\mathcal{A}}(L)$. Let $\mathcal{M}_{m,\tau}(\varepsilon, \widehat{\mathcal{A}}(L))$ be the cardinality of such a minimal (m, ε)-cover.

The (spatial density of) topological entropy is defined as follows:

$$h_{\mathrm{top}} := \limsup_{\varepsilon\to 0} \limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \lim_{m\to\infty} \frac{1}{m\tau} \log \mathcal{M}_{m,\tau}(\varepsilon, \widehat{\mathcal{A}}(L)) . \tag{3.1}$$

The existence of the first limit in (3.1) can be proved by a subadditivity argument; see [4, 6, 14]. A useful way of computing a lower bound on the topological entropy is by measuring the volume expansion rate (see section 5.2).

DEFINITION 3.5. *Let $L \mapsto \mathcal{D}(L)$ be a family of ℓ-dimensional $\mathcal{C}^\infty$ submanifolds of the absorbing ball $\widehat{B}$. We define $\mathcal{V}$, the volume expansion rate, by*

$$\mathcal{V} := \limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \limsup_{m\to\infty} \frac{1}{m\tau} \log \mathrm{Vol}_\ell\big(\widehat{\Phi}^{m\tau}(\mathcal{D}(L))\big) \ ,$$

where Vol_ℓ *is the ℓ-dimensional (Euclidean) volume.*

4. Upper bound on the ε-entropy. We now work towards proving our main result which is a bound on the ε-entropy. First we discuss a preliminary result on the smoothing property of the semiflow which is proved in Appendix A.

4.1. Smoothing property of the semiflow. We consider here differences $w = u - v$ of two orbits u and v of the semiflow $\widehat{\Phi}^t$ of Assumption 3.1. We define functions G_1 and G_2 in such a way that w satisfies

$$\partial_t w \ = \ \big(\gamma + \nu\Delta\big)w + \mathrm{P}^N\big(G_1(u,v)w + G_2(u,v)\overline{w}\big) \tag{4.1}$$

for continuous time and

$$w\big((n+1)h\big) \ = \ \mathcal{K}(h) \star \big(w(nh) + h\mathrm{P}^N\big(G_1(nh)w(nh) + G_2(nh)\overline{w}(nh)\big)\big) \tag{4.2}$$

for discrete time. From now on we view G_1 and G_2 as functions of x and t (rather than of u and v), and we use the following consequence of Assumption 3.1.

LEMMA 4.1. *There exists $\alpha > 0$ and $R > 0$, both independent of N, L, and t, such that $w(t)$, $G_1(t)$, and $G_2(t)$ all belong to $[\mathcal{G}_\alpha(R)]^2$ for all $t > 0$ (and $t/h \in \mathbb{N}$ for (4.2)).*

Remark. We may assume without loss of generality that the R and the α of Assumption 3.1 and Lemma 4.1 are equal and that they are also equal for the fully continuous, semidiscrete, and fully discrete equations.

We compute bounds on the weighted L^2 norm of w shifted in the complex plane over a finite time interval. Instead of taking the usual (flat) L^2 norm over $[-L\pi, L\pi]^d$, which would not behave well in the limit $L \to \infty$, we take a norm over the whole of $\mathbb{R}^d$ weighted with the function φ from Definition 1.1. Therefore, L disappears completely from our estimates. However, in Definition 3.3, we chose to work with the L^∞ topology. We therefore use the following bootstrap argument. From a bound in L^∞ at time $t = 0$, we get a bound in weighted L^2 at time $t = 0$. Using the next lemma we deduce a bound at $t = 1$ in a weighted L^2 space on a strip of the complex plane. This is in turn combined with Lemma C.1 and provides an L^∞ bound at $t = 1$.

LEMMA 4.2. *There is a constant $b > 0$ such that, for any $\beta \in (-\alpha, \alpha)$, any N, and any L, the following bound holds on w a solution of (4.1) (or (4.2)) as long as $t \le 1$ (and $t/h \in \mathbb{N}$ in the case of a fully discrete scheme):*

$$\sup_{|y|\le L\pi} \int \varphi(x-y)|w(x+i\beta t,t)|^2\,dx \ \le \ e^{2bt} \sup_{|y|\le L\pi} \int \varphi(x-y)|w(x,0)|^2\,dx \ . \tag{4.3}$$

The proof of Lemma 4.2 is given in Appendix A.

These L^2 norms shifted in the complex plane can be understood in terms of the classical Gevrey norms. First consider $\varphi \equiv 1$. Then, using Fourier series and taking $\beta > 0$, we see that

$$\int \big(|f(x+2i\beta)|^2 + |f(x-2i\beta)|^2\big)\,dx \ = \ \Big\|\Gamma e^{\beta(-\Delta)^{1/2}} f\Big\|_2^2 \ , \tag{4.4}$$

where Γ is the bounded invertible operator defined by

$$\big(\mathcal{T}(\Gamma f)\big)_n \;=\; (1+e^{-2\beta|n|/L})(\mathcal{T}f)_n \;.$$

This means that the left-hand side of (4.4) is equivalent to a Gevrey norm. (Similar norms have been used in [11, 12].) We apply a nonconstant weight function φ to this norm in order to get estimates which are independent of L and take the sup over $|\beta| \le \alpha$ to be able to use Lemma C.1. Similar issues have been raised in the paper [21], but our approach is different in that we never explicitly work in Fourier space. We note also that the norms used in [21] grow with the domain size (due to the embedding constant), a problem we avoid here by using the cutoff φ.

4.2. Proof of the upper bound. We next show that the ε-entropy H_ε (Definition 3.3) is of order $\log \varepsilon^{-1}$ at most.

THEOREM 4.3. *There exists a constant $C < \infty$, independent of ε, such that*

$$H_\varepsilon \;\le\; C \log\left(\frac{R}{\varepsilon}\right) \;,$$

where R is the radius of the absorbing ball $\widehat{B}(L)$ in Assumption 3.1.

The proof is based on the following lemma.

LEMMA 4.4. *There is a constant $C > 0$ such that, for all $\varepsilon > 0$, the following holds:*

$$H_\varepsilon \;\le\; H_{2\varepsilon} + C \;.$$

Proof. The proof is a consequence of the smoothness result of the previous section. We give the proof for the time continuous cases (1.1), (2.3). The time discrete case (2.5) is similar; it requires only restricting t to multiples of h.

Suppose we are given a 2ε-cover $\{U_1, \dots, U_{\mathcal{N}}\}$ of $\widehat{\mathcal{A}}(L)$. Then by invariance of $\widehat{\mathcal{A}}$ the set

$$\{\widehat{\Phi}^t(U_1), \dots, \widehat{\Phi}^t(U_{\mathcal{N}})\}$$

is a cover of $\widehat{\mathcal{A}}(L)$ for all $t > 0$. Moreover, if $u, v \in U_1$, by Lemma C.1 combined with Lemma 4.2, we have

$$\sup_{|x|\le L\pi\,,\,2|y|\le\alpha} |\big(\widehat{\Phi}^1(u) - \widehat{\Phi}^1(v)\big)(x+iy)| \;\le\; C\varepsilon \;.$$

That is, if we let $w = \widehat{\Phi}^1(u) - \widehat{\Phi}^1(v)$, then $w \in [\mathcal{G}_{\alpha/2}(C\varepsilon)]^2$ with C independent of L and ε.

We now use an argument due to Tikhomirov [28], discussed in [15, section 8, Theorem XXII]. By Proposition D.2 w can be written as

$$w(z) \;=\; \sum_{n\in\mathbb{Z}^d} e^{-\alpha|n|/2} e^{in\cdot z} w_n(z) \;, \tag{4.5}$$

with w_n in the Bernstein class $[\mathcal{B}_2(C'\varepsilon)]^2$. (See Appendix D for the definition of $\mathcal{B}_2$.) Thus, splitting the sum in (4.5) in two, we can find a K independent of ε and L, and a $\widetilde{w} \in [\mathcal{B}_K(C'\varepsilon)]^2$, such that

$$\|w - \widetilde{w}\|_\infty \;\le\; \frac{\varepsilon}{2} \;.$$

If $\widetilde{w} \in [\mathcal{B}_K(C\varepsilon)]^2$, then, by Proposition D.3,

$$\widetilde{w}(x) \;=\; \sum_{n\in\mathbb{Z}^d} \widetilde{w}\big(x_K(n)\big)\mathcal{F}_K\big(x - x_K(n)\big)\;,$$

and hence there is a $\delta > 0$ depending only on K such that $\|\widetilde{w}\|_\infty \le \varepsilon/2$ if $|\widetilde{w}(x_K(n))| \le \delta\varepsilon$ for all $n \in \mathbb{Z}^d$ for which $x_K(n) = (n\pi)/(3K) \in [-L\pi, L\pi]^d$. There are $c(K)(2L\pi)^d$ such points, and hence at most

$$\left(\frac{C\varepsilon}{\delta\varepsilon}\right)^{c(K)(2L\pi)^d} \;=:\; C_*^{(2L\pi)^d}$$

balls of radius $\varepsilon/2$ will be needed to cover $[\mathcal{B}_K(C\varepsilon)]^2$. This covers all the functions $\widetilde{w}$ obtained from the set $\widehat{\Phi}^1(U_1)$ by the above construction. Consequently, $\widehat{\Phi}^1(U_1)$ can be covered with the same number of balls of diameter ε.

Repeating the operation with each one of the $\mathcal{N}(2\varepsilon, \widehat{\mathcal{A}}(L))$ sets of diameter 2ε of the original cover $\{U_1, \dots, U_{\mathcal{N}}\}$, we obtain a cover with at most

$$\mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L)) \;\le\; \mathcal{N}(2\varepsilon, \widehat{\mathcal{A}}(L)) C_*^{(2L\pi)^d}$$

elements. Taking the logarithm, dividing by $(2L\pi)^d$, and passing to the limit $L \to \infty$, we obtain Lemma 4.4. □

Proof of Theorem 4.3. It trivially holds that $H_R = 0$, because $\mathcal{N}(R, \widehat{\mathcal{A}}(L)) = 1$ by Assumption 3.1. Let k be the smallest integer larger than $\log(R/\varepsilon)/\log 2$; then by Lemma 4.4 we have

$$H_\varepsilon \;\le\; H_{2\varepsilon} + C \;\le\; \cdots \;\le\; H_{2^k\varepsilon} + Ck \;\le\; C' \log R/\varepsilon\;.$$

This proves Theorem 4.3. □

5. The topological entropy.

5.1. Upper bound by the dimension. In this section, we prove that the topological entropy of the attractors $\widehat{\mathcal{A}}$ is bounded by a multiple of the upper density of dimension, a quantity related to the ε-entropy. The corresponding inequality for finite-dimensional dynamical systems is well known; see [14].

THEOREM 5.1. *There is a $b < \infty$ such that*

$$h_{\text{top}} \;\le\; b d_{\text{up}} \;<\; \infty\;. \tag{5.1}$$

Proof. The right-hand inequality is a direct consequence of Theorem 4.3. The left-hand inequality follows from the arguments in [4, 14] that we summarize here. Let $\rho > 0$ be such that $H_\varepsilon \le \big(d_{\text{up}} + \rho\big)\log 1/\varepsilon$ for all $\varepsilon < \varepsilon_0$ and then let $L_0 = L_0(\varepsilon, \rho)$ be such that, for all $L > L_0$,

$$\frac{\log \mathcal{N}(\varepsilon, \widehat{\mathcal{A}}(L))}{(2L\pi)^d} \;\le\; H_\varepsilon + \rho \;\le\; \big(d_{\text{up}} + \rho\big)\log\frac{1}{\varepsilon} + \rho\;.$$

By iterating Lemma C.1 and Lemma 4.2, there is a $b > 0$ such that, for all L and all (sufficiently small) $\varepsilon > 0$, if $\|u - v\|_\infty \le \varepsilon$, then, for $t > 0$,

$$\|\widehat{\Phi}^t(u) - \widehat{\Phi}^t(v)\|_\infty \;\le\; e^{bt}\varepsilon\;.$$

Let $\varepsilon' = \exp(-bT)\varepsilon$. Let an ε'-cover of $\widehat{\mathcal{A}}(L)$ (in the sense of Definition 3.3) be given. Then it is also a $(T/\tau, \varepsilon)$-cover (in the sense of Definition 3.4), and hence

$$\mathcal{M}_{T/\tau,\tau}(\varepsilon, \widehat{\mathcal{A}}(L)) \le \mathcal{N}(\varepsilon', \widehat{\mathcal{A}}(L)) .$$

It follows that

$$\begin{aligned} h_{\text{top}} &= \limsup_{\varepsilon\to 0} \limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \lim_{T\to\infty} \frac{1}{T} \log \mathcal{M}_{T/\tau,\tau}(\varepsilon, \widehat{\mathcal{A}}(L)) \\ &= \limsup_{\varepsilon\to 0} \limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \inf_{T} \frac{1}{T} \log \mathcal{M}_{T/\tau,\tau}(\varepsilon, \widehat{\mathcal{A}}(L)) \\ &\le \limsup_{\varepsilon\to 0} \limsup_{L\to\infty} \frac{1}{T} \frac{\log \mathcal{N}(\varepsilon', \widehat{\mathcal{A}}(L))}{(2L\pi)^d} \\ &\le \limsup_{\varepsilon\to 0} \limsup_{L\to\infty} \frac{1}{T} \left((d_{\text{up}} + \rho) \log \frac{1}{\varepsilon'} + \rho \right) . \end{aligned}$$

Since $\log 1/\varepsilon' = bT + \log 1/\varepsilon$, the limit $T \to \infty$ and $\rho \to 0$ leaves only bd_{up} on the right-hand side above. ∎

5.2. Lower bound by the expansion rate. We provide here a way of computing a lower bound on the topological entropy (hence on the upper dimension d_{up} by Theorem 5.1), based on Yomdin's theorem [30], an account of which may be found in [22].

THEOREM 5.2. *Let h_{top} be as in Definition* 3.4. *Then for all choices of $\mathcal{D}(L)$ in Definition* 3.5,

$$\mathcal{V} \le h_{\text{top}} .$$

Remark. The lower bound in [5] is in the same spirit. An adequate sequence of submanifolds is chosen (small balls around the trivial solution). The volume expansion rate of that sequence can be controlled, yielding a lower bound on the (ε-)entropy.

Proof. The proof follows from the argument by Yomdin [30] and Gromov [10]. By a lemma of Gromov [10], there exists a $C > 0$ such that if $\widehat{\Phi}^\tau$ is $\mathcal{C}^r$, then

$$\text{Vol}_\ell\big(\widehat{\Phi}^{m\tau}(\mathcal{D}(L))\big) \le \mathcal{M}_{m,\tau}(\varepsilon, \widehat{\mathcal{A}}(L))(C\|D\widehat{\Phi}^\tau\|_\infty)^{m\ell/r} , \text{ and}$$

hence

$$\begin{aligned} &\limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \limsup_{m\to\infty} \frac{1}{m\tau} \log \text{Vol}_\ell\big(\widehat{\Phi}^{m\tau}(\mathcal{D}(L))\big) \\ &\le \limsup_{L\to\infty} \frac{1}{(2L\pi)^d} \limsup_{m\to\infty} \frac{1}{m\tau} \log \mathcal{M}_{m,\tau}(\varepsilon, \widehat{\mathcal{A}}(L)) \\ &\quad + \limsup_{L\to\infty} \frac{\ell/r}{(2L\pi)^d} \log\big(C^{1/\tau}\|D\widehat{\Phi}^\tau\|_\infty^{1/\tau}\big) . \end{aligned}$$

Since τ can be arbitrarily large, the constant C drops out, and, since $\widehat{\Phi}^\tau$ is $\mathcal{C}^\infty$, the second term is arbitrarily small by letting $r \to \infty$. The first term tends to h_{top} upon letting $\varepsilon \to 0$. ∎

6. Upper semicontinuity of the infinite volume attractors. In this section we discuss four different invariant sets and their mutual relationship. The first two invariant sets are $\mathcal{A}_{N,h}(L)$ and $\mathcal{A}(L)$ from Definition 3.2. Then we also introduce two large volume limits:

$$\mathcal{A}_{N,h}(\infty) := \overline{\bigcup_{L\in\mathbb{N}} \mathcal{A}_{N,h}(L)}\ , \qquad \mathcal{A}(\infty) := \overline{\bigcup_{L\in\mathbb{N}} \mathcal{A}(L)}\ , \tag{6.1}$$

where the closure is taken in the uniformly local topology of [19]. We define the distance between a point and a set and between two sets in the standard way:

$$\begin{aligned} \operatorname{dist}\big(U,\mathcal{V}\big) &:= \inf_{V\in\mathcal{V}} \|U-V\|_{L^\infty([-L\pi,L\pi]^d)}\ , \\ \operatorname{dist}\big(\mathcal{U},\mathcal{V}\big) &:= \sup_{U\in\mathcal{U}} \operatorname{dist}\big(U,\mathcal{V}\big)\ . \end{aligned}$$

We claim that

$$\lim_{N\to\infty,h\to 0} \operatorname{dist}\big(\mathcal{A}_{N,h}(L),\mathcal{A}(L)\big) \,=\, 0\ , \qquad \lim_{N\to\infty,h\to 0} \operatorname{dist}\big(\mathcal{A}_{N,h}(\infty),\mathcal{A}(\infty)\big) \,=\, 0\ , \tag{6.2}$$

and the following relations are straightforward from (6.1):

$$\begin{aligned} \lim_{L\to\infty} \operatorname{dist}\big(\mathcal{A}_{N,h}(L),\mathcal{A}_{N,h}(\infty)\big) &= 0\ , \\ \lim_{L\to\infty} \operatorname{dist}\big(\mathcal{A}(L),\mathcal{A}(\infty)\big) &= 0\ . \end{aligned}$$

Hence we obtain the following diagram, in which each arrow represents a relation of upper semicontinuity:

$$\begin{array}{ccc} \mathcal{A}_{N,h}(L) & \xrightarrow[h\to 0]{N\to\infty} & \mathcal{A}(L) \\ {\scriptstyle L\to\infty}\big\downarrow & & \big\downarrow{\scriptstyle L\to\infty} \\ \mathcal{A}_{N,h}(\infty) & \xrightarrow[h\to 0]{N\to\infty} & \mathcal{A}(\infty)\ . \end{array}$$

The relation (6.3) is a consequence of the following (see, e.g., [13, 17, 19, 20]).

THEOREM 6.1. *For all $\varepsilon > 0$, there is a T_1, an h_1, and an N_1 such that if $h < h_1$ and $N > N_1$, then, for all $L \in \mathbb{N}$,*

$$\Phi^T_{L,N,h}\left(B_{N,h}(L)\right) \subset \mathcal{U}_\varepsilon(\mathcal{A}(L)) \qquad \forall T > T_1\ ,$$

where $\mathcal{U}_\varepsilon(\mathcal{A}(L))$ is the ε-neighborhood of $\mathcal{A}(L)$ in L^∞.

Proof. The proof is by induction using the attracting property of the attractor and a finite time error estimate.

By the attraction property of $\mathcal{A}(L)$, there exists a T such that, for all $T > T_1$,

$$\Phi^T_L\left(B(L)\cup B_{N,h}(L)\right) \subset \mathcal{U}_{\varepsilon/2}(\mathcal{A}(L))$$

for all $L \in \mathbb{N}$. Hence for any $u_0 \in B_{N,h}(L)$ we have

$$\begin{aligned} \operatorname{dist}\big(\Phi^{nh}_{L,N,h}(u_0),\mathcal{A}(L)\big) &= \inf_{u\in\mathcal{A}(L)} \|\Phi^{nh}_{L,N,h}(u_0)-u\|_\infty \\ &\le \inf_{u\in\mathcal{A}(L)} \|\Phi^{nh}_{L}(u_0)-u\|_\infty + \|\Phi^{nh}_{L,N,h}(u_0)-\Phi^{nh}_{L}(u_0)\|_\infty \\ &\le \frac{\varepsilon}{2} + \|\Phi^{nh}_{L,N,h}(u_0)-\Phi^{nh}_{L}(u_0)\|_\infty\ , \end{aligned} \tag{6.3}$$

provided $nh > T$.

We next show that N, h can be chosen in such a way that the second term above is smaller than $\varepsilon/2$ for all $T \in (0, 2T_1]$.

Let $v(t) = \Phi_L^t(u_0)$ and $w(nh+s) = \Phi_{\mathrm{Lin}}^s \Phi_{L,N,h}^{nh}(u_0)$, where Φ_{Lin}^s is the solution semiflow of (2.6). We thus have for $s < h$

$$\begin{aligned}\partial_t\big(v(nh+s) - w(nh+s)\big) &= \big(\gamma + \nu\Delta\big)\big(v(nh+s) - w(nh+s)\big) + F\big(v(nh+s)\big) \\ &= \big(\gamma + \nu\Delta\big)\big(v(nh+s) - w(nh+s)\big) + \mathrm{P}^N\left(F\big(v(nh+s)\big) - F\big(w(nh+s)\big)\right) \\ &\quad -\mathrm{P}^N F\big(w(nh+s)\big) + \mathrm{Q}^N F\big(v(nh+s)\big) \ .\end{aligned}$$

Using Proposition D.2 we see that

$$\sup_{s<h} \|\mathrm{Q}^N F\big(v(nh+s)\big)\|_\infty \ \le\ C(R) e^{-\alpha N} \ .$$

It is also quite easy (using Fourier transforms) to see that

$$\left\| \int_0^h \left(\mathrm{P}^N F\big(w((n+1)h)\big) - \mathcal{K}(s) \star \mathrm{P}^N F\big(w((n+1)h - s)\big)\right) ds \right\|_\infty \ \le\ C(R) h \ .$$

Hence, using the same analysis as in the proof of Lemma 4.2, we obtain

$$\|v((n+1)h) - w((n+1)h)\|_\infty \ \le\ e^{ch} \|v(nh) - w(nh)\|_\infty + C(R)h\big(1 + e^{-\alpha N}\big) \ .$$

By iteration, we obtain

$$\|v(nh) - w(nh)\|_\infty \ \le\ e^{cnh} \|v(0) - w(0)\|_\infty + C(R) e^{cnh} h (1 + e^{-\alpha N}) \ . \tag{6.4}$$

Taking h small enough, we can make the second term of (6.3) smaller than $\varepsilon/2$ for all $T \in (0, 2T_1]$.

To complete the induction we note that the absorbing ball is forward invariant, and so we can repeat the argument for $T > 2T_1$. □

7. Discussion: The complex Ginzburg–Landau equation. An interesting example to which our results apply is the (cubic) complex Ginzburg–Landau (CGL) equation in $d = 1$ space dimension

$$\partial_t u(x,t) \ =\ (1+ia)\partial_x^2 u(x,t) + u(x,t) - (1+ib)|u(x,t)|^2 u(x,t) \ . \tag{7.1}$$

In terms of the notations of (1.1), we have

$$d \ =\ 1 \ , \quad \nu \ =\ 1 + ia \ , \quad \gamma \ =\ 1 \ , \quad F(u) \ =\ -(1+ib)|u|^2 u \ .$$

Remark that the equation for the difference $w = u - v$ of two solutions u and v that we use in section 4.1 admits a simple expression:

$$\begin{aligned}\partial_t w(x,t) = (1+ia)\partial_x^2 w(x,t) + w(x,t) \\ + \int \xi_N(x-y)\big(G_1(y,t) w(y,t) + G_2(y,t)\overline{w}(y,t)\big) dy \ ,\end{aligned}$$

where

$$G_1(x,t) \ =\ -(1+ib)\big(|u(x,t)|^2 + |v(x,t)|^2\big) \ , \qquad G_2(x,t) \ =\ -(1+ib) u(x,t) v(x,t) \ .$$

The CGL equation (7.1) arises as a "normal form" in certain types of bifurcation with continuous spectrum; see [1, 3]. Assumption 3.1 for the continuous case follows from the works [2, 1, 25]. In particular, the following results have been proved.

THEOREM 7.1. *Equation* (7.1) *defines a semiflow* Φ^t *on* $L^\infty(\mathbb{R})$ *which has an absorbing ball* B *in* $\mathcal{G}_\alpha(C)$ *for some* $C > 0$ *and* $\alpha > 0$ *(see Appendix D). The attractor* $\mathcal{A} = \cap_{t>0}\Phi^t(B)$ *exists and is compact in* $L^\infty([-L, L])$ *for any* $L > 0$.

Remark that these results hold on the whole space without boundary conditions, but they obviously remain true on the set of spatially periodic solutions, which is invariant under the time evolution.

The following rigorous upper and lower bounds on the ε-entropy in unbounded volumes were obtained in [5].

THEOREM 7.2. *Let* $\mathcal{A}$ *be the attractor of* (7.1) *for general initial conditions in* $L^\infty(\mathbb{R})$ *and let* $\mathcal{N}(\varepsilon, \mathcal{A})$ *be the minimum the number of balls in an* ε*-cover of* $\mathcal{A}$ *in the topology of* $L^\infty([-L, L])$. *There is a* $C > 0$ *for which*

$$C^{-1}\log(1/\varepsilon) \,\le\, H_\varepsilon(\mathcal{A}) \,=\, \lim_{L\to\infty} \frac{\log \mathcal{N}(\varepsilon, \mathcal{A})}{2L} \,\le\, C\log(1/\varepsilon)\ .$$

In particular, the limit exists.

The discretization (2.5) in the particular case of the CGL equation is

$$u_m^N\big((n+1)h\big) \,=\, e^{(1-(1+ia)m^2)nh}\Big(1 - h(1+ib)\widehat{\xi}_N(m/L)|u_m^N(nh)|^2\Big)u_m^N(nh)\ , \tag{7.2}$$

where $n = 0, 1, \ldots$ is the time index and $m = -N, \ldots, N$ is the Fourier index.

A closely related time discretization was considered in [26]. Although there is no formal proof of existence of a semiflow and global attractor for the modified Galerkin scheme considered here, this can be seen to be true by considering the error bound (6.4) and the results of Theorem 7.1 over a finite time interval $[0, T]$. This suffices to prove that the discretized evolution is well defined and solutions stay bounded on that time interval. Iterating over $[qT, (q+1)T]$ for all $q > 0$ we obtain the existence of a global semiflow. The proof of existence of the absorbing balls of Assumption 3.1 is sketched in Appendix B. This implies that the following theorem holds as a special case of Theorem 4.3.

THEOREM 7.3. *Consider the CGL equation* (7.2). *There exists a constant* $C < \infty$, *independent of* ε, *such that*

$$H_\varepsilon \,\le\, C\log\left(\frac{R}{\varepsilon}\right)\ ,$$

where R *is the radius of the absorbing ball* B *in Gevrey space for* (7.2), *and* H_ε *is defined in Definition* 3.3.

Appendix A. Proof of Lemma 4.2. We first consider the time continuous case (4.1). We write the analytic extension of w as a vector-valued function with components $w_{\rm r}$ and $w_{\rm i}$ (each of which is complex-valued), and its complex argument $x + iy$ is also written as a vector of reals. Namely,

$$w(x + iy, t) \,=\, \big(w_{\rm r}(x, y; t), w_{\rm i}(x, y; t)\big)\ .$$

As a preparation for the proof, we estimate the following expression:

$$\begin{aligned}&\operatorname{Re}\nu \int \varphi(x)(\overline{w}_{\rm r}(x, y; t)\Delta_x w_{\rm r}(x, y; t) + \overline{w}_{\rm i}(x, y; t)\Delta_x w_{\rm i}(x, y; t))\,dx\\ &+ \operatorname{Re} i\beta \int \varphi(x)(\overline{w}_{\rm r}(x, y; t)\nabla_y w_{\rm r}(x, y; t) + \overline{w}_{\rm i}(x, y; t)\nabla_y w_{\rm i}(x, y; t))\,dx.\end{aligned} \tag{A.1}$$

By using the Cauchy–Riemann equations ($|\nabla_y u_{\mathrm{r,i}}| = |\nabla_x u_{\mathrm{r,i}}|$), we obtain

$$\begin{aligned}
&\operatorname{Re}\nu \int \varphi\Big(\overline{w}_{\mathrm{r}}\Delta_x w_{\mathrm{r}} + \overline{w}_{\mathrm{i}}\Delta_x w_{\mathrm{i}}\Big)\,dx + \operatorname{Re} i\beta \int \varphi(\overline{w}_{\mathrm{r}}\nabla_y w_{\mathrm{r}} + \overline{w}_{\mathrm{i}}\nabla_y w_{\mathrm{i}})\,dx \\
&= \ -\operatorname{Re}\nu \int \varphi\big(|\nabla_x w_{\mathrm{r}}|^2 + |\nabla_x w_{\mathrm{i}}|^2\big)\,dx - \operatorname{Re}\nu \int \nabla_x\varphi(\overline{w}_{\mathrm{r}}\nabla_x w_{\mathrm{r}} + \overline{w}_{\mathrm{i}}\nabla_x w_{\mathrm{i}})\,dx \\
&\quad + \operatorname{Re} i\beta \int \varphi(\overline{w}_{\mathrm{r}}\nabla_y w_{\mathrm{r}} + \overline{w}_{\mathrm{i}}\nabla_y w_{\mathrm{i}})\,dx \\
&\le -\operatorname{Re}\nu \int \varphi\big(|\nabla_x w_{\mathrm{r}}|^2 + |\nabla_x w_{\mathrm{i}}|^2\big)\,dx \\
&\quad + |\nu| \left\| \frac{\nabla\varphi}{\varphi} \right\|_\infty \int \varphi(|w_{\mathrm{r}}||\nabla_x w_{\mathrm{r}}| + |w_{\mathrm{i}}||\nabla_x w_{\mathrm{i}}|)\,dx \\
&\quad + |\beta| \int \varphi(|w_{\mathrm{r}}||\nabla_x w_{\mathrm{r}}| + |w_{\mathrm{i}}||\nabla_x w_{\mathrm{i}}|)\,dx \\
&\le \frac{|\beta|^2 + |\nu|^2\|\nabla\varphi/\varphi\|_\infty^2}{2\mathrm{Re}\nu} \int \varphi\big(|w_{\mathrm{r}}|^2 + |w_{\mathrm{i}}|^2\big)\,dx \\
&=: b_0 \int \varphi\big(|w_{\mathrm{r}}|^2 + |w_{\mathrm{i}}|^2\big)\,dx\ . \qquad \text{(A.2)}
\end{aligned}$$

Define

$$\varphi_y(x) := \varphi(x-y)\ , \qquad \xi_y^*(x) := \xi_N(x-y)\ ,$$

where φ and ξ_N are as in Definitions 1.1–1.2. We next compute the time derivative of the left-hand side of (4.3). The expression (A.1) is the linear part of the time derivative, and hence we simply insert the bound (A.2) and compute the nonlinear part:

$$\begin{aligned}
&\frac{1}{2}\partial_t \sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \ \le\ \frac{1}{2}\sup_y \partial_t \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \\
&\le \big(\gamma + b_0\big)\sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \\
&\quad + \operatorname{Re}\sup_y \Big| \int \varphi_y(x)\overline{w}(x+i\beta t,t) \\
&\quad \times \left(\int \xi_x^*(z)(G_1(z+i\beta t,t)\overline{w}(z+i\beta t,t) + G_2(z+i\beta t,t)w(z+i\beta t,t))dz \right) dx \Big| \\
&\le \big(\gamma + b_0\big)\sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \\
&\quad + \sup_y \int \varphi_y(x)|w(x+i\beta t,t)| \\
&\quad \times \left(\int \frac{|\xi_x^*(z)|}{\sqrt{\varphi_x(z)}}\sqrt{\varphi_x(z)}(|G_1(z+i\beta t,t)| + |G_2(z+i\beta t,t)|)|w(z+i\beta t,t)|dz \right) dx\ .
\end{aligned}$$

At this point, we apply the Cauchy–Schwarz inequality to each of the two integrals on the right-hand side. Using Lemma 4.1 we know that

$$\sup_{|\beta|\le\alpha}\ \sup_{t\le 1}\ \sup_{x\in\mathbb{R}^d} \big(|G_1(x+i\beta,t)| + |G_2(x+i\beta,t)|\big) \ \le\ 2R\ .$$

This gives

$$\begin{aligned}
\frac{1}{2}\partial_t \sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \;&\le\; \big(\gamma+b_0\big)\sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \\
&+\sup_y \left(\int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx\right)^{1/2}\left(\int \varphi(x)\,dx\int \frac{\xi_N^2(z)}{\varphi(z)}\,dz\right)^{1/2} \\
&\times 2R\left(\sup_x \int \varphi_x(z)|w(z+i\beta t,t)|^2\,dz\right)^{1/2} \\
\le\; &\left(\gamma+b_0+2R\left(\int \frac{\xi_N^2}{\varphi}\right)^{1/2}\right)\sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx \\
=:\; & b\sup_y \int \varphi_y(x)|w(x+i\beta t,t)|^2\,dx\ ,
\end{aligned}$$

where we used that, by Definition 1.1, $\int \xi_N^2/\varphi<\infty$ because ξ_N^2 is a Schwartz function and $1/\varphi$ is a Schwartz distribution. Equation (4.3) now follows from Gronwall's lemma.

In the discrete case, we solve the linear differential equation (see (2.6))

$$\partial_t w(nh+t) \;=\; \big(\gamma+\nu\Delta\big)w(nh+t)$$

for $t\in[0,h)$ with initial condition $w(nh)+h\xi_N\star\big(G_1(nh)w(nh)+G_2(nh)\overline{w}(nh)\big)$, and then we iterate for $n=0$ to $n=[1/h]+1$. Over one time step, the same calculations as in the continuous case give

$$\begin{aligned}
&\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x+i\beta(n+1)h,(n+1)h)|^2\,dx \\
&\le e^{2bh}\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x+i\beta nh,nh)|^2\,dx\ ,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\sup_{|y|\le L\pi}\int \varphi(x-y)|h\xi_N\star\big(G_1((n+1)h)w((n+1)h)+G_2((n+1)h)\overline{w}((n+1)h)\big)|^2\,dx \\
&\le (2Rh)^2\left(\int\frac{\xi_N^2}{\varphi}\right)\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x+i\beta(n+1)h,(n+1)h)|^2 \\
&\le e^{Ch}\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x+i\beta(n+1)h,(n+1)h)|^2\ ,
\end{aligned}$$

and hence we can iterate

$$\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x+i\beta nh,nh)|^2\,dx \;\le\; e^{2bnh}\sup_{|y|\le L\pi}\int \varphi(x-y)|w(x,0)|^2\,dx\ .$$

This completes the proof of Lemma 4.2.

Appendix B. Analyticity for the fully discrete scheme. The full discretization discussed in section 2.2 is similar to that introduced in [26], where Gevrey regularity is proved. We give here another simple and direct proof that the semigroup generated by (2.5) maps into $\mathcal{G}_\alpha(C)$ (see Appendix D) for some α and C independent

of N and L. Our proof is in the spirit of Collet [1] or Takáč et al. [25]. We assume that the solution $u(x, nh)$ of (2.5) has reached an absorbing ball in L^∞, and hence there is an $R > 0$ such that $\|u(nh)\|_\infty \le R$ irrespective of u_0 and n. We then use a contraction argument to show that, for small T, for $nh \in [0, T]$, there is a unique solution to (2.5) in the metric space of functions satisfying $|||u||| \le R$, where

$$|||f||| \;=\; \max_{nh\in[0,T]} \sup_{|x|\le L\pi} |f(x + i\sqrt{nh}, nh)| \;.$$

Remark that, if $T < h$, there is nothing to prove. (The solutions are entire functions anyway.) The purpose of this section is to provide bounds on the radius of analyticity which are independent of h and N, and hence we may assume h to be small.

We seek a solution to the equation $u(nh) = \mathcal{Y}\big(u, u_0\big)(nh)$ with $\mathcal{Y}$ defined by

$$\mathcal{Y}\big(f, f_0\big)(nh) \;=\; \mathcal{K}(nh) \star f_0 + \sum_{j=0}^{n-1} h\mathcal{K}\big(h(n-j)\big) \star \mathrm{P}^N F(f(jh)) \;,$$

where $\mathcal{K}$ is given by (2.4).

It is easy to see that, for small $T > 0$, $\mathcal{Y}(\cdot, f_0)$ is a contraction:

$$\begin{aligned}
&\big|\mathcal{Y}\big(f, f_0\big)(x + i\sqrt{nh}, nh) - \mathcal{Y}\big(g, f_0\big)(x + i\sqrt{nh}, nh)\big| \\
&\le \sum_{j=0}^{n-1} \int h|\mathrm{P}^N \mathcal{K}\big(y - z + i(\sqrt{nh} - \sqrt{jh}), h(n-j)\big)| \\
&\quad \times \Big|F\big(f(z + i\sqrt{jh}, jh)\big) - F\big(g(z + i\sqrt{jh}, jh)\big)\Big| dz \\
&\le \mathrm{Lip}(F, R)\, |||f - g||| \sum_{j=0}^{[T/h]} \int h|\mathrm{P}^N \mathcal{K}\big(x + i(\sqrt{nh} - \sqrt{jh}), h(n-j)\big)|\, dx.
\end{aligned}$$

Here $\mathrm{Lip}(F, R)$ is the Lipschitz constant of F in the ball of radius R, and hence by taking T small enough (depending on $\mathrm{Lip}(F, R)$ only) the solution to the fixed point problem exists and is unique. Since u belongs to an absorbing ball of L^∞, the argument can be iterated indefinitely, and hence u is analytic for all times thereafter.

Appendix C. Uniform bounds on complex analytic functions. In this section we show that an L^p bound in a strip of the complex plane provides an L^∞ bound in a smaller strip.

LEMMA C.1. *Let $p \ge 1$. There is a constant $C = C(\varphi, \delta)$ such that any function f analytic in $|\mathrm{Im}(x)| \le \delta$ satisfies*

$$|f(y + iz)|^p \;\le\; C \sup_{|\gamma|\le\delta} \int \varphi(x - y)|f(x + i\gamma)|^p \, dx$$

for all $y \in \mathbb{R}^d$ and $|z| \le \delta/2$.

Proof. We take $y = 0$ and $\delta = 1$ for simplicity. The general case is obtained by translation and scaling. Since analytic functions are harmonic the following mean value property holds (see [16]). Let $\mathcal{D}$ be the unit ball centered at 0 in the n-dimensional complex space; then

$$f(0) \;=\; \frac{1}{\mathrm{Vol}(\mathcal{D})} \int_{\mathcal{D}} f(x + i\gamma)\, dx\, d\gamma \;.$$

We apply Jensen's inequality and use that there is a C for which

$$\inf_{|x|\le 1} C\varphi(x) \;\ge\; 1$$

(see Definition 1.1) to obtain

$$\begin{aligned}|f(0)|^p &\le \frac{1}{\mathrm{Vol}(\mathcal{D})}\int_{\mathcal{D}} |f(x+i\gamma)|^p\,dx\,d\gamma \\ &\le \frac{1}{\mathrm{Vol}(\mathcal{D})}\sup_{|\gamma|\le 1}\int_{|x|\le 1} |f(x+i\gamma)|^p\,dx \\ &\le \frac{C}{\mathrm{Vol}(\mathcal{D})}\sup_{|\gamma|\le 1}\int \varphi(x)|f(x+i\gamma)|^p\,dx\ . \qquad \square\end{aligned}$$

Appendix D. Gevrey and Bernstein classes of analytic functions. We introduce here the metric spaces $\mathcal{B}_\sigma(C)$ (the Bernstein class) and $\mathcal{G}_\alpha(C)$ (the Gevrey class) and recall two properties of functions belonging to these spaces (see [7, 15, 18] for details).

DEFINITION D.1. *The Bernstein class $\mathcal{B}_\sigma(C)$ is the set of all functions f having an analytic extension to the whole of $\mathbb{C}^d$ with exponential growth along the imaginary directions:*

$$|f(x+iy)| \le Ce^{\sigma|y|}\ \forall (x,y)\in\mathbb{R}^d\times\mathbb{R}^d\ .$$

The Gevrey class $\mathcal{G}_\alpha(C)$ is the set of all functions f admitting an analytic extension to a strip of width 2α around the real axes and which are uniformly bounded in this strip:

$$|f(x+iy)| \le C\ \forall (x,y)\in\mathbb{R}^d\times[-\alpha,\alpha]^d\ .$$

The first result states that any function in $\mathcal{G}_\alpha(C)$ can be written as a sum of entire functions.

PROPOSITION D.2. *Let $f\in\mathcal{G}_\alpha(C)$. Then there exists a C' depending on C only such that*

$$f(z) = \sum_{n\in\mathbb{Z}^d} e^{-\alpha|n|}e^{in\cdot z}f_n(z)\ ,$$

with $f_n\in\mathcal{B}_2(C')$.

The second result is a classical sampling formula (see [7] or [15] where it is called the Cartwright formula).

PROPOSITION D.3. *For all $f\in\mathcal{B}_\sigma(C)$, the following identity holds:*

$$f(z) = \sum_{n\in\mathbb{Z}^d} f\big(x_\sigma(n)\big)\mathcal{F}_\sigma\big(z-x_\sigma(n)\big)\ ,$$

where

$$x_\sigma(n) = \frac{n\pi}{3\sigma}\ , \qquad \mathcal{F}_\sigma(x) = \frac{\sin(3\sigma x)\sin(\sigma x)}{3\sigma^2x^2}\ .$$

Acknowledgment. We are grateful to Jan Kristensen for useful discussions, especially in relation to Lemma C.1.

REFERENCES

[1] P. COLLET, *Nonlinear parabolic evolutions in unbounded domains*, in Dynamics, Bifurcation and Symmetry (Cargèse, 1993), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 97–104.

[2] P. Collet, *Thermodynamic limit of the Ginzburg-Landau equations*, Nonlinearity, 7 (1994), pp. 1175–1190.
[3] P. Collet, *Extended dynamical systems*, Doc. Math., Extra Vol. III (1998), pp. 123–132 (electronic).
[4] P. Collet and J.-P. Eckmann, *The definition and measurement of the topological entropy per unit volume in parabolic PDEs*, Nonlinearity, 12 (1999), pp. 451–473.
[5] P. Collet and J.-P. Eckmann, *Extensive properties of the complex Ginzburg-Landau equation*, Comm. Math. Phys., 200 (1999), pp. 699–722.
[6] P. Collet and J.-P. Eckmann, *Topological entropy and ε-entropy for damped hyperbolic equations*, Ann. Henri Poincaré, 1 (2000), pp. 715–752.
[7] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
[8] C. R. Doering, J. D. Gibbon, D. D. Holm, and B. Nicolaenko, *Low-dimensional behaviour in the complex Ginzburg-Landau equation*, Nonlinearity, 1 (1988), pp. 279–309.
[9] J.-M. Ghidaglia and B. Héron, *Dimension of the attractors associated to the Ginzburg-Landau partial differential equation*, Phys. D, 28 (1987), pp. 282–304.
[10] M. Gromov, *Entropy, homology and semialgebraic geometry*, Astérisque, 145-146 (1987), pp. 225–240, Séminaire Bourbaki, Vol. 1985/86.
[11] Z. Grujić and I. Kukavica, *Space analyticity for the Navier-Stokes and related equations with initial data in L^p*, J. Funct. Anal., 152 (1998), pp. 447–466.
[12] Z. Grujić and I. Kukavica, *Space analyticity for the nonlinear heat equation in a bounded domain*, J. Differential Equations, 154 (1999), pp. 42–54.
[13] J. Hale, X.-B. Lin, and G. Raugel, *Upper–semicontinuity of attractors for approximations of semigroups and partial differential equations*, Math. Comp., 50 (1988), pp. 89–123.
[14] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.
[15] A. N. Kolmogorov and V. M. Tikhomirov, *ε-entropy and ε-capacity of sets in functional space*, in Selected Works of A. N. Kolmogorov, Vol. III, A. N. Shiryayev, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 86–170.
[16] S. G. Krantz, *Function Theory of Several Complex Variables*, 2nd ed., Wadsworth and Brooks/Cole, Pacific Grove, CA, 1992.
[17] G. J. Lord and A. M. Stuart, *Discrete Gevrey regularity, attractors and upper-semicontinuity for a finite difference approximation to the complex Ginzburg-Landau equation*, Numer. Funct. Anal. Optim., 16 (1995), pp. 1003–1049.
[18] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, Cambridge, UK, 1992.
[19] A. Mielke, *The complex Ginzburg-Landau equation on large and unbounded domains: Sharper bounds and attractors*, Nonlinearity, 10 (1997), pp. 199–222.
[20] A. Mielke and G. Schneider, *Attractors for modulation equations on unbounded domains—existence and comparison*, Nonlinearity, 8 (1995), pp. 743–768.
[21] M. Oliver and E. S. Titi, *On the Domain of Analyticity for Solutions of Second Order Analytic Nonlinear Differential Equations*, preprint, 1999.
[22] M. Pollicott, *Lectures on Ergodic Theory and Pesin Theory on Compact Manifolds*, Cambridge University Press, Cambridge, UK, 1993.
[23] L. Schwartz, *Théorie des distributions*, Hermann, Paris, 1966.
[24] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.
[25] P. Takáč, P. Bollerman, A. Doelman, A. van Harten, and E. S. Titi, *Analyticity of essentially bounded solutions to semilinear parabolic systems and validity of the Ginzburg–Landau equation*, SIAM J. Math. Anal., 27 (1996), pp. 424–448.
[26] P. Takáč and A. Jüngel, *A nonstiff Euler discretization of the complex Ginzburg–Landau equation in one space dimension*, SIAM J. Numer. Anal., 38 (2000), pp. 292–328.
[27] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Springer-Verlag, New York, 1997.
[28] V. M. Tikhomirov, *On the ε-entropy of some classes of analytic functions*, Dokl. Akad. Nauk SSSR (N.S.), 117 (1957), pp. 191–194.
[29] Y. Yan, *Dimensions of attractors for discretizations for Navier-Stokes equations*, J. Dynam. Differential Equations, 4 (1992), pp. 275–340.
[30] Y. Yomdin, *Volume growth and entropy*, Israel J. Math., 57 (1987), pp. 285–300.
[31] S. V. Zelik, *An attractor of a nonlinear system of reaction-diffusion equations in $\boldsymbol{R}^n$ and estimates for its ε-entropy*, Mat. Zametki, 65 (1999), pp. 941–944.
[32] S. V. Zelik, *The attractor of a quasilinear hyperbolic equation with dissipation in $\boldsymbol{R}^n$: dimension and ε-entropy*, Mat. Zametki, 67 (2000), pp. 304–308.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1330–1353

COMPUTING THE EFFECTIVE HAMILTONIAN IN THE MAJDA–SOUGANIDIS MODEL OF TURBULENT PREMIXED FLAMES*

BOUALEM KHOUIDER† AND ANNE BOURLIOUX†

Abstract. Turbulence enhances the speed of propagation of a premixed flame front. According to the Majda–Souganidis model, the procedure to predict this enhancement involves computing the effective Hamiltonian in a small-scale nonlinear cell-problem. We first discuss how to transform this problem into computing the steady-state solution of a system of conservation laws whose vector solution represents the gradient of the eigenfunction associated with the effective Hamiltonian. Theoretical arguments as well as numerical evidence are presented to emphasize the importance of enforcing the constraint that the vector solution must effectively be the gradient of a scalar function. We introduce a scheme that satisfies this constraint exactly by relying on staggered grids for the gradient components. Also discussed is the issue of selecting a time integrator to achieve fast convergence to a steady state. Validation is performed by examining convergence under grid refinement and by comparison with analytical results when available.

Key words. gradient-preserving scheme, essentially nonoscillatory, staggered grid, Hamilton–Jacobi equation, conservation laws, steady state

AMS subject classifications. 65M06, 65M12, 65M25

PII. S003614290138872X

1. Introduction.

The flamelet regime in premixed combustion is characterized by a very thin reaction zone that separates burnt and fresh gas so that, for all practical matters, it can be viewed as an infinitely thin flame front, propagating normal to itself due to burning and advection. The speed of propagation of that interface can be easily predicted in the laminar case, where advection plays a trivial role. It is, however, much more difficult to predict its enhancement due to turbulence, when the front is wrinkled by a multiple scale advecting flow field. A rigorous asymptotic strategy to predict this enhancement has been developed by Majda and Souganidis [17] for a flow field with separate scales. According to the theory, the procedure to compute the enhanced burning speed involves minimizing a function of the effective Hamiltonian for the flame; the effective Hamiltonian must be computed as the eigenvalue of a nonlinear cell-problem. In [5], this procedure was implemented for the simple case of a one-dimensional shear layer; for that case, the solution can be expressed mostly through explicit formulas. The method presented in this paper extends the procedure for more general small-scale turbulent-like flows such as steady arrays of eddies or combinations of eddies and shears; for such cases, explicit formulas are no longer available and the problem must be solved numerically. Solving the cell-problem

*Received by the editors April 26, 2001; accepted for publication (in revised form) February 28, 2002; published electronically September 27, 2002. This work is part of the first author's Ph.D. thesis with support by scholarship grants from FCAR (Quebec Government) and the Institut des Sciences Mathématiques (Montreal).

http://www.siam.org/journals/sinum/40-4/38872.html

†Departement de Mathématiques et Statistiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montreal, Quebec, Canada, H3C2J7 (khouider@dms.umontreal.ca, Anne.Bourlioux@UMontreal.ca). The first author was also partially supported as a postdoctoral researcher at the Courant Institute by a grant from the U.S. Army Research Office (ARO-DAADI9-01-10810) during the final stages of this work. The research of the second author was supported by the Natural Sciences and Engineering Research Council of Canada and from the U.S. Army Research Office (ARO DAAG55-98-1-0220).

is the hardest part: it requires computing the eigenvalue of a Hamilton–Jacobi equation. Numerical methods for Hamilton–Jacobi equations (without eigenvalues) are well known [4, 18], and so are iterative methods to compute the eigenvalues for large linear systems. The present case combines both problems and the challenge from a numerical point of view is to provide a scheme capable of a robust handling of the nonlinearity of Hamilton–Jacobi equations and of an efficient search for the eigenvalue of the resulting discretized equations (hence, a large nonlinear system). The method presented in this paper tackles this challenge by reformulating the problem so that the eigenvalue is effectively eliminated from the preliminary phase of the computation by differentiation of the eigenvalue problem. This strategy leads to a very robust and practical scheme for two main reasons: the entire eigenvalue search is replaced by a simple algebraic postprocessing of the results instead of the typical iterative process; the system obtained by differentiation leads to a system of conservation laws which is a class of problems for which a well-established numerical machinery is available.

The paper is organized as follows. In section 2, the Majda–Souganidis asymptotic model equations are stated as well as the reformulation of the eigenvalue problem as one of finding a pseudotime steady-state solution for the eigenfunction gradient. In sections 3 and 4, some key theoretical properties of the equations are discussed to motivate the strategy to design the scheme. In section 3, the lack of strong hyperbolicity of the gradient equation is established. Yet, in section 4, convergence results for the equivalent time marching problem for the eigenfunction itself are exploited to formulate the principal constraint to guarantee that a discrete solution of the gradient problem will also converge to a steady state. This constraint is that the discrete vector solution be in some sense the discrete gradient of a scalar function. In section 5, we describe a novel *gradient-preserving* scheme, i.e., a scheme that explicitly preserves the gradient structure of the initial data throughout the computation, in some appropriate discrete sense. The fact that the gradient-preserving property is essential for convergence is further demonstrated by numerical experiments in section 6, where the performance of the gradient-preserving scheme is contrasted with that of other schemes which do not quite satisfy that constraint. A second order spatially accurate version of the scheme is presented in section 7, along with a discussion on how to select the time integrator to accelerate convergence to a steady state. In section 8, the performance of the method is validated systematically by comparison with the reference solutions for the case of a simple shear layer [5, 6].

2. Homogenization theory formulation.

2.1. The Majda–Souganidis asymptotic model. Here we simply state the model equations to be solved numerically; details regarding the derivation of the model can be found in [5, 17]. Assuming that the heat release due to combustion is weak and that temperature and all the relevant chemical species diffuse at the same rate (Lewis number unity), the flame propagation can be described using a single advection-diffusion-reaction equation for temperature. The homogenization theory that leads to the model equations (2.1) and (2.2) below applies under the following additional assumptions:

- The reaction zone is thin, as a result of the balance between very weak diffusion and very fast reaction.
- The incompressible advecting velocity field includes two separate scales: one large scale and a scale intermediate between the large scale and the flame thickness.
- The reaction rate is of the Kolmogorov–Petrovskii–Piskunov type. A typical

example of such reaction rate is given by $f(T) = \overline{K}\,T(1-T)$, where the temperature T has been normalized between $T = 0$ on the cold (unburnt) side and $T = 1$ on the hot (burnt) side and $\overline{K} > 0$ is the reaction rate constant.

The intermediate scale velocity field causes the flame front to wrinkle: qualitatively, the increase in flame area due to this wrinkling leads to an overall burning speed enhancement. The objective of the homogenization theory is to predict this enhanced speed of propagation of the flame. The equations in the form stated below describe how to do this assuming that, at large scales, the advecting velocity field is constant and the flame front is planar.

Given the flame front unit normal $\mathbf{n} = (\cos\theta, \sin\theta)$, the flame speed $F(\mathbf{n})$ in that direction must be computed as

$$F(\mathbf{n}) = \min_{r>0} \frac{H(r\mathbf{n}) + \bar{K}}{r}, \tag{2.1}$$

where $\overline{K} = f'(0)$ is the positive constant used to define the reaction rate above and H is the effective Hamiltonian of the flame, computed as the unique eigenvalue of the following so-called *cell-problem*:

$$-|\mathbf{p} + \mathbf{D}w|^2 + \mathbf{V}(\mathbf{y}) \cdot (\mathbf{p} + \mathbf{D}w) = -H(\mathbf{p}). \tag{2.2}$$

(At least for steady flows, it is trivial from a numerical point of view to deal with the unsteady terms—they are linear—so we will not discuss this issue in this paper; see [14] for examples with unsteady flows.) The eigenfunction $w(\mathbf{y})$ must be of zero mean and biperiodic with respect to the spatial variables $\mathbf{y} = (x, y)$. (One can always assume that the biperiodic domain has been rescaled to a unit square.) $\mathbf{D}w$ represents the spatial gradient of the eigenfunction. The velocity field $\mathbf{V}(\mathbf{y})$ is assumed to combine a large-scale constant flow and the smaller scale "turbulent" flow:

$$\mathbf{V} = \bar{\mathbf{v}} + \lambda\mathbf{v} = \bar{\lambda}(\cos\bar{\theta}, \sin\bar{\theta}) + \lambda\mathbf{v}(\mathbf{y}).$$

Here, $\bar{\lambda}$ and λ represent the magnitude of the velocity field, respectively, at large and intermediate scales, while $\mathbf{v}(\mathbf{y})$ is the intermediate scale velocity field defined over the unit periodic box. It is also assumed to have a zero mean and to be biperiodic as well as to be incompressible. For example, in section 8 below and elsewhere [2, 15], we use our procedure on velocity fields obtained from the Childress–Soward stream function $\psi(\mathbf{y})$:

$$\psi(\mathbf{y}) = \psi(x, y) = \sin(2\pi x)\sin(2\pi y) + \delta\cos(2\pi x)\cos(2\pi y), \qquad 0 \le \delta \le 1. \tag{2.3}$$

Streamlines for $\delta = 0, 0.5, 1$ are shown in Figure 2.1.

In summary, the input data are

- the front angle θ;
- δ, λ: the parameters that define the "turbulent" velocity field responsible for the burning speed enhancement;
- $\overline{\lambda}$, $\overline{\theta}$: the parameters that define the large-scale (constant) velocity field.

Results are typically presented in terms of the flame speed enhancement F_e defined as

$$F_e = F + \overline{\mathbf{v}} \cdot \mathbf{n} - S_L \tag{2.4}$$

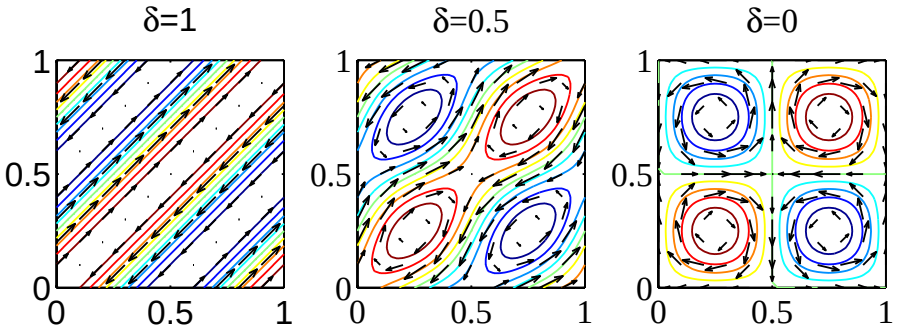


FIG. 2.1. *Streamlines for the Childress–Soward flow.* $\delta = 1$*: simple shear tilted at* 45 *degrees,* $\delta = 0.5$*: combination of eddies and shear layers, and* $\delta = 0$*: periodic array of eddies.*

with $S_L = 2/\sqrt{\overline{\overline{K}}}$ the laminar burning speed of the flame. In the rest of the paper, the problem is rescaled so that $S_L = 1$; F_e represents the increase in burning speed of the flame compared to the laminar case. A qualitative understanding of the wrinkling of the front can be achieved by looking at the isolevels of the eigenfunction w: they can be related to successive realizations of the wrinkled flame front as it moves across the periodic cell due to burning and advection.

The main challenge in performing the optimization over the variable r in (2.1) is to provide the effective Hamiltonian $H(r\mathbf{n})$, which must be computed as the eigenvalue in (2.2.) The ultimate objective here is to design a method sufficiently efficient to allow for the extensive tabulation of F_e as a function of all the input data. Each entry in such a table requires iterating numerically over many values of r to perform the minimization in (2.1), hence requiring multiple evaluations of H; the main issue is therefore to design an efficient, robust, fully automated algorithm to compute this effective Hamiltonian for a wide range of parameters.

Equation (2.2) is both nonlinear and contains an eigenvalue (the effective Hamiltonian H): it is very challenging numerically to deal simultaneously with both difficulties, and it is therefore very tempting to try to avoid dealing directly with at least one of them. One possible strategy along that line would be to linearize the problem, hence eliminating the nonlinearity aspect and retaining the eigenvalue; this could be done by adding a small viscous term of order ϵ and doing the following transformation: $T^\epsilon(x, y) = \exp(Z(x, y)/\epsilon)$ with $Z(x, y) = \mathbf{p} \cdot (x, y) + w(x, y)$; formulating the problem in terms of T^ϵ instead of w basically undoes one of the steps in the homogenization procedure in [17]. The main difficulty with this approach is that the amount of viscosity ϵ needed to lead to a system of equations that can be safely discretized by a centered scheme is not known a priori but is solution-dependent. Another difficulty is that the discretized equations would lead to a very large linear eigenvalue problem, requiring a costly iterative procedure for its numerical solution along with the need of generating an adequate initial guess for the eigenvalue.

In that sense, the strategy to be discussed in the rest of the paper is much more robust. It is the eigenvalue aspect that is eliminated from the formulation by differentiating the cell-problem, resulting in a system of equations for the eigenfunction gradient that is still nonlinear but that no longer contains the eigenvalue explicitly. This strategy is described next.

2.2. Reformulation as a gradient problem. By differentiating (2.2) with respect to $\mathbf{y}$, we eliminate the eigenvalue and obtain the following nonlinear equation for the eigenfunction gradient $\mathbf{D}w$:

$$\mathbf{D}\{-|\mathbf{D}w|^2 - 2\mathbf{p}\cdot\mathbf{D}w + (\overline{\mathbf{v}} + \lambda\mathbf{v}(\mathbf{y}))\cdot\mathbf{D}w + \lambda\mathbf{v}(\mathbf{y})\cdot\mathbf{p}\} = 0. \tag{2.5}$$

Integrating (2.2) over one periodic cell and using the divergence theorem, one obtains a formula that relates algebraically H and $\mathbf{D}w$:

$$H(\mathbf{p}) = |\mathbf{p}|^2 + \langle|\mathbf{D}w|^2\rangle - \overline{\mathbf{v}}\cdot\mathbf{p}, \tag{2.6}$$

where $\langle\cdot\rangle$ represents the average over the periodic cell. Therefore, assuming that indeed one is able to solve (2.2), the effective Hamiltonian H can be computed very economically a posteriori by simple postprocessing of the solution for $\mathbf{D}w$, hence avoiding entirely any eigenvalue iterative procedure.

A practical approach to compute the solution of (2.5) is to view it as the steady state of the following system:

$$\begin{cases} \partial_s u + \partial_x K(u,v,x,y) = 0, \\ \partial_s v + \partial_y K(u,v,x,y) = 0, \end{cases} \tag{2.7}$$

where K is given by

$$K(\mathbf{u},x,y) = -|\mathbf{u}|^2 - 2\mathbf{p}\cdot\mathbf{u} + (\overline{\mathbf{v}} + \lambda\mathbf{v}(x,y))\cdot\mathbf{u} + \lambda\mathbf{v}(x,y)\cdot\mathbf{p} \tag{2.8}$$

and $\mathbf{u} = (u,v) = \mathbf{D}w$ is the eigenfunction gradient.

The pseudotime marching method to solve (2.7) to a steady state is described in detail in section 5 below. Before describing the numerical method, however, some results regarding the equation before discretization are reported in section 3 (lack of strong hyperbolicity) and section 4 (effective convergence to a steady state); those results provide essential insight on the type of constraints to be taken into account in order to design a successful numerical method, with a particular concern for the convergence property of the algorithm toward a steady-state solution.

3. Lack of strong hyperbolicity. Here we prove that (2.7) is not strongly hyperbolic. Set $F = (F_1, F_2)$ with

$$F_1(\mathbf{u},x,y) = \begin{pmatrix} K(\mathbf{u},x,y) \\ 0 \end{pmatrix} \text{ and } \quad F_2(\mathbf{u},x,y) = \begin{pmatrix} 0 \\ K(\mathbf{u},x,y) \end{pmatrix}.$$

The associated system is strongly hyperbolic if for all reals α and β the matrix

$$A = \alpha\frac{\partial F_1}{\partial \mathbf{u}} + \beta\frac{\partial F_2}{\partial \mathbf{u}},$$

where

$$\frac{\partial F_1}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial K}{\partial u} & \frac{\partial K}{\partial v} \\ 0 & 0 \end{bmatrix}$$

and

$$\frac{\partial F_2}{\partial \mathbf{u}} = \begin{bmatrix} 0 & 0 \\ \frac{\partial K}{\partial u} & \frac{\partial K}{\partial v} \end{bmatrix},$$

has two real eigenvalues and two linearly independent eigenvectors [10].

The eigenvalues of the matrix A are $\nu = \alpha\frac{\partial K}{\partial u} + \beta\frac{\partial K}{\partial v}$ and 0. Clearly, if $\frac{\partial K}{\partial v} \neq 0$ or $\frac{\partial K}{\partial v} \neq 0$ one can choose the constants α and β such that ν is zero and the matrix A is not identically zero. This is equivalent to setting

$$\begin{cases} \dfrac{\partial K}{\partial u} \neq 0 \text{ or } \dfrac{\partial K}{\partial v} \neq 0, \\ \alpha\dfrac{\partial K}{\partial u} + \beta\dfrac{\partial K}{\partial v} = 0. \end{cases} \tag{3.1}$$

In this case, the matrix A has only one free eigenvector associated with the double eigenvalue 0; i.e., A is equivalent to a Jordan block. Thus, (2.7) is not strongly hyperbolic. The lack of strong hyperbolicity for the pseudotime marching equation has important consequences regarding the possibility of reaching numerically a steady state by long time marching:

1. Without taking into account other specific properties of the system studied here (as will be done in section 4 below), the lack of strong hyperbolicity means that there is no guarantee that the solution will converge to a steady state [8].
2. Standard numerical methods for conservation laws might not work, as most rely on the strong hyperbolic nature of the equations.

Remark 1. The vector solution to be computed here is actually a gradient, a property which was not taken into account in the discussion above. It is interesting to notice that imposing this additional constraint on the solution is not, per se, sufficient to recover strong hyperbolicity.

Take $\mathbf{u} = \mathbf{D}\phi$ and $\mathbf{V} = \overline{\mathbf{v}} + (\psi_y, -\psi_x)$ for some periodic functions ϕ and ψ with zero mean, a mean flow $\overline{\mathbf{v}} = (\overline{v}_1, \overline{v}_2)$, and a mean flame gradient $\mathbf{p} = (p_1, p_2)$. The conditions of nonhyperbolicity (3.1) are equivalent to

$$\begin{cases} 2\phi_x + 2p_1 - \overline{v}_1 - \psi_y \neq 0 \text{ or } 2\phi_y + 2p_2 - \overline{v}_2 + \psi_x \neq 0, \\ \alpha(2\phi_x + 2p_1 - \overline{v}_1 - \psi_y) + \beta(2\phi_y + 2p_2 - \overline{v}_2 + \psi_x) = 0. \end{cases}$$

It is easy to construct examples that would satisfy those nonhyperbolicity conditions. For example, set $\phi(x, y) = \psi(-y, x)/2$ so that the conditions

$$2\phi_x = \psi_y \text{ and } 2\phi_y = -\psi_x$$

are satisfied on the curve $y = -x$ and choose the constant parameters such that

$$2p_1 - \overline{v}_1 \neq 0 \text{ or } 2p_2 - \overline{v}_2 \neq 0$$

and

$$\alpha(2p_1 - \overline{v}_1) + \beta(2p_2 - \overline{v}_2) = 0.$$

4. Convergence to a steady state. The main conclusion from section 3 is that if one views the system of equations in (2.7) as a system of conservation laws for a vector solution, there is no guaranty that, starting from general initial data, convergence to a steady state can be achieved by pseudotime marching because of the lack of strong hyperbolicity of the system of equations (2.7).

Such a convergence result, however, can be recovered by first considering the convergence of the following pseudotime marching equation for the eigenfunction itself:

$$w_s - |\mathbf{p} + \mathbf{D}w|^2 + \mathbf{V}(\mathbf{y}) \cdot (\mathbf{p} + \mathbf{D}w) = -H(\mathbf{p}), \tag{4.1}$$

with H the eigenvalue of the cell-problem in (2.2). Convergence of $w + Hs$ to a steady solution was proved by Barles and Souganidis [1] for general initial data. This long-time convergence property automatically implies convergence, in the weak sense, of the gradient $\mathbf{D}w$, therefore establishing that the pseudotime iterations on (2.7) should converge, at least theoretically (i.e., before numerical discretization). The key implication of this remark relevant to the design of a scheme to solve those equations numerically is that one way to recover convergence for the gradient system is to ensure that the vector solution of (2.7) is effectively the gradient, in some appropriate discrete sense, of a scalar function.

For smooth functions, a vector function is a gradient if it is curl-free. We state next an equivalent definition to the curl-free condition which does not involve derivatives and hence constitutes a useful generalization to weak derivatives. This equivalent definition will also turn out to be very useful in dealing with discrete data, as is done in the next section. Let (u, v) be an integrable vector function:

- (u, v) is the gradient of a function w if and only if, given a reference point (a, b), we have the double equality

$$w(x, y) = \int_a^x u(\xi, y)\, d\xi + w(a, y) = \int_b^y v(x, \xi)\, d\xi + w(x, b)$$

for a.e. each point (x, y). However, similarly,

$$w(a, y) = \int_b^y v(a, \xi)\, d\xi + w(a, b)$$

and

$$w(x, b) = \int_a^x u(\xi, b)\, d\xi + w(a, b);$$

hence, without referring to any primitive, we can state that $(u(x, y), v(x, y))$ is a gradient in the weak sense if and only if for a.e. (x, y) we have

$$\int_b^y v(a, \xi)\, d\xi + \int_a^x u(\xi, y)\, d\xi = \int_a^x u(\xi, b)\, d\xi + \int_b^y v(x, \xi)\, d\xi. \tag{4.2}$$

- In addition, the primitive function is biperiodic of period T if and only if for a.e. (x, y)

$$\int_a^{a+T} u(\xi, y)\, d\xi = 0, \tag{4.3}$$

$$\int_b^{b+T} v(x, \xi)\, d\xi = 0. \tag{4.4}$$

5. Gradient-preserving scheme. Brute force attempts at numerically solving (2.5) to a steady state with standard numerical schemes for conservation laws failed to converge, and this comes as no surprise given the considerations in sections 3 and 4 above. Instead, the results in section 4 suggest that convergence could be guaranteed only if the discrete vector solution is actually a gradient. The method we propose here has precisely this property: the scheme is *gradient-preserving* in the discrete sense of formula (4.2) inasmuch as, given initial vector data which are a discrete gradient, the solution will remain a discrete gradient at all later discrete times. (In practice, the simplest such initial data are identically zero.) The scheme is actually based on a fairly standard conservative formulation except for the staggered discretization grids to be described first.

5.1. Staggered grids. Let $h = 1/n$ define the mesh discretization of the interval $[0,1]$ with $x_i = ih$, $x_{i+1/2} = x_i + h/2$, $y_j = jh$, and $y_{j+1/2} = y_j + h/2$ for any integers $0 \le i, j \le n$. We consider the staggered grid obtained by the superposition of the two grids (x_i, y_j) and $(x_{i+1/2}, y_{j+1/2})$ (see Figure 5.1).

The first component, u, of the gradient solution is defined at the nodes (x_i, y_j), while the second component, v, is defined at the nodes $(x_{i+1/2}, y_{j+1/2})$. The scalar primitive (the eigenfunction) w itself is defined at the hybrid nodes $(x_{i+1/2}, y_j)$.

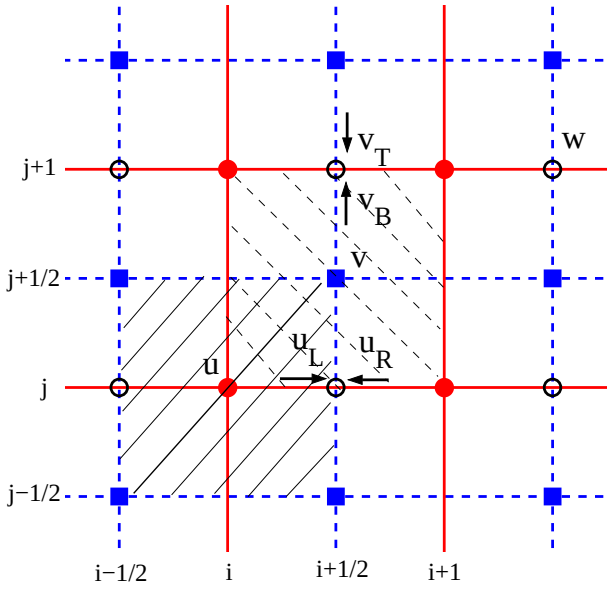


FIG. 5.1. *Staggered grid for a gradient-preserving scheme: u is defined on the vertices (i,j) by its cell averages on $[i-1/2, i+1/2]\times[j-1/2, j+1/2]$ (filled circles), v on the vertices $(i+1/2, j+1/2)$ by its cell averages on $[i, i+1]\times[j, j+1]$ (filled squares), and the primitive w is obtained on $(i+1/2, j)$ (empty circles). The shaded squares show the control volumes for u and v, respectively, and the arrows point out the associated Riemann problems.*

5.2. Conservative formulation. Using the notation $\mathbf{V} = (V_1, V_2)$ and $\mathbf{p} = (p_1, p_2)$, the starting point of the strategy is to notice the symmetry in the flux function K in (2.8), which can be split into two independent parts, $K(u, v, x, y) = f(u, x, y) + g(v, x, y)$, with

$$f = f(u, x, y) = -u^2 - 2p_1 u + V_1(x, y)u + V_1(x, y)p_1$$

and

$$g = g(v, x, y) = -v^2 - 2p_2 v + V_2(x, y)v + V_2(x, y)p_2.$$

The conservative formulation is obtained by integrating the conservation laws over each u- and v-control volume. Taking the integral of

$$\partial_s u + \partial_x f(u, x, y) + \partial_x g(v, x, y) = 0$$

over the cell $[x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}] \times [s_n, s_{n+1}]$ and the integral of

$$\partial_s v + \partial_y f(u, x, y) + \partial_y g(v, x, y) = 0,$$

over $[x_{i-1}, x_i] \times [y_{j-1}, y_j] \times [s_n, s_{n+1}]$ and then dividing by the control volume area h^2, one obtains the following equations in conservative form:

$$\begin{aligned} \overline{u}_{i,j}^{n+1} = \overline{u}_{i,j}^{n} &- \frac{1}{h^2} \int_{s_n}^{s_{n+1}} \int_{y_{j-1/2}}^{y_{j+1/2}} \left\{ f(u, x_{i+1/2}, y) + g(v, x_{i+1/2}, y) \right\} dy\, ds \\ &+ \frac{1}{h^2} \int_{s_n}^{s_{n+1}} \int_{y_{j-1/2}}^{y_{j+1/2}} \left\{ f(u, x_{i-1/2}, y) + g(v, x_{i-1/2}, y) \right\} dy\, ds \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} \overline{v}_{i-1/2,j-1/2}^{n+1} = \overline{v}_{i-1/2,j-1/2}^{n} &- \frac{1}{h^2} \int_{s_n}^{s_{n+1}} \int_{x_{i-1}}^{x_i} \left\{ f(u, x, y_j) + g(v, x, y_j) \right\} dx\, ds \\ &+ \frac{1}{h^2} \int_{s_n}^{s_{n+1}} \int_{x_{i-1}}^{x_i} \left\{ f(u, x, y_{j-1}) + g(v, x, y_{j-1}) \right\} dx\, ds, \end{aligned} \tag{5.2}$$

where

$$\overline{u}_{i,j}^{k} = \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \int_{y_{j-1/2}}^{y_{j+1/2}} u(x, y, s_k) dx dy$$

is the cell average of u with the obvious corresponding considerations for v. This can be expressed simply as

$$\overline{u}_{i,j}^{n+1} = \overline{u}_{i,j}^{n} - \frac{\Delta s}{h} \left(F_{i+1/2,j}^{n,1} - F_{i-1/2,j}^{n,1} \right) \tag{5.3}$$

and

$$\overline{v}_{i-1/2,j-1/2}^{n+1} = \overline{v}_{i-1/2,j-1/2}^{n} - \frac{\Delta s}{h} \left(F_{i-1/2,j+1}^{n,2} - F_{i-1/2,j}^{n,2} \right), \tag{5.4}$$

where the flux $F_{i-1/2,j}^{n,1}$ corresponds to the double space-time integral in (5.1) over one time-step and over the vertical edge (of a u-cell) with the edge center located at $(x_{i-1/2}, y_j)$:

$$F_{i-1/2,j}^{n,1} = \frac{1}{h\,\Delta s} \int_{s_n}^{s_{n+1}} \int_{y_{j-1/2}}^{y_{j+1/2}} \left\{ f(u, x_{i+1/2}, y) + g(v, x_{i+1/2}, y) \right\} dy\, ds, \tag{5.5}$$

while the flux $F^{n,2}_{i-1/2,j}$ corresponds to the double space-time integral in (5.2) over one time-step and over the horizontal edge (of a v-cell) with the edge center also located at $(x_{i-1/2}, y_j)$:

$$F^{n,2}_{i-1/2,j} = \frac{1}{h\,\Delta s}\int_{s_n}^{s_{n+1}}\int_{x_{i-1}}^{x_i}\{f(u,x,y_{j-1}) + g(v,x,y_{j-1})\}\,dx\,ds. \tag{5.6}$$

As will be shown in Proposition 5.1 below, the key in designing a gradient-preserving scheme is to use the same numerical value for $F^{n,1}_{i-1/2,j}$ and $F^{n,2}_{i-1/2,j}$. This can be justified as follows. In evaluating $F^{n,1}_{i-1/2,j}$, for instance, it is clear that the f-component of the flux (corresponding to the integration of $f(u,x,y)$) should be estimated using a standard Riemann solver, with left and right states corresponding to $\bar{u}^n_{i-1,j}$ and $\bar{u}^n_{i,j}$, respectively (for a first order method at least). The g-component of the flux, however, can be specified somewhat more arbitrarily because it depends only on v and is independent of u. A second order accurate choice is to estimate that portion of the integral by its value at the midpoint of the edge $(x_{i-1/2}, y_j)$, in which case this contribution can be shown to be equal to second order to the g-component of the $F^{n,2}_{i-1/2,j}$ flux, to be estimated by resorting again to a Riemann solver, this time with top and bottom states given by $\bar{v}^n_{i-1/2,j}$ and $\bar{v}^n_{i-1/2,j-1}$, respectively. Specifically,

$$F^{n,1}_{i-1/2,j} = F^{n,2}_{i-1/2,j} = F^n_{i-1/2,j} = f(\Re^1_{i-1/2,j}, x_{i-1/2}, y_j) + g(\Re^2_{i-1/2,j}, x_{i-1/2}, y_j), \tag{5.7}$$

where $\Re^1_{i-1/2,j}$ is the solution of the Riemann problem for the flux f associated with the vertical edge $(i-1/2, j)$ and the left and right states u_L and u_R ($\overline{u}_{i-1,j}$ and $\overline{u}_{i,j}$ for the first order scheme) and $\Re^2_{i-1/2,j}$ is the solution for the Riemann problem for the flux g at the horizontal edge $(i-1/2, j)$ for the bottom and top states v_B and v_T ($\overline{v}_{i-1/2,j-1/2}$ and $\overline{v}_{i-1/2,j+1/2}$ for the first order scheme); see Figure 5.1. The scheme can be rewritten as

$$\begin{aligned}\overline{u}^{n+1}_{i,j} &= \overline{u}^n_{i,j} - \frac{\Delta s}{h}\left\{F^n_{i+1/2,j} - F^n_{i-1/2,j}\right\},\\ \overline{v}^{n+1}_{i-1/2,j-1/2} &= \overline{v}^n_{i-1/2,j-1/2} - \frac{\Delta s}{h}\left\{F^n_{i-1/2,j} - F^n_{i-1/2,j-1}\right\}.\end{aligned} \tag{5.8}$$

PROPOSITION 5.1. *The scheme in* (5.8) *is gradient preserving in the sense that if the numerical vector solution satisfies the condition* (4.2) *at some given initial time, s_0, then this condition will be satisfied at any latter time, $s_n > s_0$. Furthermore, if the primitive function is periodic at the initial time, then it remains periodic; i.e.,* (4.3) *and* (4.4) *are also satisfied at time s_n if they were satisfied at time s_0.*

Proof. It is straightforward to verify the second part of the proposition by exploiting the conservative formulation. To prove the first claim, recall that the discrete values representing the numerical solution for (u, v) are cell averages, so, given these values at any time, s_n, we can get the solution primitive at the hybrid vertices $((i-1/2), j)$, without any further approximation. Taking $(a, b) = (x_{1/2}, y_0)$ and $(x, y) = (x_{i-1/2}, y_j)$ in (4.2) leads to

$$\sum_{k=1}^{i-1}\overline{u}_{k,0} + \sum_{k=1}^{j}\overline{v}_{i-1/2,k-1/2} = \sum_{k=1}^{j}\overline{v}_{1/2,k-1/2} + \sum_{k=1}^{i-1}\overline{u}_{k,j}. \tag{5.9}$$

To show that the scheme is gradient preserving, we must show that if the discrete solution at time s_n verifies (5.9), then so does the solution at time $s_{n+1} = s_n + \Delta s$, which is equivalent to showing that the same condition is satisfied by the difference of the two vector solutions in (5.8), i.e.,

$$\sum_{k=1}^{i-1} F^n_{k+1/2,0} - F^n_{k-1/2,0} + \sum_{k=1}^{j} {F^n}_{i-1/2,k} - {F^n}_{i-1/2,k-1}$$
$$= \sum_{k=1}^{j} {F^n}_{1/2,k} - {F^n}_{1/2,k-1} + \sum_{k=1}^{i-1} {F^n}_{k+1/2,j} - {F^n}_{k-1/2,j},$$

and the two sides of the equality collapse to their common value

$${F^n}_{i-1/2,j} - {F^n}_{1/2,0};$$

thus, (5.8) is gradient preserving. □

Remark 2. A standard von Neumann stability analysis [11, 12] applied to the model problem

$$\begin{cases} u_s + a u_x + b v_y = 0, \\ v_s + a u_x + b v_y = 0 \end{cases} \tag{5.10}$$

(with $a, b > 0$) shows that the gradient-preserving scheme in (5.8) is linearly stable under the CFL condition

$$\lambda_1 + \lambda_2 \leq 1,$$

where $\lambda_1 = a\Delta s/h$ and $\lambda_2 = b\Delta s/h$ [13].

6. Failure of non-gradient-preserving schemes. The constraint of formulating a gradient-preserving scheme was motivated theoretically in sections 3 and 4. Here, we provide more practical motivations by reporting the results from failed numerical experiments with three schemes that do not quite satisfy that constraint.

6.1. Roe's scheme (GNPS1). The first scheme is the standard Roe's Riemann solver approximation scheme for hyperbolic systems [16]. The scheme relies heavily on hyperbolic features, so it obviously may not work for our system because of the lack of strong hyperbolicity.

6.2. Direction-splitting on a single uniform grid (GNPS2). The second scheme is obtained by giving up the staggered grid, instead using the same uniform grid for the two components (u, v). The direction-splitting leads to the following expressions for the edge fluxes:

$$F^{n,1}_{i-1/2,j} = f(\Re^1_{i-1/2,j}, x_{i-1/2}, y_j) + g(v^n_{i-1/2,j}, x_{i-1/2}, y_j),$$
$$F^{n,2}_{i,j-1/2} = f(u^n_{i,j-1/2}, x_i, y_{j-1/2}) + g(\Re^2_{i,j-1/2}, x_i, y_{j-1/2}),$$

where one can simply estimate

$$v^n_{i-1/2,j} = (v^n_{i-1,j} + v^n_{i,j})/2$$

and a similar expression for $u^n_{i,j-1/2}$. Those centered approximations are adequate here because of the directional-splitting.

6.3. Staggered grid without flux approximation (GNPS3). For the third scheme, we retain the staggered grid in Figure 5.1 and the conservative formulation in (5.1) and (5.2) but give up the centered quadrature approximation for the integrals, using instead the exact integration in space-time of the corresponding Riemann problem. For instance,

$$F^{n,1}_{i-1/2,j} = f(\Re^1_{i-1/2,j}, x_{i-1/2}, y_j) + \frac{1}{h\,\Delta s}\int_{s_n}^{s_{n+1}}\int_{y_{j-1/2}}^{y_{j+1/2}} g(v_{\Re^2}(y,s), x_{i-1/2}, y)dy\,ds,$$

where $v_{\Re^2}(y,s)$ represents the detailed solution in space-time of the Riemann problem with initial states (v_B, v_T). The gradient-preserving scheme simply replaces the detailed expression for $v(y,s)$ in the integral by the constant value $v(y = y_j, s) = \Re^2_{i-1/2,j}$.

Except for the fact that they do not automatically preserve gradients, the last two test schemes are actually very similar to the gradient-preserving scheme (5.8), and one could expect them to perform similarly. In Figure 6.1, however, the difference in performance is striking. The eigenvalue is seen to grow indefinitely with time for each one of the three alternative test schemes and converges to a steady state only in the case of the gradient-preserving scheme. This confirms the theoretical intuition that, to achieve convergence, the gradient-preserving property is essential to compensate for the lack of strong hyperbolicity in the system of conservation laws for the gradient.

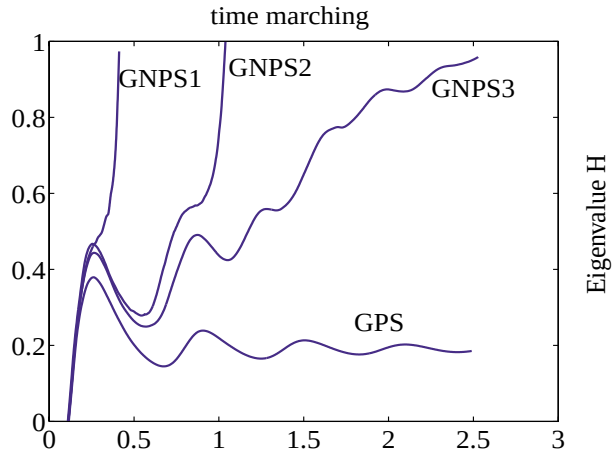


FIG. 6.1. *Effective Hamiltonian H(r) as a function of pseudotime; comparison of the performance of the gradient-preserving scheme (GPS) with the three non-gradient-preserving schemes (GNPS) described in section* 6. *Spatial resolution* 40×40, 800 *pseudotime iterations with CFL*=0.45*;* $\lambda = 1$, $\overline{\lambda} = 2$, $\theta = \overline{\theta} = \delta = 0$.

7. Efficient second order scheme. A second order version of the scheme is designed by resorting to the essentially nonoscillatory (ENO) interpolation strategy in space, coupled with a Runge–Kutta time integrator [12, 16].

Increasing the spatial order of accuracy to second order will be shown in section 8 to improve significantly the efficiency of the method, as much less resolution is required to achieve a given accuracy with the higher order method than with the first order scheme. However, pseudotime accuracy is not needed here: the main consideration for an efficient scheme is that a converged state be reached in as few pseudotime steps as possible, so as to minimize the overall cost of the calculation.

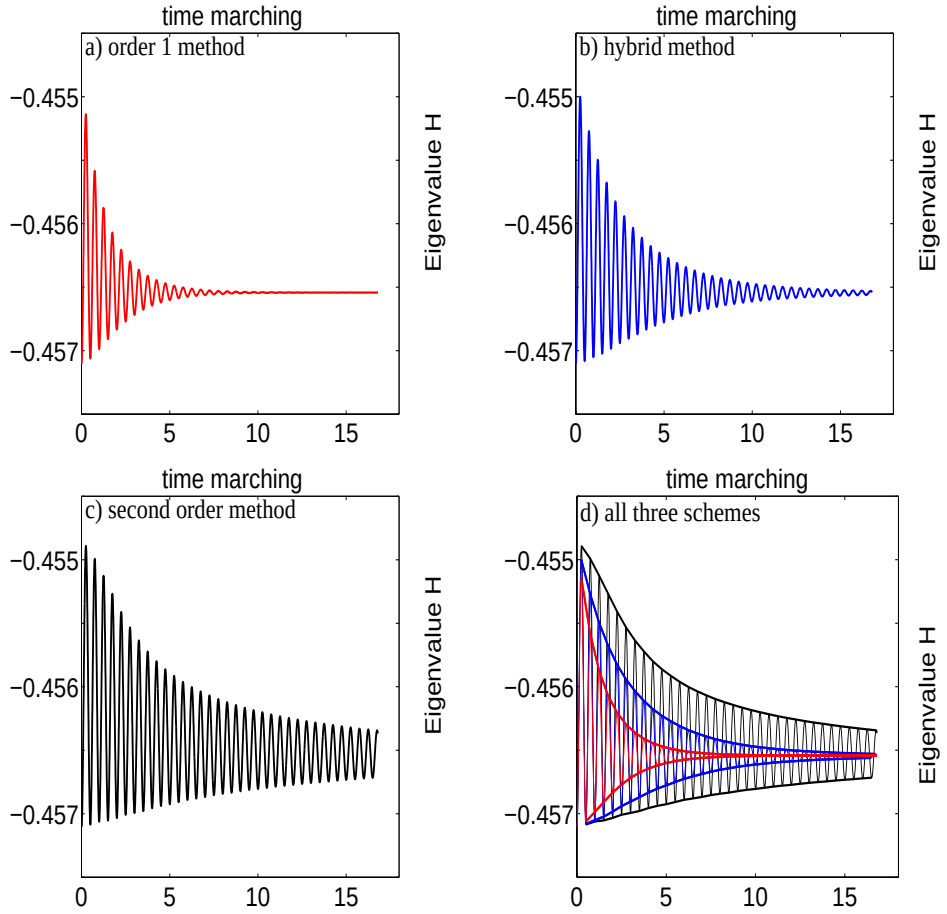


FIG. 7.1. *Effective Hamiltonian H as a function of pseudotime, comparing the performance of the hybrid scheme (second order in space, first order in time) compared with the overall second order scheme and with the first order scheme. Spatial resolution* 40×40*, 3200 time iterations with* $CFL = 0.45$*;* $\overline{\lambda} = 2$*,* $\theta = \pi/4$*,* $\delta = 0.5$ *and* $\lambda = 0.1$*.*

Numerical experiments such as the one reported in Figure 7.1 (a) and (c) reveal that the ENO second order scheme coupled with a standard Runge–Kutta method of the same order does not converge to a steady state as rapidly as the first order scheme previously discussed. This is a serious drawback if the scheme is to be used repeatedly for tabulation. In the rest of this section, we gain some insight on the time-convergence properties of various two-step Runge–Kutta time integrators by studying their linear stability for two limit versions of the ENO scheme: the two limit schemes are obtained by artificially freezing the discretization stencils and correspond, respectively, to the best (upwind) and worst (downwind) case scenario. We use this insight to identify constraints on the coefficients of the Runge–Kutta integrator likely to lead to good damping properties and fast convergence to a steady state.

7.1. Order of accuracy. To analyze the stability and accuracy properties of a generic explicit two-step Runge–Kutta scheme for our set of equations, we introduce first the difference operator $\mathcal{A}$ associated with the original first order scheme (5.8);

i.e.,

$$\mathcal{A}\begin{pmatrix} \overline{u}_{i,j} \\ \overline{v}_{i-1/2,j-1/2} \end{pmatrix} = \frac{1}{h}\begin{pmatrix} F_{i+1/2,j} - F_{i-1/2,j} \\ F_{i-1/2,j} - F_{i-1/2,j-1} \end{pmatrix}$$

for which the left and right states in the Riemann problems are reconstructed by the primitive functions according to the ENO second order interpolation. Set

$$U^n = \begin{pmatrix} \overline{u}^n_{i,j} \\ \overline{v}^n_{i-1/2,j-1/2} \end{pmatrix}$$

so that the Euler version of our scheme can be written

$$U^{n+1} = U^n - \Delta s \mathcal{A}(U^n).$$

We consider the Runge–Kutta two-step method

$$\begin{aligned} U^{n+1/2} &= U^n - \beta_0 \Delta s \mathcal{A}(U^n), \\ U^{n+1} &= \alpha_1 U^n + \alpha_2 U^{n+1/2} - \Delta s \beta_1 \mathcal{A}(U^n) - \Delta s \beta_2 \mathcal{A}(U^{n+1/2}) \end{aligned}$$

or

$$U^{n+1} = (\alpha_1 + \alpha_2)U^n - (\alpha_2\beta_0 + \beta_1)\Delta s \mathcal{A}(U^n) - \beta_2 \Delta s \mathcal{A}(U^n - \beta_0 \Delta s \mathcal{A}(U^n)).$$

A standard accuracy analysis by matching terms in the Taylor expansion of the exact solution leads to the usual *conditions for first order accuracy*

$$\alpha_1 + \alpha_2 = 1, \tag{7.1}$$

$$\alpha_2\beta_0 + \beta_1 + \beta_2 = 1 \tag{7.2}$$

and the usual *additional condition for second order accuracy*

$$\beta_2\beta_0 = \frac{1}{2}. \tag{7.3}$$

If the coefficients of the time integrator satisfy those constraints, the complete scheme coupling this integrator with the second order ENO scheme in space will lead to an overall convergent scheme with second order accuracy in smooth regions, as long as the time-step is subjected to the appropriate CFL condition [16].

7.2. Linear stability of frozen-stencil schemes. Taking into account the constraint for a first order accuracy in (7.1), the scheme can be written as

$$U^{n+1} = U^n - (1 - \beta_2)\Delta s \mathcal{A}(U^n) - \Delta s \beta_2 \mathcal{A}(U^n - \Delta s \beta_0 \mathcal{A}(U^n)). \tag{7.4}$$

We apply the discretization in (7.4) to the linear system in (5.10) where we take the gradient-preserving scheme in (5.8) as the spatial discretization operator $\mathcal{A}$, with a second order ENO interpolation strategy at the interfaces of the associated Riemann problems:

$$\mathcal{A}\begin{pmatrix} u^n_{i,j} \\ \\ \\ v^n_{i+1/2,j+1/2} \end{pmatrix} = \frac{1}{h}\begin{pmatrix} \{a(u^n_{i,j} + S^n_{i,j} - u^n_{i-1,j} - S^n_{i-1,j}) + b(v^n_{i+1/2,j-1/2} \\ \qquad + T^n_{i+1/2,j-1/2} - v^n_{i-1/2,j-1/2} - T^n_{i-1/2,j-1/2})\} \\ \\ \{a(u^n_{i,j+1} + S^n_{i,j+1} - u^n_{i,j} - S^n_{i,j}) + b(v^n_{i+1/2,j+1/2} \\ \qquad + T^n_{i+1/2,j+1/2} - v^n_{i+1/2,j-1/2} - T^n_{i+1/2,j-1/2})\} \end{pmatrix}, \tag{7.5}$$

where $S_{i,j}$ and $T_{i+1/2,j+1/2}$ are the second order corrections associated with the ENO reconstruction:

$$S_{i,j} = \begin{cases} \dfrac{u_{i+1,j} - u_{i,j}}{2} & \text{if } |u_{i+1,j} - u_{i,j}| \le |u_{i,j} - u_{i-1,j}|, \\ \dfrac{u_{i,j} - u_{i-1,j}}{2} & \text{otherwise,} \end{cases}$$

$$T_{i+1/2,j+1/2} = \begin{cases} \dfrac{v_{i+1/2,j+3/2} - v_{i+1/2,j+1/2}}{2} & \text{if } |v_{i+1/2,j+3/2} - v_{i+1/2,j+1/2}| \\ & \le |v_{i+1/2,j+1/2} - v_{i+1/2,j-1/2}|, \\ \dfrac{v_{i+1/2,j+1/2} - v_{i+1/2,j-1/2}}{2} & \text{otherwise.} \end{cases}$$

The expressions for the corrections $S_{i,j}$ and $T_{i+1/2,j+1/2}$ involve discretization stencils which are solution-dependent so that the discretization operator $\mathcal{A}$ corresponding to an ENO scheme has nonconstant coefficients. To gain insight into the behavior of the time integrator, "frozen-stencil" variations will be studied next. For such schemes, one of the two choices for the corrections S and T is systematically used for the entire domain, for all time-iterations, regardless of the computed solution. As a result, the corresponding operator $\mathcal{A}$ is linear with constant coefficients, and a standard Fourier analysis is feasible. It is clear that such frozen-stencil schemes would not converge numerically for a general nonlinear problem, unlike the original ENO scheme; the motivation for studying such an unpractical scheme is given at the end of the section. When the stencil is frozen, the discrete operator $\mathcal{A}$ is constant and (7.4) can be rewritten as

$$U^{n+1} = U^n - \Delta s \mathcal{A}(U^n) + (\Delta s)^2 \beta_2 \beta_0 \mathcal{A}^2(U^n). \tag{7.6}$$

Let G_2 be the amplification matrix associated with the operator $-\Delta s\mathcal{A}$ and I_d the identity matrix; then the amplification matrix $\mathcal{G}$ associated with (7.6) is given by

$$\mathcal{G} = I_d + G_2 + \beta_0 \beta_2 G_2 o G_2. \tag{7.7}$$

Note that μ is an eigenvalue for G_2 if and only if $1+\mu+\beta_0\beta_2\mu^2$ is an eigenvalue for $\mathcal{G}$. So one needs only to compute G_2 and its eigenvalues. Next, we analyze the spectral radius of G_2 for two particular choices of frozen stencils.

1. *Worst case scenario with frozen stencil: Downwind scheme.* Intuitively, the worst case scenario as far as stability is concerned corresponds to the case where the stencil in both directions includes systematically downwind information. With the advecting velocities $a, b > 0$, this corresponds to the choices

$$S_{i,j} = \frac{u_{i+1,j} - u_{i,j}}{2} \quad \text{and} \quad T_{i+1/2,j+1/2} = \frac{v_{i+1/2,j+3/2} - v_{i+1/2,j+1/2}}{2}.$$

Replacing $(u_{i,j}, v_{i+1/2,j+1/2})$ in (7.5) by a single Fourier harmonic leads to the amplification matrix

$$G_2 = \begin{bmatrix} -\lambda_1 I \sin(\phi_{k_1}) & -\lambda_2 I \left(\sin(\frac{\phi_{k_1}+\phi_{k_2}}{2}) + \sin(\frac{\phi_{k_1}-\phi_{k_2}}{2}) \right) \\ -\lambda_1 I \left(\sin(\frac{\phi_{k_1}+\phi_{k_2}}{2}) - \sin(\frac{\phi_{k_1}-\phi_{k_2}}{2}) \right) & -\lambda_2 I \sin(\phi_{k_2}) \end{bmatrix}$$

with $\phi_k = 2\pi k h$ and $I = \sqrt{-1}$. The eigenvalues of G_2 are $\mu_1 = 0$ and $\mu_2 = -(\lambda_1 \sin(\phi_{k_1}) + \lambda_2 \sin(\phi_{k_2}))I = -\Phi I$. Hence, the spectral radius of the matrix $\mathcal{G}$, given in (7.7), is less than or equal to one if and only if

$$|1 - \Phi I - \beta_0 \beta_2 \Phi^2| \le 1 \tag{7.8}$$

$$\iff \sqrt{1+\left(\beta_0\beta_2\Phi^2\right)^2-(2\beta_0\beta_2-1)\Phi^2}\le 1 \tag{7.9}$$

$$\implies (2\beta_0\beta_2-1)>0. \tag{7.10}$$

Taking into account that $|\phi|\le\lambda_1+\lambda_2$, if this last inequality is verified, then the stability condition for this case is given by the following CFL condition:

$$\lambda_1+\lambda_2\le\frac{\sqrt{2\beta_0\beta_2-1}}{\beta_0\beta_2}. \tag{7.11}$$

2. *Best case scenario with frozen stencil: Upwind scheme.*

$$S_{i,j}=\frac{u_{i,j}-u_{i-1,j}}{2}\quad\text{and}\quad T_{i+1/2,j+1/2}=\frac{v_{i+1/2,j+1/2}-v_{i+1/2,j-1/2}}{2}.$$

This case corresponds to a standard second order upwind scheme. A sufficient condition for stability is that each substep satisfies a classical CFL condition:

$$\begin{aligned}|\beta_0|(\lambda_1+\lambda_2)&\le 1,\\ (|\beta_1\alpha_2|+|\beta_2|)(\lambda_1+\lambda_2)&\le 1.\end{aligned} \tag{7.12}$$

A detailed analysis would lead to similar CFL conditions with less restrictive constants; however, the exact expressions will not be needed here. An interesting observation at this stage is that, as to be expected, the stability condition for the downwind scheme (7.10) and the CFL stability conditions (7.11)–(7.12) show opposite trends. For instance, one way to stabilize the downwind scheme is to pick β_0,β_2 large, which implies taking an intermediate time-step which is actually larger than the final time-step. Such a choice, however, would lead to a very severe final time-step restriction because of the CFL conditions, in particular those of the upwind scheme.

The motivation to study the frozen-stencil schemes is to get some insight on how to achieve a steady state as efficiently as possible. The heuristic in selecting coefficients for an efficient Runge–Kutta integrator is that a scheme that satisfies all the stability constraints from the two frozen-stencil limit schemes (in addition to at least the first order accuracy conditions (7.1)–(7.2)) must have excellent damping properties as it stabilizes even the particularly unstable downwind scheme. Therefore, it is expected that such an integrator would lead to an efficient pseudotime marching scheme to a steady state by damping numerically the oscillations faster than a time-accurate scheme would.

A first observation is that the condition in (7.10) for stability of the downwind frozen-stencil scheme is incompatible with the condition in (7.3) for second order accuracy in pseudotime. It is also trivial to verify that, as one should expect, the standard one-step forward Euler scheme (with $\beta_0=0$) cannot possibly satisfy the frozen-stencil stability condition for the downwind scheme.

Good damping per time-step is expected to be achieved by selecting a time integrator with coefficients that minimize the spectral radius in (7.9) (with a similar expression for the upwind scheme). One could attempt to find an optimal set of coefficients to minimize the spectral radii in a systematic search. However, such a procedure would be costly and probably not very useful because the results might not

be relevant to the actual ENO scheme with variable stencil as used for the nonlinear cell-problem. Instead, we limited the search to sampling a small number of combinations for the time-integrator coefficients and compared their performance (in conjunction with the second order variable-stencil ENO scheme) in numerical experiments for the actual cell-problem. Among the combinations we tested were the standard explicit Runge–Kutta integrators of order one and two mentioned earlier—that they were outperformed is consistent with the heuristic analysis above.

The best performance observed in our limited search was achieved by the following combination, whose coefficients satisfy all the frozen-stencil stability constraints:

$$\alpha_1 = \frac{1}{3}, \qquad \alpha_2 = \frac{2}{3}, \qquad \beta_0 = \frac{3}{2}, \qquad \beta_1 = -\frac{1}{2}, \qquad \beta_2 = \frac{1}{2}$$

with the following CFL condition:

$$\lambda_1 + \lambda_2 \leq \frac{2}{3}. \tag{7.13}$$

Even though this linearly stable scheme (in the frozen-stencil sense) is only first order accurate with respect to the pseudotime variable, numerical experiments such as the one reported in Figure 7.1 demonstrate that it significantly improves the convergence to a steady state compared to the second order time integrator. In that example, the eigenvalue oscillates rapidly in the pseudotime with the oscillation amplitude decaying to zero when the pseudotime grows, as predicted theoretically. The oscillations are rapidly damped with the first order scheme, (a), and with the hybrid second order in space and first order in time scheme, but are much less so with the overall second order scheme, (c). Because we are not interested in an accurate prediction of the time evolution but only in the steady state, the new scheme has the advantage of converging to a steady state almost as efficiently as the first order scheme, while at the same time achieving second order accuracy in space (at least in cases with smooth eigenfunctions) as will be demonstrated in the next section by analyzing the convergence of the results under systematic mesh refinement.

8. Validation.

8.1. Small-scale shears. To validate the method, we first consider the response of the flame to velocity fields generated with the Childress–Soward flow (see the stream function in (2.3)) with $\delta = 1$. Then, the velocity field is given by

$$v_1(x,y) = v_2(x,y) = \frac{\lambda}{\sqrt{2}} \left(-\sin(2\pi x)\cos(2\pi y) + \cos(2\pi x)\sin(2\pi y) \right).$$

This flow field actually represents a simple sine shear titled at 45 degrees. The problem can be reduced to a one-dimensional problem by aligning the coordinate system with the shearing direction, and the results in [5] and [17] can be applied directly to provide us with reference data. (Here, the stream function is scaled so that λ represents the maximum shear intensity.) We solve the problem numerically in the original coordinate system as a two-dimensional case using the gradient-preserving scheme. The combustion speed enhancement is then compared with the reference value.

The numerical procedure consists of minimizing $F(r)$ in (2.1) as a function of r, which requires solving the cell-problem corresponding to each trial value for r. The minimization is performed using a standard routine [3] to a specified tolerance; all our numerical experiments showed that there were no numerical difficulties associated

TABLE 8.1
Convergence of speed enhancement with $\delta = 1$, $\theta = \pi/4$, $\overline{\lambda} = 0$.

λ	Reference F_e	Grid	F_e order 1	(Error)	F_e order 2	(Error)
0.4	0.4	16×16	0.24902	0.15098	0.39540	0.00460
		32×32	0.31928	0.08072	0.39939	0.00060
		64×64	0.35815	0.04184	0.39989	0.00011
1.6	1.6	16×16	1.06888	0.53112	1.58045	0.01955
		32×32	1.30876	0.29124	1.59761	0.00239
		64×64	1.44679	0.15320	1.59959	0.00040
6.4	6.4	16×16	3.39313	3.00687	6.27134	0.12866
		32×32	4.69028	1.70972	6.38566	0.01434
		64×64	5.50819	0.89182	6.39792	0.00208

with the minimization routine itself, the key numerical issue being able to provide a sufficiently accurate value for the effective Hamiltonian $H(r)$.

The test cases reported below correspond to the front angle $\theta = \pi/4$ (i.e., the front normal is aligned with the shearing direction) and no mean flow $\overline{\lambda} = \overline{\theta} = 0$. With those parameters, it is particularly straightforward to predict analytically the burning speed enhancement as those set-ups are known theoretically to achieve the upper bound $F_e = \lambda$. Table 8.1 reports the computed values of F_e for three values of λ: small turbulence intensity $\lambda = 0.4$, medium intensity $\lambda = 1.6$, and large turbulence intensity $\lambda = 6.4$. (Recall that in all the test cases here, velocities are normalized with the laminar burning speed $S_L = 1$.) Figure 8.1 shows sequences of the corresponding wrinkled fronts: they represent a flame propagating from the right upper corner towards the left lower corner of the domain. Both the order 1 and the order 2 methods are seen in Table 8.1 to converge under mesh refinement, with the expected order of convergence: in particular, the gain in accuracy going from a first order to a second order method is significant at low resolutions—for practical purposes, it is possible, with the hybrid second order scheme, to predict F_e within one or two percent with the very coarse resolution of 16×16 ! Notice that the eigenfunctions computed here are not smooth (for instance, see the cusps in the flame fronts in Figure 8.1, corresponding to shocks in the gradients). As a consequence, one would expect a detailed numerical convergence study of the eigenfunction to show a reduction to first order convergence, even with the second order method. However, we are interested here only in the enhanced speed, obtained by processing the effective Hamiltonian, which itself is obtained by integration of the square of the norm of the eigenfunction gradient over the domain: this processing is sufficient to recover second order accuracy (see the last column of Table 8.1) even if the eigenfunction is locally first order accurate in the vicinity of the cusps.

8.2. Other flows. In the test cases with $\delta = 1$ just described, the small-scale flow is a simple shear aligned with the normal to the large-scale front so that the wrinkled flame front has a very simple topology, traveling without changing shape from the upper right corner into the unburnt mixture in the lower left corner, at a constant velocity. Such a simple flame pattern could have been computed using explicit formulas [5] and it was considered here only for the sake of validation. However, the gradient-preserving scheme is very robust and is designed to handle much more complex flame fronts behaviors. For instance, selecting $\delta < 1$ in (2.3) leads to more interesting turbulent-like flows. For $\delta = 0$ the flow corresponds to an array of eddies and for $\delta = 0.5$ we have a combination of eddies and shears; see Figure 2.1. Results for those two cases are reported in Tables 8.2 and 8.3 as well as in Figures 8.2 and

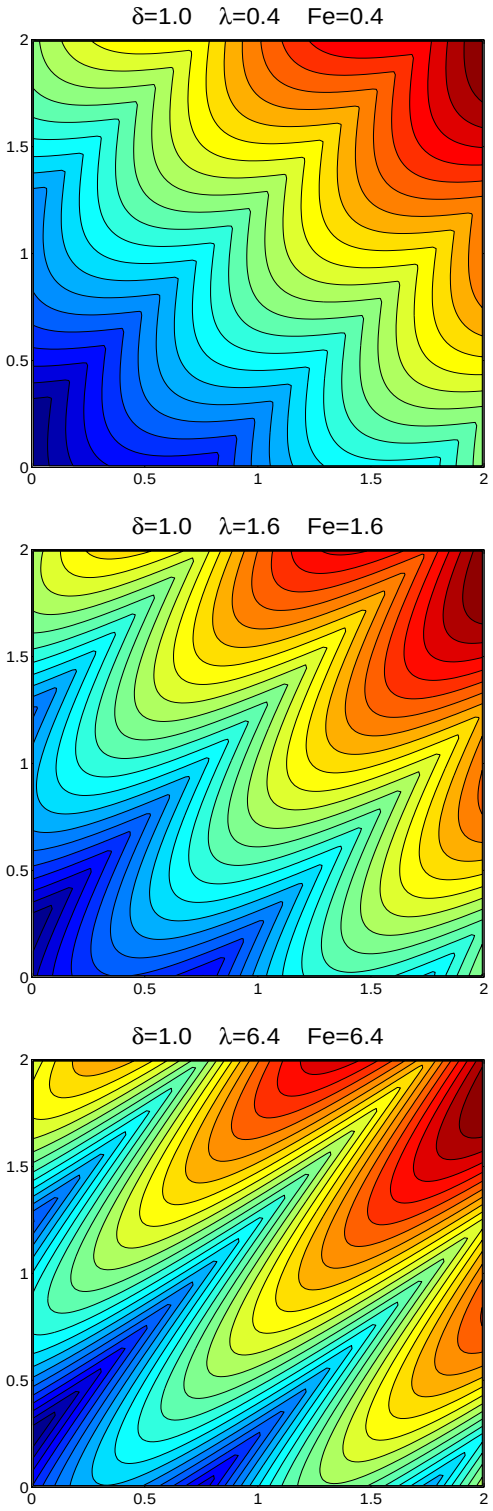


FIG. 8.1. *Instantaneous flame fronts with* $\delta = 1$*; same data as in Table* 8.1. *(Hybrid) second order method, resolution* 64×64 *(an array of* 2×2 *cells is shown).*

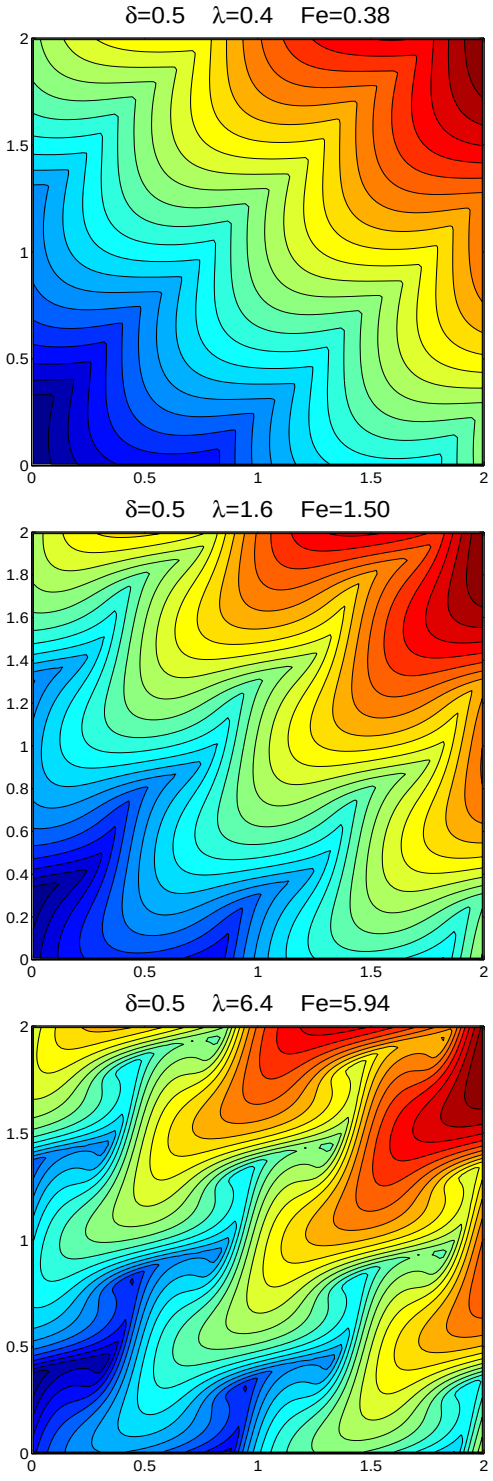


FIG. 8.2. *Same as Figure* 8.1, *with* $\delta = 0.5$.

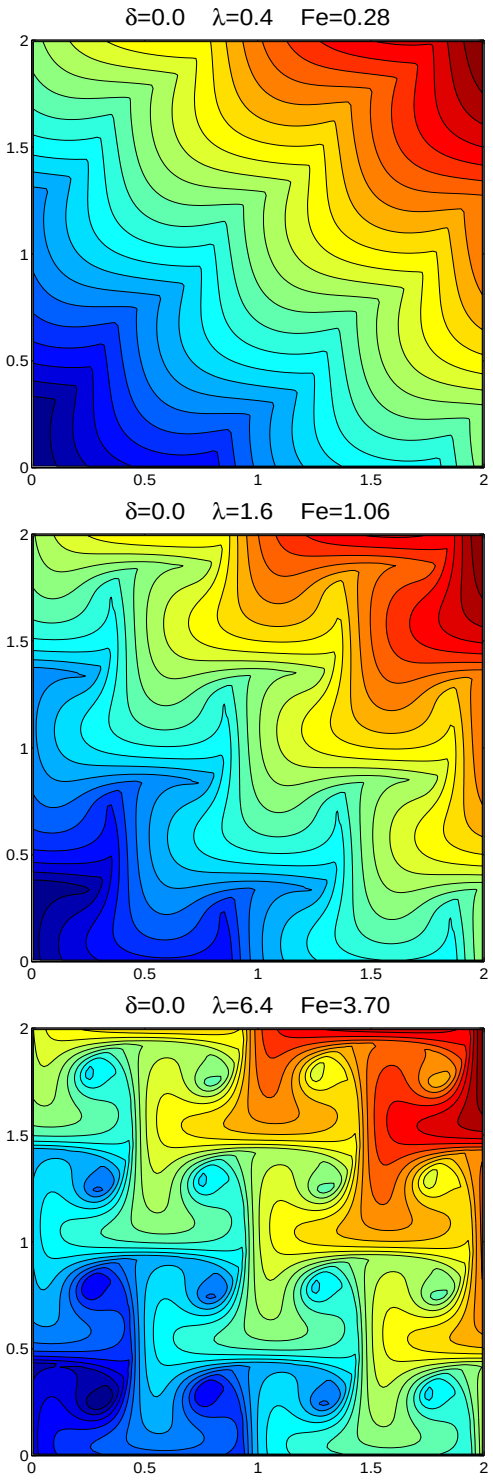


FIG. 8.3. *Same as Figure* 8.1, *with* $\delta = 0.0$.

TABLE 8.2
Convergence of speed enhancement with $\delta = 0.5$, $\theta = \pi/4$, $\overline{\lambda} = 0$.

λ	Reference F_e	Grid	F_e order 1	(Error)	F_e order 2	(Error)
1.6	1.50539	16×16	1.00988	0.49551	1.45519	0.05019
		32×32	1.23407	0.27132	1.49656	0.00882
		64×64	1.36283	0.14255	1.50331	0.00208
		128×128	1.43227	0.07311	1.50494	0.00044

TABLE 8.3
Convergence of speed enhancement with $\delta = 0.0$, $\theta = \pi/4$, $\overline{\lambda} = 0$.

λ	Reference F_e	Grid	F_e order 1	(Error)	F_e order 2	(Error)
1.6	1.06534	16×16	0.77352	0.33186	0.99950	0.06588
		32×32	0.88433	0.18106	1.04836	0.01702
		64×64	0.97049	0.09485	1.06186	0.00347
		128×128	1.01676	0.04858	1.06485	0.00048

8.3. Again, the data are $\theta = \pi/4$, $\overline{\lambda} = \overline{\theta} = 0$. The results of the mesh refinement reported in the two tables for the intermediate turbulence intensity $\lambda = 1.6$ confirm the predicted order of accuracy and the significant gain in accuracy with the second order method. When $\delta \neq 1$, there are no analytical predictions available for F_e: instead, a reference value is obtained here by extrapolation of the data with the second order method—this should not affect the error analysis, except maybe for the estimation of the error on the finest grid. At low turbulence intensity $\lambda = 0.4$, the flame patterns at the top of Figures 8.2 and 8.3 appear to be very similar to the simple shear case from Figure 8.1. (Notice, however, that F_e decreases with δ.) At larger intensities, however, in particular when $\lambda = 6.4$, the flame patterns become much more complex, some portions of the front overlap other portions, and there are even topological changes associated with pockets of unburnt gas lagging behind the leading front. Many more computations of this type can be found in [14, 2], along with a more detailed analysis of the parameterization of F_e as a function of the characteristics of the flow.

9. Conclusion. A numerical method has been introduced to solve the nonlinear eigenvalue cell-problem arising in the homogenization theory of turbulent premixed flame fronts [17]. The scheme allows for the efficient second order accurate computation of both the effective Hamiltonian (i.e., the eigenvalue) and the eigenfunction (related to successive realizations of the wrinkled flame front). The eigenvalue problem is solved using pseudotime marching to the steady state of a system of conservation laws for the eigenfunction gradient. Theoretical arguments are presented regarding the importance of satisfying the constraint that the steady-state vector solution be effectively the gradient of a scalar function. Exploiting the symmetry of the problem, a novel staggered grid formulation is shown to automatically satisfy the gradient-structure constraint in some appropriate discrete sense if the initial data did. Numerical experiments with variations of the scheme confirm the theoretical predictions by demonstrating that (i) the gradient-preserving property is necessary to guarantee convergence to steady state (Figure 6.1), (ii) time integrators with good damping properties can be achieved by studying their behavior for frozen-stencil variations of the scheme (Figure 7.1), and (iii) the scheme does achieve the predicted second order accuracy, with a significant gain compared to the first order scheme at low resolution (Tables 8.1, 8.2, and 8.3).

The idea of converting a multidimensional Hamilton–Jacobi equation into a sys-

tem of conservation laws for the gradient of the solution w was also used by Jin and Xin [9]. Their scheme also enforces to some extent the gradient condition in a formulation that has many advantages for general cases. The scheme presented here is different in several respects: it exploits the specific structure of the Hamiltonian to enforce the gradient condition exactly, not in a relaxation sense as in [9]; the two components of the discrete vector solution are represented on staggered grids, whereas in [9] they are collocated, only w is staggered; the scheme presented here is geared toward the efficient computation of a steady-state solution for the effective Hamiltonian, a quantity that involves only $\mathbf{D}w$ and not w itself.

The scheme presented here was used successfully in [14, 2] to study systematically the parameterization of the turbulent enhancement of the flame speed for a variety of small-scale flows. The numerical data were used to identify two distinct scalings regimes, similar to those observed in experiments with real flames. The transition between the two regimes was shown to depend essentially on a nondimensional "flame residence time" that relates an intrinsic flame response time to a time scale related to the flame passage-time through a periodic cell; this qualitative analysis inspired by the numerical data was explained theoretically via a formal asymptotic analysis.

Ultimately, one objective is to use the asymptotic speed enhancement as a basis for a subgrid-scale model in large eddy simulations of turbulent flames, where the effect of the unresolved turbulent flow scales must be accounted for as a modelled enhanced burning speed. The feasibility of such a strategy is demonstrated in [15] for an idealized case: the scheme introduced in this paper is used to generate a complete database or "flamelet library" which can then be used repeatedly as an input to a level-set formulation for the flame front at large scales. Results of such computation are shown in [15] to be in excellent agreement with detailed direct simulation predictions for the wrinkled flames, with the large eddy simulations requiring only a small fraction of the computational cost of the detailed simulations.

Another potential application for the gradient-preserving scheme described in this paper would be to solve the quadratic nonlinear eigenvalue cell-problem arising in the homogenization of the stationary Schrödinger equation [7]. The method described in this paper exploits the specific structure of the quadratic flux in the cell-problem that splits very naturally into two distinct one-dimensional Riemann problems. One natural extension of the present scheme is to consider cases where such splitting is not possible.

Acknowledgment. The authors are grateful to A. J. Majda for his helpful suggestions regarding this project.

REFERENCES

[1] G. Barles and P. E. Souganidis, *On the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.

[2] A. Bourlioux, B. Khouider, and A. J. Majda, *Parameterizing the burning speed enhancement by small-scale periodic flows:* II. *Application to jets and eddies,* Combust. Theory Model., to be submitted.

[3] R. Brent, *Module DFMIN in NMS*, http://gams.nist.gov/serve.cgi/Module/NMS/DFMIN/5671/.

[4] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[5] P. F. Embid, A. J. Majda, and P. E. Souganidis, *Effective geometric front dynamics for premixed turbulent combustion with separated velocity scales*, Combust. Sci. and Tech., 103 (1994), pp. 85–115.

[6] P. F. Embid, A. J. Majda, and P. E. Souganidis, *Comparison of turbulent flame speeds from complete averaging and the G-equation*, Phys. Fluids, 7 (1995), pp. 2052–2060.
[7] L. C. Evans, *Effective Hamiltonians and Quantum States*, Seminaire Équations aux Derivées Partielles, Ecole Polytechnique, Palaiseau Cedex, France, 2001.
[8] J. Glimm and P. D. Lax, *Decay of Solutions of Systems of Nonlinear Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc. 101, AMS, Providence, RI, 1970.
[9] S. Jin and Z. Xin, *Numerical passage from systems of conservation laws to Hamilton–Jacobi equations, and relaxation schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 2385–2404.
[10] E. Godlewski and P.-A. Raviart, *Hyperbolic systems of conservation laws* SMAI, Math. Appl. 3/4, Ellipses, Paris, 1991.
[11] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.
[12] C. Hirsch, *Numerical Computation of Internal and External Flows, Volume* 1*: Fundamentals of Numerical Discretization*, John Wiley and Sons, New York, 1988.
[13] B. Khouider, *Asymptotic Modeling for Large Eddy Simulations of Turbulent Premixed Combustion*, Ph.D. thesis, University of Montreal, Montreal, Quebec, Canada, 2002.
[14] B. Khouider, A. Bourlioux, and A. J. Majda, *Parameterizing the burning speed enhancement by small-scale periodic flows:* I. *Unsteady shears, flame residence time and bending,* Combust. Theory Model., 5 (2001), pp. 295–318.
[15] B. Khouider and A. Bourlioux, *A rigorous asymptotic perspective on the large scale simulations of turbulent premixed flame fronts*, Multiscale Model. Simul., to be submitted.
[16] R. J. LeVeque, *Numerical methods for conservation laws*, Lectures Math. ETH Zurich, Birkhäuser Verlag, Basel, 1992.
[17] A. J. Majda and P. E. Souganidis, *Large scale front dynamics for turbulent reaction-diffusion equations with separated velocity scales*, Nonlinearity, 7 (1994), pp. 1–30.
[18] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1354–1368

CONSTRUCTION OF SMOOTH REFINABLE FUNCTION VECTORS BY CASCADE ALGORITHMS*

DI-RONG CHEN†

Abstract. This paper establishes an equivalent relation between the convergence of a cascade algorithm in Sobolev space and the convergence of an associated cascade algorithm in L_p space. It reduces the convergence in Sobolev space to that in L_p space. On the other hand, by the equivalence we present an algorithm for construction of refinement masks which generate convergent cascade algorithms in Sobolev space. It is very easy to implement the algorithm. Examples are given to illustrate the theory.

Key words. refinable function vector, Sobolev space, cascade algorithm, factorization of mask, joint spectral radius

AMS subject classifications. 39B12, 41A25, 42C15, 65F15

PII. S0036142901392614

1. Introduction.

Cascade algorithms give an iterative method for the generation of refinable function vectors which are the main building blocks for the construction of multiwavelets. The main purpose of this paper is to present an algorithm for constructing cascade algorithms that converge in Sobolev space.

Suppose that $a = (a(\alpha))_{\alpha\in\mathbb{Z}}$ is a sequence of $r \times r$ matrices and satisfies, for a positive integer $N, a(\alpha) = 0\ \forall\alpha \notin \{0, 1, \ldots, N\}$. A vector $\Phi = (\phi_1, \ldots, \phi_r)^T$ of compactly supported distributions is said to be *refinable associated with* a if it satisfies the following refinement equation:

$$\Phi = \sum_{\alpha=0}^{N} a(\alpha)\Phi(2\cdot -\alpha). \tag{1.1}$$

The sequence a is called the *refinement mask*. In terms of Fourier transform, we can rewrite (1.1) as

$$\widehat{\Phi}(2\omega) = H_a(\omega)\widehat{\Phi}(\omega), \qquad \omega \in \mathbb{R}, \tag{1.2}$$

where $H_a(\omega) := 1/2\sum_{\alpha=0}^{N} a(\alpha)e^{-i\alpha\omega}$.

It is known that if the matrix $H_a(0)$ has 1 as a simple eigenvalue and, for any positive integer $\beta, 2^\beta$ is not an eigenvalue of $H_a(0)$, then there exists a unique vector Φ of compactly supported distributions such that Φ satisfies (1.1) and the first nonzero component of $\widehat{\Phi}(0)$ is 1. This vector is referred as to the *normalized solution* of the refinement equation with mask a and denoted by Φ_a.

Let $W_p(\mathbb{R})$ denote $L_p(\mathbb{R}^s)$ for $1 \le p < \infty$ and the space $C_u(\mathbb{R})$ of uniformly continuous and bounded function on $\mathbb{R}$ for $p = \infty$. The Sobolev space $W_p^s(\mathbb{R})$ is defined as

$$W_p^s(\mathbb{R}) := \{\, f :\ D^\mu f \in W_p(\mathbb{R}),\ 0 \le \mu \le s\},$$

*Received by the editors July 19, 2001; accepted for publication (in revised form) April 1, 2002; published electronically September 27, 2002. This research was supported in part by NSF of China under grant 10171007 and City University of Hong Kong under grant 7001029.
http://www.siam.org/journals/sinum/40-4/39261.html

†Department of Applied Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100083, The People's Republic of China and Hubei Institute of Nationalities, Enshi, Hubei 44500, The People's Republic of China (matdrchen@sina.com).

where $D^\mu = \frac{d^\mu}{dx^\mu}$ is the differential operator. Throughout, s is a nonnegative integer. The space $W_p^s(\mathbb{R})$ is a Banach space with the norm

$$||f||_{W_p^s(\mathbb{R})} := \max\{||D^\mu f||_p : 0 \le \mu \le s\}.$$

Let $(W_p^s(\mathbb{R}))^r$ be the Banach space consisting of all vectors $f := (f_1, \dots, f_r)^T$ with $f_j \in W_p^s(\mathbb{R})$ equipped with norm

$$\|f\|_{(W_p(\mathbb{R}))^r} := \max\{||f_j||_{W_p^s(\mathbb{R})} : 1 \le j \le r\}.$$

For a vector F_0 of compactly supported functions, we construct a sequence $(F_k)_{k\ge 0}$ by iteration $F_k = Q_a F_{k-1}, k = 1, 2, \dots$, where Q_a is the *cascade operator* associated with a defined as

$$Q_a F = \sum_{\alpha=0}^{N} a(\alpha) F(M \cdot -\alpha). \tag{1.3}$$

We say that the cascade algorithm associated with mask a (or generated by Q_a) converges on F_0 in $(W_p^s(\mathbb{R}))^r$ norm if

$$\lim_{n\to\infty} \|Q_a^n F_0 - \Phi_a\|_{(W_p^s(\mathbb{R}))^r} = 0. \tag{1.4}$$

Let $\mathbb{C}^{r\times m}$ be the space of $r \times m$ matrices with entries being complex numbers. Denote by $(\ell(\mathbb{Z}))^{r\times m}$ the set of all sequences $\lambda = (\lambda(\alpha))_{\alpha\in\mathbb{Z}}$ with $\lambda(\alpha) \in \mathbb{C}^{r\times m}$. Furthermore, let $(\ell_0(\mathbb{Z}))^{r\times m}$ denote the set of all sequences in $(\ell(\mathbb{Z}))^{r\times m}$ with finite supports. The *subdivision operator* S_a, associated with a mask $a \in (\ell_0(\mathbb{Z}))^{r\times r}$, is an operator defined on the space $(\ell(\mathbb{Z}^s))^{m\times r}$ as follows:

$$S_a c(\alpha) := \sum_{\beta\in\mathbb{Z}^s} c(\beta) a(\alpha - 2\beta), \qquad \alpha \in \mathbb{Z}^s. \tag{1.5}$$

For convenience, let

$$a_n = S_a^{n-1} a, \qquad n = 1, 2, \dots .$$

The iteration Q_a^n is related with S_a by the equality

$$Q_a^n F = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha) F(2^n \cdot -\alpha).$$

The convergence of cascade algorithms is related intimately with the properties of refinable function vectors. There has been a comprehensive study of cascade algorithms. When $s = 0$ and $r = 1$, see [1] for $p = \infty$, [11] for $1 \le p \le \infty$ and [10] for $1 \le p \le \infty$, and the dilation matrix being an arbitrary isotropic matrix. For general r, see [14]. When $s > 0$, $p = 2$, and $r = 1$, see [9] and [12]. We studied this problem in the general setting in [3]. In [4], the effect of perturbation of refinement masks to the convergence was discussed.

While the papers mentioned above focused on the convergence in a fixed space, we clarify in this paper the relationship between the convergence of a cascade algorithm in Sobolev space and that of an associated cascade algorithm in L_p space. To this end we make use of the methods for factoring masks in [16] and [18]. An equivalence between the convergence in different spaces is established. By that equivalence we

present an algorithm for construction of masks which generate a convergent cascade algorithm in Sobolev space. It is somewhat like the construction of [18], of refinable function vectors with arbitrary approximation order and smoothness order. Roughly speaking, based on a convergent cascade algorithm in $(W_p^s(\mathbb{R}))^r$, we can construct a convergent cascade algorithm in $(W_p^{s+1}(\mathbb{R}))^r$ by one step of two scale similarity transform.

Although it is accepted that there is a direct relation between the smoothness of refinable function vectors and the factorization of masks, there has so far not been any equivalence without the stability condition. The present paper is related to [6], [15], and [17]. In [6] it was shown that the factorization of masks can lead to decay of the Fourier transform of the refinable function vectors. For a refinable function vector, which is a limit of a convergent cascade algorithm, its smoothness was characterized in terms of the factorization of mask [15, Theorem 4.2]. Under the stability condition, a relation between the spectral radius concerning the factorized masks and the smoothness of refinable function vectors is established in [15] by an approach different from ours. A specific factorization technique based on superfunction theory was presented in [17]. The importance of that technique for the study of the smoothness lies in determining *separately* the smoothness of each component of the refinable vectors. However, our aim is to provide a method for constructing smooth refinable function vectors by use of the factorization technique of masks in [16] and [18].

Here is a brief outline of the paper. In section 2, a formula for spectral radius of subdivision operators concerning factorization of masks is established. Moreover, we connect the formula with the notion of p-joint spectral radii. In section 3, an invariant subspace associated with p-joint spectral radius is characterized explicitly in terms of the factorization of the mask. The main results of the paper are presented in section 4. In this section we derive the equivalence between the convergence of a cascade algorithm Q_a in Sobolev norm and the convergence of another cascade algorithm generated by Q_{a_0} in L_p, where a_0 is determined by a. As a corollary, we give an algorithm for construction of masks which generate convergent cascade algorithms in Sobolev space. In section 5, we illustrate our theory by two examples.

2. Factorization of masks and spectral radii of subdivision operators. In this section we first give a formula, in terms of some limits, for spectral radii of subdivision operators concerning the factorization of masks. Then a relation between those limits and joint spectral radii is established.

Assume now that $\mathbb{C}^{r\times m}$ is equipped with a norm $||\cdot||$ on $\mathbb{C}^{r\times m}$. Here and elsewhere, all norms on a finite dimension space are denoted by $\|\cdot\|$. Let $(\ell_p(\mathbb{Z}))^{r\times m}$ be the normed linear space of sequences $\lambda = (\lambda(\alpha))_{\alpha\in\mathbb{Z}}, \lambda(\alpha) \in \mathbb{C}^{r\times m}$, such that $||\lambda||_p < \infty$, where

$$||\lambda||_p := \left(\sum_{\alpha\in\mathbb{Z}} \|\lambda(\alpha)\|^p\right)^{1/p}, \qquad 1 \le p < \infty,$$

and $\|\lambda\|_\infty := \sup_{\alpha\in\mathbb{Z}} ||\lambda(\alpha)||$. It is easily seen that, for any mask $b \in (\ell_0(\mathbb{Z}))^{r\times r}$, the subdivision operator S_b is a bounded operator on $(\ell_p(\mathbb{Z}))^{m\times r}$. Clearly, the set $(\ell_p(\mathbb{Z}))^{r\times m}$ is independent of the choice of norm $||\cdot||$ on $\mathbb{C}^{r\times m}$, and any two norms on $\mathbb{C}^{r\times m}$ induce two equivalent norms on $(\ell_p(\mathbb{Z}))^{r\times m}$. It is easily seen that, for any $b \in (\ell_0(\mathbb{Z}))^{r\times r}, S_b$ is a bounded operator on $(\ell_p(\mathbb{Z}))^{m\times r}$.

The convolution of $c * d = (c * d(\alpha))_{\alpha\in\mathbb{Z}} \in (\ell(\mathbb{Z}))^{l\times k}$, for two sequences $c \in$

$(\ell(\mathbb{Z}))^{l\times m}$ and $d\in(\ell(\mathbb{Z}))^{m\times k}$, is defined by

$$c*d(\alpha)=\sum_{\beta\in\mathbb{Z}}c(\alpha-\beta)d(\beta)\qquad\forall\alpha\in\mathbb{Z},$$

provided that the series converges for any entry and any $\alpha\in\mathbb{Z}$.

If $K\subseteq\mathbb{Z}$ is a finite set, we denote $(\ell(K))^{r\times m}$ the subspace consisting of all $c\in(\ell_0(\mathbb{Z}))^{r\times m}$ with supp$c\subseteq K$. Let

$$K_N:=\begin{cases}\{0,1\} & \text{for } N=1,\\ \{0,1,\dots,N-1\} & \text{for } N\geq 2.\end{cases}\tag{2.1}$$

The following interesting result concerning the spectral radius of S_b was established in [5].

THEOREM 2.1 (see [5]). *Let $b\in(\ell_0(\mathbb{Z}))^{r\times r}$. Suppose that $c\in(\ell_p(\mathbb{Z}))^{r\times r}$ satisfies*

$$c*t\neq 0\qquad\forall t\in(\ell(K_N))^{r\times 1}\backslash\{0\}.\tag{2.2}$$

We have

$$\rho_p(S_b)=\lim_{n\to\infty}||S_b^nc||_p^{\frac{1}{n}}.$$

We will extend the above result to the case when the mask a is factorable. This means that H_a satisfies

$$H_aC=C(2\cdot)H_b\tag{2.3}$$

for some mask $b\in(\ell_0(\mathbb{Z}))^{r\times r}$ and a 2π periodic $r\times r$ matrix $C(\omega)$.

For $\eta\in(\ell(\mathbb{Z}))^{m\times n}$, the Fourier transform $\widehat{\eta}$ of η is defined at least formally by

$$\widehat{\eta}(\omega)=\sum_{\alpha\in\mathbb{Z}}\eta(\alpha)e^{-i\alpha\omega}.$$

If there is $c\in(\ell(\mathbb{Z}))^{r\times r}$ such that $\widehat{c}=C$, then (2.3) is equivalent to

$$a*c=S_bc.\tag{2.4}$$

Theorem 2.1 is generalized as follows, which is the main result of this section and will be needed in section 4.

THEOREM 2.2. *Suppose that $a,b\in(\ell_0(\mathbb{Z}))^{r\times r}$ and $c\in(\ell_p(\mathbb{Z}))^{r\times r}$ satisfy* (2.4). *If c satisfies conditions of Theorem* 2.1*, then*

$$\rho_p(S_b)=\lim_{n\to\infty}||a_n*c||_p^{\frac{1}{n}}.$$

Proof. We claim that, for any positive integer n,

$$S_b^nc=a_n*c.\tag{2.5}$$

For $n=1$, (2.5) reduces to (2.4). Let us check (2.5) for $n=2$. In fact, it follows from (2.4) that

$$S_b^2c(\gamma)=\sum_{\alpha\in\mathbb{Z}}\sum_{\beta\in\mathbb{Z}}a(\beta)c(\alpha-\beta)b(\gamma-2\alpha)=\sum_{\beta\in\mathbb{Z}}a(\beta)S_bc(\gamma-2\beta)\qquad\forall\gamma\in\mathbb{Z}.$$

Using (2.4) again we obtain

$$S_b^2 c(\gamma) = \sum_{\beta\in\mathbb{Z}} a(\beta) a * c(\gamma - 2\beta) = \sum_{\beta\in\mathbb{Z}} a_2(\gamma-\beta)c(\beta) \qquad \forall \gamma \in \mathbb{Z},$$

as desired. The verification of (2.5) for general n may proceed inductively. The proof of our theorem is now complete by (2.5) and Theorem 2.1. □

While Theorem 2.2 connects the spectral radius of a subdivision operator with the limit $\lim_{n\to\infty} ||a_n * c||^{\frac{1}{n}}_p$, in the rest of this section we shall represent the limit in terms of the so-called p-joint spectral radius.

For $p = \infty$, the p-joint spectral radius, called the uniform joint spectral radius, was introduced in [19] and was employed to investigate regularity of refinable function in [7]. When $p = 1$, the p-joint spectral radius was introduced in [20] and was referred to as the mean joint spectral radius there. For $1 < p < \infty$, the p-joint spectral radius was introduced in [11] and applied to the study of L_p convergence of cascade algorithms. We recall from [11] the definition of p-joint spectral radius.

If V is a finite-dimensional space, let $\mathcal{B}(V)$ denote the collection of linear operators on V. Suppose that $\mathcal{A} \subseteq \mathcal{B}(V)$ is a finite set. For a positive integer n we denote by $\mathcal{A}^n$ the Cartesian power of $\mathcal{A}$:

$$\mathcal{A}^n = \big\{(A_1, \dots, A_n) : A_1, \dots, A_n \in \mathcal{A}\big\}.$$

Suppose that $\mathcal{B}(V)$ is equipped with the operator norm $||\cdot||$. For $1 \le p < \infty$, define a number $||\mathcal{A}^n||_p$ by

$$||\mathcal{A}^n||_p^p = \sum_{(A_1,\dots,A_n)\in\mathcal{A}^n} ||A_1 \cdots A_n||^p,$$

and, for $p = \infty$, define

$$||\mathcal{A}^n||_\infty = \max\big\{\, ||A_1 \cdots A_n|| : (A_1, \dots, A_n) \in \mathcal{A}^n \,\big\}.$$

The p-joint spectral radius of $\mathcal{A}$, for $1 \le p \le \infty$, is defined to be

$$\rho_p(\mathcal{A}) := \lim_{n\to\infty} ||\mathcal{A}^n||_p^{\frac{1}{n}}. \tag{2.6}$$

We know from [11] that this limit indeed exists and

$$\lim_{n\to\infty} ||\mathcal{A}^n||_p^{\frac{1}{n}} = \inf_{n\ge1} ||\mathcal{A}^n||_p^{\frac{1}{n}}.$$

For our purposes, we are mainly concerned with the operators A_ε, $\varepsilon = 0, 1$, on $(\ell_0(\mathbb{Z}))^{r\times m}$ defined as follows:

$$A_\varepsilon c(\alpha) = \sum_{\beta\in\mathbb{Z}} a(2\alpha + \varepsilon - \beta)c(\beta) \qquad \forall c \in (\ell_0(\mathbb{Z}))^{r\times m} \text{ and } \alpha \in \mathbb{Z}. \tag{2.7}$$

Let $\mathcal{A} := \{A_0, A_1\}$. A subspace $V \subseteq (\ell_0(\mathbb{Z}))^{r\times m}$ is called an $\mathcal{A}$ *invariant subspace* if

$$A_\varepsilon c \in V \qquad \forall \varepsilon = 0, 1 \text{ and } c \in V.$$

Suppose that M_1 and M_2 are two constants. It is easy to check that, for any $c \in (\ell_0(\mathbb{Z}))^{r\times m}$,

$$\operatorname{supp} c \subseteq [-M_1, M_2] \Longrightarrow \operatorname{supp} A_\varepsilon c \subseteq \left[\frac{-M_1 - 1}{2}, \frac{N + M_2}{2}\right], \qquad \varepsilon = 0, 1. \tag{2.8}$$

For any $c \in (\ell_0(\mathbb{Z}))^{r\times m}$, let $V(c)$ be the subspace spanned by

$$A_{\varepsilon_j}\cdots A_{\varepsilon_1}c,\ j=0,1,\ldots.$$

Clearly, $V(c)$ is an $\mathcal{A}$ invariant subspace, called the $\mathcal{A}$ invariant subspace *generated by c*. Moreover, we know by (2.8) that $V(c)$ is a finite-dimensional subspace. It follows from (2.8) that, for any $c \in (\ell_0(\mathbb{Z}))^{r\times m}$, there exists a finite set $K \subseteq \mathbb{Z}$ such that $V(c) \subseteq (\ell(K))^{r\times m}$. In particular, by (2.8), $(\ell(K_N))^{r\times m}$ is an invariant subspace of $\mathcal{A}$, where K_N is given as in (2.1).

For any $\alpha \in \mathbb{Z}$, there are unique $\varepsilon_1,\ldots,\varepsilon_n \in \{0,1\}$ and a $\gamma \in \mathbb{Z}$ such that $\alpha = 2^n\gamma+2^{n-1}\varepsilon_n\cdots+\varepsilon_1$. It is not difficult to check that (see, e.g., [3]), for $c \in (\ell_0(\mathbb{Z}))^{r\times m}$,

$$a_n * c(\alpha) = A_{\varepsilon_n}\cdots A_{\varepsilon_1}c(\gamma).$$

As is known, there is a finite set $K \subseteq \mathbb{Z}$ such that $V(c) \subseteq (\ell(K))^{r\times m}$. Then it follows from the last equality that

$$\|a_n * c\|_p = \sum_{\varepsilon_1,\ldots,\varepsilon_n=0,1} ||A_{\varepsilon_n}\cdots A_{\varepsilon_1}c||^p \qquad \forall c \in (\ell_0(\mathbb{Z}))^{r\times m}, \tag{2.9}$$

where $||\cdot||$ in the right-hand side is a norm on $(\ell(K))^{r\times m}$. With this equality we can derive for any $c \in (\ell_0(\mathbb{Z}))^{r\times m}$

$$\lim_{n\to\infty} ||a_n * c||_p^{\frac{1}{n}} = \rho_p\big(\{A_0|_{V(c)},\ A_1|_{V(c)}\}\big).$$

Consequently, for a finite dimension $\mathcal{A}$-invariant subspace $V \subseteq (\ell(K))^{r\times m}$ we have

$$\max\Big\{\lim_{n\to\infty} ||a_n * \lambda||_p^{\frac{1}{n}} :\ \lambda \in V\Big\} = \rho_p\big(\{A_0|_V,\ A_1|_V\}\big). \tag{2.10}$$

In fact, it was established in [10] for $m = r = 1$, and in [14] for $m = 1$ and general r. We refer to [10] and [14] for the details. By the same methods and (2.9), we can establish (2.10) for our setting.

For a sequence $c \in (\ell_0(\mathbb{Z}))^{r\times r}$, we define a subspace of $(\ell_0(\mathbb{Z}))^{r\times 1}$ as follows:

$$V(c*) := \big\{\lambda \in (\ell_0(\mathbb{Z}))^{r\times 1} : \lambda = c * \eta,\ \eta \in (\ell_0(\mathbb{Z}))^{r\times 1}\big\}. \tag{2.11}$$

With the help of (2.9) we can prove the following result.

THEOREM 2.3. *Recall that K_N is given as in* (2.1). *Let $c \in (\ell(K_N))^{r\times r}$. If $V(c*)$ is an $\mathcal{A}$ invariant subspace, we have*

$$\lim_{n\to\infty} ||a_n * c||_p^{\frac{1}{n}} = \rho_p\big(\{A_0|_{V(c*)\cap(\ell(K_N))^{r\times 1}}, A_1|_{V(c*)\cap(\ell(K_N))^{r\times 1}}\}\big).$$

Proof. Let $V := V(c*)\cap(\ell(K_N))^{r\times 1}$. As mentioned, $(\ell(K_N))^{r\times 1}$ is an $\mathcal{A}$ invariant subspace. Therefore, V is an $\mathcal{A}$ invariant subspace by the assumption.

For any $\lambda \in V(c*)$, there is an $\eta \in (\ell_0(\mathbb{Z}))^{r\times 1}$ such that $\lambda = c*\eta$. Thus, $||a_n*\lambda||_p \le \kappa_\lambda||a_n * c||_p, n = 1,2,\ldots,$ for some constant κ_λ. Thus,

$$\lim_{n\to\infty} ||a_n * \lambda||_p^{\frac{1}{n}} \le \lim_{n\to\infty} ||a_n * c||_p^{\frac{1}{n}} \qquad \forall \lambda \in V(c*).$$

On the other hand, we define η_j by setting $\eta_j(\alpha) = 0\ \forall\alpha \ne 0$ and $\eta_j(0)$ the jth-column of $r \times r$ identity matrix, $1 \le j \le r$. Moreover, let $\lambda_j = c * \eta_j, 1 \le j \le r$. Obviously, $a_n * \lambda_j$ is just the jth-column of $a_n * c, 1 \le j \le r$. It follows that

$$\lim_{n\to\infty} ||a_n * c||_p^{\frac{1}{n}} \le \max_{1\le j\le r} \lim_{n\to\infty} ||a_n * \lambda_j||_p^{\frac{1}{n}}.$$

Note that $\lambda_j \in V, 1 \le j \le r$. We have by the last two inequalities

$$\lim_{n\to\infty} ||a_n * c||_p^{\frac{1}{n}} = \max_{\lambda\in V} \lim_{n\to\infty} ||a_n * \lambda||_p^{\frac{1}{n}}.$$

The proof is complete by (2.10). □

Combining Theorems 2.2 and 2.3 we have the following result.

COROLLARY 2.4. *Suppose that a, b, and c are in $(\ell_0(\mathbb{Z}))^{r\times r}$ and satisfy the conditions of Theorem* 2.2. *In addition, suppose that $V(c*)$ is an invariant subspace of A_0 and A_1. Then*

$$\rho_p(S_b) = \rho_p\big(\{A_0|_{V(c*)\cap(\ell(K_N))^{r\times 1}}, A_1|_{V(c*)\cap(\ell(K_N))^{r\times 1}}\}\big).$$

3. Structures of some invariant subspaces. The main purpose of this section is to represent explicitly some invariant subspaces of both A_0 and A_1 in terms of convolution. These subspaces are important to the study of convergence of cascade algorithms.

In the study of convergence of cascade algorithms and regularities of function vectors, we usually restrict ourselves to a class of masks a satisfying some conditions on the modulii of eigenvalues of $H_a(0)$. The reason will soon be clear.

Suppose that s is a nonnegative integer. Recall that $H_a(\omega)$ is defined in section 1. Denote by E_{s+1} the set of all sequences $a \in (\ell_0(\mathbb{Z}))^{r\times r}$ satisfying the following conditions:

(1) 1 is a simple eigenvalue of $H_a(0)$.

(2) The other eigenvalues of $H_a(0)$ are of modulus less than 2^{-s}.

It was proved in [3] that $a \in E_{s+1}$ provided that the refinable function vector $\Phi_a \in (W_p^s(\mathbb{R}))^r$ satisfies $\widehat{\Phi}_a(0) \ne 0$ and the independence condition

$$\operatorname{span}\Big\{\widehat{\Phi}_a(2\alpha\pi) :\ \alpha \in \mathbb{Z}\Big\} = \mathbb{C}^r.$$

If $a \in E_{s+1}$, there exists a unique, up to a constant factor, $1 \times r$ vector $t_a(\omega)$ of trigonometric polynomials with spectrum contained in $\{\mu : 0 \le \mu \le s\}$ satisfying

$$D^\mu\big(t_a(2\cdot)H_a\big)(0) = D^\mu t_a(0), \qquad 0 \le \mu \le s. \tag{3.1}$$

We say a sequence $a \in (\ell_0(\mathbb{Z}))^{r\times r}$ satisfies *sum rules of order $s+1$ with t_a* if there is a vector t_a of trigonometric polynomials with $t_a(0) \ne 0$ such that (3.1) and the following equalities are true:

$$D^\mu\big(t_a(2\cdot)H_a\big)(\pi) = 0, \qquad 0 \le \mu \le s.$$

Let $\mathcal{S}_{s+1}$ denote the set of all sequences satisfying sum rules of order $s+1$. It is known that there is a close relation between the order of sum rules and the approximation order provided by the refinable function vector; see, for example, [16], [13], and [2].

For a mask $a \in E_{s+1}$, we define

$$V_a^s := \Big\{\lambda : \lambda \in (\ell_0(\mathbb{Z}))^{r\times 1}, D^\mu\big(t_a\widehat{\lambda}\big)(0) = 0,\ 0 \le \mu \le s\Big\}. \tag{3.2}$$

The importance of V_a^s lies in the following result.

LEMMA 3.1 (see [3]). *Let $\mathcal{A} = \{A_0, A_1\}$. Assume that $a \in E_{s+1}$ and V_a^s is given as above. Then V_a^s is an $\mathcal{A}$ invariant subspace if and only if a satisfies sum rules of order $s+1$ with t_a.*

For later use we will identify V_a^s with a subspace $V(c*)$ for some finitely supported sequence c determined by a. To this end, we need the results of [16] and [18] about factorizations of masks which satisfy sum rules of appropriate order.

Let $\mathcal{C}$ be the set of $r\times r$ matrices $C(\omega)$ with entries being 2π periodic trigonometric polynomials such that

(1) $C(0)$ has a simple eigenvalue 0;

(2) $\det C(\omega)=\kappa(1-e^{-i\omega})e^{-i\gamma\omega}$, where $\kappa\neq 0$ and $\gamma\in\mathbb{Z}$ are constants.

For any $x\in\mathbb{C}^r\backslash\{0\}$, we denote by C_x a matrix in $\mathcal{C}$ such that x is a left eigenvector of $C_x(0)$ with eigenvalue 0. We note that such a matrix is not unique. A special form of such matrices can be found in [16]. Furthermore, any two matrices C_x^1 and C_x^2 in $\mathcal{C}$ that have the same x as a left eigenvector with eigenvalue 0 are related by $C_x^1(\omega)M(\omega)=C_x^2(\omega)$, where the $r\times r$ matrix $M(\omega)$ is invertible for any ω and its entries are 2π periodic trigonometric polynomials.

By the definition of $\mathcal{C}$ we know there is a $r\times r$ matrix $G_x(\omega)$ with entries being 2π periodic trigonometric polynomials such that the following conditions are satisfied:

$$C_x(\omega)^{-1}=\frac{G_x(\omega)}{1-e^{-i\omega}}\qquad\forall\omega\notin 2\pi\mathbb{Z}. \tag{3.3}$$

The set $\mathcal{C}$ plays an important role in the factorization of refinement masks. The following result was established in [16] with special form of C_x and in [18] for the general case.

LEMMA 3.2 (see [16] and [18]). *Assume that $a\in E_{s+1}$. Then a satisfies sum rules of order $s+1$ with t_a if and only if there are nonzero vectors $x_0,x_1,\dots,x_s\in\mathbb{C}^r$ and a sequence $\tilde{a}\in(\ell_0(\mathbb{Z}))^{r\times r}$ such that the following conditions are satisfied:*

(1) *a can be factorized as follows:*

$$\widehat{a}C_{x_s}\cdots C_{x_0}=\frac{1}{2^{s+1}}C_{x_s}(2\cdot)\cdots C_{x_0}(2\cdot)\widehat{\tilde{a}}. \tag{3.4}$$

(2) *a^μ, $\mu=0,1,\dots,s-1$, satisfies sum rules of order $\mu+1$, where $a^\mu\in(\ell_0(\mathbb{Z}))^{r\times r}$ is defined by its Fourier transform*

$$\widehat{a^\mu}(\omega)=\frac{1}{2^{\mu+1}}C_{x_\mu}(2\omega)\cdots C_{x_0}(2\omega)\widehat{\tilde{a}(\omega)}C_{x_0}^{-1}(\omega)\cdots C_{x_\mu}^{-1}(\omega)\qquad\forall\omega\notin 2\pi\mathbb{Z}.$$

Suppose that $a\in E_{s+1}$ satisfies sum rules of order $s+1$. Then the factorization as in Lemma 3.2 holds. It was proved in [6] that 1 is a simple eigenvalue of $H_{a^j}(0), 0\le j\le s$, where $a^s=a$. Moreover, each $x_j, 0\le j\le s$, is a left eigenvector of $H_{a^j}(0)$ with the eigenvalue 1. In particular,

$$x_s=t_a(0)^T. \tag{3.5}$$

We recall that $a\in(\ell(K_N))^{r\times r}$. Suppose that a satisfies the conditions of Lemma 3.2. Without loss of generality we assume that $a^\mu, 0\le\mu\le s-1$, determined by condition (2) of Lemma 3.2, satisfy $a^\mu\in(\ell(K_N))^{r\times r}$ as well.

For our consideration we restate a result of Plonka and Strela in the following slightly different form. As in Lemma 3.2, it was proved in [16] with special C_x and in [18] for the general case.

LEMMA 3.3 (see [18, Theorem 2.4]). *Suppose that $a\in E_{s+1}$ and a satisfies sum rules of order $s+1$ with t_a. Let a^{s-1} and x_s be given as in condition (2) of Lemma 3.2. Then a^{s-1} satisfies sum rules of order s with $t_{a^{s-1}}$ verifying*

$$D^\mu t_{a^{s-1}}(0)=\frac{i}{\mu+1}D^{\mu+1}(t_aC_{x_s})(0),\qquad 0\le\mu\le s-1.$$

We are in a position to characterize the structure of V_a^s in terms of the factorization of mask which satisfies sum rules of order $s+1$.

THEOREM 3.4. *Suppose that a satisfies the hypotheses of Lemma 3.2 and, consequently, (3.4) is true for some $C_{x_s},\dots,C_{x_0}$. Let $C_s = C_{x_s}\cdots C_{x_0}$. Then a vector $\lambda \in (\ell_0(\mathbb{Z}))^{r\times 1}$ belongs to V_a^s if and only if there is a $\xi \in (\ell_0(\mathbb{Z}))^{r\times 1}$ such that*

$$\widehat{\lambda} = C_s\widehat{\xi}. \tag{3.6}$$

Proof. Let a satisfy sum rules of order $s+1$ with t_a. Recall that a^{s-1} is given as in (2) of Lemma 3.2. We first claim that

$$V_a^s = \left\{\lambda : \widehat{\lambda} = C_{x_s}\widehat{\eta},\ \eta \in V_{a^{s-1}}^{s-1}\right\}. \tag{3.7}$$

In fact, suppose first that $\lambda \in (\ell_0(\mathbb{Z}))^{r\times r}$ satisfies $\widehat{\lambda} = C_{x_s}\eta$ for some $\eta \in V_{a^{s-1}}^{s-1}$. By the definition of $C_{x_s}(\omega)$ and equality (3.5), it is easily seen that $t_a(0)C_{x_s}(0) = 0$ and, consequently,

$$t_a(0)\widehat{\lambda}(0) = 0.$$

Moreover, it follows from this equality and Lemma 3.3 that

$$\begin{aligned} D^{\mu+1}(t_a\widehat{\lambda})(0) &= \sum_{\nu\le\mu+1}\binom{\mu+1}{\nu}D^{\mu+1-\nu}(t_aC_{x_s})(0)D^\nu\widehat{\eta}(0)\\ &= \frac{\mu+1}{i}\sum_{\nu\le\mu}\binom{\mu}{\nu}D^{\mu-\nu}t_{a^{s-1}}(0)D^\nu\widehat{\eta}(0). \end{aligned}$$

We thus obtain

$$D^{\mu+1}(t_a\widehat{\lambda})(0) = \frac{\mu+1}{i}D^\mu(t_{a^{s-1}}\widehat{\eta})(0), \qquad 0\le\mu\le s-1. \tag{3.8}$$

Since $\eta \in V_{a^{s-1}}^{s-1}, D^{\mu+1}(t_a\widehat{\lambda})(0) = 0$ by (3.8), $0\le\mu\le s-1$. This proves $\lambda \in V_a^s$. Conversely, let $\lambda \in V_a^s$. Since $\operatorname{rank}C_{x_s}(0) = r-1$, it follows from (3.3) that $\operatorname{rank}G_{x_s}(0) = 1$. Furthermore, each row of $G_{x_s}(0)$ is a multiple of x_s^T due to the fact that x_s^T is a left eigenvector of $H_{x_s}(0)$ with the simple eigenvalue 1. Note that $t_a(0) = x_s^T$. Consequently, we have $G_{x_s}(0)\widehat{\lambda}(0) = 0$. This implies that there is an $\eta \in (\ell_0(\mathbb{Z}))^{r\times 1}$ such that $G_{x_s}(\omega)\widehat{\lambda}(\omega) = (1-e^{-i\omega})\widehat{\eta}(\omega)$, $\omega\in\mathbb{R}$. It follows from (3.3) that $\widehat{\eta}(\omega) = C_{x_s}(\omega)^{-1}\widehat{\lambda}(\omega)$, $\omega\in\mathbb{R}$. Substituting it into the right-hand side of (3.8) and appealing to Lemma 3.3 we conclude that (3.8) is true for any $0\le\mu\le s-1$. This together with $\lambda\in V_a^s$ tells us $\eta\in V_{a^{s-1}}^{s-1}$. Therefore we have proved (3.7), as claimed.

By replacing V_a^s with $V_{a^{s-1}}^{s-1},\dots,V_{a^1}^1$ recursively in (3.7), we know that $\lambda\in V_a^s$ if and only if there is an $\eta\in V_{a^0}^0$ such that

$$\widehat{\lambda} = C_{x_s}\cdots C_{x_1}\widehat{\eta}.$$

However, as is known, $a^0\in E_1$ and a^0 satisfies sum rules of order 1 with

$$t_{a^0}(\omega) = x_0^T \qquad \forall\omega\in\mathbb{R}.$$

Therefore, by the definitions of $V_{a^0}^0$ and $C_{x_0}(\omega)$, we have

$$V_{a^0}^0 = \left\{\eta\in(\ell_0(\mathbb{Z}))^{r\times 1} : \widehat{\eta} = C_{x_0}\widehat{\xi}, \xi\in(\ell_0(\mathbb{Z}))^{r\times 1}\right\}.$$

Thus (3.6) is true. The proof is complete. □

Let $c_s \in (\ell_0(\mathbb{Z}))^{r\times r}$ be given by its Fourier transform as follows:

$$\widehat{c}_s = C_{x_s} \cdots C_{x_0}. \tag{3.9}$$

Then we restate Theorem 3.4 in the following form.

COROLLARY 3.5. *Under the conditions of Theorem* 3.4 *and with the notations as above we have*

$$V_a^s = V(c_s *).$$

4. Convergence of cascade algorithms. In this section we establish the equivalence between the convergence of a cascade algorithm in Sobolev space and the convergence of an associated cascade algorithm in L_p space. Therefore, the problem of convergence of cascade algorithms in Sobolev norm may reduce to that in L_p space. On the other hand, an algorithm for construction of refinement masks which generate convergent cascade algorithms in Sobolev space is presented. The algorithm is easy to implement.

Assume as before that $a \in (\ell_0(\mathbb{Z}))^{r\times r}$ and $a \in E_{s+1}$. As mentioned in section 3, there is a unique vector t_a, up to a constant factor, such that (3.1) holds. Using $t_a(\omega)$, we define W_a^s to be the set of vectors F of compactly supported functions in $W_p^s(\mathbb{R})$ satisfying

$$t_a(0)\widehat{F}(0) = t_a(0)\widehat{\Phi}_a(0) \qquad \text{and} \qquad D^\mu(t_a\widehat{F})(2\alpha\pi) = 0 \ \ \forall \alpha \neq 0,\ 0 \le \mu \le s. \tag{4.1}$$

It had been proved in [3] that if

$$\lim_{n\to\infty} ||Q_a^n F_0 - \Phi_a||_{(W_p^s(\mathbb{R}))^r} = 0, \tag{4.2}$$

for some vector F_0 of compactly supported functions, then $F_0 \in W_a^s$. Therefore, the notion of convergence was defined in [3] as follows. Let $a \in E_{s+1}$. We say that the cascade algorithm generated by Q_a *converges in* $(W_p^s(\mathbb{R}))^r$ *norm* if (4.2) holds for any $F_0 \in W_a^s$.

The characterization of the convergence of cascade algorithm in terms of the p-joint spectral radius is given as follows.

THEOREM 4.1 (see [3]). *Assume that $a \in E_{s+1}$. Suppose that $H_a(\omega)$ and $t_a(\omega)$ satisfy* (3.1). *Let V_a^s be defined in* (3.2). *Then the cascade algorithm generated by Q_a converges in $(W_p^s(\mathbb{R}))^r$ norm if and only if the following conditions are satisfied:*

(1) *V_a^s is invariant under $A_\varepsilon, \varepsilon = 0, 1$.*

(2) *$\rho_p\left(\{A_0|_{V_N}, A_1|_{V_N}\}\right) < 2^{-s+1/p}$, where $V_N = V_a^s \cap (\ell(K_N))^{r\times 1}$.*

We are in a position to establish an equivalence between the convergence of a cascade algorithm in Sobolev space on one hand and the convergence of an associated cascade algorithm in L_p norm on the other hand.

THEOREM 4.2. *Assume that $a \in E_{s+1}$. Then the cascade algorithm generated by Q_a converges in $(W_p^s(\mathbb{R}))^r$ if and only if the following conditions are satisfied:*

(1) *There are $s+1$ nonzero vectors $x_0, x_1, \ldots, x_s \in \mathbb{C}^r$ and a sequence $\tilde{a} \in (\ell_0(\mathbb{Z}))^{r\times r}$ such that a satisfies conditions* (1) *and* (2) *of Lemma* 3.2.

(2) *The cascade algorithm corresponding to a^0 converges in $(W_p(\mathbb{R}))^r$, where a^0 is given in* (2) *of Lemma* 3.2.

If this is the case, then $\widehat{\Phi}_{a^0}(0) \neq 0$ and the refinable function vectors Φ_a and Φ_{a^0} are related by

$$(i\omega)^s\widehat{\Phi}_a = \kappa C_{x_s} \cdots C_{x_{s-1}}\widehat{\Phi}_{a^0}, \tag{4.3}$$

where κ is a constant.

Proof. Suppose that the cascade algorithm generated by Q_a converges in $(W_p^s(\mathbb{R}))^r$. Then $s+1 \le N$ and $a \in \mathcal{S}_{s+1}$ by [3]. Therefore, conditions (1) and (2) of Lemma 3.2 are true for some vectors $x_0, \dots, x_s$ and a sequence $\tilde{a} \in (\ell_0(\mathbb{Z}))^{r\times r}$.

Let $c_s \in (\ell_0(\mathbb{Z}))^{r\times r}$ be given as in (3.9). Then $V(c_s*) = V_a^s$ by Corollary 3.5 and, consequently, $V(c_s*)$ is invariant under $A_\varepsilon, \varepsilon = 0, 1$, by Theorem 4.1.

On the other hand, we may rewrite (3.4) as

$$a * c_s = 2^{-s-1} S_{\tilde{a}}.$$

By the requirements of $\mathcal{C}$, the Fourier transform $\widehat{c}_s$ of c_s is invertible for any $\omega \notin 2\pi\mathbb{Z}$. Consequently, c_s satisfies condition (2.2) on c. It follows from Corollary 2.4 that

$$\rho_p\big(\{A_0|_{V_a^s \cap (\ell(K_N))^{r\times 1}}, A_1|_{V_a^s \cap (\ell(K_N))^{r\times 1}}\}\big) = 2^{-s-1}\rho_p(S_{\tilde{a}}). \tag{4.4}$$

Let a^0 be defined in condition (2) of Lemma 3.2. As is known, $a^0 \in E_1$ and a^0 satisfies sum rules of order 1 with a constant vector $t_{a^0}(\omega) = x_0^T \ \forall \omega \in \mathbb{R}$. Let $A_\varepsilon^0, \varepsilon = 0, 1$, be defined in (2.7) by replacing a with a^0. Then $V_{a^0}^0$ is invariant under $A_\varepsilon^0, \varepsilon = 0, 1$, where $V_{a^0}^0$, corresponding to a^0, is given by (3.2). Similar to the proof of (4.4) we can establish

$$\rho_p\left(\left\{A_0^0|_{V_{a^0}^0 \cap (\ell(K_N))^{r\times 1}}, A_1^0|_{V_{a^0}^0 \cap (\ell(K_N))^{r\times 1}}\right\}\right) = 2^{-1}\rho_p(S_{\tilde{a}}). \tag{4.5}$$

Furthermore, by condition (2) of Theorem 4.1 and equality (4.4) we have

$$\rho_p(S_{\tilde{a}}) < 2^{1+1/p}. \tag{4.6}$$

It in turn implies by (4.5) that

$$\rho_p\left(\left\{A_0^0|_{V_{a^0}^0 \cap (\ell(K_N))^{r\times 1}}, A_1^0|_{V_{a^0}^0 \cap (\ell(K_N))^{r\times 1}}\right\}\right) < 2^{1/p}. \tag{4.7}$$

Therefore, the cascade algorithm associated with a^0 converges in $(W_p(\mathbb{R}))^r$ norm by Theorem 4.1. This proves the necessity of (1) and (2).

Assume now that conditions (1) and (2) are true. Then V_a^s as above is an invariant subspace of A_0 and A_1, thereby verifying condition (1) of Theorem 4.1. Moreover, by applying Theorem 4.1 to the cascade algorithm generated by Q_{a^0} and $s = 0$ we get (4.7). It is easy to deduce condition (2) of Theorem 4.1 from (4.4), (4.5), and (4.7). This proves the sufficiency of the theorem.

Finally, if the conditions of the theorem are satisfied, the relation (4.3) between Φ_a and Φ_{a^0} may be found, e.g., in [18]. Moreover, $\widehat{\Phi}_{a^0}(0) \neq 0$ since it is a right eigenvector of $H_{a^0}(0)$. The proof is complete. □

From the proof of Theorem 4.2 we know that, when a mask a is factorized as in condition (1) of Lemma 3.2, the cascade algorithm generated by Q_a converges in $(W_p^s(\mathbb{R}))^r$ norm if and only if the corresponding sequence $\tilde{a}$ satisfies (4.6).

While the matrices C_x are determined by a left eigenvector of $C_x(0)$ to the eigenvalue 0, it is convenient sometimes, as observed in [18], to identity the matrices with the help of right eigenvectors to the same eigenvalue. More precisely, we let y be a right eigenvector of $C_x(0)$ with eigenvalue 0, and we set $M_y(\omega) = C_x(\omega)$. Therefore, for any finitely supported sequence $a \in E_{s+1}$, $a \in \mathcal{S}_{s+1}$ is equivalent to the factorization

$$\widehat{a} M_{y_s} \cdots M_{y_0} = \frac{1}{2^{s+1}} M_{y_s}(2\cdot) \cdots M_{y_0}(2\cdot)\widehat{\tilde{a}}, \tag{4.8}$$

where y_j is a right eigenvector of $H_{a^{j-1}}(0)$, $0 \le j \le s$, to the eigenvalue 0, the sequences $a^j \in (\ell_0(\mathbb{Z}))^{r\times r}$ are given by

$$H_{a^j}(\omega) = \frac{1}{2} M_{y_j}(2\omega) H_{a^{j-1}}(\omega) M_{y_j}(\omega)^{-1}, \quad \omega \notin 2\pi\mathbb{Z}, \quad 0 \le j \le s-1,$$

and $a^{-1} = \tilde{a}$. We refer to Theorem 2.7 and Corollary 2.8 of [18] for the details.

At the end of this section we present an algorithm for construction of refinement masks which generate convergent cascade algorithms in Sobolev space.

ALGORITHM 4.3. *Start with a finitely supported sequence $\tilde{a}$ such that $H_{\tilde{a}}(0)$ has 1 as a simple eigenvalue and $\rho(H_{\tilde{a}}(0)) < 2$. Suppose that $\rho_p(S_{\tilde{a}}) < 2^{1+1/p}$. Let y_0 be a right eigenvector of $H_{\tilde{a}}(0)$ associated with eigenvalue 1. Choose a matrix $M_{y_0} \in \mathcal{C}$ satisfying $M_{y_0}(0)y_0 = 0$. Define a finitely supported sequence a^0 by*

$$H_{a^0}(\omega) = \frac{1}{2} M_{y_0}(2\omega) H_{\tilde{a}}(\omega) M_{y_0}(\omega)^{-1} \qquad \forall \omega \notin 2\pi\mathbb{Z}. \tag{4.9}$$

(1) *Find a right eigenvector y_{j+1} of $H_{a^j}(0)$ associated with eigenvalue 1.*
(2) *Construct a finitely supported sequence a^{j+1} by*

$$H_{a^{j+1}}(\omega) = \frac{1}{2} M_{y_{j+1}}(2\omega) H_{a^j}(\omega) M_{y_{j+1}}(\omega)^{-1} \qquad \forall \omega \notin 2\pi\mathbb{Z},$$

where $M_{y_{j+1}} \in \mathcal{C}$ satisfies $M_{y_{j+1}}(0)y_{j+1} = 0$.

(3) *Repeat steps 1 and 2 as many times as needed.*

s cycles of Algorithm (steps 1, 2, and 3) yield a refinement mask a^s generating a convergent cascade algorithm in Sobolev space $(W_p^s(\mathbb{R}))^r$ norm.

Let us justify our algorithm. By the assumptions on $\tilde{a}$ we know that the spectrum of $H_{\tilde{a}}(0)$ is $\{1, \mu_1, \dots, \mu_{r-1}\}$ with $|\mu_j| < 2, j = 1, \dots, r-1$. It follows from [6] that the spectrum of $H_{a^j}(0)$ is of form $\{1, 2^{-j-1}\mu_1, \dots, 2^{-j-1}\mu_{r-1}\}$. Consequently, 1 is a simple eigenvalue of $H_{a^j}(0)$. So step 3 is consistent. Clearly, $a^s \in E_{s+1}$. Moreover, $a^s \in \mathcal{S}_{s+1}$ by Theorem 2.7 and Corollary 2.8 in [18].

Finally, since $S_{\tilde{a}}$ satisfies (4.6), the cascade algorithm generated by Q_{a^s} converges in Sobolev space $(W_p^s(\mathbb{R}))^r$ norm by what is mentioned in the paragraph. ∎

5. Examples. The following examples will illustrate our theory.

EXAMPLE 5.1. *We consider a refinable function vector taken from* [8]. *Let $a \subseteq (\ell(\{0,1,2,3\}))^{2\times 2}$. Its Fourier transform $\widehat{a}$ is*

$$\begin{pmatrix} -\frac{(t^2-4t-3)(1+z)}{2(t+2)} & 1 \\ -\frac{3(t-1)(t+1)((t^2-3t-1)(1+z^3)+(t^2-t+3)(z+z^2))}{4(t+2)^2} & \frac{(3t^2+t-1)(1+z^2)}{2(t+2)} + z \end{pmatrix}, \qquad z = e^{-i\omega}.$$

Then, if $|t| < 1/2$, the cascade algorithm generated by Q_a converges in $(W_p^1(\mathbb{R}))^2$ norm for $1 \le p < \infty$. If $1/2 < |t| < 1$, it converges in $(W_p(\mathbb{R}))^2$ norm for $1 \le p \le \infty$.

Proof. It is known from [6] that (3.4) holds for $s = 1$ with $x_0 = (1,1)^T$, $x_1 = (-3(t^2-1)/(t+2), 1)^T$, and

$$\widehat{\tilde{a}}(\omega) = \begin{pmatrix} 2 & 0 \\ \frac{(t^2-3t-1)z^2+(-10t^2-8t+6)z+(t^2-3t-1)}{(t+2)} & 4t(1+z) \end{pmatrix}, \qquad z = e^{-i\omega}.$$

Let us first compute the spectral radius $\rho_p(S_{\tilde{a}})$ of $S_{\tilde{a}}$. To this end we cite the following formula of [14]:

$$\rho_p(S_{\tilde{a}}) = \rho_p\left(\left\{\tilde{A}_0|_{(\ell(\{0,1\}))^{2\times 1}}, \tilde{A}_1|_{(\ell(\{0,1\}))^{2\times 1}}\right\}\right), \tag{5.1}$$

where $\tilde{A}_\varepsilon$ is defined as in section 2 by replacing a with $\tilde{a}, \varepsilon = 0, 1$. Moreover, we identify any $\lambda \in (\ell(\{0,1\}))^{2\times 1}$ with a vector

$$\begin{pmatrix}\lambda(0)\\ \lambda(1)\end{pmatrix} \in \mathbb{C}^4$$

and, therefore, $\tilde{A}_\varepsilon|_{(\ell(\{0,1\}))^{2\times 1}}$, $\varepsilon = 0, 1$, with operators on $\mathbb{C}^4$. Choose a basis of $\mathbb{C}^4$ as follows. $e_1 = (0,0,1,0)^T$, $e_2 = (1,0,0,0)^T$, $e_3 = (0,0,0,1)^T$, $e_4 = (0,1,0,0)^T$. Denote by T_ε the representing matrix of $\tilde{A}_\varepsilon|_{(\ell(\{0,1\})^{2\times 1}}$ on this basis. Then by a simple computation we know

$$T_0 = \begin{pmatrix} 4t & 0 & \frac{-1-3t+t^2}{2+t} & 0\\ 0 & 4t & \frac{-1-3t+t^2}{2+t} & \frac{-2(-3+4t+5t^2)}{2+t}\\ 0 & 0 & 2 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$T_1 = \begin{pmatrix} 4t & 4t & \frac{-2(-3+4t+5t^2)}{2+t} & \frac{-1-3t+t^2}{2+t}\\ 0 & 0 & 0 & \frac{-1-3t+t^2}{2+t}\\ 0 & 0 & 0 & 2\\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Appealing to (4.3) of [14] we have

$$\rho_p\left(\left\{\tilde{A}_0, \tilde{A}_1\right\}\right) = 4\max\{1/2, 2^{1/p}|t|\}, \qquad 1 \le p \le \infty.$$

It follows from (5.1) that $\rho_p(S_{\tilde{a}}) = 4\max\{1/2, 2^{1/p}|t|\}$.

As mentioned in the paragraph following Theorem 4.2, the cascade algorithm generated by Q_a converges in $(W_p^1(\mathbb{R}))^2$ norm if and only if (4.6) holds. Therefore, if $|t| < 1/2$, the cascade algorithm generated by Q_a converges in $(W_p^1(\mathbb{R}))^2$ norm for $1 \le p < \infty$ and *does not* converge in $(W_\infty^1(\mathbb{R}))^2$ norm.

For $1/2 < |t| < 1$ and $1 \le p \le \infty$, the cascade algorithm converges in $(W_p(\mathbb{R}))^2$ norm by Theorem 4.2. □

EXAMPLE 5.2. *Let* $y^1 = (\sqrt{2}, 1)^T$ *and* $a^0 \subseteq (\ell(\{0,1,2,3\}))^{2\times 2}$ *be given by*

$$H_{a^0}(\omega) = \frac{1}{20}\begin{pmatrix} 6+6e^{-i\omega} & 8\sqrt{2}\\ (-1+9e^{-i\omega}+9e^{-2i\omega}-e^{-3i\omega})/\sqrt{2} & -3+10e^{-i\omega}-3e^{-2i\omega}\end{pmatrix}.$$

If we define a^1 *by*

$$H_{a^1}(\omega) = \frac{1}{2}M_{y_1}(2\omega)H_{a^0}(\omega)M_{y_1}(\omega)^{-1}, \qquad \omega \notin 2\pi\mathbb{Z},$$

then the cascade algorithm generated by Q_{a^1} *converges in* $(W_\infty^1(\mathbb{R}))^2$ *norm.*

Proof. It is known (see [18]) that the normalized refinable function vector $\Phi_{a^0} = (\phi_1, \phi_2)^T$ has orthogonal shifts. Furthermore, ϕ_1 and ϕ_2 are continuous. It follows from [14] that the cascade algorithm generated by Q_{a^0} converges in $(W_\infty(\mathbb{R}))^2$ norm.

Therefore, by the proof of Theorem 4.2, a^0 can be factorized as in (4.9) for some $\tilde{a}$ and y_0, and $\tilde{a}$ satisfies the requirements of Algorithm 4.3. In order to construct a mask which generates a cascade algorithm converging in $(W_\infty^1(\mathbb{R}))^2$ norm, we need only to use one cycle of Algorithm 4.3.

In fact, it is easy to check that $y_1 = (\sqrt{2}, 1)^T$ is a right eigenvector of $H_{a^0}(0)$ with eigenvalue 1. Therefore, the conclusion is true by Algorithm 4.3.

In particular, if we set as in [18]

$$M_{y_1}(\omega) = \begin{pmatrix} 1 + e^{-i\omega} & -2\sqrt{2} \\ 1 - e^{-i\omega} & 0 \end{pmatrix},$$

then the mask a^1 is given by

$$H_{a^1}(\omega) = \frac{1}{40} \begin{pmatrix} -7 + 10e^{-i\omega} - 7e^{-2i\omega} & 15(1 - e^{-2i\omega} \\ -4(1 - e^{-2i\omega}) & 10(1 + e^{-i\omega})^2 \end{pmatrix}.$$

In this case, the components of Φ_{a^1} are symmetric [18]. □

Acknowledgments. The author thanks Prof. Rong-Qing Jia and Dr. Ding-Xuan Zhou for their very helpful comments and the anonymous referees for their valuable comments which helped to improve the presentation of the paper.

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[2] D. R. Chen, *Algebraic properties of subdivision operators with matrix mask and their applications*, J. Approx. Theory, 97 (1999), pp. 294–310.

[3] D. R. Chen, R. Q. Jia, and S. D. Riemenschneider, *Vector subdivision schemes in Sobolev spaces*, Appl. Comput. Harmon. Anal., 12 (2001), pp. 128–149.

[4] D. R. Chen and G. Plonka, *Convergence of cascade algorithms in Sobolev space for perturbed refinement masks*, J. Approx. Theory, to appear.

[5] D. R. Chen and X. B. Zheng, *Spectral Radii and Eigenvalues of Subdivision Operators*, manuscript, 2000.

[6] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[7] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[8] G. C. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal, 27 (1996), pp. 1158–1192.

[9] T. N. T. Goodman and S. L. Lee, *Convergence of cascade algorithms*, in Mathematical Methods for Curves and Surfaces II, M. Daehlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 191–212.

[10] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[11] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[12] R. Q. Jia, Q. T. Jiang, and S. L. Lee, *Convergence of cascade algorithms in Sobolev spaces and integrals of wavelets*, Numer. Math., 91 (2002), pp. 453–473.

[13] R. Q. Jia, S. D. Riemenschneider, and D. X. Zhou, *Approximation by multiple refinable functions*, Canad. J. Math., 49 (1998), pp. 944–962.

[14] R. Q. Jia, S. D. Riemenschneider, and D. X. Zhou, *Vector subdivision schemes and multiple wavelets*, Math. Comp., 67 (1998), pp. 1533–1563.

[15] C. A. Micchelli and T. Sauer, *Regularity of multiwavelets*, Adv. Comp. Math., 7 (1997), pp. 455–545.

[16] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.

[17] G. Plonka and A. Ron, *A new factorization technique of matrix mask of univariate refinable functions*, Numer. Math., 87 (2001), pp. 555–595.

[18] G. Plonka and V. Strela, *Construction of multiscaling functions with approximation and symmetry*, SIAM J. Math. Anal., 29 (1998), pp. 481–510.
[19] G.-C. Rota and G. Strang, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.
[20] Y. Wang, *Two scale dilation equations and the mean spectral radius*, Random Comput. Dynam., 4 (1996), pp. 49–72.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1369–1404

ENO-WAVELET TRANSFORMS FOR PIECEWISE SMOOTH FUNCTIONS*

TONY F. CHAN† AND H. M. ZHOU‡

Abstract. We have designed an adaptive essentially nonoscillatory (ENO)-wavelet transform for approximating discontinuous functions without oscillations near the discontinuities. Our approach is to apply the main idea from ENO schemes for numerical shock capturing to standard wavelet transforms. The crucial point is that the wavelet coefficients are computed without differencing function values across jumps. However, we accomplish this in a different way than in the standard ENO schemes. Whereas in the standard ENO schemes the stencils are adaptively chosen, in the ENO-wavelet transforms we adaptively change the function and use the same uniform stencils. The ENO-wavelet transform retains the essential properties and advantages of standard wavelet transforms such as concentrating the energy to the low frequencies, obtaining maximum accuracy, maintained up to the discontinuities, and having a multiresolution framework and fast algorithms, all without any edge artifacts. We have obtained a rigorous approximation error bound which shows that the error in the ENO-wavelet approximation depends only on the size of the derivative of the function away from the discontinuities. We will show some numerical examples to illustrate this error estimate.

Key words. ENO, wavelet, image compression, image denoising, signal processing

AMS subject classifications. 65D15, 65T60, 68P30, 94A08

PII. S0036142900370915

1. Introduction.

In this paper, we develop new wavelet algorithms to approximate piecewise continuous functions, for instance piecewise smooth functions connected by large jumps. It is well known that wavelet linear approximation (i.e., truncating the high frequencies) can approximate smooth functions very efficiently: It can achieve high order accuracy by selecting appropriate wavelet basis; it can concentrate the large wavelet coefficients in the low frequencies; and it has a multiresolution framework and associated fast transform algorithms.

Standard wavelet linear approximation techniques cannot achieve similar results for functions which are not smooth, for example piecewise smooth functions with large jumps in function value or in its derivatives. Several problems arise near jumps, primarily caused by the well-known Gibbs phenomenon. The jumps generate large high frequency wavelet coefficients and thus linear approximation cannot get the same high accuracy near the points of discontinuity as in the smooth region. In fact, the jump points generate oscillations which cannot be removed by mesh refinement.

To overcome these problems within the standard wavelet transform framework, nonlinear data-dependent approximations, which selectively retain certain high frequency coefficients, are often used, e.g., hard and soft thresholding techniques; see [6], [15], [19], [18], [27], and corresponding references therein. The main idea of these thresholding approximations is to truncate both low and high frequency wavelet coefficients by their magnitudes, not frequencies. For instance, hard thresholding sets all coefficients whose magnitudes are less than a given tolerance to zero and retains the

*Received by the editors April 18, 2000; accepted for publication (in revised form) February 26, 2002; published electronically October 23, 2002. This research was supported in part by grants ONR-N00017-96-1-0277 and NSF DMS-96-26755.

http://www.siam.org/journals/sinum/40-4/37091.html

†Department of Mathematics, The University of California, Los Angeles, CA 90095-1555 (chan@math.ucla.edu).

‡Department of Applied and Computational Mathematics, Mail Code 217-50, California Institute of Technology, Pasadena, CA 91125 (hmzhou@acm.caltech.edu).

other coefficients unchanged. It has been verified through many research efforts that such nonlinear processes can effectively reduce Gibbs oscillations, and therefore they have been widely used in many applications such as image compression and denoise, and even computation of partial differential equations. However, these techniques often require more complicated data structure to record the location of the retained wavelet coefficients and still cannot remove the effects of the Gibbs phenomenon completely unless all jump-related coefficients are preserved.

Another fundamental approach is to modify the wavelet transform to not generate large high frequency wavelet coefficients near jumps. A few papers in the literature have discussed this approach. Claypoole et al. [12] proposed an adaptive lifting scheme which lowers the order of approximation near jumps, thus minimizing the Gibbs effect. Consequently, this scheme suffers from reduced approximation accuracy near jumps, and some residual Gibbs phenomenon still exists. Another way due to Donoho is to construct orthonormal basis such as wedgelets [16] and ridgelets [7], [17] to represent the discontinuities.

In this paper, we develop a new wavelet algorithm by borrowing the well-developed essentially nonoscillatory (ENO) technique for shock capturing in computational fluid dynamics (e.g., see [23] and [29]) to modify the standard wavelet transform near discontinuities in order to overcome the above-mentioned difficulties. ENO schemes are systematic ways of adaptively defining piecewise polynomial approximations of the given functions according to their smoothness. There are two crucial points in designing ENO schemes. The first is to use one-sided information near jumps and never differencing across the discontinuities. The second is to adaptively form the divided difference table and select the smoothest *stencil* (the support of the basis) for every grid point. ENO schemes lead to uniform high accuracy approximations for each smooth piece of the function. We will use only the first point in our design of the ENO-wavelet transforms. Preliminary results of this work have been reported in [8].

Combining the ENO idea with the multiresolution data representation is a natural way to avoid oscillations in the approximations. In fact, it has been explored by Harten in his general framework of multiresolution in [20], [21], and [22]. (The lifting scheme of Sweldens [31] uses a similar idea.) Recent studies of his general framework and its application in data compression can be found in [2], [3], [4], and [9]. Harten's approach is to directly blend the two ideas and to fully implement the ENO schemes at every point. This consists of using the adaptive ENO finite difference table to select the stencil and then compute the decomposition as well as the reconstruction process. However, his method cannot be directly applied to the more generally used pyramidal filtering algorithms which the standard wavelet transforms are implemented in because in this context we have to work only with fixed size and fixed value filters, and these rigid filters cannot be directly used to compute the adaptive divided difference tables at each grid point.

Our goal is to design a more direct functional replacement of the standard wavelet transforms such that there are no oscillations at the discontinuities in the approximations. We want to stick with the classical pyramidal filtering framework because they are easy to use and have been successfully applied in many applications. Compared to Harten's multiresolution approach, which is more flexible and easier to adaptively implement than the ENO idea, the standard wavelet transforms are more regular and rigid in algorithmic structure; therefore, directly applying the ENO idea would lead to a more drastic perturbation of the underlying pyramidal filtering algorithms. This is the challenge we face.

Conceptually, the ENO-wavelet transforms that we will introduce in this paper are closely related to the ENO implementation of Harten's multiresolution framework. Both methods share the one-sided information idea, which computes the decomposition and reconstruction from smooth data. However, we achieve this in a different manner. The way we accomplish this is to not change the wavelet transforms or the filter coefficients, which most data-dependent multiresolution algorithms do, but instead locally change the function near the discontinuities in such a way that the standard filters are applied only to smooth data. By recording how the changes are made, the original discontinuous function can be exactly recovered by using the original inverse filters. Indeed, by applying the idea of using one-sided information near the discontinuities, we directly extend the functions from both sides of the discontinuities, thus we can apply the standard wavelet transforms on these extended values such that there are no large coefficients generated in the high frequencies and the low frequency approximations are essentially nonoscillatory, and therefore the Gibbs phenomenon can be completely avoided.

In addition, in this modified wavelet transform, the low frequency part preserves the piecewise smoothness of the original function. In particular, the jumps in the low frequency part is not spread widely as in the standard transform. Therefore, the same ENO idea can be recursively used for the coarser levels of the low pass coefficients. By doing so, the multiresolution framework also can be kept.

We show that the resulting wavelet transform retains all the desirable properties of the standard transform: It can have uniformly maximum accuracy, maintained up to the discontinuities (with a rigorous uniform order of the error bound); it concentrates the large coefficients to the low frequencies; it preserves the multiresolution framework and fast transform algorithms; and it is easy to implement. Furthermore, since we do not fully adopt the ENO schemes, in particular we do not build the divided difference table and compare the smoothness of all possible stencils at every point, the extra cost (in floating point operations) required by the modified ENO-wavelet transforms is insignificant. In fact, it is of the order $O(dl)$, where d is the number of discontinuities and $l+1$ the stencil length. Compared to the cost of the standard wavelet transform, which is of the order $O(nl)$, where n is the size of the data, the ratio of the extra cost over that of the standard transform is of the order $O(\frac{d}{n})$ which is independent of l and negligible when n is large.

Besides, since the designed ENO-wavelet transforms play the same role as the standard wavelet transforms in the applications, in principle, any of the numerous existing algorithms for postprocessing wavelet coefficients can also be used in conjunction with the ENO-wavelet coefficients. For example, ENO-wavelet transforms can be used in conjunction with the standard adaptive nonlinear techniques such as hard and soft thresholding, tree structured (e.g., Shapiro's EZW [28]) coders in image compression, and Coifman and Donoho's translation invariant algorithm [10] in denoising. However, in this paper we focus on the construction of ENO-wavelet transforms, and we will not discuss those applications in detail. Instead, we show a numerical example which illustrates the advantages of using the combination of ENO-wavelet transforms with hard thresholding in section 5.

The arrangement of the paper is as follows. In section 2 , we review the standard continuous and discrete wavelet transforms. In section 3, we give a general algorithm to implement the ENO-wavelet transform discretely. In section 4, we prove an error bound for the ENO-wavelet approximation which shows that the error in the ENO-wavelet approximation depends only on the size of the derivative of the function *away*

from the discontinuities. Finally, in section 5, we give some numerical examples to illustrate the main advantage of the ENO-wavelet transforms, including some two-dimensional (2-D) examples.

2. Wavelet transforms. Before we introduce the adaptive ENO-wavelet transforms, we briefly review the standard wavelet transforms; e.g., see [5], [11], [13], [14], [25], [26], [27], and [30]. We use Daubechies orthonormal wavelets as the framework in all discussion in this paper. We will go over both continuous and discrete wavelet transforms, because we will present our ENO-wavelet transforms in the discrete form and prove the approximation error bound by using the continuous form.

First, we review the standard wavelet transforms. To simplify the notation, we assume zeros have been padded to the data at the boundaries.

The standard wavelet transforms are based on translation and dilation. Suppose $\phi(x)$ and $\psi(x)$ are the scaling function and the corresponding wavelet, respectively, with finite support $[0, l]$, where l is a positive integer. It is well known that $\phi(x)$ satisfies the basic dilation equation

$$\phi(x) = \sqrt{2} \sum_{s=0}^{l} c_s \phi(2x - s) \tag{1}$$

and $\psi(x)$ satisfies the corresponding wavelet equation

$$\psi(x) = \sqrt{2} \sum_{s=0}^{l} h_s \phi(2x - s), \tag{2}$$

where the c_s's and h_s's are constants called low pass and high pass filter coefficients, respectively.

We assume that $\psi(x)$ has p vanishing moments

$$\int \psi(x) x^j dx = 0 \qquad \text{for} \quad j = 0, 1, \ldots, p-1. \tag{3}$$

We will use the following standard notations:

$$\phi_{j,i}(x) = 2^{\frac{j}{2}} \phi(2^j x - i) \tag{4}$$

and

$$\psi_{j,i}(x) = 2^{\frac{j}{2}} \phi(2^j x - i). \tag{5}$$

Consider the subspace V_j of L^2 defined by

$$V_j = Span\{\phi_{j,i}(x), i \in Z\}$$

and the subspace W_j of L^2 defined by

$$W_j = Span\{\psi_{j,i}(x), i \in Z\}.$$

The subspaces V_j's, $-\infty < j < \infty$, form a multiresolution of L^2 with the subspace W_j being the difference between V_j and V_{j+1}. In fact, the L^2 space has an orthonormal decomposition as

$$L^2 = V_J \oplus \sum_{j=J}^{\infty} W_j. \tag{6}$$

The projection of a L^2 function $f(x)$ onto the subspace V_j is defined by

$$f_j(x) = \sum_i \alpha_{j,i}\phi_{j,i}(x), \tag{7}$$

where

$$\alpha_{j,i} = \int f(x)\phi_{j,i}(x)dx, \qquad i = \cdots, -1, 0, 1, \ldots, \tag{8}$$

which we call low frequency wavelet coefficients. (They are often called scaling coefficients in many literatures.) Similarly, we can project $f(x)$ onto W_j by

$$w_j(x) = \sum_i \beta_{j,i}\psi_{j,i}(x), \tag{9}$$

where

$$\beta_{j,i} = \int f(x)\psi_{j,i}(x)dx, \quad i = \cdots, -1, 0, 1, \ldots, \tag{10}$$

which we call high frequency wavelet coefficients (often called wavelet coefficients in many literatures). In this paper, we refer to wavelet coefficients as both low and high frequency coefficients. Therefore, the function $f(x)$ can be decomposed by

$$f(x) = f_j(x) + \sum_{t=j}^{\infty} w_t(x). \tag{11}$$

The projection $f_j(x)$ is called the linear approximation of the function $f(x)$ in the subspace V_j.

From (4) and (5), the projection coefficients $\alpha_{j,i}$ and $\beta_{j,i}$ of $f(x)$ in the subspaces V_j and W_j can be easily computed by the so-called fast wavelet transform

$$\alpha_{j,i} = \sum_{s=0}^{l} c_s \alpha_{j+1,2i+s} \tag{12}$$

and

$$\beta_{j,i} = \sum_{s=0}^{l} h_s \alpha_{j+1,2i+s}. \tag{13}$$

In practice, discrete wavelet transforms are often directly used with a set of discrete numbers which are the low frequency coefficients of the L^2 function $f(x)$ at a fine level subspace V_{j+1}. In many applications, this set of numbers are sample values of the function $f(x)$ on a fine grid (although in [30] this is called a "wavelet crime").

Let us define the following matrices:

$$L = \begin{pmatrix} c_0 & c_1 & \cdots & c_l & & & & & \\ & & c_0 & c_1 & \cdots & c_l & & & \\ & & & & \cdots & \cdots & \cdots & & \\ & & & & & c_0 & c_1 & \cdots & c_l \end{pmatrix}$$

and

$$H = \begin{pmatrix} h_0 & h_1 & \cdots & h_l & & & & \\ & & h_0 & h_1 & \cdots & h_l & & \\ & & & & \cdots & \cdots & \cdots & \\ & & & & & h_0 & h_1 & \cdots & h_l \end{pmatrix}.$$

We also denote $\vec{\alpha}_j = (\ldots, \alpha_{j,i}, \alpha_{j,i+1}, \ldots)^T$ and $\vec{\beta}_j = (\ldots, \beta_{j,i}, \beta_{j,i+1}, \ldots)^T$.

By using matrix and vector forms, the fast wavelet transform equations (12) and (13) can be written as

$$\vec{\alpha}_j = L\vec{\alpha}_{j+1} \tag{14}$$

and

$$\vec{\beta}_j = H\vec{\alpha}_{j+1}. \tag{15}$$

It is well known that the wavelet transform matrices L and H are orthogonal:

$$L^*L + H^*H = I. \tag{16}$$

It follows that the inverse wavelet transform is simply

$$\vec{\alpha}_{j+1} = L^*\vec{\alpha}_j + H^*\vec{\beta}_j. \tag{17}$$

The standard linear wavelet approximation achieves maximum accuracy away from discontinuities, but it oscillates near the jumps. The reason for the oscillations is that some stencils cross jumps and cause the corresponding high frequency coefficients to becoming large and therefore more information is lost when the high frequency coefficients are discarded.

In Figure 1, we display a piecewise continuous function (left) and its DB4 wavelet coefficients (right) with low frequencies at the left end and high frequencies at the right end. From the right picture, we see that most of the high frequency coefficients are zero, except for a few large coefficients which are computed near jumps. Figure 2 displays the linear approximation (solid line) compared to the initial function (dotted line). The right picture is the zoom-in to show the approximation behavior near a jump. In this figure, we clearly see oscillations (people call them the Gibbs phenomenon) near discontinuities.

Since the oscillations are generated by discarding large high frequency coefficients which are computed on the stencils crossing discontinuities, to get rid of the oscillations, we want to avoid stencils crossing discontinuities. This motivates us to apply the ENO idea to avoid stencils crossing jumps.

Before we introduce the ENO-wavelet transforms, we give the following definition which we will use in the later sections. Given a function $f(x)$ which has discontinuous set D, then

$$D = \{x_i : f(x) \quad \text{is discontinuous at} \quad x_i\}.$$

Denote t as the closest distance between any two discontinuous points, i.e.,

$$t = \inf\{|x_i - x_j| : x_i, x_j \in D\}.$$

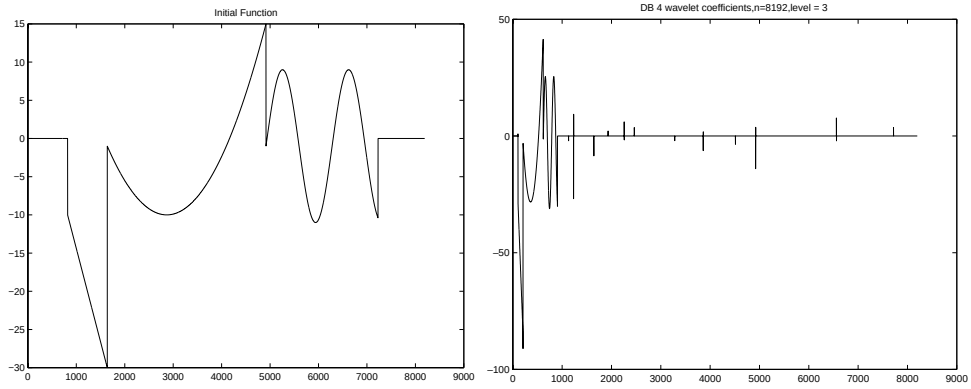


FIG. 1. *The initial function (left) and its DB4 coefficients (right). Most of the high frequency coefficients (right part) are zero except for a few large coefficients computed near the jumps.*

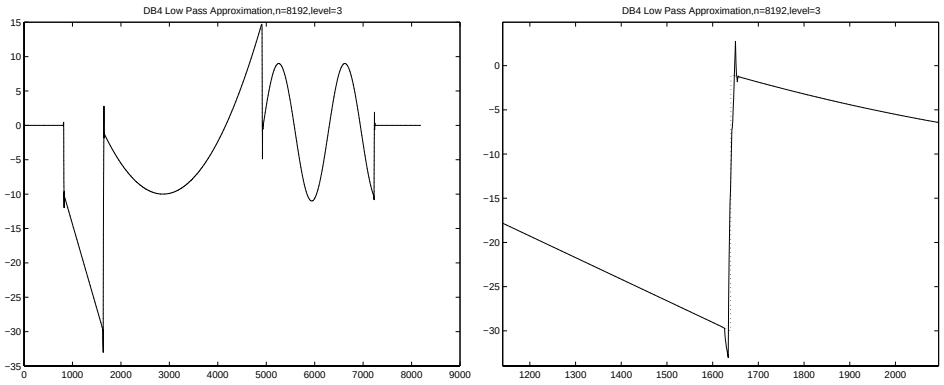


FIG. 2. *The approximation function (left) and its zoom-in (right). Oscillations are generated near the discontinuities in the linear approximation.*

DEFINITION 1. *For a given wavelet filter with stencil length $l+1$, we say a projection of $f(x)$ in space V_j with spatial step $\Delta x = 2^{-j}$ satisfies the* discontinuity separation property *(DSP) if $(l+2)\Delta x < t$.*

A projection satisfying the DSP implies that any one discontinuity is located at least one stencil and two data points away from other discontinuities. In other words, there are no two consecutive stencils containing two discontinuities. We assume that all projections we consider in this paper satisfy the DSP. Since our ENO-wavelet transform is essentially using ENO techniques to modify the standard wavelet transform near discontinuities, this property will avoid the modifications near one discontinuity interacting with the modifications near other discontinuities.

Remark. For any piecewise discontinuous function, a projection will satisfy this DSP if j is sufficiently large, i.e., if the discretization is fine enough. On the other hand, at the place where the DSP is invalid, the approximations produced by the ENO-wavelet transforms are comparable to that by the standard wavelet transforms. We will show numerical examples in section 5 illustrating this point.

3. ENO-wavelet transforms. In this section, we design the ENO-wavelet transforms. In addition to the standard wavelet transforms, our ENO-wavelet

transforms are composed of two phases: locating the jumps and forming the approximations at the discontinuities. First, assuming that the location of the jumps are known, we give the ENO-wavelet approximations at the discontinuities by using one-sided information to avoid oscillations. Then we give the methods to detect the exact subinterval on the next finer grid at which the discontinuity is located.

3.1. ENO-wavelet approximation at discontinuities. In this subsection, we assume that the exact subintervals on the next finer grid at which the discontinuities is located are known. We want to modify the standard wavelet transforms near the jumps such that oscillations can be avoided in the approximation. From ENO schemes, we borrow the idea of using one-sided information to form the approximation and avoid applying the wavelet filters crossing the discontinuities. Since we assume the DSP is satisfied by the given projection of the function $f(x)$, we can just consider the local modification near one jump. The main tool which we use to modify the standard wavelet transforms at the discontinuities is function extrapolation in the function spaces or in the wavelet spaces.

Direct function extrapolation. The first way is to extend the function directly at the discontinuity by extrapolation from both sides. Then we can apply the standard wavelet transforms on the extended functions and avoid computing wavelet coefficients using information from both sides.

To maintain the same approximation accuracy near the discontinuity as that for away from the discontinuity, the extrapolation has to be pth order accurate if the wavelet functions have p vanishing moments. For instance, we use constant extrapolation for Haar wavelets and $(p-1)$th order extrapolation for Daubechies-$2p$ orthogonal wavelets which have p vanishing moments.

We use the diagram in Figure 3 to show how to extend the function and compute the ENO-wavelet coefficients.

As shown in Figure 3, the discontinuity is located between $\{x(2i+l-2), x(2i+l-1)\}$. We extend the function from both sides of the discontinuity using $(p-1)$th order extrapolation; i.e., we use the information from the left side of the jump to extrapolate the function over $\hat{x}(2i+l-1), \ldots, \hat{x}(2i+2l-2)$ and use the information from the right side to extrapolate the function over $\bar{x}(2i), \ldots, \bar{x}(2i+l-2)$. And then for $i \le m \le i+k-2$, where $l = 2k-1$, we can compute the wavelet coefficients $\hat{\alpha}_{j,m}$ and $\hat{\beta}_{j,m}$ from the left side, and compute $\bar{\alpha}_{j,m}$ and $\bar{\beta}_{j,m}$ from the right side by using the standard wavelet transforms, respectively.

In general, we have the low frequency wavelet coefficients on the finer levels instead of knowing the function values themselves near the discontinuities. We extrapolate these finer level coefficients from both sides of the discontinuities to obtain the values of $\hat{\alpha}_{j+1,m}$ and $\bar{\alpha}_{j+1,m}$, and use the fast wavelet transforms (12) and (13) to compute the coarser level coefficients. For instance, we can compute $\hat{\alpha}_{j,i}$ and $\hat{\beta}_{j,i}$ by

$$\begin{aligned}\begin{pmatrix} \hat{\alpha}_{j,i} \\ \hat{\beta}_{j,i} \end{pmatrix} &= \begin{pmatrix} \sum_{s=0}^{l-2} c_s \alpha_{j+1,2i+s} + c_{l-1}\hat{\alpha}_{j+1,2i+l-1} + c_l \hat{\alpha}_{j+1,2i+l} \\ \sum_{s=0}^{l-2} h_s \alpha_{j+1,2i+s} + h_{l-1}\hat{\alpha}_{j+1,2i+l-1} + h_l \hat{\alpha}_{j+1,2i+l} \end{pmatrix} \\ &\equiv \begin{pmatrix} \delta_{j,i} \\ \gamma_{j,i} \end{pmatrix} + A \begin{pmatrix} \hat{\alpha}_{j+1,2i+l-1} \\ \hat{\alpha}_{j+1,2i+l} \end{pmatrix}, \end{aligned} \tag{18}$$

where $\delta_{j,i}$ and $\gamma_{j,i}$ are $\sum_{s=0}^{l-2} c_s \alpha_{j+1,2i+s}$ and $\sum_{s=0}^{l-2} h_s \alpha_{j+1,2i+s}$, respectively, and depend only on the unextrapolated values of $\alpha_{j+1,m}$, and A a 2×2 matrix defined by

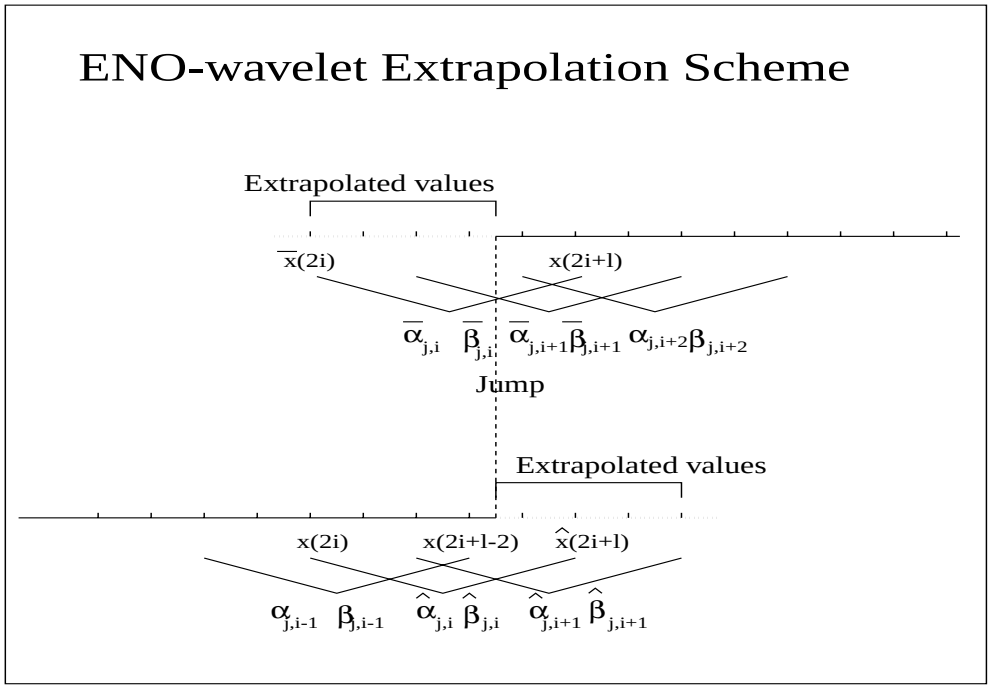


FIG. 3. *Coarse level extrapolation illustration. From the left side of the discontinuity, we extrapolate the low frequency coefficients $\hat{\alpha}_{j,m}$ to determine corresponding high frequency coefficients $\hat{\beta}_{j,m}$ and store them. From the right side of the discontinuity, we extend the high frequency coefficients $\bar{\beta}_{j,m}$ to determine and store the low frequency coefficients $\bar{\alpha}_{j,m}$.*

the filter coefficients as

$$A = \begin{pmatrix} c_{l-1} & c_l \\ h_{l-1} & h_l \end{pmatrix}.$$

In computing $\hat{\alpha}_{j,m}$ and $\hat{\beta}_{j,m}$ by the fast wavelet transforms, the number of extrapolated values we must use is 2 for $m = i$, 4 for $m = i + 1$, and so on. Those extrapolated values are determined from the smooth side of the discontinuity; then the high frequency coefficients $\hat{\beta}_{j,m}$ remain as small values as those of the smooth stencils.

By symmetry, we can compute $\bar{\alpha}_{j,m}$'s and $\bar{\beta}_{j,m}$'s from the right side in a similar way.

There are many methods to extrapolate the extended values. For example, a straightforward way is to use p-point polynomial extrapolation such as Lagrange polynomials or Taylor expansion polynomials. In our numerical experiments in this paper, we use Lagrange polynomial extrapolation. Least square extrapolation can be used too [33], especially for noisy data.

There is a storage problem for this direct function extrapolation. Indeed, it doubles the number of the wavelet coefficients near every discontinuity. To retain the perfect invertible property, we need to store the ENO-wavelet coefficients $\hat{\alpha}_{j,m}$ and $\hat{\beta}_{j,m}$ from the left side, also $\bar{\alpha}_{j,m}$ and $\bar{\beta}_{j,m}$ from the right side. Thus, the output sequences are no longer the same size as the input sequences. In many applications, such as image compression, this extra storage requirement definitely needs to be avoided.

Remark. In the least square extrapolation case, it is possible to reduce the demands of the extra storage because not all the wavelet coefficients $\hat{\alpha}_{j,m}$, $\hat{\beta}_{j,m}$, $\bar{\alpha}_{j,m}$, and $\bar{\beta}_{j,m}$ are linearly independent [33]. However, this requires complicated extra computation.

Our approach: Coarse level extrapolation. To avoid computing the wavelet coefficients using the information from both sides of the discontinuities, to maintain the same high order accuracy near the discontinuities as away from the discontinuities, and also to keep the size of the output sequences the same as that of the input sequences without significant extra computation, we introduce the coarse level extrapolation schemes. The idea is to extrapolate the coarser level wavelet coefficients near the discontinuities instead of the function values or the finer level wavelet coefficients.

We still use Figure 3 to illustrate these schemes. We consider the left side of the jump first.

In the direct function extrapolation case, the computation process is to directly extrapolate the finer level wavelet coefficients $\hat{\alpha}_{j+1,m}$, $(2i+l-1) \le m \le (2i+2l-2)$, and then compute the extended coarser level wavelet coefficients $\hat{\alpha}_{j,m}$ and $\hat{\beta}_{j,m}$, $i \le m \le (i+k-2)$ using the standard filters. We reverse the order of this process in our coarse level extrapolation. More precisely, we extrapolate the coarser level low frequency coefficients $\hat{\alpha}_{j,m}$ using the known low frequency coefficients from the left, and extend the coarser level high frequency coefficients $\hat{\beta}_{j,m}$ to zero (or some predefined values), and then determine the extended finer level wavelet coefficients. For example, in the direct function extrapolation, we extrapolate finer level values $\hat{\alpha}_{j+1,m}$ and then compute the coarser level coefficients $\hat{\alpha}_{j,i}$ and $\hat{\beta}_{j,i}$ by (18). On the contrary, we can first extend the coarser level coefficients $\hat{\alpha}_{j,i}$ and $\hat{\beta}_{j,i}$ and then determine the finer level values. Indeed, if the matrix A is nonsingular, we can uniquely determine the finer level values by solving (18). In this case, we can prescribe both the coarser level coefficients simultaneously. However, in Daubechies orthogonal wavelet transforms, the matrix A is singular, because

$$\frac{h_{l-1}}{c_{l-1}} = \frac{h_l}{c_l}. \tag{19}$$

Thus, in order to have a solution of (18), we must extend the coarser level coefficients $\hat{\alpha}_{j,i}$ and $\hat{\beta}_{j,i}$ in such a way that they satisfy

$$\begin{pmatrix} \hat{\alpha}_{j,i} \\ \hat{\beta}_{j,i} \end{pmatrix} - \begin{pmatrix} \delta_{j,i} \\ \gamma_{j,i} \end{pmatrix} \in R(A),$$

where $R(A)$ is the range space of A. This requirement implies that

$$\begin{pmatrix} -1 & \dfrac{c_l}{h_l} \end{pmatrix} \left[\begin{pmatrix} \hat{\alpha}_{j,i} \\ \hat{\beta}_{j,i} \end{pmatrix} - \begin{pmatrix} \delta_{j,i} \\ \gamma_{j,i} \end{pmatrix} \right] = 0,$$

which we can also rewrite as

$$\hat{\beta}_{j,i} = \gamma_{j,i} + \frac{h_l}{c_l}(\hat{\alpha}_{j,i} - \delta_{j,i}) \tag{20}$$

or

$$\hat{\alpha}_{j,i} = \delta_{j,i} + \frac{c_l}{h_l}(\hat{\beta}_{j,i} - \gamma_{j,i}). \tag{21}$$

Therefore, we cannot prescribe both $\hat{\alpha}_{j,i}$ and $\hat{\beta}_{j,i}$ simultaneously. Thus we have two choices:

(1) We can extrapolate the low frequency coefficients $\hat{\alpha}_{j,i}$ first and then determine the corresponding high frequency coefficients $\hat{\beta}_{j,i}$ by (20).
(2) Or we can extend $\hat{\beta}_{j,i}$ to zero first and then determine the corresponding $\hat{\alpha}_{j,i}$ by (21).

Once coefficients $\hat{\beta}_{j,i}$ and $\hat{\alpha}_{j,i}$ are obtained, we can determine the finer level coefficients $\hat{\alpha}_{j+1,2i+l-1}$ and $\hat{\alpha}_{j+1,2i+l}$. Since A is not invertible for Daubechies wavelets, $\hat{\alpha}_{j+1,2i+l-1}$ and $\hat{\alpha}_{j+1,2i+l}$ cannot be uniquely determined by $\hat{\beta}_{j,i}$ and $\hat{\alpha}_{j,i}$. There is one more freedom left to use. (In the case of the discontinuity being located between $\alpha_{j+1,2i+l-1}$ and $\alpha_{j+1,2i+l}$, $\hat{\alpha}_{j+1,2i+l}$ can be uniquely determined.) Indeed, there are many ways to completely determine the values of $\hat{\alpha}_{j+1,2i+l-1}$ and $\hat{\alpha}_{j+1,2i+l}$. For instance, one can simply extend $\hat{\alpha}_{j+1,2i+l-1}$ by any extrapolation technique, such as $(p-1)$th order polynomial extrapolation for smooth data or averaging extrapolation techniques for noisy data (we use them in our numerical experiments in section 5), and then determine $\hat{\alpha}_{j+1,2i+l}$ by $\hat{\beta}_{j,i}$ or $\hat{\alpha}_{j,i}$. Another possible way to uniquely extend the coefficients $\hat{\alpha}_{j+1,2i+l-1}$ and $\hat{\alpha}_{j+1,2i+l}$ on the finer level is to leave this extra freedom to be used in the next stencil by requiring some special desire properties in the next extended coarser level coefficients. This involves slightly more complicated formulation which we will not exploit further in this paper. Thereafter, the above procedure can be repeatedly used to the next stencil to compute $\hat{\beta}_{j,i+1}$ and $\hat{\alpha}_{j,i+1}$ by treating $\hat{\alpha}_{j+1,2i+l-1}$ and $\hat{\alpha}_{j+1,2i+l}$ as known values. By the same principle, all extended coefficients $\hat{\beta}_{j,m}$ and $\hat{\alpha}_{j,m}$ can be calculated.

Remark. We notice that in both cases (20) and (21) the coefficients are computed by applying the standard filters to the extended data which is smooth. This implies that there are no large coefficients generated by them.

Again by symmetry, we have two analogous choices for the right side of the jump.

Using this coarse level extrapolation technique, we can easily solve the storage problem which we have in the direct function extrapolation. In fact, we just need to store the high frequency coefficients $\hat{\beta}_{j,m}$ for choice (1) and the low frequency coefficients $\hat{\alpha}_{j,m}$ for choice (2). In our implementation, we use choice (1) for the left side of the jumps and choice (2) for the right side of the jumps; therefore we store $\hat{\beta}_{j,m}$ and $\bar{\alpha}_{j,m}$ for every m. This satisfies the standard wavelet storage scheme, i.e., storing one low frequency and one high frequency coefficient for every stencil.

Remark. We select choice (1) from the left side of the jumps and choice (2) from the right side because we want to keep half of the output sequence to be α's and half to be β's. It is possible to select choice (1) or choice (2) for both sides of the jumps, but that will not give equal number of α's and β's in the output sequence; also, it may destroy the data structure for the next level decomposition.

Since we know the way we extend the data at the discontinuities, we can easily extrapolate the low frequency coefficients $\hat{\alpha}_{j,m}$ from the left sides of the discontinuities. Using them together with the stored high frequency coefficients $\hat{\beta}_{j,m}$, we can exactly recover data at the left sides by applying the standard inverse filters. Similarly, the data at the right sides of the discontinuities can also be exactly restored.

For each stencil crossing a jump, an extra cost (in floating point operation) is required in the extrapolation of low frequency coefficients, which is of the order $O(1)$ per stencil, and in the computation of the corresponding high and low frequency coefficients by (20) and (21), which is of the order $O(l)$ per stencil. Overall, the extra cost over the standard wavelet transform is of the order $O(dl)$, where d is the number of discontinuities. Compared to the cost of the standard wavelet transform, which is

of the order $O(nl)$ where n is the size of data, the ratio of the extra cost over that of the standard transform is $O(\frac{d}{n})$, which is independent of l and negligible when n is large.

3.2. Locating the discontinuities. In the previous subsection, we showed how to modify the standard wavelet transforms at the discontinuities to avoid oscillations if we know the exact subinterval on the next finer grid at which the jumps are located. In this subsection, we introduce the methods to detect those exact subintervals for discontinuities for piecewise smooth functions with and without noise. First we give a method for smooth data.

Piecewise smooth functions. Our purpose is to avoid wavelet stencils crossing discontinuities. Theoretically, a discontinuity can be characterized by comparing the left and right limit of the derivatives $f^{(m)}(x)$ at the given point x; i.e., we call a point x a discontinuity if for some $m < p$ we have

$$f^{(m)}(x-) \neq f^{(m)}(x+).$$

We define the intensity of a jump in the mth derivative at x as

$$[f^{(m)}(x)] = |f^{(m)}(x+) - f^{(m)}(x-)|.$$

It is well known that the high pass filters in wavelet transforms measure the smoothness of functions: they produce smaller values at smoother regions and larger values at rougher regions. In fact, it has been shown in [1], [24], and [32] that if a function $f(x)$ is Lipschitz $\gamma \leq p$ at x, i.e., $|f(x+\delta) - f(x)| \leq \delta^\gamma$ for any small δ, the corresponding high frequency wavelet coefficients are of the order of $O(\Delta x^\gamma)$. From this, it is easy to obtain that at smooth regions the magnitudes of high frequency coefficients $|\beta_{j,i}|$ have the order of $|f^{(p)}(x)|O(\Delta x^p)$. On the other hand, if a stencil contains a discontinuity, no matter if it is a discontinuity in function value ($m = 0$) or in its mth derivative, the magnitude of the corresponding high frequency coefficient $|\beta_{j,i}|$ is of the order of $O(\Delta x^{(m)})$, which is at least one order lower than that at the smooth regions. Therefore, instead of fully adopting the ENO comparison idea which compares the magnitudes of divided differences on all possible stencils, we can use the magnitudes of the high frequency coefficients as our criterion to identify the discontinuities.

The obvious way, also the cheapest way, to identify the discontinuities is to compare the magnitudes of the high frequency coefficients on the current standard stencils $|\beta_{j,i}|$ with that on the previous standard stencils $|\beta_{j,i-1}|$. Since for smooth functions we have $|\beta_{j,i}| = |f^{(p)}(x)|O(\Delta x^p)$, this implies that at smooth regions, by Taylor expansion, we have

$$|\beta_{j,i}| = (1 + O(\Delta x))|\beta_{j,i-1}|, \tag{22}$$

where the constant in the term $O(\Delta x)$ depends on the size of higher order derivatives of $f(x)$ such as $\max_x |f^{(p+1)}(x)|$. In contrast, the magnitudes of high frequency coefficients $|\beta_{j,i}|$ based on the stencils containing the discontinuities are at least one order lower than that at the smooth regions. More precisely, if we assume function $f(x)$ has a jump in its mth derivatives at point $x_0 \in (i\Delta x, (i+l)\Delta x)$ for some integer i, using Taylor expansion, in a small neighborhood of x_0, we can write this function as

$$f(x) = g(x) + \begin{cases} f^{(m)}(x_0-)(x-x_0)^m + O(x-x_0)^{(m+1)}, & x \leq x_0, \\ f^{(m)}(x_0+)(x-x_0)^m + O(x-x_0)^{(m+1)}, & x_0 < x, \end{cases}$$

where $g(x)$ is its Taylor polynomial of order $m-1$ near x_0. Then the wavelet coefficients $\beta_{j,i}$ is estimated by using the vanishing moments property as

$$\begin{aligned}|\beta_{j,i}| &= \left|\int f(x)\psi_{j,i}(x)dx\right| \\ &= \left|\int_{i\Delta x}^{x_0}(f^{(m)}(x_0-)(x-x_0)^m + O(x-x_0)^{m+1})\psi_{j,i}(x)dx\right. \\ &\quad \left. + \int_{x_0}^{(i+l)\Delta x}(f^{(m)}(x_0+)(x-x_0)^m + O(x-x_0)^{m+1})\psi_{j,i}(x)dx\right| \\ &= |[f^{(m)}(x_0)]|O(\Delta x^m).\end{aligned}$$

It depends on the Δx^m and also on the intensity of the jump.

Thus, we can design a method to detect the discontinuities as follows: For each standard stencil, suppose we know that the previous standard stencil does not contain any discontinuities, if we have $|\beta_{j,i}| \le a|\beta_{j,i-1}|$, where $a > 1$ is a given constant, and then we treat the current stencil as a smooth stencil. Otherwise, we conclude that there are discontinuities contained in it.

The choice of constant a depends on the grid size Δx and also on the intensity of the jumps. In fact, the ratio between a high frequency coefficient at the rough regions and that at the smooth regions is of the order of $|[f^{(m)}(x)]|O(\Delta x^{(m-p)})$. When Δx becomes small, this ratio is large. We can choose a as any number such that

$$(1 + O(\Delta x)) \le a \le \min_x\{|[f^{(m)}(x)]|O(\Delta x^{(m-p)})\}, \tag{23}$$

provided the above minimal number is larger than $1+O(\Delta x)$. This is always true for piecewise smooth functions with small enough grid size Δx.

Remark. When a jump in the mth derivative has very small intensity less than $O(\Delta x^{(p-m)})$, this jump cannot be detected by the above-described method. However, the error caused by missing this jump is also very small, which is at the same order of the error bound we will give in section 4. In practice, especially when we care only about the jumps in function values, we have a large range to select a.

To completely avoid oscillations, we also need to know the exact locations of the discontinuities so that we can avoid computing the wavelet coefficients crossing them. In fact, the above comparison method based on the magnitudes of high frequency coefficients can also help us to locate the exact positions of the discontinuities. We will use the diagram in Figure 4 to explain how to find the exact jump positions.

Assume we consider the wavelet filters with length $(l+1)$. We compare the magnitude of the high frequency coefficient $|\beta_{j,i}|$ on the current stencil, which starts at $x(2i)$ with $|\beta_{j,i-1}|$ on the previous stencil. If we have $|\beta_{j,i}| > a|\beta_{j,i-1}|$, we identify the discontinuity lying in the current stencil. Since there are no discontinuities in the previous stencils, we know that this discontinuity must be located between $\{x(2i+l-2), x(2i+l)\}$, where it has only two possible positions: between $\{x(2i+l-2), x(2i+l-1)\}$ or between $\{x(2i+l-1), x(2i+l)\}$. In fact, we can determine the exact position of the jump by continuing to compare the subsequent values of $\beta_{j,m}$. As shown in Figure 4, we must have at least $(k-1)$ consecutive "large" $\beta_{j,m}$, $i \le m \le (i+k-2)$, because the subsequent $(k-1)$ stencils also include the discontinuity. We compute $\beta_{j,i+k-1}$ and $\beta_{j,i+k}$ on the corresponding standard stencils, if we have $|\beta_{j,i+k-1}| > a|\beta_{j,i+k}|$, and then we have k consecutive stencils containing the discontinuity, which implies that the discontinuity is located between $\{x(2i+l-1), x(2i+2l-1)\}$ (see Figure 4(a)).

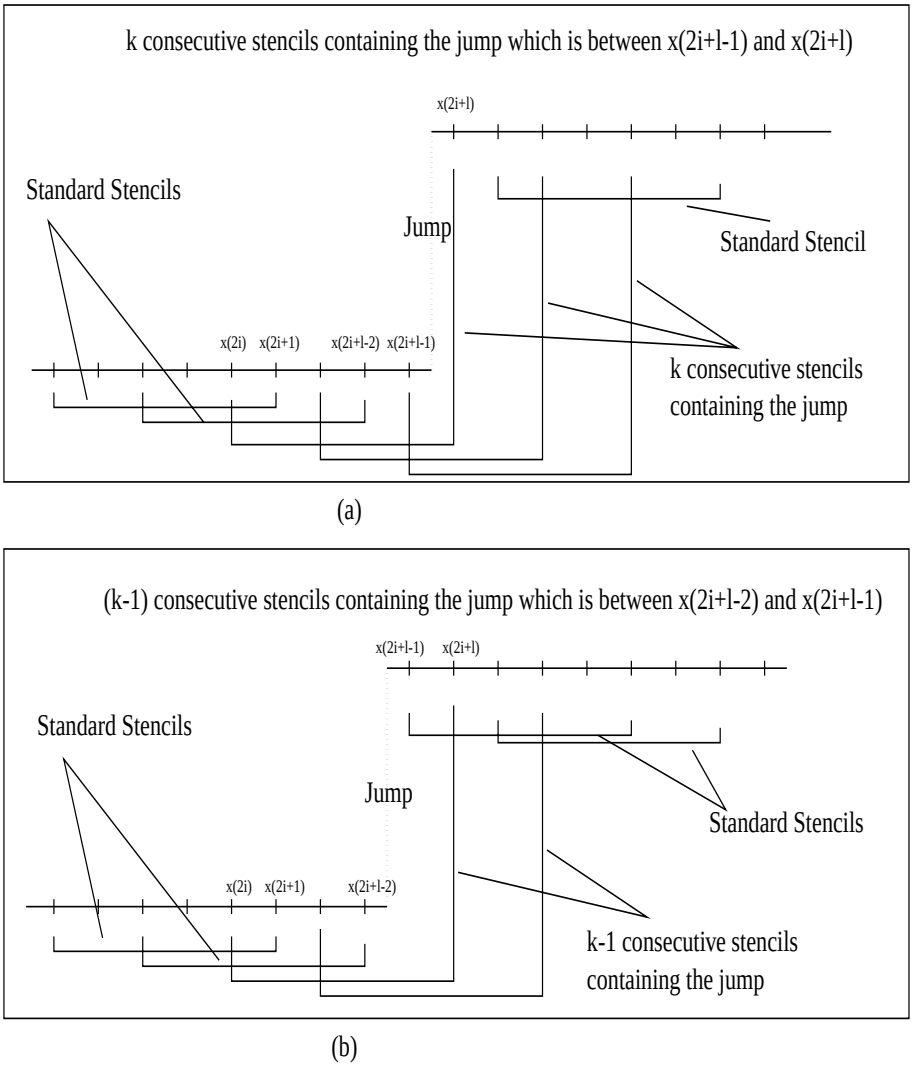


FIG. 4. *Locating the exact position of the jump by counting the number of consecutive stencils containing the jump.* (a) *If k stencils contain the jump, then the jump position is between $x(2i+l-1)$ and $x(2i+l)$.* (b) *If $(k-1)$ consecutive stencils contain the jump, then the jump is located between $x(2i+l-2)$ and $x(2i+l-1)$.*

If we have exactly $(k-1)$ consecutive standard stencils containing the discontinuity, then this implies that the jump must be located between $\{x(2i+l-2),\, x(2i+l-1)\}$ (see Figure 4(b)). We summarize the above arguments in the following proposition.

PROPOSITION 1. *Consider the wavelet filters with length $l+1$, where $l = 2k-1$. For a given index i, assume we have $|\beta_{j,i-1}| \le a|\beta_{j,i-2}|$ but $|\beta_{j,i}| > a|\beta_{j,i-1}|$. Then*

1. *if $|\beta_{j,i+k-1}| > a|\beta_{j,i+k}|$, which means there are k consecutive standard stencils containing the jump, then the discontinuity is located between $\{x(2i+l-1),\, x(2i+l)\}$;*
2. *or else we have $|\beta_{j,i+k-1}| \le a|\beta_{j,i+k}|$, which implies that there are $(k-1)$ consecutive standard stencils containing the jump, and then the discontinuity*

is located between $\{x(2i+l-2),\ x(2i+l-1)\}$.

The extra cost introduced by this comparison jump identification method over the standard wavelet transforms is just the comparison $|\beta_{j,i}| > a|\beta_{j,i-1}|$ for each stencil. In section 5, we use this detection method for all noise-free numerical examples.

Noisy data. The above-described detection method may not be reliable if the function is polluted by noise, especially when the noise is "large." This is because the high frequency coefficients β's may not be able to measure the correct order of smoothness of the functions. Indeed, the high frequency coefficients have the order $\|f^{(p)}(x)+\sigma n^{(p)}(x)\|O(\Delta x^p)$, where $n(x)$ is the random noise and σ a positive number indicating the noise level. In general, the derivatives of the noise $n^{(p)}(x)$ have large values. The noise term $\sigma n^{(p)}(x)$ can dominate the function term $f^{(p)}(x)$ if the noise level σ is large and thus the high frequency coefficients β's may not be able to detect certain discontinuities, e.g., if the jump is small or the discontinuity is in the higher derivatives. In this situation, we need to use heuristics to locate the exact position of the essential discontinuities. Here, we give a simple method to detect the significant large jumps in function values in noisy data.

In many applications, such as in image processing, large discontinuities in function value are the most significant features. Using the standard wavelet transforms, these large discontinuities will generate high frequency coefficients which can be much larger than those generated by the noise. (This is also the fundamental principle in the design of wavelet thresholding.) A simple way to detect these kinds of discontinuities is to look for these large magnitude high frequency coefficients and then compare the data values in the corresponding stencils to locate the exact jump positions. For example, we can look for the places which have the largest difference between two adjacent data values within the stencils. In our numerical experiments, we found that this simple way works very well in practice. In section 5, we will show an example using this method.

Remark. Other jump detection methods can be used for noisy data. As long as the exact subintervals of the discontinuities on the next finer grid are correctly determined, the coarse level ENO-wavelet approximations can be formed at the discontinuities, and our experience shows that it is not sensitive to the presence of noise.

In the ENO-wavelet transforms, to retain the perfect invertibility property, we need to store the adaptive information near every discontinuity, i.e., the exact location of the jump. The reason can also be illustrated by using Figure 4. If there is a jump in the low frequency coefficients (after down sampling) on the coarser level, one can predict a jump in the finer level coefficients. One can further identify the jump existing, for example, between $\{x(2i+l-2),\ x(2i+l)\}$ due to the down sampling. However, as shown in the diagram in Figure 4, for each identified jump, there are two possible locations, i.e., between $\{x(2i+l-2),\ x(2i+l-1)\}$ or between $\{x(2i+l-1),\ x(2i+l)\}$, in the finer level coefficients. Therefore, in order to achieve the perfect reconstruction, the exact locations of discontinuities have to be recorded. In our implementation, we just use one extra bit for each stencil near the discontinuities to indicate it contains a discontinuity. In the application of compression, which aims to reduce the total storage of representing an image, these extra bits need to be taken into account carefully. However, this is beyond the scope of this paper, and we will not discuss it here.

3.3. Forward and inverse transform algorithms. In this subsection, we explicitly present the complete one level forward and inverse ENO-wavelet transform algorithms for the noise-free piecewise smooth data.

We consider the forward transform algorithms first. We denote by $\{c_0, \ldots, c_l\}$ and $\{h_0, \ldots, h_l\}$ the standard wavelet filter coefficients, and by $\{r_0, \ldots, r_l\}$ and $\{d_0, \ldots, d_l\}$ the corresponding inverse filter coefficients. In this paper, since we consider Daubechies orthonormal wavelets, these inverse filter coefficients are defined as $r_s = (-1)^{s+1}h_s$, and $d_s = (-1)^s c_s$, for $s = 0, 1, \ldots, l$. We use a one-bit variable s_i to indicate whether a stencil contains a jump in our algorithms.

Forward Transform Algorithm.

For each i,

(i) compute $\beta_{j,i}$ by (13).

(ii) If $|\beta_{j,i}| \geq a|\beta_{j,i-1}|$ and $|\beta_{j,i}| \geq \epsilon$, then

- compute $\beta_{j,i+k-1}$ and $\beta_{j,i+k}$ by (13).
- Find the exact subinterval of the jump by Proposition 1. For $i \leq m \leq i+k$ or $i \leq m \leq i+k-1$,
  - for the left side of the jump, compute $\hat{\alpha}_{j,m}$ by extrapolation, compute $\hat{\beta}_{j,m}$ by (20), and then set
$$\beta_{j,m} = \hat{\beta}_{j,m}, s_i = 1;$$
  - for the right side of the jump, set $\bar{\beta}_{j,m} = 0$ and compute $\bar{\alpha}_{j,m}$ by (21), and set
$$\alpha_{j,m} = \bar{\alpha}_{j,m}.$$

(iii) Otherwise, compute $\alpha_{j,i}$ by (12). Set $s_i = 0$.

In the algorithm, ϵ is a predefined small positive number which is used to prevent the numerical instability caused by small $\beta_{j,i}$. More precisely, if both $\beta_{j,i}$ and $\beta_{j,i-1}$ are less than the given tolerance ϵ, we treat the current standard stencil as a smooth stencil.

In step (ii), it is possible to use any extrapolation techniques to handle the discontinuities.

Here, we just described the algorithm for one level ENO-wavelet transform with input data sequence $\alpha_{j+1,i}$, and output data $\alpha_{j,i}$ and $\beta_{j,i}$. The coefficient sequences $\alpha_{j,i}$ and $\beta_{j,i}$ have the same size, and their combined size is the same as the input data size at level $j+1$. The multiresolution transform algorithms can be constructed straightforwardly by recursively applying the one level transform to the low frequency coefficients $\alpha_{j,i}$. This is accomplished in the same way as that of the standard multiresolution algorithms. We do not explicitly include them in this paper. Similarly, we present the one level inverse transforms next.

Inverse Transform Algorithm.

For each i,

(i) if $s_i = 0$ and $s_j = 0$, $j = i-k, \ldots, i-1$, then the standard inverse wavelet transforms are applied:

$$\alpha_{j+1,2i} = \sum_{s=0}^{l} (r_{2s+1}\alpha_{j,i-s} + d_{2s+1}\beta_{j,i-s}) \tag{24}$$

and

$$\alpha_{j+1,2i+1} = \sum_{s=0}^{l} (r_{2s}\alpha_{j,i-s} + d_{2s}\beta_{j,i-s}). \tag{25}$$

(ii) If $s_j = 1$, $i-k \le j \le i$ or $i-k+1 \le j \le i$,
- use Proposition 1 to locate the position of the jump by counting the number of consecutive $s_i = 1$;
- extrapolate $\hat{\alpha}_{j,i}$ from the left side of the jump;
- set $\bar{\beta}_{j,i}$ as zero for the right side of the jump;
- use $\hat{\alpha}_{j,k}$ and $\beta_{j,k}$ to restore the left side by (24) and (25);
- use $\alpha_{j,k}$ and $\bar{\beta}_{j,k}$ to restore the right side of the jump by (24) and (25).

Two simple examples. We give two simple examples in the ENO-Haar and ENO-DB4 cases to illustrate the algorithms. First, we consider computing the transform coefficients of the following initial data:

$$\begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 \end{pmatrix}.$$

The standard Haar produces the low and high frequency coefficients

$$\alpha = \begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{3}{\sqrt{2}} & \frac{4}{\sqrt{2}} \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2}} & 0 \end{pmatrix}.$$

The corresponding linear approximation is

$$\begin{pmatrix} 1 & 1 & \frac{3}{2} & \frac{3}{2} & 2 & 2 \end{pmatrix},$$

which cannot recover the discontinuity correctly.

Using the ENO-Haar wavelet, we break the initial data sequence into two smooth pieces as shown in the following two rows:

$$\begin{pmatrix} & & y & 2 & 2 & 2 \\ 1 & 1 & 1 & x & & \end{pmatrix},$$

where x and y are some smooth extensions of the corresponding pieces. In fact, we extend x in a way such that the low frequency coefficient $\hat{\alpha}_2$ (boxed in (26)) based on the stencil $(1, x)$ is the same as the previous α_1, which is based on the stencil $(1, 1)$ giving $x = 1$. Similarly, we extend y in a way such that the high frequency coefficient $\bar{\beta}_2$ (boxed in (26)) is zero giving $y = 2$. Therefore we compute the high frequency coefficients $\hat{\beta}_2$ based on stencil $(1, x)$ and the low frequency coefficients $\bar{\alpha}_2$ based on stencil $(y, 2)$ by using the corresponding standard filters giving $\hat{\beta}_2 = 0$ and $\bar{\alpha}_2 = \frac{4}{\sqrt{2}}$. Thus we have the coefficients

$$\alpha = \begin{pmatrix} & \frac{4}{\sqrt{2}} & \frac{4}{\sqrt{2}} \\ \frac{2}{\sqrt{2}} & \boxed{\frac{2}{\sqrt{2}}} & \end{pmatrix}, \qquad \beta = \begin{pmatrix} & \boxed{0} & 0 \\ 0 & 0 & \end{pmatrix}. \tag{26}$$

Since we know how we extended $\hat{\alpha}_2$ and $\bar{\beta}_2$, we do not need to store them. In fact, we just need to store the low and high frequency coefficients as

$$\alpha = \begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{4}{\sqrt{2}} & \frac{4}{\sqrt{2}} \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix},$$

which have the same storage schemes as the standard Haar wavelet transform.

When we reconstruct the linear approximation, we can first recover $\hat{\alpha}_2$ and $\bar{\beta}_2$ the same way as in the forward transform and then apply the standard inverse filters to the smooth data to build the approximation. In fact, in this case the linear approximation is exactly the initial data.

In the next example, we show a similar example in which the ENO-DB4 linear approximation is not exactly the same as the initial data, but it still preserves the jump well. The initial data is given as

$$a = \begin{pmatrix} 0 & 1 & 2 & 3 & 4.1 & 5 & 20 & 21 & 22 & 23 \end{pmatrix}.$$

To better demonstrate the coarse level extrapolation idea, we ignore the boundary extension at two ends of the array. We leave out the coefficients based on the boundary extension at the two ends and display only the coefficients corresponding to the middle part of the array. The DB4 filters are given by the low pass filter $(0.4830 \quad 0.8365 \quad 0.2241 \quad -0.1294)$ and the high pass filter $(-0.1294 \quad -0.2241 \quad 0.8365 \quad -0.4830)$. The standard DB4 low and high frequency coefficients (α_2 to α_5 and β_2 to β_5) are

$$\alpha = \begin{pmatrix} 0.8966 & 3.7474 & 7.9280 & 29.1808 \end{pmatrix}$$

and

$$\beta = \begin{pmatrix} -0.0000 & 0.0837 & 4.9368 & -0.0000 \end{pmatrix}.$$

Notice that in this case we have a large high frequency coefficient β_3 which corresponds to the discontinuity between $a_6 = 5$ and $a_7 = 20$ in the array. If we discard the high frequency part, the corresponding linear approximation for the central part of the array around the jump (from $a_3 = 2$ to $a_8 = 21$) is

$$\begin{pmatrix} 2.0108 & 3.0187 & 4.6689 & 6.1470 & 15.8703 & 23.3843 \end{pmatrix},$$

and the discontinuity cannot be preserved.

Using the ENO-DB4 wavelet, we break the initial data sequence into two smooth pieces as shown in the following two rows:

$$\begin{pmatrix} & & & & u & v & 20 & 21 & 22 & 23 \\ 0 & 1 & 2 & 3 & 4.1 & 5 & x & y & & \end{pmatrix},$$

where (x, y) and (u, v) are some smooth extensions of the corresponding pieces. In fact, we extend (x, y) in such a way that the low frequency coefficient $\hat{\alpha}_3 = 6.5983$ (boxed in (27)) based on the stencil $(4.1, 5, x, y)$ is the linear extension of the previous α_1 and α_2. Similarly, we extend (u, v) in such a way that the high frequency coefficient $\bar{\beta}_3$ (boxed in (28)) is zero. Therefore we compute the high frequency coefficients $\hat{\beta}_3$ based on stencil $(4.1, 5, x, y)$ by (20) giving $\hat{\beta}_3 = -0.0259$ and the low frequency coefficients $\bar{\alpha}_3$ based on stencil $(u, v, 20, 21)$ by the analogy of (20) at the right side of a jump giving $\bar{\alpha}_3 = 26.3524$. Thus we have the coefficients

$$\alpha = \begin{pmatrix} & & 26.3524 & 29.1808 \\ 0.8966 & 3.7474 & \boxed{6.5983} & \end{pmatrix} \tag{27}$$

and

$$\beta = \begin{pmatrix} & & \boxed{0} & 0 \\ 0 & 0.0837 & -0.0259 & \end{pmatrix}. \tag{28}$$

Since we know how we extended $\hat{\alpha}_3$ and $\bar{\beta}_3$, we do not need to store them. In fact, we just need to store the low and high frequency coefficients as

$$\alpha = \begin{pmatrix} 0.8966 & 3.7474 & 26.3524 & 29.1808 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0 & 0.0837 & -0.0259 & 0 \end{pmatrix},$$

which have the same storage schemes as the standard DB4 wavelet transform.

The recovered linear approximation for a_3 to a_8 is

$$\begin{pmatrix} 2.0108 & 3.0187 & 4.0267 & 5.0346 & 20.0000 & 21.0000 \end{pmatrix}.$$

In this case, although the linear approximation is not the same as the initial data, it forms a much more accurate approximation than that of the standard DB4 transform. More importantly, this approximation preserves the discontinuity sharply in contrast to the standard DB4 wavelet which smears the discontinuity.

Remarks.

(i) The ENO-wavelet transforms are just simple modifications of the standard wavelet transforms near discontinuities. The computational complexity of the algorithms remains $O(n)$, and they are relatively easy to implement.

(ii) In the transform algorithms and the corresponding inverse algorithms, the extended low frequency coefficients $\hat{\alpha}_{j,m}$ and the high frequency coefficients $\bar{\beta}_{j,m}$ can be computed by other extrapolation schemes such as least square extrapolation. This may be more robust, especially for noisy data.

(iii) The adaptive ENO-wavelet idea can also be used for other kind of wavelets. They do not necessarily have to be orthogonal wavelets. For instance, one can apply it to the biorthonormal wavelets.

(iv) Like other wavelet transforms, 2-D or even higher-dimensional transforms can be formed by tensor products. In the numerical example section, we will give a 2-D example.

(v) The adaptive ENO-wavelet idea can be recursively used even if the projections do not satisfy the DSP. In such a case, of course we will not get the nice error bound (see section 4), but the approximation errors are comparable to that of the standard wavelet transforms. Also, it is easy to modify the algorithms such that the standard wavelet transforms are applied at the place where the DSP is invalid.

4. Approximation error. In this section, we consider the ENO-wavelet approximation error for piecewise continuous functions.

Given a function $f(x)$ in L^2, in standard wavelet theory [27], [14], [30], it can be linearly approximated by its projection $f_j(x)$ in V_j as in (7) and (8). This linear approximation has a standard error estimate which we state in the following theorem; see also [30].

THEOREM 1. *Suppose the wavelet $\psi(x)$ generated by scaling function $\phi(x)$ has p vanishing moments and $f_j(x)$ is the approximation of $f(x)$, which has bounded pth order derivative, in V_j with basis $\phi_{j,k}(x)$; then,*

$$\|f(x) - f_j(x)\| \le C(\Delta x)^p \|f^{(p)}(x)\|, \tag{29}$$

where $\Delta x = 2^{-j}$ and C is a constant which is independent of j.

This theorem holds for the L^2 norm in general. Moreover, if the scaling function and its wavelet have finite support, then it also holds for the L^∞ norm.

In this theorem, we can see that the approximation error is controlled by two factors. One is the pth power of the spatial step Δx; the other is the norm of the pth derivative of the function. This error bound does not hold if the function does not have the finite pth derivative. This implies that the approximation could be poor for irregular functions even if the spatial step Δx is small. For piecewise continuous functions, especially functions with large jumps, the approximation error cannot be controlled as smooth functions. In fact, in the standard approximation function $f_j(x)$,

oscillations are generated near the discontinuous points, and they will not disappear even if the spatial step size is reduced (the Gibbs phenomenon).

In contrast, in our ENO-wavelet transforms, since no approximation coefficients are computed using information from both sides of the discontinuities, we can obtain a similar error estimate without taking derivatives across the jumps. In the next theorem, we state the estimation and prove it in the rest of this section.

THEOREM 2. *Suppose the scaling function $\phi(x)$ and its $\psi(x)$ have finite support in $[0, l]$, $\psi(x)$ has p vanishing moments, $f(x)$ is a piecewise continuous function in $[a, b]$ with bounded pth derivatives in each piece of smooth regions, and $f_j(x)$ is its jth level ENO-wavelet projection obtained by using any one of the three extrapolation methods given in section* 2.4 *with the choice of a satisfying* (23). *If the projection $f_{j+1}(x)$ satisfies the DSP, then*

$$\|f(x) - f_j(x)\| \le C(\Delta x)^p \|f^{(p)}(x)\|_{(a,b)\setminus D}, \tag{30}$$

where $\Delta x = 2^{-j}$ and D is the set where $f(x)$ has jumps in the function value or up to the pth derivatives. The norm $\|\cdot\|$ can be either the L^2 or the L^∞ norm.

Proof. We prove the inequality (30) under the L^∞ norm, and the L^2 result can be obtained in a similar way.

According to section 3.2, with the choice of a, all jumps in set D will be detected by the algorithms described for the piecewise smooth data unless the intensity of the jump is less than $O(\Delta x^{(p-m)})$, where the jump is in the mth derivative. In the latter case, the error caused by missing the jump is of the order of $O(\Delta x^p)$, which can be absorbed by the right-hand side of (30).

The DSP allows us to separate the discontinuities and individually consider a small neighborhood around each jump. Therefore, to simplify the discussion without loss of generality, we consider a piecewise function $f(x)$ with one jump at the origin. In other words,

$$f(x) = \begin{cases} f_1(x), & a \le x < 0, \\ f_2(x), & 0 \le x \le b, \end{cases}$$

where $f_1(x) \in C^p[a, 0]$ and $f_2(x) \in C^p[0, b]$. Because both $\phi_{j,i}(x)$ and $\psi_{j,i}(x)$ have support $[i, (l+i)\Delta x]$, the small neighborhood affected by the ENO decision is $[-l\Delta x, l\Delta x]$. In fact, the ENO-wavelet coefficients depend only on one-sided information and therefore, by symmetry, we just need to prove (30) in $[-l\Delta x, 0]$.

Before we prove that (30) holds for the three types of extrapolation methods, namely direct function extrapolation and the two choices of coarse level extension ((1) and (2) in section 3.1), we give some notations which we will frequently use in the proof.

Denote by $g_1(x)$ the $(p-1)$th order polynomial which is the first p term of the Taylor expansion of $f_1(x)$ at the origin, i.e.,

$$f_1(x) = g_1(x) + \frac{f_1^{(p)}(\xi)}{p!} x^p, \tag{31}$$

where ξ is in interval $[-l\Delta x, 0]$. Also denote by $\alpha_{j,m}$ and $\beta_{j,m}$ the ENO-wavelet low and high frequency coefficients, respectively, and $\bar{\alpha}_{j,m}$ the low frequency coefficients of the polynomial $g_1(x)$, i.e.,

$$\bar{\alpha}_{j,m} = \int g_1(x)\phi_{j,m}(x)dx$$

and

$$g_{1,j}(x) = \sum_m \bar{\alpha}_{j,m}\phi_{j,m}(x).$$

As we mentioned in section 3.1, different techniques can be used for extrapolation. Here we select the extrapolation by Taylor expansion as our starting point throughout the proof because of its simplicity. For other types of extrapolation techniques, the proof can be directly generalized by taking into account the difference between that type of extrapolation and the Taylor expansion extrapolation. For instance, the classical approximation result shows us that the difference between the Lagrange extrapolation that we use in the numerical experiments in this paper and Taylor expansion extrapolation is of the order of $O(\Delta x^p)$, which will be absorbed into the right-hand side of the estimate (30).

Now we are ready to prove that (30) holds for the three types of extrapolation methods. We first prove (30) for direct function extrapolation.

Direct function extrapolation. The direct function extrapolation extends $f_1(x)$ to interval $[0, l\Delta x]$ by defining

$$f_d(x) = \begin{cases} f_1(x), & -l\Delta x \le x < 0, \\ g_1(x), & 0 \le x \le l\Delta x. \end{cases}$$

The corresponding ENO-wavelet low frequency coefficients $\alpha_{j,m}$ are computed by

$$\alpha_{j,m} = \int f_d(x)\phi_{j,m}(x)dx, \tag{32}$$

and the approximation function is defined as

$$f_{d,j}(x) = \sum_m \alpha_{j,m}\phi_{j,m}(x), \qquad x \in [-l\Delta x, 0]. \tag{33}$$

For any point $x_0 \in [-l\Delta x, 0]$, by using (31) and the fact that since $g_1(x)$ is a $(p-1)$th order polynomial, $g_{1,j}(x) = g_1(x)$, we have

$$\begin{aligned} |f_1(x_0) - f_{d,j}(x_0)| &\le |f_1(x_0) - g_1(x_0)| + |g_{1,j}(x_0) - f_{d,j}(x_0)| \\ &\le C(\Delta x)^p \|f_1^{(p)}\| + |g_{1,j}(x_0) - f_{d,j}(x_0)|. \end{aligned} \tag{34}$$

Let q be an integer in $[-l, 0]$ such that $x_0 \in [q\Delta x, (q+1)\Delta x)$; then the last term of (34) can be bounded by

$$\begin{aligned} |g_{1,j}(x_0) - f_{d,j}(x_0)| &= \left| \sum_m (\bar{\alpha}_{j,m} - \alpha_{j,m})\phi_{j,m}(x_0) \right| \\ &\le \sum_{q-l \le m \le q} |\bar{\alpha}_{j,m} - \alpha_{j,m}||\phi_{j,m}(x_0)| \\ &= \sum_{q-l \le m \le q} |\bar{\alpha}_{j,m} - \alpha_{j,m}||(\Delta x)^{-\frac{1}{2}}\phi(2^j x_0 - m)|. \end{aligned} \tag{35}$$

To prove (30), we now need to estimate $|\bar{\alpha}_{j,m} - \alpha_{j,m}|$. Since when $m \le -l$ the coefficients are computed in the standard manner, i.e., $\bar{\alpha}_{j,m} = \alpha_{j,m}$, we just need to

consider all m with $-l+1 \le m \le 0$. In fact, we have

$$\begin{aligned}|\bar{\alpha}_{j,m} - \alpha_{j,m}| &= \left| \int (f_d(x) - g_1(x))\phi_{j,m}(x)dx \right| \\ &\le \left| \int_{m\Delta x}^{0} (f_d(x) - g_1(x))\phi_{j,m}(x)dx \right| \\ &\quad + \left| \int_{0}^{(m+l)\Delta x} (f_d(x) - g_1(x))\phi_{j,m}(x)dx \right|.\end{aligned}$$

Because $f_d(x)$ is the same as $g_1(x)$ in $[0, (m+l)\Delta x]$, using (31), we have

$$\begin{aligned}|\bar{\alpha}_{j,m} - \alpha_{j,m}| &= \left| \int_{m\Delta x}^{0} (f_1(x) - g_1(x))\phi_{j,m}(x)dx \right| \\ &\le \left(\int_{m\Delta x}^{0} |f_1(x) - g_1(x)|^2 dx \right)^{\frac{1}{2}} \left(\int_{m\Delta x}^{0} |\phi_{j,m}(x)|^2 dx \right)^{\frac{1}{2}} \\ &\le C(\Delta x)^p \|f^{(p)}\| (\Delta x)^{\frac{1}{2}} \\ &\le C(\Delta x)^{p+\frac{1}{2}} \|f^{(p)}\|.\end{aligned}$$

Therefore, combining this with (35), we have

$$|g_{1,j}(x_0) - f_{d,j}(x_0)| \le C(\Delta x)^p \|f^{(p)}\|.$$

This and (34) complete the proof of (30) for the case of direct function extrapolation.

Coarse level extrapolation. As described in section 3.1, there are two ways of extrapolating coefficients on the coarse level. One way is to set the extended high frequencies to zero. The other way is to extrapolate the low frequency coefficients by a $(p-1)$th order polynomial in wavelet space. In the following part of the proof, we consider them separately.

We consider the high frequency zero extension first.

Similar to the direct function extrapolation, we also extend $f_1(x)$ to the interval $[0, l\Delta x]$ and denote it by

$$f_h(x) = \begin{cases} f_1(x), & x \in [-l\Delta x, 0], \\ g_h(x), & x \in (0, l\Delta x], \end{cases}$$

where $g_h(x)$ is implicitly defined such that it makes $f_h(x)$ satisfy

$$\int f_h(x)\psi_{j,m}(x)dx = 0, \qquad -l+1 \le m \le 0, \tag{36}$$

and

$$\int f_h(x)\phi_{j,m}(x)dx = \alpha_{j,m}, \qquad -l+1 \le m \le 0. \tag{37}$$

The difference between $f_d(x)$ and $f_h(x)$ is that in the direct function extrapolation $f_d(x)$ we know that $g_1(x)$ is the $(p-1)$th order polynomial, but in this case $g_h(x)$ is unknown.

Formally following the proof of (30) for the direct function extrapolation, (34) and (35) also hold for this case. Therefore, we need only to estimate $|\bar{\alpha}_{j,m} - \alpha_{j,m}|, -l+1 \le$

$m \le 0$. We consider $m = -l+1$ first. Unlike in the direct function extrapolation, where $|\bar{\alpha}_{j,-l+1} - \alpha_{j,-l+1}|$ can be computed directly by the Taylor expansion, here we cannot bound $|\bar{\alpha}_{j,-l+1} - \alpha_{j,-l+1}|$ in the same way. Instead, we use the following trick to obtain the estimate we need.

From the dilation equation (1) and the wavelet equation (2), we have the following relationships:

$$\phi_{j,m}(x) = \sum_{s=0}^{l} c_s \phi_{j+1,s+2m}(x) \tag{38}$$

and

$$\psi_{j,m}(x) = \sum_{s=0}^{l} h_s \phi_{j+1,s+2m}(x). \tag{39}$$

Using (38), $\bar{\alpha}_{j,-l+1}$ and $\alpha_{j,-l+1}$ can be computed by

$$\alpha_{j,-l+1} = \int f_h(x)\phi_{j,-l+1}(x)dx = \sum_{s=0}^{l} c_s \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} f_h(x)\phi_{j+1,s-2l+2}(x)dx$$

and

$$\bar{\alpha}_{j,-l+1} = \int g_1(x)\phi_{j,-l+1}(x)dx = \sum_{s=0}^{l} c_s \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} g_1(x)\phi_{j+1,s-2l+2}(x)dx.$$

Therefore, we have

$$\begin{aligned} |\bar{\alpha}_{j,-l+1} - \alpha_{j,-l+1}| \le & \left| \sum_{s=0}^{l-2} c_s \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} (g_1(x) - f_h(x))\phi_{j+1,s-2l+2}(x)dx \right| \\ & + \left| c_{l-1} \int_{\frac{\Delta x}{2}(-l+1)}^{\frac{\Delta x}{2}} (g_1(x) - f_h(x))\phi_{j+1,1-l}(x)dx \right. \\ & \left. + c_l \int_{\frac{\Delta x}{2}(-l+2)}^{\Delta x} (g_1(x) - f_h(x))\phi_{j+1,2-l}(x)dx \right|. \end{aligned} \tag{40}$$

We know that only the last two terms involve the value of $f_h(x)$ in $[0, \Delta x]$. The other terms use $f_h(x)$ in $[-l\Delta x, 0]$, which is $f_1(x)$. Then, by Taylor expansion and Schwartz inequality,

$$\left| \sum_{s=0}^{l-2} c_s \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} (g_1(x) - f_h(x))\phi_{j+1,s-2l+2}(x)dx \right| \le C(\Delta x)^p \|f_1^{(p)}\| 2^{\frac{-(j+1)}{2}}. \tag{41}$$

Thus, to bound $|\bar{\alpha}_{j,-l+1} - \alpha_{j,-l+1}|$, the only remaining task is to estimate the last two terms in (40).

Considering that $g_1(x)$ is a $(p-1)$th order polynomial, we obtain

$$\int f_h(x)\psi_{j,-l+1}(x)dx = 0 = \int g_1(x)\psi_{j,-l+1}(x)dx.$$

Substituting the wavelet equation (39) into the above equation, we have

$$\sum_{s=0}^{l} h_s \int (f_h(x) - g_1(x))\phi_{j+1,s-2l+2}(x)dx = 0.$$

We can rewrite this equation in the following form:

$$\begin{aligned} & h_{l-1} \int_{\frac{\Delta x}{2}(-l+1)}^{\frac{\Delta x}{2}} (f_h(x) - g_1(x))\phi_{j+1,1-l}(x)dx \\ & + h_l \int_{\frac{\Delta x}{2}(-l+2)}^{\Delta x} (f_h(x) - g_1(x))\phi_{j+1,2-l}(x)dx \\ & \qquad = -\sum_{s=0}^{l-2} h_s \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} (f_h(x) - g_1(x))\phi_{j+1,s-2l+2}(x)dx. \end{aligned}$$

Notice that we have $\frac{h_{l-1}}{c_{l-1}} = \frac{h_l}{c_l}$. We find that the left-hand side contains the term we need to estimate, whereas the right-hand side uses only $f_h(x)$ at the left side of the origin and thus can be controlled again by Taylor expansion. This means that we have

$$\begin{aligned} & \left| c_{l-1} \int_{\frac{\Delta x}{2}(-l+1)}^{\frac{\Delta x}{2}} (f_h(x) - g_1(x))\phi_{j+1,1-l}(x)dx \right. \\ & \left. + c_l \int_{\frac{\Delta x}{2}(-l+2)}^{\Delta x} (f_h(x) - g_1(x))\phi_{j+1,2-l}(x)dx \right| \\ & \qquad \le \left| \frac{c_l}{h_l} \right| \sum_{s=0}^{l-2} |h_s| \int_{\frac{\Delta x}{2}(s-2l+2)}^{\frac{\Delta x}{2}(s-l+2)} |f_1(x) - g_1(x)||\phi_{j+1,s-2l+2}(x)|dx \\ & \qquad \le C(\Delta x)^p \|f_1^{(p)}\| 2^{\frac{-(j+1)}{2}}. \end{aligned} \tag{42}$$

Combining (40), (41), and (42), we have

$$|\bar{\alpha}_{j,-l+1} - \alpha_{j,-l+1}| \le C(\Delta x)^p \|f_1^{(p)}\| 2^{-\frac{(j+1)}{2}}.$$

Similarly, we can prove that, for all $m, -l+1 < m \le 0$,

$$|\bar{\alpha}_{j,m} - \alpha_{j,m}| \le C(\Delta x)^p \|f_1^{(p)}\| 2^{-\frac{(j+1)}{2}}.$$

Substituting them into (35), we prove that (30) holds for the high frequency extension case.

The last case we need to consider is the coarse level extrapolation of low frequency coefficients. To prove (30), we use the result obtained for the direct function extrapolation.

We denote by $\alpha_{j,m}^d$ the low frequency coefficients for $f_d(x)$. The jth level low frequency extrapolation approximation $f_{l,j}(x)$ is defined as

$$f_{l,j}(x) = \sum_m \alpha_{j,m}\phi_{j,m}(x).$$

For any point $x_0 \in [q\Delta x, (q+1)\Delta x) \subset [-l\Delta x, 0]$, we have

$$|f_1(x_0) - f_{l,j}(x_0)| \le |f_1(x_0) - f_{d,j}(x_0)| + |f_{d,j}(x_0) - f_{l,j}(x_0)|. \tag{43}$$

Using (30) for the direct function extrapolation case, we know that

$$|f_1(x_0) - f_{d,j}(x_0)| \le C(\Delta x)^p \|f_1^{(p)}\|. \tag{44}$$

And the remaining term can be bounded by

$$|f_{d,j}(x_0) - f_{l,j}(x_0)| \le \sum_{q-l\le m\le q} |\alpha_{j,m}^d - \alpha_{j,m}| |2^{\frac{j}{2}} \phi(2^j x_0 - m)|. \tag{45}$$

Again, we need to estimate $|\alpha_{j,m}^d - \alpha_{j,m}|$.

Unlike the previous two cases where the low frequency coefficients $\alpha_{j,m}$ are computed by integration (32) or (37), in this case $\alpha_{j,m}$ are determined by the low frequency extrapolation on the coarse level in wavelet space. So, to estimate $|\alpha_{j,m}^d - \alpha_{j,m}|$, we need to consider them in wavelet space. We introduce the following operator notations first.

Define the continuous wavelet transform (WT) of any function $f(x)$ in space V_j by

$$WT(f)(s) = \int f(x)\phi_{j,s}(x)dx = 2^{\frac{j}{2}} \int f(x)\phi(2^j x - s)dx.$$

Also define the following Taylor extrapolation operator (EX) of $f(x)$:

$$EX(f)(x) = \begin{cases} f(x), & x \le 0, \\ g(x), & x > 0, \end{cases}$$

where $g(x)$ is the $(p-1)$th order Taylor polynomial of $f(x)$. Using these notations, we can represent the low frequency wavelet coefficients

$$\alpha_{j,m} = EX_w(WT(f_1))(m), \quad \text{for} \quad -l+1 \le m \le 0,$$

and

$$\alpha_{j,m}^d = WT(EX_f(f_1))(m), \quad \text{for} \quad -l+1 \le m \le 0,$$

where EX_w and EX_f represent the extrapolation operator EX in the wavelet and physical space, respectively.

Instead of estimating $|\alpha_{j,m}^d - \alpha_{j,m}|$ directly, we prove the following more general result.

LEMMA 1. *Given a smooth function $g(x)$, let $g_{we}(s) = WT(EX_f(g))(s)$ and $g_{ew}(s) = EX_w(WT(g))(s)$; then*

$$|g_{we}(s) - g_{ew}(s)| \le C(\Delta x)^p \|g^{(p)}\| 2^{-\frac{j}{2}}.$$

Using this lemma, we obtain the desired bounds for $|\alpha_{j,m}^d - \alpha_{j,m}|$ easily by taking $s = m$. Combining them with (44) and (45), we prove that (30) holds for the low frequency coefficient extrapolation case.

Proof. Denote $\bar{g}(x) = EX_f(g)(x)$, and then

$$\begin{aligned} g_{we}(s) &= 2^{\frac{j}{2}} \int \bar{g}(x)\phi(2^j x - s)dx \\ &= 2^{-\frac{j}{2}} \int \bar{g}(2^{-j}(y-s))\phi(y)dy. \end{aligned}$$

By changing variable $z = 2^{-j}s$, and denoting

$$e_j(z) = \int \bar{g}(2^{-j}y - z)\phi(y)dy,$$

we have

$$g_{we}(s) = 2^{-\frac{j}{2}} e_j(2^{-j}s).$$

We also know that $e_j(z)$ is a smooth function and, by differentiating p times, we have

$$\|e_j^{(p)}\| = \left\| \int (-1)^p \bar{g}^{(p)}(2^{-j}y - z)\phi(y)dy \right\| \le C\|g^{(p)}\| \left\| \int \phi(y)dy \right\| \le C\|g^{(p)}\|. \tag{46}$$

Taking the $(p-1)$th order Taylor expansion of $e_j(z)$ at $z = -l\Delta x$, we have

$$e_j(z) = \hat{e}_j(z) + e_j^{(p)}(\xi)\frac{(z + l\Delta x)^p}{p!},$$

where $\hat{e}_j(z)$ is the $(p-1)$th order Taylor polynomial and $\xi \in [2l, 0]$. Since $g_{ew}(s)$ is the same as $g_{we}(s)$ if $s \le -l$, it is defined as the Taylor polynomial for $s > -l$ according to the definition of EX; i.e., we have

$$g_{ew}(s) = \begin{cases} 2^{-\frac{j}{2}} e_j(2^{-j}s), & s \le -l, \\ 2^{-\frac{j}{2}} \hat{e}_j(2^{-j}s), & s > -l. \end{cases}$$

Therefore,

$$\begin{aligned} |g_{we}(s) - g_{ew}(s)| &\le 2^{-\frac{j}{2}} |e_j(2^{-j}s) - \hat{e}_j(2^{-j}s)| \\ &\le C(\Delta x)^p \|g^{(p)}\| 2^{-\frac{j}{2}}. \end{aligned}$$

This completes the proof of Lemma 1 and also completes the proof of Theorem 2.

5. Numerical examples. In this section, we give some one-dimensional (1-D) and 2-D numerical examples by using the ENO-wavelet transforms. In particular, we show results of the ENO-Haar, ENO-DB4, and ENO-DB6 wavelet transforms.

In all examples, for simplicity, we just consider functions with zero values at the boundary. For nonzero boundary functions, we can easily extend the function by zero and treat the boundaries as discontinuities.

To illustrate the performance of ENO-wavelet transforms, we show picture comparisons of the standard wavelet approximations and corresponding ENO-wavelet approximations. In addition, we compare the L_∞ and L_2 errors of the standard wavelet approximations and the ENO-wavelet approximations at different levels by measuring $E_{\infty,j} = \inf_x \|f(x) - f_j(x)\|$, which is computed by finding the largest difference on the finest grid, and $E_{2,j} = \|f(x) - f_j(x)\|_2$. Using them, we compute the orders of accuracy defined by

$$Order_\infty = \log_2 \frac{E_{\infty,i}}{E_{\infty,i-1}}$$

and

$$Order_2 = \log_2 \frac{E_{2,i}}{E_{2,i-1}},$$

TABLE 1

Comparison of the maximum error of the standard Haar and the ENO-Haar wavelet approximation for the smooth function $\sin(x)$. *We see that they have the same approximation error for the smooth functions.*

Level	Haar E_∞	ENO-Haar E_∞	$Order_\infty$
4	0.0919	0.0919	
3	0.0430	0.0430	1.070
2	0.0184	0.0184	1.202
1	0.0061	0.0061	1.585

TABLE 2

Comparison of the maximum error of the standard DB4 and the ENO-DB4 approximations for the smooth function $f(x) = \exp[-(\frac{1}{x} + \frac{1}{1-x})], 0 < x < 1$. *They have the same error and both achieve second order accuracy which agrees with the results in Theorem* 1 *for the smooth functions.*

Level	DB4 E_∞	ENO-DB4 E_∞	$Order_\infty$
4	3.316e-5	3.316e-5	
3	7.650e-6	7.650e-6	2.104
2	1.590e-6	1.590e-6	2.232
1	2.972e-7	2.973e-7	2.406

which indicates the order of accuracy of the approximation in the L_∞ norm and L_2 norm, respectively.

For all noise-free examples, we use the method described in section 3.2 to locate the exact positions of the discontinuities. And we select $a = 2$ and $\epsilon = 0.0001$ (as used in the algorithms in section 3.3) for all 1-D examples.

First, we compare the approximations for smooth functions. Table 1 is the comparison of Haar and ENO-Haar approximations for the smooth function $f(x) = \sin(x), 0 \le x \le 2\pi$ at different levels, and Table 2 is the comparison of DB4 and ENO-DB4 approximations for the function $f(x) = \exp[-(\frac{1}{x} + \frac{1}{1-x})], 0 < x < 1$.

We see from these tables that for smooth functions the ENO-wavelet transforms have exactly the same approximation error as the standard wavelet transforms. Both of them maintain the approximation order 1 and 2 for Haar and DB4, respectively, which agree with the results in Theorem 1. In fact, we notice that in this situation no singularity is detected, and the ENO-wavelet algorithms perform the standard transforms for completely smooth functions as we expected.

Next, we consider a piecewise smooth function defined by

$$f(x) = \begin{cases} 0, & 0 \le x < 0.2, \\ -50x - 5, & 0.2 \le x < 0.4, \\ 10\sin(4\pi x + 0.8\pi) - 1, & 0.4 \le x < 1.1, \\ 5e^{2x} - 100, & 1.1 \le x < 1.6, \\ 0, & 1.6 \le x \le 2. \end{cases}$$

We apply Haar and ENO-Haar, DB4 and ENO-DB4, and DB6 and ENO-DB6 transforms to this function and compare the approximation error. Figure 5 shows the comparison of the order of accuracy in the L_∞ and L_2 norm. It is clear that both L_∞ and L_2 order of accuracy for ENO-wavelet transforms are of the order 1, 2, and 3 for ENO-Haar, ENO-DB4, and ENO-DB6, respectively, and they agree with the results in Theorem 2. In contrast, standard wavelet transforms do not retain the corresponding order of accuracy for piecewise smooth functions.

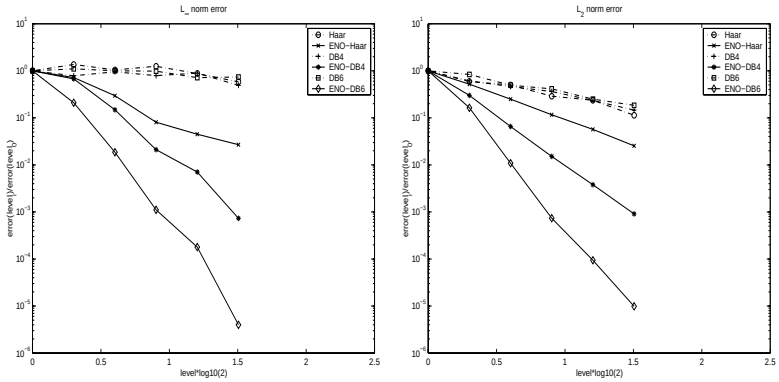


FIG. 5. *The approximation accuracy comparison of ENO-wavelet and wavelet transforms. Both L_∞ (left) and L_2 (right) order of accuracy show that ENO-wavelet transforms maintain the order* 1, 2, *and* 3 *for ENO-Haar, ENO-DB*4, *and ENO-DB*6, *respectively, and they agree with the results of Theorem* 2. *In contrast, standard wavelet transforms do not retain the order of accuracy for piecewise smooth functions.*

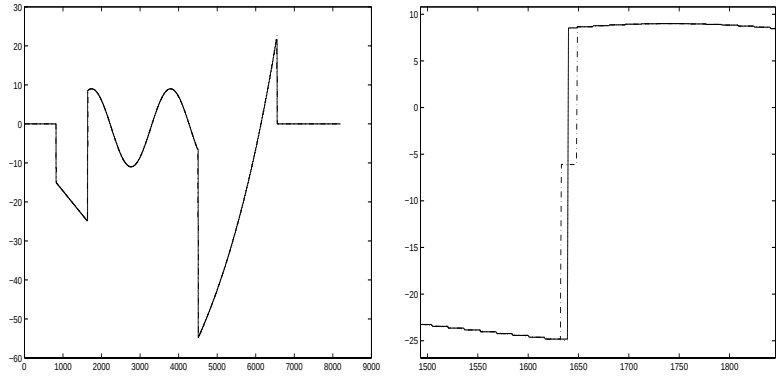


FIG. 6. *The* 4*-level ENO-Haar and Haar approximation. The left picture shows the original function (dotted line), the standard Haar approximation (dash-dotted line) and the ENO-Haar approximation (solid line). The right picture is a zoom-in near a discontinuity. We see the Gibbs phenomenon (staircase) in the standard Haar approximation but not in the ENO-Haar approximation.*

To see the Gibbs oscillations, we display the 4-level ENO-wavelet and standard wavelet approximations in Figures 6, 7, and 8 for ENO-Haar, ENO-DB4, and ENO-DB6 approximations, respectively. In the left column, we show the original function (dotted line), the standard wavelet linear approximations (dash-dotted), and the ENO-wavelet approximations (solid line). The right pictures are zoom-ins of the left pictures near a discontinuity. We clearly see the Gibbs oscillations in the standard approximations; in contrast, the ENO-wavelet approximations preserve the jump accurately.

In Figures 9, 10, and 11, we also present the standard Haar, DB4, and DB6 wavelet coefficients (dotted line) and the ENO-Haar, ENO-DB4, and ENO-DB6 wavelet coefficients (solid line), respectively. The left part corresponds to the low frequency coefficients and the right part to the high frequency coefficients. We notice that there are some large standard high frequency coefficients near the discontinuities. On the other hand, no large high frequency coefficients are present in the ENO-wavelet

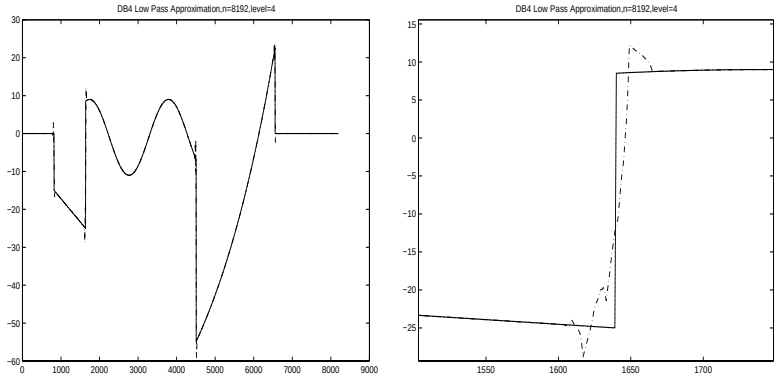


FIG. 7. *The 4-level ENO-DB4 and the standard DB4 approximations. The original discontinuous function (dotted line), the standard DB4 approximation (dash-dotted line), and the ENO-DB4 approximation (solid line) are displayed. The Gibbs phenomenon is clearly seen for the standard DB4 approximation but not for the ENO-DB4 approximation.*

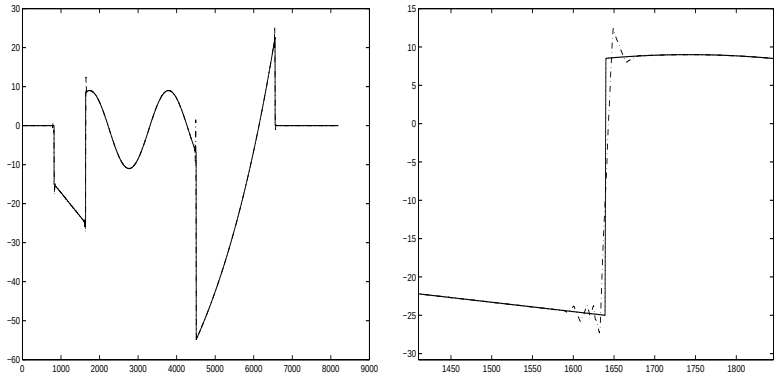


FIG. 8. *The 4-level ENO-DB6 (solid line) and the standard DB6 (dash-dotted line) approximation. The standard DB6 generates oscillations near discontinuities, but the ENO-DB6 does not.*

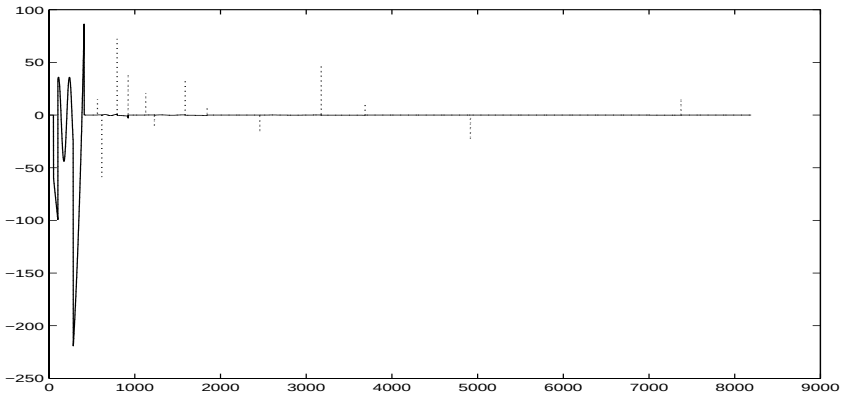


FIG. 9. *The 4-level ENO-Haar (solid line) and the standard Haar coefficients (dotted line). The left part corresponds to the low frequencies, the right part to the high frequencies. In the standard Haar coefficients, large high frequency coefficients present near discontinuities, while in the ENO-Haar case there are no large high frequency coefficients.*

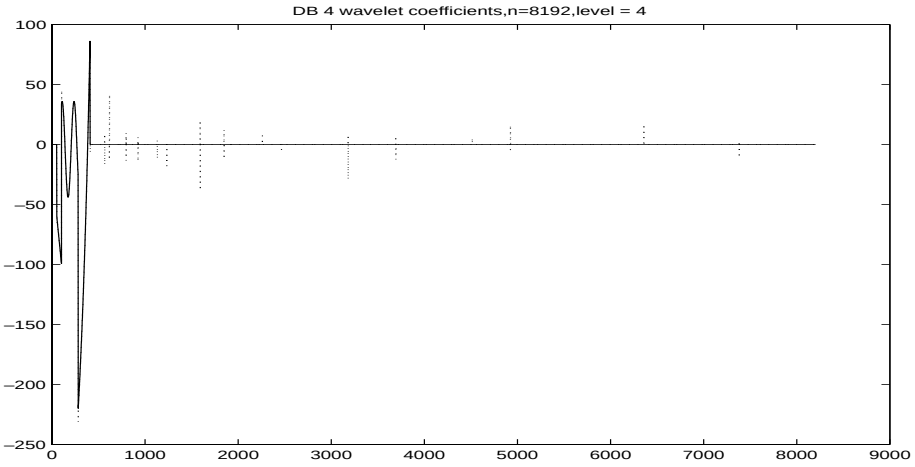


FIG. 10. *The* 4*-level ENO-DB*4 *coefficients (solid line) and the standard DB*4 *coefficients (dotted line). There are large high frequency coefficients (right part) near the discontinuities in the standard DB*4 *transform but not in the ENO-DB*4 *transform.*

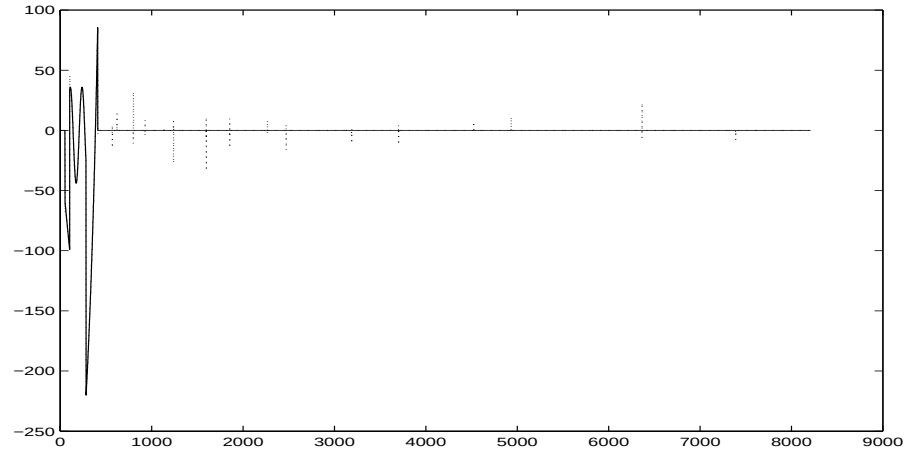


FIG. 11. *The* 4*-level ENO-DB*6 *coefficients (solid line) and the standard DB*6 *coefficients (dotted line). There are large high frequency coefficients near the discontinuities in the standard DB*6 *transform but not in the ENO-DB*6 *transform.*

coefficients. This illustrates that the ENO-wavelet coefficients have better distribution than standard wavelet coefficients; i.e., they have no large coefficients in the high frequencies, and the energy is concentrated in the low frequency end.

The next 1-D example we present here (Figure 12) is a comparison of the standard DB6 and the ENO-DB6 transforms to illustrate the performance at places where the DSP is not valid and also at jumps in the derivative. The original data (circles) has two discontinuities (the middle bump) which violate the DSP assumption, which requires that there are at least eight data points between any pair of discontinuities. Although the ENO-DB6 approximation (solid line) does not preserve this pair of discontinuities exactly, its approximation error is still comparable (actually better in this case) to that of the standard DB6 approximation (dotted line). At the left bump where the DSP holds, the ENO-DB6 does preserve the discontinuities exactly as we

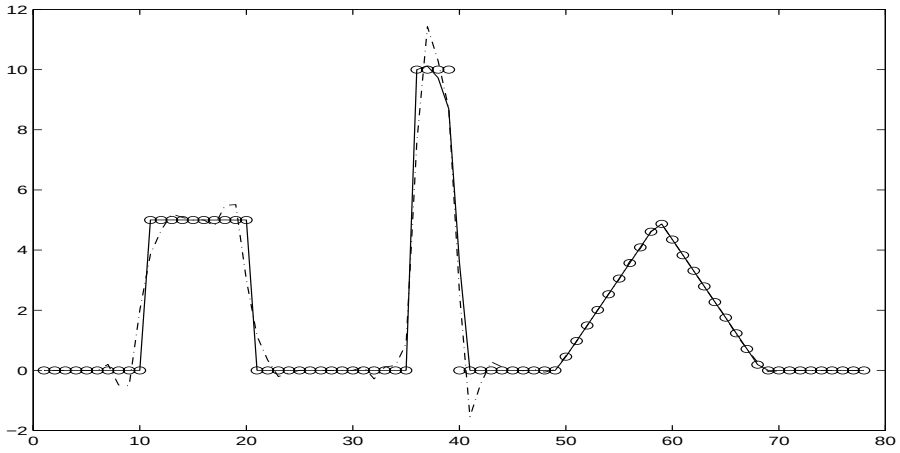


FIG. 12. *The level-*1 *approximation comparison of the ENO-DB*6 *and the standard DB*6 *wavelets at places where the DSP is invalid (the middle bump). The initial data (circles) has two close discontinuities. The ENO-DB*6 *approximation (solid line) error is comparable to that of the standard DB*6 *approximation (dotted line). The left bump satisfies the DSP and therefore the ENO-DB*6 *exactly recovers it. The right kink is a discontinuity in the first derivative, and the standard DB*6 *still generates oscillations although their magnitudes are not significant. The ENO-DB*6 *restores it perfectly, We display a zoom-in picture of this kink in Figure* 13.

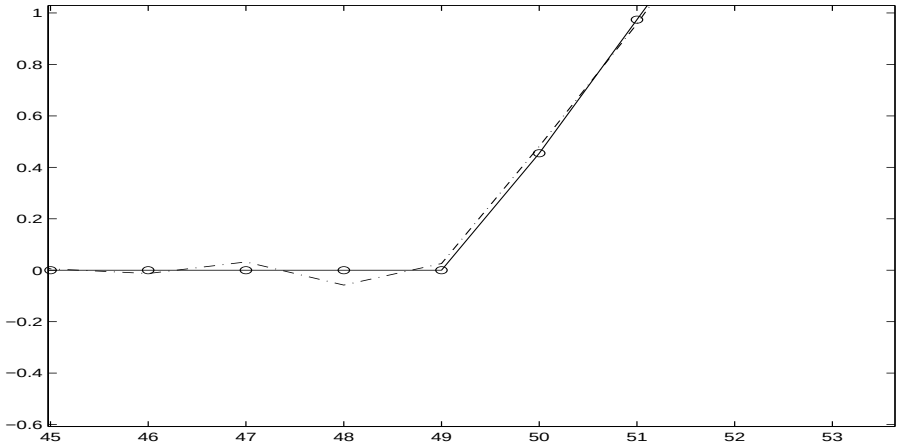


FIG. 13. *The zoom-in of Figure* 12 *at the kink where there is a discontinuity in its derivative. The ENO-DB*6 *(solid line) can recover it perfectly, but the standard DB*6 *(dash-dotted line) generates oscillations.*

expected. In the same example, we also display the comparison of the ENO-DB6 and the standard DB6 approximations at the right kink, which is not a discontinuity in function values but in its first order derivative. The standard DB6 approximation has oscillations, although their magnitudes are small, but the ENO-DB6 restores it exactly (see Figure 13).

The last 1-D example is applying the ENO-DB6 wavelet transform to a piecewise constant function polluted by Gaussian random noise (see Figure 14). For this example, the jump detection method corresponding to Lemma 1 does not work. Instead, we use the simple method given in section 3.2, which detects jumps by looking for

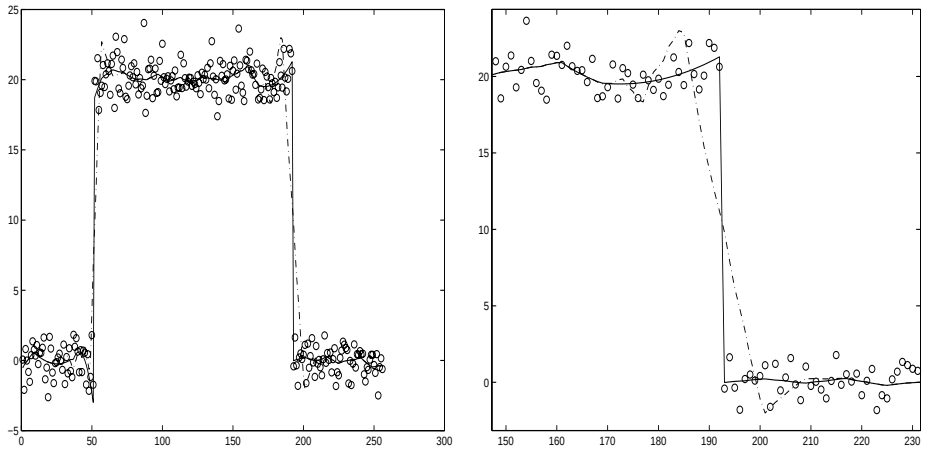


FIG. 14. *Left: The comparison of the* 3*-level ENO-DB*6 *approximation (solid line) with the standard DB*6 *approximation (dash-dotted line) for noisy initial data (circles). The ENO-DB*6 *approximation retains the sharp jumps, but the standard DB*6 *approximation does not (right picture). Right: A zoom-in of the left example at the discontinuities.*

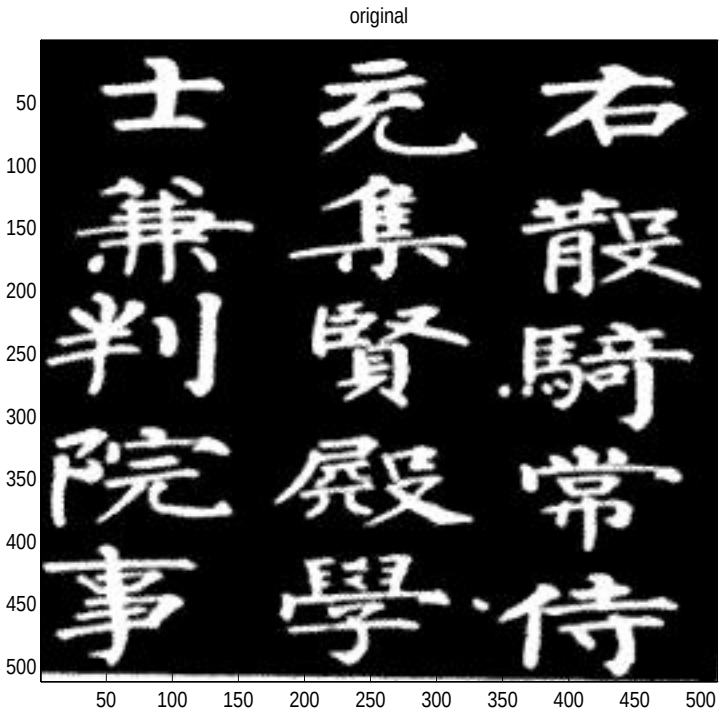


FIG. 15. *Original* 2*-D function.*

stencils with significant larger high frequency coefficients than their neighbors and then locates the exact jump locations by directly comparing the differences between two adjacent function values within the stencil. Despite the presence of noise in the initial data (circles), the level-3 ENO-DB6 approximation (solid line) still retains the sharp edges (see zoom-in in the right picture in Figure 14) compared to the stan-

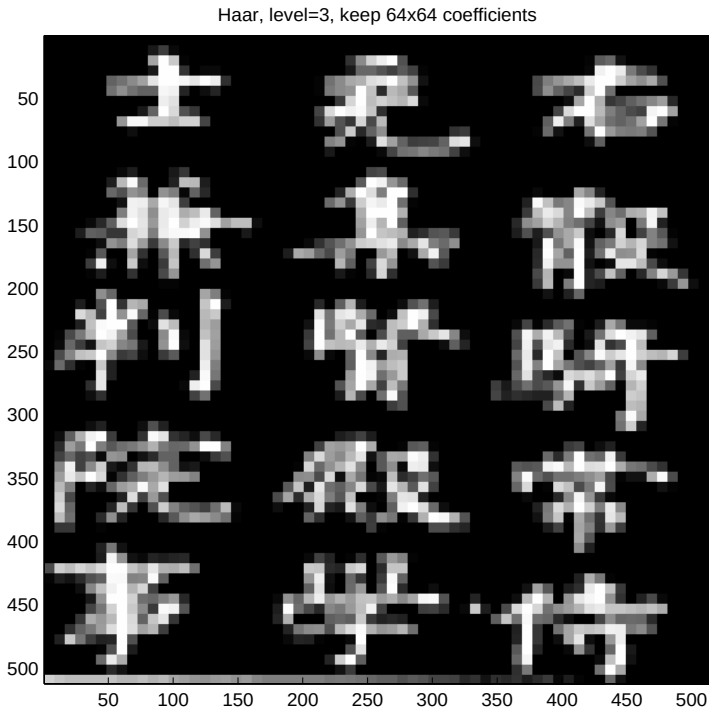


FIG. 16. *The 3-level standard Haar approximation: The reconstructions are obtained from low frequency coefficients α_{J-3} only, where α_J is the original image. The edges are fuzzier than those in the next picture.*

dard DB6 approximation (dash-dotted line) which not only has oscillations at the discontinuities but also smears them.

Finally, we give a 2-D testing image example to compare the standard Haar and the ENO-Haar approximations. Here we use tensor products of 1-D transforms. The original picture is shown in Figure 15. Figure 16 is the 3-level standard Haar approximation and Figure 17 is the 3-level ENO-Haar approximation. Both use low frequency approximations (the reconstructions are obtained from low frequency coefficients α_{J-3} only, where α_J is the original image) and store the same number of coefficients ($\frac{1}{64}$ of the original data). It is clear that in the standard Haar case the function becomes fuzzier than in the ENO-Haar case. This illustrates that the ENO-Haar approximation can reduce the edge oscillations for 2-D functions. In addition, as we mentioned in the introduction, we designed ENO-wavelet transforms not to replace the standard nonlinear adaptive wavelet techniques; rather we think it would be beneficial to use them in conjunction with the standard adaptive nonlinear techniques. For instance, we can combine ENO-wavelets with hard thresholding techniques as one can do it for the standard wavelet transforms. We show the standard hard thresholding approximation image by retaining the largest 64×64 coefficients in Figure 18, and we note that sharper edges are recovered comparing to the linear approximations. Similarly, we can apply the same thresholding techniques to the ENO-wavelet transforms. In Figure 19, we give the approximate image by using the ENO-Haar hard thresholding technique by keeping the largest 3506 ENO-Haar coefficients, which is 70% of number of coefficients retained in the previous image. In this image, edges are almost perfectly recovered.

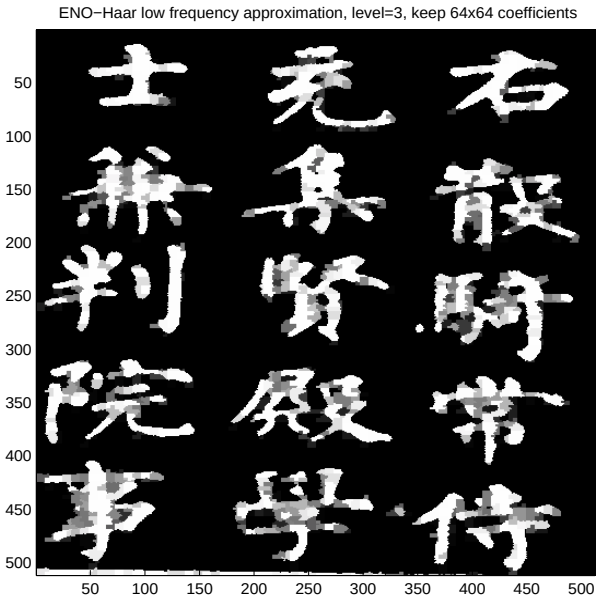


FIG. 17. *The* 3*-level ENO-Haar approximation: Similar to Figure* 16*, the reconstruction is obtained from low frequency coefficients* α_{J-3} *only. Both the edges and the interior of the characters are clearer than those in the standard Haar linear approximation.*

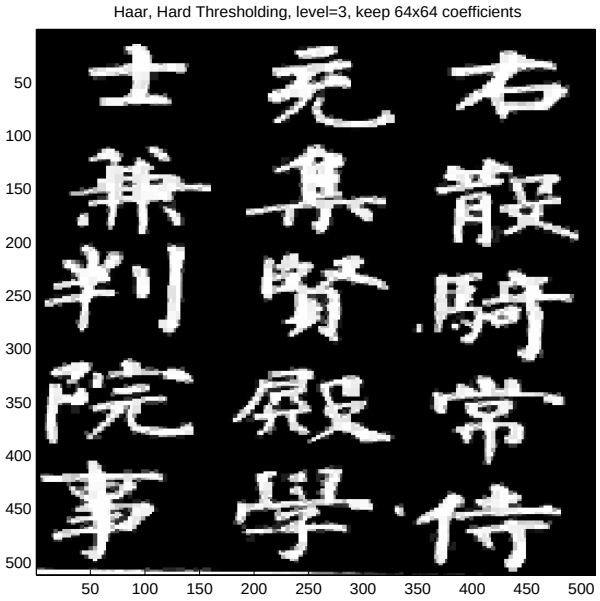


FIG. 18. *The* 3*-level standard Haar hard thresholding approximation: The image is reconstructed from the largest* 64×64 *wavelet coefficients (including* $\alpha_{J-3}, \beta_{J-3}, \beta_{J-2}, \beta_{J-1}$*). The edge artifacts are less severe than the standard linear approximation. On the other hard, the picture is comparable to the ENO-Haar low frequency approximation.*

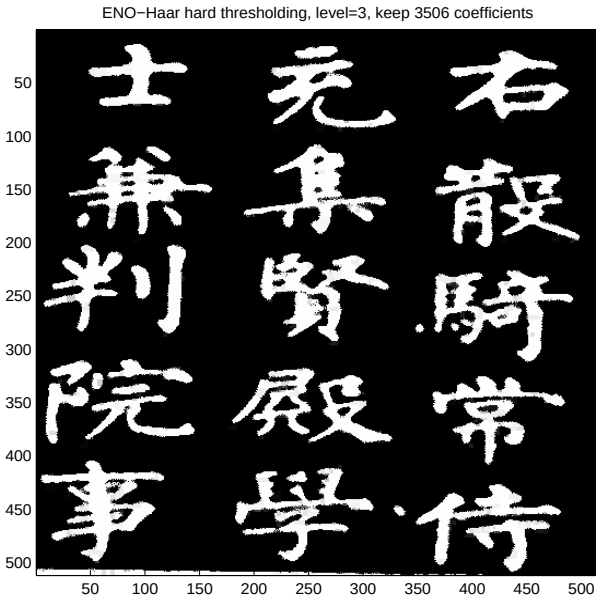


FIG. 19. *The 3-level ENO-Haar hard thresholding approximation: Similar to Figure 18, the image is reconstructed from the largest* 64×64 *ENO-wavelet coefficients (including* $\alpha_{J-3}, \beta_{J-3}, \beta_{J-2}, \beta_{J-1}$*). Less severe edge artifacts are generated compared to the previous images.*

REFERENCES

[1] A. ARNEODO, *Wavelet analysis of fractals: From the mathematical concepts to experimental reality*, in Wavelets: Theory and Applications, G. Erlebacher, M. Hussaini, and L. Jameson, eds., Oxford University Press, New York, 1996.

[2] S. AMAT, F. ARANDIGA, A. COHEN, R. DONAT, G. GARCIA, AND M. VON OEHSEN, *Data compression with ENO schemes: A case study*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 273–288.

[3] S. AMAT, F. ARANDIGA, A. COHEN, AND R. DONAT, *Tensor product multiresolution analysis with error control for compact image representations*, Signal Process., to appear.

[4] F. ARANDIGA AND R. DONAT, *Nonlinear multiscale decompositions: The approach of A. Harten*, Numer. Algorithms, 23 (2000), pp. 175–216.

[5] G. BEYLKIN, R. COIFMAN, AND V. ROKHLIN, *Fast wavelet transforms and numerical algorithms*, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[6] A. CHAMBOLLE, R. DEVORE, N. LEE, AND B. LUCIER, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.

[7] E. CANDES AND D. DONOHO, *Ridgelets: A Key to higher-dimensional intermittency?*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 2495–2509.

[8] T. F. CHAN AND H. M. ZHOU, *Adaptive ENO-wavelet Transforms for Discontinuous Functions*, CAM Report 99-21, UCLA, Los Angeles, CA, 1999.

[9] A. COHEN AND B. MATEI, *Compact representations of images by edge adapted multiscale transforms*, in Proceedings of the IEEE ICIP Conference, Tessaloniki, Greenland, 2001, to appear.

[10] R. COIFMAN AND D. DONOHO, *Translation invariant de-noising*, in Wavelets and Statistics, A. Antoniadis and G. Oppenheim, eds., Springer-Verlag, New York, 1995, pp. 125–150.

[11] C. K. CHUI, *Wavelet: A Mathematical Tool for Signal Analysis*, SIAM Monogr. Math. Model. Comput. 1, SIAM, Philadelphia, 1997.

[12] P. CLAYPOOLE, G. DAVIS, W. SWELDENS, AND R. BARANIUK, *Nonlinear Wavelet Transforms for Image Coding*, preprint, 1999, IEEE Trans. Image Process., submitted.

[13] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[14] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.
[15] D. Donoho, *De-noising by soft thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
[16] D. Donoho, *Wedgelets: Nearly-Minimax Estimation of Edges*, Technical report, Department of Statistics, Stanford University, Stanford, CA, 1997.
[17] D. Donoho, *Orthonormal Ridgelets and Linear Singularities*, Technical report, Department of Statistics, Stanford University, Stanford, CA, 1998.
[18] D. Donoho, I. Daubechies, R. DeVore, and M. Vetterli, *Data Compression and Harmonic Analysis*, preprint, Stanford University, Stanford, CA, 1998.
[19] D. Donoho and I. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.
[20] A. Harten, *Discrete multi-resolution analysis and generalized wavelet*, Appl. Numer. Math., 12 (1993), pp. 153–192.
[21] A. Harten, *Multiresolution Representation of Data,* II. *General Framework*, CAM Report 94-10, UCLA, Los Angeles, CA, 1994.
[22] A. Harten, *Multiresolution Representation of Cell-Averaged Data*, CAM Report 94-21, UCLA, Los Angeles, CA, 1994.
[23] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy, *Uniformly high order essentially non-oscillatory schemes,* III, J. Comput. Phys., 71 (1987), pp. 231–303.
[24] S. Jaffard, *Exposants de Hölder en des points donnés et coefficients d'ondelettes*, C. R. Acad. Sci. Paris Sér. I Math., 308 (1989), pp. 79–81.
[25] S. Mallat, *Multiresolution approximation and wavelet orthonormal bases of* $L^2(R)$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.
[26] S. Mallat, *A theory of multiresolution signal decomposition: The wavelet representation*, IEEE Trans. PAMI, 11 (1989), pp. 674–693.
[27] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.
[28] J. Shapiro, *Embedded image coding using zerotrees of wavelet coefficients*, IEEE Trans. Signal Process., 41 (1993), pp. 3445–3462.
[29] C.W. Shu, *High order ENO and WENO schemes for computational fluid dynamics*, in High-Order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, T. Barth and H. Deconinck, eds., Springer, Berlin, 1999, pp 439–582.
[30] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.
[31] W. Sweldens, *The lifting scheme: A construction of second generation wavelets*, SIAM J. Math. Anal., 29 (1998), pp. 511–546.
[32] P. Tchamitchian, *Wavelets, Functions, and Operators*, in Wavelets: Theory and Applications, G. Erlebacher, M. Hussaini, and L. Jameson, eds., Oxford University Press, New York, 1996.
[33] J. R. Williams and K. Amaratunga, *A Discrete Wavelet Transform without Edge Effects*, IESL Technical report 95-02, MIT, Cambridge, MA, 1995.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1405–1420

A NEW NUMERICAL METHOD FOR BACKWARD PARABOLIC PROBLEMS IN THE MAXIMUM-NORM SETTING*

J. M. MARBÁN† AND C. PALENCIA‡

Abstract. A new method for solving numerically backward parabolic problems is proposed. As usual for this kind of ill posed problems, it is assumed that an a priori bound for the solution is available. The algorithm consists of two basic steps. First, a standard forward integration is performed, in order to approximate the solution at suitable future time levels. Second, a holomorphic recovery procedure is carried out, providing the required approximations for the preceding times. The analysis is valid in the maximum-norm setting, and rigorous estimates are derived. Among other advantages, the method can also be applied to nonlinear problems, and it produces a continuous output. Some numerical illustrations are presented.

Key words. ill posed problems, backward parabolic problems, maximum-norm, holomorphic recovery, harmonic measure, Chebyshev nodes, least squares method

AMS subject classifications. 65J10, 65J20, 65M15, 65M20, 65M30

PII. S0036142901386422

1. Introduction.

The present paper is devoted to introducing and analyzing a new numerical method for backward parabolic problems. Our main estimate relies on an assumption (see (2.7) in section 2) that, for standard discretizations of classical parabolic problems, holds only in the maximum-norm setting. However, in order to outline the main difficulties associated with the kind of ill-posed problems we have in mind, let us start by adopting an abstract point of view. Thus, let X be a complex Banach space and let $A : D(A) \subset X \to X$ be the infinitesimal generator of a holomorphic semigroup $S(t)$, $t \geq 0$, of linear and bounded operators in X. We do not assume that A is densely defined so that $S(t)$ might fail to be continuous at $t = 0$ (see, e.g., [48]). In particular, this allows us to consider diffusion problems in the context of $X = L^\infty$. It is well known that the forward Cauchy problem

$$\begin{cases} w'(t) = Aw(t), & t \geq 0, \\ w(0) = w_0 \in X, \end{cases}$$

is well posed, being that its (generalized) solution is given by $w(t) = S(t)w_0$, $t \geq 0$. Moreover, there exists an angle $0 < \theta < \pi/2$ such that S admits a holomorphic extension, of exponential growth, to the sector

$$\Sigma_\theta = \{\, z \in \mathbb{C} \,:\, |\arg(z)| \leq \theta \,\}.$$

Without loss of generality, after performing an appropriate shift of A if necessary, we can assume that there exists $C_\theta > 0$ such that

$$\|S(z)\| \leq C_\theta, \qquad z \in \Sigma_\theta. \tag{1.1}$$

*Received by the editors March 14, 2001; accepted for publication (in revised form) March 6, 2002; published electronically October 23, 2002. This research was supported by project MCYT BFM2001-2013.

http://www.siam.org/journals/sinum/40-4/38642.html

†Departamento de Matemática Aplicada Fundamental, Universidad de Valladolid, 47005 Valladolid, Spain (josema@mac.cie.uva.es). This author was partly supported by grant JCYL VA025/01.

‡Departamento de Matemática Aplicada y Computación, Universidad de Valladolid, 47005 Valladolid, Spain (palencia@mac.cie.uva.es).

We are interested in the corresponding backward Cauchy problem (BCP)

$$\left\{ \begin{array}{l} w'(t) = Aw(t), \qquad 0 \le t \le T, \\ w(T) = w_T \in \mathcal{R}_T, \end{array} \right. \tag{1.2}$$

where $T > 0$ and w_T are given. Here $\mathcal{R}_T$ stands for the image of the operator $S(T)$. Problems of this nature arise in different contexts. Beyond their interest in connection with standard diffusion problems (then A is usually the laplacian operator Δ), they also appear, for instance, in some deconvolution problems, such as deblurring processes [10, 11, 12]. (Now $-A$ is often a fractional power of $-\Delta$.)

Fix $u_T \in \mathcal{R}_T$. Certainly, problem (1.2) with $w_T = u_T$ has got a unique solution [18] which hereon is denoted by $u : [0, +\infty) \to X$. However, in practice u_T is not available but rather some approximation $u_T + \delta\, u_T \in X$. This gives rise to two difficulties: (i) for unbounded A the available final datum $u_T + \delta\, u_T$ might not belong to $\mathcal{R}_T$, and (ii) the uncertainty due to $\delta\, u_T$ might propagate uncontrolled for $0 \le t < T$. Therefore, problems of this nature are typically ill posed. These difficulties can be partly overcome by incorporating some a priori information on the solution. In this paper we assume that a number $M_0 > 0$ is known in such a way that

$$\|u(t)\| \le M_0, \qquad 0 \le t \le T. \tag{1.3}$$

In the applications, the underlying physical problem usually provides reasonable values for M_0. However, what really matters for our method is an a priori bound

$$\|u(z)\| \le M, \qquad z \in \Sigma_\theta. \tag{1.4}$$

Notice that (1.1) and (1.3) give (1.4) with $M = C_\theta M_0$, though sharper estimates of M might be available in particular cases. This additional information renders problem (1.2) well posed, in the sense that the difference between two solutions satisfying (1.4) depends continuously on their difference at time T. Before stating this precisely, for the convenience of the reader, we recall the notion of harmonic measure [19, 21, 44]: Given a bounded, open domain $\Omega \subset \mathbb{C}$ whose boundary Γ consists of two disjoint, piecewise smooth arcs Γ_1 and Γ_2, the harmonic measure of Γ_1 with respect to Ω is the solution $\omega : \Omega \cup \Gamma \to \mathbb{R}$ of the Dirichlet problem

$$\left\{ \begin{array}{lll} \Delta\, \omega = 0 & \text{in} & \Omega, \\ \omega \ = 1 & \text{on} & \Gamma_1, \\ \omega \ = 0 & \text{on} & \Gamma_2. \end{array} \right.$$

Notice that, by virtue of the maximum principle, there holds $0 < \omega(z) < 1$ for $z \in \Omega$.

Now we are in a position to state the following theorem, whose proof, based on the *two-constants theorem* [44], is analogous to those of Theorem 6.2.1 in [18] and the main result in [36]. Recall that C_θ stands for a constant fulfilling (1.1).

THEOREM 1.1. *Let v_1, $v_2 : [0, T] \to X$ be two mappings satisfying the differential equation in* (1.2) *and assumption* (1.4) *for some $M > 0$. For $0 < \theta' \le \theta < \pi/2$, set*

$$\Omega = \{\, z \notin T + \Sigma_\theta \,:\, |\arg z| < \theta' \,\}$$

(see Figure 1*) and let Γ_1 be the part of the boundary of Ω lying in the sector $T + \Sigma_\theta$. Then*

$$\|v_2(t) - v_1(t)\| \le C_\theta^{\omega(t)} (2M)^{1-\omega(t)} \|v_2(T) - v_1(T)\|^{\omega(t)}, \qquad 0 \le t \le T, \tag{1.5}$$

where ω is the harmonic measure of Γ_1 with respect to Ω.

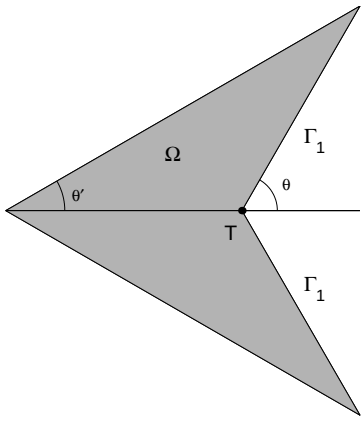


FIG. 1. *Domain* Ω *in Theorem* 1.1 *for* $\theta' < \theta$.

Thus, as we mentioned, problem (1.2) is well posed in a Hölder sense when restricted to solutions satisfying (1.4). Note that the Hölder exponent $\omega(t)$ degrades as t approaches $0+$ and that the sharper estimate of M yields the better bound in (1.5). It is also worth mentioning that the Hölder exponent $\omega(t)$, $0 < t < T$, depends on the ratio t/T rather than on t itself. Finally, notice that for a nonpositive self-adjoint generator A in a Hilbert space we can take the limits $\theta' \to \pi/2-$ and $\theta \to \pi/2-$ in (1.5). Then, since $C_{\pi/2}$ can be chosen equal to 1, we can take $M = M_0$, and (1.5) reads

$$\|v_2(t) - v_1(t)\| \le (2M_0)^{1-t/T} \|v_2(T) - v_1(T)\|^{t/T}, \qquad 0 \le t \le T,$$

a well-known result that can also be proved directly by using the logarithmic convexity of $\|v_2(t) - v_1(t)\|$ [18, 42].

A large variety of numerical methods have been proposed for (1.2). Some of them [13, 17, 30, 47] are based on the idea of quasi reversibility, while others use either some kind of regularization or filtering processes [1, 15, 16, 20, 46, 52]. Another interesting approach is presented in [8, 9], where (1.2) is transformed into a second order in time boundary value problem. In all these references X is assumed to be a Hilbert space and A a nonpositive self-adjoint operator. In [24] the backward heat equation, with constant coefficients, is dealt in L_p, $1 < p \le +\infty$, by considering an approach based on mollification and filtering. However, the numerical aspects are not developed, and this task does not seem to be straightforward. Pointwise estimates for this equation can also be found in [28]. More recently, a method relying on stochastic arguments, valid on Banach spaces, has been suggested in [7]. Nevertheless, the stochastic approach presents the drawback that the restriction on the step-size is too demanding.

Henceforth we restrict ourselves to classical parabolic problems in the maximum-norm setting. Thus, A is assumed to be a second order elliptic operator, $D(A)$ incorporates the boundary conditions, and X is taken to be the closure of $D(A)$ with respect to the maximum norm. It is known [49, 50] that A generates a holomorphic semigroup in X. Consideration of L^∞ spaces is also possible [48]. Our goal is to approximate the solution u of the problem

$$\begin{cases} u'(t) = Au(t), & 0 \le t \le T, \\ u(T) = u_T \in \mathcal{R}_T \end{cases} \tag{1.6}$$

from knowledge of a perturbed final datum $u_T + \delta\, u_T \in X$ with $\|\delta\, u_T\| \leq \delta$ and under assumption (1.4), where δ, $M > 0$ are given. Notice that $u_T + \delta\, u_T$ is not expected to belong to $\mathcal{R}_T$ so that, in general, this datum does not correspond to any solution of the equation in (1.2). Moreover, even in cases where $u_T + \delta\, u_T \in \mathcal{R}_T$, (1.4) is likely no longer satisfied by the solution v with final datum $u_T + \delta\, u_T$. Therefore, even though Theorem 1.1 provides the kind of estimates we can expect for the errors, in practice it cannot be applied.

Our algorithm is based on the realistic assumption that (1.6) can be integrated forward in time, i.e., for $t \geq T$, and it consists of two basic steps: (i) a standard forward integration in order to get fully discrete approximations $\hat{u}_h^n$ (belonging to discrete spaces X_h) to the values $u(t_n)$ of the solution at suitable future nodes $t_n \geq T$, $1 \leq n \leq N$, and (ii) the numerical holomorphic recovery of $u(t)$, $0 < t \leq T$, based on the previous approximations [33]. Let ε be the total uncertainty in the first step, i.e., the contributions of $\|\delta\, u_T\|$ and the error of the forward integration procedure. Our main result (see Theorem 3.1 below) essentially states that with a fairly moderate number of future nodes $N = O(|\ln \varepsilon|)$, the solution $u(t)$, $0 < t \leq T$, is approximated within an order $O\left((\varepsilon |\ln \varepsilon|)^{\omega^*(t)}\right)$, where ω^* is an appropriate harmonic measure. Among other advantages, let us point out that the method can also be applied to nonlinear problems and that it provides a continuous output $U_h(t)$ so that $u(t)$ can be readily approximated at any required time $0 < t \leq T$.

2. Notation and preliminaries. Recall that we are considering classical parabolic problems in the maximum norm, thus A is a second order elliptic operator and X is a Banach space formed by bounded mappings, endowed with the maximum norm. As we mentioned in the introduction, our numerical method is based on the realistic assumption that the forward parabolic problem

$$\left\{ \begin{array}{ll} v'(t) = Av(t), & t \geq 0, \\ v(0) = v_0 \in X & \end{array} \right. \tag{2.1}$$

can be fully discretized efficiently. Though the nature of this discretization is not essential to our algorithm, for the sake of convenience we adopt the standard approach of the method of lines. We introduce a well-known abstract setting (see, e.g., [27, 29, 40]) that covers either finite differences or finite elements, for the discretization in space, combined with a rational method for the discretization in time.

Let X_h, $0 < h \leq h_0$, be a family of finite dimensional normed spaces. All the norms used hereon, including operator norms, are denoted by $\|\cdot\|$. For $0 < h \leq h_0$, let $P_h : X \to X_h$ and $A_h : X_h \to X_h$ be linear operators, with P_h bounded. First, the solution of problem (2.1) is approximated by the solution $v_h : [0, +\infty) \to X_h$ of the semidiscrete problem

$$\left\{ \begin{array}{ll} v_h'(t) = A_h v_h(t), & t \geq 0, \\ v_h(0) = v_{h,0} := P_h v_0 \in X_h. & \end{array} \right. \tag{2.2}$$

Second, problem (2.2) is integrated in time by a rational method [6, 14, 23, 26, 37, 51] based on a rational approximation $r(z)$ to e^z of order $q \geq 1$. This approximation is assumed to be stable for variable step-sizes. This means the following:

(a) There exists $\bar{k} > 0$ (independent of h) such that the spectrum of kA_h, $0 < k \leq \bar{k}$, does not contain any pole of $r(z)$. Thus, for $0 < k \leq \bar{k}$, $r(kA_h)$ is a bounded operator in X_h, $0 < h \leq h_0$.

(b) For $\bar{t} > 0$ there exists C_s (independent of h) such that

$$\left\| \prod_{l=1}^{L} r(k_l A_h) \right\| \leq C_s, \tag{2.3}$$

for any finite sequence $\{k_l\}_{l=1}^{L}$ of step-sizes with $0 < k_l \leq \bar{k}$, $1 \leq l \leq L$, and with $\sum_{l=1}^{L} k_l \leq \bar{t}$.

It is known that the operators A_h arising when using either finite differences or finite elements are uniformly sectorial in h [3, 5]. Thus, by the abstract results in [4, 38], it turns out that (2.3) is satisfied in these important situations.

Given time levels $0 = \tau_0 < \tau_1 < \cdots < \tau_L = \bar{t}$, corresponding to step-sizes $k_{l+1} = \tau_{l+1} - \tau_l \leq \bar{k}$, the rational method applied to (2.2) provides the fully discrete approximations $v_h^l \simeq v_h(\tau_l)$ defined by

$$v_h^{l+1} = r(k_{l+1} A_h) v_h^l, \qquad 0 \leq l \leq L-1, \quad 0 < h \leq h_0.$$

The error is assumed to be of order $\varphi(h)$ in space, where $\varphi : (0, h_0] \to (0, +\infty)$ satisfies $\varphi(h) \to 0$ as $h \to 0+$, and of order q in time. To be precise, we assume that for given $\bar{t} > 0$ there exist $C_d > 0$ and $\nu > 0$ (independent of h) such that for a smooth enough initial datum $v_0 \in D((-A)^\nu)$ the error of the fully discrete scheme is governed by (see, e.g., [39, 51])

$$\|P_h v(\tau_l) - v_h^l\| \leq C_d \left(\|v_0\| + \|(-A)^\nu v_0\| \right) \cdot \left(\varphi(h) + k^q \right),$$

where $k = \max_{1 \leq l \leq L} k_l$.

Let us return to problem (1.6). For the analysis of our method it will be important to consider the forward problem

$$\begin{cases} u'(t) = Au(t), & t \geq T, \\ u(T) = u_T. \end{cases} \tag{2.4}$$

Since $u_T \in \mathcal{R}_T \subset D((-A)^\nu)$, given future nodes $\{T + \tau_l\}_{l=0}^{L}$, with $\{\tau_l\}_{l=0}^{L}$ as above, the fully discrete approximations to $u(T + \tau_l)$, denoted by u_h^l, satisfy

$$\|P_h u(T + \tau_l) - u_h^l\| \leq C_p(\varphi(h) + k^q), \qquad 1 \leq l \leq L, \quad 0 < h \leq h_0, \tag{2.5}$$

where $C_p = C_p(u_T)$.

For practical calculations, suitable bases $\{\chi_{h,j}\}_{j=1}^{J(h)}$ of X_h, $0 < h \leq h_0$, are required. Our hypothesis is that these bases are well conditioned in the following sense: there exist k_*, $k^* > 0$, independent of h, such that, for any $x_h \in X_h$,

$$x_h = \sum_{j=1}^{J(h)} x_{h,j} \chi_{h,j}, \tag{2.6}$$

there holds

$$k_* \cdot \max_{1 \leq j \leq J(h)} |x_{h,j}| \leq \|x_h\| \leq k^* \cdot \max_{1 \leq j \leq J(h)} |x_{h,j}|. \tag{2.7}$$

The coefficients in the expansion (2.6) will be denoted by

$$x_{h,j} =: \langle x_h, \chi_{h,j} \rangle, \qquad x_h \in X_h, \quad 1 \leq j \leq J(h).$$

To end this section we briefly describe the numerical algorithm in [33] for the recovery of holomorphic mappings.

Fix $0 < r < 1$ and set $I = [-r, r]$. For $N \geq 1$, let s_n, $1 \leq n \leq N$, be the Chebyshev nodes of first kind over I:

$$s_n = -r \cos\left(\frac{(2n-1)\pi}{2N}\right), \qquad 1 \leq n \leq N. \tag{2.8}$$

Let $D \subset \mathbb{C}$ be the unit disc and let $\tilde{f} : D \to \mathbb{C}$ be a holomorphic mapping. Set $\mathbf{w} = \{w_n\}_{n=1}^N := \{\tilde{f}(s_n)\}_{n=1}^N \in \mathbb{C}^N$ and assume that we are given perturbed nodal values $\mathbf{w} + \delta\,\mathbf{w} = \{w_n + \delta\, w_n\}_{n=1}^N \in \mathbb{C}^N$. The goal is to recover $\tilde{f}$ from knowledge of the approximate values $\mathbf{w}+\delta\,\mathbf{w}$ (see [33]). We also assume that $|\delta\, w_n| \leq \rho$, $1 \leq n \leq N$, and that $|\tilde{f}(s)| \leq H$, $s \in D$, where ρ and H are a priori known.

Let $\mathcal{S}_N$ be the linear space generated by the Cauchy kernels

$$K_n(s) = \frac{1}{1 - s_n s}, \qquad 1 \leq n \leq N.$$

The recovery of $\tilde{f}$ is provided by the least squares method (LSM) [33, 34] as $\tilde{F} \in \mathcal{S}_N$,

$$\tilde{F}(s) = \sum_{n=1}^{N} \gamma_n K_n(s),$$

by solving the constrained minimization problem

$$\begin{cases} \displaystyle\min_{G \in \mathcal{S}_N} \sum_{n=1}^{N} |G(s_n) - (w_n + \delta\, w_n)|^2 \\ \text{subject to} \\ \displaystyle\|G\|_2^2 := \sup_{0<\sigma<1} \frac{1}{2\pi} \int_0^{2\pi} |G(\sigma \mathrm{e}^{i\phi})|^2\, d\phi \leq H^2. \end{cases} \tag{2.9}$$

Set

$$\tau = \frac{r}{1 + \sqrt{1 - r^2}}, \qquad H^* = \frac{4H(1+r)}{1 - r^2}, \qquad \xi = 1 + \frac{\ln(\rho/H^*)}{\ln \tau}.$$

Corollary 2.1 in [33] shows that if $\rho < H^*$, then for $N = [\xi]$ and $s \in D$ we have

$$|\tilde{f}(s) - \tilde{F}(s)| \leq (2H\tilde{\gamma}(s))^{1-\tilde{\omega}(s)} \left(1 + \left(1 + \frac{2\ln \xi}{\pi}\right)\left(1 + \sqrt{\xi}\right)\right)^{\tilde{\omega}(s)} \rho^{\tilde{\omega}(s)}, \tag{2.10}$$

where

$$\tilde{\gamma}(s) = (1 + |s|)(1 - |s|)^{-1}(1 - \tilde{\omega}(s))^{-1} \tag{2.11}$$

and $\tilde{\omega} : \mathrm{cl}\,(D) \to [0, +\infty)$ is the harmonic measure of I with respect to $D \setminus I$ (see, e.g., [18, 44]), i.e., the continuous mapping in $\mathrm{cl}\,(D)$ that is harmonic in $D \setminus I$ and such that $\tilde{\omega}(s) = 1$ for $s \in I$, $\tilde{\omega}(s) = 0$ for $|s| = 1$.

In the present paper we rather use the simplified bound (easily derived from (2.10))

$$\left|\tilde{f}(s) - \tilde{F}(s)\right| \leq (2H\tilde{\gamma}(s))^{1-\tilde{\omega}(s)} (3N\rho)^{\tilde{\omega}(s)}, \qquad s \in D. \tag{2.12}$$

In practice, to solve the constrained minimization problem (2.9) an orthonormal basis $\{g_n\}_{n=1}^N$ of $\mathcal{S}_N$ is employed so that the matrix for the quadratic inequality constraint (see [22, section 12.1.1]) becomes the identity. We adopt the basis defined by (see [41])

$$g_1(s) = \frac{\sqrt{1-s_1^2}}{1-s_1 s}$$

and

$$g_n(s) = \frac{\sqrt{1-s_n^2}}{1-s_n s} \prod_{j=1}^{n-1} \frac{s-s_j}{1-s_j s}, \qquad 2 \le n \le N.$$

Working with this basis, problem (2.9) reduces to a matrix format which can be efficiently solved by means of the SVD (see [22, 33]). Finally, since

$$\max_{|s|\le 1} |g_n(s)| \le B_r := (1+r)^{1/2}(1-r)^{-1/2}, \qquad 1 \le n \le N,$$

Theorem 2.2 in [33] shows that we also have

$$|\tilde{f}(s) - \tilde{F}(s)| \le (H(1+B_r N^{1/2}))^{1-\tilde{\omega}(s)} (3N\rho)^{\tilde{\omega}(s)}, \qquad s \in D,$$

an estimate that can be advantageous when $|s| \to 1^-$.

3. The numerical algorithm for the BCP. For given integration parameters h and k (see section 2), set

$$\varepsilon = C_p\,(\varphi(h) + k^q) + C_s \|P_h\| \delta.$$

Fix R with $T < R < +\infty$ and set

$$\Sigma = \{\, z \in \Sigma_\theta \,:\, |z| \le R \,\}.$$

Let $\Psi : \Sigma \to \mathrm{cl}\,(D)$ be the conformal transformation (see Figure 2) with $\Psi(0) = -1$, $\Psi(R) = 1$, and $\Psi(T) = -r$ $(0 < r < 1)$, namely

$$\Psi(z) = \frac{a + \zeta - \zeta^{-1}}{a - \zeta + \zeta^{-1}},$$

where $\zeta = (z/R)^\sigma$, $a = b(1-r)/(1+r)$, $b = (T/R)^{-\sigma} - (T/R)^\sigma$, $\sigma = \pi/(2\theta)$.

Once M, r, and ε are fixed, set $\rho = \varepsilon/k_*$ and $H = M\|P_h\|/k_*$ and compute (see section 2)

$$N = [\xi], \qquad \xi = 1 + \frac{\ln(\rho/H^*)}{\ln \tau} = 1 + \frac{\ln(\varepsilon/(M^*\|P_h\|))}{\ln \tau}, \tag{3.1}$$

where $M^* = 4M(1+r)/(1-r^2)$. Let $-r < s_1 < \cdots < s_N < r$ be the corresponding Chebyshev nodes (2.8). Finally, set $T' = \Psi^{-1}(r)$ and $t_n = \Psi^{-1}(s_n)$, $1 \le n \le N$. Notice that the evaluation of $t = \Psi^{-1}(s)$, $-r \le s \le r$, reduces to solve the quadratic equation

$$(s+1)\zeta^2 + a(1-s)\zeta - (1+s) = 0, \qquad \zeta = (t/R)^\sigma.$$

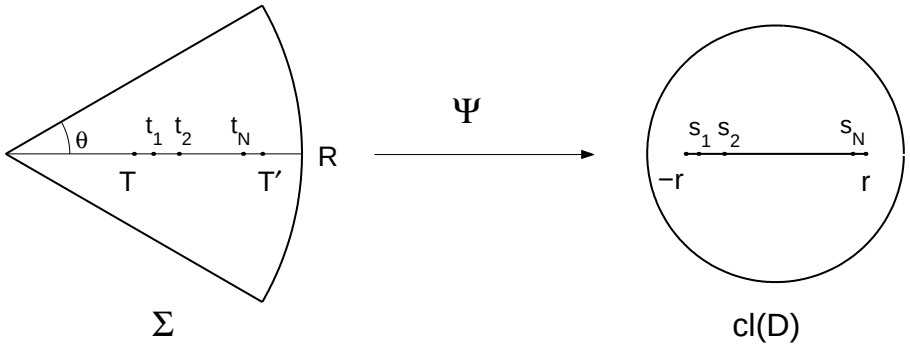


FIG. 2. *The conformal mapping* Ψ.

The numerical algorithm consists of two steps:

(1) Starting from the available approximate final datum $u_T + \delta\, u_T$, we integrate forward the problem

$$\begin{cases} \bar{u}'(t) = A\bar{u}(t), & T \le t \le T', \\ \bar{u}(T) = u_T + \delta\, u_T, \end{cases} \tag{3.2}$$

with the selected parameters h and k. The time levels τ_l, $1 \le l \le L$, must include the future nodes t_n, $1 \le n \le N$; i.e., $t_n = \tau_{l_n}$ for some $1 \le l_1 < l_2 < \cdots < l_N \le L$. For simplicity, the discrete values $\bar{u}_h^{l_n}$ are denoted by $\hat{u}_h^n$, $1 \le n \le N$. These vectors are stored.

(2) For each $1 \le j \le J(h)$, the holomorphic mapping

$$\tilde{f}_{h,j}(s) = \langle P_h u(\Psi^{-1}(s)), \chi_{h,j} \rangle, \qquad s \in D,$$

is recovered by the LSM algorithm from knowledge of the approximate values

$$\tilde{f}_{h,j}(s_n) \simeq \langle \hat{u}_h^n, \chi_{h,j} \rangle, \qquad 1 \le n \le N.$$

Let $\tilde{F}_{h,j} : D \to \mathbb{C}$ be the resulting recovery of $\tilde{f}_{h,j}$ provided by the LSM (2.9). Then the method adopts the approximations $U_h(t)$ to $P_h u(t)$, $0 \le t \le T$, defined by

$$P_h u(t) \simeq U_h(t) := \sum_{j=1}^{J(h)} F_{h,j}(t) \chi_{h,j},$$

where $F_{h,j} = \tilde{F}_{h,j} \circ \Psi$, $1 \le j \le J(h)$.

The following theorem provides a rigorous estimate for the error.

THEOREM 3.1. *Under the above conditions, for* $0 \le t \le T$ *there holds*

$$\|P_h u(t) - U_h(t)\| \le \left(\frac{k^*}{k_*}\right) (2\|P_h\| M\gamma(t))^{1-\omega(t)}\, (3N\varepsilon)^{\omega(t)}, \tag{3.3}$$

where $\gamma = \tilde{\gamma} \circ \Psi$ *(see* (2.11)*),* $\omega = \tilde{\omega} \circ \Psi$*, and* k_*, k^* *satisfy* (2.7).

Notice that ω is the harmonic measure of $[T, T']$ relative to $\Sigma \setminus [T, T']$. Thus, by the maximum principle, $\omega(t) < \omega_{\theta'}(t)$, $0 < t < T$, where now $\omega_{\theta'}$ stands for the harmonic measure used in Theorem 1.1. In fact, (3.3) with $\omega_{\theta'}$ instead of ω would

somehow be a quasi-optimal estimate. This suggests that a forward integration along the complex segments of Γ_1 in Figure 1 could provide such a quasi-optimal result, something left for future research. Notice also that the freedom in the choice of R and $T' = T'(R, r)$ results in different harmonic measures ω. Increasing R and T' provides better estimates (3.3) but requires a higher computational effort. On the other hand, T' very close to T requires less computation but leads to worse estimates (3.3). Let us also point out that the value $\omega(t)$ depends solely on the nondimensional quantities t/T, T'/T, and R/T (cf. Theorem 1.1). Finally, notice that, as for (1.5), estimate (3.3) becomes useless for times t with t/T close to 0.

Before proceeding with the proof of the theorem, several remarks are in order.

Remark 3.1. Notice that only the vectors $\hat{u}_h^n = \bar{u}_h^{l_n}$, corresponding to the future nodes t_n, $1 \le n \le N$, must be stored. Since typically $\|P_h\| = O(1)$ or $\|P_h\| = O(|\ln h|)$ (see, e.g., [45]), (3.1) shows that essentially $N = O(|\ln \varepsilon|)$. Thus, the algorithm requires a moderate amount of memory and accurately solving the required least square problems (see [33]) is rather cheap. Therefore, the efficiency of the new method relies on that of the numerical integration of the underlying forward problem.

Remark 3.2. Since constants C_p and C_s are unknown in most cases, it turns out that in general neither ε nor N could be determined. Set $\varepsilon_r = \varphi(h) + k^q + \delta$ (notice that this is the accessible part of the error) and $C_r = \max\{1, C_p, C_s\|P_h\|\}$ so that $\varepsilon \le C_r \cdot \varepsilon_r$. Then it is possible to prove that (3.3) remains valid by taking $N = [\xi_r]$, where ξ_r is the computable quantity

$$\xi_r := 1 + \frac{\ln(\varepsilon_r/(M^*\|P_h\|))}{\ln \tau}.$$

Remark 3.3. The algorithm provides the coefficients $F_{h,j}$ of the approximations

$$U_h(t) = \sum_{j=1}^{J(h)} F_{h,j}(t)\chi_{h,j}, \qquad 0 < t \le T,$$

where $F_{h,j}$ are continuous outputs; i.e., we are in a position to evaluate $U_h(t)$ at any $0 < t \le T$.

Remark 3.4. Notice that the different coefficients $F_{h,j}$ can be obtained in parallel from the future values $\hat{u}_h^n$. Moreover, when the basis $\chi_{h,j}$ is localized in space the recovery can be limited to indexes j affecting a given region of interest.

Remark 3.5. In some cases the aim could be the approximation of a certain functional of $u(t)$

$$\Lambda(u(t)), \qquad 0 < t \le T,$$

where $\Lambda \in X^*$, rather than $u(t)$ itself. Then, denoting by $\Lambda_h \in X_h^*$ a suitable approximation to Λ, the problem is reduced to compute

$$f_h(t) = \Lambda_h(P_h u(t)), \qquad 0 < t \le T.$$

Therefore, it is enough to recover the mapping f_h from knowledge of its nodal approximations

$$f_h(t_n) \simeq \Lambda_h(\hat{u}_h^n), \qquad 1 \le n \le N;$$

i.e., the recovery of all the individual coefficients $F_{h,j}$ of U_h is not required. Combining this idea with the previous remark, in practice we could first approximate u on a coarser grid and later refine up to the original grid where required.

Remark 3.6. Usually, an extension operator $E_h : X_h \to X$ such that $P_h E_h = I$ on X_h is available. In these cases we can adopt $E_h U_h(t)$ as the numerical aproximation to $u(t)$, rather than $U_h(t)$. We have

$$\|u(t) - E_h U_h(t)\| \leq \|(I - E_h P_h)u(t)\| + \|E_h\| \cdot \|P_h u(t) - U_h(t)\|.$$

The last term on the right-hand side has already been analyzed, while the first one depends on the approximation properties of X_h. Typically an estimate of the form (see, e.g., [39, 45, 51])

$$\|(I - E_h P_h)u(t)\| \leq C\varphi(h)\left(\|u(t)\| + \|(-A)^{\mu} u(t)\|\right)$$

for some $C, \mu > 0$ is known. Thus, when the above estimate holds, it turns out [43] that this term behaves like $O(\varphi(h)t^{-\mu})$ (or even like $O(\varphi(h))$ in cases where $u_0 \in D((-A)^{\mu})$). Hence, the bound (3.3) remains valid for $\|u(t) - E_h U_h(t)\|$.

Remark 3.7. Finally, it is worth mentioning that the linearity of problem (1.2) is not essential. In fact, as long as the solution $u(t)$ is holomorphic in a sector Σ, some a priori bound is available there, the errors in the forward integration behave like ε, and the full discretization is stable, our algorithm can be applied to backward, nonlinear evolution problems.

Usually, the infinitesimal generator of an abstract, parabolic nonlinear flux (see, e.g., [25, 31]) is the sum of a linear operator A and a dominated, nonlinear term $\mathcal{N}$. Then, the standard approach based on the variation-of-constants formula enables us to show that if $\mathcal{N}$ is holomorphic, then it is the nonlinear flux. This approach also yields estimates for the solution on sectors Σ where e^{zA} is holomorphic in terms of its initial datum. Thus, this kind of problems can be solved backward in time by means of the new method whenever a forward numerical integration can be performed.

Proof. Fix $1 \leq j \leq J(h)$. For $s \in D$ we have

$$|\tilde{f}_{h,j}(s)| = |\langle P_h u(\Psi^{-1}(s)), \chi_{h,j}\rangle| \leq k_*^{-1}\|P_h\| \cdot \|u(\Psi^{-1}(s))\| \leq k_*^{-1}\|P_h\| M.$$

Moreover, by (2.5), for $1 \leq n \leq N$

$$\|P_h u(t_n) - u_n^{l_n}\| \leq C_p\left(\varphi(h) + k^q\right)$$

and by (2.3)

$$\|u_h^{l_n} - \hat{u}_h^n\| = \|u_h^{l_n} - \bar{u}_h^{l_n}\| \leq C_s\|P_h\|\delta.$$

Therefore, by combining these inequalities, we get

$$\begin{aligned} |\tilde{f}_{h,j}(s_n) - \langle \hat{u}_h^n, \chi_{h,j}\rangle| &= |\langle P_h u(t_n) - \hat{u}_h^n, \chi_{h,j}\rangle| \\ &\leq k_*^{-1}\|P_h u(t_n) - \hat{u}_h^n\| \\ &\leq k_*^{-1} C_p\left(\varphi(h) + k^q\right) + k_*^{-1} C_s\|P_h\|\delta \\ &= k_*^{-1}\varepsilon. \end{aligned}$$

Now estimate (2.12) with $\rho = k_*^{-1}\varepsilon$ and $H = k_*^{-1}\|P_h\|M$ shows that for $s \in D$

$$|\tilde{f}_{h,j}(s) - \tilde{F}_{h,j}(s)| \leq k_*^{-1}\left(2\|P_h\|M\tilde{\gamma}(s)\right)^{1-\tilde{\omega}(s)}\left(3N\varepsilon\right)^{\tilde{\omega}(s)}.$$

Finally, for $t \in \Sigma$, setting $s = \Psi(t)$,

$$\|P_h u(t) - U_h(t)\| \leq k^* \sup_{1\leq j\leq J(h)} |\langle P_h u(t) - U_h(t), \chi_{h,j}\rangle| \leq k^* \sup_{1\leq j\leq J(h)} |\tilde{f}_{h,j}(s) - \tilde{F}_{h,j}(s)|,$$

and the theorem is proved since $\tilde{\omega}(s) = \omega(t)$. ∎

4. Numerical experiments. In this final section we present some numerical experiments in order to illustrate the new method and the theoretical results.

Experiment 1. Consider the backward problem

$$\left\{\begin{array}{ll} u_t(t,x) = u_{xx}(t,x), & 0 \le x \le 1, \quad t \ge 0, \\ u(t,0) = 0, & t \ge 0, \\ u(t,1) = 0, & t \ge 0, \\ u(T,x) = u_T(x) + \delta\, u_T(x) \in X = C_0[0,1], & \end{array}\right. \tag{4.1}$$

where $C_0[0,1]$ stands for the space formed by all the continuous mappings $f : [0,1] \to \mathbb{C}$ such that $f(0) = f(1) = 0$, endowed with the maximum norm.

We take $T = 1/8192$ and let $u_T \in C_0[0,1]$ be the value at time T of the exact solution of the forward heat equation with initial datum $u_0(x) = \sin(25\pi x^2)$ and homogeneous Dirichlet boundary conditions. Since no analytic expression for u_T is known we proceed as follows: First we integrate numerically the initial value problem, starting from u_0, in order to approximate u_T as well as $u(3T/4)$, $u(T/2)$, $u(T/4)$, $u(T/8)$ and $u(T/16)$. (These values are used for evaluating the errors shown in Table 1.) This is performed by combining central differences over a uniform mesh $x_j = jh$, $1 \le j \le J-1$, of size $h = 1/J$, for the discretization in space together with the well known MATLAB routine ODE23s, which is based on a modified Rosenbrock method [23], for the integration in time, in such a way that the total error is below 10^{-7}. Notice that in this situation $X_h = \mathbb{R}^{J-1}$, endowed with the maximum norm. The corresponding operator P_h brings each mapping f into the vector given by its grid values so that $k_* = k^* = \|P_h\| = 1$. Consideration of L^∞ spaces is also possible but requires the formalism introduced, for instance, in [2].

The final time T has been chosen so that the size of the oscillations in u_T close to the right endpoint is below 0.1, i.e., less than 10 percent of their initial size. Despite the fact that T is rather small, notice that it is the ratio t/T which is relevant to the experiment. Once we have accurately computed u_T, we introduce the perturbation $\delta\, u_T$ as a pseudorandom term, uniformly distributed on $[-\delta, \delta]$ with $\delta = 10^{-6}$ (see Figure 3) .

We take $\theta = \pi/2.2$, $R = 4.1T$, and $r = 0.3$. An easy calculation (based on the classical representation of the solution of the heat equation in terms of the Gaussian kernel) shows that $C_\theta = \sec^{1/2}\theta$ so that we can set $M = C_\theta \|u_0\|_\infty = \sec^{1/2}\theta$ in Theorem 3.1. Now we integrate forward (4.1) tuning h and the time step-sizes in such a way that $\varepsilon = 3\,\delta$. We are now in a position to calculate the number of nodes given by (3.1), which turns out to be $N = [9.75] = 9$. The resulting approximations at the required future nodes t_n, $1 \le n \le 9$, are displayed in Figure 4. From these approximations we construct the continuous outputs $U_h(t) \in \mathbb{R}^{J-1}$, $0 < t \le T$.

Table 1 shows the errors in the maximum norm at the past times $3T/4$, $T/2$, $T/4$, $T/8$, and $T/16$.

It is noteworthy how the solution is fairly well reproduced despite its severely oscillatory behavior and the loss of information at time T for points x close to 1. Note that even at $T/16$, i.e., about 95 percent of the way back, the relative error is only 7.8 percent. In fact, $U_h(t)$ and $P_h u(t)$, $t = T/2$, $T/4$, $T/8$, $T/16$, are hard to distinguish when plotted together (see Figure 5). A zoom corresponding to $t = T/16$ and the zone where the worst errors occur is displayed in Figure 6.

Finally, in order to compare errors in Table 1 with those predicted by Theorem 3.1, the corresponding harmonic measure ω must be numerically computed (for instance, by means of the D03EAF NAG FORTRAN library routine). It turns out that in this example the computed errors in Table 1 are smaller by a factor of 1/30 than those

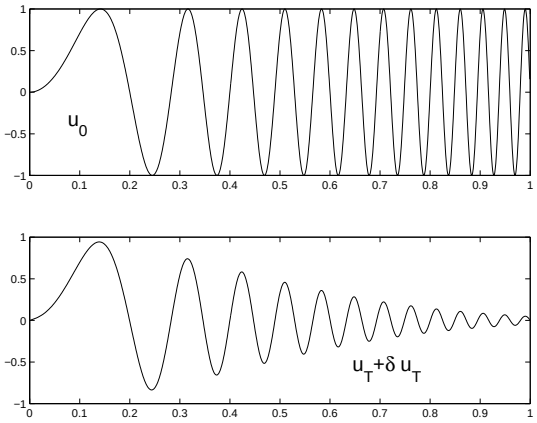


FIG. 3. *Initial datum u_0 and perturbed u_T.*

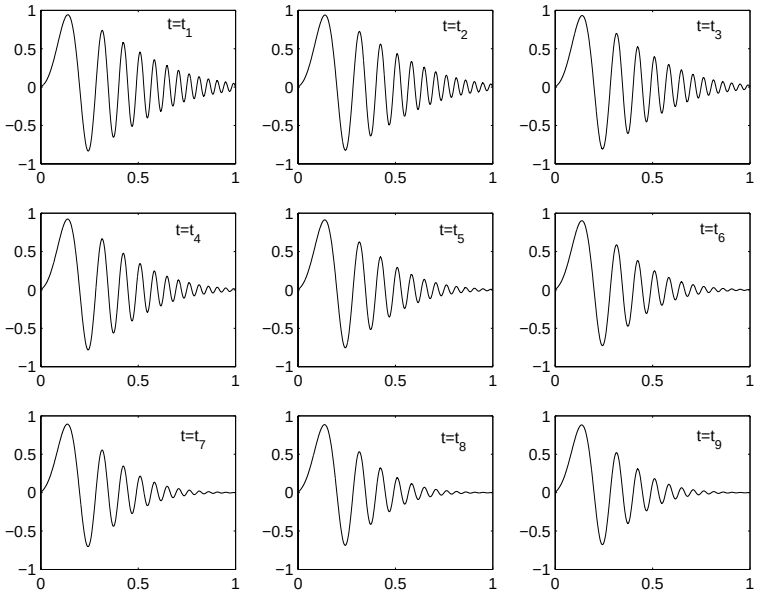


FIG. 4. *Numerical solution at future nodes.*

TABLE 1

Evaluation time	Absolute error
$t = T/16$	7.765457e-02
$t = T/8$	4.437583e-02
$t = T/4$	1.403523e-02
$t = T/2$	1.093864e-03
$t = 3T/4$	6.247916e-05

predicted by Theorem 3.1. (This factor is similar to the one found in the experiments in [33] when testing the LSM (2.9).)

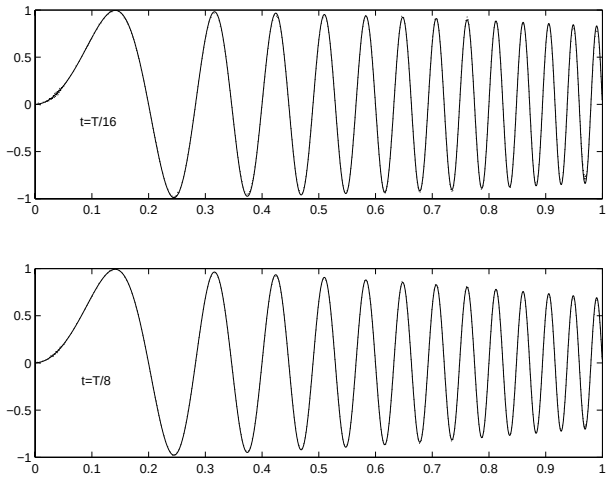


FIG. 5. *Numerical solution (dotted line) and exact solution (solid line) are hard to distinguish.*

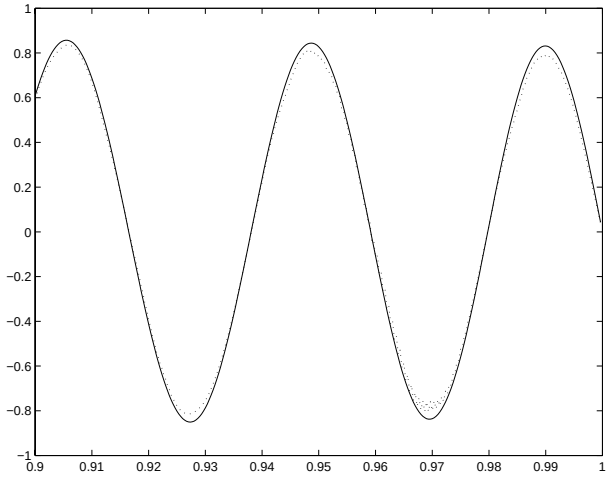


FIG. 6. *Zoom of the numerical solution (dotted line) and exact solution (solid line) at* $T/16$.

Experiment 2. Let us now consider the following nonlinear problem:

$$\begin{cases} u_t(t,x) = u_{xx}(t,x) + u(t,x)\,u_x(t,x) - \frac{1}{2}e^{-2t}\sin(2x), & 0 \le x \le \pi, \quad t \ge 0, \\ u(t,0) = 0 \qquad t \ge 0, \\ u(t,\pi) = 0 \qquad t \ge 0, \\ u(T,x) = u_T(x) + \delta\, u_T(x) \in X = C_0[0,\pi]. \end{cases}$$

The final time is taken to be $T = 1$ and $u_T(x) = e^{-1}\sin(x)$, $0 \le x \le \pi$. The source term is chosen in order for $u(t,x) = e^{-t}\sin(x)$ to be the solution, in the absence of perturbation. For the computations we take again $\delta\, u_T$ as a pseudorandom term of amplitude $\delta = 10^{-6}$. We maintain the values of r and θ. In this example, in the view of the solution, we take $M = 1$. The space and time discretizations are as in the

TABLE 2

Evaluation past times	Error for $M = 1$ ($N = 8$)	Error for $M = 7$ ($N = 9$)
$t = T/16$	1.324940e-02	6.238772e-02
$t = T/8$	8.925938e-03	3.698770e-02
$t = T/4$	3.251045e-03	1.220574e-02
$t = T/2$	2.424795e-04	9.937696e-04
$t = 3T/4$	7.889683e-06	3.986467e-05

TABLE 3

Evaluation past times	Absolute error
$t = T/16$	9.889366e-02
$t = T/8$	6.833551e-02
$t = T/4$	3.432584e-02
$t = T/2$	8.022480e-03
$t = 3T/4$	1.184737e-03

previous example. Now (3.1) provides $N = [8.71] = 8$. The errors in maximum norm are shown in Table 2. We repeat the experiment but now with $M = 7$. Then it turns out that $N = 9$. The corresponding errors are also shown in Table 2. This illustrates how a sharper a priori information yields better results.

Experiment 3. Finally we consider the two dimensional backward heat equation in the square $\Omega = [1, 1] \times [-1, 1]$:

$$\begin{cases} u_t(t,x,y) = (1/\pi)\Delta u(t,x,y), & (x,y) \in \Omega, \quad t \geq 0, \\ u(t,x,y) = 0, & (x,y) \in \partial\Omega, \quad t \geq 0, \\ u(T,x,y) = u_T(x,y) + \delta\, u_T(x,y) \in X = L_\infty([-1,1] \times [-1,1]), \end{cases}$$

with $T = 1/4$ and $u_T(x, y) = e^{-2\pi T} \sin(\pi x) \sin(\pi y)$. The solution for the unperturbed problem is clearly $u(t, x, y) = e^{-2\pi t} \sin(\pi x) \sin(\pi y)$. The term $\delta\, u_T$ is once more randomly generated but now with amplitude $\delta = 10^{-4}$. The discretization in space is performed by linear finite elements on a quasi-regular mesh of diameter $h = 10^{-2}$, while the time integration is carried out by the MATLAB ODE23S routine with $k = 10^{-2}$. Since $\varphi(h) = h^2 |\ln h|$ [45, 51], it turns out that $\varepsilon_r = 6.60517 \cdot 10^{-4}$. Moreover, M is assigned its exact value; i.e., $M = 1$. Now Remark 3.2 yields $N = [5.83] = 5$, and the errors at indicated times are collected in Table 3. Less accurate estimates of M lead to higher values of N and bigger errors.

Acknowledgment. The authors wish to express their gratitude to the anonymous referees for their helpful and valuable comments and suggestions.

REFERENCES

[1] K. A. AMES AND J. F. EPPERSON, *A kernel-based method for the approximate solution of backward parabolic problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1357–1390.

[2] R. ANSORGE, Differenzenapproximationen partieller Anfangswertaufgaben, Teubner, Stuttgart, 1978.

[3] N. YU. BAKAEV, *Resolvent estimates for a multidimensional elliptic difference operator*, Funct.-Diff. Eqs., Perm', PPI (1991), pp. 118–126 (in Russian).

[4] N. YU. BAKAEV, *On variable stepsize Runge–Kutta approximations of a Cauchy problem for the evolution equation*, BIT, 38 (1998), pp. 462–485.

[5] N. Yu. Bakaev, V. Thomée, and L. B. Wahlbin, *Maximum-Norm Estimates for Resolvents of Elliptic Finite Elements Operators*, preprint, Department of Mathematics, Chalmers University of Technology, Göteborg University, Göteborg, Sweden, 2001.
[6] P. Brenner and V. Thomée, *On rational approximations of semigroups*, SIAM J. Numer. Anal., 16 (1979), pp. 683–694.
[7] K. Burrage and S. Piskarev, *Stochastic methods for ill-posed problems*, BIT, 40 (2000), pp. 226–240.
[8] B. L. Buzbee and A. Carasso, *On the numerical computation of parabolic problems for preceding times*, Math. Comp., 27 (1973), pp. 237–265.
[9] A. Carasso, *The backward beam equation: Two A-stable schemes for parabolic problems*, SIAM J. Numer. Anal., 9 (1972), pp. 406–434.
[10] A. S. Carasso, *Overcoming Hölder continuity in ill-posed continuation problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1535–1557.
[11] A. S. Carasso, *Error bounds in nonsmooth image deblurring*, SIAM J. Math. Anal., 28 (1997), pp. 656–668.
[12] A. S. Carasso, J. G. Sanderson, and J. M. Hyman, *Digital removal of random media image degradations by solving the diffusion equation backwards in time*, SIAM J. Numer. Anal., 15 (1978), pp. 344–367.
[13] D. Colton and J. Wimp, *The construction of solutions to heat equation backward in time*, Math. Meth. Appl. Sci., 1 (1979), pp. 32–39.
[14] M. Crouzeix, S. Larsson, S. Piskarev, and V. Thomée, *The stability of rational approximations of analytic semigroups*, BIT, 33 (1993), pp. 74–84.
[15] L. Elden, *Time discretization in the backward solution of parabolic equations,* I, Math. Comp., 39 (1982), pp. 53–68.
[16] L. Elden, *Time discretization in the backward solution of parabolic equations,* II, Math. Comp., 39 (1982), pp. 69–84.
[17] R. E. Ewing, *The approximation of certain parabolic equations backward in time by Sobolev equations*, SIAM J. Math. Anal., 6 (1975), pp. 283–294.
[18] H. O. Fattorini, *The Cauchy Problem, Encyclopedia of Mathematics and Its Applications,* Addison-Wesley, Reading, MA, 1983.
[19] S. D. Fisher, *Function Theory in Planar Domains,* John Wiley, New York, 1983.
[20] J. N. Franklin, *On Tikhonov's method for ill-posed problems*, Math. Comp., 18 (1974), pp. 889–907.
[21] D. Gaier, *Lectures on Complex Approximation,* Birkhäuser, Boston, 1987.
[22] G. Golub and C. F. Van Loan, *Matrix Computations,* 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.
[23] E. Hairer and G. Wanner, *Solving Ordinary Equations* II. *Stiff and Differential-Algebraic Problems,* 2nd ed., Springer-Verlag, Berlin, 1996.
[24] D. N. Háo, *A mollification method for ill-posed problems*, Numer. Math., 68 (1994), pp. 469–506.
[25] D. Henry, *Geometric Theory of Semilinear Parabolic Equations,* Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1998.
[26] R. Hersh and T. Kato, *High-accuracy stable difference schemes for well-posed initial-value problems*, SIAM J. Numer. Anal., 16 (1979), pp. 670–682.
[27] K. Ito and F. Kappel, *The Trotter-Kato theorem and approximations of PDEs*, Math. Comp., 67 (1998), pp. 21–44.
[28] F. John, *Numerical solution of the equation of heat conduction for preceding times*, Ann. Mat. Pura Appl., 40 (1955), pp. 129–142.
[29] T. Kato, *Theory for Linear Operators,* Springer-Verlag, New York, 1966.
[30] R. Lattes and J. L. Lions, *Méthode de Quasi-réversibilité et Applications,* Dunod, Paris, 1967.
[31] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems,* Birkhäuser, Basel, 1995.
[32] P. Manselli and K. Miller, *Dimensionality reduction methods for efficient numerical solution, backward in time, of parabolic equations with variable coefficients*, SIAM J. Math. Anal., 11 (1980), pp. 147–159.
[33] J. M. Marbán and C. Palencia, *On the numerical recovery of a holomorphic mapping from a finite set of approximate values*, Numer. Math., 91 (2002), pp. 57–75.
[34] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.
[35] K. Miller, *Efficient numerical methods for backward solution of parabolic problems with variable coefficients*, in Improperly Posed Boundary Value Problems, Res. Notes Math. 1, A. Carasso and A. P. Stone, eds., Pitman, Bath, 1975, pp. 54–64.

[36] K. Miller, *Logarithmic convexity results for holomorphic semigroups*, Pacific J. Math., 58 (1975), pp. 549–551.
[37] C. Palencia, *A stability result for sectorial operators in Banach spaces*, SIAM J. Numer. Anal., 30 (1993), pp. 1373–1384.
[38] C. Palencia, *On the stability of variable stepsize approximations of holomorphic semigroups*, Math. Comp., 62 (1994), pp. 93–103.
[39] C. Palencia, *Maximum norm analysis of completely discrete finite element methods for parabolic problems*, SIAM J. Numer. Anal., 33 (1996), pp. 1654–1668.
[40] C. Palencia and J. M. Sanz-Serna, *An extension of the Lax-Richtmyer theory*, Numer. Math., 44 (1984), pp. 279–283.
[41] J. R. Partington, *Interpolation, Identification and Sampling,* London Math. Soc. Monogr. (N.S.) 17, Oxford University Press, New York, 1997.
[42] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, CBMS-NSF Reg. Conf. Ser. in Appl. Math. 22, SIAM, Philadelphia, 1975.
[43] A. Pazy, *Semigroups of Operators and Applications to Partial Differential Equations,* Springer-Verlag, Berlin, 1983.
[44] T. Ransford, *Potential Theory in the Complex Plane,* London Math. Soc. Stud. Texts 28, Cambridge University Press, Cambridge, UK, 1995.
[45] A. H. Schatz and L. B. Wahlbin, *On the quasi-optimality in L_∞ of the $\overset{0}{H}_1$-projection into finite element spaces*, Math. Comp., 38 (1982), pp. 1–21.
[46] T. I. Seidman, *Optimal filtering for the backward heat equation*, SIAM J. Numer. Anal., 33 (1996), pp. 162–170.
[47] R. E. Showalter, *Quasi-reversibility of first and second order parabolic evolution equations*, in Improperly Posed Boundary Value Problems, Res. Notes Math. 1, A. Carasso and A. P. Stone, eds., Pitman, Bath, 1975, pp. 76–84.
[48] E. Sinestrari, *On the abstract Cauchy problem of parabolic type in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16–65.
[49] H. B. Stewart, *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 199 (1974), pp. 141–162.
[50] H. B. Stewart, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.
[51] V. Thomée, *Galerkin Finite Elements for Parabolic Problems*, Springer-Verlag, Berlin, Heidelberg, New York, 1997.
[52] A. N. Tikhonov, *Regularization of incorrectly posed problems*, Dokl. Akad. Nauk. SSSR, 153 (1963), pp. 49–52 (in Russian); Soviet. Math. Dokl., 4 (1963), pp. 1624–1627 (in English).

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1421–1445

NUMERICAL APPROXIMATION OF SOME LINEAR STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS DRIVEN BY SPECIAL ADDITIVE NOISES*

QIANG DU† AND TIANYU ZHANG‡

Abstract. This paper is concerned with the numerical approximation of some linear stochastic partial differential equations with additive noises. A special representation of the noise is considered, and it is compared with general representations of noises in the infinite dimensional setting. Convergence analysis and error estimates are presented for the numerical solution based on the standard finite difference and finite element methods. The effects of the noises on the accuracy of the approximations are illustrated. Results of the numerical experiments are provided.

Key words. stochastic partial differential equation, additive noise, finite difference method, finite element method, convergence, error estimate

AMS subject classifications. 65M65, 65C30, 35R60, 60H15

PII. S0036142901387956

1. Introduction.

In recent years, it has been increasingly acceptable to adopt SDE models as an essential component in the analysis of complex phenomena such as wave propagation [19], climate change [22], turbulence [21, 24], and phase transition [9, 16, 18]. The initial value and boundary value problems of stochastic partial differential equations (SPDEs) have been studied theoretically in, for example, [5, 6, 8, 10, 33]. Various numerical methods and approximation schemes for SDEs have also been developed, analyzed, and tested [1, 2, 4, 7, 12, 13, 14, 15, 20, 25, 27, 29, 28, 31, 34, 35].

For a given physical system, many different stochastic perturbations may be considered. Generically speaking, noise may enter the physical system either as temporal fluctuations of internal degrees of freedom or as random variations of some external control parameters; internal randomness often reflects itself in *additive* noise terms, while external fluctuations gives rise to *multiplicative* noise terms [18]. The main aim of this paper is to study the properties of some standard numerical approximations to the linear SPDEs for the random field $u = u(x,t)$ driven by an additive noise:

$$du = Au\,dt + dW, \quad x \in \Omega, \quad t > 0. \tag{1.1}$$

Here, Ω is a bounded spatial domain and A is a linear second order elliptic operator with deterministic coefficients, which is defined on a space of functions satisfying certain boundary conditions. W represents an infinite dimensional Brownian motion. We also consider the related time-independent equation

$$-Au = g + \dot{W}, \quad x \in \Omega, \tag{1.2}$$

*Received by the editors April 13, 2001; accepted for publication (in revised form) April 25, 2002; published electronically October 23, 2002. This work was partially supported by the State Major Basic Research Project G199903280 and by NSF grant DMS-0196522.
http://www.siam.org/journals/sinum/40-4/38795.html
†Department of Mathematics, Penn State University, University Park, PA 16802, and Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong (qdu@math.psu.edu).
‡Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong. Current address: Department of Mathematics, University of Minnesota, Minneapolis, MN 55455 (tyzhang@math.umn.edu).

where g is a given deterministic function and $\dot{W}$ denote a one-parameter family noise. The additive noises may appear in various forms, ranging from the space time white noise to colored noises generated by some infinite dimensional Brownian motion with a prescribed covariance operator [6, 28]. Once the equation is reformulated into a weak form [5], the usual Galerkin finite element methods can be constructed and also analyzed using standard techniques. A priori error estimates of the numerical solution depend on the regularity of the solutions of the original SPDE. Such regularity results are often much harder to establish than their deterministic counterpart [5, 33]. In fact, if dW corresponds to the Brownian white noise, then the regularity estimates are usually very weak, and they lead to very low order error estimates [1, 7, 13]. On the other hand, if the noise is more regular, then it becomes possible to get higher order of error estimates for the numerical solution. In recent years, studies of models with colored noises and their numerical approximation have started to receive more attention; see [28] for an example of physical application and the recent works [26, 14] for works related to stochastic ordinary differential equations (SODEs) and the time discretization. In the present work, we provide the connections between the discrete realizations of noises in different formulations of some SPDEs. Moreover, we illustrate how the error analysis of the standard finite element and finite difference approximations depends on the noises used in the model and the approximation. In order to present a simple analysis, in this paper we focus on the case $\Omega = (0, 1)$ and $Au = u_{xx} - bu$ with the homogeneous Dirichlet boundary condition and b being a deterministic coefficient, though much of our results can be readily extended to higher spatial dimensions and more general second order elliptic operators. For most of the discussion, we also try to present our results in simple finite element terminology that is familiar to people working on the numerical approximations of deterministic PDEs so that it is easy to be understood even for readers who are not necessarily experts on SDEs.

The paper is organized as follows. We first describe the various forms of the noises and their discrete representations. Next, we discuss some convergence results for standard finite element and finite difference approximations. The models used are one dimensional linear stochastic elliptic and parabolic equations, and the results are established for noises given in general forms, which include the spatial or space time white noises as special cases. Then numerical results are presented to support the theoretical analysis. Finally, some concluding remarks are given. The details of the proofs are provided in the appendix.

2. The representation of random noises. To study the accuracy of the discrete approximations, it is useful to first consider the properties of the noises which drive the stochastic equations and the discrete representations of the noises.

Following [1], we *regularize* the noise through discretization. Let $\{x_i = ih\}_0^n$ be a partition of $[0, 1]$ with $h = 1/n$. We begin with $\dot{W}(x)$ being the standard one-parameter family Brownian white noise that satisfies

$$E(\dot{W}(x) \cdot \dot{W}(x')) = \delta(x - x'), \tag{2.1}$$

where δ denote the usual Dirac δ-function and E the expectation. A piecewise constant approximation of the one-parameter white noise is given by [1]

$$\frac{d\widehat{W}_n(x)}{dx} = c_n \sum_{j=1}^{n} \eta_j \chi_j(x), \tag{2.2}$$

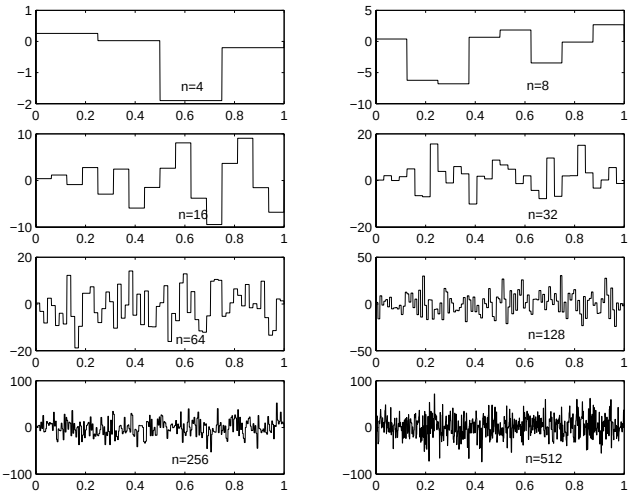


FIG. 2.1. *Piecewise constant approximation for the noise* $d\widehat{W}_n(x)/dx = (1/\sqrt{h})\sum_{j=1}^{n}\eta_j\chi_j(x)$.

where $c_n = h^{-1/2} = \sqrt{n}$ and, for $j = 1, 2, \ldots, N$, $\eta_j \in N(0,1)$ is independently and identically distributed (iid),

$$\sqrt{h}\eta_j = \int_{x_j}^{x_{j+1}} dW(x)\,, \quad \text{and} \quad \chi_j(x) = \begin{cases} 1, & x_j \le x < x_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

The discrete analogue of (2.1) for the piecewise constant approximation is given by

$$E\left(\frac{d\widehat{W}_n(x)}{dx}\cdot\frac{d\widehat{W}_n(x')}{dx}\right) = \begin{cases} h^{-1} & \text{if } x_j \le x,\ x' < x_{j+1} \text{ for some } j, \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\lim_{n\to\infty} E\left(\frac{d\widehat{W}_n(x)}{dx}\cdot\frac{d\widehat{W}_n(x')}{dx}\right) = \delta(x - x').$$

In Figure 2.1, some sample realizations of the piecewise constant approximation of one-parameter white noise are illustrated for various values of n. (The random numbers are generated using MATLAB.) We note that similar discussions can be easily generalized to the space time two-parameter family white noises.

2.1. Noises in abstract forms. The SPDEs driven by the white noise often have poor regularity estimates. In the physical world, to take into account the short and long range correlations of the stochastic effects, both white noise and colored noises may be considered. There are many situations where colored noises model the reality more closely, and there are also instances where the important stochastic effects are the noises acting on a few selected frequencies.

In general, we may use an abstract formulation of the infinite dimensional noise:

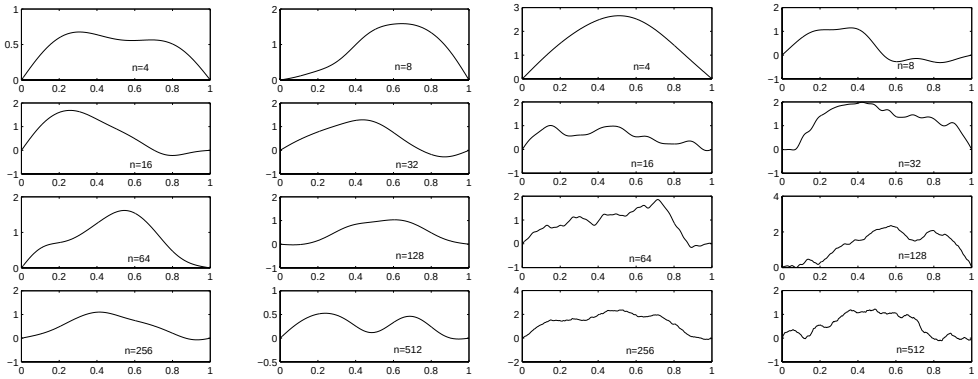


FIG. 2.2. *Noises by Fourier modes* $\sum_{k=1}^{n} \sigma_k \eta_k \sqrt{2} \sin k\pi x$ *with* $\sigma_k = \frac{1}{2^k}$ *(left) and* $\sigma_k = \frac{1}{k^{3/2}}$ *(right).*

$$\dot{W}(x) = \sum_{k=1}^{\infty} \sigma_k \eta_k \psi_k(x), \tag{2.3}$$

where the random variable $\eta_k \sim N(0,1)$ is iid for any k, the deterministic functions $\{\psi_k(x)\}$ form an orthonormal basis of $L_2(0,1)$ or its subspace, and the coefficients $\{\sigma_k\}$ are to be chosen to ascertain the convergence of the series in the mean square sense with respect to some suitable norms.

One of the examples is given by the Fourier modes $\psi_k(x) = \sqrt{2}\sin k\pi x$ which forms a basis of $H_0^1(0,1)$. According to the different decay rates of the coefficients, the noises may display quite different pictures. The pictures in Figure 2.2 and the left two columns of Figure 2.3 provide sample realizations of noises having forms (2.3) in the Fourier basis with coefficients $\sigma_k = 2^{-k}$, $k^{-3/2}$, and $k^{-1/2}$, respectively. Clearly, the realizations give trajectories that look smoother than the ones for the white noise. It can also be seen that the faster the coefficients σ_k decay, the smoother the noise trajectory dW_n/dx looks, which reflects stronger spatial correlation since the noises are heavily concentrated near a few low frequencies. On the other hand, if the coefficients decay sufficiently slowly, then the trajectory can clearly resemble that of a white noise away from the boundary. In fact, it is well known that for spatially uncorrelated white noises, their Fourier coefficients are independent of the frequencies, and they stay at a constant value.

In the analysis and numerical examples given in later sections, the noises given in terms of the Fourier modes are used. The Fourier modes provide one of many possible representations of noises where the smoothness of the noise trajectories are related to the decay of the coefficients in the representation. Another illustrative example is to define the noise in terms of the lowest order wavelet basis. We include the discussion here for comparison. Let ψ be the *wavelet function* and ϕ be the *scaling function* [32]. Let j denote the *dilation index* and k denote the *translation index*, and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. The discrete noise formulated in the wavelet basis is given as

$$\dot{W}_J(x) = c\gamma\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}\eta_{jk}\psi_{j,k}(x). \tag{2.4}$$

Here, J is the highest level to be considered, and $\gamma, \eta_{jk} \in N(0,1)$ are iid. In the

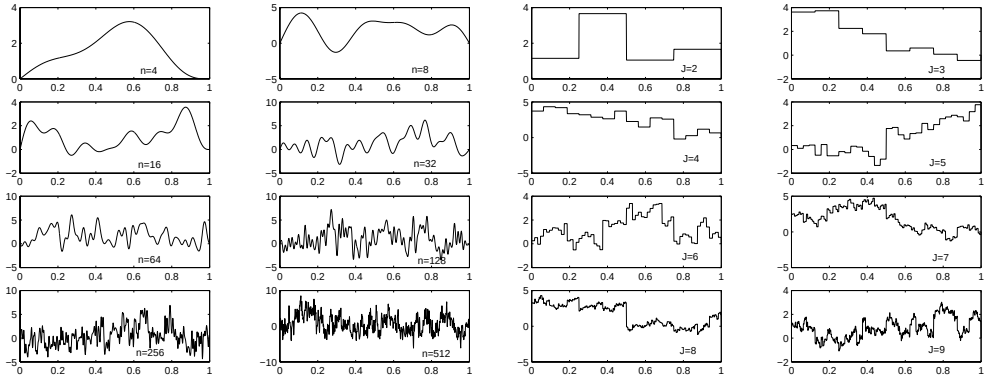


FIG. 2.3. *Noises by* $\sum_{k=1}^{n} \frac{1}{k^{1/2}} \eta_k \sqrt{2} \sin k\pi x$ *(left) and* $\gamma\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \frac{1}{2^j} \eta_{jk} \psi_{j,k}(x)$ *(right).*

simplest case, we may take the *Haar* wavelet

$$\psi(x) = \begin{cases} 1, & 0 \le x < 1/2, \\ -1, & 1/2 \le x < 1, \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \phi(x) = \begin{cases} 1, & 0 \le x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The right two columns of Figure 2.3 show sample realizations of noises taking the form (2.4) with $c = 1$, $d_{j,k} = 2^{-j}$. The correlation of the noise (2.4) $E(\frac{dW_J(x)}{dx} \cdot \frac{dW_J(x')}{dx})$ is given by

$$E\left(\frac{dW_J(x)}{dx} \cdot \frac{dW_J(x')}{dx}\right) = c^2\phi(x)\phi(x') + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}^2 \psi_{j,k}(x)\psi_{j,k}(x') \,. \tag{2.5}$$

If in (2.2) $n = 2^J$ and $h = 2^{-J}$, then the piecewise constant approximation of the white noise may also be represented using the wavelet Haar basis. In fact, let $\chi_k(x)$ be characteristic function of interval $[kh, (k+1)h]$; then

$$\frac{d\widehat{W}_n(x)}{dx} = \frac{1}{\sqrt{h}} \sum_{k=0}^{2^J-1} \eta_k \chi_k(x) = \gamma\phi(x) + \sum_{j=0}^{J-1} \sum_{l=0}^{2^j-1} \gamma_{j,l} \psi_{j,l}(x).$$

Here, $\gamma = 2^{-J/2} \sum_{k=0}^{2^J-1} \eta_k \sim N(0,1)$ and

$$\gamma_{j,l} = 2^{(j-J)/2} \left(\sum_{k=l2^{J-j}}^{(l+1)2^{J-j}-1} (-1)^{[k/2^{J-j-1}]} \eta_k \right) \sim N(0,1)$$

are iid. Corresponding to (2.5), $c = d_{j,k} = 1$ so that (2.5) leads again to (2.1). Naturally, when higher order wavelets are used [32, 30], we may expect to have discrete noises that are smoother spatially than the ones represented by the Haar basis when the high frequency coefficients enjoy fast decay properties. Comparing with Fourier modes, wavelet functions may also have compact support; thus, on the one hand, the noises in wavelet basis can closely resemble spatially uncorrelated white noises, while on the other hand they can also be used conveniently to simulate noise more concentrated on certain frequencies as well as certain spatial regions.

In summary, different forms to represent the various noises are discussed in this section. Similar discussion can be carried out in more than one space dimension and for noises parameterized by both time and space variables. Such discussions are relevant to the numerical study of SDEs as the solutions of the stochastic equations that use noises with better regularity become more regular themselves and thus may allow higher order numerical approximations.

3. Numerical method and error analysis. In [1], approximations of SPDEs with the additive space time white noise term discretized by the piecewise constant random process have been studied. Here, we follow roughly the same route, though more general types of noises are used. We show how the accuracy is affected by the correlation or the *smoothness* of the noises.

We divide the discussion into two parts, starting with the simplest one dimensional elliptic equation (boundary value problem of a SODE) and then moving to a parabolic equation in one space dimension and in time (initial boundary value problem of a SPDE). In the set-up of the problems, noises represented in general basis are used, but in the analysis we specialize in using the Fourier modes as the basis of choice to simplify the discussion.

3.1. One dimensional elliptic equation with noise. We now consider the SDE (1.2); that is,

$$\begin{cases} -\Delta u(x) + bu(x) = g(x) + \dot{W}(x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{3.1}$$

where $\dot{W}(x)$ denotes the noise, $g(x)$ is a given deterministic term, and $b = b(x)$ is a given deterministic coefficient.

As in [1], we may first replace $\dot{W}(x)$ by a finite dimensional noise $\dot{W}_n(x)$ and let u_n denote the solution of the corresponding equation. We then numerically approximate the equation associated with $\dot{W}_n(x)$ and let u_n^h denote the numerical solution.

If the noise $\dot{W}(x)$ in (3.1) is the white noise, $\dot{W}_n(x)$ is the piecewise constant approximation (2.2), and the Galerkin finite element method with piecewise constant basis is applied to (3.1), the error estimate is given by [1]

$$\begin{aligned} &E\|u - u_n\|_{L_2} \le C\ h, \\ &E\|u_n - u_n^h\|_{L_2} \le C\ h^{3/2}, \\ &E\|u - u_n^h\|_{L_2} \le C\ h. \end{aligned}$$

Due to the poor regularity of the solution, it is seen that, even with higher order finite elements, the order of error estimates does not improve. With colored noises, the order of approximation may increase with better regularity on the solution and the use of higher order elements. As an illustration, we consider the following noise:

$$\dot{W}(x) = \sum_{k=1}^{\infty} \sigma_k \eta_k \psi_k(x), \tag{3.2}$$

where $\{\eta_k\}$ are random variables satisfying

$$\eta_k \sim N(0,1) \ \text{ and } \ cov(\eta_k, \eta_l) = E(\eta_k \eta_l) = q_{kl},$$

with $\{\sigma_k\}$ to be chosen.

Let $\{\sigma_k^n\}_{k=1}^{\infty}$ approach $\{\sigma_k\}_{k=1}^{\infty}$ as $n \to \infty$ in some appropriate sense; then an approximation of $\dot{W}(x)$ is

$$\dot{W}_n(x) = \sum_{k=1}^{\infty} \sqrt{2}\sigma_k^n \eta_k \psi_k(x) \sin k\pi x.$$

The definition of noise term leads to the following stochastic integral for $f \in L_2(0,1)$:

$$S = \int_0^1 f(x)dW(x) = \sum_{k=1}^{\infty} \sigma_k f_k \eta_k,$$

$$S_n = \int_0^1 f(x)dW_n(x) = \sum_{k=1}^{\infty} \sigma_k^n f_k \eta_k,$$

where $f_k = \int_0^1 f(x)\psi_k(x)dx$. That is, S and S_n are random variables having the distribution

$$S \sim N\left(0, \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \sigma_k \sigma_l f_k f_l q_{kl}\right),$$

$$S_n \sim N\left(0, \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \sigma_k^n \sigma_l^n f_k f_l q_{kl}\right),$$

provided the double sum is convergent.

For convenience, we introduce the following notation:

$$\overrightarrow{\sigma^n} = (\sigma_1^n, \sigma_2^n, \dots, \sigma_k^n, \dots)^T,$$
$$\vec{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_k, \dots)^T$$

are infinite column vectors. For two vectors $\overrightarrow{\sigma^n}$ and $\overrightarrow{f}$, we use $\overrightarrow{\sigma^n f}$ to denote the componentwise product

$$\overrightarrow{\sigma^n f} = (\sigma_1^n f_1, \sigma_2^n f_2, \dots, \sigma_k^n f_k, \dots)^T.$$

Let Q be the covariance matrix of random fields $\{\eta_k\}$, namely, Q is the infinite matrix (operator) with entries $Q = (q_{kl})_{k,l=1}^{\infty}$. For an integer s, let Q_s be the infinite matrix with entries $Q_s = ((kl)^s q_{kl})_{k,l=1}^{\infty}$. It is easy to see both Q and Q_s are positive semidefinite. Define the weighted semi-inner products of the vectors $\vec{\sigma}$ and $\vec{\delta}$ as

$$\langle \vec{\sigma}, \vec{\delta} \rangle_Q = \vec{\sigma}^T \cdot Q \cdot \vec{\delta} = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \sigma_k \delta_l q_{kl},$$

$$\langle \vec{\sigma}, \vec{\delta} \rangle_{Q_s} = \vec{\sigma}^T \cdot Q_s \cdot \vec{\delta} = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \sigma_k \delta_l (kl)^s q_{kl}.$$

The seminorms induced by the above semi-inner products are

$$\|\vec{\sigma}\|_Q^2 = \langle \vec{\sigma}, \vec{\sigma} \rangle_Q \quad \text{and} \quad \|\vec{\sigma}\|_{Q_s}^2 = \langle \vec{\sigma}, \vec{\sigma} \rangle_{Q_s}.$$

Note that $Q_0 = Q$. Using the above notation,

$$S \sim N\left(0, \|\overrightarrow{\sigma f}\|_Q^2\right), \qquad S_n \sim N\left(0, \|\overrightarrow{\sigma^n f}\|_Q^2\right).$$

The difference between S and S_n is given by

$$E|S - S_n|^2 = E|\sum_{k=1}^{\infty}(\sigma_k^n - \sigma_k) f_k \eta_k|^2 = \left\|\overrightarrow{\sigma f} - \overrightarrow{\sigma^n f}\right\|_Q^2.$$

Equation (3.1) can be written in a weak form or an integral form. Both forms are equivalent as shown in [3]. In fact, the solution of (3.1) is a stochastic process $u = u(x)$ which satisfies the weak formulation

$$-\int_0^1 u(x)\Delta\phi(x)dx + \int_0^1 bu(x)\phi(x)dx = \int_0^1 g(x)\phi(x)dx + \int_0^1 \phi(x)dW(x) \tag{3.3}$$

for $\phi \in C^2(0,1) \cap C_0(0,1)$. The integral form is

$$u(x) + \int_0^1 b\,k(x,y)\,u(y)dy = \int_0^1 k(x,y)g(y)dy + \int_0^1 k(x,y)dW(y). \tag{3.4}$$

Here, $k(x,y) = x \wedge y - xy$ is the Green's function associated with the elliptic equation $-\Delta v(x) = \phi(x), v(0) = v(1) = 0$ so that $v(x) = \int_0^1 k(x,y)\phi(y)dy$. ($x \wedge y$ means the smaller one of x and y.) In the present investigation, it is assumed the coefficient b is small enough so that $\lambda^2 = \int_0^1 \int_0^1 b^2 k^2(x,y)dxdy < 1$. We note that this condition is primarily needed in the case of $b < 0$; such a restriction can be lifted for $b > 0$, and the conclusions given later remain valid.

We now substitute $dW(y)$ by $dW_n(y)$ in (3.4) to obtain the following equation:

$$u_n(x) + \int_0^1 b\,k(x,y)\,u_n(y)dy = \int_0^1 k(x,y)g(y)dy + \int_0^1 k(x,y)dW_n(y). \tag{3.5}$$

Thus, $u_n(x)$ satisfy the two-point boundary value problem

$$-\Delta u_n(x) + bu_n(x) = g(x) + \dot{W}_n(x), \qquad u_n(0) = u_n(1) = 0. \tag{3.6}$$

The following theorem shows that u_n indeed approximates u, the solution of (3.4). In order to illustrate the higher order of convergence for more regular noises, we specialize our discussion to the choice of $\{\psi_k(x) = \sqrt{2}\sin k\pi x\}$, that is, noises represented by the Fourier modes.

THEOREM 3.1. *For $\dot{W}_n(x) = \sum_{k=1}^{\infty} \sigma_k^n \eta_k \psi_k(x)$ and $\psi_k(x) = \sqrt{2}\sin k\pi x$, if u_n and u are the solutions of* (3.5) *and* (3.4)*, respectively, then, for some constant $C > 0$,*

$$E\|u - u_n\|_{L_2} \le \frac{C}{1-\lambda}\left\|\overrightarrow{\sigma^n} - \vec{\sigma}\right\|_{Q_{-1}},$$

where $\lambda < 1$ is defined as before.

Proof. Let $e_n(x) = u(x) - u_n(x)$ and

$$F(x) = \int_0^1 k(x,y)dW(y) - \int_0^1 k(x,y)dW_n(y).$$

Subtracting (3.5) from (3.4), we have

$$e_n(x) = -\int_0^1 b\,k(x,y)\,e_n(y)dy + F(x).$$

By Hölder's inequality, it is easy to show that

$$\int_0^1 e_n^2(x)dx \le \lambda^2 \int_0^1 e_n^2(y)dy + 2\lambda \left(\int_0^1 F^2(x)dx\right)^{1/2} \left(\int_0^1 e_n^2(y)dy\right)^{1/2} + \int_0^1 F^2(x)dx,$$

where $\lambda^2 = \int_0^1 \int_0^1 b^2k^2(x,y)dxdy$ and it is assumed that $\lambda < 1$. Taking expectations on both sides, letting $\hat{e}_n = E(\int_0^1 e_n^2(x)dx)$ and $\hat{G}_n = E(\int_0^1 F^2(x)dx)$ and using the Burkholder–Gundy-type inequality $(EX)^2 \le E(X^2)$, we get

$$\hat{e}_n(1-\lambda^2) - 2\lambda\sqrt{\hat{e}_n}\sqrt{\hat{G}_n} - \hat{G}_n \le 0. \tag{3.7}$$

This implies

$$\sqrt{\hat{e}_n} \le \sqrt{\hat{G}_n}(1-\lambda). \tag{3.8}$$

Now let us estimate $\hat{G}_n$.

$$\begin{aligned}\hat{G}_n &= E\left(\int_0^1 F^2(x)dx\right) = \int_0^1 E\left(\sum_{k=1}^{\infty}(\sigma_k^n - \sigma_k)f_k(x)\eta_k\right)^2 dx \\ &= \int_0^1 \left\|\overrightarrow{\sigma f(x)} - \overrightarrow{\sigma^n f(x)}\right\|_Q^2 dx,\end{aligned}$$

where $\overrightarrow{f(x)} = (f_1(x), f_2(x), \dots, f_k(x), \dots)^T$ and $f_k(x) = \int_0^1 k(x,y)\psi_k(y)dy$. Since $k(x,y) = x \wedge y - xy$, direct calculation gives that, for any $x \in [0,1]$,

$$|f_k(x)| = \left|\int_0^1 k(x,y)\psi_k(y)dy\right| = \left|\int_0^1 k(x,y)\sqrt{2}\sin k\pi y dy\right| \le \frac{c}{k},$$

which implies that, for $x \in [0,1]$,

$$\left\|\overrightarrow{\sigma f(x)} - \overrightarrow{\sigma^n f(x)}\right\|_Q \le C\left\|\overrightarrow{\sigma} - \overrightarrow{\sigma^n}\right\|_{Q_{-1}}$$

for some constant $C > 0$. Hence,

$$\hat{G}_n \le C\left\|\vec{\sigma} - \overrightarrow{\sigma^n}\right\|_{Q_{-1}}^2.$$

Combining the above inequality with (3.8), we get

$$E\|u - u_n\|_{L_2} \le \sqrt{E\|u-u_n\|_{L_2}^2} = \sqrt{\hat{e}_n} \le \frac{C}{1-\lambda}\left\|\overrightarrow{\sigma^n} - \vec{\sigma}\right\|_{Q_{-1}}.$$

This proves the theorem. □

We now state a bound on $\dot{W}_n(x)$ in the following lemma.

LEMMA 3.1. *For* $\dot{W}_n(x) = \sum_{k=1}^{\infty} \sigma_k^n \eta_k \psi_k(x)$ *and* $\psi_k(x) = \sqrt{2}\sin k\pi x$, *if* $s \geq 0$ *is an integer, then*

$$E\|\dot{W}_n\|_{H^s} \leq C\left(\sum_{k=1}^{\infty}(\sigma_k^n k^s)^2\right)^{1/2},$$

provided that the right-hand side is convergent.

Proof. First,

$$\frac{d^s}{dx^s}\left(\frac{dW_n}{dx}\right) = \sum_{k=1}^{\infty}\sqrt{2}\sigma_k^n\eta_k(k\pi)^s \sin\left(s\frac{\pi}{2} + k\pi x\right).$$

Since $\{\sin(s\frac{\pi}{2} + k\pi x)\}$ are orthogonal on $[0,1]$, we have

$$\begin{aligned} E\left\|\frac{d^s}{dx^s}\left(\frac{dW_n}{dx}\right)\right\|_{L_2}^2 &= E\int_0^1\left(\sum_{k=1}^{\infty}\sqrt{2}\sigma_k^n\eta_k(k\pi)^s\sin\left(s\frac{\pi}{2}+k\pi x\right)\right)^2 dx \\ &= E\sum_{k=1}^{\infty}(\sigma_k^n)^2\eta_k^2(k\pi)^{2s} \leq c\sum_{k=1}^{\infty}(\sigma_k^n\cdot k^s)^2 \end{aligned}$$

for some constant $c > 0$. The above inequality also implies that, for any $r \leq s$,

$$E\left\|\frac{d^r}{dx^r}\left(\frac{dW_n}{dx}\right)\right\|_{L_2}^2 \leq E\left\|\frac{d^s}{dx^s}\left(\frac{dW_n}{dx}\right)\right\|_{L_2}^2.$$

Hence,

$$E\|\dot{W}_n\|_{H^s} \leq \sqrt{E\|\dot{W}_n\|_{H^s}^2} \leq C\left(\sum_{k=1}^{\infty}(\sigma_k^n k^s)^2\right)^{1/2}$$

for some constant $C > 0$. ∎

Concerning the above lemma, we note that similar lower bound can also be established. Moreover, the results may be established for the case $s < 0$ as well.

We now consider a standard finite element approximation of u_n. From the weak formulation (3.3), u_n satisfies

$$\int_0^1 u_n'\phi'(x)dx + b\int_0^1 u_n(x)\phi(x)dx = \int_0^1 g(x)\phi(x)dx + \int_0^1 \phi(x)dW_n(x) \tag{3.9}$$

for $\phi(x) \in H_0^1(0,1)$. By the Lax–Milgram theorem, there exists a unique solution $u_n \in H_0^1(0,1)$ to (3.9). For convenience, we consider the same partition of $[0,1]$: $0 = x_1 < x_2 < \cdots < x_{n+1} = 1$ with $x_i = (i-1)h$ and $h = 1/n$. If $V_0^h(0,1)$ denotes the finite element subspace of $H_0^1(0,1)$, and $\{\phi_j(x)\}_{j=1}^N$ forms a basis of $V_0^h(0,1)$, the finite element solution of (3.9) is $u_n^h \in V_0^h(0.1)$ that satisfies (3.9) for all $\phi(x) \in V_0^h(0,1)$. Thus, $u_n^h(x) = \sum_{l=1}^N u_l\phi_l(x)$ satisfies the following linear system for $j = 1, 2, \ldots, N$:

$$\begin{aligned} &\sum_{l=1}^N u_l\int_0^1 \phi_l'(x)\phi_j'(x) + b\sum_{l=1}^N u_l\int_0^1\phi_l(x)\phi_j(x)dx \\ &= \int_0^1 g(x)\phi_j(x)dx + \sum_{k=1}^{\infty}\sigma_k^n\eta_k\int_0^1\phi_j(x)\psi_k(x)dx, \end{aligned} \tag{3.10}$$

where $\eta_k \in N(0,1)$. The solution u_n^h is clearly well defined.

The following lemma gives the standard finite element error estimates of (3.9) in the pathwise sense.

LEMMA 3.2. *If $V_0^h(0,1)$ contain all piecewise polynomials of degree r in $H_0^1(0,1)$, and $u_n \in H_0^1(0,1) \cap H^{r+1}(0,1)$, then*

$$\|u_n - u_n^h\|_{L^2} + h|u_n - u_n^h|_{H^1} \le Ch^{r+1}\|u_n\|_{H^{r+1}} \le Ch^{r+1}\|g + \dot{W}_n\|_{H^{r-1}} \tag{3.11}$$

for some constant $C > 0$. □

Furthermore, combining Theorem 3.1 and Lemma 3.2, an estimate on $E(\|u - u_n^h\|_{L_2})$ follows from the triangle inequality.

THEOREM 3.2. *Let u and u_n^h be the solution of* (3.3) *and* (3.10)*, respectively; if the hypothesis in Lemma* 3.2 *is satisfied, then the error estimate is*

$$\begin{aligned} E\|u - u_n^h\|_{L_2} &\le C\left\{\left\|\overrightarrow{\sigma^n} - \vec{\sigma}\right\|_{Q_{-1}} + h^{r+1}\|g\|_{H^{r-1}} + h^{r+1}E\|\dot{W}_n\|_{H^{r-1}}\right\} \\ &\le C\left\{\left\|\overrightarrow{\sigma^n} - \vec{\sigma}\right\|_{Q_{-1}} + h^{r+1}\|g\|_{H^{r-1}} + h^{r+1}\left[\sum_{k=1}^{\infty}(\sigma_k^n k^{r-1})^2\right]^{1/2}\right\} \end{aligned} \tag{3.12}$$

for some generic constant $C > 0$. □

Numerical examples are given in a later section to provide an illustration of the specific order of error estimates one can get based on the above theorem.

Remark 3.1. The same idea can be applied to two dimensional elliptic equations in a rectangular domain, namely, by representing the two dimensional noise as the combinations of the tensor products of $\psi_k(x)$, similar to how Theorem 3.2 can be obtained.

3.2. Parabolic equation in one spatial dimension. Let $\frac{\partial^2 W}{\partial t \partial x}$ denote a space time noise term and g be a deterministic function; we now consider the linear stochastic equations of the form

$$\begin{cases} \frac{\partial u}{\partial t}(t,x) - \frac{\partial^2 u}{\partial x^2}(t,x) + bu(t,x) = \frac{\partial^2 W}{\partial t \partial x}(t,x) + g(t,x),\ t > 0, \\ u(0,x) = u_0(x),\ 0 \le x \le 1, \\ u(t,0) = u(t,1) = 0,\ t \ge 0, \end{cases} \tag{3.13}$$

where the coefficient b, for simplicity, is assumed to be a constant.

The weak formulation of (3.13) is

$$\begin{aligned} &\int_0^1 u(t,x)\phi(x)dx - \int_0^t\int_0^1 u(s,x)\frac{d^2\phi}{dx^2}dxds + \int_0^t\int_0^1 bu(s,x)\phi(x)dxds \\ &= \int_0^1 u_0(x)\phi(x)dx + \int_0^t\int_0^1 \phi(x)dW(s,x) + \int_0^t\int_0^1 g(s,x)\phi(x)dxds \end{aligned} \tag{3.14}$$

for $\phi \in C^2[0,1] \cap C_0[0,1]$. The integral formulation of (3.13) is

$$\begin{aligned} u(t,x) + \int_0^t\int_0^1 G_{t-s}(x,y)bu(x,y)dyds &= \int_0^1 G_t(x,y)u_0(y)dy \\ &+ \int_0^t\int_0^1 G_{t-s}(x,y)dW(s,y) + \int_0^t\int_0^1 G_{t-s}(x,y)g(s,y)dyds, \end{aligned} \tag{3.15}$$

where $G_t(x,y) = 2\sum_{m=1}^{\infty} \sin m\pi x \sin m\pi y e^{-(m\pi)^2 t}$ is the fundamental solution of

$$v_t(t,x) - v_{xx}(t,x) = 0, \quad v(0,x) = \phi(x), \quad v(t,0) = v(t,1) = 0,$$

so that $v(t,x) = \int_0^1 G_t(x,y)\phi(y)dy$.

Using the same idea as that in the previous section, we represent the noise as

$$\frac{\partial^2 W}{\partial t \partial x} = \sum_{k=1}^{\infty} \sigma_k(t)\dot{\eta_k}(t)\psi_k(x), \tag{3.16}$$

where $\sigma_k(t)$ is a continuous function, $\dot{\eta_k}(t)$ is the derivative of standard Wiener process, and $\psi_k(x) = \sqrt{2}\sin k\pi x$. Now define a partition of $[0,T] \times [0,1]$ by rectangles $[t_i, t_{i+1}] \times [x_j, x_{j+1}]$ for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, n$, where $t_i = (i-1)\Delta t$, $x_j = (j-1)h$, $\Delta t = T/I$, and $h = 1/n$. A sequence of noise which approximates the noise is defined as

$$\frac{\partial^2 W_n}{\partial t \partial x} = \sum_{k=1}^{\infty} \sigma_k^n(t)\psi_k(x) \sum_{i=1}^{I} \frac{1}{\sqrt{\Delta t}} \eta_{ki}\chi_i(t), \tag{3.17}$$

where $\chi_i(t)$ is the characteristic function for the ith time subinterval and

$$\eta_{ki} = \frac{1}{\sqrt{\Delta t}} \int_{t_i}^{t_{i+1}} d\eta_k(t) \ \sim \ N(0,1).$$

Replacing $\sigma_k(t)$ by $\sigma_k^n(t)$, we get the discretization in the x-direction, and replacing $\dot{\eta_k}(t)$ by $\sum_{i=1}^{I} \frac{1}{\sqrt{\Delta t}} \eta_{ki}\chi_i(t)$ we get the discretization in the t-direction. Then $\frac{\partial^2 W_n}{\partial t \partial x}$ is substituted for $\frac{\partial^2 W}{\partial t \partial x}$ in (3.15) to get the following equation:

$$\begin{aligned} u_n(t,x) + \int_0^t \int_0^1 G_{t-s}(x,y) b u_n(s,y) dy ds &= \int_0^1 G_t(x,y) u_0(y) dy \\ + \int_0^t \int_0^1 G_{t-s}(x,y) dW_n(s,y) &+ \int_0^t \int_0^1 G_{t-s}(x,y) g(s,y) dy ds; \end{aligned} \tag{3.18}$$

that is, u_n is the solution of the equation

$$\begin{cases} \frac{\partial u_n}{\partial t}(t,x) - \frac{\partial^2 u_n}{\partial x^2}(t,x) + b u_n(t,x) = \frac{\partial^2 W_n}{\partial t \partial x}(t,x) + g(t,x), \ t > 0, \\ u_n(0,x) = u_0(x), \ 0 \le x \le 1, \\ u_n(t,0) = u_n(t,1) = 0, \ t \ge 0. \end{cases} \tag{3.19}$$

Now we assume that

$$\int_0^T \int_0^1 \int_0^t \int_0^1 G_{t-s}^2(x,y) b^2 dy ds dx dt = \bar{\lambda}^2 < 1.$$

Then, under proper assumptions on $\{\sigma_k(t)\}$ and $\{\sigma_k^n(t)\}$, u_n approximates u, the solution of (3.15), as illustrated in the next theorem.

THEOREM 3.3. *Let* $\{\sigma_k(t)\}$ *and its derivative be uniformly bounded by*

$$|\sigma_k(t)| \le \beta_k, \ \ |\sigma_k'(t)| \le \gamma_k \ \ \forall t \in [0,T],$$

and the coefficients $\{\sigma_k^n(t)\}$ *are constructed such that*

$$|\sigma_k(t) - \sigma_k^n(t)| \le \alpha_k^n, \quad |\sigma_k^n(t)| \le \beta_k^n, \quad |\sigma_k^{n\prime}(t)| \le \gamma_k^n \quad \forall t \in [0,T]$$

with positive sequences $\{\alpha_k^n\}$ *being arbitrarily chosen,* $\{\beta_k^n\}$ *and* $\{\gamma_k^n\}$ *being related to* $\{\alpha_k^n\,\beta_k\}$ *and* $\{\gamma_k\}$. *Let* $u_n(t,x)$ *and* $u(t,x)$ *be the solution of* (3.18) *and* (3.15), *respectively; then, for some constants* $C > 0$, *independent of* Δt *and* h,

$$E\|u - u_n\|_{L_2}^2 \le \frac{C}{(1-\bar{\lambda})^2} \sum_{k=1}^{\infty} \left(\frac{(\alpha_k^n)^2}{2(k\pi)^2} + [k^4(\beta_k^n)^2 + (\gamma_k^n)^2](\Delta t)^2 \right), \tag{3.20}$$

provided that the infinite series are all convergent.

The proof of Theorem 3.3 is given in the appendix.

Remark 3.2. The assumption on $\bar{\lambda}$ being small is not crucial; some generalizations can be made without this assumption, for example when $b < 0$.

Now we consider the approximation of u_n. In particular, we use a finite element discretization with respect to the x variable and an implicit difference method in the t variable. Since u_n satisfies the weak formulation,

$$\begin{aligned} &\int_0^1 u_n(t,x)\phi(x)dx + \int_0^t \int_0^1 \frac{\partial u_n}{\partial x}(s,x)\frac{d\phi}{dx}(x)dxds + \int_0^t \int_0^1 bu_n(s,x)\phi(x)dxds \\ &= \int_0^1 u_0(x)\phi(x)dx + \int_0^t \int_0^1 \phi(x)dW_n(s,x) + \int_0^t \int_0^1 g(s,x)\phi(x)dxds \end{aligned} \tag{3.21}$$

for $\phi \in H_0^1(0,1)$. Meanwhile, the semidiscretization in space leads only to the following problem: find $u_n(t,\cdot) \in H_0^1(0,1)$, $t \in (0,T)$, such that

$$\int_0^1 \frac{\partial u_n}{\partial t}\phi dx + \int_0^1 \frac{\partial u_n}{\partial x}\frac{\partial \phi}{\partial x}dx + \int_0^1 bu_n\phi dx = \int_0^1 \left(g + \frac{\partial^2 W_n}{\partial t \partial x} \right)\phi dx \tag{3.22}$$

with

$$\int_0^1 u_n(0,x)\phi(x)dx = \int_0^1 u_0(x)\phi(x)dx$$

for all $\phi \in H_0^1(0,1), t \in (0,T)$.

The finite element discretization of (3.22) is to find $\bar{u}_n^h(t,\cdot) \in V_0^h(0,1)$, $t \in (0,T)$, such that

$$\int_0^1 \frac{\partial \bar{u}_n^h}{\partial t}\phi dx + \int_0^1 \frac{\partial \bar{u}_n^h}{\partial x}\frac{\partial \phi}{\partial x}dx + \int_0^1 b\bar{u}_n^h\phi dx = \int_0^1 \left(g + \frac{\partial^2 W_n}{\partial t \partial x} \right)\phi dx \tag{3.23}$$

with

$$\int_0^1 \bar{u}_n^h(0,x)\phi(x)dx = \int_0^1 u_0(x)\phi(x)dx$$

for all $\phi \in V_0^h(0,1)$, $t \in (0,T)$. Here, $V_0^h(0,1)$ denote the finite element subspace of $H_0^1(0,1)$. By using the expression

$$\bar{u}_n^h(t,x) = \sum_{l=1}^{n-1} u_l(t)\phi_l(x), \qquad t \in (0,T),$$

(3.23) leads to a system of ODEs for $u_l(t)$, $l = 1, \ldots, n-1$. Using the backward-Euler method to solve this ODE system yields the following numerical scheme:

$$\begin{aligned}
&\sum_{l=1}^{n-1}(u_{i+1,l} - u_{i,l}) \int_0^1 \phi_l(x)\phi_j(x)dx + \Delta t \sum_{l=1}^{n-1} u_{i+1,l} \int_0^1 \phi_l'(x)\phi_j'(x)dx \\
&\qquad + b\Delta t \sum_{l=1}^{n-1} u_{i+1,l} \int_0^1 \phi_l(x)\phi_j(x)dx \\
&= \int_{t_i}^{t_{i+1}} \int_0^1 g(s,x)\phi_j(x)dxds + \int_{t_i}^{t_{i+1}} \int_0^1 \phi_j(x)dW_n(s,x)
\end{aligned} \tag{3.24}$$

for $j = 1, 2, \ldots, n-1, i = 1, 2, \ldots, I$ where $u_{i,l} \approx u_l(t_i)$. Let

$$u_n^h(t_i, x) = \sum_{l=1}^{n-1} u_{i,l}\phi_l(x).$$

For simplicity, we now focus on the case of using the continuous piecewise linear finite element in the spatial discretization. The following pathwise error estimate can be found in Theorem 8.2 of [17]:

$$\begin{aligned}
&\|u_n(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2} \\
&\le C\sqrt{1 + \log\frac{t_m}{\Delta t}} \left(\max_{i\le m} \int_{t_{i-1}}^{t_i} \left\| \frac{\partial u_n}{\partial t}(\tau,\cdot) \right\|_{L_2} d\tau + \max_{t\le t_m} h^2 \|u_n(t,\cdot)\|_{H^2} \right).
\end{aligned} \tag{3.25}$$

The following lemma gives estimates of the terms on the right-hand side of (3.25).

LEMMA 3.3. *Let u_n be the solution of* (3.15) *with $g \in C^2([0,T]\times[0,1])$, $u_0 \in C^2[0,1]$, and $\sigma_k^n(t)$ has the bound given in Theorem* 3.3. *Let the constant b be suitably small. Then, if $\delta t \le 1/(2|b|)$, the following inequalities hold for some constant c, independent of Δt and h:*

$$E \int_{t_{i-1}}^{t_i} \left\| \frac{\partial u_n}{\partial t}(\tau,\cdot) \right\|_{L_2} d\tau \le c \left((\Delta t)^2 + \Delta t \sum_k k^2 (\beta_k^n)^2 + \sum_k (\Delta t \beta_k^n)^2 \right)^{1/2} \tag{3.26}$$

and

$$E\|u_n(t,\cdot)\|_{H^2} \le c \left(1 + \frac{1}{\Delta t} \sum_k k^2 (\beta_k^n)^2 \right)^{1/2}. \tag{3.27}$$

The proof of Lemma 3.3 is given in the appendix.

Combining Lemma 3.3 and inequality (3.25), we have the following theorem.

THEOREM 3.4. *Assume that the conditions in Lemma* 3.3 *hold; then*

$$\begin{aligned}
E\|u_n(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2} &\le c \left(1 + \log\frac{t_m}{\Delta t} \right)^{1/2} \\
&\times \left((\Delta t)^2 + \Delta t \sum_k k^2 (\beta_k^n)^2 + \sum_k (\Delta t \beta_k^n)^2 + \frac{h^4}{\Delta t} \sum_k k^2 (\beta_k^n)^2 \right)^{1/2}
\end{aligned}$$

for some constant c. ∎

The error $E\|u(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2}$ can be obtained by applying the triangle inequality to the results of Theorems 3.3 and 3.4.

Remark 3.3. Note that when applied to the case of white noise, that is, $\sigma_k(t) = 1$ for all k, we may take $\beta_k^n = \sigma_k^n = 1$, $\alpha_k^n = 0$ for $k \le N$, and $\beta_k^n = \sigma_k^n = 0$, $\alpha_k^n = 1$ for $k > N$, where $N \to \infty$ as $n \to \infty$; then, after simplification, the estimates in the above theorems give

$$E\|u(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2} \le c\left(1 + \log\frac{t_m}{\Delta t}\right)^{1/2}\left\{\frac{1}{N^{1/2}} + (\Delta t)^{1/2}N^{3/2} + \frac{h^2N^{3/2}}{(\Delta)^{1/2}}\right\}$$

so that $h = O(\Delta t)^{1/2}$ and $N = O(h^{-1/2}) = O((\Delta t)^{-1/4})$ give a best order of $(\Delta t)^{1/8}$ or $h^{1/4}$, up to a logarithmic factor, for $E\|u(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2}$. This is indeed a very low order convergence estimate as was expected [1]. In the next section, however, we present a few examples with colored noises for which the above theorems allow much better estimates on the order of the approximations.

Remark 3.4. The estimate on the order of convergence in the time step size is seen to be at best $O(\sqrt{\Delta t})$, which is largely due to the fact that we restricted our attention to the case where $\{\dot\eta_k(t)\}$ in (3.16) correspond to the derivatives of the Wiener process with t being the parameter. In many physical applications, other processes may also be used [11]. One may also naturally consider more general formulation for the noise terms $\{\dot\eta_k(t)\}$ like what is used for dW/dx in (3.2). In the case where $\{\dot\eta_k\}$ are more regular in time, better error estimates may be obtained using similar techniques.

Discussions and extensions to higher space dimensions can be found in [36].

4. Numerical results for some model equations.

4.1. One dimensional elliptic equation. We now study two cases of the one dimensional elliptic equation with noise described in the previous section. We demonstrate that for different forms of coefficient $\{\sigma_k^n\}$, different rates of convergence are to be obtained.

Case 1. Let the random variables $\{\eta_k\}$ be iid, namely,

$$q_{kl} = E(\eta_k\eta_l) = \delta_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \ne l, \end{cases} \qquad \sigma_k = \frac{1}{k^{3/2}}, \quad \sigma_k^n = \begin{cases} \sigma_k, & k \le n, \\ 0, & k > n. \end{cases}$$

Then

$$\left\|\overrightarrow{\sigma^n} - \vec\sigma\right\|_{Q_{-1}} = \left(\sum_{k=n+1}^{\infty}\left(\frac{1}{k^{3/2}}\cdot\frac{1}{k}\right)^2\right)^{1/2} \le \frac{1}{n^2}.$$

From Lemma 3.1, we have, for some generic constant $C > 0$,

$$E\|\dot W_n\|_{L_2} \le C\left(\sum_{k=1}^{\infty}(\sigma_k^n)^2\right)^{1/2} \le C\left(\sum_{k=1}^{\infty}\frac{1}{k^3}\right)^{1/2} = C.$$

In other words, $\dot W_n \in L_2(0,1)$; this means that, in Theorem 3.2, $r = 1$. If the piecewise linear finite element basis is used, and $g \in L_2(0,1)$, the following error estimate yields

$$E(\|u - u_n^h\|_{L_2}) \le C(n^{-2} + h^2\|g + \dot W_n\|_{L_2}) \le C\,h^2\,.$$

Thus, asymptotically, we have a second order convergence rate in h for the expectation of the L^2 error.

Case 2. Now let us consider using different coefficients $\{\sigma_k^n\}$ which yield high order convergence results for high order finite element spaces. Still let

$$q_{kl} = E(\eta_k \eta_l) = \delta_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l, \end{cases} \qquad \sigma_k = \frac{1}{k^{7/2}}, \quad \sigma_k^n = \begin{cases} \sigma_k, & k \leq n, \\ 0, & k > n. \end{cases}$$

Then

$$\left\| \overrightarrow{\sigma^n} - \vec{\sigma} \right\|_{Q_{-1}} = \left(\sum_{k=n+1}^{\infty} \left(\frac{1}{k^{7/2}} \cdot \frac{1}{k} \right)^2 \right)^{1/2} \leq \frac{1}{n^4}.$$

From Lemma 3.1, we have

$$E\|\dot{W}_n\|_{H^2} \leq C \left(\sum_{k=1}^{\infty} (\sigma_k^n k^2)^2 \right)^{1/2} \leq C \left(\sum_{k=1}^{\infty} \left(\frac{1}{k^{7/2}} k^2 \right)^2 \right)^{1/2} = C.$$

In other words, $\dot{W}_n \in H^2(0,1)$; this means that, in Theorem 3.2, $r = 3$. If we use the cubic spline finite element basis, and assume that g is bounded in $H^2(0,1)$, the following error estimate yields

$$E(\|u - u_n^h\|_{L_2}) \leq C(n^{-4} + h^4 \|g + \dot{W}_n\|_{H^2}) \leq C\, h^4$$

for some constant C that depends only on g. Note that such a high order cannot be achieved if we have adopted a white noise [1].

The finite element method (3.10) is implemented for (3.1) with $g(x) = 2 + bx - bx^2$ and the noise $\dot{W}$ as defined in section 3. The exact solution of (3.1) is given by $u = u_d + u_s$, where u_d and u_s correspond to the deterministic and the stochastic parts. Moreover, $u_d(x) = x(1-x)$ and

$$u_s(x) = \sum_{k=1}^{\infty} \frac{\sqrt{2}\sigma_k}{b + (k\pi)^2} \eta_k \sin k\pi x \ .$$

The numerical solution is calculated for $n = 4, 8, 16, 32, 64, 128$ ($h = 1/n$ being the length of the subintervals). For each n, 10,000 runs are performed with different samples of the noise, $\|u - u_n^h\|_{L_2}$ is calculated for each sample, and the averaged value $E\|u - u_n^h\|_{L_2}$ is calculated.

For Case 1, we let $b = 0.5$, $\sigma_k = k^{-3/2}$, and we use the continuous piecewise linear finite element space. The left picture in Figure 4.1 gives the decay of error. The horizontal axis denotes $\log_{10} n$, and the vertical axis denotes $\log_{10} E\|u - u_n^h\|_{L_2}$. The slope of the error curve is nearly -2, in agreement with the theoretical result.

As for Case 2, we let $b = 0.5$, $\sigma_k = k^{-7/2}$, and we use the finite element space consisting of piecewise cubic splines. The right picture in Figure 4.1 gives the decay of error. The slope of the error curve is now nearly -4, also in agreement with the theoretical result.

4.2. Parabolic equation in one spatial dimension. Now consider a special case of parabolic equation described in the previous section. Let

$$\sigma_k(t) = \frac{\cos t}{k^3}, \qquad \sigma_k^n(t) = \begin{cases} \sigma_k(t), & k \leq n, \\ 0, & k > n, \end{cases}$$

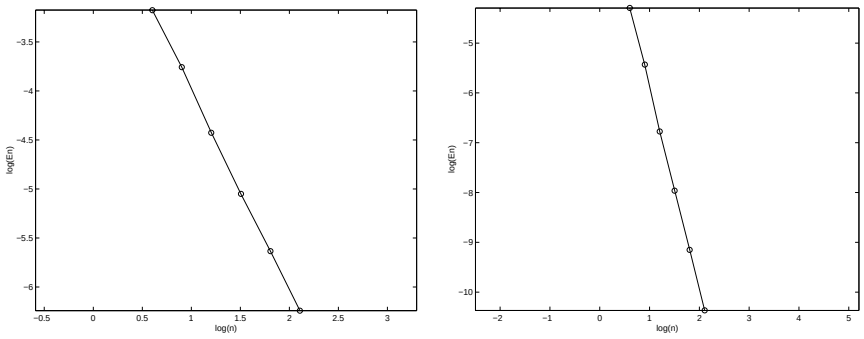


FIG. 4.1. *The error decay with* $\sigma_k = k^{-3/2}$ *and* $k^{-7/2}$.

and the upper bounds $\alpha_k^n, \beta_k^n, \gamma_k^n$ given in Theorem 3.3 can be chosen as

$$\alpha_k^n = \begin{cases} 0, & k \le n, \\ \frac{1}{k^3}, & k > n, \end{cases} \qquad \beta_k^n = \gamma_k^n = \frac{1}{k^3}.$$

Backward-Euler in time with the piecewise linear finite element in space approximation (3.24) was tested for the numerical solution of problem (3.14) with

$$g(t,x) = 10(1+b)x^2(1-x)^2e^t - 10(2-12x+12x^2)e^t \ .$$

We use $b = 0.5$ and $T = 1$. In the absence of noise term, the exact solution is

$$u(t,x) = u_d(t,x) = 10e^t x^2(1-x)^2 \ , \quad \text{with} \quad u_0(x) = 10x^2(1-x)^2 \ .$$

The exact value of $Eu(1, 0.5)$ is about 1.699.

In theory, using the above definitions, we have

$$\sum_{k=1}^{\infty} \frac{(\alpha_k^n)^2}{2(k\pi)^2} \le \sum_{k=n+1}^{\infty} \frac{1}{k^8} \le \frac{1}{n^7} = h^7, \tag{4.1}$$

$$\sum_{k=1}^{\infty} k^4(\beta_k^n)^2 + (\gamma_k^n)^2) \le \sum_{k=1}^{\infty} \left(\frac{1}{k^2} + \frac{1}{k^3} \right) \le C, \tag{4.2}$$

$$\sum_{k=1}^{\infty} (\beta_k^n)^2 \le \sum_{k=1}^{\infty} (k\beta_k^n)^2 \le C \ . \tag{4.3}$$

From Theorems 3.3 and 3.4, we have

$$E\|u - u_n\|_{L_2} \le c(h^7 + (\Delta t)^2)^{1/2},$$

$$E\|u_n(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2} \le c \left(1 + \log \frac{t_m}{\Delta t}\right)^{1/2} \left((\Delta t)^{1/2} + \frac{h^2}{(\Delta t)^{1/2}} \right).$$

Hence,

$$E\|u(t_m,\cdot) - u_n^h(t_m,\cdot)\|_{L_2} \le c \left(1 + \log \frac{t_m}{\Delta t}\right)^{1/2} \left((\Delta t)^{1/2} + \frac{h^2}{(\Delta t)^{1/2}} \right). \tag{4.4}$$

TABLE 4.1
$E(u_n^h(1,0.5))$ and $E(u_n^h(1,0.5))^2$ by the backward-Euler finite element scheme.

h	Δt	$E(u_n^h(1,0.5))$	$E(u_n^h(1,0.5))^2$	$E(\eta_{n/2,I})$	$var(\eta_{n/2,I})$
.25	.25	1.5268	2.3495	.0061	.9830
.25	.125	1.6147	2.6301	-.0217	1.0141
.25	.0625	1.6599	2.7826	-.0166	1.0079
.25	.03125	1.6821	2.8586	.0083	.9908
.25	.01563	1.6976	2.9142	-.0086	.9750
.125	.25	1.5198	2.3283	.0045	.9697
.125	.125	1.6071	2.6059	-.0014	1.0097
.125	.0625	1.6529	2.7569	-.0238	.9780
.125	.03125	1.6777	2.8432	.0002	.9829
.125	.01563	1.6912	2.8910	.0006	.9687
.0625	.25	1.5193	2.3263	-.0006	1.0182
.0625	.125	1.6043	2.5963	-.0069	.9886
.0625	.0625	1.6519	2.7539	.0124	.9852
.0625	.03125	1.6758	2.8372	-.0110	.9908
.0625	.01563	1.6888	2.8825	.0069	.9962
.03125	.25	1.5198	2.3277	-.0163	.9650
.03125	.125	1.6044	2.5971	-.0217	.9527
.03125	.0625	1.6503	2.7497	-.0071	.9984
.03125	.03125	1.6731	2.8281	.0044	.9765
.03125	.01563	1.6855	2.8724	-.0101	1.0479
.01563	.25	1.5181	2.3230	-.0166	.9918
.01563	.125	1.6067	2.6041	-.0134	.9667
.01563	.0625	1.6500	2.7482	-.0114	1.0067
.01563	.03125	1.6749	2.8336	-.0170	1.0365
.01563	.01563	1.6851	2.8704	-.0067	.9872

In the actual implementation, different values of Δt and h were used. For each pair $\{\Delta t, h\}$, 10,000 runs are performed with different sample of noise, and the ensemble averages are calculated. The numerical results of $E(u_n^h(1,0.5))$ and $E(u_n^h(1,0.5))^2$ are presented in Table 4.1.

The computational results converge as Δt and h approach to 0. From the table, it can be observed that, for fixed h, the results converge faster as Δt decreases, but for fixed Δt the convergence is less transparent as h decreases. This can be explained by the error estimate (4.4), which is bounded by $(\Delta t)^{1/2} + h^2(\Delta t)^{-1/2}$. If Δt and h are of the same order, the Δt term dominates in the estimate.

The numerical accuracy is also affected by the random number generators used in the different realizations. (The particular generator used in our implementation is obtained using MATLAB.) For comparison, the last two columns of Table 4.1 list the mean and variance of $\eta_{n/2,I}$. We see that, for the relatively larger magnitude of $E(\eta_{n/2,I})$, the error of $E(u_n^h(1,0.5))$ turns out to be larger as well.

Additional numerical examples can be found in [36].

5. Conclusion. In this paper, the numerical approximations of SDEs with different noise realizations are considered. In many instances of stochastic modeling, the noises may indeed be represented in various forms, with some emphasis on the correla-

tion in space and time, while others exhibit the correlation in frequency or spectrum. Our study indicates that the accuracy of the numerical approximation depends on the form of the underlying noise. Both rigorous error estimates and experimental results are provided in our paper.

Throughout our discussion, simple linear equations in one space dimension are used for the purpose of illustrations. We note that much of our consideration can be generalized to stochastic elliptic and parabolic equations in higher space dimensions. For the case of a simple two dimensional square domain, related discussions have been provided in [36]. By confining the theoretical analysis to the one space dimension here, some tedious technical details and complicated expressions are avoided.

Naturally, it will be very interesting to study the similar problems for nonlinear SDEs, which actually motivated the present investigation. It is hopeful that such studies may lead to a better understanding of the behaviors of the discretization error and the modeling error in conducting numerical simulations of nonlinear stochastic dynamics for practical problems [9, 18, 28].

Appendix.

Proof of Theorem 3.3.

Step 1. First, we verify the existence of such $\{\sigma_k^n(t)\}$. Since $\{\sigma_k'(t)\}$ are continuous on interval $[0,T]$, by the Weierstrass approximation theorem, for an arbitrary sequence α_k^n, where n is a fixed number, $k = 1, 2, \ldots$, there exists a sequence of polynomial $\{P_k^n(t)\}$ such that

$$|\sigma_k'(t) - P_k^n(t)| \le \frac{\alpha_k^n}{T} \ \forall t \in [0,T] \ .$$

Let

$$\sigma_k^n(t) = \int_0^t P_k^n(s)ds + \sigma_k(0),$$

and we have

$$|\sigma_k(t) - \sigma_k^n(t)| = \left| \int_0^t (\sigma_k'(s) - P_k^n(s))ds \right| \le \alpha_k^n.$$

By the triangle inequality,

$$|\sigma_k^n(t)| \le |\sigma_k(t)| + \alpha_k^n \le \beta_k + \alpha_k^n = \beta_k^n,$$

$$|\sigma_k^{n\prime}(t)| = |P_k^n(t)| \le |\sigma_k'(t)| + \frac{\alpha_k^n}{T} \le \gamma_k + \frac{\alpha_k^n}{T} = \gamma_k^n.$$

Step 2. Let $\bar{e}_n(t,x) = u(t,x) - u_n(t,x)$ and

$$F(t,x) = \int_0^t \int_0^1 G_{t-s}(x,y)dW(s,y) - \int_0^t \int_0^1 G_{t-s}(x,y)dW_n(s,y),$$

$$\bar{e}_n = E \int_0^T \int_0^1 \bar{e}_n^2(t,x)dxdt,$$

$$\bar{F}_n = E \int_0^T \int_0^1 F^2(t,x)dxdt.$$

Subtracting (3.18) from (3.15), and applying similar manipulation as that in section 3, we get

$$E\|u-u_n\|_{L_2}^2=\bar{e}_n\le\frac{\bar{F}_n}{(1-\bar{\lambda})^2}.$$

To estimate $\bar{F}_n$, we introduce an intermediate noise form

$$\frac{\partial^2\overline{W}_n}{\partial t\partial x}=\sum_{k=1}^{\infty}\sigma_k^n(t)\dot{\eta}_k(t)\psi_k(x),$$

that is, a noise discretized only in the x-direction. Let

$$F_1(t,x)=\int_0^t\int_0^1 G_{t-s}(x,y)dW(s,y)-\int_0^t\int_0^1 G_{t-s}(x,y)d\overline{W}_n(s,y),$$

$$F_2(t,x)=\int_0^t\int_0^1 G_{t-s}(x,y)d\overline{W}_n(s,y)-\int_0^t\int_0^1 G_{t-s}(x,y)dW_n(s,y);$$

then

$$F(t,x)=F_1(t,x)+F_2(t,x),$$

$$\bar{F}_n=E\int_0^T\int_0^1 F^2(t,x)dxdt\le 2\left(E\int_0^T\int_0^1 F_1^2(t,x)dxdt+E\int_0^T\int_0^1 F_2^2(t,x)dxdt\right).$$

Taking advantage of the orthogonality of $\{\sin k\pi x\}$ on the interval $[0,1]$, we have

$$F_1(t,x)=\sum_{k=1}^{\infty}\sqrt{2}\sin k\pi x e^{-(k\pi)^2t}\int_0^t(\sigma_k(s)-\sigma_k^n(s))e^{(k\pi)^2s}d\eta_k(s).$$

Since $\eta_k(t)$ is the standard Wiener process,

$$E\int_0^T\int_0^1 F_1^2(t,x)dxdt=\sum_{k=1}^{\infty}\int_0^T e^{-2(k\pi)^2t}\left(\int_0^t e^{2(k\pi)^2s}(\sigma_k(s)-\sigma_k^n(s))^2ds\right)dt$$

$$\le\sum_{k=1}^{\infty}(\alpha_k^n)^2\int_0^T e^{-2(k\pi)^2t}\left(\int_0^t e^{2(k\pi)^2s}ds\right)dt\le C_1\sum_{k=1}^{\infty}\frac{(\alpha_k^n)^2}{2(k\pi)^2}.$$

Using

$$\eta_{ki}=\frac{1}{\sqrt{\Delta t}}\int_{t_i}^{t_{i+1}}d\eta_k(t),$$

we have

$$F_2(t,x)=\sum_{k=1}^{\infty}\psi_k e^{-(k\pi)^2t}\left[\int_0^t e^{(k\pi)^2s}\sigma_k^n(s)d\eta_k(s)\right.$$

$$\left.-\int_0^t e^{(k\pi)^2s}\sigma_k^n(s)\sum_{i=1}^{I}\frac{1}{\sqrt{\Delta t}}\eta_{ki}\chi_i(s)ds\right]$$

$$=\sum_{k=1}^{\infty}\psi_k e^{-(k\pi)^2t}\left[\sum_{i=1}^{I_t-1}\int_{t_i}^{t_{i+1}}\left(e^{(k\pi)^2s}\sigma_k^n(s)-\frac{1}{\Delta t}\int_{t_i}^{t_{i+1}}e^{(k\pi)^2\tilde{s}}\sigma_k^n(\tilde{s})d\tilde{s}\right)d\eta_k(s)\right.$$

$$\left.+\int_{t_{I_t}}^{t}\left(e^{(k\pi)^2s}\sigma_k^n(s)-\frac{1}{t-t_{I_t}}\int_{t_{I_t}}^{t}e^{(k\pi)^2\tilde{s}}\sigma_k^n(\tilde{s})d\tilde{s}\right)d\eta_k(s)\right],$$

where I_t is the integer such that $t_{I_t} < t \le t_{I_t+1}$. Then

$$\begin{aligned}
&E\int_0^T\int_0^1 F_2^2(t,x)dxdt\\
&=\sum_{k=1}^{\infty}\int_0^T \frac{e^{-2(k\pi)^2t}}{(\Delta t)^2}\left(\sum_{i=1}^{I_t-1}\int_{t_i}^{t_{i+1}}\left(\int_{t_i}^{t_{i+1}}(e^{(k\pi)^2 s}\sigma_k^n(s)-e^{(k\pi)^2\tilde{s}}\sigma_k^n(\tilde{s}))d\tilde{s}\right)^2 ds\right.\\
&\quad\left.+\int_{t_{I_t}}^{t}(\frac{\Delta t}{t-t_{I_t}})^2\left(\int_{t_i}^{t_{i+1}}(e^{(k\pi)^2 s}\sigma_k^n(s)-e^{(k\pi)^2\tilde{s}}\sigma_k^n(\tilde{s}))d\tilde{s}\right)^2 ds\right)dt.
\end{aligned}$$

For s, $\tilde{s}\in[t_i,t_{i+1}]$, using the smoothness assumption on $\sigma_k(t)$, we get

$$\begin{aligned}
&|e^{(k\pi)^2 s}\sigma_k^n(s)-e^{(k\pi)^2\tilde{s}}\sigma_k^n(\tilde{s})|\\
&\le |e^{(k\pi)^2 s}-e^{(k\pi)^2\tilde{s}}|\sigma_k^n(s)+e^{(k\pi)^2\tilde{s}}|\sigma_k^n(s)-\sigma_k^n(\tilde{s})|\\
&\le (k\pi)^2 e^{(k\pi)^2 t_{i+1}}\sigma_k^n(s)\Delta t+e^{(k\pi)^2 t_{i+1}}\sigma_k^{n\prime}(\xi_i)\Delta t\\
&\le e^{(k\pi)^2 t_{i+1}}((k\pi)^2\beta_k^n+\gamma_k^n)\Delta t.
\end{aligned}$$

Here, $t_i\le\xi_i\le t_{i+1}$. Without loss of generality, we assume $t=t_{I_t+1}$; then

$$\begin{aligned}
&E\int_0^T\int_0^1 F_2^2(t,x)dxdt\\
&\le\sum_{k=1}^{\infty}\int_0^T e^{-2(k\pi)^2t}\left[\sum_{i=1}^{I_t}\int_{t_i}^{t_{i+1}}\frac{1}{(\Delta t)^2}(e^{(k\pi)^2 t_{i+1}}((k\pi)^2\beta_k^n+\gamma_k^n)(\Delta t)^2)^2 ds\right]dt\\
&\le C\sum_{k=1}^{\infty}\int_0^T e^{-2(k\pi)^2t}\left[\sum_{i=1}^{I_t}\int_{t_i}^{t_{i+1}}(e^{2(k\pi)^2 t_{i+1}}((k\pi)^4(\beta_k^n)^2+(\gamma_k^n)^2)(\Delta t)^2 ds\right]dt\\
&\le C\sum_{k=1}^{\infty}\sum_{i=1}^{I_t}\int_0^T e^{-2(k\pi)^2t}e^{2(k\pi)^2 t_{i+1}}dt(k^4(\beta_k^n)^2+(\gamma_k^n)^2)(\Delta t)^3\\
&\le C\sum_{k=1}^{\infty}\sum_{i=1}^{I_t}(k^4(\beta_k^n)^2+(\gamma_k^n)^2)(\Delta t)^3\\
&\le C\sum_{k=1}^{\infty}(k^4(\beta_k^n)^2+(\gamma_k^n)^2)(\Delta t)^2.
\end{aligned}$$

The last inequality comes from $\sum_{i=1}^{I_t},1\le I=1/\Delta t$. The theorem is now proved. □

Proof of Lemma 3.3. In general, by applying the same technique as in the proof of Theorem 3.1, we may first estimate

$$\begin{aligned}
E\int_0^t\int_0^1 u_n^2(t,x)dxdt&\le\frac{c}{1-\bar{\lambda}}E\int_0^t\int_0^1[(G_t(x,y)u_0(y))^2+(G_{t-s}(x,y)g(s,y))^2]dyds\\
&\quad+\frac{c}{1-\bar{\lambda}}E\int_0^t\int_0^1\left(\int_0^t\int_0^1 G_{t-s}(x,y)dW_n(s,y)\right)^2 dxdt.
\end{aligned}$$

Next, one may differentiate (3.18) to get

$$\frac{\partial u_n}{\partial t}(t,x) = -\int_0^t\int_0^1 \frac{\partial}{\partial t}G_{t-s}(x,y)bu_n(s,y)dyds + \int_0^1 \frac{\partial}{\partial t}G_t(x,y)u_0(y)dy + \int_0^t\int_0^1 \frac{\partial}{\partial t}G_{t-s}(x,y)g(s,y)dyds + \int_0^t\int_0^1 \frac{\partial}{\partial t}G_{t-s}(x,y)dW_n(s,y).$$

Then one may estimate $E\int_{t_{i-1}}^{t_i}\int_0^1(\frac{\partial u_n}{\partial t}(t,x))^2dxdt$ using the above equation. Similarly, one may estimate $E\int_0^1(\frac{\partial^2 u_n}{\partial x^2}(t,x))^2dxdt$.

Since we have assumed that b is a constant, we now provide a simpler estimate which, in spirit, is similar to the estimate derived from the integral formulation.

Let $g(t,x)=\sum_k g_k(t)\psi_k(x)$, $u_n(t,x)=\sum_k u_k^{(n)}(t)\psi_k(x)$, $u_0(x)=\sum_k u_k\psi_k(x)$; then

$$\frac{\partial}{\partial t}u_k^{(n)}(t) + (k^2\pi^2+b)u_k^{(n)}(t) = g_k(t) + \frac{\sigma_k^n(t)}{\sqrt{\Delta t}}\sum_i \eta_{ki}\chi_i(t).$$

Thus, for $t\in[t_{i-1},t_i)$,

$$u_k^{(n)}(t) = e^{-((k\pi)^2+b)t}u_k + \int_0^t e^{-((k\pi)^2+b)(t-s)}g_k(s)ds + \sum_{j=1}^{i}\int_0^t e^{-((k\pi)^2+b)(t-s)}\frac{\sigma_k^n(s)}{\sqrt{\Delta t}}\eta_{kj}\chi_j(s)ds.$$

This leads to

$$u_k^{(n)}(t) = e^{-((k\pi)^2+b)t}u_k + \int_0^t e^{-((k\pi)^2+b)(t-s)}g_k(s)ds + \sum_{j=1}^{i}\frac{\eta_{kj}}{\sqrt{\Delta t}}\int_{t_{j-1}}^{t_j^*} e^{-((k\pi)^2+b)(t-s)}\sigma_k^n(s)ds,$$

where $t_l^*=t_l$ for $l<i$ and $t_i^*=t$. It follows that

$$E\left[u_k^{(n)}(t)\right]^2 \le cu_k^2e^{-2((k\pi)^2+b)t} + cT\int_0^t e^{-2((k\pi)^2+b)(t-s)}g_k^2(s)ds + \frac{c}{\Delta t}\sum_{j=1}^{i}\left(\int_{t_{j-1}}^{t_j^*} e^{-((k\pi)^2+b)(t-s)}\sigma_k^n(s)ds\right)^2$$

for some constant c.

Since $u_0\in C^2[0,1]$ and $g\in C^2([0,T]\times[0,1])$, we have, for some constant $c>0$,

$$\sum_k (k\pi)^4\left\{u_k^2e^{-2((k\pi)^2+b)t} + \int_0^t e^{-2((k\pi)^2+b)(t-s)}g_k^2(s)ds\right\} \le c.$$

Using the bounds on σ_k^n and the fact that b is a small constant, we have

$$\begin{aligned}\sum_{l=1}^{i}\left(\int_{t_{l-1}}^{t_l^*} e^{-((k\pi)^2+b)(t-s)}\sigma_k^n(s)ds\right)^2 &\le c(\beta_k^n)^2\int_0^t e^{-2((k\pi)^2+b)(t-s)}ds\\ &\le \frac{c(\beta_k^n)^2}{(k\pi)^2+b}.\end{aligned}$$

Thus, for small b, we have

$$\begin{aligned}E\|u_n(\tau,\cdot)\|_{H^2} &\le (\,E\|u_n(\tau,\cdot)\|_{H^2}^2)^{1/2}\\ &\le c\left(1+\frac{1}{\Delta t}\sum_k\, k^2(\beta_k^n)^2\right)^{1/2}\end{aligned}$$

for some constant $c>0$. This proves the inequality (3.27).

For (3.26), we have

$$\frac{\partial}{\partial t}u_k^{(n)}(t) = -((k\pi)^2+b)u_k^{(n)}(t)+g_k(t)+\frac{\sigma_k^n(t)}{\sqrt{\Delta t}}\sum_i \eta_{ki}\chi_i(t).$$

So,

$$E\int_{t_{i-1}}^{t_i}\left(\frac{\partial}{\partial t}u_k^{(n)}(s)\right)^2 ds \le cE\int_{t_{i-1}}^{t_i}((k^2\pi^2+b)u_k^{(n)}(s))^2ds + c\int_{t_{i-1}}^{t_i}[g_k^2(s)+(\sigma_k^n(s))^2]ds.$$

Thus,

$$\begin{aligned}E\int_{t_{i-1}}^{t_i}\left\|\frac{\partial u_n}{\partial t}(\tau,\cdot)\right\|_{L_2} d\tau &\le \left(\Delta tE\int_{t_{i-1}}^{t_i}\left\|\frac{\partial u_n}{\partial t}(\tau,\cdot)\right\|_{L_2}^2 d\tau\right)^{1/2}\\ &\le c\left((\Delta t)^2+\Delta t\sum_k\, k^2(\beta_k^n)^2+\sum_k(\Delta t\beta_k^n)^2\right)^{1/2}.\end{aligned}$$

This proves (3.26). □

Acknowledgments. The authors would like to thank Weinan E, Max Gunzburger, and Zhimin Zhang for interesting discussions. The authors also want to thank them and an anonymous referee for providing useful references.

REFERENCES

[1] E. Allen, S. Novosel, and Z. Zhang, *Finite element and difference approximation of some linear stochastic partial differential equations*, Stochastics Stochastics Rep., 64 (1998), pp. 117–142.

[2] J.F. Bennaton, *Discrete time Galerkin approximation to the nonlinear filtering solution*, J. Math. Anal. Appl., 110 (1985), pp. 364–383.

[3] R. Buckdahn and E. Pardoux, *Monotonicity methods for white noise driven quasilinear SPDEs*, in Diffusion Processes and Related Problems in Analysis, I, M. Pinsky, ed., Birkhaüser Boston, Boston, MA, 1990, pp. 219–233.

[4] P.L. Chow, J.L. Jiang, and J.L. Menaldi, *Pathwise convergence of approximation solutions to Zakai's equation in a bounded domain*, in Stochastic Partial Differential Equations and Applications, G. Da Prato and L. Tubaro, eds., Longman Scientific and Technical, Harlow, UK, 1992, pp. 111–123.
[5] G. Da Prato and L. Tubaro, *Stochastic Partial Differential Equations and Applications*, Longman Scientific and Technical, Harlow, UK, 1992.
[6] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
[7] A. Davie and J. Gaines, *Convergence of numerical schemes for the solution of the parabolic stochastic partial differential equations*, Math. Comp., 70 (2001), pp. 121–134.
[8] D.A. Dawson, *Stochastic evolution equations*, Math. Biosci., 154 (1972), pp. 187–316.
[9] J. Deang, Q. Du, and M. Gunzburger, *Thermal fluctuations of superconducting vortices*, Phy. Rev. B, 64 (2001), pp. 52506–52510.
[10] A. Etheridge, *Stochastic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1995.
[11] R. Fox, *Second order algorithm for the numerical integration of colored noise problems*, Phys. Rev. A(3), 43 (1991), pp. 2649–2654.
[12] J.G. Gaines, *Numerical Experiments with S(P)DE's*, in Stochastic Partial Differential Equations, London Math. Soc. Lecture Note Ser. 216, Cambridge University Press, Cambridge, UK, 1995, pp. 55–71.
[13] I. Gyongy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise* II, Potential Anal., 11 (1999), pp. 1–37.
[14] E. Hausenblas, *Error analysis for approximation of stochastic differential equations driven by Poisson random measures*, SIAM J. Numer. Anal., 40 (2002), pp. 87–113.
[15] D.J. Higham, *Mean-square and asymptotic stability of the stochastic theta method*, SIAM J. Numer Anal, 38 (2000), pp. 753–769.
[16] M. Ibanes, J. Garcia-Ojalvo, R. Toral, and J.M. Sancho, *Noise-induced phase separation: Mean-field results*, Phys. Rev. E(3), 60 (1999), pp. 3597–3605.
[17] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1994.
[18] T. Kamppeter, F.G. Mertens, E. Moro, A. Sanchez, and A.R. Bishop, *Stochastic vortex dynamics in two-dimensional easy-plane ferromagnets: Multiplicative versus additive noise*, Phys. Rev. B, 59 (1999), pp. 11349–11357.
[19] J.B. Keller, *Stochastic equations and wave propagation in random media*, Proc. Sympos. Appl. Math., 16 (1964), pp. 145–170.
[20] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, New York, 1992.
[21] L. Machiels, and M.O. Deville, *Numerical simulation of randomly forced turbulent flows*, J. Comput. Phys., 145 (1998), pp. 246–279.
[22] A. Majda, I. Timofeyev, and E. Eijnden, *Models for stochastic climate prediction*, Proc. Natl. Acad. Sci. USA, 96 (1996), pp. 14687–14691.
[23] G.N. Milstein, E. Platen, and H. Schurz, *Balanced implicit methods for stiff stochastic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1010–1019.
[24] E.A. Novikov, *Functionals and the random-force method in turbulence theory*, Soviet Phys. JETP, 20 (1965), pp. 1290–1294.
[25] E. Platen, *An Introduction to Numerical Methods for Stochastic Differential Equations*, Acta Numer. 8, Cambridge University Press, Cambridge, UK, 1999, pp. 197–246.
[26] P. Protter and D. Talay, *The Euler scheme for Levy driven stochastic differential equations*, Ann. Probab., 25 (1997), pp. 393–423.
[27] Y. Saito and T. Mitsui, *Stability analysis of numerical schemes for stochastic differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2254–2267.
[28] J. Sancho, J. Garcia-Ojalvo, and H. Guo, *Non-equilibrium Ginzburg-Landau model driven by colored noise*, Phys. D., 113 (1998), pp. 331–337.
[29] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numer. Funct. Anal. Optim., 20 (1999), pp. 121–145.
[30] G. Strang, *Wavelets and dilation equations: A brief introduction*, SIAM Rev., 31 (1989), pp. 614–627.
[31] D. Talay, *Simulation and numerical analysis of stochastic differential systems*, in Effective Stochastic Analysis, P. Krée and W. Wedig, eds., Springer-Verlag, Berlin, 1988.
[32] R.T. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser Boston, Boston, MA, 1997.
[33] J.B. Walsh, *An Introduction to Stochastic Partial Differential Equations*, Lecture Notes in

Math. 1180, Springer-Verlag, Berlin, 1986, pp. 265–439.
[34] M. Werner and P. Drummond, *Robust algorithms for solving stochastic partial differential equations*, J. Comput. Phys., 132 (1997), pp. 312–326.
[35] H. Yoo, *Semi-discretization of stochastic partial differential equations on* $\mathbf{R}^1$ *by a finite-difference method*, Math. Comp., 69 (2000), pp. 653–666.
[36] T. Zhang, *Numerical Approximations of Stochastic Partial Differential Equations*, M. Phil thesis, Hong Kong University of Science and Technology, Hong Kong, 2000.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1446–1474

DYNAMIC ITERATION USING REDUCED ORDER MODELS: A METHOD FOR SIMULATION OF LARGE SCALE MODULAR SYSTEMS*

MURUHAN RATHINAM† AND LINDA R. PETZOLD†

Abstract. We describe a new iterative method, dynamic iteration using reduced order models (DIRM), for simulation of large scale modular systems using reduced order models that preserve the interconnection structure. This method may be compared to the waveform relaxation technique; however, unlike DIRM, waveform relaxation does not take advantage of model reduction techniques. The DIRM method involves simulating in turn each subsystem connected to model reduced versions of the other subsystems. The data from this simulation is then used to update the reduced model for that particular subsystem. We provide analytical results on convergence and accuracy of the DIRM method as well as numerical examples that demonstrate the success of DIRM and verify the analysis.

Key words. large scale systems, model reduction, dynamic iteration, proper orthogonal decomposition

AMS subject classifications. 65L99, 65M20

PII. S0036142901390494

1. Introduction.

Very large scale systems of differential and differential algebraic equations such as the U.S. power grid, very large scale integrated (VLSI) circuits, chemical reactors, and weather systems present challenges in computing. Usually such large scale systems consist of many interacting subsystems, which may in some problems be governed by very different physical laws. On such systems conventional methods of direct numerical integration of the full system may not be feasible without massive computing resources.

An iterative approach known as *waveform relaxation* (WR), where smaller subsystems are simulated separately and then the couplings are accounted for through iteration, was developed by researchers for the simulation of VLSI circuits. See Lelarasmee, Ruehli, and Sangiovanni-Vincentelli [7], Miekkala and Nevanlinna [9], and Miekkala [8] for details. The WR method is a form of *dynamic iteration* in the sense that the variable being iterated is a function (the entire solution waveform for a given time interval) and not a vector. This method has subsequently been applied by researchers to PDEs of parabolic and hyperbolic types [2]. Such a modular approach in principle has the advantage that it facilitates parallel computation, exploits the multirate nature of some problems, and offers the potential of using different numerical techniques for different subsystems. However, the WR technique has not become the mainstay in application areas. This is primarily due to the poor convergence properties of WR.

Another way to deal with models that are too complex is via model reduction. Several model reduction techniques have been studied by researchers in various fields. *Balanced truncation* has been studied by the control community (see Zhou and Doyle [15] and Lall, Marsden, and Glavaski [6], for instance), *proper orthogonal decomposition* (POD) has been applied in the study of turbulence (see Holmes, Lumley, and

*Received by the editors June 7, 2001; accepted for publication (in revised form) March 28, 2002; published electronically October 23, 2002. A preliminary version of this article appeared in *Proceedings of the IEEE Control and Decision Conference*, Sydney, Australia, 2000. This research was supported in part by grants EPRI WO-8333-06, NSF/KDI ATM-9873133, and NSF ACI-0086061.
http://www.siam.org/journals/sinum/40-4/39049.html

†Computational Science and Engineering, University of California, Santa Barbara, CA 93106 (muruhan@engineering.ucsb.edu, petzold@engineering.ucsb.edu).

Berkooz [5]), cascading failures in power grids (Parrilo et al. [11]), and control of compressors (Glavaski, Marsden, and Murray [3]), etc., and *selective modal analysis* has been developed by researchers in the electrical power field (Perèz-Arriaga et al. [12]), to name a few.

In this paper we present a method that combines the idea of dynamic iteration with the use of reduced order models. Our method also seeks to remedy some of the shortcomings of WR. Our approach, termed dynamic iteration using reduced order models (DIRM), involves simulation of each subsystem in turn while it is connected to reduced order models of the rest of the subsystems. The simulation results are then used to update the reduced order model for that particular subsystem. If the reduced order models are small enough, then the combination of an unreduced subsystem with the rest of the reduced subsystems results in a system small enough not to pose insurmountable computational difficulties. In principle any model reduction method that uses data from trajectories could be used in this iteration. In this paper we use POD (also known as *Karhunen–Loève decomposition*) for the model reduction.

Even though theoretically the WR method has good asymptotic convergence, in practice there may be large initial overheads. For example, consider a one-dimensional (1D) PDE with the spatial domain divided into 10 subsystems of adjacent regions. It will take nine iterations before the first subsystem "sees" the last subsystem. In the DIRM method, by contrast, every subsystem is connected to all other (reduced versions of) subsystems, and one may not expect such overheads. This is possible only because of the fact that DIRM incorporates reduced order models.

This paper is organized as follows. In section 2 we review the POD method of model reduction and comment on its application to modular systems. In section 3 we describe DIRM in detail and also provide a brief account of the WR technique. In section 4 we provide an analysis of the DIRM method as applied to a linear time invariant system consisting of two subsystems and present results on the accuracy and convergence behavior of DIRM. Section 5 describes several numerical examples. These include a nonlinear power grid simulation and some reaction diffusion problems described by PDEs, with comparison to WR. We also give some examples highlighting certain special cases, which include situations where DIRM has difficulty converging as predicted by the analysis, and show how to modify DIRM to fix this problem. Finally, in section 6 we present conclusions and discuss future research.

2. Model reduction using POD. The POD technique for model reduction consists of first finding a subspace in the full phase space of a given dynamical system and then constructing an approximating dynamical system in that subspace. The original dynamical system may be nonlinear, and in that case the resulting lower dimensional model will also typically be nonlinear.

2.1. POD. POD, also known as Karhunen–Loève decomposition or principal component analysis, provides a method for finding the best approximating subspace to a given set of data. Originally POD was used as a data representation technique. For model reduction of dynamical systems POD may be used on data points obtained from system trajectories obtained via experiments, numerical simulations, or analytical derivations. For more information see Rathinam and Petzold [13], Holmes, Lumley, and Berkooz [5], Moore [10], Lall, Marsden, and Glavaski [6], Glavaski, Marsden, and Murray [3], and references therein.

Given a set of data points $x^{(\alpha)} \in \mathbb{R}^n$, POD seeks a subspace $S \subset \mathbb{R}^n$ so that the total square distance

$$D = \sum_{\alpha=1}^{N} \left\| x^{(\alpha)} - \rho_S x^{(\alpha)} \right\|^2,$$

where ρ_S is the orthogonal projection onto the subspace S, is minimized. The norm considered is the 2-norm. (Thus we assume that the phase space comes equipped with a notion of inner product.) The solution to this problem may be stated in terms of the *correlation matrix* defined by

$$R = \sum_{\alpha=1}^{N} x^{(\alpha)} \left(x^{(\alpha)} \right)^T.$$

Note that R is $n \times n$ and symmetric positive semidefinite. Let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n \geq 0$ be the ordered eigenvalues of R. Then the minimum value of D over all $k(\leq n)$ dimensional subspaces S is given by $\sum_{j=k+1}^{n} \lambda_j$ [5]. In addition, the S that minimizes D is the invariant subspace corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$. In practice one need not compute R. Instead, it is efficient to use the $n \times N$ matrix X whose columns are $x^{(\alpha)}$. Then $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ are the singular values of X (assuming $n \leq N$), and S is the span of the left singular vectors of X corresponding to k the largest singular values. Note that $R = XX^T$.

Often it may be best to find an affine subspace as opposed to a linear subspace. This requires first to find the mean value of the data points

$$\bar{x} = \frac{1}{N} \sum_{\alpha=1}^{N} x^{(\alpha)}$$

and then construct the *covariance matrix* $\bar{R}$ given by

$$\bar{R} = \sum_{\alpha=1}^{N} \left(x^{(\alpha)} - \bar{x} \right) \left(x^{(\alpha)} - \bar{x} \right)^T.$$

Let S_0 be the invariant subspace of the k largest eigenvalues of $\bar{R}$. Then the best approximating affine subspace S passes through $\bar{x}$ and is obtained by shifting S_0 by $\bar{x}$. Algebraically the projection onto the subspace S is given by

$$z = \rho(x - \bar{x}), \tag{2.1}$$

where $z \in \mathbb{R}^k$ are coordinates in the subspace S, $x \in \mathbb{R}^n$ are coordinates in the original coordinate system in $\mathbb{R}^n$, and the matrix ρ of the projection consists of row vectors ϕ_i^T $(i = 1, \ldots, k)$, where ϕ_i are the unit eigenvectors corresponding to the largest k eigenvalues of $\bar{R}$. Note that given any point $p \in S$ with coordinates $z \in \mathbb{R}^k$ the coordinates $x \in \mathbb{R}^n$ of the same point in the original coordinate system are given by

$$x = \rho^T z + \bar{x}.$$

2.2. Galerkin projection. Having found the approximating subspace for our system data, our next task is to construct a vector-field on this subspace that represents the reduced order model. This procedure is known as Galerkin projection and has been widely used in reducing PDEs to ODEs by projecting onto appropriate basis functions that describe the spatial variations in the solution. The procedure is applicable to any subspace; the subspace need not be obtained from the POD method. See [5] for more details.

Suppose the original dynamical system in $\mathbb{R}^n$ is given by a vector-field f,

$$\dot{x} = f(x,t).$$

Let $S \subset \mathbb{R}^n$ be the best k dimensional approximating affine subspace with projection given by (2.1). A vector-field f_a in the subspace S is constructed by the following rule: for any point $p \in S$ compute the vector-field $f(p,t)$ and take the projection $\rho f(p,t)$ onto the subspace S to be the value of $f_a(p,t)$. If z are the subspace coordinates of p, then $f_a(z,t) = \rho f(\rho^T z + \bar{x}, t)$. Thus we obtain the following reduced model:

$$\dot{z} = f_a(z,t) = \rho f(\rho^T z + \bar{x}, t). \tag{2.2}$$

If we are solving an initial value problem with $x(0) = x_0$, then in the reduced model one has the initial condition $z(0) = z_0$, where

$$z_0 = \rho(x_0 - \bar{x}).$$

Hence the approximating solution $\hat{x}(t)$ in the original coordinates in $\mathbb{R}^n$ is given by

$$\hat{x}(t) = \rho^T z(t) + \bar{x}.$$

From the above it is easy to see that the approximating solution $\hat{x}(t)$ is the solution to the following initial value problem:

$$\dot{\hat{x}} = Pf(\hat{x},t), \qquad \hat{x}(0) = \hat{x}_0 = P(x_0 - \bar{x}) + \bar{x}, \tag{2.3}$$

where $P = \rho^T \rho \in \mathbb{R}^{n \times n}$ is the matrix of the projection expressed in the original coordinate system in $\mathbb{R}^n$. Also note that $\hat{x}_0$ is just the projection of x_0 onto the affine subspace S.

2.3. Modular model reduction. In this paper *modular system* shall mean any system expressed in the form

$$\dot{x}_i = f_i(x_1, \ldots, x_m, t), \qquad i = 1, \ldots, m. \tag{2.4}$$

Note that any system $\dot{x} = f(x,t)$ can be written in this form. All that is involved is a partitioning of the states $x = (x_1, \ldots, x_m)$, where $x_i \in \mathbb{R}^{n_i}$ are vectors. This partitioning may arise naturally from the physical interpretation of the system, as in the power grid example presented later, or may be introduced according to some optimal criteria for the simulation problem at hand. In this paper we consider situations where the overall system is very large; hence we modularize the system by breaking it into manageable smaller parts. The POD method can be made to "respect" the partitioning by forming separate covariance matrices for each of the subsystem states $x_i \in \mathbb{R}^{n_i}$,

$$\bar{R}_i = \sum_{\alpha=1}^{N} \left(x_i^{(\alpha)} - \bar{x}_i \right) \left(x_i^{(\alpha)} - \bar{x}_i \right)^T,$$

and computing separate projections $\rho_i \in \mathbb{R}^{k_i \times n_i}$ that operate within the state space of a subsystem. Thus the reduced model will be

$$\dot{z}_i = \rho_i f_i(\rho_1^T z_1 + \bar{x}_1, \ldots, \rho_m^T z_m + \bar{x}_m, t), \qquad i = 1, \ldots, m.$$

3. DIRM. In this section we describe the DIRM method of simulating a large scale modular system of the form (2.4). We first describe the WR method in order to put our method in context.

The basic idea behind WR as applied to system (2.4) may be explained as follows. Start with an initial approximation for solutions of each of the subsystem trajectories: $x_1^{(0)}(t), \ldots, x_m^{(0)}(t)$. At the kth iteration, simulate each subsystem separately:

$$\dot{x}_i^{(k)} = f_i(x_1^{(k-1)}(t), \ldots, x_{i-1}^{(k-1)}(t), x_i^{(k)}, x_{i+1}^{(k-1)}(t), \ldots, x_m^{(k-1)}(t), t), \quad i = 1, \ldots, m.$$

This is a simplified explanation of the method. For a detailed exposition and analysis we refer to [7], [9], and [8]. It has been shown that this iteration converges for ODE systems in finite interval simulations under some mild conditions. However, WR may suffer from slow convergence. Overlapping techniques are often used to speed up the convergence [8].

The DIRM method also simulates each subsystem in turn, but not in isolation. Instead, the unreduced model of the subsystem is connected to reduced order models of the other subsystems. If the reduced order models are small enough, then the overall size of the resulting system is still of manageable dimensions. Consider the modular system (2.4) with initial conditions $x_i(0) = x_{i,0}$ and suppose we are interested in a simulation interval $[0, T]$. The DIRM method is described as follows.

Start with some initial reduced model for each subsystem. In the POD approach a reduced model for subsystem i is characterized by the projection matrix ρ_i and the mean data value $\bar{x}_i$. Let the initial reduced models be $(\rho_i^{(0)}, \bar{x}_i^{(0)})$. One way to generate these is to simulate each subsystem in isolation (in the given interval), setting the states of the other subsystems to some constant values, for instance, the initial conditions. In other words, simulate the following equations:

$$\dot{x}_i = f_i(x_{1,0}, x_{2,0}, \ldots, x_{i-1,0}, x_i, x_{i+1,0}, \ldots, x_{m,0}, t), \qquad i = 1, \ldots, m,$$

with initial conditions $x_i = x_{i,0}$. The resulting solutions $x_i(t)$ may be used to compute the covariance matrices $\bar{R}_i$:

$$\begin{aligned} \bar{x}_i &= \frac{1}{T} \int_0^T x_i(t) dt, \\ \bar{R}_i &= \int_0^T (x_i(t) - \bar{x}_i)(x_i(t) - \bar{x}_i)^T dt. \end{aligned} \tag{3.1}$$

At the jth step in the iteration we have the reduced models from the previous step $(\rho_i^{(j-1)}, \bar{x}_i^{(j-1)})$. We also have the trajectories $x_i^{(j-1)}(t), t \in [0, T]$, which were used in constructing these reduced models. Now for $i = 1, \ldots, m$ connect the unreduced subsystem i with the reduced versions of all other subsystems and simulate the resulting system

$$\begin{aligned} \dot{x}_i &= f_i(X, t), \\ \dot{z}_l &= \rho_l^{(j-1)} f_l(X, t), \qquad l = 1, \ldots, i-1, i+1, \ldots, m, \end{aligned} \tag{3.2}$$

where X is the following list of vector arguments:

$$X = \left(\rho_1^{(j-1)}\right)^T z_1 + \bar{x}_1^{(j-1)}, \ldots, \left(\rho_{i-1}^{(j-1)}\right)^T z_{i-1} + \bar{x}_{i-1}^{(j-1)}, \\ x_i, \left(\rho_{i+1}^{(j-1)}\right)^T z_{i+1} + \bar{x}_{i+1}^{(j-1)}, \ldots, \left(\rho_m^{(j-1)}\right)^T z_m + \bar{x}_m^{(j-1)}.$$

Use the resulting trajectory for the ith subsystem $x_i^{(j)}(t)$ to compute an updated reduced order model for the ith subsystem $(\rho_i^{(j)}, \bar{x}_i^{(j)})$. The iteration is terminated when

$$\sup_{t\in[0,T]} \left\{ \|x_i^{(j)}(t) - x_i^{(j-1)}(t)\| \right\} \leq tol, \qquad i = 1, \ldots, m, \tag{3.3}$$

where *tol* is some specified tolerance.

Remark 3.1. In (3.2) the trajectories $z_l^{(j)}(t)$ correspond to reduced models, while the trajectory $x_i^{(j)}(t)$ corresponds to the full model. In situations when the coupling between subsystems is "weak," $x_i^{(j)}(t)$ will be more accurate than $z_l^{(j)}(t)$. Since the reduced models are computed directly from $x_i^{(j)}(t)$, the simulation for the next iteration $j+1$ is directly affected by $x_i^{(j)}(t)$ and only indirectly by $z_l^{(j)}(t)$. This helps keep the effect of errors due to model reduction small. Also note that the final solution comes directly from $x_i^{(j)}(t)$, and the $z_l^{(j)}(t)$ enter only indirectly.

For any technique involving reduced order models, accuracy is an important issue. Since reduced order models are computed from the trajectories obtained from the given initial value problem, when the coupling dynamics is not very strong the situation for DIRM is reasonably close to the circumstances under which the accuracy of the POD method could be expected to be as good as possible as indicated by the error analysis of POD in [13].

We have observed from various examples, linear and nonlinear, that the DIRM method generally converges. We have also found examples where it fails to converge, but on those occasions breaking up the time interval $[0, T]$ into smaller ones $[t_i, t_{i+1}]$, $i = 0, \ldots M-1$, where $t_0 = 0$ and $t_M = T$ and running the algorithm successively in each interval achieves convergence. It is known that WR also converges better when the interval length is smaller. However, of course there is an optimal length beyond which making the intervals smaller results in higher computational effort.

Remark 3.2. The method of model reduction we use in this paper is POD, but in the overall iteration of DIRM one could in principle replace POD with any model reduction scheme that depends on simulation data. (Methods such as balanced truncation in their original form cannot be used, since they depend only on the model and not on a given set of system trajectories.)

4. Analysis for linear time invariant systems with two subsystems. The iteration operator associated with DIRM is nonlinear even if the system of ODEs is linear. This significantly complicates the convergence analysis of DIRM. In this section we provide an analysis of the DIRM method for linear time invariant systems consisting of two subsystems. We also assume that in the model reduction via POD we fit the best approximating linear subspace instead of the more general method of fitting the best approximating affine subspace. Although these assumptions are somewhat restrictive, the purpose of the analysis is to provide qualitative results rather than sharp estimates of convergence rates.

4.1. Description of Jacobi DIRM iteration operator for two subsystems. Throughout the rest of section 4 we will be concerned with the case of two subsystems each of dimension n unless stated otherwise. Suppose that the system consists of states $x = (x_1, x_2)$ with $x_i \in \mathbb{R}^n$ for $i = 1, 2$, and that the system equations are given by

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 + A_{12} x_2, \\ \dot{x}_2 &= A_{21} x_1 + A_2 x_2, \\ x_1(0) &= x_{10}, \qquad x_2(0) = x_{20}, \end{aligned} \tag{4.1}$$

and that we are interested in the finite simulation interval $[0, T]$. We shall also use the compact notation $\dot{x} = Ax,\ x(0) = x_0$ to denote the same system. We start with some approximate solution $x^{(0)}(t)$ of the system as the initial (zeroth) iterate. For instance we may use the solution of the decoupled systems given by $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$ which satisfies $\dot{x}_1^{(0)} = A_1 x_1^{(0)},\ x_1^{(0)}(0) = x_{10}$ and $\dot{x}_2^{(0)} = A_2 x_2^{(0)},\ x_2^{(0)}(0) = x_{20}$. Another approach may be to use some reduced order model solution as $x^{(0)}$. Our analysis does not depend on this initial choice.

Suppose we have trajectory $x^{(\alpha)}$ at the αth iteration. Then we find best approximating $k(\leq n)$ dimensional subspaces in $\mathbb{R}^n$ (k is fixed throughout the iterations) and the corresponding orthogonal projections $P_1^{(\alpha)}$ and $P_2^{(\alpha)}$ (both are $n \times n$ matrices) for the trajectories $x_1^{(\alpha)}$ and $x_2^{(\alpha)}$, respectively. The next iterate $x^{(\alpha+1)}$ is obtained by forming partially reduced models. We combine unreduced system 1 with reduced system 2 to obtain $x_1^{(\alpha+1)}$ and similarly for $x_2^{(\alpha+1)}$. Then we find the projections $P_1^{(\alpha+1)}$ and $P_2^{(\alpha+1)}$ corresponding to $x_1^{(\alpha+1)}$ and $x_2^{(\alpha+1)}$. Thus if

$$P^{(\alpha)} = \begin{bmatrix} P_1^{(\alpha)} & 0_{n\times n} \\ 0_{n\times n} & P_2^{(\alpha)} \end{bmatrix}$$

is the combined projection at the α th iteration, then $x^{(\alpha+1)}$ is given by

$$\begin{aligned} \dot{x}_1^{(\alpha+1)} &= A_1 x_1^{(\alpha+1)} + A_{12} \hat{x}_2^{(\alpha+1)}, \\ \dot{\hat{x}}_2^{(\alpha+1)} &= P_2^{(\alpha)} A_{21} x_1^{(\alpha+1)} + P_2^{(\alpha)} A_2 \hat{x}_2^{(\alpha+1)}, \\ x_1^{(\alpha+1)}(0) &= x_{10}, \qquad \hat{x}_2^{(\alpha+1)}(0) = P_2^{(\alpha)} x_{20} \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} \dot{\hat{x}}_1^{(\alpha+1)} &= P_1^{(\alpha)} A_1 \hat{x}_1^{(\alpha+1)} + P_1^{(\alpha)} A_{12} x_2^{(\alpha+1)}, \\ \dot{x}_2^{(\alpha+1)} &= A_{21} \hat{x}_1^{(\alpha+1)} + A_2 x_2^{(\alpha+1)}, \\ \hat{x}_1^{(\alpha+1)}(0) &= P_1^{(\alpha)} x_{10}, \qquad x_2^{(\alpha+1)}(0) = x_{20}. \end{aligned} \tag{4.3}$$

We can rewrite the above equations more compactly as

$$\begin{aligned} \dot{x}^{(\alpha+1)} &= A_d x^{(\alpha+1)} + A_o \hat{x}^{(\alpha+1)}, \\ \dot{\hat{x}}^{(\alpha+1)} &= P^{(\alpha)} A_o x^{(\alpha+1)} + P^{(\alpha)} A_d \hat{x}^{(\alpha+1)}, \\ x^{(\alpha+1)}(0) &= x_0, \qquad \hat{x}^{(\alpha+1)}(0) = P^{(\alpha)} x_0, \end{aligned} \tag{4.4}$$

where

$$P^{(\alpha)} = \begin{bmatrix} P_1^{(\alpha)} & 0_{n\times n} \\ 0_{n\times n} & P_2^{(\alpha)} \end{bmatrix}, \tag{4.5}$$

$$A_d = \begin{bmatrix} A_1 & 0_{n\times n} \\ 0_{n\times n} & A_2 \end{bmatrix}, \tag{4.6}$$

and

$$A_o = \begin{bmatrix} 0_{n\times n} & A_{12} \\ A_{21} & 0_{n\times n} \end{bmatrix}. \tag{4.7}$$

Thus $A = A_d + A_o$.

Define the iteration operator $\mathcal{I} : \mathcal{L}_2([0,T],\mathbb{R}^{2n}) \to \mathcal{L}_2([0,T],\mathbb{R}^{2n})$ as the one that maps $x^{(\alpha)}$ to $x^{(\alpha+1)}$. We are interested in the fixed points of this operator and their stability. It must be noted that $\mathcal{I}$ is essentially nonlinear and is the composition of two operators $\mathcal{I} = \mathcal{S} \circ \mathcal{R}$. Here $\mathcal{R} : \mathcal{L}_2([0,T],\mathbb{R}^{2n}) \to \mathcal{P}^2$ is the operator that maps $x^{(\alpha)}$ to $P^{(\alpha)} \in \mathcal{P}^2$, where $\mathcal{P} \subset \mathbb{R}^{n\times n}$ is the manifold of rank $k \le n$ orthogonal projections, and $\mathcal{S} : \mathcal{P}^2 \to \mathcal{L}_2([0,T],\mathbb{R}^{2n})$ maps $P^{(\alpha)}$ to $x^{(\alpha+1)}$ by (4.4). Let $X \in \mathcal{L}_2([0,T],\mathbb{R}^{2n})$ be the true solution of the original system of equations,

$$\dot{X} = AX, \qquad X(0) = x_0.$$

Let $x^* \in \mathcal{L}_2([0,T],\mathbb{R}^{2n})$ be any fixed point of $\mathcal{I}$, i.e., $\mathcal{I}x^* = x^*$. We would like x^* to be a good approximation for X. Our analysis will provide an upper bound on $\|x^* - X\|$. (All function norms are assumed to be 2-norms unless stated otherwise.) We shall show that the error depends on the norm of A_o (the off-diagonal part), the POD projection error $\|x^* - P^*x^*\|$, and the growth/decay properties of e^{At} in the time interval T.

Since $\mathcal{I}$ is nonlinear it is in general difficult to know if and how many fixed points exist. It is also difficult to determine whether $\mathcal{I}$ is globally contractive. In fact $\mathcal{I}$ is ill-defined for some trajectories x; this occurs when there are many k dimensional subspaces that best fit x in the least-square sense. However, it is clear that when A_o is the zero matrix, i.e., when the systems are decoupled, the iterations will converge after one step, and in addition there is only one fixed point. Under mild regularity conditions it can be shown that this fixed point will persist for nontrivial A_o, with $\|A_o\|$ small enough, and that this fixed point will be stable. We will provide an analysis that estimates the rate of convergence based on the linearization of $\mathcal{I}$ at a fixed point x^*. We will show that the convergence rate depends on $\|A_o\|$, norms of the exponentials of A and some related matrices, the interval length T, the error $\|x^* - P^*x^*\|$, as well as on the sensitivity of P to perturbations in x at the fixed point (x^*, P^*) which can be related to the eigenvalues of the correlation matrix of the fixed point trajectory x^*.

The rest of the subsections are organized as follows. In section 4.2 we summarize all the important results of our analysis up front. In section 4.3 we derive an estimate for the norm of the trajectory of a subsystem in a given finite time interval for a linear time invariant system with time varying inputs. In section 4.4 we show that under mild regularity conditions for sufficiently small values of $\|A_o\|$ a fixed point exists. In section 4.5 we derive an estimate for the error $\|x^* - X\|$, and in section 4.6 we provide an estimate for convergence rate of $\mathcal{I}$ and a discussion of the various factors that affect the convergence. Finally, in section 4.7 we study the behavior of DIRM for arbitrarily small time intervals.

4.2. Summary of the results of the analysis. Here we shall provide a summary of the results of the analysis from the rest of the subsections. The reader who is not interested in mathematical details and proofs may read this subsection and then skip to section 5 for numerical examples.

Result 1. For systems that are sufficiently diagonally dominant ($\|A_o\|$ small enough), under further mild regularity conditions a fixed point x^* of $\mathcal{I}$ exists. We do not provide a quantitative bound on $\|A_o\|$. This result is proven in section 4.4. See Proposition 4.5. This result holds for an arbitrary (finite) number of subsystems with possibly different dimensions.

Result 2. Assuming that a fixed point x^* of $\mathcal{I}$ exists we obtain an upper bound (4.21) for the error between the fixed point trajectory x^* and the true solution trajectory X of the system. This is shown in section 4.5. See Proposition 4.8.

Result 3. Assuming that a fixed point x^* of $\mathcal{I}$ exists we obtain an upper bound for $\|D\mathcal{I}(x^*)\|$ (the norm of the linearization of the iteration operator at the fixed point). If $\|D\mathcal{I}(x^*)\| < 1$, then DIRM will converge for all initial iterates $x^{(0)}$ that are sufficiently close to x^*. See section 4.6 and Proposition 4.9.

Result 4. We show that in the case of systems for which a fixed point x^* of DIRM exists for all small enough T, that DIRM converges to x^* for all sufficiently small T if our initial iterate $x^{(0)}$ is sufficiently close to x^*. See section 4.7 and Proposition 4.11.

Remark 4.1. The proof of Results 2 and 3 (and hence that of 4) use the equation (4.4) which holds for two subsystems. We expect "qualitatively" similar results to hold for arbitrary number of subsystems but cannot make any rigorous claims without further analysis. We have limited the analysis to two subsystems because an equation equivalent to (4.4) is combinatorially very cumbersome for the case of more than two subsystems.

The above results do not constitute a comprehensive convergence analysis. For instance we cannot make conclusions about global convergence of DIRM. But these results suggest that DIRM is likely to perform well if certain desirable conditions are met. This has been verified by numerical experiments.

4.3. Finite horizon response of a subsystem. In our analysis of errors and convergence rate we need to estimate the 2-norm of the trajectory of a subsystem in a given finite time interval in response to a forcing term (input) and nontrivial initial conditions for a linear time invariant system. In this section we introduce some relevant notation as well as estimates that will be employed in our later analysis.

Consider the system

$$\dot{x} = Ax + u$$

with input $u(t)$ and initial condition $x(0) = x_0$ in the interval $[0, T]$. We are only interested in $u \in \mathcal{L}_2([0,T], \mathbb{R}^n)$. The solution is

$$x(t) = \int_0^t e^{A(t-\tau)} u(\tau) d\tau + e^{At} x_0.$$

This may be written in the form

$$x = F(T; A)u + G(T; A)x_0, \tag{4.8}$$

where $F(T;A) : \mathcal{L}_2([0,T], \mathbb{R}^n) \to \mathcal{L}_2([0,T], \mathbb{R}^n)$ and $G(T;A) : \mathbb{R}^n \to \mathcal{L}_2([0,T], \mathbb{R}^n)$ are linear operators. It is in general very difficult to obtain sharp estimates for the

norms of $F(T;A)$ and $G(T;A)$, and in fact this basically reduces to the problem of estimating the norm of the matrix exponential. As such we shall not provide an estimate, but we will remark that these norms grow exponentially with T at a rate that is determined by the largest real part of any eigenvalue of A and in addition depend on the nonnormality of A. See [4] for an estimate of the matrix exponential. In our analysis we estimate $\|x\|$ as

$$\|x\| \le \|F(T;A)\|\|u\| + \|G(T;A)\|\|x_0\|, \tag{4.9}$$

expressing the results in terms of $\|F(T;A)\|$ and $\|G(T;A)\|$.

Remark 4.2. Note that the norms on $F(T;A)$ and $G(T;A)$ are the appropriate induced 2-norms.

We shall state and prove a simple lemma on $F(T;A)$ which will be used later.

LEMMA 4.3. $\lim_{T\to 0}\|F(T;A)\| = 0$.

Proof.

$$\begin{aligned}\|F(T;A)u\|^2 &= \int_0^T \left\| \int_0^t e^{A(t-\tau)}u(\tau)d\tau \right\|^2 dt \\ &\le \int_0^T \int_0^t \|e^{A(t-\tau)}\|^2\|u(\tau)\|^2 d\tau dt \\ &\le e^{2\|A\|T}\int_0^T \int_0^T \|u(\tau)\|^2 d\tau dt \\ &= Te^{2\|A\|T}\|u\|^2.\end{aligned}$$

So in fact, as $T \to 0$, $\|F(T;A)\| = \mathrm{O}(\sqrt{T})$. ∎

Now we will focus on a system that consists of two subsystems and obtain an estimate for one of the subsystems that relates the results with the norms of the coupling terms (the off-diagonal blocks) as well as the subsystem properties (diagonal blocks). Consider the coupled systems

$$\begin{aligned}\dot{x}_1 &= A_1x_1 + \kappa A_{12}x_2 + u_1, \\ \dot{x}_2 &= \kappa A_{21}x_1 + A_2x_2 + u_2, \\ x_1(0) = x_{10},\ x_2(0) &= x_{20}\end{aligned} \tag{4.10}$$

in the interval $[0,T]$. Here κ is a "coupling parameter" introduced to aid our analysis. The final results are all evaluated at $\kappa = 1$. We will obtain an estimate for $\|x_1\|$. Since $x_i(t)$ (for $i = 1,2$) is a (vector-valued) entire function of κ we may write it as

$$x_i(t;\kappa) = \sum_{\alpha=0}^{\infty} \kappa^\alpha \frac{\partial^\alpha x_i(t;0)}{\alpha!}, \qquad i = 1,2,$$

where $\partial = \frac{\partial}{\partial\kappa}$, and the series converges for all t and all κ. For $\alpha \ge 1$, $\partial^\alpha x_i(t;0)$ are given by the decoupled equations

$$\begin{aligned}\partial^\alpha \dot{x}_1(t;0) &= A_1\partial^\alpha x_1(t;0) + \alpha A_{12}\partial^{\alpha-1}x_2(t;0), \\ \partial^\alpha \dot{x}_2(t;0) &= A_2\partial^\alpha x_2(t;0) + \alpha A_{21}\partial^{\alpha-1}x_1(t;0), \\ \partial^\alpha x_1(0;0) = 0, &\qquad \partial^\alpha x_2(0;0) = 0.\end{aligned}$$

For the $\alpha = 0$ case we have

$$\begin{aligned}\dot{x}_1(t;0) &= A_1 x_1(t;0) + u_1(t),\\ \dot{x}_2(t;0) &= A_2 x_2(t;0) + u_2(t),\\ x_1(0;0) &= x_{10}, \qquad x_2(0;0) = x_{20}.\end{aligned}$$

Let $x_1(.\,;\kappa)$ denote the function, i.e., $x_1(.\,;\kappa) \in \mathcal{L}_2([0,T],\mathbb{R}^n)$. Setting $\kappa = 1$, from the above equations we can write $x_1(.\,;1)$ as

$$x_1(.\,;1) = \sum_{\alpha=0}^{\infty} F^\alpha x_1(.\,;0) + \sum_{\alpha=0}^{\infty} F^\alpha F_1 A_{12} x_2(.\,;0),$$

where the operators $F_1, F_2, F : \mathcal{L}_2([0,T],\mathbb{R}^n) \to \mathcal{L}_2([0,T],\mathbb{R}^n)$ are defined by $F_i = F(T;A_i)$, for $i = 1,2$, and

$$F = F_1 A_{12} F_2 A_{21}.$$

Assuming $\|F\| < 1$ (which is true for sufficiently small T by Lemma 4.3) we obtain an upper bound for $\|x_1\|$:

$$\|x_1\| \le \sum_{\alpha=0}^{\infty} \|F\|^\alpha \|x_{d1}\| + \sum_{\alpha=0}^{\infty} \|F\|^\alpha \|F_1\| \|A_{12}\| \|x_{d2}\|,$$

where we have dropped the parameter κ altogether and x_{di} for $i = 1,2$ denote the solutions of the decoupled systems: $\dot{x}_{di} = A_i x_{di} + u_i,\ x_{di}(0) = x_{i0}$. Finally, after simplifying the above bound, we obtain the result that, for sufficiently small T,

$$\|x_1\| \le \frac{(\|F_1\|\|u_1\| + \|G_1\|\|x_{10}\|)}{1 - \|F\|} + \frac{\|F_1\|\|A_{12}\|(\|F_2\|\|u_2\| + \|G_2\|\|x_{20}\|)}{1 - \|F\|}, \tag{4.11}$$

where we have used the estimates (4.9) for $\|x_{d1}\|$ and $\|x_{d2}\|$, and $G_i = G(T;A_i)$ for $i = 1,2$.

It is clear that the effect of subsystem 2 on subsystem 1 diminishes as the norm of A_{12} diminishes.

Remark 4.4. It is interesting to note that the κ series expansion mentioned here is intimately related to the Jacobi WR method. In fact the sequence of partial sums of the series for $\kappa = 1$ is the same as the sequence of iterates obtained by applying the Jacobi WR, i.e., WR with the splitting $A = A_d + A_o$, where A_d and A_o are defined by (4.6) and (4.7), respectively, starting with isolated subsystem (couplings assumed zero) solutions as the initial iterate.

4.4. Existence of fixed points of DIRM. In general it is hard to prove the existence of fixed points of the operator $\mathcal{I}$. However, under mild regularity conditions we can show that a fixed point exists for sufficiently small $\|A_o\|$. For this purpose we shall consider the iteration operator $\mathcal{J} : \mathcal{P}^2 \to \mathcal{P}^2$ that maps $P^{(\alpha)}$ to $P^{(\alpha+1)}$. Recalling that $\mathcal{I} = \mathcal{S} \circ \mathcal{R}$ from section 4.1 we see that $\mathcal{J} = \mathcal{R} \circ \mathcal{S}$. It is easy to see that x^* is a fixed point of $\mathcal{I}$ if and only if $P^* = \mathcal{R}x^*$ is a fixed point of $\mathcal{J}$ (provided $\mathcal{R}x^*$ is well defined) and similarly P^* is a fixed point of $\mathcal{J}$ if and only if $x^* = \mathcal{S}P^*$ is a fixed point of $\mathcal{I}$.

PROPOSITION 4.5. *Consider a system with a given diagonal part A_d as defined by* (4.6). *Let $x_d = (x_{d1}, x_{d2})$ be the solution of the decoupled systems; $\dot{x}_d = A_d x_d$, $x_d(0) = x_0$. Let $\nu_1^i \ge \nu_2^i \ge \cdots \ge \nu_n^i \ge 0$ be the eigenvalues of the correlation matrices*

of x_{di} for $i = 1, 2$, respectively. If $\nu_k^i > \nu_{k+1}^i$, for both $i = 1, 2$, then the operator $\mathcal{J}$ (and hence $\mathcal{I}$) has a fixed point for all off-diagonal parts A_o (as defined by (4.7)) in an open neighborhood of the origin in $\mathcal{O} \subset \mathbb{R}^{2n \times 2n}$. Here $\mathcal{O}$ is the $2n^2$ dimensional subspace of all possible off-diagonal parts.

Proof. Write the system $\dot{x} = Ax$ as

$$\dot{x} = A_d x + A_o x,$$

where A_d and A_o are defined according to (4.6) and (4.7), respectively. (Note that $A = A_d + A_o$.) We shall treat A_d as fixed and consider A_o as variable.

A fixed point

$$P^* = \begin{bmatrix} P_1^* & 0_{n\times n} \\ 0_{n\times n} & P_2^* \end{bmatrix}$$

of $\mathcal{J}$ must be such that $P_1 = P_1^*$ is a minimizer of $e_1(P_1, x_1)$ while holding x_1 fixed, and $P_2 = P_2^*$ is a minimizer of $e_2(P_2, x_2)$ while holding x_2 fixed, where $e_i(P_i, x_i)$ are defined by

$$e_i(P_i, x_i) = \int_0^T (P_i x_i(t) - x_i(t))^T (P_i x_i(t) - x_i(t)) dt, \qquad i = 1, 2.$$

Hence by applying the first order optimality conditions we see that $P = P^*$ must be a root of the following system of equations:

$$\begin{aligned} &\frac{\partial e_1}{\partial P_1}(P_1, S_1(P_2; A_o)) = 0, \\ &\frac{\partial e_2}{\partial P_2}(P_2, S_2(P_1; A_o)) = 0, \end{aligned} \tag{4.12}$$

where we have used the fact that $x_1 = x_1^*$ at the fixed point depends on P_2^* and similarly $x_2 = x_2^*$ depends on P_1^*. Here the "solution operator" S_1 maps $P_2^{(\alpha)}$ to $x_1^{(\alpha+1)}$ according to the equations (4.2). The operator S_2 is defined similarly. Note that the operators S_1 and S_2 in general both depend on A_o.

Because of the coupling, it is hard to decide if the system (4.12) has a root in general. However, when $A_o = 0 \in \mathcal{O}$, the original system of ODEs are decoupled and as such S_1 and S_2 are independent of P_2 and P_1, respectively. Hence the equations in (4.12) are decoupled. Furthermore, our assumption that $\nu_k^i > \nu_{k+1}^i$ for both $i = 1, 2$ implies that the two errors e_1 and e_2 can be minimized uniquely and independently according to the POD procedure. This proves the existence of a unique fixed point $P = P^*(A_o = 0)$ of the operator $\mathcal{J}$ for $A_o = 0$. The second order optimality conditions for unique minima imply that both $\frac{\partial^2 e_1}{\partial P_1^2}$ and $\frac{\partial^2 e_2}{\partial P_2^2}$ when evaluated at $A_o = 0$ and $P = P^*(A_o = 0)$ have full rank.

Therefore it also follows that the Jacobian

$$\begin{bmatrix} \frac{\partial^2 e_1}{\partial P_1^2} & \frac{\partial^2 e_1}{\partial P_1 \partial P_2} \\ \frac{\partial^2 e_2}{\partial P_1 \partial P_2} & \frac{\partial^2 e_2}{\partial P_2^2} \end{bmatrix}$$

is full rank for $A_o = 0$ and $P = P^*(A_o = 0)$. Hence, by the implicit function theorem, we conclude that a root $P = P^*(A_o)$ of (4.12) exists for all A_o in an open neighborhood of $0 \in \mathcal{O}$. Furthermore, by continuity it follows that $\frac{\partial^2 e_1}{\partial P_1^2}$ and $\frac{\partial^2 e_2}{\partial P_2^2}$ have

full rank for $(A_o, P = P^*(A_o))$ for all A_o in some open neighborhood of $0 \in \mathcal{O}$. This establishes $P = P^*(A_o)$ as a fixed point of $\mathcal{J}$ for all A_o in an open neighborhood of $0 \in \mathcal{O}$. ∎

COROLLARY 4.6. *Under similar assumptions Proposition* 4.5 *holds for an arbitrary number m of subsystems of possibly different dimensions.*

Proof. Follow the same line of reasoning with (4.12) replaced by

$$\frac{\partial e_i}{\partial P_i}(P_i, S_i(P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, P_m; A_o)) = 0, \qquad i = 1, \ldots, m. \tag{4.13}$$

The key point is that the operator S_i does not depend on P_i. ∎

Remark 4.7. Ideally we would like to show that for any value of $\|A_o\|$ a fixed point exists for sufficiently small T. The intuition is that when T gets arbitrarily small, the trajectories are increasingly well approximated by straight lines. However, we do not have a proof yet.

4.5. Accuracy of DIRM. We will introduce a few new variables to facilitate our analysis. Given $P^{(\alpha)}, x^{(\alpha)}, \hat{x}^{(\alpha)}$, and X as defined in section 4.1, define $v^{(\alpha)}, w^{(\alpha)}$, and $\xi^{(\alpha)}$ as follows:

$$v^{(\alpha)} = P^{(\alpha-1)}x^{(\alpha)} - x^{(\alpha)}, \tag{4.14}$$

$$w^{(\alpha)} = \hat{x}^{(\alpha)} - P^{(\alpha-1)}x^{(\alpha)}, \tag{4.15}$$

and

$$\xi^{(\alpha)} = x^{(\alpha)} - X. \tag{4.16}$$

We may think of $v^{(\alpha)}$ as a "difference" trajectory that measures the gap between $x^{(\alpha)}$ and its projection $P^{(\alpha-1)}x^{(\alpha)}$ and $w^{(\alpha)}$ as a difference trajectory that measures the gap between the reduced trajectory $\hat{x}^{(\alpha)}$ and $P^{(\alpha-1)}x^{(\alpha)}$. The trajectory $\xi^{(\alpha)}$ is the error between the true solution and the DIRM iterate at step α.

Suppose x^* is a fixed point of $\mathcal{I}$. Assume $P^* = \mathcal{R}x^*$ is well defined. Let $\hat{x}^*, v^*, w^*$, and ξ^* be the corresponding fixed point trajectories. Note that the error in using DIRM is ξ^*. We will provide an estimate of $\|\xi^*\|$. Substituting $v^{(\alpha)} = v^*$, $P^{(\alpha-1)} = P^*$, and $x^{(\alpha)} = x^*$ in (4.14), we obtain

$$v^* = P^*x^* - x^*.$$

Similarly we obtain $w^* = \hat{x}^* - P^*x^*$ from (4.15). Note that these two relations imply that $\hat{x}^* - x^* = v^* + w^*$. Differentiating $w^* = \hat{x}^* - P^*x^*$ with respect to time, and using (4.4), we obtain

$$\begin{aligned} \dot{w}^* &= P^*A_d\hat{x}^* + P^*A_ox^* - P^*A_dx^* - P^*A_o\hat{x}^* \\ &= P^*(A_d - A_o)(\hat{x}^* - x^*). \end{aligned}$$

Hence we obtain the following differential equation for w^*:

$$\dot{w}^* = P^*(A_d - A_o)w^* + P^*(A_d - A_o)v^*, \qquad w^*(0) = 0. \tag{4.17}$$

Similarly (4.16) implies $\xi^* = x^* - X$. Differentiating and using (4.4), and $\dot{X} = AX = (A_d + A_o)X$, we obtain

$$\begin{aligned}\dot{\xi}^* &= A_d x^* + A_o \hat{x}^* - AX \\ &= A(x^* - X) + A_o(\hat{x}^* - x^*).\end{aligned}$$

Using $\hat{x}^* - x^* = v^* + w^*$ we write the equation for ξ^* as

$$\dot{\xi}^* = A\xi^* + A_o(v^* + w^*), \qquad \xi^*(0) = 0. \tag{4.18}$$

From the application of the estimate (4.9) to (4.17) we obtain that

$$\|w^*\| \leq \|F(T; P^*(A_d - A_o))\|\|A_d - A_o\|\|v^*\|. \tag{4.19}$$

Applying the estimate (4.9) to (4.18) and using the above equation we obtain

$$\|\xi^*\| \leq \|F(T; A)\|\|A_o\| \left\{1 + F(T; P^*(A_d - A_o))\|A_d - A_o\|\right\} \|v^*\|. \tag{4.20}$$

The quantity $\|v^*\|$ is the sum of the POD projection errors $\|P_i^* x_i - x_i\|$ of both the subsystems. This quantity is the same as the square root of the sum of the eigenvalues of the neglected modes summed over both the subsystems. Note that if the POD projection error of the fixed point trajectory is zero, then the error of the converged DIRM solution is zero. It is also clear that the error depends on the norm of the off-diagonal blocks A_o, on the norm of the exponentials of A and $P^*(A_d - A_o)$, as well as on the time interval T.

We have thus proved the following proposition.

PROPOSITION 4.8. *Let x^* be a fixed point of $\mathcal{I}$ and suppose that $P^* = \mathcal{R}x^*$ is well defined. Let X be the true solution: $\dot{X} = AX$; $X(0) = x_0$. Then the error $\|x^* - X\|$ (2-norm) satisfies*

$$\|x^* - X\| \leq \|F(T; A)\|\|A_o\| \left\{1 + F(T; P^*(A_d - A_o))\|A_d - A_o\|\right\} \|v^*\|, \tag{4.21}$$

where $\|v^*\| = \|P^* x^* - x^*\|$ is the projection error of the fixed point trajectory.

4.6. Rate of convergence. In this section, we will compute an upper bound for $\|D\mathcal{I}(x^*)\|$, the norm of the linearization of the iteration $\mathcal{I}$ at a fixed point x^*. First, we will compute $\delta x^{(\alpha+1)} = D\mathcal{I}(x^{(\alpha)})(\delta x^{(\alpha)})$, which is the variation in $x^{(\alpha+1)}$ due to a variation $\delta x^{(\alpha)}$ in $x^{(\alpha)}$. The notation $D\mathcal{I}(x)(\delta x)$ denotes the directional derivative of the operator $\mathcal{I}$ evaluated at $x \in \mathcal{L}_2([0,T], \mathbb{R}^{2n})$ in the direction $\delta x \in \mathcal{L}_2([0,T], \mathbb{R}^{2n})$. The variations of all quantities will be denoted by the prefix δ, except that the variation of $P^{(\alpha)}$ will be denoted by $E^{(\alpha)}$. In our analysis the norm used for the variations of trajectories will also be the 2-norm, and the norms of matrices will be the induced 2-norm.

Again we shall make use of the difference trajectories $v^{(\alpha)}$ and $w^{(\alpha)}$ as defined by (4.14) and (4.15). Taking variations of (4.14) (with α replaced by $\alpha + 1$) it follows that

$$\delta v^{(\alpha+1)} = E^{(\alpha)} x^{(\alpha+1)} + \left(P^{(\alpha)} - 1\right) \delta x^{(\alpha+1)}. \tag{4.22}$$

Following a procedure similar to the one that was used to obtain (4.18), we obtain from (4.16) the following equation for $\xi^{(\alpha+1)}$:

$$\dot{\xi}^{(\alpha+1)} = A\xi^{(\alpha+1)} + A_o\left(v^{(\alpha+1)} + w^{(\alpha+1)}\right), \qquad \xi^{(\alpha+1)}(0) = 0.$$

From (4.16) we also see that $\delta x^{(\alpha)} = \delta\xi^{(\alpha)}$. Therefore, taking variations of the above equation, we get

$$\delta\dot{x}^{(\alpha+1)} = A\delta x^{(\alpha+1)} + A_o\delta v^{(\alpha+1)} + A_o\delta w^{(\alpha+1)}, \qquad \delta x^{(\alpha+1)}(0) = 0. \tag{4.23}$$

Following a procedure similar to the one that was used to obtain (4.17), we obtain from (4.15) the following equation for $w^{(\alpha+1)}$:

$$\begin{aligned} \dot{w}^{(\alpha+1)} &= P^{(\alpha)}(A_d - A_o)w^{(\alpha+1)} + P^{(\alpha)}(A_d - A_o)v^{(\alpha+1)}, \\ w^{(\alpha+1)}(0) &= 0. \end{aligned} \tag{4.24}$$

Hence the variation $\delta w^{(\alpha+1)}$ is given by

$$\begin{aligned} \delta\dot{w}^{(\alpha+1)} &= P^{(\alpha)}(A_d - A_o)\delta w^{(\alpha+1)} + P^{(\alpha)}(A_d - A_o)\delta v^{(\alpha+1)} \\ &\quad + E^{(\alpha)}(A_d - A_o)\left(w^{(\alpha+1)} + v^{(\alpha+1)}\right), \quad \delta w^{(\alpha+1)}(0) = 0. \end{aligned} \tag{4.25}$$

Substituting for $\delta v^{(\alpha+1)}$ from (4.22) into (4.23) and (4.25) we get a coupled system of equations for $\delta x^{(\alpha+1)}$ and $\delta w^{(\alpha+1)}$:

$$\begin{aligned} \delta\dot{x}^{(\alpha+1)} &= (A + A_o(P^{(\alpha)} - 1))\delta x^{(\alpha+1)} + A_o\delta w^{(\alpha+1)} + A_oE^{(\alpha)}x^{(\alpha+1)}, \\ \delta\dot{w}^{(\alpha+1)} &= P^{(\alpha)}(A_d - A_o)\left(P^{(\alpha)} - 1\right)\delta x^{(\alpha+1)} + P^{(\alpha)}(A_d - A_o)\delta w^{(\alpha+1)} \\ &\quad + P^{(\alpha)}(A_d - A_o)E^{(\alpha)}x^{(\alpha+1)} + E^{(\alpha)}(A_d - A_o)\left(w^{(\alpha+1)} + v^{(\alpha+1)}\right), \\ \delta x^{(\alpha+1)}(0) &= 0, \qquad \delta w^{(\alpha+1)}(0) = 0. \end{aligned} \tag{4.26}$$

Note that the above system is driven by terms containing $x^{(\alpha+1)}$, $v^{(\alpha+1)}$, and $w^{(\alpha+1)}$. Since we are interested in perturbations about the fixed point x^*, we set

$$\begin{aligned} x^{(\alpha+1)} &= x^{(\alpha)} = x^*, \\ v^{(\alpha+1)} &= v^{(\alpha)} = v^*, \\ w^{(\alpha+1)} &= w^{(\alpha)} = w^*. \end{aligned}$$

Also we shall denote $\delta x^{(\alpha)} = \delta x$ and write $E^{(\alpha)}$ and $\delta x^{(\alpha+1)}$ as

$$\begin{aligned} E^{(\alpha)} &= \frac{dP}{dx}(x^*)(\delta x), \\ \delta x^{(\alpha+1)} &= D\mathcal{I}(x^*)(\delta x). \end{aligned}$$

Note that $\frac{dP}{dx}(x^*)(\delta x)$, which may also be written as $D\mathcal{R}(x^*)(\delta x)$, is the directional derivative of $P = \mathcal{R}(x)$ at $x = x^*$ in the direction δx. For sufficiently small T we apply the estimate (4.11) to (4.26) at a fixed point and obtain

$$\begin{aligned} \|D\mathcal{I}(x^*)(\delta x)\| &\le \frac{\|F_1\|}{(1 - \|F\|)}\left\|A_oP^*\frac{dP}{dx}(x^*)(\delta x)x^*\right\| \\ &+ \frac{\|F_1\|\|F_2\|\|A_o\|}{(1 - \|F\|)}\left\|P^*(A_d - A_o)\frac{dP}{dx}(x^*)(\delta x)x^* + \frac{dP}{dx}(x^*)(\delta x)(A_d - A_o)(w^* + v^*)\right\|, \end{aligned}$$

where

$$\begin{aligned} F_1 &= F(T; A + A_o(P^* - 1)), \\ F_2 &= F(T; P^*(A_d - A_o)), \\ F &= F_1A_oF_2P^*(A_d - A_o)(P^* - 1). \end{aligned} \tag{4.27}$$

We simplify further and write

$$\|D\mathcal{I}(x^*)(\delta x)\| \le \frac{\|F_1\|\|A_o\|}{(1-\|F\|)}H,$$

where the term H is given by

$$H = \left\|\frac{dP}{dx}(x^*)(\delta x)x^*\right\| + \|F_2\|\|A_d - A_o\|\left(\left\|\frac{dP}{dx}(x^*)(\delta x)x^*\right\| + \left\|\frac{dP}{dx}(x^*)(\delta x)\right\|\|v^* + w^*\|\right).$$

Using the estimate (4.19) one can obtain an upper bound for H which does not contain w^*. After rearranging some terms we obtain the upper bound

$$\|D\mathcal{I}(x^*)(\delta x)\| \le \frac{\|F_1\|\|A_o\|(1+\|F_2\|\|A_d - A_o\|)}{(1-\|F\|)} \times \left(\left\|\frac{dP}{dx}(x^*)(\delta x)x^*\right\| + \left\|\frac{dP}{dx}(x^*)(\delta x)\right\|\|F_2\|\|A_d - A_o\|\|v^*\|\right).$$

Taking the supremum over all unit norm variations δx, we obtain the bound

$$\|D\mathcal{I}(x^*)\| \le \frac{\|F_1\|\|A_o\|(1+\|F_2\|\|A_d - A_o\|)}{(1-\|F\|)} \times \left(\sup_{\|\delta x\|=1}\left\{\left\|\frac{dP}{dx}(x^*)(\delta x)x^*\right\|\right\} + \left\|\frac{dP}{dx}(x^*)\right\|\|F_2\|\|A_d - A_o\|\|v^*\|\right).$$

The sensitivity of the POD projection matrix $P(x)$ to perturbations in the trajectory x has been studied and quantified in [13]. It follows directly from the results in [13] that

$$\left\|\frac{dP}{dx}(x^*)\right\|_{\mathrm{F}} = \max_{i\le k,\, j\le n-k}\sqrt{2}\frac{\sqrt{\lambda_i^1+\lambda_{j+k}^1}}{\lambda_i^1-\lambda_{j+k}^1} + \max_{i\le k,\, j\le n-k}\sqrt{2}\frac{\sqrt{\lambda_i^2+\lambda_{j+k}^2}}{\lambda_i^2-\lambda_{j+k}^2}, \tag{4.28}$$

$$\sup_{\|\delta x\|=1}\left\{\left\|\frac{dP}{dx}(x^*)(\delta x)x^*\right\|\right\} = \left(\frac{\lambda_k^1+\lambda_{k+1}^1}{\lambda_k^1-\lambda_{k+1}^1}\right) + \left(\frac{\lambda_k^2+\lambda_{k+1}^2}{\lambda_k^2-\lambda_{k+1}^2}\right),$$

where $\lambda_1^i \ge \lambda_2^i \ge \cdots \ge \lambda_n^i$ are the ordered eigenvalues of the correlation matrix of the fixed point trajectories x_i^*, for $i = 1, 2$, and the subscript "F" denotes the Frobenius norm. Note that we have assumed that $\lambda_k^i > \lambda_{k+1}^i$, for $i = 1, 2$, in order for P^* to be well defined. Also note that $\|v^*\|$ is given by

$$\|v^*\| = \sqrt{\lambda_{k+1}^1 + \cdots + \lambda_n^1 + \lambda_{k+1}^2 + \cdots + \lambda_n^2}.$$

Since $\|\frac{dP}{dx}(x^*)\|_2 \le \|\frac{dP}{dx}(x^*)\|_{\mathrm{F}}$, using the expression for $\|v^*\|$ we may obtain an upper bound for $\|D\mathcal{I}(x^*)\|$ which we shall state as a proposition.

PROPOSITION 4.9. *Suppose $x^* = (x_1^*, x_2^*)$ is a fixed point of $\mathcal{I}$ and assume that $P^* = \mathcal{R}x^*$ is well defined. Then, for sufficiently small T, $\|D\mathcal{I}(x^*)\|$ satisfies*

$$\|D\mathcal{I}(x^*)\| \le \frac{\|F_1\|\|A_o\|}{(1-\|F\|)}(1+\|F_2\|\|A_d - A_o\|)\{\mathcal{C}_1(\lambda) + \mathcal{C}_2(\lambda)\|F_2\|\|A_d - A_o\|\}, \tag{4.29}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are functions of the eigenvalues λ of the correlation matrices of the fixed point trajectories x_1^* and x_2^* given by

$$\begin{aligned}
\mathcal{C}_1 &= \left(\frac{\lambda_k^1+\lambda_{k+1}^1}{\lambda_k^1-\lambda_{k+1}^1}\right)+\left(\frac{\lambda_k^2+\lambda_{k+1}^2}{\lambda_k^2-\lambda_{k+1}^2}\right),\\
\mathcal{C}_2 &= \left(\max_{i\le k,\, j\le n-k}\sqrt{2}\frac{\sqrt{\lambda_i^1+\lambda_{j+k}^1}}{\lambda_i^1-\lambda_{j+k}^1}+\max_{i\le k,\, j\le n-k}\sqrt{2}\frac{\sqrt{\lambda_i^2+\lambda_{j+k}^2}}{\lambda_i^2-\lambda_{j+k}^2}\right)\\
&\quad\times\sqrt{\lambda_{k+1}^1+\cdots+\lambda_n^1+\lambda_{k+1}^2+\cdots+\lambda_n^2},
\end{aligned}\tag{4.30}$$

and F_1, F_2, and F are as defined in (4.27).

It can be seen from the above results that the convergence rate becomes arbitrarily fast as $\|A_o\|$ becomes arbitrarily small, which agrees with intuition. We also see that the convergence is faster when the POD error is small. However, when the POD error is zero ($\lambda_{k+1}^1+\cdots+\lambda_n^1+\lambda_{k+1}^2+\cdots+\lambda_n^2=0$), the above expression does not predict arbitrarily fast convergence. This may seem counterintuitive. However, one needs to consider this more carefully. The quantity $\|v^*\|=\lambda_{k+1}^1+\cdots+\lambda_n^1+\lambda_{k+1}^2+\cdots+\lambda_n^2$ is the POD projection error of the fixed point trajectory. Even if this is zero, $\|v\|$ corresponding to a nearby trajectory x need not be. Convergence depends on the behavior of nearby trajectories as well. This is evident from the numerical example in section 5.4. If, however, all subsystem trajectories always lie in some k dimensional subspace, then it is clear that DIRM will converge after one iteration. In terms of the above analysis this corresponds to $P^{(\alpha)}$ being constant and $v^{(\alpha+1)}=0$ for all α. Hence $E^{(\alpha)}=0$ and $\delta v^{(\alpha+1)}=0$. Then from (4.24) we see that $w^{(\alpha+1)}=0$ as well. All these together and (4.26) imply that $\delta x^{(\alpha+1)}=0$, indicating immediate convergence.

The analysis also suggests that when the eigenvalues λ_k^1 and λ_{k+1}^1 (or λ_k^2 and λ_{k+1}^2) are very close to each other we may expect difficulties in convergence, since both $\mathcal{C}_1$ and $\mathcal{C}_2$ become very large. This is numerically evident from the example of section 5.3.

4.7. Small time interval case. In this section we consider the convergence behavior of DIRM as $T\to 0$. First let us state and prove the following lemma.

LEMMA 4.10. *Let $z:[0,T_0]\to\mathbb{R}^n$ be a C^n-smooth trajectory. Let $0<T<T_0$ and let $\lambda_1\ge\lambda_2\cdots\ge\lambda_n\ge 0$ be the eigenvalues of the correlation matrix of $z:[0,T]\to\mathbb{R}^n$. Hence λ_i are functions of T. Furthermore, suppose the values of z and its first $n-1$ derivatives at $t=0$ form a linearly independent set. Note that this assumption is generically true in the sense that it holds for an open and dense subset (in $\mathbb{R}^{n(n-1)}$) of possible values of $z(t)$ and its first $n-1$ derivatives at $t=0$. Then, as $T\to 0$, $\lambda_{j+1}/\lambda_j\to 0$ for any $1\le j\le n-1$.*

Proof. In order to keep the proof concise, we shall prove for the case when $z(t)$ is analytic. The proof for C^n is similar. By assumption the set

$$\{z(0), z^{(1)}(0),\ldots,z^{(n-1)}(0)\}$$

is linearly independent, where $z^{(j)}(t)$ denotes the jth derivative of $z(t)$. Since the correlation matrix $R=\int_0^T z(t)z^T(t)dt$ is analytic in T, using Taylor expansion we may write it as

$$R=\sum_{j=1}^{\infty} r_j T^j,$$

where after simplification

$$r_j = \frac{1}{j}\sum_{l=1}^{j} \frac{z^{(l-1)}(0)}{(l-1)!}\frac{(z^{(j-l)}(0))^T}{(j-l)!}, \qquad j = 1, 2, \ldots.$$

Note that when $z(t)$ is C^n, R is C^{n+1} smooth in T, and we should use Taylor's theorem with the remainder term which is $O(T^{n+1})$.

It is clear that $\text{Image}(r_j) = \text{span}\{z(0), z^{(1)}(0), \ldots, z^{(j-1)}(0)\}$ and that $\text{Null}(r_j) = \text{Image}(r_j)^{\perp}$. Hence by our assumption it follows that $\text{rank}\,(r_j) = j$ for $j = 1, \ldots, n$. Let $\mu_1^j \geq \mu_2^j \geq \cdots \geq \mu_n^j \geq 0$ be the eigenvalues of r_j for all j. Note that $\mu_j^j > 0$ and $\mu_l^j = 0$ for $l = j+1, \ldots, n$, and $j = 1, \ldots, n$.

Define partial sums

$$R_j = \sum_{l=1}^{j} r_l T^l$$

for $j = 1, \ldots, n$. Let $\nu_1^j \geq \nu_2^j \geq \cdots \geq \nu_n^j \geq 0$ be the eigenvalues of R_j. Since $\text{Image}(r_{j-1}) \subset \text{Image}(r_j)$, for $j = 1, \ldots, n$, it follows that $\text{rank}(R_j) = \text{rank}(r_j) = j$ for $j = 1, \ldots, n$. Hence $\nu_j^j > 0$ and $\nu_l^j = 0$ for $l = j+1, \ldots, n$ and $j = 1, \ldots, n$. Since $R_j = R_{j-1} + T^j r_j$, for $j = 1, \ldots, n$, using Theorem 8.1.5 of [4] on symmetric eigenvalue perturbations, we see that $\nu_j^j \leq \nu_j^{j-1} + T^j \mu_1^j$. The same theorem also yields $\nu_n^{j-1} + T^j \mu_j^j \leq \nu_j^j$. Since $\nu_n^{j-1} = \nu_j^{j-1} = 0$, we get

$$0 < T^j \mu_j^j \leq \nu_j^j \leq T^j \mu_1^j, \qquad j = 1, \ldots, n. \tag{4.31}$$

For $j = 1, \ldots, n$ we may write R as

$$R = R_j + T^{j+1} N_{j+1},$$

where $N_{j+1} = \sum_{l=1}^{\infty} r_{j+l} T^{l-1}$. Let $\tilde{\nu}_1^j \geq \tilde{\nu}_2^j \geq \cdots \geq \tilde{\nu}_n^j \geq 0$ be the eigenvalues of N_j for $j = 2, \ldots, n+1$. Application of Theorem 8.1.5 of [4] to $R = R_j + T^{j+1} N_{j+1}$ yields

$$\nu_j^j + T^{j+1}\tilde{\nu}_n^{j+1} \leq \lambda_j \leq \nu_j^j + T^{j+1}\tilde{\nu}_1^{j+1}, \qquad j = 1, \ldots, n.$$

This together with (4.31) implies

$$0 < T^j \mu_j^j \leq \lambda_j \leq T^j \mu_1^j + T^{j+1}\tilde{\nu}_1^{j+1}, \qquad j = 1, \ldots, n. \tag{4.32}$$

This yields

$$\frac{\lambda_{j+1}}{\lambda_j} \leq T\left(\frac{\mu_1^{j+1}}{\mu_j^j} + T\tilde{\nu}_1^{j+2}\right), \qquad j = 1, \ldots, n-1.$$

Since $\lim_{T\to 0} N_{j+2} = r_{j+2}$, by continuity $\lim_{T\to 0} \tilde{\nu}_1^{j+2} = \mu_1^{j+2}$ which is finite. Hence $\lim_{T\to 0} \frac{\lambda_{j+1}}{\lambda_j} = 0$ for $j = 1, \ldots, n-1$. ∎

Now we state the following proposition about convergence of DIRM in the limit $T \to 0$.

PROPOSITION 4.11. *Suppose there exists a T_1 such that for all $T < T_1$ a fixed point x^* of $\mathcal{I}$ exists and that $P^* = \mathcal{R}x^*$ is well defined. Further suppose x^* satisfies the*

conditions of Lemma 4.10. *Let* X *be the true solution. Then there exists a* $T_0 < T_1$ *such that DIRM converges to* x^**, for all* $T < T_0$*, for initial guesses* $x^{(0)}$ *that are sufficiently close to* x^*.

Proof. Application of Lemma 4.10 immediately yields that

$$\lim_{T\to 0} \mathcal{C}_1 = 2$$

and

$$\lim_{T\to 0} \mathcal{C}_2 = 0.$$

Also, as $T \to 0$, both $\|F_1\|$ and $\|F_2\| \to 0$ (Lemma 4.3), and hence $\|F\| \to 0$ as well. So $\|D\mathcal{I}(x^*)\| \to 0$ as $T \to 0$, and in particular $\|D\mathcal{I}(x^*)\| < 1$ for all small enough T. Hence DIRM will converge for all initial guesses $x^{(0)}$ that are sufficiently close to x^*. ∎

5. Examples.

5.1. Example: Nonlinear power grid. In this section we present a power grid model and apply the DIRM method to simulate the transient behavior. The model has been taken from the paper [11]. The model equations we use represent the coupling between power flows and frequency variations across power networks and are known as the *swing equations*. Swing dynamics potentially interact with protection mechanisms and may lead to cascading failures. See [11] for more details.

We used a power grid consisting of 36 nodes arranged in a 6×6 square grid. Each node is either a generator or a load. For general types of load situations we would need DAEs to represent the system. Here we assume that the loads are all synchronous motors. In that case the swing equations involve the variables δ_i, where indices $i = 1, \ldots, N$ denote the nodes ($N = 36$ in our example). Physically at node i, δ_i stands for the generator or motor rotor angle with respect to a synchronously rotating reference frame. The equations are then given by

$$M_i\ddot{\delta}_i + D_i\dot{\delta}_i = P_{mi} - P_{gi}, \qquad i = 1, \ldots, N, \tag{5.1}$$

where M_i and D_i are inertia and damping terms for the generator or motor at the ith node and P_{mi} is the mechanical power input to the generator or the mechanical power output (negative) of the motor at the ith node, and P_{gi} is the electrical power output from the ith node. It is assumed that the voltage magnitudes at the nodes are maintained fixed by regulators. The electrical power P_{gi} is given by

$$\begin{aligned} P_{gi} = Re(V_i^* I_i) &= Re\left(V_i^* \sum_{j=1}^{N} Y_{ij} V_j\right) \\ &= -\sum_{j=1}^{N} |V_i||V_j| b_{ij} \sin(\delta_i(t) - \delta_j(t)), \qquad i = 1, \ldots, N, \end{aligned} \tag{5.2}$$

where $V_i = |V_i|e^{i\delta_i}$, and $Y = G + iB$ is the admittance matrix for the network connections (with some of the i denoting $\sqrt{-1}$). We assume that the lines are lossless ($G = 0$). The b_{ij} are the terms of the *susceptance* matrix B. The diagonal entrees b_{ii} are all zero. If a line is not present between nodes i and j, then $b_{ij} = 0$. We chose $b_{ij} = 1$ for all connected lines.

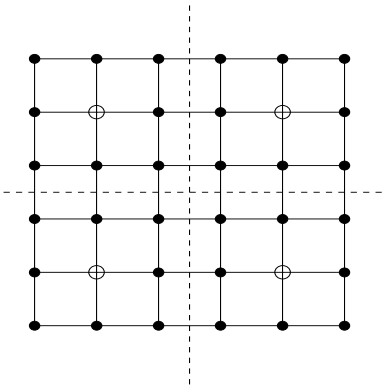


FIG. 5.1. *Power grid: generators—open circles; loads—filled circles.*

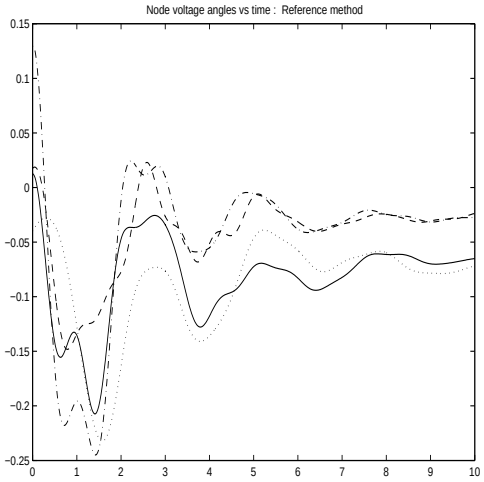


FIG. 5.2. δ_i *vs. time: reference method.*

The grid we chose is shown in Figure 5.1. The generators are marked by open circles, and all other nodes (filled circles) are motors. All physical parameters chosen were nondimensional. We chose all the voltage magnitudes to be the same (constant) value of 1. The mechanical powers P_{mi} were chosen to be -0.0880 for the motor nodes, and generator powers were all 0.7040 so that the total sums to zero. The parameter values $M_i = 2.0$ and $D_i = 0.8$ were chosen for the generators, while values $M_i = 0.1$ and $D_i = 0.1$ were chosen for all load motors. Given these parameters the swing equations have a nontrivial (not all δ_i are zero) steady state solution. We picked a random initial condition (Gaussian with zero mean and 0.1 standard deviation for both δ_i and $\dot{\delta}_i$ for all i). First, we used the MATLAB ODE solver `ode15s` to solve the equations in the interval $[0, 10]$, which indicated that the steady state was more or less reached in that time interval. A plot of the "reference solution" for δ_i for four of the nodes is shown in Figure 5.2. One motor node from each subsystem was chosen for this plot. In order to apply the DIRM method we modularized the system so that the square grid of 6×6 nodes was split into four subsystems, each consisting of a

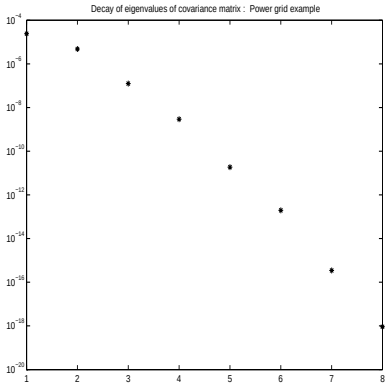


FIG. 5.3. *Decay of POD mode energies: power grid example.*

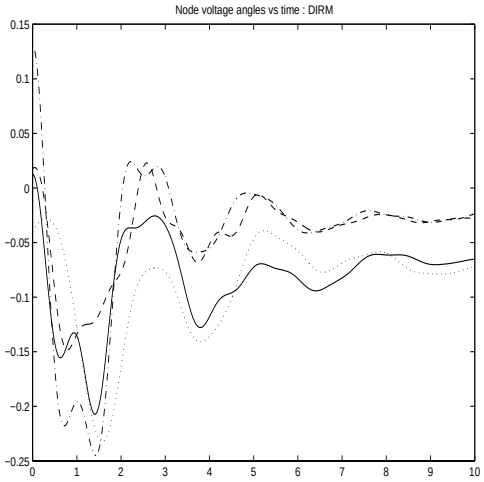


FIG. 5.4. δ_i *vs time: DIRM.*

square subgrid of 3×3 nodes. The broken lines in Figure 5.1 show this partitioning. DIRM did not converge when the simulation interval was $[0, 10]$. However, breaking the interval into smaller intervals achieved convergence. Using intervals of length 0.5 (i.e. $[0, 0.5]$, $[0.51]$, $\ldots$, $[9.5, 10]$) achieved convergence. We used the same MATLAB solver `ode15s` as the underlying solver. Within each interval at most seven iterations were required. The final value of the solution from one interval was used as the initial condition for the next interval. Initial reduced models were obtained by simulating each subsystem in isolation (all other subsystem states were assumed to remain zero) for the time interval under consideration. The reduced order models of all subsystems were chosen to be of dimension 3. (The full dimension of each subsystem is 18.) Figure 5.3 shows the largest eight eigenvalues of the DIRM solution for the first subsystem trajectory in the final simulation interval $[9.5, 10]$. The decay was similar for the other subsystems as well.

The solution obtained using DIRM is shown in Figure 5.4 for the same nodes as in Figure 5.2. The solutions of the reference method and DIRM are visually

indistinguishable and so we plotted them on separate figures. We also computed the maximum relative error e_r defined by

$$e_r = \sup_i \frac{\sup_{t\in[0,10]} |\hat{x}_i(t) - x_i(t)|}{\sup_{t\in[0,10]} |x_i(t)|}, \tag{5.3}$$

where i indexed over all states (node voltage phases and velocities). It was $e_r = 0.0037$; i.e., the relative error for any state variable was less than 0.37%.

An important point during swing dynamic transients is that large deviations of δ_i and $\dot{\delta}_i$ can trigger protection mechanisms that can shut down a line or a generator, and this changes the system parameters discontinuously (some b_{ij} change to 0, for instance) which in turn leads to further transients and so on (see [11]). Since our scheme predicts the solutions accurately, we expect it to predict the first failure (location and time) accurately. However, for accurate prediction beyond the first failure, we have to restart our iteration for a time interval beginning at the failure. Numerical simulations done by Cao and Petzold [1] revealed that the reduced model formed from trajectories obtained before a failure was not accurate after the failure and could not be used to predict further failures. If we include failures in our model, our iterative method may need to be modified. During the iterative simulations, if a subsystem indicates failure, then the time interval of simulation may be shortened so that the reduced models remain more accurate. We have not yet numerically investigated this type of scenario. This is a subject of future research.

5.2. PDE Example: Comparison with WR. We considered the reaction-convection-diffusion equation

$$x_t = \nu x_{ss} + a x_s + b x \tag{5.4}$$

in one spatial variable s in the interval $s \in [0, 6]$ with spatial discretization giving 100 equally spaced interior points. Both first and second spatial derivatives were approximated by centered differences. We used the triangular function

$$\begin{aligned} &x(0,s) = s/3, \qquad 0 \le s \le 3, \\ &x(0,s) = 1 - (s-3)/3, \qquad 3 \le s \le 6, \end{aligned} \tag{5.5}$$

as initial condition, and the boundary conditions were zero. This results in a tridiagonal linear system (ODE) of the form

$$\dot{x} = Ax,$$

where $x \in \mathbb{R}^{100}$.

For DIRM we divided the system into 10 subsystems each of size 10 consisting of adjacent grid points. For WR we used the splitting $A = A_d + A_o$, where A_d is block diagonal with 10×10 blocks and A_o the remaining off-diagonal coupling part. This corresponds to the Jacobi version of WR. We also used overlapping for WR in order to improve its convergence. It is known from the work of Miekkala that in the case of tridiagonal systems the order of convergence of WR is $\omega = 1$ (as defined by [8]), which is very slow. However, if we overlap by o variables, then $\omega = \frac{1}{o+1}$, and this should improve the asymptotic convergence [8].

Simulations of three different sets of parameter values are discussed here. In all cases we computed the solutions using the MATLAB ODE solver `ode15s` to provide a benchmark. The same solver was also used within DIRM and WR. In all simulations of DIRM and WR we used the convergence tolerance (see (3.3)) $tol = 0.001$ and a

TABLE 5.1
Convergence and accuracy of DIRM and WR for the 1D PDE with convection and diffusion. Parameter values $\nu = 0.1, a = 1, b = 0$. Simulation interval $[0, 10]$. Subsystem reduced model dimension $k = 3$.

	Number of iterations	Maximum error over subsystems
DIRM	3	1.3112×10^{-3}
WR overlap $o = 3$	21	12.2107×10^{-3}
WR (no overlap)	> 30 (did not converge)	241.8189×10^{-3}

TABLE 5.2
Convergence and accuracy of DIRM and WR for the 1D PDE with reaction, convection, and diffusion. Parameter values $\nu = 0.1, a = 6, b = 6$. Simulation interval $[0, 1.2]$. Subsystem reduced model dimension $k = 3$.

	Number of iterations	Maximum error over subsystems
DIRM	11	0.5284
WR overlap $o = 3$	16	0.7861
WR (no overlap)	> 30	39.5314

TABLE 5.3
Convergence and accuracy of DIRM and WR for the 1D PDE with pure diffusion. Parameter values $\nu = 0.1, a = 0, b = 0$. Simulation interval $[0, 10]$. Subsystem reduced model dimension $k = 3$.

	Number of iterations	Maximum error over subsystems
DIRM	2	0.4511×10^{-3}
WR overlap $o = 5$	21	1.9688×10^{-3}

maximum iteration count of 30. All the subsystem reduced models in DIRM were of dimension $k = 3$. The convergence and error results are summarized in Tables 5.1, 5.2, and 5.3. The error is measured by $\sup\{\|\hat{x}_i(t) - x_i(t)\|_2 : t \in [0, T],\ i = 1, \ldots, 10\}$, where $x_i(t)$ and $\hat{x}_i(t)$ are the benchmark solution and the iterative method (DIRM or WR) solution of the ith subsystem, respectively. The WR method in general needed some overlapping to achieve convergence. Especially the pure diffusion case required a greater overlap without which WR did not converge at all.

Note that couplings exist only between adjacent subsystems. This is essentially a property of 1D PDEs. Thus for the WR method we expect at least 10 iterations before the first subsystem "sees" the last subsystem. This is true even when overlapping is used. Hence we may expect at least 10 iterations (initial overhead) before we see convergence of WR. In DIRM, the entire system is always being simulated, albeit in a partially reduced form. Hence we do not expect this adverse effect. In order to test this hypothesis we considered the system with $\nu = 0.1, a = 1$, and $b = 0$. This gives rise to a system with diffusion as well as convection propagating towards the decreasing s direction ($a > 0$). This can be seen from Figure 5.5, where the initial condition as well as the solution (benchmark solver) at $t = 5$ are shown. This may be thought of as the initial triangle diffusing and propagating to the left at the same time. Thus there is more "information flow" from right to left than from left to right. In order to capture the waveform for the first subsystem (the leftmost 10 grid points) accurately we need to capture the information flow from the other subsystems. Hence we could expect that the WR method would take at least 10 iterations for convergence and also expect it to converge slower for the first (leftmost) subsystem than for the last (rightmost) subsystem. We picked a simulation interval $[0, 10]$ so that the entire system decayed to zero. We found that WR did not converge even after 30 iterations, and hence used overlapping by three variables, keeping the number of subsystems the

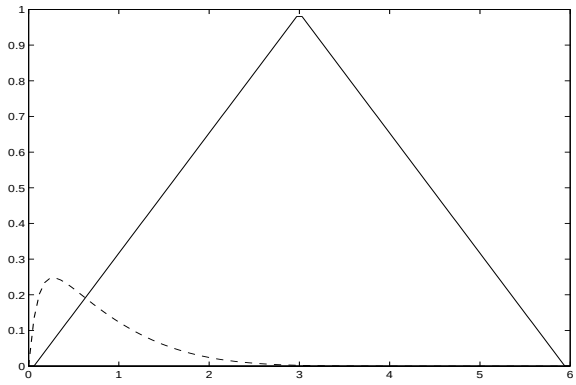


FIG. 5.5. *Plot of the benchmark solution $x(t,s)$ versus s at time $t=0$ (solid) and at time $t=5$ (dashed) for the convection-diffusion case with $\nu=0.1, a=1, b=0$.*

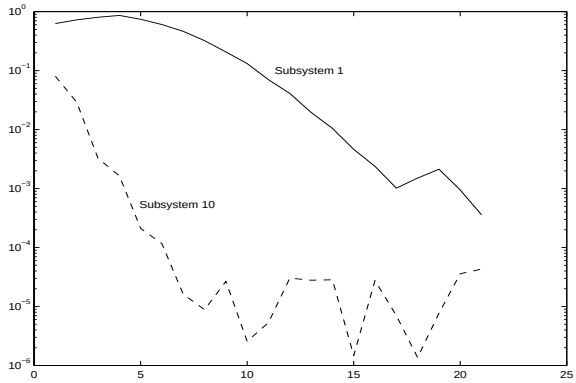


FIG. 5.6. *Convergence of subsystems 1 and 10 for WR with overlap $o=3$ for the convection-diffusion case with $\nu=0.1, a=1, b=0$. Plot of the difference $\sup\{\|x_i^{(j)}(t)-x_i^{(j-1)}(t)\|_2 : t\in[0,T]\}$ between successive iterates versus the iteration step j for subsystems $i=1$ and $i=10$.*

same. This resulted in the first nine subsystems being of size 13 and the last one being of size 10. The convergence plots for the overlapped WR for the first and last subsystems are shown in Figure 5.6, which confirms our intuition. The DIRM method did not experience this overhead and converged in three iterations. Figures 5.7 and 5.8 compare the convergence rates and Figure 5.9 compares the subsystem errors for DIRM and WR.

Second, we considered the system with $\nu=0.1, a=6$, and $b=6$. This system has a dominant reactive term. We picked the simulation interval $[0,1.2]$ in which the system had an explosive reaction after which it decayed to zero. The benchmark simulation showed that subsystem 1 (leftmost) underwent the most explosive change; see Figure 5.10. DIRM and WR with overlapping ($o=3$) performed comparably. Even though the maximum error (Table 5.2) seems large for both DIRM and WR with overlap, it occurred in subsystem 1 (which underwent the most explosive growth), and it is small compared to the peak value of the subsystem trajectory. In fact it was hard to visually distinguish the trajectory plots of subsystem 1 for DIRM and overlapped WR from those of the benchmark solution in Figure 5.10.

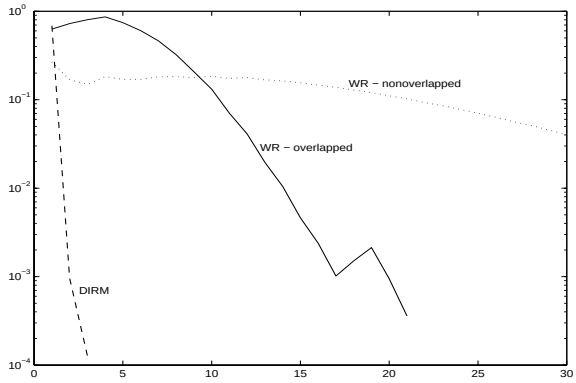


FIG. 5.7. *Convergence of subsystem* 1 *for WR with overlap* $o = 3$*, WR without overlap, and DIRM for the convection-diffusion case with* $\nu = 0.1, a = 1, b = 0$*. Plot of the difference* $\sup\{\|x_i^{(j)}(t) - x_i^{(j-1)}(t)\|_2 : t \in [0,T]\}$ *between successive iterates versus the iteration step* j *for the subsystem* $i = 1$*.*

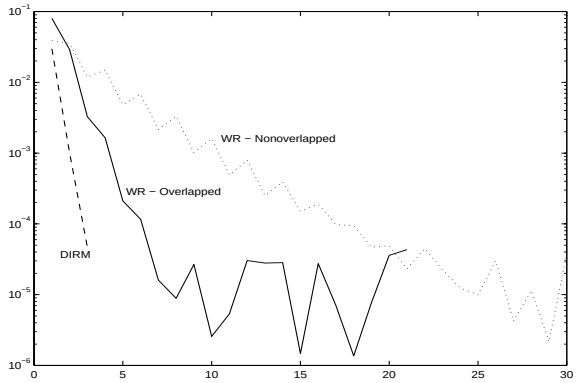


FIG. 5.8. *Convergence of subsystem* 10 *for WR with overlap* $o = 3$*, WR without overlap, and DIRM for the convection-diffusion case with* $\nu = 0.1, a = 1, b = 0$*. Plot of the difference* $\sup\{\|x_i^{(j)}(t) - x_i^{(j-1)}(t)\|_2 : t \in [0,T]\}$ *between successive iterates versus the iteration step* j *for the subsystem* $i = 10$*.*

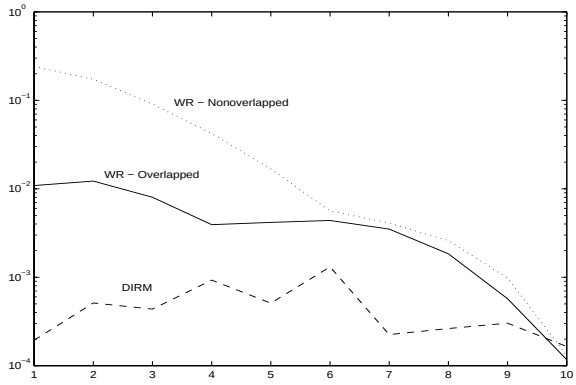


FIG. 5.9. *Errors of each subsystem for WR with overlap* $o = 3$*, WR without overlap, and DIRM for the convection-diffusion case with* $\nu = 0.1, a = 1, b = 0$*. Plot of the error in* i*th subsystem given by* $\sup\{\|\hat{x}_i(t) - x_i(t)\|_2 : t \in [0,T]\}$*, where* $\hat{x}_i$ *is the iterative method solution and* x_i *is the benchmark solution versus the subsystem* i*.*

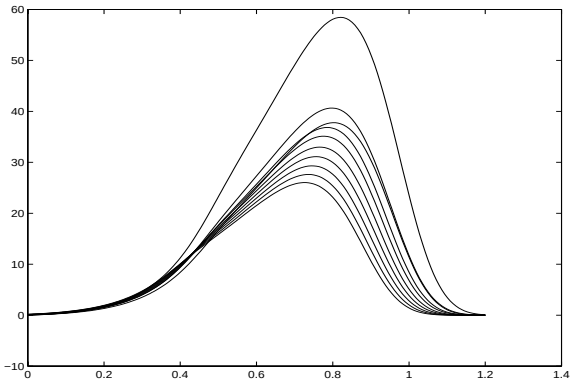


FIG. 5.10. *Plot of the benchmark solution of all* 10 *of the states of subsystem* 1 *for the reaction-convection-diffusion case with* $\nu = 0.1, a = 6, b = 6$.

Finally, we simulated the purely diffusive system with $\nu = 0.1, a = 0$, and $b = 0$. In this case WR performed worst, and DIRM performed best, showing a very clear advantage over WR. Overlapping by $o = 3$ was not sufficient to achieve convergence of WR. However, a bigger overlap of $o = 5$ achieved convergence.

In summary, DIRM seems to converge better than WR for both the convection-diffusion and pure diffusion cases, with the pure diffusion case being the strongest point for DIRM. For the reaction-convection-diffusion case there is no clear winner. Since the error in the POD method is large when system trajectories undergo explosive growth in the interval of interest ([13]), it is not surprising that DIRM performed worst for this type of equation. In contrast, WR seems to have performed best for the reaction-convection-diffusion case.

5.3. Example: Almost coincident eigenvalues $\lambda_k \approx \lambda_{k+1}$ and the modified DIRM method. The convergence analysis of DIRM in section 4.6 predicts difficulties in convergence if the eigenvalues λ_k and λ_{k+1} of the covariance matrix of the fixed point trajectory of any subsystem are very close. To generate such an example, we employ the equation

$$\dot{x} = A(x - f(t)) + f'(t), \tag{5.6}$$

which for any choice of A has $x = f(t)$ as the solution corresponding to the initial condition $x(0) = f(0)$.

Consider the following trajectory $g(t) \in \mathbb{R}^3$ such that the covariance matrix for the interval $[0, 1]$ has prescribed eigenvalues λ_1, λ_2, and λ_3:

$$g(t) = \left(\sqrt{2\lambda_1}\sin(2\pi t), \sqrt{2\lambda_2}\cos(2\pi t), \sqrt{2\lambda_3}\sin(4\pi t)\right).$$

Note that $g(0) = (0, \sqrt{2\lambda_2}, 0)$. We may construct a system of the form (5.6) which is six dimensional, and when split into two subsystems (each of dimension 3) the subsystem solution trajectories are both $g(t)$. According to the analysis of section 4.5, if A is diagonally dominant, and its fundamental modes do not grow substantially in the interval of simulation $[0, 1]$, and the POD projection error is small, then we

expect the fixed point trajectories of the subsystems to be close to the true solution $g(t)$. Thus for the system in (5.6), if we set $f(t) = (g(t), g(t))$, $\lambda_3 = \lambda_2$, and choose an A with the above properties and the initial condition $x(0) = (0, \sqrt{2\lambda_2}, 0, 0, \sqrt{2\lambda_2}, 0)$, then we can expect difficulties in convergence.

Numerical experiments were conducted with A obtained from discretizing the PDE in section 5.2 with $\nu = 1$, $a = b = 0$, and six interior points. The choice of $\lambda_1 = 0.5, \lambda_2 = \lambda_3 = 0.2$ lead to eight iterations for convergence; the choice of $\lambda_1 = 0.5, \lambda_2 = 0.2, \lambda_3 = 0.199$ required 12 iterations; and the choice $\lambda_1 = 0.5, \lambda_2, \lambda_3 = 0.19$ converged in four iterations. In other cases, where λ_2 and λ_3 were further separated, convergence took a similar number of (five or less) iterations.

Remark 5.1. Note that the $\lambda_2 = 0.2, \lambda_3 = 0.19$ case was worse than the $\lambda_2 = \lambda_3 = 0.2$ case because the fixed point trajectory is slightly different from the true solution trajectory. As a verification of our analysis we computed the quantities $\mathcal{C}_1$ and $\mathcal{C}_2$ of equations (4.30) for the fixed point trajectories of the above examples and observed that the larger they were the more the iterations needed for convergence.

In order to make DIRM robust against such situations we tried the following "modified DIRM" method. In modified DIRM, reduced models are computed from a covariance matrix and mean value that are a weighted combination of the covariance matrix and mean value of the current trajectory and some precomputed covariance matrix and mean which are typically obtained from an ensemble of trajectories such as those with initial conditions uniformly chosen from the unit sphere. Thus at the αth iteration, given trajectory $x^{(\alpha)}$ of a subsystem we compute the mean $\bar{x}^{(\alpha)}$ and the covariance matrix $R^{(\alpha)}$ of that trajectory as usual. However, instead of using $\bar{x}^{(\alpha)}$ and $R^{(\alpha)}$ to obtain the reduced model we compute

$$\tilde{R}^{(\alpha)} = \beta R^{(\alpha)} + (1-\beta) R_0,$$
$$\tilde{\bar{x}}^{(\alpha)} = \beta \bar{x}^{(\alpha)} + (1-\beta)\bar{x}_0,$$

where $0 < \beta \leq 1$, and then compute the projection $\tilde{P}^{(\alpha)}$ corresponding to $\tilde{R}^{(\alpha)}$. The new reduced model is given by $\tilde{\bar{x}}^{(\alpha)}$ and $\tilde{P}^{(\alpha)}$.

Here R_0 and $\bar{x}_0$ are precomputed from some ensemble of trajectories with various initial conditions and do not change from iteration to iteration. This method changes the fixed point of DIRM. Typically if β is small you would expect the method not to be so accurate.

When we applied the modified DIRM with $\beta = 0.9$, for the worst case above ($\lambda_1 = 0.5, \lambda_2 = 0.2, \lambda_3 = 0.199$), we converged in four iterations, with more or less the same accuracy as the unmodified DIRM. It is important to note that the precomputed covariance matrix R_0 should have reasonable separation between its kth and $k+1$st eigenvalues. Then for sufficiently small β we are guaranteed to have sufficient separation of λ_k and λ_{k+1}, the eigenvalues of $\tilde{R}$. This example confirms our analysis, and the modified DIRM method provides a safeguard against the situation of coincident eigenvalues.

Remark 5.2. In the modified DIRM method, it is easy to see that when $\beta = 0$ we have immediate convergence, since the reduced models do not change. However, its accuracy is not as good as that of the regular DIRM method. When $\beta = 1$ we have the regular DIRM method, which is more accurate, but does not always have good convergence behavior. So one may expect that by choosing an appropriate β we could achieve a good compromise between convergence and accuracy. However, the numerical experiments with the PDE example of section 5.2 showed that the convergence rate did not depend monotonically on β. For the parameter values

($\nu = 0.1, a = 6, b = 6$), subsystem reduced model dimension $k = 2$, and the time interval $[0, 0.5]$, convergence took 12 iterations for $\beta = 1$, 10 iterations for $\beta = 0.8$, 13 iterations for $\beta = 0.5$, and for the $\beta = 0$ case 1 iteration (as expected). This could be because there is an intermediate regime of β values where the loss of accuracy affects the convergence negatively. It should be noted that the WR method took more than 15 iterations.

5.4. Example: Fixed point trajectory with zero POD projection error. For the example in section 5.3, if we choose $\lambda_1 = 0.5, \lambda_2 = 0.2$, and $\lambda_3 = 0$, we have a solution trajectory which has zero POD projection error. The fixed point trajectory being close to the true solution also had almost zero POD projection error. Yet DIRM took four iterations to converge, in agreement with the convergence analysis (see section 4.6).

When trajectories of all the subsystems always lie in a k dimensional subspace (where k is the dimension of all the reduced order models), the POD projection error is always zero, and in this case convergence is immediate (in agreement with the convergence analysis). This can be numerically observed from an example $\dot{x} = Ax$, where rows of A corresponding to each subsystem are of rank k or less. When the row rank is strictly less than k we have zero POD projection error as well as coincident eigenvalues ($\lambda_k = \lambda_{k+1} = 0$). In this case DIRM still converges immediately, even though the projection matrices computed at each step may not converge.

6. Conclusions and future work. We have presented a new dynamic iteration called DIRM for simulation of large scale interconnected systems. DIRM uses reduced order models (obtained here via POD) of subsystems which are also refined during the iterations. We provided an analysis of the DIRM method as applied to linear time invariant systems of ODEs consisting of two subsystems, giving results on accuracy and convergence of DIRM. We also presented numerical examples, including some special cases chosen to test the validity of the analysis and to illustrate some special situations, and two realistic examples: a nonlinear power grid model and a discretized linear reaction-convection-diffusion type PDE in one dimension. Both the power grid and the PDE examples demonstrated the success of DIRM. In the PDE example we also provided comparisons with WR. This example showed that DIRM has clear advantages over WR for pure diffusion and convection-diffusion-type equations. DIRM performed worst for systems showing explosive reactions, for which WR performed best.

Future work will include DAEs as well as hybrid systems such as the power grid with failure models resulting in discontinuous changes in system parameters. The complementary nature of DIRM and WR seen in the PDE example suggests that the development of an approach that combines the two methods in an optimal manner might prove valuable in parallel computation of large scale systems. The framework of DIRM allows for the use of model reduction techniques other than POD provided that they are data driven. Since the POD reduced models may not achieve considerable savings for nonlinear banded Jacobian systems [13], it might be advantageous to explore the use of other model reduction methods.

Acknowledgments. We would like to thank John C. Doyle for suggesting the general notion of iterating reduced order models of subsystems. We would also like to thank Pablo Parrilo for providing us with the software for the power grid generation.

REFERENCES

[1] Y. Cao and L. Petzold, *A Note on Model Reduction for Analysis of Cascading Failures in Power Systems*, manuscript.

[2] M. Gander, *A waveform relaxation algorithm with overlapping splitting for reaction diffusion equations*, Numer. Linear Algebra Appl., 6 (1999), pp. 125–145.

[3] S. Glavaski, J. Marsden, and R. Murray, *Model reduction, centering, and the Karhunen-Loève expansion*, in Proceedings of the IEEE Control and Decision Conference, Tampa, FL, 1998, pp. 2071–2076.

[4] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, London, 1996.

[5] P. Holmes, J. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1996.

[6] S. Lall, J. Marsden, and S. Glavaski, *Empirical model reduction of controlled nonlinear systems*, in Proceedings of the 14th IFAC World Congress, Beijing, People's Republic of China, 1999, pp. 473–478.

[7] E. Lelarasmee, A. Ruehli, and A. Sangiovanni-Vincentelli, *The waveform relaxation method for time-domain analysis of large scale integrated circuits*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 1 (1982), pp. 131–145.

[8] U. Miekkala, *Remarks on waveform relaxation method with overlapping splittings*, J. Comput. Appl. Math., 88 (1997), pp. 349–361.

[9] U. Miekkala and O. Nevanlinna, *Quasinilpotency of the operators in Picard-Lindelöf iteration*, Numer. Funct. Anal. Optim., 13 (1992), pp. 203–221.

[10] B. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–31.

[11] P. Parrilo, F. Paganini, G. Verghese, B. Lesieutre, and J. Marsden, *Model reduction for analysis of cascading failures in power systems*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 4028–4212.

[12] I. Pérez-Arriaga, G. Verghese, F. Pagola, J. Sancha, and F. Schweppe, *Developments in selective modal analysis of small-signal stability in electric power systems*, Automatica, 26 (1990), pp. 215–231.

[13] M. Rathinam and L. Petzold, *A new look at proper orthogonal decomposition,* SIAM J. Numer. Anal., submitted.

[14] M. Rathinam and L. Petzold, *An iterative method for simulation of large scale modular systems using reduced order models*, in Proceedings of the IEEE Control and Decision Conference, Sydney, Australia, 2000.

[15] K. Zhou and J. Doyle, *Essentials of Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1475–1490

AN H^1-GALERKIN MIXED FINITE ELEMENT METHOD FOR AN EVOLUTION EQUATION WITH A POSITIVE-TYPE MEMORY TERM*

AMIYA K. PANI† AND GRAEME FAIRWEATHER‡

Abstract. An H^1-Galerkin mixed finite element method is analyzed for a class of evolution equations with memory. When a classical mixed method is applied to such problems, it has not been possible to obtain any estimate for the flux. However, the proposed approach yields optimal order convergence without the LBB consistency condition and quasi uniformity of the finite element mesh. Compared to the results proved for one space variable, the L^2 estimate of the flux is not optimal for problems in two and three space dimensions. Therefore, a modification of the method is proposed and analyzed. A maximum norm estimate is also derived in one and two space variables. A backward Euler approximation of the modified method is also analyzed.

Key words. evolution equation, positive-type memory, mixed finite element method, H^1-Galerkin, LBB condition, semidiscrete schemes, backward Euler method, optimal error estimates, flux estimates, several space variables

AMS subject classifications. Primary, 65M12, 65M15, 65M60; Secondary, 45K05

PII. S0036142900372318

1. Introduction.

Consider the initial boundary value problem

$$\begin{aligned} u_t - \int_0^t \beta(t-s)\nabla\cdot(a(x)\nabla u(s))\,ds &= f(x,t), \quad (x,t)\in\Omega\times(0,T], \\ u(x,t) &= 0, \quad (x,t)\in\partial\Omega\times(0,T], \\ u(x,0) &= u_0(x), \quad x\in\Omega, \end{aligned} \tag{1.1}$$

where $u_t = \partial u/\partial t$, Ω is a bounded domain in R^d $(d = 1, 2, 3)$ with boundary $\partial\Omega$, and $a_1 \geq a(x) \geq a_0 > 0$, $x \in \Omega$, for positive constants a_1 and a_0. The kernel β is assumed to be positive definite, i.e., for each $t \in (0, T]$, $\beta \in L^1_{loc}(0,\infty)$, and

$$\int_0^t \left(\int_0^s \beta(s-\tau)v(\tau)\right) v(s)\,d\tau\,ds \geq 0, \quad v\in C[0,t]. \tag{1.2}$$

(Note that (1.2) holds if and only if

$$\Re\hat{\beta}(iy) = \int_0^\infty \beta(t)\cos(yt)\,dt \geq 0, \quad y\in R,$$

because, for any $\psi \in L^2(R_+)$,

$$\int_0^\infty \psi(t)\int_0^t \beta(t-s)\psi(s)\,ds\,dt = \frac{1}{2\pi}\int_{-\infty}^\infty \Re\hat{\beta}(iy)|\hat{\psi}(iy)|^2\,dy,$$

*Received by the editors May 11, 2000; accepted for publication (in revised form) April 1, 2002; published electronically October 23, 2002.

http://www.siam.org/journals/sinum/40-4/37231.html

†Department of Mathematics, Industrial Mathematics Group, Indian Institute of Technology, Bombay, Powai, Mumbai-400076, India (akp@math.iitb.ac.in). The research of this author was supported by the Department of Mathematical and Computer Sciences, Colorado School of Mines.

‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (gfairwea@mines.edu). The research of this author was supported by National Science Foundation grant DMS-9805827.

where $\hat{\psi}$ is the Laplace transform of ψ.) Problems of the type (1.1) and a nonlinear version thereof occur in viscoelasticity and heat conduction in materials with memory; see, for example, [8, 18].

Over the last decade, various numerical methods have been applied to (1.1) for both smooth and nonsmooth kernels. In [19], a spectral method is employed for the one dimensional form of (1.1) with a weakly singular kernel, while a finite difference scheme is used to discretize the spatial variable in [7]. Error estimates for methods based on an orthogonal spline collocation discretization in space combined with backward Euler or Crank–Nicolson-type time stepping schemes for (1.1) with $a = 1$ are derived in [3], where a modified spline collocation method for a one dimensional problem is also discussed. In [22], the Laplace transform is used in time with orthogonal spline collocation in space. In [9] and [10], finite element Galerkin methods are applied to discretize (1.1) in space, with finite difference schemes in time for both smooth and nonsmooth kernels. Convergence is discussed for smooth initial data and constant time step size in [9], and variable time stepping schemes are examined in [10]. In all of these papers, the positivity of the kernel plays a key role in the stability and convergence analyses.

To approximate both u and its flux $\boldsymbol{\sigma} := a\nabla u$ accurately, we reformulate (1.1) as the first order system

$$\boldsymbol{\sigma} = a\nabla u, \quad u_t - \int_0^t \beta(t-s)\nabla\cdot\boldsymbol{\sigma}(s)\,ds = f. \tag{1.3}$$

A standard procedure is to apply a classical mixed method to this system, but for this method it has not been possible to obtain an error estimate for the flux. The purpose of the present paper is to apply to (1.3) an H^1-Galerkin mixed method based on that proposed in [14] and to derive optimal error estimates for u in $L^\infty(L^2)$ and $L^\infty(H^1)$ and for $\boldsymbol{\sigma}$ in $L^\infty(L^2)$. It is observed that, compared to the results proved for problem (1.1) in one space variable, the $L^\infty(L^2)$-norm estimate of the flux is not optimal for the problem in two or three space variables. Therefore, a modification is considered in which a term containing the curl of the flux is added. This modification leads to a strong coercivity property of $\boldsymbol{\sigma}$ and facilitates the derivation of an optimal estimate in $L^\infty(L^2)$ for the flux.

As a consequence of our analysis, we obtain a maximum norm estimate for u in one and two space variables. We note that this has not been derived previously for the standard Galerkin method.

The proposed methods have several attractive features. First, they are not subject to the LBB consistency condition. The finite element spaces V_h (for approximating u) and $\mathbf{W_h}$ (for approximating the flux $\boldsymbol{\sigma}$) may be of differing polynomial degrees. Moreover, the L^2- and H^1-error estimates do not require the finite element mesh to be quasi-uniform. Related studies on least squares mixed methods for elliptic equations can be found in [4, 6, 11, 12, 13, 16] and references therein.

In the past, attempts have been made to circumvent the LBB-consistency condition by adding least square like terms, which are generally mesh dependent, to the classical mixed formulations for elliptic problems (see, for example, [1, 5]). These methods are known as stabilizing mixed finite element methods. In contrast to the stabilizing methods, the present formulation deals directly with the original equation and is not based on classical methods. In fact, it is not possible to apply even the classical mixed methods to derive any estimate of the flux for problem (1.1). The approach proposed in this article is not a stabilizing procedure and does not use mesh

dependent parameters. The reason for adding the curl term is simply to obtain a strong coercivity property.

A brief outline of this paper is as follows. In section 2, error estimates are derived for the semidiscrete case in one space variable, while in section 3 error bounds are obtained for (1.1) in two and three space variables, again for the semidiscrete case. Since the estimates are not sharp compared to the one dimensional case, in section 4, a modified H^1-Galerkin mixed finite element method is formulated, and error estimates are established for the semidiscrete case. In section 5, a discrete time backward Euler scheme is considered for the time discretization of the modified H^1-Galerkin mixed method, and error estimates are derived. Finally, in section 6, the key results of the paper are summarized and a possible extension of them indicated.

Throughout this paper, C denotes a generic positive constant which does not depend on the spatial and time discretization parameters h and Δt but may depend on T and $\|\beta\|_{L^1(0,T)}$. Also, we make use of the following: for a function $\phi \in H^1(0,T)$,

$$\phi(t) = \phi(0) + \int_0^t \phi_t(s)\, ds. \tag{1.4}$$

2. Error estimates for the semidiscrete case in one space variable. Consider the one dimensional form of (1.1),

$$\begin{aligned} u_t - \int_0^t \beta(t-s)(a(x)u_x(s))_x\, ds &= f(x,t), \quad (x,t) \in \Omega \times (0,T], \\ u(0,t) = u(1,t) &= 0, \quad t \in (0,T], \\ u(x,0) &= u_0(x), \quad x \in \Omega, \end{aligned} \tag{2.1}$$

where $\Omega = (0,1)$, $0 < T < \infty$, $a_1 \geq a(x) \geq a_0 > 0$, $x \in \Omega$, for positive constants a_1 and a_0. Setting $\sigma = au_x$, we rewrite (2.1) as the first order system

$$a(x)u_x = \sigma, \quad u_t - \int_0^t \beta(t-s)\sigma_x(s)\, ds = f. \tag{2.2}$$

We denote the natural inner product in $L^2(\Omega)$ by $(\cdot,\cdot)$ and the norm by $\|\cdot\|$, and let $H_0^1 = \{v \in H^1(\Omega) : v(0) = v(1) = 0\}$. Further, we use the classical Sobolev spaces $W^{m,p}(\Omega)$, $1 \leq p \leq \infty$, denoted by $W^{m,p}$, with norm $\|\cdot\|_{m,p}$. When $p = 2$, we simply write $W^{m,p}$ as H^m with norm $\|\cdot\|_m$. To derive the H^1-Galerkin mixed finite element method, we consider the following weak formulation of (2.2): find $\{u, \sigma\}$ satisfying

$$\begin{aligned} (au_x, v_x) &= (\sigma, v_x), \quad v \in H_0^1, \\ (\alpha\sigma_t, w) + \int_0^t \beta(t-s)\,(\sigma_x(s), w_x)\, ds &= -(f, w_x), \quad w \in H^1, \end{aligned} \tag{2.3}$$

where

$$\alpha = \frac{1}{a}. \tag{2.4}$$

Let V_h and W_h be finite dimensional subspaces of H_0^1 and H^1, respectively, with the following approximation properties: for $1 \leq p \leq \infty$ and positive integers k and r,

$$\inf_{v_h \in V_h} \{\|v - v_h\|_{L^p} + h\|v - v_h\|_{W^{1,p}}\} \leq Ch^{k+1}\|v\|_{W^{k+1,p}}, \quad v \in H_0^1 \cap W^{k+1,p},$$

and

$$\inf_{w_h \in W_h} \{\|w - w_h\|_{L^p} + h\|w - w_h\|_{W^{1,p}}\} \le Ch^{r+1}\|w\|_{W^{r+1,p}}, \quad w \in W^{r+1,p}.$$

With the finite element spaces V_h and W_h, we define the semidiscrete H^1-Galerkin mixed finite element approximation $\{u_h, \sigma_h\} : [0,T] \mapsto V_h \times W_h$ by

$$\begin{aligned} &(au_{hx}, v_{hx}) = (\sigma_h, v_{hx}), \quad v_h \in V_h, \\ &(\alpha\sigma_{ht}, w_h) + \int_0^t \beta(t-s)\,(\sigma_{hx}(s), w_{hx})\,ds = -(f, w_{hx}), \quad w_h \in W_h, \end{aligned} \tag{2.5}$$

with $u_h(0)$ and $\sigma_h(0)$ specified later. Since V_h and W_h are finite dimensional subspaces, the problem (2.5) leads to a linear system of integro-differential equations combined with algebraic equations of index one, as the stiffness matrix associated with $(\alpha u_{hx}, v_{hx})$ is invertible. Using Picard's iteration, it is easy to prove the existence of a unique pair of solutions to the system (2.5).

For use in the error analysis, we introduce the projections $\{\tilde{u}_h, \tilde{\sigma}_h\}$ defined by

$$\begin{aligned} &(a(u_x - \tilde{u}_{hx}), v_{hx}) = 0, \quad v_h \in V_h, \\ &A(\sigma - \tilde{\sigma}_h, w_h) = 0, \quad w_h \in W_h, \end{aligned} \tag{2.6}$$

where $A(\phi, \chi) = (\phi_x, \chi_x) + (\phi, \chi)$.

With $\eta = u - \tilde{u}_h$ and $\rho = \sigma - \tilde{\sigma}_h$, the following estimates are well known [20]: for $j = 0, 1$,

$$\|\eta\|_j \le Ch^{k+1-j}\|u\|_{k+1} \tag{2.7}$$

and

$$\|\rho\|_j + \|\rho_t\|_j \le Ch^{r+1-j}\left(\|\sigma\|_{r+1} + \|\sigma_t\|_{r+1}\right). \tag{2.8}$$

Moreover, for $j = 0, 1$, and $1 \le p \le \infty$, we have

$$\|\eta\|_{W^{j,p}} \le Ch^{k+1-j}\|u\|_{W^{k+1,p}}. \tag{2.9}$$

For the maximum norm estimate (i.e., when $p = \infty$), the finite element mesh is required to be quasi-uniform.

Using the projections $\{\tilde{u}_h, \tilde{\sigma}_h\}$, we write $u - u_h = (u - \tilde{u}_h) + (\tilde{u}_h - u_h) := \eta + \zeta$ and $\sigma - \sigma_h = (\sigma - \tilde{\sigma}_h) + (\tilde{\sigma}_h - \sigma_h) := \rho + \xi$. From (2.3), (2.5), and (2.6), we then obtain

$$(a\zeta_x, v_{hx}) = (\rho, v_{hx}) + (\xi, v_{hx}), \quad v_h \in V_h, \tag{2.10}$$

and

$$\begin{aligned} &(\alpha\xi_t, w_h) + \int_0^t \beta(t-s)\,(\xi_x(s), w_{hx})\,ds \\ &\qquad = -(\alpha\rho_t, w_h) + \int_0^t \beta(t-s)\,(\rho, w_h), \quad w_h \in W_h. \end{aligned} \tag{2.11}$$

THEOREM 2.1. *Assume that* $\sigma_0 = au_{0x}$ *and* $\sigma_h(0) = \tilde{\sigma}_h(0)$. *Then*

$$\|(\sigma - \sigma_h)(t)\| \le Ch^{r+1}\left[\|\sigma_0\|_{r+1} + \|\sigma_t\|_{L^1(H^{r+1})}\right], \quad t \in (0, T].$$

Moreover, for $1 < p \le \infty$,

$$\|(u - u_h)(t)\|_{L^p} \le Ch^{\min(k+1,r+1)}\left[\|\sigma_0\|_{r+1} + \|u\|_{L^\infty(W^{k+1,p})} + \|\sigma_t\|_{L^1(H^{r+1})}\right]$$

and

$$\|(u - u_h)(t)\|_1 \le Ch^{\min(k,r+1)}\left[\|\sigma_0\|_{r+1} + \|u\|_{L^\infty(H^{k+1})} + \|\sigma_t\|_{L^1(H^{r+1})}\right].$$

Proof. Since estimates of η and ρ are given by (2.7) and (2.8), respectively, it is sufficient to estimate ζ and ξ. To this end, set $v_h = \zeta$ in (2.10) and use $a \ge a_0 > 0$ to obtain

$$\|\zeta_x\| \le C(\|\rho\| + \|\xi\|). \tag{2.12}$$

Further, choose $w_h = \xi(t)$ in (2.11) and apply the Cauchy–Schwarz inequality with boundedness property of α to obtain

$$\begin{aligned} &\frac{d}{dt}\|\alpha^{\frac{1}{2}}\xi\|^2 + 2\int_0^t \beta(t-s)\,(\xi_x(s), \xi_x(t))\,ds \\ &\qquad \le C\left[\|\rho_t\| + \int_0^t \beta(t-s)\|\rho(s)\|\,ds\right]\|\xi(t)\|. \end{aligned} \tag{2.13}$$

On integrating (2.13) with respect to time, and using (1.2), the positive definiteness of β, and the fact that from (2.4) α is bounded below, we obtain

$$\|\xi(t)\|^2 \le C\left[\|\xi(0)\| + \int_0^t \|\rho_t(s)\|\,ds + \int_0^t\int_0^s \beta(s-\tau)\|\rho(\tau)\|\,d\tau\,ds\right]\max_{0\le s\le t}\|\xi(s)\|.$$

Since $\sigma_h(0) = \tilde{\sigma}_h(0)$, it follows that $\xi(0) = 0$. Then, taking the maximum of both sides over $[0, t]$, we have

$$\|\xi(t)\| \le \max_{0\le s\le t}\|\xi(s)\| \le C\left[\|\xi(0)\| + \int_0^t \|\rho_t(s)\|\,ds + \int_0^t B(t-s)\|\rho(s)\|\,ds\right].$$

Here

$$B(t) = \int_0^t \beta(s)\,ds, \tag{2.14}$$

and we have used

$$\int_0^t\int_0^s \beta(s-\tau)\|\rho(\tau)\|\,d\tau\,ds = \int_0^t B(t-s)\|\rho(s)\|\,ds.$$

From (2.8), the first estimate follows. Using the Sobolev embedding theorem and Poincaré inequality, it follows that $\|\zeta(t)\|_{L^p} \le C\|\zeta_x(t)\|$, $\zeta(t) \in H_0^1$. Finally, using (2.7)–(2.9), (2.12), and (1.4) appropriately, we apply the triangle inequality to complete the proof. ∎

Remark 2.1. (i) From the proof of Theorem 2.1, it is clear that we can choose $\sigma_h(0)$ as the L^2 projection of σ_0 into W_h instead of the elliptic projection $\tilde{\sigma}_h(0)$. Note

that the results presented in Theorem 2.1 are optimal with respect to the approximation property but not with respect to the regularity of the solution.

(ii) The estimates in Theorem 2.1 are also valid for a weakly singular kernel $\beta(t)$, i.e.,

$$\beta(t) = \frac{t^{\mu-1}}{\Gamma(\mu)}, \quad 0 < \mu < 1.$$

When $k = r$, the regularity results needed for optimal $L^\infty(L^2)$ estimates of u and $\boldsymbol{\sigma}$ are $u_0 \in H^{r+2}$, $u \in L^\infty(H^{r+1}) \cap W^{1,1}(H^{r+2})$, and these results can be easily derived under some compatibility conditions following the analysis of [9] (see Lemmas 5.1–5.6 of [9]).

3. Error estimates for problems in two and three space variables. Let $W = \left\{\mathbf{q} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{q} \in L^2(\Omega)\right\}$ with norm

$$\|\mathbf{q}\|_{\mathbf{H}(\mathrm{div},\Omega)} = (\|\mathbf{q}\|^2 + \|\nabla \cdot \mathbf{q}\|^2)^{\frac{1}{2}}.$$

Then the weak formulation of (1.1) for $d = 2, 3$ is the following: find $\{u(t), \boldsymbol{\sigma}(t)\} \in H_0^1 \times \mathbf{W}$ satisfying

$$(a\nabla u, \nabla v) = (\boldsymbol{\sigma}, \nabla v), \quad v \in H_0^1,$$

$$(\alpha\boldsymbol{\sigma}_t, \mathbf{w}) + \int_0^t \beta(t-s)\,(\nabla\cdot\boldsymbol{\sigma}(s), \nabla\cdot\mathbf{w})\,ds = -(f, \nabla\cdot\mathbf{w}), \quad \mathbf{w} \in \mathbf{W}. \tag{3.1}$$

In the analysis of this problem, we employ the classical Hilbert Sobolev spaces $H^m(\Omega)$, denoted by H^m, with norm $\|\cdot\|_m$. Let $(H^m)^d = \mathbf{H}^m$ denote the corresponding product space with the usual product norm.

To define the semidiscrete H^1-Galerkin mixed finite element procedure, let $\mathcal{T}_h$ be a partition of Ω into a finite number of elements called simplices; i.e., $\Omega = \cup_{K\in\mathcal{T}_h} K$ with $h = \max\{\mathrm{diam}(K) : K \in \mathcal{T}_h\}$. Let V_h and $\mathbf{W}_h$ be finite dimensional subspaces of H_0^1 and $\mathbf{W}$, respectively, satisfying the following approximation properties: for nonnegative integers k and r,

$$\inf_{v_h\in V_h} \{\|v - v_h\| + h\|v - v_h\|_1\} \le Ch^{k+1}\|v\|_{k+1},\ v \in H^{k+1} \cap H_0^1 \tag{3.2}$$

and

$$\inf_{\mathbf{q}_h\in\mathbf{W}_h} \{\|\mathbf{q} - \mathbf{q}_h\| + h\|\mathbf{q} - \mathbf{q}_h\|_{\mathbf{H}(\mathrm{div};\Omega)}\} \le Ch^{r+1}\|\mathbf{q}\|_{r+1},\ \mathbf{q} \in \mathbf{H}^{r+1}. \tag{3.3}$$

Standard examples of such spaces are

$$V_h = \left\{v_h \in \mathcal{C}^0(\Omega) : v_h|_K \in P_k(K)\ \forall K \in \mathcal{T}_h, v_h = 0\ \text{ on }\ \partial\Omega\right\} \tag{3.4}$$

and

$$\mathbf{W}_h = \{\mathbf{q}_h \in \mathbf{W} : (\mathbf{q}_h)_i|_K \in P_r(K), i = 1, 2, \ldots, d\ \forall K \in \mathcal{T}_h\},$$

where $P_s(K)$ is the space of polynomials of degree $\le s$ on K. Other examples of approximating spaces can be found in [2] and [17]. Note that we also allow the use of isoparametric elements.

The semidiscrete H^1-Galerkin mixed finite element approximation $\{u_h, \boldsymbol{\sigma}_h\} : [0,T] \mapsto V_h \times \mathbf{W}_h$ for (3.1) is determined by

$$(a\nabla u_h, \nabla v_h) = (\boldsymbol{\sigma}_h, \nabla v_h), \quad v_h \in V_h, \tag{3.5}$$

and

$$(\alpha\boldsymbol{\sigma}_{ht}, \mathbf{w}_h) + \int_0^t \beta(t-s)\,(\nabla\cdot\boldsymbol{\sigma}_h(s), \nabla\cdot\mathbf{w}_h)\,ds = -(f, \nabla\cdot\mathbf{w}_h), \quad \mathbf{w}_h \in \mathbf{W}_h, \tag{3.6}$$

with $u_h(0)$ and $\boldsymbol{\sigma}_h(0)$ specified later.

Corresponding to (2.6), we define the projections $\tilde{u}_h \in V_h$ and $\tilde{\boldsymbol{\sigma}}_h \in \mathbf{W}_h$ by

$$(\nabla(u - \tilde{u}_h), \nabla v_h) = 0, \quad v_h \in V_h, \tag{3.7}$$

and

$$A(\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h, \mathbf{w}_h) = 0, \quad \mathbf{w}_h \in \mathbf{W}_h, \tag{3.8}$$

where $A(\mathbf{w}, \mathbf{w}_h) = (\nabla\cdot\mathbf{w}, \nabla\cdot\mathbf{w}_h) + (\mathbf{w}, \mathbf{w}_h)$.

With $\boldsymbol{\rho} = \boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h$ and $\eta = u - \tilde{u}_h$, the following estimates are easy to obtain (cf., [21]):

$$\|\eta(t)\| + h\|\nabla\eta(t)\| \le Ch^{k+1}\|u(t)\|_{k+1} \tag{3.9}$$

and

$$\|\boldsymbol{\rho}(t)\|_{\mathbf{H}(\mathrm{div},\Omega)} + \|\boldsymbol{\rho}_t(t)\|_{\mathbf{H}(\mathrm{div},\Omega)} \le Ch^r\left(\|\boldsymbol{\sigma}\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{r+1}\right). \tag{3.10}$$

To determine the desired error estimates, we write $u - u_h := (u - \tilde{u}_h) + (\tilde{u}_h - u_h) = \eta + \zeta$ and $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h := (\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h) + (\tilde{\boldsymbol{\sigma}}_h - \boldsymbol{\sigma}_h) = \boldsymbol{\rho} + \boldsymbol{\xi}$. From (3.1), (3.5)–(3.8), we obtain

$$(a\nabla\zeta, \nabla v_h) = ((\boldsymbol{\rho} + \boldsymbol{\xi}), \nabla v_h), \quad v_h \in V_h, \tag{3.11}$$

and

$$\begin{aligned}(\alpha\boldsymbol{\xi}_t, \mathbf{w}_h) + \int_0^t \beta(t-s)\,(\nabla\cdot\boldsymbol{\xi}(s), \nabla\cdot\mathbf{w}_h)\,ds \\ = -(\alpha\boldsymbol{\rho}_t, \mathbf{w}_h) + \int_0^t \beta(t-s)\,(\boldsymbol{\rho}(s), \mathbf{w}_h)\,ds, \quad \mathbf{w}_h \in \mathbf{W}_h.\end{aligned} \tag{3.12}$$

THEOREM 3.1. *With $\boldsymbol{\sigma}_0 = a\nabla u_0$, assume that*

$$\|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_{0h}\|_{\mathbf{H}(\mathrm{div},\Omega)} \le Ch^r\|\boldsymbol{\sigma}_0\|_{r+1}.$$

Then

$$\|(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)(t)\| \le Ch^r\left[\|\boldsymbol{\sigma}_0\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right] \tag{3.13}$$

and

$$\begin{aligned}\|(u - u_h)(t)\| + h\|(u - u_h)\|_1 \\ \le Ch^{\min(k+1,r)}\left[\|\boldsymbol{\sigma}_0\|_{r+1} + \|u\|_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right].\end{aligned}$$

Proof. Choose $v_h = \zeta$ in (3.11) to obtain

$$\|\nabla\zeta\| \le C(\|\boldsymbol{\rho}\| + \|\xi\|). \tag{3.14}$$

Next, set $\mathbf{w}_h = \boldsymbol{\xi}(t)$ in (3.12) and use the Cauchy–Schwarz inequality to obtain

$$\frac{1}{2}\frac{d}{dt}\|\alpha^{\frac{1}{2}}\boldsymbol{\xi}\|^2 + \int_0^t \beta(t-s)\,(\nabla\cdot\boldsymbol{\xi}(s), \nabla\cdot\boldsymbol{\xi}(t))\,ds \le \left[\|\boldsymbol{\rho}_t\| + \int_0^t \beta(t-s)\|\boldsymbol{\rho}(s)\|\,ds\right]\|\boldsymbol{\xi}(t)\|.$$

On integrating from 0 to t and using the positivity property (1.2) of β, the second term on the left-hand side of the resulting equation is nonnegative. As in the one dimensional case, on taking the maximum over $[0,t]$, we have

$$\|\boldsymbol{\xi}(t)\| \le C\left[\|\boldsymbol{\xi}(0)\| + \int_0^t \|\boldsymbol{\rho}_t(s)\|\,ds + \int_0^t B(t-s)\|\boldsymbol{\rho}(s)\|\,ds\right].$$

For the L^2-norm estimate of ζ, we use the Poincaré inequality, $\|\zeta\| \le C\|\nabla\zeta\|$. Since the $\mathbf{L}^2$ norm is dominated by the $\mathbf{H}(\mathrm{div},\Omega)$ norm, the use of the triangle inequality with (3.9)–(3.10) and (3.14) completes the proof. ∎

Remark 3.1. (i) In Theorem 3.1, the estimate (3.13) does not depend on the approximation properties of V_h, and hence the degree k of V_h does not influence the estimate of $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$.

(ii) When $r = k+1$, we obtain the optimal order of convergence for $u - u_h$ in the L^2 norm, that is, optimality with respect to the approximation property, but not with respect to the regularity of the solution.

(iii) Estimate (3.13) indicates that the error estimate $\|(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)(t)\|$ is not optimal in the L^2 norm. This is primarily due to the fact that the bounds in the L^2 norm for $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_t$ are not of optimal order. Note that, for optimal L^2 estimates, the use of the Aubin–Nitsche trick requires an $\mathbf{H}^2$ regularity result for the adjoint operator

$$-\nabla(\nabla\cdot\boldsymbol{\Phi}) + \boldsymbol{\Phi}$$

associated with the bilinear form $A(\cdot,\cdot)$. In general, this is difficult to obtain as the operator is not coercive in the $\mathbf{H}^1$ norm. Since $-\Delta\boldsymbol{\Phi} = -\nabla(\nabla\cdot\boldsymbol{\Phi}) + \nabla\times\nabla\times\boldsymbol{\Phi}$, we therefore add a curl term to modify the formulation, and that is the theme of the next section.

4. Modified H^1-Galerkin mixed finite element method. Compared to the estimates obtained for the one dimensional case in section 2, the L^2 estimates derived in section 3 for $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$ and $u - u_h$ are not optimal. Therefore, in this section, we propose a modification of the H^1-Galerkin mixed finite element method (3.5)–(3.6) so that optimal error estimates in the L^2 norm can be determined. In order to accomplish this, we use the fact that $\nabla\times\nabla v = \mathbf{0}$, and hence we add $\nabla\times(\alpha\boldsymbol{\sigma}) = \mathbf{0}$ to the first order system (1.3). More precisely, with $\alpha\boldsymbol{\sigma} = \nabla u$, we write (1.1) as

$$\begin{aligned} u_t - \int_0^t \beta(t-s)\nabla\cdot\boldsymbol{\sigma}(s)\,ds &= f(x,t), \quad x\in\Omega, \quad t\in(0,T], \\ \nabla\times(\alpha\boldsymbol{\sigma}) &= \mathbf{0}, \quad (x,t)\in\Omega\times(0,T], \\ \mathbf{n}\wedge\alpha\boldsymbol{\sigma} &= 0, \quad (x,t)\in\partial\Omega\times(0,T], \\ u(x,0) &= u_0(x), \quad x\in\Omega, \end{aligned}$$

where $\mathbf{n}$ is the outward normal and $\wedge$ denotes the exterior product. If

$$\mathbf{W} = \{\mathbf{w} \in (H^1(\Omega))^d : \mathbf{n} \wedge \alpha\mathbf{w} = 0 \text{ on } \partial\Omega\}, \quad d = 2, 3,$$

then the weak formulation of (1.1) is the following: find $\{u, \boldsymbol{\sigma}\} : [0, T] \mapsto H_0^1 \times \mathbf{W}$ such that

$$(\nabla u, \nabla v) = (\alpha\boldsymbol{\sigma}, \nabla v), \quad v \in H_0^1,$$

$$(\alpha\boldsymbol{\sigma}_t, \mathbf{w}) + \int_0^t \beta(t-s)\mathcal{A}(\boldsymbol{\sigma}(s), \mathbf{w})\, ds = (-f, \nabla \cdot \mathbf{w}), \ \mathbf{w} \in \mathbf{H}^1, \tag{4.1}$$

where

$$\mathcal{A}(\phi, \mathbf{w}) = (\nabla \cdot \phi, \nabla \cdot \mathbf{w}) + (\nabla \times \alpha\phi, \nabla \times \alpha\mathbf{w}).$$

For the modified H^1-Galerkin mixed finite element method, we take V_h as in (3.4) and define

$$\begin{aligned}\mathbf{W}_h = \{&\mathbf{w}_h \in C(\bar{\Omega})^d : (\mathbf{w}_h)_i \,|_K \in P_r(K),\ i = 1, \ldots, d\ \forall K \in \mathcal{T}_h,\\ &(\mathbf{n} \wedge \alpha\mathbf{w}_h) = 0 \text{ at the nodes on } \partial\Omega\}.\end{aligned}$$

Since $\mathbf{n} \wedge \alpha\mathbf{w}_h = 0$ only at the boundary nodes, the finite element space $\mathbf{W}_h$ is not a subspace of $\mathbf{W}$, and hence we have a mildly nonconforming method. Note that the finite dimensional spaces V_h and $\mathbf{W}_h$ have the approximation properties (3.2) and (3.3), respectively. The modified H^1-Galerkin mixed finite element method consists of determining the pair $\{u_h, \boldsymbol{\sigma}_h\} : [0, T] \mapsto V_h \times \mathbf{W}_h$ such that

$$(\nabla u_h, \nabla v_h) = (\alpha\boldsymbol{\sigma}_h, \nabla v_h), \quad v \in V_h,$$

$$(\alpha\boldsymbol{\sigma}_{ht}, \mathbf{w}_h) + \int_0^t \beta(t-s)\mathcal{A}(\boldsymbol{\sigma}_h, \mathbf{w}_h)\, ds = (-f, \nabla \cdot \mathbf{w}_h), \quad \mathbf{w}_h \in \mathbf{W}_h, \tag{4.2}$$

with $u_h(0)$ and $\boldsymbol{\sigma}_h(0)$ specified later.

For the error analysis, we define the projections $\{\tilde{u}_h, \tilde{\boldsymbol{\sigma}}_h\} : [0, T] \mapsto V_h \times \mathbf{W}_h$ by

$$\begin{aligned}&(\nabla(u - \tilde{u}_h), \nabla v_h) = 0, \quad v_h \in V_h,\\ &A_1(\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h, \mathbf{w}_h) = 0, \quad \mathbf{w}_h \in \mathbf{W}_h,\end{aligned} \tag{4.3}$$

where

$$A_1(\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h, \mathbf{w}_h) = \mathcal{A}(\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h, \mathbf{w}_h) + (\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h, \mathbf{w}_h).$$

When the domain Ω is convex or the boundary $\partial\Omega$ is of class $C^{1,1}$ or Ω is a curvilinear polygon (or polytope) of class $C^{1,1}$ with no concave angles, then there is a positive constant μ_0, independent of h, such that

$$\|\mathbf{q}_h\|^2_{\mathbf{H}(\mathrm{div},\Omega)} + \|\nabla \times (\alpha\mathbf{q}_h)\|^2 \geq \mu_0 \|\mathbf{q}_h\|^2_{\mathbf{H}^1(\Omega)} \tag{4.4}$$

for all $\mathbf{q}_h \in \mathbf{W}_h$ and for small h; see pp. 509–510 of [15]. Thus, $A_1(\cdot, \cdot)$ satisfies the coercivity condition

$$A_1(\boldsymbol{\phi}_h, \boldsymbol{\phi}_h) \geq \mu_0 \|\boldsymbol{\phi}_h\|_1^2, \quad \boldsymbol{\phi}_h \in \mathbf{W}_h.$$

Let $u-\tilde{u}_h=\eta$ and $\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}_h=\boldsymbol{\rho}$. With an appropriate modification of the analysis of [15] (see also [11, 12, 13]), it is easy to obtain the following estimates for $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_t$:

$$\|\boldsymbol{\rho}\|_j+\|\boldsymbol{\rho}_t\|_j\le Ch^{r+1-j}[\|\boldsymbol{\sigma}\|_{r+1}+\|\boldsymbol{\sigma}_t\|_{r+1}],\quad j=0,1. \tag{4.5}$$

Note that the difficulties due to nonconformity enter the error analysis of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_t$. As before, we write $u-u_h:=(u-\tilde{u}_h)+(\tilde{u}_h-u_h)=\eta+\zeta$ and $\boldsymbol{\sigma}-\boldsymbol{\sigma}_h:=(\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}_h)+(\tilde{\boldsymbol{\sigma}}_h-\boldsymbol{\sigma}_h)=\boldsymbol{\rho}+\boldsymbol{\xi}$. From (4.1)–(4.3), we have

$$(\nabla\zeta,\nabla v_h)=(\alpha(\boldsymbol{\rho}+\boldsymbol{\xi}),\nabla v_h),\quad v_h\in V_h, \tag{4.6}$$

and

$$\begin{aligned}&(\alpha\boldsymbol{\xi}_t,\mathbf{w}_h)+\int_0^t\beta(t-s)\mathcal{A}(\boldsymbol{\xi},\mathbf{w}_h)\,ds\\&\quad=-(\alpha\boldsymbol{\rho}_t,\mathbf{w}_h)+\int_0^t\beta(t-s)(\boldsymbol{\rho}(s),\mathbf{w}_h)\,ds,\ \mathbf{w}_h\in\mathbf{W}_h.\end{aligned} \tag{4.7}$$

We now prove the main theorem in this section.

THEOREM 4.1. *Assume that* $\boldsymbol{\sigma}_h(0)=\tilde{\boldsymbol{\sigma}}_h(0)$ *with* $\boldsymbol{\sigma}_0=a\nabla u_0$. *Then*

$$\|(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)(t)\|\le Ch^{r+1}\left[\|\boldsymbol{\sigma}_0\|_{r+1}+\|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right] \tag{4.8}$$

and

$$\begin{aligned}&\|(u-u_h)(t)\|+h\|(u-u_h)(t)\|_1\\&\quad\le Ch^{\min(k+1,r+1)}\left[\|u\|_{L^\infty(H^{k+1})}+\|\boldsymbol{\sigma}_0\|_{r+1}+\|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right].\end{aligned}$$

Proof. Choose $v_h=\zeta(t)$ in (4.6) to obtain

$$\|\nabla\zeta\|\le C\left(\|\boldsymbol{\rho}\|+\|\boldsymbol{\xi}\|\right). \tag{4.9}$$

Further, setting $\mathbf{w}_h=\boldsymbol{\xi}(t)$ in (4.7) gives

$$\frac{d}{dt}\|\alpha^{\frac{1}{2}}\boldsymbol{\xi}\|^2+2\int_0^t\beta(t-s)\mathcal{A}(\boldsymbol{\xi}(s),\boldsymbol{\xi}(t))\,ds\le\left(\|\boldsymbol{\rho}_t\|+\int_0^t\beta(t-s)\|\boldsymbol{\rho}(s)\|\,ds\right)\|\boldsymbol{\xi}\|.$$

Integrate from 0 to t and use (1.2) to obtain

$$\|\boldsymbol{\xi}(t)\|^2\le C[\|\boldsymbol{\xi}(0)\|+\int_0^t\|\boldsymbol{\rho}_t(s)\|\,ds+\int_0^t B(t-s)\|\boldsymbol{\rho}(s)\|\,ds]\max_{0\le s\le t}\|\boldsymbol{\xi}(s)\|,$$

where B is given by (2.14). Taking the maximum over $[0,t]$, it is easy to see that

$$\|\boldsymbol{\xi}(t)\|\le\max_{0\le s\le t}\|\boldsymbol{\xi}(s)\|\le C[\|\boldsymbol{\xi}(0)\|+\int_0^t\|\boldsymbol{\rho}_t(s)\|\,ds+\int_0^t B(t-s)\|\boldsymbol{\rho}(s)\|\,ds].$$

Since $\boldsymbol{\sigma}(0)=\tilde{\boldsymbol{\sigma}}_h(0)$, then $\boldsymbol{\xi}(0)=\mathbf{0}$. Using the triangle inequality and (4.5), we obtain (4.8). On substituting $\|\boldsymbol{\xi}(t)\|$ in (4.9), it follows that

$$\|\nabla\zeta(t)\|\le Ch^{r+1}\left(\|\boldsymbol{\sigma}\|_{L^\infty(\mathbf{H}^{r+1})}+\|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right). \tag{4.10}$$

Using (3.9), (1.4) with ϕ replaced by $\boldsymbol{\sigma}$ and the triangle inequality, we complete the proof. ∎

Remark 4.1. (i) With $r = k$, we have $\|u - u_h\|_{L^\infty(L^2)} = O(h^{k+1})$.

(ii) The present analysis yields better results than those of section 3 for the L^2 estimate of $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$ but with additional regularity assumptions on the exact solution. Note that, from Theorem 4.1, $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| = O(h^{r+1})$ even if $k < r$, and hence the degree of V_h does not influence the estimate of $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$.

COROLLARY 4.2. *Assume that $d = 2$, that is, $\Omega \subset R^2$, and $\boldsymbol{\sigma}_h(0) = \tilde{\boldsymbol{\sigma}}_h(0)$. Then*

$$\|(u - u_h)(t)\|_{L^\infty} \leq C \left(\log \frac{1}{h}\right) h^{\min(k+1,r+1)} [\|u\|_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_0\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}],$$

provided the finite element mesh is quasi-uniform.

Proof. Using the Sobolev imbedding theorem for $d = 2$, the inverse hypothesis, and the superconvergence estimate (4.10) for $\|\nabla\zeta\|$, we obtain

$$\|\zeta(t)\|_{L^\infty} \leq Ch^{\min(k+1,r+1)} \left(\log \frac{1}{h}\right)^{\frac{1}{2}} [\|u\|_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_0\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}].$$

From [2, 20], it follows that the error η in the elliptic projection satisfies

$$\|\eta(t)\|_{L^\infty} \leq \begin{cases} C \left(\log \dfrac{1}{h}\right) h^2 \|u\|_{L^\infty(W^{k+1,\infty})}, & k = 1, \\ Ch^{k+1} \|u\|_{L^\infty(W^{k+1,\infty})}, & k > 1. \end{cases}$$

The use of the triangle inequality completes the proof. ∎

5. The backward Euler method for the modified method. In this section, we briefly describe the backward Euler method for approximating $\{u, \boldsymbol{\sigma}\}$ of (4.1) and discuss the related error estimates. Since the error analysis for higher order time stepping schemes such as the Crank–Nicolson and second order backward difference methods is similar, we shall not pursue these methods further in this paper (see [9] for related results).

To describe the backward Euler method, let $\Delta t = T/M$, for some positive integer M, and set $t^n = n\Delta t$, $n = 0, \ldots, M$. For a smooth function ϕ on [0,T], define

$$\phi^n = \phi(t_n), \quad \partial_t \phi^n = \frac{\phi^n - \phi^{n-1}}{\Delta t}.$$

To approximate the integral, we introduce the right rectangle quadrature rule

$$q^n(\phi) = \Delta t \sum_{j=1}^{n} \beta_{n-j} \phi^j \approx \int_0^{t_n} \beta(t_n - s)\phi(s)\, ds,$$

where $\beta_{n-j} = \beta(t_n - t_j)$. This quadrature rule is positive [3, 9] in the sense that

$$\sum_{n=1}^{J} q^n(\phi)\phi^n = \Delta t \sum_{n=1}^{J} \sum_{j=1}^{n} \beta_{n-j} \phi^j \phi^n \geq 0, \ J = 1, \ldots, M.$$

The quadrature error

$$\epsilon^n(\phi) := q^n(\phi) - \int_0^{t_n} \beta(t_n - s)\phi(s)\, ds$$

satisfies

$$|\epsilon^n(\phi)| \le C\Delta t \int_0^{t_0} \left(|\phi(s)| + |\phi_t(s)|\right)\, ds,$$

provided $\beta,\ \phi \in C^1[0,T]$.

Let U^n and $\mathbf{Z}^n$ be approximations to u and $\boldsymbol{\sigma}$ at $t = t_n$, respectively, defined as follows. Given $\{U^{n-1}, \mathbf{Z}^{n-1}\} \in V_h \times \mathbf{W}_h$, determine $\{U^n, \mathbf{Z}^n\}$ in $V_h \times \mathbf{W}_h$ satisfying

$$\begin{aligned}(\nabla U^n, \nabla v_h) &= (\alpha \mathbf{Z}^n, \nabla v_h)\,, \quad v \in V_h,\\ (\alpha \partial_t \mathbf{Z}^n, \mathbf{w}_h) + q^n_{\mathcal{A}}(\mathbf{Z})(\mathbf{w}_h) &= (-f^n, \nabla\cdot \mathbf{w}_h)\,, \quad \mathbf{w}_h \in \mathbf{W}_h,\end{aligned}$$

with $U^0 = u_{0,h}$ specified later. Here,

$$q^n_{\mathcal{A}}(\mathbf{Z})(\mathbf{w}_h) = \Delta t \sum_{j=1}^{n} \beta_{n-j} \mathcal{A}\left(\mathbf{Z}^j, \mathbf{w}_h\right).$$

To determine the desired error estimates, we write $u^n - U^n := (u^n - \tilde{u}^n_h) + (\tilde{u}^n_h - U^n) = \eta^n + \zeta^n$ and $\boldsymbol{\sigma}^n - \mathbf{Z}^n := (\boldsymbol{\sigma}^n - \tilde{\boldsymbol{\sigma}}^n_h) + (\tilde{\boldsymbol{\sigma}}^n_h - \mathbf{Z}^n) = \boldsymbol{\rho}^n + \boldsymbol{\xi}^n$. Since estimates of η^n and $\boldsymbol{\rho}^n$ are given by (3.9) and (4.5) at $t = t_n$, it is sufficient to estimate ζ^n and $\boldsymbol{\xi}^n$. Note that the equations for ζ^n and $\boldsymbol{\xi}^n$ may be written as

$$(\nabla \zeta^n, \nabla v_h) = (\alpha(\boldsymbol{\rho}^n + \boldsymbol{\xi}^n), \nabla v_h)\,, \quad v_h \in V_h, \tag{5.1}$$

and

$$\begin{aligned}(\alpha \partial_t \boldsymbol{\xi}^n, \mathbf{w}_h) + q^n_{\mathcal{A}}(\boldsymbol{\xi})(\mathbf{w}_h) = &-(\alpha \partial_t \boldsymbol{\rho}^n + \alpha \boldsymbol{\tau}^n, \mathbf{w}_h)\\ &+\epsilon^n_{\mathcal{A}}(\boldsymbol{\sigma})(\mathbf{w}_h) + (q^n(\boldsymbol{\rho}), \mathbf{w}_h)\,,\ \mathbf{w}_h \in \mathbf{W}_h,\end{aligned} \tag{5.2}$$

where $\boldsymbol{\tau}^n = \boldsymbol{\sigma}_t(t_n) - \partial_t \boldsymbol{\sigma}(t_n)$ and

$$\epsilon^n_{\mathcal{A}}(\boldsymbol{\sigma})(\mathbf{w}_h) = q^n_{\mathcal{A}}(\boldsymbol{\sigma})(\mathbf{w}_h) - \int_0^{t_n} \beta(t_n - s)\mathcal{A}\left(\boldsymbol{\sigma}, \mathbf{w}_h\right) ds.$$

THEOREM 5.1. *Assume that* $\mathbf{Z}^0 = \tilde{\boldsymbol{\sigma}}_h(0)$ *with* $\boldsymbol{\sigma}_0 = a\nabla u_0$ *and* $\beta \in C^1[0,T]$. *Then*

$$\begin{aligned}\|\boldsymbol{\sigma}^J - \mathbf{Z}^J\| \le C \Big\{ & h^{r+1}\left[\|\boldsymbol{\sigma}_0\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right]\\ &+\Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2\right)\, ds\Big\}.\end{aligned}$$

Further,

$$\begin{aligned}\|u^J - U^J\| + h\|u^J - U^J\|_1 &\\ \le C \Big\{ & h^{\min(k+1,r+1)}\left[\|u\|_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_0\|_{r+1} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^{r+1})}\right]\\ &+ \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2\right)\, ds\Big\}.\end{aligned}$$

Proof. Choose $v_h = \zeta^n$ in (5.1) to obtain, for $n = 0, 1, \ldots, M$,

$$\|\nabla \zeta^n\| \le C(\|\boldsymbol{\rho}^n\| + \|\boldsymbol{\xi}^n\|). \tag{5.3}$$

Set $\mathbf{w}_h = \boldsymbol{\xi}^n$ in (5.2) and use the Cauchy–Schwarz inequality and Young's inequality to obtain

$$\frac{1}{2}\partial_t \|\alpha^{\frac{1}{2}} \boldsymbol{\xi}^n\|^2 + q_A^n(\boldsymbol{\xi})(\boldsymbol{\xi}^n) \le C\left[\|\partial_t \boldsymbol{\rho}^n\| + \|\epsilon_A^n(\boldsymbol{\sigma})\| + \|\boldsymbol{\tau}^n\| + \|q^n(\boldsymbol{\rho})\|\right] \|\boldsymbol{\xi}^n\|. \tag{5.4}$$

Note that

$$\begin{aligned}
\Delta t \sum_{n=1}^{J} \|\partial_t \boldsymbol{\rho}^n\| &\le C h^{r+1} \int_0^{t_J} \|\boldsymbol{\sigma}_t(s)\|_{r+1}\, ds, \\
\Delta t \sum_{n=1}^{J} \|\epsilon_A^n(\boldsymbol{\sigma})\| &\le C \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2\right)\, ds, \\
\Delta t \sum_{n=1}^{J} \|\boldsymbol{\tau}^n\| &\le C \Delta t \int_0^{t_J} \|\boldsymbol{\sigma}_{tt}(s)\|\, ds.
\end{aligned}$$

On substituting these estimates in (5.4) after summing from $n = 1, \ldots, J$, the second term on the left-hand side of the resulting inequality is nonnegative. Thus, we obtain

$$\begin{aligned}
\|\boldsymbol{\xi}^J\| \le \max_{1 \le n \le J} \|\boldsymbol{\xi}^n\| \le C \Bigg\{ &\|\boldsymbol{\xi}^0\| + h^{r+1} \left(\|\boldsymbol{\sigma}\|_{L^\infty(\mathbf{H}^{r+1})} + \int_0^{t_J} \|\boldsymbol{\sigma}_t(s)\|_{r+1}\, ds \right) \\
&+ \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2\right) ds \Bigg\}.
\end{aligned} \tag{5.5}$$

Note that $\boldsymbol{\xi}^0 = 0$. Using (5.5) in (5.3), we obtain the superconvergence result

$$\begin{aligned}
\|\nabla \zeta^J\| \le C \Bigg\{ &h^{r+1} [\|\boldsymbol{\sigma}\|_{L^\infty(\mathbf{H}^{r+1})} + \int_0^{t_J} \|\boldsymbol{\sigma}_t(s)\|_{r+1}\, ds] \\
&+ \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2\right) ds] \Bigg\}.
\end{aligned}$$

The use of the triangle inequality with the estimates of $\boldsymbol{\rho}^J$ and η^J completes the proof. ∎

In the remainder of the section, we relax the regularity assumptions on the kernel, requiring only $\beta \in L^1(0, T)$. We still assume that β is positive definite in the sense of (1.2). We now consider the backward Euler method using product integration [7, 9] to approximate the integral:

$$q^n(\phi) = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \beta(t_n - s)\phi^j\, ds = \sum_{j=1}^{n} \kappa_{n-j}\phi^j \approx \int_0^{t_n} \beta(t_n - s)\phi(s)\, ds, \tag{5.6}$$

where

$$\kappa_j = \int_{t_j}^{t_{j+1}} \beta(s)\, ds.$$

When

$$\beta \in L^1_{loc}(0,\infty) \cap C^2(0,\infty) \ \text{ and } \ (-1)^j \beta^{(j)}(t) \geq 0, \ \text{ for } \ t > 0, \ j = 0, 1, 2, \tag{5.7}$$

the product rule (5.6) is positive and the quadrature error satisfies [9]

$$|\epsilon^n(\phi)| \leq C\Delta t \int_0^{t_n} |\phi_t(s)|\, ds, \quad t_n \leq T. \tag{5.8}$$

For the commonly occurring kernel $\beta(t) = t^{\mu-1}/\Gamma(\mu), \quad 0 < \mu < 1$, (5.7) is satisfied, and the proof of the following theorem may be obtained by modifying the arguments used in the proof of Theorem 5.1 and by using (5.8).

THEOREM 5.2. *Assume that $\beta(t) = t^{\mu-1}/\Gamma(\mu)$, $0 < \mu < 1$, and that the quadrature q^n is the product rule of* (5.6). *Then, for $k = 1$, $r = 1$, and $J = 0, 1, \ldots, M$,*

$$\|\boldsymbol{\sigma}^J - \mathbf{Z}^J\| \leq C \left\{ h^2 \left[\|\boldsymbol{\sigma}\|_{L^\infty(\mathbf{H}^2)} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^2)} \right] + \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2 \right) ds \right\}.$$

Further,

$$\|u^J - U^J\| + h\|u^J - U^J\|_1 \leq C \left\{ h^2 \left[\|u\|_{L^\infty(H^2)} + \|\boldsymbol{\sigma}\|_{L^\infty(\mathbf{H}^2)} + \|\boldsymbol{\sigma}_t\|_{L^1(\mathbf{H}^2)} \right] + \Delta t \int_0^{t_J} \left(\|\boldsymbol{\sigma}_{tt}(s)\| + \|\boldsymbol{\sigma}(s)\|_2 + \|\boldsymbol{\sigma}_t(s)\|_2 \right) ds \right\}.$$

6. Concluding remarks. In this paper, a priori error estimates are derived for an H^1-Galerkin mixed finite element method without the LBB consistency condition and also without a quasi-uniformity assumption on the finite element mesh. Since the estimate for $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|$ derived in section 3 is not optimal in two and three space dimensions, a modification of the method is proposed and analyzed in section 4 to establish an optimal estimate in the L^2 norm as in the one dimensional case of section 2. Another notable advantage of the present method is that it allows the use of two different finite element spaces for approximating u and its flux $\boldsymbol{\sigma}$. In particular, use of piecewise linear polynomial spaces yields $O(h^2)$ of convergence in both $u - u_h$ and $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$ in the L^2 norm. Moreover, in two dimensions, we obtain a quasi-optimal maximum norm estimate for $u - u_h$.

The results presented in this paper can be easily extended to the initial and boundary value problem

$$\begin{aligned} u_t - \int_0^t \beta(t-s) \left[\nabla \cdot (a(x)\nabla u(s)) - b(x)u(s) \right] ds &= f(x,t), \quad (x,t) \in \Omega \times (0,T], \\ u(x,t) &= 0, \quad (x,t) \in \partial\Omega \times (0,T], \\ u(x,0) &= u_0(x), \quad x \in \Omega, \end{aligned}$$

where $b = b(x) \geq 0$, $x \in \Omega$. For the mixed formulation, this equation is rewritten as

$$\begin{aligned} \boldsymbol{\sigma} &= a\nabla u, \\ u_t - \int_0^t \beta(t-s) \left(\nabla \cdot \boldsymbol{\sigma}(s) - bu(s) \right) ds &= f. \end{aligned}$$

The corresponding H^1-Galerkin mixed finite element method is based on the weak formulation

$$
\begin{aligned}
&(a\nabla u, \nabla v) = (\boldsymbol{\sigma}, \nabla v), \quad v \in H_0^1, \\
&(\alpha \boldsymbol{\sigma}_t, \mathbf{w}) + \int_0^t \beta(t-s)\left[(\nabla\cdot\boldsymbol{\sigma}(s), \nabla\cdot\mathbf{w}) + (b\boldsymbol{\sigma}(s), \mathbf{w})\right] ds \\
&\qquad = -(f, \nabla\cdot\mathbf{w}) - \int_0^t \beta(t-s)\,(u(s)\nabla b, \mathbf{w}), \quad \mathbf{w} \in \mathbf{W}.
\end{aligned} \tag{6.1}
$$

Unlike (3.1), the system (6.1) is now strongly coupled with unknowns $(u, \boldsymbol{\sigma})$. Since the error estimates closely follow the proof techniques of the present paper, we shall not pursue these further.

Acknowledgments. The first author thanks the Department of Mathematical and Computer Sciences, Colorado School of Mines for financial support and hospitality during his sabbatical leave from the Indian Institute of Technology, Bombay. The authors thank the referees for their valuable comments and suggestions.

REFERENCES

[1] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[2] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[3] G. Fairweather, *Spline collocation methods for a class of hyperbolic partial integro-differential equations*, SIAM J. Numer. Anal. 31 (1994), pp. 444–460.

[4] G. J. Fix, M. D. Gunzburger, and R. A. Nicolaides, *On the mixed finite element methods for first order elliptic systems*, Numer. Math., 37 (1981), pp. 29–48.

[5] L. P. Franca and T. J. R. Hughes, *Two classes of mixed finite element methods*, Comput. Methods Appl. Mech. Engrg., 69 (1988), pp. 89–129.

[6] J. Haslinger and P. Neittaanmäki, *On different finite element methods for approximating the gradient of the solution to the Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 42 (1984), pp. 131–148.

[7] J. C. López-Marcos, *A difference scheme for a nonlinear partial integrodifferential equation*, SIAM J. Numer. Anal., 27 (1990), pp. 20–31.

[8] R. C. MacCamy, *An integro-differential equation with application to heat flow*, Quart. Appl. Math., 35 (1977), pp. 21–33.

[9] W. McLean and V. Thomée, *Numerical solution of an evolution equation with a positive type memory term*, J. Austral. Math. Soc. Ser. B, 35 (1993), pp. 23–70.

[10] W. McLean, V. Thomée, and L. B. Wahlbin, *Discretization with variable time steps of an evolution equation with a positive type memory term*, J. Comput. Appl. Math., 69 (1996), pp. 49–69.

[11] P. Neittaanmäki and J. Saranen, *On finite element approximation of the gradient for solution of Poisson equation*, Numer. Math., 37 (1981), pp. 333–337.

[12] P. Neittaanmäki and J. Saranen, *On the finite element approximation of vector fields by curl and divergence*, Math. Methods Appl. Sci., 3 (1981), pp. 328–335.

[13] P. Neittaanmäki and J. Saranen, *A mixed finite element method for the heat flow problem*, BIT, 21 (1981), pp. 342–346.

[14] A. K. Pani, *An H^1-Galerkin mixed finite element method for parabolic partial differential equations*, SIAM J. Numer. Anal., 35 (1998), pp. 721–727.

[15] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 499–516.

[16] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[17] P. A. Raviart and J. M. Thomas, *A Mixed Finite Element Method for Second Order Elliptic Problems*, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 293–315.

[18] M. Renardy, W. Hrusa, and J. Nohel, *Mathematical Problems in Viscoelasticity*, Pitman Monogr. Surveys Pure Appl. Math. 35, Wiley, New York, 1987.
[19] J. M. Sanz-Serna, *A numerical method for a partial integro-differential equation*, SIAM J. Numer. Anal., 25 (1988), pp. 319–327.
[20] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. in Comput. Math. 25, Springer-Verlag, New York, 1997.
[21] M. F. Wheeler, *A priori L^2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–749.
[22] Y. Yan and G. Fairweather, *Orthogonal spline collocation methods for some partial integrodifferential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 755–768.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1491–1515

ERROR ANALYSIS FOR CHARACTERISTICS-BASED METHODS FOR DEGENERATE PARABOLIC PROBLEMS*

ZHANGXIN CHEN†, RICHARD E. EWING‡, QIAOYUAN JIANG†, AND ANNA M. SPAGNUOLO‡

Abstract. We consider characteristics-based finite element methods for solving nonlinear, degenerate, advection-diffusion equations. These equations have applications in the simulation of petroleum reservoirs and groundwater aquifers and in the modeling of free boundary problems. Standard finite element Galerkin methods have been studied for these equations. In this paper, we analyze the characteristics-based finite element methods for them. The main difficulty in the analysis is that the equations are degenerate and the solution lacks regularity. Here we develop a technique that respects the degeneracy and the known minimal regularity. This technique is based on the Green operator for standard elliptic equations and is developed directly for the degenerate advection-diffusion equations. We concentrate our analysis on the modified method of characteristics (MMOC) and one of its variants, the modified method of characteristics with adjusted advection (MMOCAA), which conserves mass. We derive error estimates in various norms. The extension to other variants is discussed. The present technique is also applied to nondegenerate problems; error estimates previously obtained for the MMOC are derived under much weaker regularity assumptions on the solution, and the error estimates for the MMOCAA appear new even in the nondegenerate case. Finally, numerical results are presented to show the sharpness of the error estimates derived.

Key words. degeneracy, advection-diffusion equations, characteristics-based finite elements, error estimates

AMS subject classifications. 35K60, 35K65, 76S05, 76T05

PII. S003614290037068X

1. Introduction.

Let $\Omega \subset \Re^d$, $d \leq 3$, be a bounded domain and $J = (0, T]$, with $T > 0$. We consider and analyze characteristics-based finite element methods for the advection-diffusion equation in $u(x, t)$:

$$c\partial_t u + b \cdot \nabla u - \nabla \cdot (a(u)\nabla u) = f \qquad \text{in } \Omega \times J, \tag{1.1}$$

where $c = c(x, t)$, $b = b(x, t)$ (a vector), $a(u) = \alpha(x, t)\beta(u)$, and $f = f(x, t)$, with α being a $d \times d$ symmetric, positive-definite matrix and β a nonnegative function in u. Because β can be zero in u, (1.1) is generally degenerate in this variable.

Problem (1.1) arises in many applications. It appears in petroleum reservoir simulation that often requires the numerical solution for similar problems of multiphase fluid flow in porous media [6, 38], in groundwater aquifer modeling for the transport of a solute in porous media with an equilibrium adsorption reaction [26, 29, 31], and in the solution of parabolic free boundary problems [30, 35], for example. Standard finite element Galerkin methods have been studied for these applications with the degeneracy taken into account; see [11, 14, 15, 16, 17, 28, 34, 40, 43] in the reservoir

*Received by the editors April 11, 2000; accepted for publication (in revised form) March 29, 2002; published electronically October 23, 2002. This work was supported in part by National Science Foundation grants DMS-9626179, DMS-9972147, and INT-9901498 and by a gift grant from Mobil Technology Company.

http://www.siam.org/journals/sinum/40-4/37068.html

†Department of Mathematics, Box 750156, Southern Methodist University, Dallas, TX 75275-0156 (zchen@mail.smu.edu, qjiang@mail.smu.edu).

‡Department of Mathematics and Institute for Scientific Computation, Texas A&M University, College Station, TX 77843-3404 (ewing@isc.tamu.edu, annas@math.tamu.edu).

simulation, [5, 13, 19] for the groundwater modeling, and [32, 36] for the free boundary problems, for instance.

In the petroleum and groundwater areas, finite difference methods are most often used to solve equations analogous to (1.1). It is known [42] that certain finite difference methods are actually equivalent to mixed finite element methods of the lowest order on rectangles [7], combined with special quadrature rules. The mixed finite element methods for equations similar to (1.1) have been analyzed in [3].

It is well known that advection-diffusion equations often present serious numerical difficulties. Standard finite element and finite difference methods usually exhibit some combination of nonphysical oscillation and excessive numerical dispersion [27, 33]. Many numerical methods have been developed to overcome these difficulties. In this paper, we consider and analyze the modified method of characteristics (MMOC) (or the Eulerian–Lagrangian method) [25, 39] for (1.1). Because of the Lagrangian nature of the advection term, this method treats this term by a characteristic tracking scheme. It has many advantages and one fundamental flaw, the failure to preserve as an algebraic identity a desired conservation law associated with the underlying physical problem. Recently, a variant of the MMOC, called the modified method of characteristics with adjusted advection (MMOCAA) has been introduced [22]. The MMOCAA does preserve the desired conservation property and also the conceptual and computational advantages of the MMOC. For this reason, we also carry out a formal analysis for the MMOCAA procedure. The extension of the analysis to other recently developed characteristics-based methods, such as the Eulerian–Lagrangian localized adjoint method (ELLAM) [8] and the characteristic-mixed finite element method [2], is discussed.

The main difficulty in the error analysis is that problem (1.1) is degenerate and its solution lacks regularity. Here we develop a technique that respects the degeneracy and the known minimal regularity. This technique is based on the Green operator for standard elliptic equations and is developed directly for the degenerate advection-diffusion equation. We first establish sharp error estimates in various norms for the MMOC and MMOCAA in the degenerate case. We then use the present technique to obtain optimal error estimates for nondegenerate problems. The degenerate case is analyzed for the first time for these two characteristic methods, while the error estimates for the nondegenerate case are derived under much weaker regularity assumptions on the solution than those previously used.

The rest of the paper is outlined as follows. In the next section, we review the theoretical results available for (1.1). The MMOC and MMOCAA procedures are analyzed in the third and fourth sections, respectively. Numerical results are presented in the final section.

2. Preliminaries. The usual Sobolev spaces $W^{l,\pi}(\Omega)$ with the norm $\|\cdot\|_{W^{l,\pi}(\Omega)}$ will be used, where l is a nonnegative integer and $0 \le \pi \le \infty$. When $\pi = 2$, we simply write $H^l(\Omega) = W^{l,2}(\Omega)$. When $l = 0$, we have $L^2(\Omega) = H^0(\Omega)$. The spaces $L^2(J; W^{l,\pi}(\Omega))$ and $L^\infty(J; W^{l,\pi}(\Omega))$ will also be used, with the norms $\|\cdot\|_{L^2(J;W^{l,\pi}(\Omega))}$ and $\|\cdot\|_{L^\infty(J;W^{l,\pi}(\Omega))}$, respectively. Below $(\cdot,\cdot)_Q$ denotes the $L^2(Q)$ inner product (or sometimes the duality pairing); Q is omitted if $Q = \Omega$. Finally, set $\Omega_T = \Omega \times J$.

It follows from [1] that under appropriate boundary and initial conditions and reasonable assumptions on the data problem (1.1) has at least the regularity results

$$u \in L^\infty(J; L^1(\Omega)), \quad \partial_t u \in L^2(J; H^{-1}(\Omega)), \quad \beta(u)\nabla u \in L^2(J; (L^2(\Omega))^d), \tag{2.1}$$

where $H^{-1}(\Omega)$ is the dual to $H^1(\Omega)$. In general, we can only expect the above

regularity to hold for $\partial_t u$. Under the assumption that (1.1) is physically consistent so that a maximum principle holds, u can be shown to be bounded [1]:

$$u \in L^\infty(\Omega_T). \tag{2.2}$$

With (2.2), we introduce the Kirchhoff transformation

$$\theta = \int_{u_r}^{u} \beta(\xi)\mathrm{d}\xi,$$

where u_r is any reference element in a neighborhood of the solution u. For notational convenience later, take $u_r = 0$; i.e., θ is defined by

$$\theta = \int_{0}^{u} \beta(\xi)\mathrm{d}\xi. \tag{2.3}$$

A main assumption in the later analysis is that there is a constant β^*, independent of time, such that

$$\|\theta_1 - \theta_2\|^2_{L^2(\Omega)} \le \beta^*(\theta_1 - \theta_2, u_1 - u_2), \qquad u_1,\ u_2 \in L^2(\Omega), \tag{2.4}$$

where θ_i corresponds to u_i through (2.3), $i = 1, 2$. A sufficient condition for (2.4) to hold is

$$0 \le \beta(v) \le \beta^* < \infty \quad \text{ in a neighborhood of } u. \tag{2.5}$$

This assumption will be tacitly made later and is physically reasonable [14].

Also, we rewrite the advection-diffusion equation (1.1) in the form

$$c(x)\partial_t u + b(x)\cdot\nabla u - \nabla\cdot(\alpha(x)\nabla\theta) = f(x,t) \qquad \text{in } \Omega_T, \tag{2.6}$$

with θ given in (2.3). By lagging the coefficients in time in the first and second terms of (1.1), we easily obtain (2.6). Also, in physical applications (1.1) is often coupled with other equations that determine the coefficients c and b. In the petroleum application, for example, c is the porosity and b is a velocity field. These coefficients can be calculated independent of u, so (1.1) reduces to (2.6). Finally, the present analysis can be easily extended to the case where b and c explicitly depend on t. Note that (2.6) is still degenerate and nonlinear in u since $\nabla\theta = \beta(u)\nabla u$.

Next, to avoid the difficulty associated with the boundary conditions, we assume that (2.6) is Ω-periodic; i.e., all functions in (2.6) are spatially Ω-periodic. This is physically reasonable because no-flow boundaries are usually handled by reflection and interior flow behavior is often much more important than boundary effects.

Finally, we focus on the two space dimensions; i.e., $d = 2$. An analysis can be done in the same fashion for the three dimensions [41]. Therefore, we give an analysis of error estimates in the case where Ω is a rectangular domain.

3. Analysis for the MMOC.

3.1. The MMOC procedure.

We assume that the coefficients c and b satisfy

$$0 < c_* \le c(x) \le c^* < \infty, \quad \left|\frac{b(x)}{c(x)}\right| + \left|\nabla\cdot\left(\frac{b(x)}{c(x)}\right)\right| \le C, \quad x \in \Omega. \tag{3.1}$$

We also assume that $\alpha(x)$ is a 2×2 symmetric matrix that is uniformly positive definite with respect to $x \in \bar{\Omega}$; i.e., with $\alpha = (\alpha_{ij})$,

$$0 < \alpha_* \le |\xi|^{-2} \sum_{i,j=1}^{d} \alpha_{ij}(x)\xi_i\xi_j \le \alpha^* < \infty, \qquad x \in \bar{\Omega},\ \xi \neq 0 \in \Re^2. \tag{3.2}$$

Let

$$\psi(x) = \left(c^2(x) + |b(x)|^2\right)^{1/2},$$

and let the characteristic direction associated with the operator $c\partial_t u + b \cdot \nabla u$ be denoted by $\tau(x)$, where

$$\partial_\tau = \frac{c(x)}{\psi(x)}\partial_t + \frac{b(x)}{\psi(x)} \cdot \nabla.$$

Then (2.6) can be written as follows:

$$\psi(x)\partial_\tau u - \nabla \cdot (\alpha(x)\nabla\theta) = f(x,t) \qquad \text{in } \Omega_T, \tag{3.3}$$

where u is related to θ through (2.3): $u = \mathcal{U}(\theta)$, with $\mathcal{U}$ being the inverse of (2.3). The initial condition is given by

$$u(x,0) = u_0(x) \qquad \text{in } \Omega. \tag{3.4}$$

For $0 < h$, let $M_h \subset H^1(\Omega)$ be a standard C^0-finite element space such that the approximation property holds:

$$\inf_{v_h \in M_h} \|v - v_h\|_{H^1(\Omega)} \le Ch\|v\|_{H^2(\Omega)}, \tag{3.5}$$

where and below C (with or without a subscript) indicates a generic constant independent of h, which will probably take on different values in different occurrences. In this paper, we consider only lowest-order C^0-finite elements such that (3.5) is satisfied; due to lacking regularity on the true solution, no improvement in the asymptotic convergence rate results from taking higher-order finite element spaces. We denote by P_h the L^2-projection into M_h, which satisfies that

$$\|v - P_h v\|_{H^{-1}(\Omega)} \le Ch\|v\|_{L^2(\Omega)} \qquad \forall v \in L^2(\Omega). \tag{3.6}$$

For each positive integer N, let $0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of J into subintervals $J^n = (t^{n-1}, t^n]$, with length $\Delta t^n = t^n - t^{n-1}$, $1 \le n \le N$, and let $\Delta t = \max_{1 \le n \le N} \Delta t^n$. Also, set $v^n = v(\cdot, t^n)$ and $\bar{v}^{n-1} = v(x - b(x)\Delta t^n/c(x), t^{n-1})$. The standard MMOC procedure is the determination of the map $\theta_h : \{t^1, \ldots, t^N\} \to M_h$ satisfying

$$\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, v\right) + (\alpha\nabla\theta_h^n, \nabla v) = (f^n, v) \qquad \forall v \in M_h,\ 1 \le n \le N, \tag{3.7}$$

where $u_h^n = \mathcal{U}(\theta_h^n)$, $1 \le n \le N$, and

$$u_h^0 = P_h u_0. \tag{3.8}$$

It can be shown that (3.7) determines $\{\theta_h\}$ and $\{u_h\}$ uniquely in terms of the data u_0 and f [12].

3.2. The Green operator. We introduce the bilinear form $a(\cdot,\cdot)$ on $H^1(\Omega)$,

$$a(v,w) = (\alpha\nabla v, \nabla w) + (v,w) \qquad \forall v,\ w \in H^1(\Omega),$$

and define the Green operator $G : H^{-1}(\Omega) \to H^1(\Omega)$ by

$$a(Gv,w) = (cv,w) \qquad \forall w \in H^1(\Omega),\ v \in H^{-1}(\Omega). \tag{3.9}$$

We assume that G is regular; i.e.,

$$\|Gv\|_{H^2(\Omega)} \le C\|v\|_{L^2(\Omega)} \qquad \forall v \in L^2(\Omega). \tag{3.10}$$

This assumption is satisfied with (the boundary of Ω) $\partial\Omega \in C^{1,1}$ (or, optionally, Ω being polygonal and convex). The norm in $H^{-1}(\Omega)$ can be represented in terms of G:

$$\|v\|_{H^{-1}(\Omega)} = a^{1/2}(Gv,Gv) = (cv,Gv)^{1/2} \qquad \forall v \in H^{-1}(\Omega). \tag{3.11}$$

The discrete Green operator $G_h : H^{-1}(\Omega) \to M_h$ is given by

$$a(G_h v,w) = (cv,w) \qquad \forall w \in M_h,\ v \in H^{-1}(\Omega). \tag{3.12}$$

By the regularity of G and (3.5), we have the approximation property [18]

$$\|(G-G_h)v\|_{H^l(\Omega)} \le Ch^{2-(l+\pi)}\|v\|_{B^{-\pi}(\Omega)}, \qquad 0 \le l,\pi \le 1, \tag{3.13}$$

where $B^{-\pi}(\Omega) = [L^2(\Omega), H^{-1}(\Omega)]_\pi$ is the interpolation space. Moreover, it follows from (3.12) that

$$a(G_h v, G_h v) \le C\|v\|^2_{H^{-1}(\Omega)} \qquad \forall v \in H^{-1}(\Omega). \tag{3.14}$$

3.3. Stability. In addition to assumptions (2.5), (3.1), and (3.2), we also assume that the norms

$$\|f\|_{L^\infty(J;H^{-1}(\Omega))} \text{ and } \|u_0\|_{L^2(\Omega)} \text{ are bounded.} \tag{3.15}$$

The proof of the next lemma can be found in [25] (see also Lemma 3.4 below).

LEMMA 3.1. *If $\eta \in L^2(\Omega)$ is Ω-periodic and $\bar\eta = \eta(x - g(x)\Delta t)$, where g and $\nabla\cdot g$ are bounded, then*

$$\|\eta - \bar\eta\|_{H^{-1}(\Omega)} \le C\Delta t\|\eta\|_{L^2(\Omega)}.$$

We now establish some stability results, which will be utilized in the subsequent error analysis. Note that inequality (2.5) needs to hold in the range of the numerical solution, so we extend β in some reasonable way [14].

LEMMA 3.2. *Under assumptions* (2.5), (3.1), (3.2), *and* (3.15), *we have*

$$\max_{1\le n\le N}\{\|u_h^n\|^2_{L^2(\Omega)} + \|\theta_h^n\|^2_{L^2(\Omega)}\} + \sum_{n=1}^N \|\nabla\theta_h^n\|^2_{L^2(\Omega)}\Delta t^n \le C.$$

Proof. Take $v = \theta_h^n$ in (3.7) to see that

$$\left(c\frac{u_h^n - \bar u_h^{n-1}}{\Delta t^n}, \theta_h^n\right) + (\alpha\nabla\theta_h^n, \nabla\theta_h^n) = (f^n, \theta_h^n);$$

after summing over n from 1 to N, this equation can be put in the form

$$\sum_{n=1}^{N}\left\{(c[u_h^n-u_h^{n-1}],\theta_h^n)+(\alpha\nabla\theta_h^n,\nabla\theta_h^n)\Delta t^n\right\} \tag{3.16}$$

$$=\sum_{n=1}^{N}\left\{(f^n,\theta_h^n)\Delta t^n+(c[\bar{u}_h^{n-1}-u_h^{n-1}],\theta_h^n)\right\}.$$

Define

$$\Phi(s)=\int_0^s\theta(\xi)\mathrm{d}\xi,$$

where

$$\theta(\xi)=\int_0^\xi\beta(\zeta)\mathrm{d}\zeta,$$

as defined in (2.3). Then we see that

$$(u_h^n-u_h^{n-1})\theta_h^n\geq\Phi(u_h^n)-\Phi(u_h^{n-1}),$$

so

$$\sum_{n=1}^{N}(c[u_h^n-u_h^{n-1}],\theta_h^n)\geq\sum_{n=1}^{N}\left(c[\Phi(u_h^n)-\Phi(u_h^{n-1})],1\right)=\left(c[\Phi(u_h^N)-\Phi(u_h^0)],1\right).$$

Using (2.5), we have

$$\Phi(u_h^N)\geq\frac{1}{2\beta^*}(\theta_h^N)^2\quad\text{ and }\quad\Phi(u_h^0)\leq\frac{\beta^*}{2}(u_h^0)^2.$$

Thus, by (3.1), we obtain

$$\sum_{n=1}^{N}(c[u_h^n-u_h^{n-1}],\theta_h^n)\geq\frac{c_*}{2\beta^*}\|\theta_h^N\|_{L^2(\Omega)}^2-\frac{c^*\beta^*}{2}\|u_h^0\|_{L^2(\Omega)}^2.$$

With this, (3.2), and the Schwarz inequality, (3.16) can be written as

$$\frac{c_*}{2\beta^*}\|\theta_h^N\|_{L^2(\Omega)}^2-\frac{c^*\beta^*}{2}\|u_h^0\|_{L^2(\Omega)}^2+\alpha_*\sum_{n=1}^{N}\|\nabla\theta_h^n\|_{L^2(\Omega)}^2\Delta t^n$$

$$\leq C\sum_{n=1}^{N}\left\{\|f^n\|_{H^{-1}(\Omega)}\Delta t^n+\|\bar{u}_h^{n-1}-u_h^{n-1}\|_{H^{-1}(\Omega)}\right\}\|\theta_h^n\|_{H^1(\Omega)},$$

which, together with Lemma 3.1, the discrete Gronwall inequality, the fact that $|\mathcal{U}(s)|\leq C_1|s|+C_2$, and a kick-back argument, yields the desired result. ∎

LEMMA 3.3. *With the same assumptions as in Lemma* 3.2, *if* $h=O(\Delta t)$, *we have*

$$\sum_{n=1}^{N}\left\|\frac{u_h^n-\bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}^2\Delta t^n\leq C.$$

Proof. Choose $v = G_h(u_h^n - \bar{u}_h^{n-1})/\Delta t^n$ in (3.7) to see that

$$\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) + \left(\alpha\nabla\theta_h^n, \nabla\frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) = \left(f^n, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right).$$

By the definition of G_h in (3.12), this equation becomes

$$\begin{aligned}\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) + \left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \theta_h^n\right) - \left(\frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}, \theta_h^n\right) \\ = \left(f^n, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right).\end{aligned} \tag{3.17}$$

Using (3.14), we have

$$\begin{aligned}&\left|\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \theta_h^n\right) - \left(\frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}, \theta_h^n\right)\right| \le C\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)} \|\theta_h^n\|_{H^1(\Omega)}, \\ &\left|\left(f^n, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right)\right| \le C\|f^n\|_{H^{-1}(\Omega)}\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}.\end{aligned} \tag{3.18}$$

Also, note that

$$\begin{aligned}&\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) \\ &\quad = \left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{G(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) + \left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{(G_h - G)(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right),\end{aligned}$$

so, by (3.13),

$$\begin{aligned}&\left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, \frac{G_h(u_h^n - \bar{u}_h^{n-1})}{\Delta t^n}\right) \\ &\quad \ge C\left(\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}^2 - \frac{1}{(\Delta t^n)^2}\|(G_h - G)(u_h^n - \bar{u}_h^{n-1})\|_{H^1(\Omega)}^2\right) \\ &\quad \ge C\left(\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}^2 - \frac{h^2}{(\Delta t^n)^2}\|u_h^n - \bar{u}_h^{n-1}\|_{L^2(\Omega)}^2\right).\end{aligned} \tag{3.19}$$

Substitute (3.18) and (3.19) into (3.17) and use Lemma 3.2 to obtain the desired result. ∎

3.4. Error analysis I. For the next lemma, we need the assumption

$$c \in W^{1,\infty}(\Omega) \quad \text{and} \quad \frac{b}{c} \in \left(W^{1,\infty}(\Omega)\right)^2. \tag{3.20}$$

The proof of the following lemma follows a similar treatment of Lemma 3.1 in [20]. However, the argument is simpler and the assumptions are weakened.

LEMMA 3.4. *With assumptions* (3.1) *and* (3.20)*, we have, for* Δt *sufficiently small,*

$$(c\bar{v}, G\bar{v}) - (cv, Gv) \le C\Delta t(cv, Gv) \qquad \forall v \in L^2(\Omega),$$

where $\bar{v} = v(x - b(x)\Delta t/c(x))$.

Proof. Let $y = x - b(x)\Delta t/c(x) \equiv H(x)$ for each $x \in \Omega$. The Jacobian of this transformation is

$$J(H(x)) = \begin{pmatrix} 1 - \partial_{x_1}\left(\frac{b_1}{c}\right)\Delta t & -\partial_{x_2}\left(\frac{b_1}{c}\right)\Delta t \\ -\partial_{x_1}\left(\frac{b_2}{c}\right)\Delta t & 1 - \partial_{x_2}\left(\frac{b_2}{c}\right)\Delta t \end{pmatrix},$$

so its determinant equals

$$|J(H(x))| = 1 - \nabla \cdot \left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right).$$

Note that, for Δt sufficiently small, $|J(H(x))| > 0$. Also, since H maps the periodic domain Ω onto itself, a change of variable leads to

$$\begin{aligned}(c\bar{v}, G\bar{v}) &= \int_\Omega c(x)v(y)Gv(y)\mathrm{d}x \\ &= \int_\Omega c(x)v(y)Gv(y)\frac{1}{|J(H(x))|}\mathrm{d}y \\ &= \int_\Omega c(x)v(y)Gv(y)\left(1 + \nabla \cdot \left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right)\right)\mathrm{d}y.\end{aligned}$$

Subtracting (cv, Gv) from both sides of this equation, we see that

$$\begin{aligned}(c\bar{v}, G\bar{v}) - (cv, Gv) &= \int_\Omega [c(x) - c(y)]v(y)Gv(y)\left(1 + \nabla \cdot \left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right)\right)\mathrm{d}y \\ &\quad + \int_\Omega c(y)v(y)Gv(y)\left(\nabla \cdot \left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right)\right)\mathrm{d}y \\ &\equiv A_1 + A_2.\end{aligned}$$

Note that

$$|c(x) - c(y)| \le \|\nabla c\|_{L^\infty(\Omega)}|x - y| \le \|\nabla c\|_{L^\infty(\Omega)}\left\|\frac{b}{c}\right\|_{L^\infty(\Omega)}\Delta t,$$

so that, by (3.1) and (3.20),

$$A_1 \le C\Delta t(cv, Gv).$$

Also, by (3.1), it is obvious that

$$A_2 \le C\Delta t(cv, Gv).$$

The bounds for A_1 and A_2 complete the proof. ∎

3.4.1. Case A: Under assumption (2.5). We first derive error estimates under assumption (2.5); i.e., (2.6) is degenerate. Note that, by the periodicity assumption, (3.3) (or (2.6)) can be written in the weak form

$$(\psi(x)\partial_\tau u, v) + (\alpha(x)\nabla\theta, \nabla v) = (f, v) \qquad \forall v \in H^1(\Omega). \tag{3.21}$$

Set

$$u_h(\cdot, t) = u_h^n(\cdot) \quad \text{and} \quad \theta_h(\cdot, t) = \theta_h^n(\cdot) \quad \text{for } t \in J^n, \quad n = 1, \dots, N.$$

Also, let

$$e_u(t) = u(t) - u_h(t), \quad e_\theta(t) = \theta(t) - \theta_h(t), \quad \text{and} \quad e_u^n = e_u(t^n).$$

Below ϵ is a positive constant independent of h and Δt, as small as we please. Also, whenever Lemma 3.4 is used, we will implicitly require that Δt be sufficiently small.

THEOREM 3.5. *Under assumptions* (2.1), (2.2), (2.5), (3.1), (3.2), (3.10), (3.15), (3.20), *and* $h = O(\Delta t)$, *we have the error estimate*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^N \|\theta^n - \theta_h^n\|^2_{L^2(\Omega)} \Delta t^n \le C\Delta t(1 + \Delta t \, \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}). \tag{3.22}$$

Proof. Take $v = Ge_u^n$ in (3.21) with $t = t^n$ and $v = G_h e_u^n$ in (3.7), subtract the resulting two equations, use (3.9) and (3.12), and sum over n from 1 to N to have the error equation

$$\begin{aligned}
&\sum_{n=1}^N \left(c[e_u^n - \bar{e}_u^{n-1}], Ge_u^n\right) + \sum_{n=1}^N (ce_\theta^n, e_u^n)\Delta t^n \\
&= \sum_{n=1}^N (f^n, Ge_u^n - G_h e_u^n)\Delta t^n + \sum_{n=1}^N \left\{(Ge_u^n, \theta^n) - (G_h e_u^n, \theta_h^n)\right\} \Delta t^n \\
&\quad - \sum_{n=1}^N \left(\psi\partial_\tau u^n - c\frac{u^n - \bar{u}^{n-1}}{\Delta t^n}, Ge_u^n\right) \Delta t^n - \sum_{n=1}^N \left(c\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}, Ge_u^n - G_h e_u^n\right) \Delta t^n.
\end{aligned} \tag{3.23}$$

With the obvious definition of I_i, $i = 1, \dots, 6$, we express (3.23) simply by

$$I_1 + I_2 = I_3 + I_4 + I_5 + I_6.$$

First, by (3.9) and (3.11), note that

$$\begin{aligned}
I_1 &= \sum_{n=1}^N a\left(G[e_u^n - \bar{e}_u^{n-1}], Ge_u^n\right) \\
&\ge \frac{1}{2}\sum_{n=1}^N \left\{a(Ge_u^n, Ge_u^n) - a(G\bar{e}_u^{n-1}, G\bar{e}_u^{n-1})\right\} \\
&= \frac{1}{2}\sum_{n=1}^N \{\|e_u^n\|^2_{H^{-1}(\Omega)} - \|e_u^{n-1}\|^2_{H^{-1}(\Omega)}\} + \frac{1}{2}\sum_{n=1}^N \{\|e_u^{n-1}\|^2_{H^{-1}(\Omega)} - \|\bar{e}_u^{n-1}\|^2_{H^{-1}(\Omega)}\},
\end{aligned}$$

so that, using Lemma 3.4,

$$I_1 \geq \frac{1}{2}\|e_u^N\|^2_{H^{-1}(\Omega)} - \frac{1}{2}\|e_u^0\|^2_{H^{-1}(\Omega)} - C\sum_{n=1}^{N}\|e_u^{n-1}\|^2_{H^{-1}(\Omega)}\Delta t^{n-1}.$$

Second, by (2.5) (i.e., (2.4)) and (3.1), we see that

$$I_2 \geq C\sum_{n=1}^{N}\|e_\theta^n\|^2_{L^2(\Omega)}\Delta t^n.$$

Third, apply (3.13) to have

$$\begin{aligned}|I_3| &\leq \sum_{n=1}^{N}\|f^n\|_{H^{-1}(\Omega)}\|Ge_u^n - G_he_u^n\|_{H^1(\Omega)}\Delta t^n\\ &\leq Ch\sum_{n=1}^{N}\|f^n\|_{H^{-1}(\Omega)}\|e_u^n\|_{L^2(\Omega)}\Delta t^n\\ &\leq Ch\left(\sum_{n=1}^{N}\|f^n\|^2_{H^{-1}(\Omega)}\Delta t^n + \sum_{n=1}^{N}\|e_u^n\|^2_{L^2(\Omega)}\Delta t^n\right).\end{aligned}$$

Fourth, it follows from (3.13) and (3.14) that

$$|I_4| \leq \epsilon\sum_{n=1}^{N}\|e_\theta^n\|^2_{L^2(\Omega)}\Delta t^n + C\sum_{n=1}^{N}(\|e_u^n\|^2_{H^{-1}(\Omega)} + h^2\|\theta^n\|^2_{L^2(\Omega)})\Delta t^n.$$

Fifth, exploit the standard backward (in the characteristic direction) difference analysis [25] and (3.11) to see that

$$\begin{aligned}|I_5| &\leq \sum_{n=1}^{N}\left\|\psi\partial_\tau u^n - c\frac{u^n - \bar{u}^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}\|Ge_u^n\|_{H^1(\Omega)}\Delta t^n\\ &\leq C\left((\Delta t)^2\left\|\partial_{\tau\tau}u\right\|^2_{L^2(J;H^{-1}(\Omega))} + \sum_{n=1}^{N}\|e_u^n\|^2_{H^{-1}(\Omega)}\Delta t^n\right).\end{aligned}$$

Sixth, by (3.13), we have

$$\begin{aligned}|I_6| &\leq \sum_{n=1}^{N}\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}\|Ge_u^n - G_he_u^n\|_{H^1(\Omega)}\Delta t^n\\ &\leq Ch\sum_{n=1}^{N}\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|_{H^{-1}(\Omega)}\|e_u^n\|_{L^2(\Omega)}\Delta t^n\\ &\leq Ch\left(\sum_{n=1}^{N}\left\|\frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n}\right\|^2_{H^{-1}(\Omega)}\Delta t^n + \sum_{n=1}^{N}\|e_u^n\|^2_{L^2(\Omega)}\Delta t^n\right).\end{aligned}$$

Finally, apply the bounds of I_i, $i = 1, \ldots, 6$, to (3.23) and use Lemmas 3.2 and 3.3, the discrete Gronwall lemma, (3.6), and (3.8) to obtain the desired result. ∎

Note that if the norm

$$\|\partial_{\tau\tau} u\|_{L^2(J;H^{-1}(\Omega))} \quad \text{is bounded,} \tag{3.24}$$

the left-hand side of (3.22) is bounded by $C\Delta t$. Assumption (3.24) appears physically reasonable, since the solution is much smoother along the characteristic direction. An error estimate without this assumption will be obtained in section 3.5. To check the optimality of the error estimate in Theorem 3.5, notice that we have utilized the whole regularity at our disposal. The estimate seems sharp under the present assumptions (2.5), (3.1), and (3.15) on the data; see the numerical example in section 5. Finally, we have derived an estimate for the error $u - u_h$ in the $H^{-1}(\Omega)$ norm. With the present assumption and technique, we are not able to obtain it in the $L^2(\Omega)$ norm due to the minimum regularity on the solution and the degeneracy of (1.1). For an estimate in this norm, see the next two subsections.

3.4.2. Case B: A nondegenerate case. For completeness, we also consider a nondegenerate case and derive error estimates under the minimal regularity on the solution. That is, instead of (2.5), (only) in this subsection we assume that

$$0 < \beta_* \le \beta(v) \le \beta^* < \infty \qquad \forall\, v \in \Re. \tag{3.25}$$

As mentioned before, (3.25) needs to hold only in a neighborhood of the solution. With this assumption, we now prove the next result.

THEOREM 3.6. *Under assumptions* (2.1), (2.2), (3.1), (3.2), (3.10), (3.15), (3.20), (3.25), *and* $h = O(\Delta t)$, *we have*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N} \{\|u^n - u_h^n\|^2_{L^2(\Omega)} + \|\theta^n - \theta_h^n\|^2_{L^2(\Omega)}\}\Delta t^n \le C(\Delta t)^2(1 + \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

Proof. The proof of Theorem 3.5 can be modified as follows. The error equation (3.23) and the estimate on I_1 and I_5 are the same as before. I_2, I_3, I_4, and I_6 are now estimated in a different manner:

$$\begin{aligned}
I_2 &\ge C \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)} \Delta t^n,\\
|I_3| &\le Ch^2 \sum_{n=1}^{N} \|f^n\|^2_{H^{-1}(\Omega)} \Delta t^n + \epsilon \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)} \Delta t^n,\\
|I_4| &\le \epsilon \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)} \Delta t^n + C \sum_{n=1}^{N} (\|e_u^n\|^2_{H^{-1}(\Omega)} + h^2 \|\theta^n\|^2_{L^2(\Omega)}) \Delta t^n,\\
|I_6| &\le Ch^2 \sum_{n=1}^{N} \left\| \frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n} \right\|^2_{H^{-1}(\Omega)} \Delta t^n + \epsilon \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)} \Delta t^n,
\end{aligned}$$

where (3.25) has been used. With these modifications, the desired result follows in the same fashion as for Theorem 3.5. ∎

The error estimate in the $L^2(\Omega)$ norm in this theorem in the nondegenerate case has been previously obtained in [25]; it is derived here, however, under much weaker regularity assumptions on the solution. Also, note that the estimate for $u - u_h$ in $L^2(J; H^1(\Omega))$ is optimal.

3.4.3. Case C: Another degenerate case. Note that assumption (2.5) implies that the diffusion coefficient in (2.6) (or in (3.3)) can be zero in the unknown u. We now consider another case where the following inequality holds:

$$0 \le \partial_\theta \mathcal{U}(\eta) \le \mathcal{U}^* < \infty \qquad \forall\, \eta \in \Re. \tag{3.26}$$

Recall that $\mathcal{U}$ is the inverse of (2.3). Again, (3.26) needs to hold only in a neighborhood of θ. This assumption says that the coefficient in the time differentiation term of (2.6) (or (3.3)) can be zero in θ. This case is sometimes referred to as the singular case [30, 35].

In this subsection we derive the error estimate under (3.26). Toward that end, we need another assumption. Let V be a subspace of $H^1(\Omega)$, and assume that

$$\|\nabla v\|_{L^2(\Omega)} \text{ is a norm in } V \text{ and is equivalent to } \|v\|_{H^1(\Omega)}. \tag{3.27}$$

In turn, this requires that the Poincare inequality holds in V. For example, if (2.6) is equipped with the Dirichlet boundary condition on the part of the boundary that has a positive Hausdorff measure, then we can define V in an appropriate way so that (3.27) holds. With (3.27), we can define the Green operator $G : H^{-1}(\Omega) \to V$ in terms of the bilinear form

$$a(v, w) = (\alpha \nabla v, \nabla w) \qquad \forall v,\ w \in V.$$

For future use (see the discussion on the extension of the present analysis for the characteristics-based methods to other boundary conditions later), we now state a result under assumptions (3.26) and (3.27).

THEOREM 3.7. *Under assumptions* (2.1), (2.2), (3.1), (3.2), (3.10), (3.15), (3.20), (3.26), (3.27), $h = O(\Delta t)$, *and* $M_h \subset V$, *we have*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N} \|u^n - u_h^n\|^2_{L^2(\Omega)} \Delta t^n \le C(\Delta t)^2 (1 + \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

Because of the definition of $a(\cdot,\cdot)$ in the present case, $I_4 \equiv 0$ in the error equation (3.23). Then the proof of Theorem 3.6 applies here.

3.5. Error analysis II. In this subsection, we derive error estimates without assumption (3.24).

LEMMA 3.8. *For* $v \in H^1(\Omega)$, *let* $\bar v(x) = v(\bar x)$, *where* $\bar x = x - b(x)\Delta t/c(x)$. *Then, under* (3.1) *and* (3.20), *we have*

$$\|v - \bar v\|_{L^2(\Omega)} \le C\Delta t \|\nabla v\|_{L^2(\Omega)}.$$

Proof. Let $z(x)$ be the unit vector in the direction of $x - \bar x$; i.e., $z(x) = x - \bar x/|x - \bar x|$. Then

$$v(x) - \bar v(x) = \int_0^1 |x - \bar x| \partial_z v(x + \zeta z) \mathrm{d}\zeta,$$

so

$$\int_\Omega |v(x) - \bar v(x)|^2 \mathrm{d}x = \int_\Omega |x - \bar x|^2 \left(\int_0^1 \partial_z v \mathrm{d}\zeta \right)^2 \mathrm{d}x.$$

Applying the facts that $|x-\bar{x}| \le C\Delta t$ and the determinant of the Jacobian of the transformation $x \to \bar{x}$ is $1+O(\Delta t)$ by (3.1) and (3.20), we see that

$$\int_\Omega |v(x)-\bar{v}(x)|^2 \mathrm{d}x \le C(\Delta t)^2 \|\nabla v\|^2_{L^2(\Omega)}.$$

Consequently, the desired result follows. □

In the present case, we need a slightly stronger assumption on the data:

$$b,\ \frac{b}{c} \in (H^2(\Omega))^2. \tag{3.28}$$

Also, whenever we use $1/|b(x)|$ below, we will assume that $|b(x)| \neq 0$ there.

LEMMA 3.9. *Under assumptions* (2.1), (2.2), (3.1), (3.10), (3.20), *and* (3.28), *we have*

$$\int_{J^n} \left(b\cdot\nabla u - c\frac{u^{n-1}-\bar{u}^{n-1}}{\Delta t^n}, Ge_u^n \right) dt \le C\Delta t^n (\|\partial_t u\|^2_{L^2(J^n;H^{-1}(\Omega))} + \Delta t^n), \quad 1 \le n \le N.$$

Proof. Let $z(x)$ be the unit vector in the direction of $b(x)$; i.e., $z(x) = b(x)/|b(x)|$. Then we see that

$$c(x)\frac{u^{n-1}-\bar{u}^{n-1}}{\Delta t^n} = \int_0^1 |b(x)| \partial_z u \left(x - \frac{b(x)}{c(x)}\Delta t^n \zeta, t^{n-1} \right) \mathrm{d}\zeta = \int_0^1 b(x)\cdot\nabla_y u^{n-1}(y)\mathrm{d}\zeta,$$

where $y = x - b(x)\Delta t^n \zeta / c(x)$. We denote this transformation by $y = H_\zeta(x)$ for each fixed $\zeta \in [0,1]$. Its Jacobian and determinant are, respectively,

$$J(H_\zeta(x)) = \begin{pmatrix} 1-\partial_{x_1}\left(\frac{b_1}{c}\right)\Delta t^n\zeta & -\partial_{x_2}\left(\frac{b_1}{c}\right)\Delta t^n\zeta \\ -\partial_{x_1}\left(\frac{b_2}{c}\right)\Delta t^n\zeta & 1-\partial_{x_2}\left(\frac{b_2}{c}\right)\Delta t^n\zeta \end{pmatrix}$$

and

$$|J(H_\zeta(x))| = 1 - \nabla\cdot\left(\frac{b}{c}(x)\right)\Delta t^n\zeta + O\left((\Delta t^n)^2\right).$$

Let

$$F_\zeta(x) = 1 + \nabla\cdot\left(\frac{b}{c}(x)\right)\Delta t^n\zeta + O\left((\Delta t^n)^2\right).$$

Then, as in the proof of Lemma 3.4, we have

$$\int_\Omega b(x)\cdot\nabla_y u^{n-1}(y) G e_u^n(x)\mathrm{d}x = \int_\Omega b(x)\cdot\nabla_y u^{n-1}(y) G e_u^n(x) F_\zeta(x)\mathrm{d}y,$$

so, by the periodicity assumption and the Green formula,

$$\begin{aligned}\int_\Omega b(x)\cdot\nabla_y u^{n-1}(y) G e_u^n(x)\mathrm{d}x = &-\int_\Omega \nabla_y\cdot b(x) u^{n-1}(y) G e_u^n(x) F_\zeta(x)\mathrm{d}y \\ &-\int_\Omega b(x)\cdot\nabla_y\left(G e_u^n(x)\right) u^{n-1}(y) F_\zeta(x)\mathrm{d}y \\ &-\int_\Omega u^{n-1}(y) G e_u^n(x) b(x)\cdot\nabla_y F_\zeta(x)\mathrm{d}y.\end{aligned}$$

Apply the periodicity assumption and the Green formula again to see that

$$\int_\Omega b(y)\cdot\nabla_y u(y)Ge_u^n(y)\mathrm{d}y = -\int_\Omega \nabla_y\cdot b(y)u(y)Ge_u^n(y)\mathrm{d}y - \int_\Omega b(y)\cdot\nabla_y\left(Ge_u^n(y)\right)u(y)\mathrm{d}y.$$

Subtract these two equations to obtain

$$\begin{aligned}
&\int_\Omega b(y)\cdot\nabla_y u(y)Ge_u^n(y)\mathrm{d}y - \int_\Omega b(x)\cdot\nabla_y u^{n-1}(y)Ge_u^n(x)\mathrm{d}x\\
&\quad= -\left\{\int_\Omega \nabla_y\cdot b(y)u(y)Ge_u^n(y)\mathrm{d}y - \int_\Omega \nabla_y\cdot b(x)u^{n-1}(y)Ge_u^n(x)F_\zeta(x)\mathrm{d}y\right\}\\
&\qquad-\left\{\int_\Omega b(y)\cdot\nabla_y\left(Ge_u^n(y)\right)u(y)\mathrm{d}y - \int_\Omega b(x)\cdot\nabla_y\left(Ge_u^n(x)\right)u^{n-1}(y)F_\zeta(x)\mathrm{d}y\right\}\\
&\qquad+\int_\Omega u^{n-1}(y)Ge_u^n(x)b(x)\cdot\nabla_y F_\zeta(x)\mathrm{d}y\\
&\quad\equiv II_1+II_2+II_3.
\end{aligned}$$

Observe that

$$\begin{aligned}
II_1 &= -\int_\Omega \nabla_y\cdot b(y)Ge_u^n(y)[u(y)-u^{n-1}(y)]\mathrm{d}y\\
&\quad-\left\{\int_\Omega \nabla_y\cdot b(y)u^{n-1}(y)Ge_u^n(y)\mathrm{d}y - \int_\Omega \nabla_y\cdot b(x)u^{n-1}(y)Ge_u^n(y)F_\zeta(x)\mathrm{d}y\right\}\\
&\quad-\left\{\int_\Omega \nabla_y\cdot b(x)u^{n-1}(y)Ge_u^n(y)F_\zeta(x)\mathrm{d}y - \int_\Omega \nabla_y\cdot b(x)u^{n-1}(y)Ge_u^n(x)F_\zeta(x)\mathrm{d}y\right\}\\
&\equiv II_{11}+II_{12}+II_{13}.
\end{aligned}$$

For $t\in J^n$, apply the Schwarz inequality, (2.1), (3.1), and (3.10) to have

$$\begin{aligned}
|II_{11}| &= \left|\int_\Omega\left\{\int_{t^{n-1}}^t \partial_t u(y,\xi)\mathrm{d}\xi\right\}\nabla_y\cdot b(y)Ge_u^n(y)\mathrm{d}y\right|\\
&\le C(\|\partial_t u\|^2_{L^2(J^n;H^{-1}(\Omega))}+\Delta t^n).
\end{aligned}$$

With a similar argument as in Lemma 3.4 and an application of Lemma 3.8 on $\nabla\cdot b$ and Ge_u^n, we can show that

$$|II_{12}|+|II_{13}|\le C\Delta t^n.$$

Analogously, with the expression

$$\begin{aligned}
II_2 &= -\int_\Omega b(y)\cdot\nabla_y Ge_u^n(y)[u(y)-u^{n-1}(y)]\mathrm{d}y\\
&\quad-\left\{\int_\Omega b(y)\cdot\nabla_y Ge_u^n(y)u^{n-1}(y)\mathrm{d}y - \int_\Omega b(x)\cdot\nabla_y Ge_u^n(y)u^{n-1}(y)F_\zeta(x)\mathrm{d}y\right\}\\
&\quad-\left\{\int_\Omega b(x)\cdot\nabla_y Ge_u^n(y)u^{n-1}(y)F_\zeta(x)\mathrm{d}y - \int_\Omega b(x)\cdot\nabla_y Ge_u^n(x)u^{n-1}(y)F_\zeta(x)\mathrm{d}y\right\},
\end{aligned}$$

we see that

$$|II_2|\le C(\|\partial_t u\|^2_{L^2(J^n;H^{-1}(\Omega))}+\Delta t^n).$$

Obviously, by (3.28), we have

$$|II_3| \le C\Delta t^n.$$

Finally, note that

$$\begin{aligned}&\int_{J^n}\left(b\cdot\nabla u - c\frac{u^{n-1}-\bar u^{n-1}}{\Delta t^n}, Ge_u^n\right)\mathrm{d}t\\ &\quad=\int_{J^n}\int_0^1\left\{\int_\Omega b(y)\cdot\nabla_y u(y)Ge_u^n(y)\mathrm{d}y - \int_\Omega b(x)\cdot\nabla_y u^{n-1}(y)Ge_u^n(x)\mathrm{d}x\right\}\mathrm{d}\zeta\mathrm{d}t.\end{aligned}$$

Consequently, the desired result comes from the bounds for II_i, $i=1,2,3$. ∎

We are now in a position to derive error estimates. As an example, we consider only Case A; i.e., we derive the error estimates under assumption (2.5). Cases B and C can be similarly handled.

For each $1\le n\le N$, we integrate (3.21) over J^n:

$$\int_{J^n}(\psi(x)\partial_\tau u, v)\mathrm{d}t + \int_{J^n}(\alpha(x)\nabla\theta,\nabla v)\mathrm{d}t = \int_{J^n}(f,v)\mathrm{d}t \qquad \forall v\in H^1(\Omega). \tag{3.29}$$

THEOREM 3.10. *Under assumptions* (2.1), (2.2), (2.5), (3.1), (3.2), (3.10), (3.15), (3.20), (3.28), *and* $h=O(\Delta t)$, *we have the error estimate*

$$\max_{1\le n\le N}\|u^n-u_h^n\|_{H^{-1}(\Omega)} + \|\theta-\theta_h\|_{L^2(\Omega_T)} \le C(\Delta t)^{1/2},$$

provided that $\|\partial_t f\|_{L^2(J;H^{-1}(\Omega))}$ *is bounded.*

Proof. Choose $v=Ge_u^n$ in (3.29) and $v=G_he_u^n$ in (3.7), subtract the resulting two equations, use (3.9) and (3.12), and sum over n from 1 to N to have the error equation

$$\begin{aligned}&\sum_{n=1}^N\left(c[e_u^n-\bar e_u^{n-1}], Ge_u^n\right) + \sum_{n=1}^N\int_{J^n}(c[\theta(t)-\theta_h^n], e_u^n)\,\mathrm{d}t\\ &=\sum_{n=1}^N\left\{\int_{J^n}(f-f^n, Ge_u^n)\mathrm{d}t + (f^n, Ge_u^n - G_he_u^n)\Delta t^n\right\}\\ &\quad+\sum_{n=1}^N\left\{\int_{J^n}(\theta, Ge_u^n)\mathrm{d}t - (\theta_h^n, G_he_u^n)\Delta t^n\right\}\\ &\quad-\sum_{n=1}^N\int_{J^n}\left(b\cdot\nabla u - c\frac{u^{n-1}-\bar u^{n-1}}{\Delta t^n}, Ge_u^n\right)\mathrm{d}t\\ &\quad-\sum_{n=1}^N\left(c\frac{u_h^n-\bar u_h^{n-1}}{\Delta t^n}, Ge_u^n - G_he_u^n\right)\Delta t^n.\end{aligned} \tag{3.30}$$

Again, with the obvious definition of I_i, $i=1,\dots,6$, we express (3.30) simply by

$$I_1+I_2 = I_3+I_4+I_5+I_6.$$

The term I_1 is estimated as in Theorem 3.5:

$$I_1 \ge \frac{1}{2}\|e_u^N\|^2_{H^{-1}(\Omega)} - \frac{1}{2}\|e_u^0\|^2_{H^{-1}(\Omega)} - C\sum_{n=1}^N\|e_u^{n-1}\|^2_{H^{-1}(\Omega)}\Delta t^{n-1}.$$

I_2 is written as

$$I_2 = \sum_{n=1}^{N} \int_{J^n} (c[\theta(t) - \theta_h^n], u(t) - u_h^n) \mathrm{d}t + \sum_{n=1}^{N} \int_{J^n} (c[\theta(t) - \theta_h^n], u^n - u(t)) \mathrm{d}t,$$

so, by (2.1), (2.5), (3.1), and Lemma 3.2,

$$I_2 \geq C\{\|\theta - \theta_h\|^2_{L^2(\Omega_T)} - \Delta t\}.$$

Next, by (3.13), we see that

$$\begin{aligned}
|I_3| &\leq \sum_{n=1}^{N} \{(\Delta t^n)^{1/2} \|\partial_t f\|_{L^2(J^n;H^{-1}(\Omega))} \|Ge_u^n\|_{H^1(\Omega)} + \|f^n\|_{H^{-1}(\Omega)} \|Ge_u^n - G_h e_u^n\|_{H^1(\Omega)}\} \Delta t^n \\
&\leq C \sum_{n=1}^{N} \{(\Delta t^n)^{1/2} \|\partial_t f\|_{L^2(J^n;H^{-1}(\Omega))} \|e_u^n\|_{H^{-1}(\Omega)} + h\|f^n\|_{H^{-1}(\Omega)} \|e_u^n\|_{L^2(\Omega)}\} \Delta t^n \\
&\leq C \left\{ (\Delta t)^2 \|\partial_t f\|^2_{L^2(J;H^{-1}(\Omega))} + \sum_{n=1}^{N} (\|e_u^n\|^2_{H^{-1}(\Omega)} + h\|f^n\|^2_{H^{-1}(\Omega)} + h\|e_u^n\|^2_{L^2(\Omega)}) \Delta t^n \right\}.
\end{aligned}$$

Also, it follows from (3.13) and (3.14) that

$$|I_4| \leq \epsilon \|e_\theta\|^2_{L^2(\Omega_T)} + C \left\{ \sum_{n=1}^{N} \|e_u^n\|^2_{H^{-1}(\Omega)} \Delta t^n + h^2 \|\theta\|^2_{L^2(\Omega_T)} \right\}.$$

Apply Lemma 3.9 to I_5 to see that

$$|I_5| \leq C\Delta t (1 + \|\partial_t u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

Finally, by (3.13), I_6 is bounded as follows:

$$\begin{aligned}
|I_6| &\leq \sum_{n=1}^{N} \left\| \frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n} \right\|_{H^{-1}(\Omega)} \|Ge_u^n - G_h e_u^n\|_{H^1(\Omega)} \Delta t^n \\
&\leq Ch \sum_{n=1}^{N} \left\| \frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n} \right\|_{H^{-1}(\Omega)} \|e_u^n\|_{L^2(\Omega)} \Delta t^n \\
&\leq Ch \left(\sum_{n=1}^{N} \left\| \frac{u_h^n - \bar{u}_h^{n-1}}{\Delta t^n} \right\|^2_{H^{-1}(\Omega)} \Delta t^n + \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)} \Delta t^n \right).
\end{aligned}$$

The rest of the proof is completed as in Theorem 3.5. ∎

We remark that the error estimate in Theorem 3.10 appears sharp under the present assumptions on the data, as mentioned before. Also, the present analysis can be extended to the more general nonlinear problem

$$c(u)\partial_t u + b(u) \cdot \nabla u - \nabla \cdot (a(u)\nabla u) = f(u) \qquad \text{in } \Omega \times J, \tag{3.31}$$

where $c(u) = c(x,t;u)$, $b(u) = b(x,t;u)$, and $f(u) = f(x,t;u)$. By linearizing b, c, and f, we can simply reduce (3.31) to (2.6). To be more accurate, we can use an extrapolation technique in the linearization of these coefficients [21].

4. Analysis for the MMOCAA. In this section, we carry out an analysis for the MMOCAA procedure.

4.1. The MMOCAA procedure. To introduce this procedure, we assume that

$$\nabla \cdot b = 0 \qquad \text{in } \Omega. \tag{4.1}$$

That is, b is divergence-free. This is physically reasonable, since b is typically a velocity field and (4.1) corresponds to the incompressibility condition. Note that, by (4.1), the periodicity assumption, and the divergence theorem, (2.6) with $f = 0$ yields the conservation law

$$\int_\Omega c(x)u(x,t)\mathrm{d}x = \int_\Omega c(x)u_0(x)\mathrm{d}x, \qquad t \in J. \tag{4.2}$$

In real applications, it is desirable to maintain at least a discrete form of this law in any numerical approximation of (2.6). However, in general, the MMOC procedure does not satisfy this property, and it creates an imbalance in mass [22].

Let M_h be defined as in the previous section, and let the initial approximation u_h^0 be defined as in (3.8). For $1 \le n \le N$, given $u_h^{n-1} \in M_h$, set

$$Q_h^{n-1} = \int_\Omega c(x)u_h^{n-1}(x)\mathrm{d}x, \qquad \bar{Q}_h^{n-1} = \int_\Omega c(x)\bar{u}_h^{n-1}(x)\mathrm{d}x.$$

As mentioned above, $Q_h^{n-1} \neq \bar{Q}_h^{n-1}$ in general. Define

$$\tilde{u}_h^{n-1}(x) = \begin{cases} \max\left\{ u_h^{n-1}\left(\bar{x} - \gamma\dfrac{b(x)}{c(x)}\left(\Delta t^n\right)^2\right), u_h^{n-1}\left(\bar{x} + \gamma\dfrac{b(x)}{c(x)}\left(\Delta t^n\right)^2\right)\right\} \\ \quad \text{if} \quad \bar{Q}_h^{n-1} < Q_h^{n-1}, \\ \min\left\{ u_h^{n-1}\left(\bar{x} - \gamma\dfrac{b(x)}{c(x)}\left(\Delta t^n\right)^2\right), u_h^{n-1}\left(\bar{x} + \gamma\dfrac{b(x)}{c(x)}\left(\Delta t^n\right)^2\right)\right\} \\ \quad \text{if} \quad \bar{Q}_h^{n-1} > Q_h^{n-1}, \end{cases}$$

and

$$\tilde{Q}_h^{n-1} = \int_\Omega c(x)\tilde{u}_h^{n-1}(x)\mathrm{d}x,$$

where γ is a fixed constant, normally chosen to be less than one [22], and $\bar{x} = x - b(x)\Delta t^n/c(x)$. If $\bar{Q}_h^{n-1} = \tilde{Q}_h^{n-1}$, we must accept that mass cannot be conserved; otherwise, find $\Lambda^{n-1} \in \Re$ such that

$$Q_h^{n-1} = \Lambda^{n-1}\bar{Q}_h^{n-1} + (1 - \Lambda^{n-1})\tilde{Q}_h^{n-1}. \tag{4.3}$$

Define

$$\hat{u}_h^{n-1} = \Lambda^{n-1}\bar{u}_h^{n-1} + (1 - \Lambda^{n-1})\tilde{u}_h^{n-1} \tag{4.4}$$

and

$$\hat{Q}_h^{n-1} = \int_\Omega c(x)\hat{u}_h^{n-1}(x)\mathrm{d}x. \tag{4.5}$$

Clearly, $\hat{Q}_h^{n-1} = Q_h^{n-1}$, so the conservation law is preserved. Now, continue in n with $\hat{u}_h^{n-1}$ in place of $\bar{u}_h^{n-1}$ in the original MMOC procedure (3.7); i.e.,

$$\left(c\frac{u_h^n - \hat{u}_h^{n-1}}{\Delta t^n}, v\right) + (\alpha\nabla\theta_h^n, \nabla v) = (f^n, v) \qquad \forall v \in M_h, \tag{4.6}$$

where $u_h^n = \mathcal{U}(\theta_h^n)$. Note that Λ^{n-1} is bounded; $0 \le \Lambda^{n-1} \le 1$ for small Δt^{n-1} [22].

4.2. Stability. The next lemma can be found in [44].

LEMMA 4.1. *Let $\eta \in L^2(\Omega)$ be Ω-periodic, and let $g_i \in (W^{1,\infty}(\Omega))^2$, $i = 1, 2$. For $\Lambda \in \Re$, define*

$$\hat{\eta}(x) = \Lambda\bar{\eta}(x) + (1-\Lambda)\tilde{\eta}(x) = \Lambda\eta(x - g_1(x)\Delta t) + (1-\Lambda)\eta(x - g_2(x)\Delta t),$$

where

$$\|g_1 - g_2\|_{L^\infty(\Omega)} \le C\Delta t.$$

Then

$$\|\eta - \hat{\eta}\|_{H^{-1}(\Omega)} \le C(\Lambda)\Delta t\|\eta\|_{L^2(\Omega)}.$$

LEMMA 4.2. *Under assumptions* (2.5), (3.1), (3.2), *and* (3.15), *the solution (u_h, θ_h) produced by the above MMOCAA procedure satisfies*

$$\max_{1\le n\le N}\{\|u_h^n\|^2_{L^2(\Omega)} + \|\theta_h^n\|^2_{L^2(\Omega)}\} + \sum_{n=1}^{N}\|\nabla\theta_h^n\|^2_{L^2(\Omega)}\Delta t^n \le C.$$

The proof of this lemma can be carried out as for Lemma 3.2. Namely, with $v = \theta_h^n$ in (4.6), we have

$$\begin{aligned}&\sum_{n=1}^{N}\left\{(c[u_h^n - u_h^{n-1}], \theta_h^n) + (\alpha\nabla\theta_h^n, \nabla\theta_h^n)\Delta t^n\right\}\\ &= \sum_{n=1}^{N}\left\{(f^n, \theta_h^n)\Delta t^n + (c[\hat{u}_h^{n-1} - u_h^{n-1}], \theta_h^n)\right\}.\end{aligned}$$

Now, applying Lemma 4.1 and the same argument as for Lemma 3.2, we can obtain the desired result.

LEMMA 4.3. *With the same assumptions as in Lemma* 4.2, *if $h = O(\Delta t)$, we have*

$$\sum_{n=1}^{N}\left\|\frac{u_h^n - \hat{u}_h^{n-1}}{\Delta t^n}\right\|^2_{H^{-1}(\Omega)}\Delta t^n \le C.$$

Note that, with $v = G_h(u_h^n - \hat{u}_h^{n-1})/\Delta t^n$ in (4.6), this equation becomes

$$\begin{aligned}&\left(c\frac{u_h^n - \hat{u}_h^{n-1}}{\Delta t^n}, \frac{G_h(u_h^n - \hat{u}_h^{n-1})}{\Delta t^n}\right) + \left(\alpha\nabla\theta_h^n, \nabla\frac{G_h(u_h^n - \hat{u}_h^{n-1})}{\Delta t^n}\right)\\ &\qquad = \left(f^n, \frac{G_h(u_h^n - \hat{u}_h^{n-1})}{\Delta t^n}\right),\end{aligned}$$

so the proof of this lemma can be completed as in Lemma 3.3.

4.3. Error analysis I. The next lemma is similar to Lemma 3.8.

LEMMA 4.4. *For $v \in H^1(\Omega)$, let $\bar{v}(x) = v(\bar{x})$ and $\tilde{v}(x) = v(\tilde{x})$, where $\bar{x} = x - b(x)\Delta t/c(x)$ and $\tilde{x} = \bar{x} - \gamma b(x)(\Delta t)^2/c(x)$ or $\tilde{x} = \bar{x} + \gamma b(x)(\Delta t)^2/c(x)$. Then, under* (3.1) *and* (3.20),

$$\|\bar{v} - \tilde{v}\|_{L^2(\Omega)} \le C(\Delta t)^2\|\nabla v\|_{L^2(\Omega)}.$$

Note that $|\bar{x} - \tilde{x}| \le C(\Delta t)^2$ by (3.1), so that this lemma can be shown as in Lemma 3.8.

LEMMA 4.5. *With assumptions* (3.1) *and* (3.20)*, we have*

$$(c\hat{v}, G\hat{v}) - (cv, Gv) \le C(\Lambda)\Delta t\{(cv, Gv) + \Delta t(v, v)^{1/2}(cv, Gv)^{1/2}\} \qquad \forall v \in L^2(\Omega),$$

where, for $\Lambda \in \Re$, $\hat{v}(x) = \Lambda\bar{v}(x) + (1-\Lambda)\tilde{v}(x)$*, with* $\bar{x}$ *and* $\tilde{x}$ *given as in Lemma* 4.4.

Proof. As in the proof of Lemma 3.4, we can show that, for $v \in L^2(\Omega)$,

$$(c\bar{v}, G\bar{v}) - (cv, Gv) \le C\Delta t(cv, Gv) \quad \text{and} \quad (c\tilde{v}, G\tilde{v}) - (cv, Gv) \le C\Delta t(cv, Gv).$$

By the definition of $\hat{v}$, we see that

$$\begin{aligned}(c\hat{v}, G\hat{v}) - (cv, Gv) = {} & \Lambda^2\left\{(c\bar{v}, G\bar{v}) - (cv, Gv)\right\} + (1-\Lambda)^2\left\{(c\tilde{v}, G\tilde{v}) - (cv, Gv)\right\} \\ & + \Lambda(1-\Lambda)\left\{(c\bar{v}, G\tilde{v}) + (c\tilde{v}, G\bar{v}) - 2(cv, Gv)\right\}.\end{aligned} \tag{4.7}$$

Consequently, it suffices to bound

$$\Lambda(1-\Lambda)\left\{(c\bar{v}, G\tilde{v}) + (c\tilde{v}, G\bar{v}) - 2(cv, Gv)\right\}.$$

Observe that

$$\begin{aligned}(c\tilde{v}, G\bar{v}) - (cv, Gv) &= \int_\Omega c(x)v(\tilde{x})Gv(\bar{x})\mathrm{d}x - \int_\Omega c(\tilde{x})v(\tilde{x})Gv(\tilde{x})\mathrm{d}\tilde{x} \\ &= \int_\Omega c(x)v(\tilde{x})[Gv(\bar{x}) - Gv(\tilde{x})]\mathrm{d}x \\ &\quad + \int_\Omega [c(x) - c(\tilde{x})]v(\tilde{x})Gv(\tilde{x})\left(1 + \nabla\cdot\left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right)\right)\mathrm{d}\tilde{x} \\ &\quad + \int_\Omega c(\tilde{x})v(\tilde{x})Gv(\tilde{x})\left(\nabla\cdot\left(\frac{b}{c}(x)\right)\Delta t + O\left((\Delta t)^2\right)\right)\mathrm{d}\tilde{x} \\ &\equiv T_1 + T_2 + T_3.\end{aligned}$$

By Lemma 4.4, (3.1), and (3.11), we see that

$$\begin{aligned}|T_1| &\le C\|v\|_{L^2(\Omega)}\|G\bar{v} - G\tilde{v}\|_{L^2(\Omega)} \\ &\le C(\Delta t)^2\|v\|_{L^2(\Omega)}\|\nabla Gv\|_{L^2(\Omega)} \\ &\le C(\Delta t)^2\|v\|_{L^2(\Omega)}\|v\|_{H^{-1}(\Omega)}.\end{aligned}$$

By (3.20), it is clear that

$$|T_2 + T_3| \le C\Delta t(cv, Gv).$$

Thus we have

$$(c\tilde{v}, G\bar{v}) - (cv, Gv) \le C\Delta t\{(cv, Gv) + \Delta t(v, v)^{1/2}(cv, Gv)^{1/2}\}. \tag{4.8}$$

In the same manner, we see that

$$(c\bar{v}, G\tilde{v}) - (cv, Gv) \le C\Delta t\{(cv, Gv) + \Delta t(v, v)^{1/2}(cv, Gv)^{1/2}\}. \tag{4.9}$$

Therefore, the lemma follows by combining (4.7)–(4.9). ∎

4.3.1. Case A: Under assumption (2.5). As in the last section, we first derive the error estimate under assumption (2.5).

THEOREM 4.6. *Under assumptions* (2.1), (2.2), (2.5), (3.1), (3.2), (3.10), (3.15), (3.20), *and* $h = O(\Delta t)$, *we have the error estimate for the MMOCAA procedure:*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N} \|\theta^n - \theta_h^n\|^2_{L^2(\Omega)} \Delta t^n \le C\Delta t(1 + \Delta t \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

Proof. Take $v = Ge_u^n$ in (3.21) with $t = t^n$ and $v = G_h e_u^n$ in (4.6), subtract the resulting two equations, use (3.9) and (3.12), and sum over n from 1 to N to have the error equation

$$\begin{aligned}
&\sum_{n=1}^{N} \left(c[e_u^n - \hat{e}_u^{n-1}], Ge_u^n\right) + \sum_{n=1}^{N} (ce_\theta^n, e_u^n)\Delta t^n \\
&= \sum_{n=1}^{N} (f^n, Ge_u^n - G_h e_u^n)\Delta t^n + \sum_{n=1}^{N} \{(Ge_u^n, \theta^n) - (G_h e_u^n, \theta_h^n)\}\, \Delta t^n \\
&\quad - \sum_{n=1}^{N} \left(\psi\partial_\tau u^n - c\frac{u^n - \hat{u}^{n-1}}{\Delta t^n}, Ge_u^n\right) \Delta t^n - \sum_{n=1}^{N} \left(c\frac{u_h^n - \hat{u}_h^{n-1}}{\Delta t^n}, Ge_u^n - G_h e_u^n\right) \Delta t^n.
\end{aligned}$$

As in Theorem 3.5, we express this equation by

$$I_1 + I_2 = I_3 + I_4 + I_5 + I_6.$$

The terms I_2–I_4 are estimated in the same way as in Theorem 3.5; I_1, I_5, and I_6 are bounded as follows. First,

$$I_1 \ge \frac{1}{2} \sum_{n=1}^{N} \{\|e_u^n\|^2_{H^{-1}(\Omega)} - \|e_u^{n-1}\|^2_{H^{-1}(\Omega)}\} + \frac{1}{2} \sum_{n=1}^{N} \{\|e_u^{n-1}\|^2_{H^{-1}(\Omega)} - \|\hat{e}_u^{n-1}\|^2_{H^{-1}(\Omega)}\},$$

so, by Lemma 4.5,

$$I_1 \ge \frac{1}{2}\|e_u^N\|^2_{H^{-1}(\Omega)} - \frac{1}{2}\|e_u^0\|^2_{H^{-1}(\Omega)} - C\sum_{n=1}^{N} (\|e_u^{n-1}\|^2_{H^{-1}(\Omega)} + \|e_u^{n-1}\|^2_{L^2(\Omega)}\Delta t^{n-1})\Delta t^{n-1}.$$

Second, note that

$$\begin{aligned}
&\psi\partial_\tau u^n - c\frac{u^n - \hat{u}^{n-1}}{\Delta t^n} \\
&\quad = \Lambda^{n-1}\psi\partial_\tau u^n - \Lambda^{n-1} c\frac{u^n - \bar{u}^{n-1}}{\Delta t^n} + (1 - \Lambda^{n-1})\psi\partial_\tau u^n - (1 - \Lambda^{n-1})c\frac{u^n - \tilde{u}^{n-1}}{\Delta t^n};
\end{aligned}$$

consequently, as in Theorem 3.5 and by the boundedness of Λ^{n-1},

$$|I_5| \le C\left((\Delta t)^2 \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))} + \sum_{n=1}^{N} \|e_u^n\|^2_{H^{-1}(\Omega)}\Delta t^n\right).$$

Finally, we see that

$$|I_6| \le Ch\left(\sum_{n=1}^{N} \left\|\frac{u_h^n - \hat{u}_h^{n-1}}{\Delta t^n}\right\|^2_{H^{-1}(\Omega)} \Delta t^n + \sum_{n=1}^{N} \|e_u^n\|^2_{L^2(\Omega)}\Delta t^n\right).$$

Apply the bounds of I_i, $i = 1, \ldots, 6$, Lemmas 4.2 and 4.3, the discrete Gronwall lemma, (3.6), and (3.8) to obtain the desired result. ∎

Again, the error estimate in Theorem 4.6 for the MMOCAA procedure seems sharp under the present assumptions on the data; see section 5.

4.3.2. Case B: A nondegenerate case. We now obtain the error estimate under assumption (3.25); i.e., we present the error analysis for the MMOCAA in the nondegenerate case. This case was analyzed in [23], but here a reduced regularity on solution is employed, as noted earlier.

THEOREM 4.7. *Under assumptions* (2.1), (2.2) (3.1), (3.2), (3.10), (3.15), (3.20), (3.25), *and* $h = O(\Delta t)$, *we have*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N} \{\|u^n - u_h^n\|^2_{L^2(\Omega)} + \|\theta^n - \theta_h^n\|^2_{L^2(\Omega)}\}\Delta t^n \le C(\Delta t)^2(1 + \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

This theorem can be shown as in Theorem 3.6 with an exception that I_1 is estimated by

$$I_1 \ge \frac{1}{2}\|e_u^N\|^2_{H^{-1}(\Omega)} - \frac{1}{2}\|e_u^0\|^2_{H^{-1}(\Omega)} - C\sum_{n=1}^{N} \|e_u^{n-1}\|^2_{H^{-1}(\Omega)}\Delta t^{n-1} - \epsilon \sum_{n=1}^{N} \|e_u^{n-1}\|^2_{L^2(\Omega)}\Delta t^{n-1}.$$

As for the MMOC, the estimate for $u - u_h$ in $L^2(J; H^1(\Omega))$ is optimal.

4.3.3. Case C: Another degenerate case. We finally state the error estimate under assumptions (3.26) and (3.27) for the MMOCAA procedure.

THEOREM 4.8. *Under assumptions* (2.1), (2.2), (3.1), (3.2), (3.10), (3.15), (3.20), (3.26), (3.27), $h = O(\Delta t)$, *and* $M_h \subset V$, *for the MMOCAA we have*

$$\max_{1\le n\le N} \|u^n - u_h^n\|^2_{H^{-1}(\Omega)} + \sum_{n=1}^{N} \|u^n - u_h^n\|^2_{L^2(\Omega)}\Delta t^n \le C(\Delta t)^2(1 + \|\partial_{\tau\tau} u\|^2_{L^2(J;H^{-1}(\Omega))}).$$

The proof is given as in Theorem 3.7.

4.4. Error analysis II. We now treat the case without assumption (3.24). As an example, we state only the result corresponding to Case A; the other two cases can be handled in a similar fashion. Also, the proof of the next theorem can be completed as in Theorems 3.10 and 4.6.

THEOREM 4.9. *Under assumptions* (2.1), (2.2), (2.5), (3.1), (3.2), (3.10), (3.15), (3.20), (3.28), *and* $h = O(\Delta t)$, *we have the error estimate for the MMOCAA procedure*

$$\max_{1\le n\le N} \|u^n - u_h^n\|_{H^{-1}(\Omega)} + \|\theta - \theta_h\|_{L^2(\Omega_T)} \le C(\Delta t)^{1/2},$$

provided that $\|\partial f/\partial t\|_{L^2(J;H^{-1}(\Omega))}$ *is bounded.*

Remark. Another variant of the MMOC procedure has been recently introduced in [8]. It is referred to as the ELLAM there. This method globally conserves mass as well. The error analysis for the ELLAM (with the degeneracy taken into account) can be done similarly as in this section for the MMOCAA. We mention that the analysis for the ELLAM in a nondegenerate case has been given in [45].

A similar variant to the ELLAM scheme has been developed in [2]; it is called the characteristic-mixed finite element method (also see [24]). In this method, the diffusion term of (2.6) is treated using the classical mixed finite element method [7], where the diffusion coefficient is assumed to be positive. The present techniques have been developed primarily for degenerate problems. For the characteristic-mixed method to be able to treat the degenerate case, we can exploit the so-called expanded mixed finite element method [9, 10] for the discretization of the diffusion term. The development and analysis of the characteristic-expanded mixed finite element method will be our future work.

5. Numerical results. In this section, we present numerical results to show the sharpness of the error estimates derived in the earlier sections. We consider the so-called porous medium equation

$$\partial_t u - \Delta u^m = 0, \qquad m > 1.$$

This equation can be equivalently rewritten in form (1.1):

$$\partial_t u - \nabla \cdot (m u^{m-1} \nabla u) = 0, \qquad m > 1, \tag{5.1}$$

so we see that the diffusion coefficient $a(u)$ is mu^{m-1} and the variable θ equals u^m. Obviously, (5.1) is degenerate at zero. Equation (5.1) often arises in the flow of a gas in porous media. To see this, ignoring certain constants, the gas flow is governed by

$$\partial_t \rho + \nabla \cdot (\rho v) = 0, \quad v = -\nabla p, \quad \rho = p^\gamma, \tag{5.2}$$

where ρ is the density, p the pressure, v the velocity, and γ a (constant) ratio of specific heats. These equations are the mass conservation, Darcy's law, and equation of state [6, 27, 29, 38], respectively. Eliminating v and p in (5.2), we see that

$$\partial_t \rho - \frac{1}{1+\gamma} \Delta(\rho^{1+1/\gamma}) = 0.$$

Rescaling t by $1/(1+\gamma)$ leads to (5.1) with $u = \rho$.

Beginning from a delta function of integral Γ at the original, the exact solution to (5.1) is of the form [4, 37]

$$u(|x|, t) = \max\left\{0, t^{-\alpha}\left(\Gamma - \frac{\alpha(m-1)}{2dm}\frac{|x|^2}{t^{2\alpha/d}}\right)^{1/(m-1)}\right\},$$

where $\alpha = 1/(m - 1 + 2/d)$. This function is radially symmetric and has compact support. Figure 1 shows an example of this solution in two dimensions.

The finite element procedure (3.7) is utilized to solve (5.1). Since we are solely interested in checking the sharpness of the error estimates in Theorem 3.5, we concentrate on the one-dimensional case, $d = 1$. Also, we take $m = 2$ in (5.1). Note that $b = 0$ in the present case, so (3.7) reduces to the standard finite element method with a backward Euler procedure for $\partial_t u$. Further, with the present choice of the initial datum, (5.1) corresponds to the flow case with a point source. The approximate solutions with different mesh sizes and at different times are presented in Figure 2. This figure shows convergence of the approximate solutions. The error bounds and convergence rates in the $L^2(\Omega_T)$ norm for u and θ with $T = 0.01$, $\Gamma = 1.0$, and $m = 2$ are given in Table 1. From this table, we observe the sharpness of the error estimates in Theorem 3.5. Similar observations can be made for Theorems 3.10, 4.6, and 4.9.

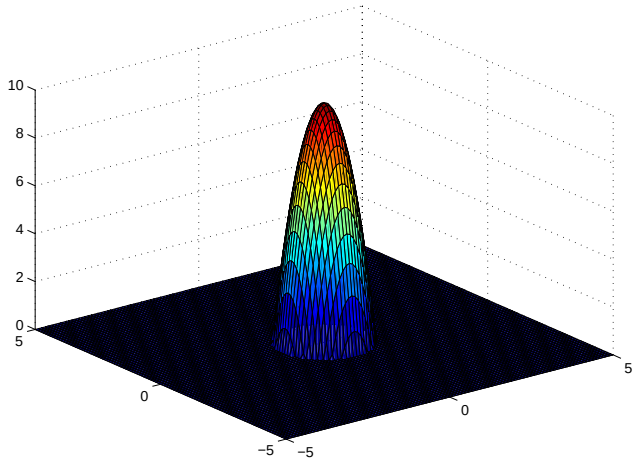


FIG. 1.

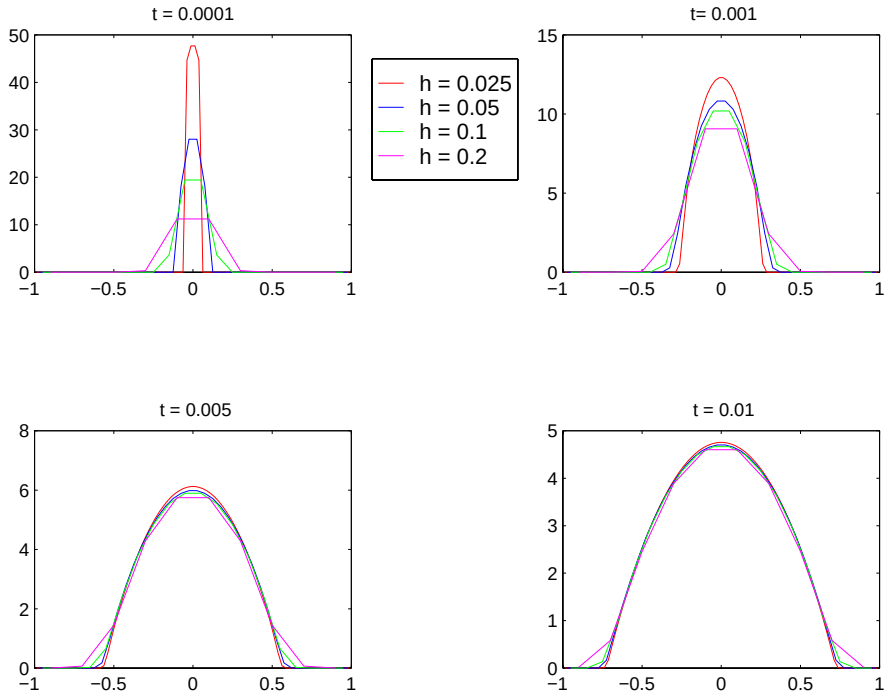


FIG. 2.

TABLE 1
The convergence rates for u and θ.

h	error for u	rate for u	error for θ	rate for θ
0.2	8.189736e-01	-	9.541634e+01	-
0.1	5.796918e-01	0.49853	6.746879e+01	0.50002
0.05	4.109090e-01	0.49647	4.770738e+01	0.50001
0.025	2.920137e-01	0.49278	3.373394e+01	0.50001

REFERENCES

[1] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations,* Math. Z., 183 (1983), pp. 311–341.

[2] T. Arbogast and M. F. Wheeler, *A characteristics-mixed finite element for advection-dominated transport problems,* SIAM J. Numer. Anal., 32 (1995), pp. 404–424.

[3] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media,* SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[4] G. I. Barenblatt, *On some unsteady motions of a liquid or a gas in a porous medium,* Akad. Nauk SSSR. Prikl. Mat. Meh., 16 (1952), pp. 67–78.

[5] J. W. Barrett and P. Knabner, *Finite element approximation of the transport of reactive solutes in porous media, Part* II. *Error estimates for equilibrium adsorption processes,* SIAM J. Numer. Anal., 34 (1997), pp. 455–479.

[6] J. Bear, *Dynamics of Fluids in Porous Media,* Dover, New York, 1972.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods,* Springer-Verlag, New York, 1991.

[8] M. A. Celia, T. F. Russell, I. Herrera, and R. E. Ewing, *An Eulerian Lagrangian localized adjoint method for the advection-diffusion equation,* Adv. in Water Res., 13 (1990), pp. 187–206.

[9] Z. Chen, *Expanded mixed finite element methods for linear second order elliptic problems* I, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 479–499.

[10] Z. Chen, *Expanded mixed finite element methods for quasilinear second order elliptic problems* II, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 501–520.

[11] Z. Chen, *Formulations and numerical methods of the black oil model in porous media,* SIAM J. Numer. Anal., 38 (2000), pp. 489–514.

[12] Z. Chen, *Degenerate two-phase incompressible flow* I: *Existence, uniqueness and regularity of a weak solution,* J. Differential Equations, 171 (2001), pp. 203–232.

[13] Z. Chen and R. E. Ewing, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology,* SIAM J. Numer. Anal., 34 (1997), pp. 2228–2253.

[14] Z. Chen and R. Ewing, *Degenerate two-phase incompressible flow* III: *Sharp error estimates,* Numer. Math., 90 (2001), pp. 215–240.

[15] Z. Chen and N. L. Khlopina, *Degenerate two-phase incompressible flow problems* I: *Regularization and numerical results,* Commun. Appl. Anal., 5 (2001), pp. 319–334.

[16] Z. Chen and N. L. Khlopina, *Degenerate two-phase incompressible flow problems* II: *Error estimates,* Commun. Appl. Anal., 5 (2001), pp. 503–521.

[17] Z. Chen and N. L. Khlopina, *Degenerate two-phase incompressible flow problems* III: *Perturbation analysis and numerical experiments,* Electronic J. Differential Equations, 2 (1999), pp. 29–46.

[18] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978.

[19] C. N. Dawson, C. J. van Duijn, and M. F. Wheeler, *Characteristic-Galerkin methods for contaminant transport with non-equilibrium adsorption kinetics,* SIAM J. Numer. Anal., 31 (1994), pp. 982–999.

[20] C. N. Dawson, T. F. Russell, and M. F. Wheeler, *Some improved error estimates for the modified method of characteristics,* SIAM J. Numer. Anal., 26 (1989), pp. 1487–1512.

[21] J. Douglas, Jr., *A survey of numerical methods for parabolic differential equations,* in Advances in Computers, Vol. 2, F. L. Alt, ed., Academic Press, New York, 1961, pp. 1–54.

[22] J. Douglas, Jr., F. Furtado, and F. Pereira, *On the numerical simulation of waterflooding of heterogeneous petroleum reservoirs,* Comput. Geosci., 1 (1997), pp. 155–190.

[23] J. Douglas, Jr., C. S. Huang, and F. Pereira, *The modified method of characteristics with adjusted advection,* Numer. Math., 83 (1999), pp. 353–369.

[24] J. Douglas, Jr., F. Pereira, and L. M. Yeh, *A Locally Conservative Eulerian-Lagrangian Numerical Method and Its Application to Nonlinear Transport in Porous Media,* Technical report 324, Center for Applied Mathematics, Purdue University, West Lafayette, IN, 1998.

[25] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures,* SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[26] C. J. van Duijn and P. Knabner, *Solute transport in porous media with equilibrium and non-equilibrium multiple-site adsorption: Travelling waves,* J. Reine Angew. Math., 415 (1991), pp. 1–49.

[27] R. Ewing, ed., *The Mathematics of Reservoir Simulation,* Frontiers Appl. Math. 1, SIAM, Philadelphia, 1984.

[28] K. Fadimba and R. Sharpley, *A priori estimates and regularization for a class of porous medium equations,* Nonlinear World, 2 (1995), pp. 13–41.
[29] C. Fetter, *Contaminant Hydrogeology,* Macmillan, New York, 1993.
[30] A. Friedman, *Variational Principles and Free-Boundary Problems,* John Wiley and Sons, New York, 1982.
[31] R. Helmig, *Multiphase Flow and Transport Processes in the Subsurface: A Contribution to the Modeling of Hydrosystems,* Springer-Verlag, Heidelberg, 1997.
[32] J. W. Jerome and M. Rose, *Error estimates for the multidimensional two-phase Stefan problem,* Math. Comp., 39 (1982), pp. 377–414.
[33] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method,* Cambridge University Press, Cambridge, UK, 1987.
[34] N. L. Khlopina, *Finite Element Methods for Degenerate Two-Phase Incompressible Flow Problems,* Ph.D. thesis, Department of Mathematics, Southern Methodist University, Dallas, TX, 1999.
[35] R. Nochetto, *Finite element methods for parabolic free boundary problem,* 1990 *Lancaster Summer School Proceedings,* in Advances in Numerical Analysis, Vol. I: Nonlinear Partial Differential Equations and Dynamical Systems, Oxford University Press, New York, 1991, pp. 34–95.
[36] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration,* SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[37] R. E. Pattle, *Diffusion from an instantaneous point source with a concentration-dependent coefficient,* Quart. J. Mech. Appl. Math., 12 (1959), pp. 407–409.
[38] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation,* Elsevier, New York, 1977.
[39] O. Pironneau, *On the transport-diffusion algorithm and its application to the Navier-Stokes equations,* Numer. Math., 38 (1982), pp. 309–332.
[40] M. Rose, *Numerical Methods for flow through porous media* I, Math. Comp., 40 (1983), pp. 437–467.
[41] T. F. Russell, *Time stepping along characteristics with incomplete iteration for a Galerkin approximation of miscible displacement in porous media,* SIAM J. Numer. Anal., 22 (1985), pp. 970–1013.
[42] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media.* in The Mathematics of Reservoir Simulation, Frontiers Appl. Math. 1, R. Ewing, ed., SIAM, Philadelphia, 1983, pp. 35–106.
[43] D. Smylie, *A Near Optimal Order Approximation to a Class of Two Sided Nonlinear Degenerate Parabolic Partial Differential Equations,* Ph.D. thesis, University of Wyoming, Laramie, WY, 1989.
[44] A. M. Spagnuolo, *Approximation of Nuclear Contaminant Transport Through Porous Media,* Ph.D. thesis, Technical report 319, Center for Applied Mathematics, Purdue University, West Lafayette, IN, 1998.
[45] H. Wang, R. E. Ewing, and T. F. Russell, *Eulerian-Lagrangian localized adjoint methods for convection-diffusion equations and their convergence analysis,* IMA J. Numer. Anal., 15 (1995), pp. 405–459.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1516–1537

PREDICTOR-CORRECTOR METHODS OF RUNGE–KUTTA TYPE FOR STOCHASTIC DIFFERENTIAL EQUATIONS*

KEVIN BURRAGE† AND TIANHAI TIAN†

Abstract. In this paper we construct predictor-corrector (PC) methods based on the trivial predictor and stochastic implicit Runge–Kutta (RK) correctors for solving stochastic differential equations. Using the colored rooted tree theory and stochastic B-series, the order condition theorem is derived for constructing stochastic RK methods based on PC implementations. We also present detailed order conditions of the PC methods using stochastic implicit RK correctors with strong global order 1.0 and 1.5. A two-stage implicit RK method with strong global order 1.0 and a four-stage implicit RK method with strong global order 1.5 used as the correctors are constructed in this paper. The mean-square stability properties and numerical results of the PC methods based on these two implicit RK correctors are reported.

Key words. stochastic differential equations, predictor-corrector methods, Runge–Kutta methods, numerical stability

AMS subject classifications. 60H10, 65L06, 65L20

PII. S0036142900372677

1. Introduction.

Runge–Kutta (RK) methods are one of the most efficient classes of methods for solving ordinary differential equations (ODEs). Certain classes of implicit RK methods have excellent stability properties and are widely used to solve stiff ODEs. In the last decade, predictor-corrector (PC) methods have been one of the major classes of methods for solving nonstiff ODEs on parallel computers (see [2], [3], [4], [5], [6], [10], [15], [25], [27], and [28]).

By comparing the Taylor series expansion of the approximation solution to the Taylor series expansion of the exact solution over one step assuming exact initial values, Butcher [13] introduced the rooted tree theory that is the key to constructing RK methods for ODEs. As the RK-type PC methods can be represented as a special class of block explicit RK methods, the rooted tree theory has been applied to RK-type PC methods. Burrage [2], [6] has developed a comprehensive theory based on the use of Butcher series which allows the analysis of the local error of any RK-type PC method and has also applied this theory to an analysis of the local behavior of two classes of PC methods, including one which is based on the trivial predictor and an implicit RK corrector.

For solving stochastic differential equations (SDEs), stochastic RK methods are an important class of numerical methods. Rümelin [22] introduced the use of traditional RK methods for SDEs. These methods resemble in their structure deterministic RK methods for ODEs. Burrage and Burrage [7], [8] and Burrage [12] established the colored rooted tree theory and stochastic B-series which is generalized from the corresponding rooted tree theory and B-series for constructing numerical methods for ODEs. Based on these theories, Burrage and Burrage present order conditions for constructing a general class of stochastic RK methods for solving Stratonovich SDEs and also construct an explicit strong global order 1.0 two-stage RK method with

*Received by the editors May 24, 2000; accepted for publication (in revised form) March 6, 2002; published electronically October 23, 2002.
http://www.siam.org/journals/sinum/40-4/37267.html
†Department of Mathematics, The University of Queensland, Brisbane, QLD 4072, Australia (kb@maths.uq.edu.au, tian@maths.uq.edu.au).

minimum principal error constants [17] and an explicit five-stage RK method with strong global order 1.5 [8]. Tian and Burrage [26] consider diagonally semi-implicit and implicit strong order 1.0 two-stage RK methods with good stability properties or good accuracy. In addition, in order to avoid the unboundedness of numerical solutions of the implicit stochastic RK methods, composite RK methods are constructed which are a combination of semi-implicit RK methods and implicit RK methods. Further research has been done by Komori, Mitsui, and Sugiura [18], in which they use the tree expansions of the true and numerical solutions to construct ROW-type schemes for SDEs.

For solving SDEs, the PC technique has been already applied to linear multistep implicit methods [1]. For weak solutions of SDEs, Kloeden and Platen [17] and Platen [21] consider families of PC methods with weak order 1.0 and 2.0. In this paper we consider PC methods using stochastic RK methods as correctors for strong solutions of SDEs. In section 2, we first give a brief review of the rooted tree theory for constructing RK methods and RK-type PC methods for ODEs and then give order conditions for constructing stochastic RK-type PC methods after a brief review of the colored rooted tree theory for constructing stochastic RK methods. In section 3, we give the detailed order conditions for two-stage RK-type PC methods with strong global order 1.0 and then construct a two-stage implicit RK method with strong global order 1.0. Similar work is done for four-stage RK-type PC methods with strong global order 1.5 in section 4. The mean-square stability properties of the RK-type PC methods using these two-stage and four-stage implicit RK correctors are considered in section 5. Numerical results are reported in section 6.

2. Order conditions for RK-type PC methods. In this section, a brief review is first given for the rooted tree theory and order conditions for constructing RK-type PC methods for ODEs. For solving the ODE

$$y'(t) = f(y(t)), \qquad y(t_0) = y_0, \qquad t \in [t_0, T], \quad y \in \mathbb{R}^m,$$

the class of s-stage RK methods is given by

$$\begin{aligned} Y_i &= y_n + h\sum_{j=1}^{s} a_{ij} f(Y_j), \qquad i = 1, 2, \dots, s, \\ y_{n+1} &= y_n + h\sum_{j=1}^{s} b_j f(Y_j), \end{aligned} \tag{2.1}$$

which can be represented by the so-called Butcher tableau

$$\begin{array}{c|c} \mathrm{c} & \mathrm{A} \\ \hline & \mathrm{b}^{\top} \end{array}, \qquad c = Ae, \qquad e = (1, \dots, 1)^{\top} \in \mathbb{R}^s. \tag{2.2}$$

In order to express derivatives of $f(y)$ systematically, Butcher [13] introduced the rooted tree theory which provides a general framework for studying order conditions of RK methods. Let T be the set of rooted trees and $t = [t_1, \dots, t_m]$ be the tree formed by joining subtrees $t_1, \dots, t_m$ each by a single branch to a common root. In addition, let ϕ denote the empty tree and τ the unique tree with one node. For each t, denote $\rho(t)$ as the number of nodes (vertices) of t, $h(t)$ as the height of t, with the height of the unique tree τ being 1, respectively. Then the elementary differential associated with $t = [t_1, \dots, t_m]$ is given by

$$F(t)y = f^{(m)}(F(t_1)y, \dots, F(t_m)y), \qquad F(\phi) = y.$$

With these definitions, the following order theorem holds for RK methods (see Burrage [6]).

THEOREM 2.1. *A RK method is of order w if and only if*

$$e(t) = 0 \qquad \forall \rho(t) \leq w,$$

where for any tree $t = [t_1, \ldots, t_m]$

$$e(\phi) = 0, \qquad e(t) = 1 - \rho(t) b^\top \prod_{j=1}^{m} k(t_j),$$

with

$$k(\phi) = e, \qquad k(t) = \rho(t) \prod_{j=1}^{m} \left(A k(t_j) \right).$$

Now consider a RK-type PC method which uses a RK corrector (2.1) and the trivial predictor based on the update value y_n, given by

$$\begin{aligned} Y^{(0)} &= (e \otimes I) y_n, \\ Y^{(k)} &= (e \otimes I) y_n + h(A \otimes I) f(Y^{(k-1)}), \qquad k = 1, 2, \ldots, l, \\ y_{n+1} &= y_n + h b^\top f(Y^{(l)}), \end{aligned}$$

where $Y = (Y_1^\top, \ldots, Y_s^\top)^\top$ and $f(Y) = (f(Y_1)^\top, \ldots, f(Y_s)^\top)^\top$. This method can be represented by a $(l+1)s$-stage explicit RK method, whose Butcher tableau is given by

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ \mathrm{c} & \mathrm{A} & 0 & & & \\ \mathrm{c} & 0 & \mathrm{A} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ \mathrm{c} & 0 & \cdots & 0 & \mathrm{A} & 0 \\ \hline & 0 & \cdots & 0 & 0 & b^\top \end{array} .$$

Applying the order conditions for RK methods (Theorem 2.1) to this $(l+1)s$-stage explicit RK method, Burrage [2], [3], [6] presents a theoretical tool for measuring the error behavior of this RK-type PC method and gives the order conditions of this method.

THEOREM 2.2. *If a RK corrector is applied to the trivial predictor with l corrections, then the local error is given by*

$$l_{n+1} = \sum_{t \in T^*} e(t) \left[F(t) \right] y(t_n) \frac{h^{\rho(t)}}{\rho(t)!},$$

where for $t = [t_1, \ldots, t_m]$

$$e(t) = 1 - \rho(t) b^\top \prod_{i=1}^{m} k_l(t_i), \tag{2.3}$$

with

$$
\begin{aligned}
&k_0(\phi) = e, \quad k_0(t) = 0, \quad \rho(t) > 0,\\
&k_{j+1}(t) = \rho(t) \prod_{i=1}^{m} \left(A k_j(t_i)\right), \qquad j = 0, 1, \ldots, l-1.
\end{aligned} \tag{2.4}
$$

By studying the behavior of the local errors of a RK-type PC method, Burrage [2], [6] has shown that each application of a corrector increases the order of the overall method by one until the order of the corrector is reached. In addition, when the number of corrections is such that the order cannot increase further, then the effect of more corrections is to shift the errors due to the predictor further away from the principal error terms.

Now we consider the order conditions of stochastic RK-type PC methods for the Stratonovich SDE driven by d-dimensional Wiener processes

$$
dy(t) = g_0(y(t))dt + \sum_{j=1}^{d} g_j(y(t)) \circ dW_j(t), \qquad y(t_0) = y_0, \quad y \in \mathbb{R}^m, \tag{2.5}
$$

where the deterministic term $g_0(y(t))$ is the drift coefficient, the stochastic term $g_j(y(t))$ $(j = 1, \ldots, d)$ are the diffusion coefficients, and $W_j(t)$ is the Wiener process, whose increment $\Delta W_j(t) = W_j(t + \Delta t) - W_j(t)$ is a Gaussian random variable $N(0, \Delta t)$.

The solution of (2.5) can be written in integral form as

$$
y(t) = y(t_0) + \int_{t_0}^{t} g_0(y(t))dt + \sum_{j=1}^{d} \int_{t_0}^{t} g_j(y(t)) \circ d_j W(t),
$$

and can also be expressed as a stochastic Taylor series, given by

$$
\begin{aligned}
y(t) = y_0 &+ \sum_{j_1=0}^{d} g_{j_1}(y_0) J_{j_1,t} + \sum_{j_1,j_2=0}^{d} L^{j_1} g_{j_2}(y_0) J_{j_1 j_2, t}\\
&+ \sum_{j_1,j_2,j_3=0}^{d} L^{j_1} L^{j_2} g_{j_3}(y_0) J_{j_1 j_2 j_3, t} + \cdots,
\end{aligned} \tag{2.6}
$$

where the Stratonovich operator is defined by

$$
L^j = \sum_{k=1}^{m} g_j^k \frac{\partial}{\partial y^k}, \qquad j = 0, 1, \ldots, d
$$

and $J_{j_1,\ldots,j_k,t}$ represents the Stratonovich multiple integral which is defined recursively by (see [16] and [17])

$$
\begin{aligned}
J_{0,t} &= \int_{t_0}^{t} dt = t - t_0,\\
J_{j,t} &= \int_{t_0}^{t} \circ dW_j(t) = \Delta W_j(t),\\
J_{j_1 j_2 \cdots j_{k-1} j_k, t} &= \int_{t_0}^{t} J_{j_1 j_2 \cdots j_{k-1}, t} dt, \qquad j_k = 0,\\
J_{j_1 j_2 \cdots j_{k-1} j_k, t} &= \int_{t_0}^{t} J_{j_1 j_2 \cdots j_{k-1}, t} \circ dW_j(t), \qquad j_k = j, \quad j = 1, \ldots, d.
\end{aligned}
$$

In order to express the stochastic Taylor series more precisely, Burrage and Burrage present the colored rooted tree theory [7] and stochastic B-series [8] which have the same structure as the corresponding rooted tree theory and B-series.

DEFINITION 2.3. *The $(d+1)$-colored rooted trees can be defined recursively by*

(i) *the elementary rooted tree is τ_k which represent the deterministic elementary rooted tree τ_0 if $k=0$ and a stochastic one τ_k if $k \in \{1,2,\dots,d\}$;*

(ii) *if $t_1,\dots,t_m$ are $(d+1)$-colored rooted trees, then $[t_1,\dots,t_m]_k$ is the $(d+1)$-colored rooted tree in which $t_1,\dots,t_m$ are each joined by a single branch to τ_k ($k \in \{1,2,\dots,d\}$).*

Similar to the rooted tree theory for ODEs, denote T_1 as the set of all $(d+1)$-colored rooted trees, $\rho(t)$ as the number of vertices of t, $\alpha(t)$ as the number of ways of labelling the vertices of t so that the labels increase outwardly along the arcs, $h(t)$ as the height of t where the height of the elementary tree is 1, and $\gamma(t)$ as the density of $t=[t_1,\dots,t_m]_k$, defined by

$$\gamma(t) = \rho(t)\prod_{j=1}^{m}\gamma(t_j)$$

and where $J(t)$ represents the corresponding J-integral associated with tree t which is defined by

$$J(t)(h) = \int_0^h \prod_{j=1}^{m} J(t_j)(s)\circ dW_k(s), \qquad J(\tau_k)(h) = W_k(h).$$

In a similar manner to the deterministic case, an elementary differential can be associated with any $t \in T_1$ such that

$$\begin{aligned} F(\tau_k)(y) &= g_k(y),\\ F(t)(y) &= g_k^{(m)}(y)[F(t_1)(y),\dots,F(t_m)(y)], \qquad t=[t_1,\dots,t_m]_k. \end{aligned}$$

With the definitions of $(d+1)$-colored rooted trees, Burrage and Burrage [7] and Burrage [12] have given the Taylor series expansion of the exact solution of an SDE.

THEOREM 2.4. *The Stratonovich–Taylor series for the actual solution of the SDE given by* (2.5) *(together with initial value $y(t_0)=y_0$) is*

$$y(t_0+h) = \sum_{t\in T_1}\frac{\gamma(t)}{\rho(t)!}J(t)\alpha(t)F(t)(y(t_0)),$$

where $F(t)(y)$ is the elementary differential defined by the structure of tree t, and $J(t)$ represents the corresponding J-integral associated with tree t.

For solving the SDE (2.5), a general class of s-stage stochastic RK method derived by Burrage and Burrage [7] and Burrage [12] is given by

$$\begin{aligned} Y_i &= y_n + \sum_{k=0}^{d}\sum_{j=1}^{s} Z_{ij}^{(k)} g_k(Y_j), \qquad i=1,\dots,s,\\ y_{n+1} &= y_n + \sum_{k=0}^{d}\sum_{j=1}^{s} z_j^{(k)} g_k(Y_j), \end{aligned} \tag{2.7}$$

where $Z_{ij}^{(k)}$ and $z_j^{(k)}$ are random variables, which are functions of h, to be determined based on order and stability analysis. Note that in the case of the deterministic parameters $Z^{(0)}$ and $z^{(0)}$, h is included implicitly in these terms.

The numerical solution obtained by the stochastic RK method (2.7) can be written as a Taylor series expansion [7], given by

$$y(t_0 + h) = \sum_{t \in T} \frac{\gamma(t)}{\rho(t)!} a(t)\alpha(t)F(t)(y(t_0)),$$

where, for $t = [t_1, \ldots, t_m]_k$, $a(t)$ is defined by

$$a(t) = z^{(k)\top}\Phi(t),$$
$$\Phi(t) = \prod_{i=1}^{m} (Z^{(k)}\Phi(t_i)), \qquad \Phi(\tau_k) = e.$$

In designing numerical schemes for solving SDEs, some criteria are needed to measure the efficiency of a numerical scheme by means of its order of convergence. There are two criteria to measure the convergence order: strong convergence and weak convergence. For problems involving direct simulations of paths, it is required that the simulated sample paths be close to the exact solution of the original SDE. This consideration leads to the strong convergence criterion (for example, see Burrage [12]).

DEFINITION 2.5. *Let y_N be the numerical approximation to $y(t_N)$ at time $T = Nh + t_0$ after N steps with constant stepsize h; then y is said to converge strongly to y with order p if $\exists C > 0$ (independent of h but dependent on the length of the time interval $T - t_0$) and $\delta > 0$ such that*

$$E(|y_N - y(t_N)|) \le Ch^p, \qquad h \in (0, \delta).$$

The local truncation error at $t = t_{n+1}$ of the stochastic RK method (2.7) can be written as

$$L_n = \sum_{t \in T_1} \frac{\gamma(t)}{\rho(t)!} \alpha(t) \left(J(t) - a(t)\right) F(t)(y(t_n)).$$

Burrage and Burrage [8] have given the following definition to measure the accuracy of the RK methods

DEFINITION 2.6. *This stochastic RK method will have strong local order p if*

$$E[|L_n|] = O(h^{p+\frac{1}{2}})$$

and will have mean local order p if

$$E(L_n) = O(h^{p+1}).$$

In addition they have proven the following theorem concerning the relationship between the local error behavior and the global error behavior (see also Milstein [19]), given by the following theorem.

THEOREM 2.7. *Let the g_j possess all necessary partial derivatives for all $y \in \mathbb{R}^m$; then if*

$$\left(E\left[||l_n||^2\right]\right)^{1/2} = O(h^{p+1/2}) \quad \forall n$$

and

$$E[l_n] = O\left(h^{p+1}\right) \quad \forall n,$$

then

$$\left(E\left[||\epsilon_N||^2\right]\right)^{1/2} = O\left(h^p\right),$$

where ϵ_N is the global error at step point t_N with the assumption of the exact initial solution of $y_0 = y(t_0)$.

Thus the stochastic RK method (2.7) is of strong global order p if it has strong local order p and mean local order p.

For solving the SDE (2.5), the stochastic RK-type PC method, which is based on a stochastic RK corrector (2.7) and the trivial predictor, is given by

$$\begin{aligned} Y^{(0)} &= (e \otimes I) y_n, \\ Y^{(i)} &= (e \otimes I) y_n + \sum_{k=0}^{d} (Z^{(k)} \otimes I) g_k(Y^{(i-1)}), \qquad i = 1, 2, \ldots, l, \\ y_{n+1} &= y_n + \sum_{k=0}^{d} (z^{(k)} \otimes I) g_k(Y^{(l)}), \end{aligned} \tag{2.8}$$

where $Y^{(i)} = (Y_1^{(i)\top}, \ldots, Y_s^{(i)\top})^\top$, $Z^{(k)} = (Z_{ij}^{(k)})_{s\times s}$, and $z^{(k)} = (z_1^{(k)}, \ldots, z_s^{(k)})$, $(k = 0, 1, \ldots, d)$. This stochastic RK-type PC method can be represented by an $(l+1)s$-stage block explicit stochastic RK method characterized by the tableau

$$\begin{array}{c|ccccc|ccccc|c} 0 & & & & & & 0 & & & & & \\ Z^{(0)} & 0 & & & & \cdots\ \cdots & Z^{(d)} & 0 & & & & \\ 0 & Z^{(0)} & 0 & & & \cdots\ \cdots & 0 & Z^{(d)} & 0 & & & \\ \vdots & \ddots & \ddots & \ddots & & \cdots\ \cdots & \vdots & \ddots & \ddots & \ddots & & \\ 0 & \cdots & 0 & Z^{(0)} & 0 & \cdots\ \cdots & 0 & \cdots & 0 & Z^{(d)} & 0 & \\ \hline 0 & \cdots & 0 & 0 & z^{(0)} & \cdots\ \cdots & 0 & \cdots & 0 & 0 & z^{(d)} & \end{array}.$$

Applying the order theorem for stochastic RK methods to (2.8), we have the main theorem on order conditions for constructing stochastic RK-type PC methods (2.8) in this paper.

THEOREM 2.8. *If a stochastic RK corrector is applied to the trivial predictor with l corrections, then the strong local error of the stochastic RK-type PC method is given by*

$$l_{n+1} = \sum_{t\in T_1} \frac{\gamma(t)}{\rho(t)!} e(t)\alpha(t)[F(t)]y(t_n),$$

where for $t = [t_1, \ldots, t_m]_k$ $e(t) = J(t) - a_l(t)$ and $a_l(t)$ is given by

$$a_l(t) = z^{(k)}\Phi_l(t),$$

and

$$\begin{aligned} \Phi_0(\tau_k) &= e, \quad \Phi_0(t) = 0, \quad \rho(t) \geq 2, \\ \Phi_{j+1}(t) &= \prod_{i=1}^{m} (Z^{(k)}\Phi_j(t_i)). \end{aligned}$$

As a special case the expressions of $a_l(t)$ are considered with a different number of corrections. When no correction is performed ($l = 0$), then

$$\begin{aligned} t_1 &= \tau_j, & a_0(t_1) &= z^{(j)}e, \\ t_2 &= \tau_0, & a_0(t_2) &= z^{(0)}e, \\ h(t) &\geq 2, & a_0(t) &= 0. \end{aligned}$$

Here for t_1, $j = 1, 2, \ldots, d$. This notation is also valid for trees $t_3 \sim t_{18}$ in the following discussion.

If one correction is performed ($l = 1$), the expressions for $a_1(t)$ associated with trees t_1 and t_2 are the same as the corresponding $a_0(t)$, namely $a_1(t_i) = a_0(t_i)$ ($i = 1, 2$). For trees with more vertices, then, assuming that the j_i are nonzero,

$$\begin{aligned} t_3 &= [\tau_{j_1}]_{j_2}, & a_1(t_3) &= z^{(j_2)}Z^{(j_1)}e, \\ t_4 &= [\tau_0]_{j_1}, & a_1(t_4) &= z^{(j_1)}Z^{(0)}e, \\ t_5 &= [\tau_{j_1}]_0, & a_1(t_5) &= z^{(0)}Z^{(j_1)}e, \\ t_6 &= [\tau_0]_0, & a_1(t_6) &= z^{(0)}Z^{(0)}e, \\ t_7 &= [\tau_{j_1}, \tau_{j_2}]_{j_3}, & a_1(t_7) &= z^{(j_3)}(Z^{(j_1)}e)(Z^{(j_2)}e), \\ t_8 &= [\tau_{j_1}, \tau_{j_2}]_0, & a_1(t_8) &= z^{(0)}(Z^{(j_1)}e)(Z^{(j_2)}e), \\ t_9 &= [\tau_{j_1}, \tau_0]_{j_2}, & a_1(t_9) &= z^{(j_2)}(Z^{(j_1)}e)(Z^{(0)}e), \\ t_{10} &= [\tau_0, \tau_{j_1}]_{j_2}, & a_1(t_{10}) &= z^{(j_2)}(Z^{(0)}e)(Z^{(j_1)}e), \\ t_{11} &= [\tau_{j_1}, \tau_{j_2}, \tau_{j_3}]_{j_4}, & a_1(t_{11}) &= z^{(j_4)}(Z^{(j_1)}e)(Z^{(j_2)}e)(Z^{(j_3)}e), \\ h(t) &\geq 3, & a_1(t) &= 0. \end{aligned}$$

If two corrections are performed ($l = 2$), the expressions for $a_2(t)$ associated with trees $t_1, \ldots, t_{11}$ are the same as the corresponding $a_1(t)$, namely $a_2(t_i) = a_1(t_i)$ ($i = 1, \ldots, 11$). For trees with more vertices, then

$$\begin{aligned} t_{12} &= [[\tau_{j_1}]_{j_2}]_{j_3}, & a_2(t_{12}) &= z^{(j_3)}Z^{(j_2)}Z^{(j_1)}e, \\ t_{13} &= [[\tau_0]_{j_1}]_{j_2}, & a_2(t_{13}) &= z^{(j_2)}Z^{(j_1)}Z^{(0)}e, \\ t_{14} &= [[\tau_{j_1}]_0]_{j_2}, & a_2(t_{14}) &= z^{(j_2)}Z^{(0)}Z^{(j_1)}e, \\ t_{15} &= [[\tau_{j_1}]_{j_2}]_0, & a_2(t_{15}) &= z^{(0)}Z^{(j_2)}Z^{(j_1)}e, \\ t_{16} &= [[\tau_{j_1}]_{j_2}, \tau_{j_3}]_{j_4}, & a_2(t_{16}) &= z^{(j_4)}(Z^{(j_2)}Z^{(j_1)}e)(Z^{(j_3)}e), \\ t_{17} &= [[\tau_{j_1}, \tau_{j_2}]_{j_3}]_{j_4}, & a_2(t_{17}) &= z^{(j_4)}Z^{(j_3)}((Z^{(j_1)}e)(Z^{(j_2)}e)), \\ h(t) &\geq 4, & a_2(t) &= 0. \end{aligned}$$

When a stochastic RK-type PC method is corrected three times, the expressions for $a_3(t)$ associated with trees t_i ($i = 1, \ldots, 17$) are the same as the corresponding $a_2(t)$, namely

$$a_3(t_i) = a_2(t_i), \qquad i = 1, \ldots, 17.$$

For the analysis of the stochastic RK-type PC methods in this paper, we need only consider additionally the expression $a_3(t)$ for the tree $[[[\tau_{j_1}]_{j_2}]_{j_3}]_{j_4}$, where none of the j_i is zero, given by

$$t_{18} = [[[\tau_{j_1}]_{j_2}]_{j_3}]_{j_4}, \qquad a_3(t_{18}) = z^{(j_4)}Z^{(j_3)}Z^{(j_2)}Z^{(j_1)}e.$$

In the following sections the order conditions associated with trees $t_1, \ldots, t_{18}$ are used to construct stochastic RK-type PC methods.

3. Strong order 1.0 RK methods. The order theory developed in section 2 will apply to the very general class of problems (2.5) with $d > 1$. However, due to spatial constraints and the extreme difficulty in solving the order conditions for the arbitrary d case, we will focus on constructing effective PC methods for $d = 1$.

The s-stage RK methods with one stochastic variable $J_1 \sim N(0, h)$ are given by

$$\begin{aligned} Y =& (e \otimes I)y_n + h(A \otimes I)g_0(Y) + J_1(B \otimes I)g_1(Y), \\ y_{n+1} =& y_n + h(\alpha^\top \otimes I)g_0(Y) + J_1(\beta^\top \otimes I)g_1(Y), \end{aligned} \tag{3.1}$$

where A and B are $s \times s$ matrices, while α and β are s-dimensional vectors. According to the theorems given by Rümelin [22] and Burrage, Burrage, and Belward [9], the maximum strong global order of these stochastic RK methods is 1.0.

For the trivial predictor, the stochastic RK-type PC method using (3.1) as the corrector is given by

$$\begin{aligned} Y^{(0)} &= (e \otimes I)y_n, \\ Y^{(i)} &= (e \otimes I)y_n + h(A \otimes I)g_0(Y^{(i-1)}) + J_1(B \otimes I)g_1(Y^{(i-1)}), \quad i = 1, \cdots, l, \\ y_{n+1} &= y_n + h(\alpha^\top \otimes I)g_0(Y^{(l)}) + J_1(\beta^\top \otimes I)g_1(Y^{(l)}). \end{aligned} \tag{3.2}$$

Now consider the order conditions of the RK-type PC method (3.2). If no correction is performed, the local truncation error of this method is given by

$$l_{10} = h(1 - \alpha^\top e)F(\tau_0)(y(t_n)) + J_1(1 - \beta^\top e)F(\tau_1)(y(t_n)) + \sum_{\rho(t)\geq 2} J(t)F(t)(y(t_n)).$$

Assuming that

$$\alpha^\top e = 1, \qquad \beta^\top e = 1, \tag{3.3}$$

this method will have strong local order 0.5, namely $E(l_{10}^2) = O(h^2)$. In this case the PC method (3.2) is equivalent in strong order to the Euler–Maruyama method, given by

$$y_{n+1} = y_n + hg_0(y_n) + J_1 g_1(y_n).$$

It is well known that the numerical solution of the Euler–Maruyama method converges to the exact solution of the corresponding Itô SDE. Thus the numerical solution of method (3.2) without any correction may not converge to the exact solution of the Stratonovich SDE (2.5).

If one correction is performed ($l = 1$), method (3.2) will have strong local order 1.0 if, in addition to (3.3),

$$e(t_3) = J(t_3) - a(t_3) = \left(\frac{1}{2} - \beta^\top Be\right) J_1^2 = 0,$$

which is equivalent to

$$\beta^\top Be = \frac{1}{2}. \tag{3.4}$$

At the same time this method will have mean local error 1.0 as

$$E(e(t_i)) = 0, \qquad i = 4, 5, 7, 12,$$

where trees t_4, t_5, t_7, and t_{12} are those associated with terms corresponding to $h^{1.5}$. Thus the stochastic RK-type PC method (3.2) will have strong global order 1.0 if one correction is applied and the order conditions (3.3) and (3.4) are satisfied at the same time.

The order conditions (3.3) and (3.4) of the stochastic RK-type PC method with strong global order 1.0 are the same as those of the stochastic RK methods (3.1) with strong global order 1.0, given in [12]. Thus the strong global order of the RK-type PC method (3.2) is 1.0 if the strong global order of the original stochastic RK method (3.1) is 1.0 and one correction is applied.

Now we construct a two-stage implicit RK method. As there are only three order conditions in (3.3) and (3.4) and 12 coefficients in this method, additional conditions can be considered. For example, we can consider the stochastic order conditions on which the terms corresponding to $h^{1.5}$ have minimum coefficients, namely the stochastic order conditions for minimum principal error coefficients. The principal error coefficients are minimized if [12]

$$\alpha^\top Be = \frac{1}{2}, \quad \beta^\top Ae = \frac{1}{2}, \quad \beta^\top (Be)^2 = \frac{1}{3}, \quad \beta^\top B(Be) = \frac{1}{6}.$$

These four conditions are called the minimum principal error conditions.

Combining the order conditions (3.3) and (3.4) and the minimum principal error conditions together and assuming that $A = B$ and $\alpha = \beta$, we have the following two-stage implicit RK corrector method with strong global order 1.0, called IRK2, given by

$$\begin{array}{c|cccc} \frac{1}{3} & \frac{1}{4} & \frac{1-\sqrt{3}}{6} & \frac{1}{3} & \frac{1-\sqrt{3}}{6} \\ \frac{1+\sqrt{3}}{6} & \frac{1}{3} & \frac{1}{4} & \frac{1+\sqrt{3}}{6} & \frac{1}{3} \\ \hline & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{array} . \tag{3.5}$$

4. Strong order 1.5 RK methods. The second special class of the stochastic RK methods (2.7) that will be discussed is those with two stochastic variables J_1 and J_{10}/h, for solving problems of the form (2.5) with $d = 1$, given by

$$\begin{aligned} Y = \quad & (e \otimes I) y_n + h(A \otimes I) g_0(Y) + \left(J_1 (B_1 \otimes I) + \frac{J_{10}}{h} (B_2 \otimes I) \right) g_1(Y), \\ y_{n+1} = \quad & y_n + h(\alpha^\top \otimes I) g_0(Y) + \left(J_1 (\beta_1^\top \otimes I) + \frac{J_{10}}{h} (\beta_2^\top \otimes I) \right) g_1(Y), \end{aligned} \tag{4.1}$$

where A, B_1, and B_2 are $s \times s$ matrices and α, β_1, and β_2 are s-dimensional vectors. Here we remind readers that on the interval $[t_n, t_{n+1}]$, J_1 and J_{10}/h are closely related. In particular, if u and v are two independent $N(0, 1)$ random variables, then

$$J_1 = u\sqrt{h}, \qquad \frac{J_{10}}{h} = \frac{\sqrt{h}}{2} \left(u + \frac{v}{\sqrt{3}} \right).$$

Burrage and Burrage [7], [8] and Burrage [12] first present this class of stochastic RK methods and study the order conditions of these methods based on the colored

rooted tree theory and stochastic B-series. A five-stage explicit stochastic RK method with strong global order 1.5 is presented in [8].

For the trivial predictor, the stochastic RK-type PC method using (4.1) as a corrector is given by

(4.2)
$$\begin{aligned} Y^{(0)} =& (e \otimes I) y_n, \\ Y^{(i)} =& (e \otimes I) y_n + h(A \otimes I) g_0(Y^{(i-1)}) + \left(J_1 (B_1 \otimes I) + \frac{J_{10}}{h} (B_2 \otimes I) \right) g_1(Y^{(i-1)}), \\ & i = 1, \dots, l, \\ y_{n+1} =& y_n + h(\alpha^\top \otimes I) g_0(Y^{(l)}) + \left(J_1 (\beta_1^\top \otimes I) + \frac{J_{10}}{h} (\beta_2^\top \otimes I) \right) g_1(Y^{(l)}). \end{aligned}$$

Now consider the order conditions for this RK-type PC method. When one correction is performed, the order conditions necessary for strong order 1.5 associated with trees $t_1, \dots, t_{11}$ are given by

$$E\left(e^2(t_i)\right) = 0, \qquad i = 1, \dots, 11.$$

Let $c = Ae$, $b = B_1 e$, and $d = B_2 e$; the above order conditions are equivalent to (see [7] and [12])

(4.3)
$$\begin{aligned} \alpha^\top (e, c, d, b) &= \left(1, \frac{1}{2}, 1, 0 \right), \\ \beta_1^\top (e, c, b, d, b^2, d^2) &= \left(1, 1, \frac{1}{2}, -\beta_2^\top b, \frac{1}{3}, -2\beta_2^\top bd \right), \\ \beta_2^\top (e, c, d, b^2, d^2) &= (0, -1, 0, -2\beta_1^\top bd, 0). \end{aligned}$$

When method (4.2) is corrected twice, the order condition associated with tree t_{12} is $E(e^2(t_{12})) = 0$, which is equivalent to (see [7] and [12])

(4.4)
$$\begin{aligned} &\beta_1^\top B_1 b = \tfrac{1}{6}, \quad \beta_2^\top B_1 b + \beta_1^\top (B_2 b + B_1 d) = 0, \\ &\beta_2^\top B_2 d = 0, \quad \beta_1^\top B_2 d + \beta_2^\top (B_2 b + B_1 d) = 0. \end{aligned}$$

In order to get mean local order 1.5, it is necessary that the following mean order conditions should be satisfied:

$$E\left(e(t_i)\right) = 0, \qquad i = 8, 9, 11, 13, 14, 15, 16, 17,$$

which are equivalent to (see [8])

$$\begin{aligned} 0 &= \alpha^\top B_1 b + \frac{1}{2} \alpha^\top (B_1 d + B_2 b) + \frac{1}{3} \alpha^\top B_2 d, \\ 0 &= \alpha^\top \left(b^2 + bd + \frac{1}{3} d^2 \right), \\ 0 &= \beta_1^\top Ab + \frac{1}{2} \left(\beta_1^\top Ad + \beta_2^\top Ab \right) + \frac{1}{3} \beta_2^\top Ad, \\ 0 &= \beta_1^\top \left(cb + \frac{1}{2} cd \right) + \beta_2^\top \left(\frac{1}{2} cb + \frac{1}{3} cd \right), \end{aligned}$$

$$
\begin{aligned}
0 &= \beta_1^\top \left(B_1 c + \frac{1}{2} B_2 c \right) + \beta_2^\top \left(\frac{1}{2} B_1 c + \frac{1}{3} B_2 c \right), \\
\frac{3}{8} &= \beta_1^\top \mathrm{Diag}(b) \left(3B_1 b + \frac{3}{2} (B_1 d + B_2 b) + \frac{5}{6} B_2 d \right) \\
&\quad + \beta_2^\top \mathrm{Diag}(d) \left(\frac{5}{6} B_1 b + \frac{1}{2} (B_1 d + B_2 b) + \frac{1}{3} B_2 d \right) \\
&\quad + \left(\beta_1^\top \mathrm{Diag}(d) + \beta_2^\top \mathrm{Diag}(b) \right) \left(\frac{3}{2} B_1 b + \frac{5}{6} (B_1 d + B_2 b) + \frac{1}{2} B_2 d \right), \\
\frac{1}{4} &= \beta_1^\top \left(B_1 \left(3b^2 + 3bd + \frac{5}{6} d^2 \right) + B_2 \left(\frac{3}{2} b^2 + \frac{5}{3} bd + \frac{1}{2} d^2 \right) \right) \\
&\quad + \beta_2^\top \left(B_1 \left(\frac{3}{2} b^2 + \frac{5}{3} bd + \frac{1}{2} d^2 \right) + B_2 \left(\frac{5}{6} b^2 + bd + \frac{1}{3} d^2 \right) \right), \\
\frac{3}{4} &= \beta_1^\top \left(3b^3 + \frac{9}{2} b^2 d + \frac{5}{2} bd^2 + \frac{1}{2} d^3 \right) + \beta_2^\top \left(\frac{3}{2} b^3 + \frac{5}{2} b^2 d + \frac{3}{2} bd^2 + \frac{1}{3} d^3 \right).
\end{aligned} \tag{4.5}
$$

It should be noticed that, for expectation in the mean, trees t_9 and t_{10} are equivalent.

However, when two corrections are performed and all of the order conditions (4.3)∼(4.5) are satisfied, the strong local order of the RK-type PC method (4.2) is 1.5, but the mean local order of this method is still 1.0 as the mean order condition associated with tree t_{18} is not satisfied, since the height of t_{18} is 4 and so $a_2(t_{18}) = 0$. In order to get a RK-type PC method with strong global order 1.5, a third correction is needed. When a third correction is performed, the mean order condition associated with tree t_{18} is given by $E\left(e(t_{18})\right) = 0$, which is equivalent to [8]

$$
\begin{aligned}
\frac{1}{8} =& \beta_1^\top \left(B_1^2 \left(3b + \frac{3}{2} d \right) + B_2^2 \left(\frac{5}{6} b + \frac{1}{2} d \right) + (B_1 B_2 + B_2 B_1) \left(\frac{3}{2} b + \frac{5}{6} d \right) \right) \\
&+ \beta_2^\top \left(B_1^2 \left(\frac{3}{2} b + \frac{5}{6} d \right) + B_2^2 \left(\frac{1}{2} b + \frac{1}{3} d \right) + (B_1 B_2 + B_2 B_1) \left(\frac{5}{6} b + \frac{1}{2} d \right) \right).
\end{aligned} \tag{4.6}
$$

The order conditions (4.3)∼(4.6) of the stochastic RK-type PC method with strong global order 1.5 are the same as those of the stochastic RK method (4.1) with strong global order 1.5, given in [8]. Thus the strong global order of the RK-type PC method (4.2) with three corrections is 1.5 if the strong global order of the original RK method (4.1) is 1.5.

Now an implicit four-stage RK method with strong global order 1.5 is constructed. In order to have small error coefficients for the deterministic terms, the following additional order conditions are considered here, given by

$$
\alpha^\top A c = \frac{1}{6}, \quad \alpha^\top A c^2 = \frac{1}{12}, \quad \alpha^\top \mathrm{Diag}(c) A c = \frac{1}{8}, \quad \alpha^\top A^2 c = \frac{1}{24}. \tag{4.7}
$$

Using Maple to solve all of the order conditions (4.3)∼(4.7), we have the following strong global order 1.5 RK corrector method, which is called IRK4, with matrices A, B_1, and B_2:

$$
A = \begin{pmatrix}
1.00436335789 & -0.56006282797 & -0.41253045082 & -0.03177007950 \\
-0.04300768840 & -0.12902306500 & -0.04300768833 & -0.04300768833 \\
2.26132980150 & -2.30000000000 & 0.11987418760 & -0.33925011871 \\
3.51937593150 & -2.30000000000 & 0.11987418760 & -0.33925011871
\end{pmatrix},
$$

$$B^{(1)} = \begin{pmatrix} 0.10566243265 & 0.03522081088 & 0.03522081088 & 0.03522081088 \\ 0.19716878372 & 0.19716878372 & 0.19716878372 & 0.19716878372 \\ -0.19879713087 & 0.53213046420 & 0.50000000000 & 0.16666666667 \\ 0.16666666667 & 0.16666666667 & 0.16666666667 & 0.50000000000 \end{pmatrix},$$

$$B^{(2)} = \begin{pmatrix} 6.2322500476 & -3.7052829175 & -3.7979991039 & 1.2710319739 \\ -4.6564739362 & 2.8209539211 & 3.0265200151 & -1.1910000000 \\ -8.0122402257 & 3.2818190506 & 3.4496446680 & -1 \\ -9.7304211738 & 5 & 5 & -2.5503553321 \end{pmatrix},$$

and weight vectors $\alpha^\top$, $\gamma^{(1)\top}$, and $\gamma^{(2)\top}$:

$$\begin{array}{rcl} \alpha^\top & = & (1.205542599,\ 0.2329045687,\ -0.7937286771,\ 0.3552815092), \\ \gamma^{(1)\top} & = & (\ 0.5,\ 0.5,\ -0.8974417060,\ 0.8974417060), \\ \gamma^{(2)\top} & = & (\ 0,\ 0,\ 0.7948834118,\ -0.7948834118). \end{array}$$

Remark. This method requires only four parallel stages and three sequential stages (cf. the strong order 1.5 explicit stochastic RK method G5 of [8] which requires five sequential stages) and so is implemented efficiently on a four processor computer.

5. Stability properties of RK-type PC methods. In this paper the following linear test equation of Stratonovich type, given by

$$dy = aydt + by \circ dW(t), \qquad y(0) = y_0, \tag{5.1}$$

is used to discuss the stability properties of stochastic RK-type PC methods.

Applying a one-step numerical scheme to (5.1), this numerical scheme is represented by

$$y_{n+1} = R(h, a, b)y_n.$$

Saito and Mitsui [24] introduced the definition of mean-square (MS) stability.

DEFINITION 5.1. *A numerical scheme is said to be MS-stable for h, a, and b if*

$$\overline{R}(h, a, b) = E(|R(h, a, b)|^2) < 1.$$

$\overline{R}(h, a, b)$ is called the MS-stability function of the numerical scheme.

Another important stability definition is that of asymptotic stability. Saito and Mitsui [23] introduced the definition of T-stability to measure asymptotic stability and give two examples on the T-stability properties of numerical methods for weak solutions. Burrage and Tian [11] present a method to measure the T-stability for strong solutions and give the definition of T(A)-stability. Here we just consider the MS-stability properties of the stochastic RK-type PC methods presented in this paper.

Applying the stochastic RK-type PC methods (2.8) to (5.1) gives

$$\begin{aligned} Y^{(0)} &= ey_n, \\ Y^{(i)} &= ey_n + aZ^{(0)}Y^{(i-1)} + bZ^{(1)}Y^{(i-1)} \\ &= \left[I + \overline{Z} + \overline{Z}^2 + \cdots + \overline{Z}^i\right] ey_n, \quad i = 1, 2, \ldots, l, \\ y_{n+1} &= y_n + az^{(0)}Y^{(l)} + bz^{(1)}Y^{(l)} \\ &= \left(1 + \overline{z}\left[I + \overline{Z} + \overline{Z}^2 + \cdots + \overline{Z}^l\right] e\right) y_n, \end{aligned}$$

where

$$\overline{z} = az^{(0)} + bz^{(1)}, \qquad \overline{Z} = aZ^{(0)} + bZ^{(1)}.$$

Let

$$R^{(l)} = 1 + \overline{z}\left[I + \overline{Z} + \overline{Z}^2 + \cdots + \overline{Z}^l\right] e; \tag{5.2}$$

then the stochastic RK-type PC methods (2.8) are MS-stable for h, a, and b if

$$\overline{R}^{(l)} = E\left(\left|R^{(l)}\right|^2\right) < 1.$$

Now we consider the MS-stability properties of the stochastic RK-type PC methods (3.2) with strong global order 1.0. Applying these methods to (5.1) gives

$$R_1^{(l)}(p, q, \overline{J}_1) = 1 + (p\alpha^\top + q\overline{J}_1\beta^\top)\left(\sum_{i=0}^{l}(pA + q\overline{J}_1 B)^i\right) e,$$

where $p = ah$, $q = b\sqrt{h}$, and $\overline{J}_1 = \frac{J_1}{\sqrt{h}} \sim N(0,1)$. For the stochastic RK-type PC method based on the two-stage implicit RK corrector IRK2 (3.5), the expressions for $R_1^{(l)}$ are given by

$$\begin{aligned}
R_1^{(1)} &= 1 + p + q\overline{J}_1 + \frac{1}{2}(p + q\overline{J}_1)^2,\\
R_1^{(2)} &= 1 + p + q\overline{J}_1 + \frac{1}{2}(p + q\overline{J}_1)^2 + \frac{1}{6}(p + q\overline{J}_1)^3,\\
R_1^{(3)} &= 1 + p + q\overline{J}_1 + \frac{1}{2}(p + q\overline{J}_1)^2 + \frac{1}{6}(p + q\overline{J}_1)^3 + \frac{1}{36}(p + q\overline{J}_1)^4,
\end{aligned}$$

and the MS-stability functions are given by

$$\begin{aligned}
\overline{R}_1^{(1)} &= 1 + 2p + 2p^2 + p^3 + \frac{1}{4}p^4 + 2q^2 + 3pq^2 + \frac{3}{2}p^2q^2 + \frac{3}{4}q^4,\\
\overline{R}_1^{(2)} &= 1 + 2p + 2p^2 + \frac{4}{3}p^3 + \frac{7}{12}p^4 + \frac{1}{6}p^5 + \frac{1}{36}p^6 + 2q^2 + 4pq^2 + \frac{7}{2}p^2q^2\\
&\quad + \frac{5}{3}p^3q^2 + \frac{5}{12}p^4q^2 + \frac{7}{4}q^4 + \frac{5}{2}pq^4 + \frac{5}{4}p^2q^4 + \frac{5}{12}q^6,\\
\overline{R}_1^{(3)} &= 1 + 2p + 2p^2 + \frac{4}{3}p^3 + \frac{23}{36}p^4 + \frac{2}{9}p^5 + \frac{1}{18}p^6 + \frac{1}{108}p^7 + \frac{1}{1296}p^8\\
&\quad + 2q^2 + 4pq^2 + \frac{23}{6}p^2q^2 + \frac{20}{9}p^3q^2 + \frac{5}{6}p^4q^2 + \frac{7}{36}p^5q^2 + \frac{7}{324}p^6q^2\\
&\quad + \frac{23}{12}q^4 + \frac{10}{3}pq^4 + \frac{5}{2}p^2q^4 + \frac{35}{36}p^3q^4 + \frac{35}{216}p^4q^4\\
&\quad + \frac{5}{6}q^6 + \frac{35}{36}pq^6 + \frac{35}{108}p^2q^6 + \frac{35}{432}q^8.
\end{aligned}$$

Here denote $\overline{R}_1^{(0)}$ as the MS-stability function of this method without any correction, namely the explicit Euler method, given by

$$\overline{R}_1^{(0)} = (1 + p)^2 + q^2.$$

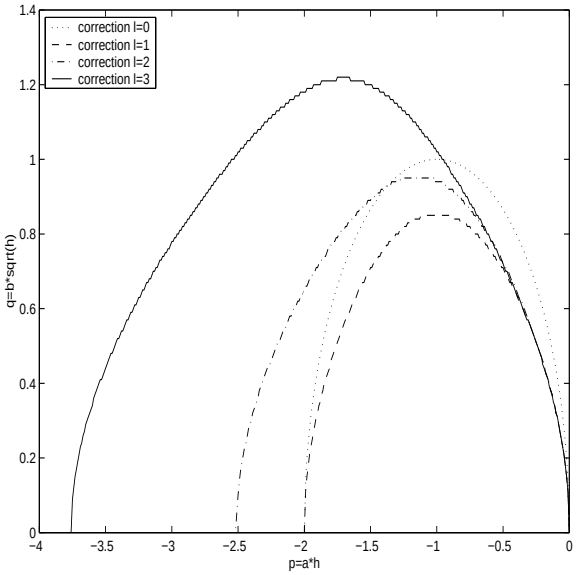


FIG. 1. *MS-stable regions of the two-stage stochastic RK-type PC method.*

Figure 1 gives the MS-stability regions of the stochastic RK-type PC method (3.2) based on IRK2 (3.5). The MS-stable regions are the areas under the plotted lines and are symmetric about the p-axis. The MS-stability properties of this method with two corrections are better than those with one correction. The MS-stability properties of this method are much improved when the third correction is performed.

6. Numerical results. Numerical results for solving SDEs driven by one Wiener process are reported in this section. Denoting $y_N^{(i)}$ as the numerical approximation to $y^{(i)}(t_N)$ at step point t_N in the ith simulation of all K simulations, we use means of MS errors MS, strong order 1 rate R_1 and strong order 1.5 rate $R_{1.5}$, defined by

$$MS = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(y_N^{(i)} - y^{(i)}(t_N))^2}, \qquad R_1 = \frac{MS}{h}, \qquad R_{1.5} = \frac{MS}{h\sqrt{h}},$$

to measure the accuracy and the convergence properties of the stochastic RK-type PC methods. All of the data in this section are based on 1000 simulated trajectories.

The first test equation is a nonlinear problem, whose Stratonovich form is

$$dy = -\alpha(1-y^2)dt + \beta(1-y^2)\circ dW(t), \quad y(0) = 0.5, \quad t \in [0,1],$$

with $\alpha = -1$ and $\beta = 1$. The exact solution of this equation is [17]

$$y(t) = \frac{(1+y_0)\exp(-2\alpha t + 2\beta W(t)) + y_0 - 1}{(1+y_0)\exp(-2\alpha t + 2\beta W(t)) - y_0 + 1}.$$

Figure 2 gives the MS errors of the two stochastic RK-type PC methods based on IRK2 and IRK4, respectively, for solving the first test equation. For the two-stage PC method based on IRK2, the implicit corrector (3.5) is applied with a different number of corrections $l = 0, 1, 2, 3, 4$. From the left figure in Figure 2, the numerical

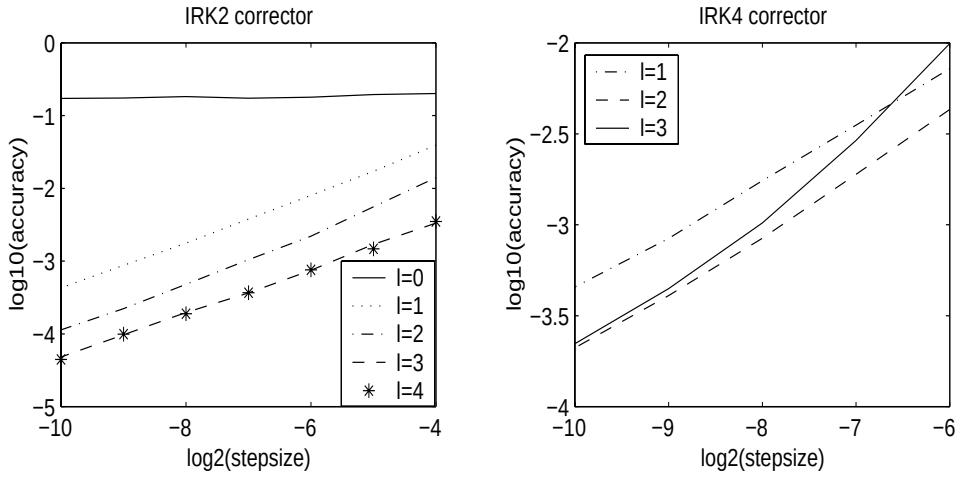


FIG. 2. *MS errors for solving the first test equation.*

solution when no correction is performed, denoted as $l = 0$, does not converge to the exact solution of the corresponding Stratonovich SDE. The strong convergence rates of this method with $l = 1$, 2, 3, or 4 are all equal to 1.0, as predicted by our theory. The averaged errors are smaller if more corrections are performed. The difference between the averaged errors of this method with three corrections and those with four corrections is small.

For the four-stage PC method based on IRK4, the implicit corrector is applied with a different number of corrections $l = 1, 2, 3$. From the right figure of Figure 2, the strong convergence rates of this method with $l = 1$ is equal to 1.0. When two corrections are performed, the strong convergence rate is between 1.0 and 1.5. The strong convergence rate of this method is 1.5 if three corrections are performed, which is again consistent with our theory.

It should be noticed that the accuracy of the stochastic RK-type PC method based on IRK4 with strong order 1.5 is not as good as that of the method based on IRK2 with strong order 1.0 when $2^{-10} \le h \le 2^{-6}$. The reason for this phenomenon is due to the eigenvalues of the method matrices. For IRK2, the eigenvalues of matrices A and B are

$$\lambda(A) = \lambda(B) = \frac{1}{3} \pm \frac{\sqrt{2}}{6} i,$$

while for IRK4 the eigenvalues of the method matrices are

$$\begin{aligned} \lambda(A) &= 0.400 \pm 0.622i, \ \ -0.072 \pm 0.253i, \\ \lambda(B_1) &= 0.096, \ 0.333, \ 0.878, \ -0.0053, \\ \lambda(B_2) &= 11.528, \ -0.335, \ -0.620 \pm 0.035i. \end{aligned}$$

The large eigenvalue of matrix B_2 causes amplifications in the errors of the PC method based on IRK4. This effect was well known in the deterministic case; see the work of Sommeijer [25].

In order to test out this supposition, we construct two methods, MIRK2 and MIRK4, which have strong order 1 and 1.5, respectively, and whose defining matrices

have smaller spectral radius. The MIRK2 method is given by

$$\begin{array}{c|cccc} \frac{1067}{3000} & \frac{933}{3000} & \frac{1067}{3000} & \frac{933}{3000} \\ -\frac{67}{1000} & \frac{67}{1000} & -\frac{67}{1000} & \frac{67}{1000} \\ \hline \frac{3}{4} & \frac{1}{4} & \frac{3}{4} & \frac{1}{4} \end{array} . \tag{6.1}$$

The eigenvalues of the defined method matrices are

$$\lambda(A) = \lambda(B) = \frac{317 \pm \sqrt{11}}{1500}.$$

The MIRK4 method is different from the IRK4 method just in the method matrices $B_M^{(1)}$ and $B_M^{(2)}$, given by

$$B_M^{(1)} = \begin{pmatrix} -0.4103843710 & 0.2113248635 & 0.2566537645 & 0.1537306083 \\ 1.1990595100 & 0 & -0.300000000 & -0.1103843746 \\ 0.2807539857 & 0.4849084469 & 0.3943375673 & -0.16000000 \\ 0.2807539819 & 0.4849084506 & 0.3943375673 & -0.160000000 \end{pmatrix},$$

$$B_M^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.7602588353 & -0.7602588353 & -0.7602588353 & 0 \\ -0.7602588353 & -0.7602588353 & -0.7602588353 & 0 \end{pmatrix},$$

whose eigenvalues are

$$\lambda(B_M^{(1)}) = -0.36 \pm 0.038i, 0.55, 0, \qquad \lambda(B_M^{(2)}) = -0.76, 0, 0, 0.$$

Using the PC methods based on the correctors MIRK2 and MIRK4, we then repeated the calculations for solving the first test equation, and the numerical results are given in Figure 3. It is clear that the MIRK2 method is more effective than the IRK2 method, whose computational results are given by Figure 2. For four-stage correctors, the accuracy of the numerical results of MIRK4 is better than that of IRK4 with stepsize $h = 2^{-6}, 2^{-7}, 2^{-8}$. When $h = 2^{-9}$ and 2^{-10}, the accuracy of MIRK4 is just slightly better than that of IRK4.

The second test equation is also a nonlinear SDE, given by

$$dy = a(1 + y^2) \circ dt, \qquad y(0) = 1, \qquad t \in [0, 1],$$

with $a = 0.1$. The exact solution is given in [17], namely

$$y = \tan(aW(t) + \arctan y_0).$$

Figure 4 gives the MS errors of the four PC methods for the second test equation. In this case the implicit corrector is applied with a different number of corrections $l = 2, 3$. It is clear that the MIRK2 and MIRK4 methods with three corrections are much more effective than the IRK2 and IRK4 methods for the second test equation.

In order to discuss the relationship between the accuracy of the numerical methods and the computational cost, we use the following explicit two-stage RK methods to solve the first test equation:

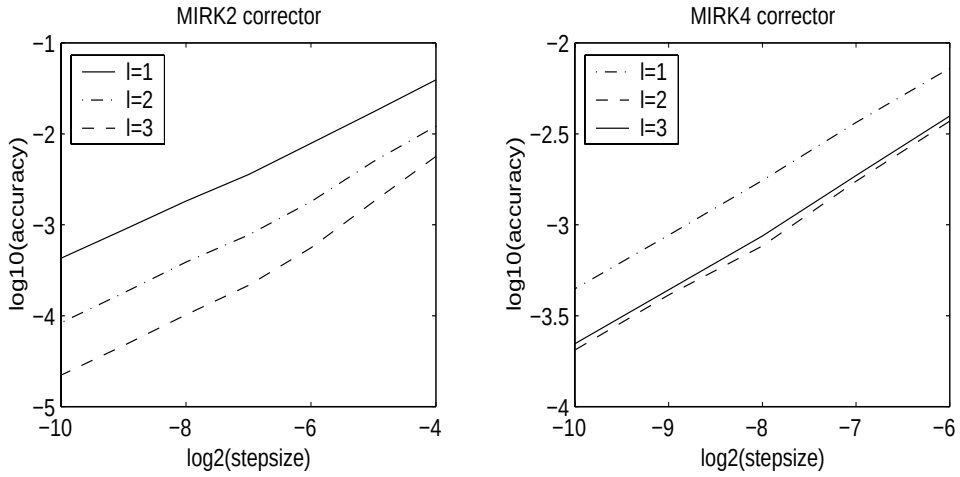


FIG. 3. *MS errors for solving the first test equation.*

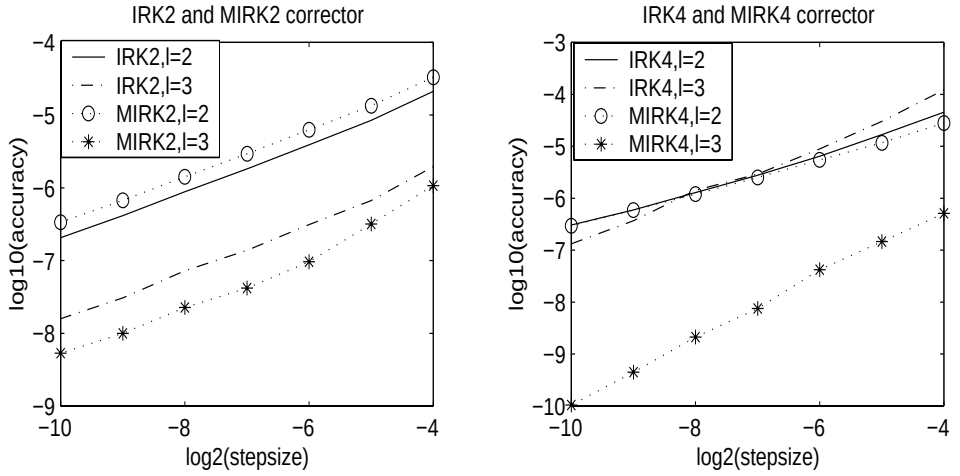


FIG. 4. *MS errors for solving the second test equation.*

(1) The Heun method [12].

$$
\begin{aligned}
Y &= y_n + hf(y_n) + \Delta W_n g(y_n),\\
y_{n+1} &= y_n + \frac{1}{2}h\left(f(y_n) + f(Y)\right) + \frac{1}{2}\Delta W_n\left(g(y_n) + g(Y)\right).
\end{aligned} \tag{6.2}
$$

(2) The Burrages scheme [12].

$$
\begin{aligned}
Y &= y_n + \frac{2}{3}hf(y_n) + \frac{2}{3}\Delta W_n g(y_n),\\
y_{n+1} &= y_n + h\left(\frac{1}{4}f(y_n) + \frac{3}{4}f(Y)\right) + \Delta W_n\left(\frac{1}{4}g(y_n) + \frac{3}{4}g(Y)\right).
\end{aligned} \tag{6.3}
$$

TABLE 1
Accuracy and computational cost of some stochastic RK methods.

	Heun	Burrages	Method 1	IRK2 ($l=3$)	MIRK2 ($l=3$)
Accuracy	4.9E-3	3.5E-3	4.7E-3	3.3E-3	5.6E-3
cost-flops	7.2E+6	7.9E+6	4.1E+6	3.0E+6	3.0E+6
Accuracy	5.9E-4	4.3E-4	5.2E-4	3.6E-4	2.2E-4
cost-flops	5.7E+7	6.4E+7	3.3E+7	2.4E+7	2.4E+7

(3) Method 1 in [26].

$$
\begin{aligned}
Y &= y_n + \frac{3}{10}hf(y_n) + \frac{58}{100}\Delta W_n g(y_n),\\
y_{n+1} &= y_n + h\left(\frac{56}{100}f(y_n) + \frac{44}{100}f(Y)\right) + \Delta W_n\left(\frac{4}{29}g(y_n) + \frac{25}{29}g(Y)\right).
\end{aligned}
\tag{6.4}
$$

Table 1 gives the accuracy and the computional cost, in terms of flops obtained by Matlab, of these explicit RK methods and those of the IRK2 and MIRK2 methods with three corrections. Clearly, both the IRK2 and MIRK2 methods with three corrections can achieve better accuracy than the other explicit methods with substantially reduced computation costs.

The third test equation is given by

$$
\begin{aligned}
dy_1 &= y_2 dt + \theta y_2 \circ dW(t),\\
dy_2 &= \mu\left((1-y_1^2)y_2 - y_1\right) + \theta\left((1-y_1^2)y_2 - y_1\right)\circ dW(t).
\end{aligned}
\tag{6.5}
$$

This equation is the ordinary Van der Pol equation [14] when $\theta = 0$. The Van der Pol equation is stiff when μ is large.

We use IRK2 with $l = 3$ to solve this equation. In Figure 5 we give four simulations of this equation. The top two simulations in Figure 5 are obtained with parameters $\mu = 1$, $\theta = 0.1$, and $\theta = 1$ and stepsize $h = 0.01$. The bottom two simulations are obtained with parameters $\mu = 10$, $\theta = 0.1$, and $\theta = 0.5$ and stepsize $h = 0.001$. The numerical simulations with $\theta = 0.1$ are similar to those of the deterministic Van der Pol equation with the same μ.

In order to discuss the efficiency of the two-stage PC methods, we use IRK2 ($l = 3$) with stepsize $h = 0.0001$ to get a numerical solution which is regarded as the "accurate" solution in the case of $\mu = 1$ and different θ. We compare this "accurate" solution with the numerical simulations obtained by the explicit RK methods (6.2), (6.3), and (6.4) and those obtained by IRK2 ($l = 3$) and MIRK2 ($l = 3$). Numerical results presented in Figure 6 are based on 100 simulations. The left figure of Figure 6 gives the accuracy of numerical solutions with $\mu = 1$ and $0.1 \le \theta \le 1.0$. The accuracy of numerical simulations of IRK2 ($l = 3$) and MIRK2 ($l = 3$) is considerably better than those of the other methods. In the right figure of Figure 6, we present the proportions of "acceptable" solutions with the standard that the averaged error is less than 1.0. It should be noticed that the proportions are dependent on the standard. We can get more "acceptable" simulations by IRK2 ($l = 3$) and MIRK2 ($l = 3$) than with the other explicit RK methods. The explicit RK methods are not suitable for solving this equation with values for $\theta > 1$.

Similar numerical results about the accuracy and the proportions of acceptable solutions can be also obtained for the case $\mu = 10$ and $\theta \in [0.1, 1.0]$. In this case a smaller stepsize, for example $h = 0.00001$, should be used for the "accurate solution."

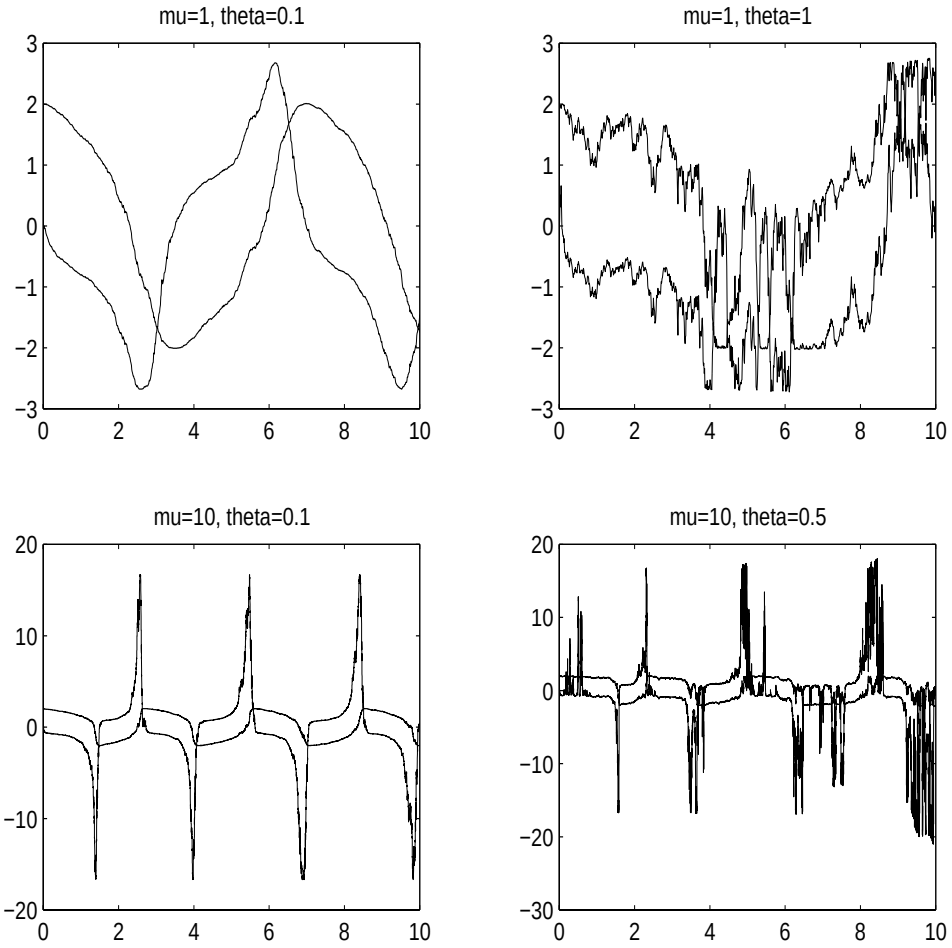


FIG. 5. *Numerical simulations of the stochastic Van der Pol equation with IRK*2 ($l = 3$).

7. Conclusions. In this paper we have constructed PC methods based on the trivial predictor and stochastic implicit RK correctors for solving SDEs. Using the colored rooted tree theory and stochastic B-series, we present an order condition theorem for constructing stochastic RK-type PC methods. We also present detailed order conditions of the stochastic RK-type PC methods with strong convergence order 1.0 and 1.5. Two two-stage implicit RK methods with strong global order 1.0 and two four-stage implicit RK methods with strong global order 1.5 are constructed in this paper. The following conclusions can be made from the stability analysis and numerical behavior of the RK-type PC methods presented in this paper.

(1) As the number of parameters is larger than the number of order conditions, additional conditions can be used to determine the coefficients of stochastic RK methods in order to get better stability properties and numerical behavior. For example, we may consider a two-stage implicit RK method which has good stability properties at infinity. Applying this method (3.1) to the linear test equation (5.1) gives

$$y_{n+1} = R(p, q)y_n,$$

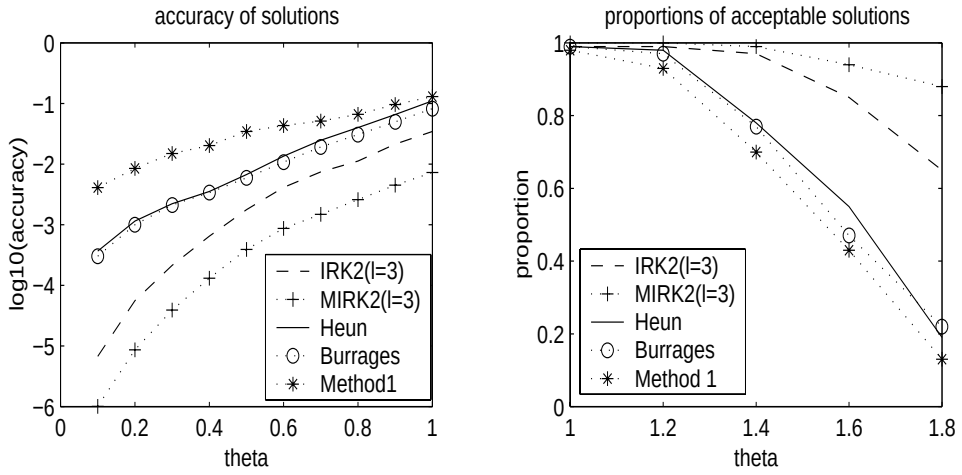


FIG. 6. *Proportions and MS accuracy of stable solutions for solving the Van der Pol problem with* $\mu = 1$ *and* θ *varying.*

where

$$R(p,q) = 1 + (p\alpha^\top + q\overline{J}_1\beta^\top)(I - pA - q\overline{J}_1 B)^{-1}e.$$

This method will have damping stability properties at infinity if

$$\alpha^\top A^{-1}e = 1, \qquad \beta^\top B^{-1}e = 1.$$

The implicit two-stage RK method (3.5) satisfies these conditions.

(2) Another possible way to improve the stability properties and the numerical behavior of stochastic RK-type PC methods is to reduce the magnitude of the eigenvalues of the matrices in the stochastic RK methods. The cue is in the expression of $R^{(l)}$ (5.2). In order to verify this supposition, we construct two methods, MIRK2 and MIRK4. Compared with IRK2 and IRK4, the eigenvalues of the method matrices in MIRK2 and MIRK4 are small in magnitude. Numerical results of the MIRK2 and MIRK4 methods are more accurate than those of IRK2 and IRK4. The effect has also been observed in the deterministic case.

(3) The stochastic RK-type PC methods are more effective than other explicit stochastic RK methods. For two-stage RK methods with strong order 1.0, the superiority of the PC method based on IRK2 or MIRK2 is due to the better stability properties (shown in Figures 1 and 6), the better accuracy, and the less computational cost (shown in Table 1 and Figure 6). For the RK methods with strong order 1.5, the PC method will be more effective than the explicit RK methods with the same order if it is implemented on a parallel computer.

Thus we may consider stochastic RK-type PC methods which have better stability properties and numerical behavior by adding additional conditions or by reducing the magnitude of the eigenvalues of the matrices in the stochastic RK methods. In addition, we can apply splitting techniques [20] to implicit RK methods to construct numerical schemes which are suitable for solving stiff SDEs. Finally, we note that these concepts can be applied to SDE problems driven by more then one Wiener process. However, spatial constraints for this work means that all of these are topics for future work.

REFERENCES

[1] L. Brugnano, K. Burrage, and P. Burrage, *Adams type methods for the numerical solution of stochastic differential equations*, BIT, 40 (2000), pp. 451–470.

[2] K. Burrage, *The error behaviour of a general class of predictor-corrector methods*, Appl. Numer. Math., 8 (1991), pp. 201–216.

[3] K. Burrage, *Efficient block predictor-corrector methods with a small number of corrections*, J. Comput. Appl. Math., 45 (1993), pp. 139–150.

[4] K. Burrage, *Parallel methods for initial value problems*, Appl. Numer. Math., 11 (1993), pp. 5–25.

[5] K. Burrage, *The search for the Holy-Grail or predictor-corrector methods for solving ODEIVPs*, Appl. Numer. Math., 11 (1993), pp. 125–141.

[6] K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*, Oxford University Press, Oxford, UK, 1995.

[7] K. Burrage and P. M. Burrage, *High strong order explicit Runge-Kutta methods for stochastic ordinary differential equations*, Appl. Numer. Math., 22 (1996), pp. 81–101.

[8] K. Burrage and P. M. Burrage, *Order conditions of stochastic Runge–Kutta methods by B-series*, SIAM J. Numer. Anal., 38 (2000), pp. 1626–1646.

[9] K. Burrage, P. M. Burrage, and J. A. Belward, *A bound on the maximum strong order of stochastic Runge-Kutta methods for stochastic ordinary differential equations*, BIT, 37 (1997), pp. 771–780.

[10] K. Burrage and T. H. Tian, *Parallel half-block methods for initial value problems*, Appl. Numer. Math., 32 (2000), pp. 255–271.

[11] K. Burrage and T. H. Tian, *A note on the stability properties of the Euler methods for solving stochastic differential equations*, New Zealand J. Math., 29 (2000), pp. 115–127.

[12] P. M. Burrage, *Runge-Kutta Methods for Stochastic Differential Equations*, Ph.D. thesis, Department of Mathematics, the University of Queensland, Brisbane, Australia, 1999.

[13] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*, John Wiley, New York, 1987.

[14] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II*, Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.

[15] K. Jackson and S. P. Nørsett, *The potential for parallelism in Runge–Kutta methods, Part* 1: *RK formulas in standard form*, SIAM J. Numer. Anal., 32 (1995), pp. 49–82.

[16] P. E. Kloeden and E. Platen, *Stratonovich and Itô stochastic Taylor expansions*, Math. Nachr., 151 (1991), pp. 33–50.

[17] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1992.

[18] Y. Komori, T. Mitsui, and H. Sugiura, *Rooted tree analysis of the order conditions of ROW-type scheme for stochastic differential equations*, BIT, 37 (1997), pp. 43–66.

[19] G. N. Milstein, *Numerical Integration of Stochastic Differential Equations*, Kluwer, Dordrecht, The Netherlands, 1995.

[20] G. N. Milstein, E. Platen, and H. Schurz, *Balanced implicit methods for stiff stochastic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1010–1019.

[21] E. Platen, *On weak implicit and predictor-corrector methods*, Math. Comput. Simulation, 38 (1995), pp. 69–76.

[22] W. Rümelin, *Numerical treatment of stochastic differential equations*, SIAM J. Numer. Anal., 19 (1982), pp. 604–613.

[23] Y. Saito and T. Mitsui, *T-stability of numerical scheme for stochastic differential equations*, World Sci. Ser. Appl. Anal., 2 (1993), pp. 333–344.

[24] Y. Saito and T. Mitsui, *Stability analysis of numerical schemes for stochastic differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2254–2267.

[25] B. P. Sommeijer, *Parallelism in the Numerical Integration of Initial Value Problems*, CWI Tract, Amsterdam, The Netherlands, 1993.

[26] T. H. Tian and K. Burrage, *Two-stage stochastic Runge-Kutta methods for stochastic differential equations*, BIT, 42 (2002), pp. 625–643.

[27] P. J. van der Houwen and N. huu Cong, *Parallel block predictor-corrector methods of Runge-Kutta type*, Appl. Numer. Math., 13 (1993), pp. 109–123.

[28] P. J. van der Houwen, B. P. Sommeijer, and J. J. B. de Swart, *Parallel predictor-corrector methods*, J. Comput. Appl. Math., 66 (1996), pp. 53–71.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1538–1560

SHARP L^2-ERROR ESTIMATES AND SUPERCONVERGENCE OF MIXED FINITE ELEMENT METHODS FOR NON-FICKIAN FLOWS IN POROUS MEDIA*

RICHARD E. EWING†, YANPING LIN‡, TONG SUN§, JUNPING WANG¶, AND SHUHUA ZHANG‖

Dedicated to Professor Zhichun Piao on the occasion of his 68th birthday

Abstract. A sharper L^2-error estimate is obtained for the non-Fickian flow of fluid in porous media by means of a mixed Ritz–Volterra projection instead of the mixed Ritz projection used in [R. E. Ewing, Y. Lin, and J. Wang, *Acta Math. Univ. Comenian.* (*N.S.*), 70 (2001), pp. 75–84]. Moreover, local L^2 superconvergence for the velocity along the Gauss lines and for the pressure at the Gauss points is derived for the mixed finite element method via the Ritz–Volterra projection, and global L^2 superconvergence for the velocity and the pressure is also investigated by virtue of an interpolation postprocessing technique. On the basis of the superconvergence estimates, some useful a posteriori error estimators are presented for this mixed finite element method.

Key words. non-Fickian flow, mixed finite element methods, mixed Ritz–Volterra projection, error estimates, superconvergence

AMS subject classifications. 76S05, 45K05, 65M12, 65M60, 65R20

PII. S0036142900378406

1. Introduction.

As mentioned in [18, 19], the non-Fickian flow of fluid in porous media is complicated by the history effect which characterizes various mixing length growth of the flow and can be modeled by an integro-differential equation: Find $u = u(x,t)$ such that

$$
\begin{aligned}
u_t &= \nabla \cdot \sigma + cu + f && \text{in } \Omega \times J,\\
\sigma &= A(t) \cdot \nabla u - \int_0^t B(t,s) \cdot \nabla u(s) ds && \text{in } \Omega \times J,\\
u &= g && \text{on } \partial\Omega \times J,\\
u &= u_0(x) && x \in \Omega,\ t = 0,
\end{aligned}
\tag{1.1}
$$

*Received by the editors September 14, 2000; accepted for publication (in revised form) February 28, 2002; published electronically October 23, 2002.
http://www.siam.org/journals/sinum/40-4/37840.html

†Institute for Scientific Computation, Texas A&M University, College Station, TX 77843-3404 (ewing@isc.tamu.edu). This author was supported in part by NSF grants DMS-9626179, DMS-9706985, DMS-9707930, NCR-9710337, DMS-9972147, INT-9901498; EPA grant 825207; two generous awards from Mobil Technology Company; and Texas Higher Education Coordinating Board Advanced Research and Technology Program grants 010366-168 and 010366-0336.

‡Department of Mathematics, University of Alberta, Edmonton, Alberta T6G 2G1, Canada (ylin@math.ualberta.ca). This author was supported in part by NSERC and ISC of Texas A & M University.

§Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403 (tsun@math.bgsu.edu).

¶Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (jwang@mines.edu). This author's research was supported in part by NSF grant DMS-9706985.

‖Department of Mathematics, Tianjin University of Finance and Economics, Tianjin 300222 and LiuHui Center for Applied Mathematics, Nankai University and Tianjin University, Tianjin 300072, People's Republic of China (shuhua@eyou.com). This author was supported in part by SRF for ROCS and LuiHui Center for Applied Mathematics, Nankai University, and Tianjin University.

where $\Omega \subset R^d$ $(d = 2, 3)$ is an open bounded domain with smooth boundary $\partial\Omega$, $J = (0, T)$ with $T > 0$, $A(t) = A(x, t)$ and $B(t, s) = B(x, t, s)$ are two 2×2 or 3×3 matrices, and A is positive definite, and c, f, g, and u_0 are known smooth functions. This kind of model can arise, e.g., from the transport of contaminants in the subsurface, which is of great interest for engineers, physicists, and mathematicians involved in porous media flows modeling. The evolution of a reactive chemical within a velocity field exhibits excitement on many scales, typically represented by using the classical Fickian dispersion theory. For instance, the evolution in such a velocity field, when modeled with Fickian-type constitutive laws, leads to a dispersion tensor dependent upon the timescales of observation. Hence, to avoid this difficulty, nonlocal Fickian models have been recently proposed, in which the dispersion term arising from integration with respect to time makes the flow non-Fickian, since it is not a pure diffusion term. For example, Chen, Ewing, and Lazarov [4, 5], Cushman [6], Cushman, Hu, and Deng [7], Cushman, Hu, and Ginn [8], and Hu, Deng, and Cushman [23] have developed a nonlocal theory and some applications for the flow of fluid in porous media. Furtado et al. [21], Glimm et al. [22], Neuman and Zhang [29], and Ewing [12, 13, 14] also studied the history effect of various mixing length growth for flow in heterogeneous porous media. In a recent laboratory experimental investigation of contaminant transport in heterogeneous porous media [32], some nonlocal behavior of dispersion tensors have been observed.

There is now sizeable literature on the numerical approximations of the problem (1.1). In [31], the method of backward Euler and Crank–Nicolson combined with a certain numerical quadrature rule is employed to deal with the time direction, which aims at reducing the computational cost and storage spaces due to the memory effect. Finite element methods have been also developed for the problem (1.1) during the past ten years [2, 3, 25, 26, 27, 28, 34], in which optimal and superconvergence can be found for the corresponding finite element approximations in various norms, such as L^p with $2 \le p \le \infty$. In particular, the method of using the Ritz–Volterra projection, discovered by Cannon and Lin [2], proved to be a powerful technique behind the analysis. In fact, in [28] the concept of Ritz–Volterra projection is proposed to unify much of the analysis of standard finite element methods for different types of problems, such as parabolic and hyperbolic integro-differential equations and Sobolev- and viscoelasticity-type equations. See [16, 17] for recent developments on finite volume element approximations, where the Ritz–Volterra projection is also employed.

However, to the best of our knowledge, there are few results except [18, 19, 24] available concerning the mathematical formulation and analysis of the mixed finite element method for (1.1). Unlike the standard finite element method, the mixed finite element method can give the numerical approximations of the velocity field and the pressure field at the same time, and also maintains the physical conservation, so that it is more favorable. Certainly, its theoretical analysis is more complicated than that of the standard finite element method. In [18, 19] the authors dealt with the general setting of the problem. However, the formulation and analysis given in [24] are valid for only a special case; i.e., the operator B is proportional to the operator A. The reader is referred to [24] for this special case. The mathematical difficulty associated with the analysis of numerical approximations to the solution of (1.1) lies on the integral term added to standard parabolic equations [33, 34]. In order to overcome this difficulty, the so-called mixed Ritz–Volterra projection will be proposed in section 2.

In the present paper we are concerned with the approximate solutions of (1.1) by mixed finite element methods. Sharper L^2-error estimates than those in [18, 19]

are obtained by employing a mixed Ritz–Volterra projection rather than the Ritz projection used in [18, 19]. In addition, local L^2 superconvergence for the velocity along the Gauss lines and for the pressure at the Gauss points is derived, and with the aid of an interpolation postprocessing method global L^2 superconvergence is also considered for the velocity and the pressure.

The paper is organized in the following way. In section 2, we give some necessary preparations, introduce the mixed Ritz–Volterra projection, and analyze its approximation properties. In section 3, we derive a sharper error estimate for the mixed finite element approximations in the L^2-norm. Sections 4 and 5 are devoted to the local and global superconvergence analysis of the mixed finite element method, respectively.

2. The mixed Ritz–Volterra-type projection. In this section, we give the mixed finite element approximate formula for the parabolic integro-differential equation (1.1) and the mixed Ritz–Volterra projection. For simplicity, the method will be presented on plane domains.

Let $W := L^2(\Omega)$ be the standard L^2 space on Ω with norm $\|\cdot\|_0$. Denote by

$$\mathbf{V} := H(\mathrm{div}, \Omega) = \left\{\sigma \in (L^2(\Omega))^2 : \ \nabla \cdot \sigma \in L^2(\Omega)\right\}$$

the Hilbert space equipped with the following norm:

$$\|\sigma\|_{\mathbf{V}} := \left(\|\sigma\|_0^2 + \|\nabla \cdot \sigma\|_0^2\right)^{\frac{1}{2}}.$$

There are several ways to discretize the problem (1.1) based on the variables σ and u; each method corresponds to a particular variational form of (1.1) [18, 19].

Let T_h be a finite element partition of Ω into triangles or quadrilaterals which is quasi-uniform. Let $\mathbf{V}_h \times W_h$ denote a pair of finite element spaces satisfying the Brezzi–Babuška condition. For example, the elements of Raviart and Thomas [30] would be a good choice for $\mathbf{V}_h$ and W_h. Although our results are based on the use of Raviart–Thomas elements of any order k, their extension to other stable elements can be discussed without any difficulty.

Let us recall from [18] that the weak mixed formulation of (1.1) is given by finding $(u, \sigma) \in W \times \mathbf{V}$ such that

$$\begin{aligned}
&(u_t, w) - (\nabla \cdot \sigma, w) - (cu, w) = (f, w) && \forall w \in W,\\
&(\alpha\sigma, \mathbf{v}) + \int_0^t (M(t,s)\sigma(s), \mathbf{v})ds + (\nabla \cdot \mathbf{v}, u) = \langle g, \mathbf{v} \cdot \mathbf{n}\rangle && \forall \mathbf{v} \in \mathbf{V},\\
&u(0,x) = u_0(x) \quad \text{in } L^2(\Omega),
\end{aligned} \tag{2.1}$$

where $\alpha = A^{-1}(t)$, $M(t,s) = R(t,s)A^{-1}(s)$, and $R(t,s)$ is the resolvent of the matrix $A^{-1}(t)B(t,s)$ and is given by

$$R(t,s) = A^{-1}(t)B(t,s) + \int_s^t A^{-1}(t)B(t,\tau)\ R(\tau,s)d\tau, \quad t > s \geq 0.$$

Here $\langle\cdot,\cdot\rangle$ indicates the L^2-inner product on $\partial\Omega$.

The corresponding semidiscrete version seeks a pair $(u_h, \sigma_h) \in W_h \times \mathbf{V}_h$ such that

$$\begin{aligned}
&(u_{h,t}, w_h) - (\nabla \cdot \sigma_h, w_h) - (cu_h, w_h) = (f, w_h) && \forall w_h \in W_h,\\
&(\alpha\sigma_h, \mathbf{v}_h) + \int_0^t (M(t,s)\sigma_h(s), \mathbf{v}_h)ds + (\nabla \cdot \mathbf{v}_h, u_h) = \langle g, \mathbf{n} \cdot \mathbf{v}_h\rangle && \forall \mathbf{v}_h \in \mathbf{V}_h.
\end{aligned} \tag{2.2}$$

The discrete initial condition $u_h(0,x) = u_{0,h}$, where $u_{0,h} \in W_h$ is some appropriately chosen approximation of the initial data $u_0(x)$, should be added to (2.2) for starting. The pair (u_h, σ_h) is a semidiscrete approximation of the true solution of (1.1) in the finite element space $W_h \times \mathbf{V}_h$ [1, 18, 19, 31], where $\sigma_h(0,x)$ is chosen to satisfy (2.2) with $t = 0$; namely, it is related to $u_{0,h}$ as follows:

$$(\alpha\sigma_h(0), \mathbf{v}_h) + (u_{0,h}, \nabla \cdot \mathbf{v}_h) = \langle g_0, \mathbf{n} \cdot \mathbf{v}_h \rangle, \tag{2.3}$$

where $g_0 = g(0,x)$ is the initial value of the boundary data.

In [18], utilizing the mixed Ritz projection we have obtained for the Raviart–Thomas element of the lowest order that

$$||u - u_h||_0^2 + ||\sigma - \sigma_h||_0^2 \le Ch^2 \left[||u_0||_1^2 + ||\sigma_0||_1^2 + \int_0^t (||u(s)||_2^2 + ||u_t(s)||_2^2) ds \right].$$

Also, we can extend easily the result to the case of any order k (≥ 1) to get

$$||u - u_h||_0^2 + ||\sigma - \sigma_h||_0^2 \le Ch^{2r} \left[||u_0||_r^2 + ||\sigma_0||_r^2 + \int_0^t (||u(s)||_{r+1}^2 + ||u_t(s)||_{r+1}^2) ds \right], \tag{2.4}$$

for $2 \le r \le k+1$. In fact, we can improve the error estimate by extending the idea from [2, 3] to introduce a new nonlocal projection incorporated with the memory effects, which allows us to obtain a sharper error estimate in regularity than that indicated in (2.4). This new projection is a natural extension of the standard Ritz–Volterra projection in the standard finite element method to the case of the mixed finite element approximations with memory. We refer the readers to [2, 3] and [28] for the analysis and applications of the Ritz–Volterra projection for standard finite element approximations to parabolic and hyperbolic integro-differential equations.

Before the mixed Ritz–Volterra projection is given, we need the following Raviart–Thomas projection [30]:

$$\Pi_h \times P_h : \mathbf{V} \times W \to \mathbf{V}_h \times W_h,$$

which has the following properties:

(i) P_h is the local $L^2(\Omega)$ projection.

(ii) Π_h and P_h satisfy

$$(\nabla \cdot (\sigma - \mathbf{\Pi}_h \sigma), w_h) = 0, \ w_h \in W_h \quad \text{and} \quad (\nabla \cdot \mathbf{v}_h, u - P_h u) = 0, \ \mathbf{v}_h \in \mathbf{V}_h. \tag{2.5}$$

(iii) The following approximation properties hold:

$$\begin{aligned} &||\sigma - \mathbf{\Pi}_h \sigma||_0 \le Ch^r ||\sigma||_r, && 1 \le r \le k+1, \\ &||\nabla \cdot (\sigma - \Pi_h \sigma)||_{-s} \le Ch^{r+s} ||\nabla \cdot \sigma||_r, && 0 \le r, \ s \le k+1, \\ &||u - P_h u||_{-s} \le Ch^{r+s} ||u||_r, && 0 \le r, \ s \le k+1. \end{aligned} \tag{2.6}$$

DEFINITION 2.1. *For $(u, \sigma) \in W \times \mathbf{V}$ we define a pair $(\bar{u}_h, \bar{\sigma}_h) : [0,T] \to W_h \times \mathbf{V}_h$ such that*

$$\begin{aligned} &\left(\alpha(\sigma - \bar{\sigma}_h) + \int_0^t M(t,s)(\sigma - \bar{\sigma}_h)(s) ds, \mathbf{v}_h \right) + (\nabla \cdot \mathbf{v}_h, u - \bar{u}_h) = 0, && \mathbf{v}_h \in \mathbf{V}_h, \\ &(\nabla \cdot (\sigma - \bar{\sigma}_h), w_h) + (c(u - \bar{u}_h), w_h) = 0, && w_h \in W_h, \end{aligned} \tag{2.7}$$

where $\alpha = A^{-1}$. The pair $(\bar{u}_h, \bar{\sigma}_h)$ is called the mixed Ritz–Volterra projection of (u, σ).

Let

$$\xi := \sigma - \bar{\sigma}_h,\ \eta := u - \bar{u}_h,\ \nu := \Pi_h\sigma - \bar{\sigma}_h,\ \tau := P_h u - \bar{u}_h,\ \rho := u - P_h u.$$

Then (2.7) becomes

$$\begin{aligned} &\left(\alpha\xi + \int_0^t M(t,s)\xi(s)ds, \mathbf{v}_h\right) + (\nabla\cdot\mathbf{v}_h, \eta) = 0, \qquad \mathbf{v}_h \in \mathbf{V}_h, \\ &(\nabla\cdot\xi, w_h) + (c\eta, w_h) = 0, \qquad w_h \in W_h, \end{aligned} \tag{2.8}$$

or, according to (2.5),

$$\begin{aligned} &(\alpha\xi, \mathbf{v}_h) + (\nabla\cdot\mathbf{v}_h, \tau) = f(\mathbf{v}_h), \qquad \mathbf{v}_h \in \mathbf{V}_h, \\ &(\nabla\cdot\xi, w_h) + (c\tau, w_h) = g(w_h), \qquad w_h \in W_h, \end{aligned} \tag{2.9}$$

where

$$f(\mathbf{v}_h) := -\left(\int_0^t M(t,s)\xi(s)ds, \mathbf{v}_h\right) \quad \text{and} \quad g(w_h) := -(c\rho, w_h).$$

In order to analyze (ξ, η), let us recall from [10] the following results.

LEMMA 2.2. *Let the index k of $\mathbf{V}_h \times W_h$ be at least one and let $0 \le s \le k-1$. Assume that Ω is $(s+2)$-regular* [10]. *Let $\xi \in \mathbf{V}$, $g \in W' = L^2(\Omega)$ and $f = \{\mathbf{f}_0, f_1\} \in \mathbf{V}'$ with $\mathbf{f}_0 \in (L^2(\Omega))^2$, $f_1 \in L^2(\Omega)$ and*

$$f(\mathbf{v}) = (\mathbf{f}_0, \mathbf{v}) + (f_1, \nabla\cdot\mathbf{v}), \quad \mathbf{v} \in \mathbf{V}.$$

If $z \in W_h$ satisfies the relations

$$\begin{aligned} &(\alpha\xi, \mathbf{v}_h) + (\nabla\cdot\mathbf{v}_h, z) = f(\mathbf{v}_h), \qquad \mathbf{v}_h \in \mathbf{V}_h, \\ &(\nabla\cdot\xi, w_h) + (cz, w_h) = g(w_h), \qquad w_h \in W_h, \end{aligned} \tag{2.10}$$

then there exists $h_0 > 0$ sufficiently small such that, for all $0 < h \le h_0$,

$$\begin{aligned} ||z||_{-s} \quad &\le C\left\{h^{s+1}||\xi||_0 + h^{s+2}||\nabla\cdot\xi||_0 + ||\mathbf{f}_0||_{-s-1} + h^{s+1}||\mathbf{f}_0||_0\right. \\ &\left. + ||f_1||_{-s} + h^s||f_1||_0 + ||g||_{-s-2} + h^{s+2}||g||_0\right\}. \end{aligned}$$

LEMMA 2.3. *Let the index k of $\mathbf{V}_h \times W_h$ be nonnegative, and let Ω be $(k+2)$-regular* [10]. *Let $\xi \in \mathbf{V}$, $g \in W' = L^2(\Omega)$ and $f = \{\mathbf{f}_0, 0\} \in \mathbf{V}'$. If $z \in W_h$ satisfies* (2.10), *then there exists $h_0 > 0$ sufficiently small such that, for all $0 < h \le h_0$,*

$$||z||_{-k} \le C\left\{h^{k+1}\left(||\xi||_0 + ||\nabla\cdot\xi||_0 + ||\mathbf{f}_0||_0 + ||g||_0\right) + ||\mathbf{f}_0||_{-k-1} + ||g||_{-k-2}\right\}.$$

Moreover, we also need the following lemma.

LEMMA 2.4. *Assume that the matrix $A(t)$ is positive definite. Then the norms $||\sigma||_0^2 := (\sigma, \sigma)$ and $||\sigma||_{A^{-1}}^2 := (A^{-1}\sigma, \sigma)$ are equivalent.*

We are now ready to state and prove our main result in this section.

THEOREM 2.5. *For $(u, \sigma) \in W \times \mathbf{V}$ its mixed Ritz–Volterra projection $(\bar{u}_h, \bar{\sigma}_h)$ defined by* (2.7) *exists and is unique. Moreover, there is a positive constant $C > 0$, independent of $h > 0$ small, such that the error $(u - \bar{u}_h, \sigma - \bar{\sigma}_h)$ can be estimated by*

$$\begin{aligned} &||u - \bar{u}_h||_0 \le C\begin{cases} h|||u(t)|||_2 & \text{if } k = 0, \\ h^r|||u(t)|||_r & \text{if } k \ge 1 \text{ and } 2 \le r \le k+1, \end{cases} \\ &||\sigma - \bar{\sigma}_h||_0 \le Ch^r|||u(t)|||_{r+1} \qquad \text{if } 1 \le r \le k+1, \\ &||\nabla\cdot(\sigma - \bar{\sigma}_h)||_0 \le Ch^r|||u(t)|||_{r+2} \quad \text{if } 0 \le r \le k+1, \end{aligned}$$

where

$$|||u(t)|||_r = ||u(t)||_r + \int_0^t ||u(s)||_r ds, \quad r \in R,\ t \geq 0.$$

Proof. We first prove the existence and uniqueness of the mixed Ritz–Volterra projection. If $M = 0$, then it follows from [1] that $(\bar{u}_h, \bar{\sigma}_h)$ exists uniquely. If M is nonzero, we see that (2.7) in fact can be written as a Volterra system for $(\bar{u}_h, \bar{\sigma}_h)$, i.e.,

$$A_h \begin{pmatrix} \bar{u}_h \\ \bar{\sigma}_h \end{pmatrix} = F_h + \int_0^t B_h(t,s) \begin{pmatrix} \bar{u}_h \\ \bar{\sigma}_h \end{pmatrix} ds,$$

where A_h and B_h are matrices with A_h nonsingular and F_h is a vector associated with the solution (u, σ). Hence, the theory of Volterra equations implies that $(\bar{u}_h, \bar{\sigma}_h)$ exists uniquely.

Next we turn our attention to error estimates. It follows from (2.6) and (2.9) that

$$\begin{aligned} &||f||_0 \leq C \int_0^t ||\xi||_0 ds, && ||f||_{-1} \leq C \int_0^t ||\xi||_{-1} ds, \\ &||g||_0 \leq C||\rho||_0, && ||g||_{-1} \leq C||\rho||_{-1}, \\ &||g||_{-2} \leq ||g||_{-1} \leq C||\rho||_{-1}, && ||\rho||_{-1} + h||\rho||_0 \leq Ch^{r+1}||u||_r. \end{aligned}$$

Now we apply either Lemma 2.2 with $s = 0$ or Lemma 2.3 with $k = 0$ to (2.9). Then, for h small and for Ω 2-regular we have for $0 \leq r \leq k+1$ that

$$\begin{aligned} ||\tau||_0 &\leq C\left\{h||\xi||_0 + h^{2-\delta_{k0}}||\nabla \cdot \xi||_0 + ||f||_{-1} + h||f||_0 + ||g||_{-2} + h||g||_0\right\} \\ &\leq C\left\{h||\xi||_0 + h^{2-\delta_{k0}}||\nabla \cdot \xi||_0 + \int_0^t (||\xi||_{-1} + h||\xi||_0) ds + (||\rho||_{-1} + h||\rho||_0)\right\} \\ &\leq C\left\{h|||\xi|||_0 + h^{2-\delta_{k0}}||\nabla \cdot \xi||_0 + \int_0^t ||\xi||_{-1} ds + h^{r+1}||u||_r\right\}, \end{aligned} \tag{2.11}$$

where

$$\delta_{k0} = \begin{cases} 1, & k = 0, \\ 0, & k \neq 0. \end{cases}$$

Letting $\varphi \in \left(H^1(\Omega)\right)^2$, then we derive from (2.5) and (2.8) that

$$\begin{aligned} &\left(\alpha\xi + \int_0^t M(t,s)\xi(s) ds, \varphi\right) + (\nabla \cdot \varphi, \eta) \\ &\quad = \left(\alpha\xi + \int_0^t M(t,s)\xi(s) ds, \varphi - \Pi_h\varphi\right) + (\nabla \cdot (\varphi - \Pi_h\varphi), \eta) \\ &\quad + \left(\alpha\xi + \int_0^t M(t,s)\xi(s) ds, \Pi_h\varphi\right) + (\nabla \cdot \Pi_h\varphi, \eta) \\ &\quad = \left(\alpha\xi + \int_0^t M(t,s)\xi(s) ds, \varphi - \Pi_h\varphi\right) + (\nabla \cdot (\varphi - \Pi_h\varphi), u) \end{aligned}$$

or

$$\begin{aligned} (\alpha\xi, \varphi) &= -\int_0^t (M(t,s)\xi(s), \varphi) ds - (\nabla \cdot \varphi, \eta) \\ &\quad + \left(\alpha\xi + \int_0^t M(t,s)\xi(s) ds, \varphi - \Pi_h\varphi\right) + (\nabla \cdot (\varphi - \Pi_h\varphi), u) \end{aligned}$$

which, together with (2.6), indicates that

$$\begin{aligned}|(\alpha\xi,\varphi)| &\le C\int_0^t ||\xi(s)||_{-1}ds||\varphi||_1 + ||\eta||_0||\varphi||_1 \\ &\quad + Ch|||\xi|||_0||\varphi||_1 + Ch||u||_1||\nabla\cdot(\varphi-\Pi_h\varphi)||_{-1} \\ &\le C\left(\int_0^t ||\varphi||_{-1}ds + ||\eta||_0 + Ch|||\xi|||_0 + Ch||u||_1\right)||\varphi||_1;\end{aligned}$$

that is,

$$||\xi||_{-1} \le C\left\{\int_0^t ||\xi(s)||_{-1}ds + ||\eta||_0 + Ch(|||\xi|||_0 + ||u||_1)\right\}.$$

This, together with Gronwall's lemma, implies that

$$||\xi||_{-1} \le C\left\{||\eta||_0 + Ch(|||\xi|||_0 + ||u||_1)\right\}. \tag{2.12}$$

Substitute (2.12) into (2.11) to obtain

$$||\tau||_0 \le C\left\{\int_0^t ||\eta(s)||_0 ds + h|||\xi|||_0 + h^{2-\delta_{k0}}||\nabla\cdot\xi||_0 + h^{r+1}||u||_r\right\}. \tag{2.13}$$

Therefore, for $0 \le r \le k+1$ we have

$$\begin{aligned}||\eta||_0 &\le ||\rho||_0 + ||\tau||_0 \\ &\le C\left\{\int_0^t ||\eta(s)||_0 ds + h|||\xi|||_0 + h^{2-\delta_{k0}}|||\nabla\cdot\xi|||_0 + h^r||u||_r\right\},\end{aligned}$$

and applying Gronwall's lemma leads to

$$||\eta||_0 \le C\left\{h|||\xi|||_0 + h^{2-\delta_{k0}}||\nabla\cdot\xi||_0 + h^r||u||_r\right\}. \tag{2.14}$$

Since, by (2.5), $(\nabla\cdot\nu, w_h) = (\nabla\cdot\xi, w_h)$ for $w_h \in W_h$, it follows from (2.8) and the choice $w_h = \nabla\cdot\nu \in W_h$ that

$$(\nabla\cdot\nu, \nabla\cdot\nu) = (\nabla\cdot\xi, \nabla\cdot\nu) = -(c\eta, \nabla\cdot\nu)$$

or

$$||\nabla\cdot\nu||_0 \le C||\eta||_0 \tag{2.15}$$

so that

$$||\nabla\cdot\xi||_0 \le ||\nabla\cdot\nu||_0 + ||\nabla\cdot(\sigma-\Pi_h\sigma)||_0 \le C(||\eta||_0 + h^q||\nabla\cdot\sigma||_q),\ 0 \le q \le k+1. \tag{2.16}$$

Also, according to (2.8) ν satisfies

$$\begin{aligned}&\left(\alpha\nu + \int_0^t M(t,s)\nu(s)ds, \nu\right) \\ &= \left(\alpha\xi + \int_0^t M(t,s)\xi(s)ds, \nu\right) + \left(\alpha(\Pi_h\sigma-\sigma) + \int_0^t M(t,s)(\Pi_h\sigma-\sigma)(s)ds, \nu\right) \\ &= -(\nabla\cdot\nu,\eta) + \left(\alpha(\Pi_h\sigma-\sigma) + \int_0^t M(t,s)(\Pi_h\sigma-\sigma)(s)ds, \nu\right) \\ &\le ||\nabla\cdot\nu||_0^2 + ||\eta||_0^2 + C|||\Pi_h\sigma-\sigma|||_0||\nu||_0.\end{aligned}$$

Then we find from Lemma 2.4, (2.15), and the ϵ-type inequality that

$$||\nu||_0^2 - C\int_0^t ||\nu(s)||_0^2 ds \le C(||\eta||_0 + |||\Pi_h\sigma - \sigma|||_0)$$

which, together with Gronwall's lemma and (2.6), implies

$$||\nu||_0 \le C(||\eta||_0 + |||\Pi_h\sigma - \sigma|||_0) \le C(||\eta||_0 + h^m|||\sigma|||_m),\ 1 \le m \le k+1, \tag{2.17}$$

and

$$||\xi||_0 \le ||\nu||_0 + ||\Pi_h\sigma - \sigma||_0 \le C(||\eta||_0 + h^m|||\sigma|||_m),\ 1 \le m \le k+1. \tag{2.18}$$

If (2.16) and (2.18) are substituted into (2.14), then for $0 \le r \le k+1$, $0 \le q \le k+1$, and $1 \le m \le k+1$ it follows that

$$||\eta||_0 \le C\left\{h|||\eta|||_0 + h^r||u||_r + h^{m+1}|||\sigma|||_m + h^{2-\delta_{k0}+q}||\nabla\cdot\sigma||_q\right\}.$$

Thus, for small h we obtain via Gronwall's inequality that

$$||\eta||_0 \le C\left\{h^r||u||_r + h^{m+1}|||\sigma|||_m + h^{2-\delta_{k0}+q}||\nabla\cdot\sigma||_q\right\},$$
$$0 \le r,\ q \le k+1,\ 1 \le m \le k+1.$$

Choose $r = m+1 = 2+q-\delta_{k0}$ to gain that

$$||\eta||_0 = \begin{cases} Ch|||u|||_2 & \text{if } k = 0, \\ Ch^r|||u|||_r & \text{if } k \ge 1 \text{ and } 2 \le r \le k+1, \end{cases}$$

since $||\sigma||_{r-1} + ||\nabla\cdot\sigma||_{r-2} \le C||u||_r$.

It then follows immediately that

$$||\xi||_0 \le Ch^r|||u|||_{r+1}, \qquad 1 \le r \le k+1,$$
$$||\nabla\cdot\xi||_0 \le Ch^r|||u|||_{r+2}, \qquad 0 \le r \le k+1.$$

Therefore, the proof of Theorem 2.5 is completed. □

THEOREM 2.6. *Let $(\bar{u}_h, \bar{\sigma}_h)$ be the mixed Ritz–Volterra projection of $(u,\sigma) \in W \times \mathbf{V}$ defined by (2.7). Then there is a positive constant $C > 0$, independent of $h > 0$ small, such that the error $(u - \bar{u}_h, \sigma - \bar{\sigma}_h)$ can be estimated for any positive integer m by*

$$||D_t^m(u - \bar{u}_h)||_0 \le C\begin{cases} h|||u(t)|||_{2,m} & \text{if } k = 0, \\ h^r|||u(t)|||_{r,m} & \text{if } k \ge 1 \text{ and } 2 \le r \le k+1, \end{cases}$$
$$||D_t^m(\sigma - \bar{\sigma}_h)||_0 \le Ch^r|||u(t)|||_{r+1,m} \quad \text{if } 1 \le r \le k+1,$$
$$||D_t^m(\nabla\cdot(\sigma - \bar{\sigma}_h))||_0 \le Ch^r|||u(t)|||_{r+2,m} \quad \text{if } 0 \le r \le k+1,$$

where

$$|||u(t)|||_{r,m} = \sum_{j=0}^m ||D_t^j u(t)||_r + \int_0^t \sum_{j=0}^m ||D_t^j u(s)||_r ds, \qquad r \in R,\ t \ge 0.$$

Proof. Differentiate (2.7), and then the result for $m = 1$ follows from the same arguments as those for Theorem 2.5.

The proof is completed by treating $m \geq 2$ inductively, using the further differentiation of (2.7). ∎

COROLLARY 2.7. *Let* $(\bar{u}_h, \bar{\sigma}_h)$ *be the mixed Ritz–Volterra projection of* $(u, \sigma) \in W \times \mathbf{V}$ *defined by* (2.7). *Then*

$$||u - \bar{u}_h||_\infty \leq Ch^r \left(||u||_{r,\infty} + |||u|||_{r+1} \right), \quad k \geq 1, \quad \textit{and} \quad 1 \leq r \leq k.$$

Proof. We easily see from (2.13) and Theorem 2.5 that

$$||\tau||_0 \leq Ch^{r+1} |||u|||_{r+1} \quad \text{for } k \geq 1 \quad \text{and} \quad 1 \leq r \leq k$$

and by the inverse inequality that

$$||\tau||_\infty \leq Ch^{-1} ||\tau||_0 \leq Ch^r |||u|||_{r+1}.$$

Thus, we have for $k \geq 1$ and $1 \leq r \leq k$ that

$$\begin{aligned} ||u - \bar{u}_h||_\infty &\leq ||u - P_h u||_\infty + ||\tau||_\infty \\ &\leq Ch^r (||u||_{r,\infty} + |||u|||_{r+1}). \end{aligned}$$ ∎

Remark 2.1. For $k = 0$ we do not have any estimate for the quantity $||u - \bar{u}_h||_\infty$. However, using the superconvergence analysis to be presented in Corollary 5.4, we have for the rectangular Raviart–Thomas elements of the lowest order,

$$||u - u_h||_\infty \leq Ch,$$

where (u, σ) and (u_h, σ_h) are the solutions of (2.1) and (2.2), respectively.

THEOREM 2.8. *Assume that* $(\bar{u}_h, \bar{\sigma}_h)$ *is the mixed Ritz–Volterra projection of* $(u, \sigma) \in W \times \mathbf{V}$ *defined by* (2.7). *Then there is a positive constant* $C_m > 0$, *independent of* $h > 0$ *small, such that for* $m \geq 0$

$$||D_t^m \bar{u}_h||_W + ||D_t^m \bar{\sigma}_h||_{\mathbf{V}} \leq C_m \left\{ \sum_{j=0}^{m} (||D_t^j \sigma||_{\mathbf{V}} + ||D_t^j u||_W) + \int_0^t (||\sigma||_{\mathbf{V}} + ||u||_W) ds \right\}. \tag{2.19}$$

Proof. Rewrite (2.7) as

$$\begin{aligned} &(\alpha \bar{\sigma}_h, \mathbf{v}_h) + (\nabla \cdot \mathbf{v}_h, \bar{u}_h) = F(\mathbf{v}_h), \qquad \mathbf{v}_h \in \mathbf{V}_h, \\ &(\nabla \cdot \bar{\sigma}_h, w_h) + (c\bar{u}_h, w_h) = G(w_h), \qquad w_h \in W_h, \end{aligned}$$

where

$$\begin{aligned} F(\mathbf{v}_h) &= \left(\alpha\sigma + \int_0^t M(t,s)(\sigma - \bar{\sigma}_h)(s) ds, \mathbf{v}_h \right) + (\nabla \cdot \mathbf{v}_h, u), \\ G(w_h) &= (\nabla \cdot \sigma, w_h) + (cu, w_h). \end{aligned}$$

$F(\mathbf{v}_h)$ and $G(w_h)$ can be considered as linear functionals of $\mathbf{v}_h$ and w_h defined on $\mathbf{V}_h$ and W_h, respectively. Thus, we have from the stability result of [1] that

$$\begin{aligned} ||\bar{\sigma}_h||_{\mathbf{V}} + ||\bar{u}_h||_W &\leq C \left\{ \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{|F(\mathbf{v}_h)|}{||\mathbf{v}_h||_{\mathbf{V}}} + \sup_{w_h \in W_h} \frac{|G(w_h)|}{||w_h||_W} \right\} \\ &\leq C \left\{ ||\sigma||_{\mathbf{V}} + \int_0^t ||\sigma||_V ds + ||u||_W + \int_0^t ||\bar{\sigma}_h||_{\mathbf{V}} ds \right\}, \end{aligned}$$

or, by Gronwall's inequality,

$$||\bar{\sigma}_h||_{\mathbf{V}} + ||\bar{u}_h||_W \le C\left\{||\sigma||_{\mathbf{V}} + \int_0^t ||\sigma||_{\mathbf{V}} ds + ||u||_W\right\},$$

which demonstrates that (2.19) is true for $m = 0$.

We can also prove (2.19) for $m \ge 1$ by differentiating (2.7) with respect to time t and repeating the same arguments above with mathematical induction. ∎

Remark 2.2. This stability result (2.19) is needed in the analysis of the backward Euler time-discretization scheme. See [19] for details.

3. Sharp L^2-error estimates. In this section, we shall show a sharper L^2-error estimate than the one indicated in (2.4) for the time-continuous approximation scheme (2.2), where the regularity requirement is one order lower than in (2.4), by means of the mixed Ritz–Volterra-type projection instead of the mixed Ritz projection used in [18] to obtain (2.4). Here, let us consider the Raviart–Thomas elements of higher order $k \ge 1$ (see [18] for the lowest-order case).

THEOREM 3.1. *Assume that* (u, σ) *and* (u_h, σ_h) *are the solutions of* (2.1) *and* (2.2), *respectively,* $||P_h u_0 - u_h(0)|| \le Ch^r ||u_0||_r$ *and* $||\Pi_h \sigma(0) - \sigma_h(0)|| \le Ch^r ||u_0||_{r+1}$. *Then we have for* $k \ge 1$ *that*

$$\begin{aligned}
&||u(t) - u_h(t)||_0^2 \\
&\quad \le Ch^{2r}\left\{||u_0||_r^2 + \int_0^t [||u(s)||_r^2 + ||u_t(s)||_r^2] ds\right\}, \qquad 2 \le r \le k+1, \\
&||\sigma(t) - \sigma_h(t)||_0^2 \\
&\quad \le Ch^{2r}\left\{||u_0||_{r+1}^2 + \int_0^t [||u(s)||_{r+1}^2 + ||u_t(s)||_{r+1}^2] ds\right\}, \qquad 1 \le r \le k+1.
\end{aligned}$$

Proof. Let $(\bar{u}_h, \bar{\sigma}_h)$ be the mixed Ritz–Volterra projection of (u, σ) defined by (2.7), and we rewrite the errors as

$$\begin{aligned}
u - u_h &= (u - \bar{u}_h) + (\bar{u}_h - u_h) := \rho + \rho_h, \\
\sigma - \sigma_h &= (\sigma - \bar{\sigma}_h) + (\bar{\sigma}_h - \sigma_h) := \theta + \theta_h.
\end{aligned}$$

Then we know from Theorems 2.5 and 2.6 that

$$\begin{aligned}
&||\rho||_0 \le Ch^r |||u(t)|||_r, && k \ge 1, \quad \text{and} \quad 2 \le r \le k+1, \\
&||\rho_t||_0 \le Ch^r \left(|||u(t)|||_r + |||u_t(t)|||_r\right), && k \ge 1, \quad \text{and} \quad 2 \le r \le k+1
\end{aligned} \tag{3.1}$$

and

$$||\theta(t)||_0 \le Ch^r |||u|||_{r+1}, \quad 1 \le r \le k+1. \tag{3.2}$$

Thus, only $||\rho_h||_0$ and $||\theta_h||_0$ need to be estimated.

It follows from (2.1)–(2.2) and (2.7) that (ρ_h, θ_h) satisfies

$$\begin{aligned}
&\left(\alpha\theta_h + \int_0^t M(t,s)\theta_h(s) ds, \mathbf{v_h}\right) + (\nabla \cdot \mathbf{v}_h, \rho_h) = 0, && \mathbf{v}_h \in \mathbf{V}_h, \\
&(\rho_{h,t}, w_h) - (\nabla \cdot \theta_h, w_h) - (c\rho_h, w_h) = -(\rho_t, w_h), && w_h \in W_h.
\end{aligned} \tag{3.3}$$

Therefore, setting $w_h = \rho_h$ and $\mathbf{v}_h = \theta_h$ in (3.3) we obtain from their sum that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}||\rho_h||_0^2 - (c\rho_h, \rho_h) + ||\theta_h||_{A^{-1}}^2 &= -\left(\int_0^t M(t,s)\theta_h(s)ds, \theta_h\right) - (\rho_t, \rho_h)\\ &\le C\int_0^t ||\theta_h(s)||_0 ds ||\theta_h||_0 + ||\rho_t||_0||\rho_h||_0\end{aligned}$$

and by means of Lemma 2.4 that

$$\frac{1}{2}\frac{d}{dt}||\rho_h||_0^2 + ||\theta_h||_{A^{-1}}^2 \le C\left(||\rho_h||_0^2 + \int_0^t ||\theta_h||_{A^{-1}}^2 ds\right) + \frac{1}{2}\left(||\theta_h||_{A^{-1}}^2 + ||\rho_t||_0^2\right).$$

Integrating from 0 to t leads to

$$||\rho_h||_0^2 + \int_0^t ||\theta_h||_{A^{-1}}^2 ds \le ||\rho_h(0)||_0^2 + \int_0^t \left[||\rho_h||_0^2 + \int_0^s ||\theta_h(s)||_{A^{-1}}^2 ds\right] + \int_0^t ||\rho_t||_0^2 ds$$

which, together with Gronwall's lemma, implies

$$||\rho_h||_0^2 + \int_0^t ||\theta_h(s)||_{A^{-1}}^2 ds \le C\left\{||\rho_h(0)||_0^2 + \int_0^t ||\rho_t||_0^2 ds\right\}. \tag{3.4}$$

It follows from (2.6), Theorem 2.5, and our initial approximation assumption that

$$\begin{aligned}||\rho_h(0)||_0^2 &= ||\bar{u}_h(0) - u_h(0)||_0^2 \le ||\bar{u}_h(0) - u_0||_0^2\\ &\quad + ||u_0 - P_h u_0||_0^2 + ||P_h u_0 - u_h(0)||_0^2\\ &\le Ch^{2r}||u_0||_r^2.\end{aligned} \tag{3.5}$$

Combining (3.1) and (3.5) with (3.4) we gain

$$||\rho_h||_0^2 \le Ch^{2r}\left\{||u_0||_r^2 + \int_0^t [||u(s)||_r^2 + ||u_t(s)||_r^2] ds\right\}. \tag{3.6}$$

In order to get the estimate for $\theta_h(t)$, we first differentiate (3.3) to obtain

$$\left(\alpha_t\theta_h + \alpha\theta_{h,t} + M(t,t)\theta_h + \int_0^t M_t(t,s)\theta_h(s)ds, \mathbf{v}_h\right) + (\nabla\cdot\mathbf{v}_h, \rho_{h,t}) = 0,\ \mathbf{v}_h \in \mathbf{V}_h,$$

and then by setting $\mathbf{v}_h = \theta_h$ in the above equation and $w_h = \rho_{h,t}$ in (3.3) we have that

$$\begin{aligned}||\rho_{h,t}||_0^2 + (\alpha\theta_{h,t}, \theta_h) + (\alpha_t\theta_h, \theta_h) &= -\left(M(t,t)\theta_h + \int_0^t M_t(t,s)\theta_h(s)ds, \theta_h\right)\\ &\quad + (c\rho_h, \rho_{h,t}) - (\rho_t, \rho_{h,t}).\end{aligned} \tag{3.7}$$

Since

$$\alpha(\theta_h^2)_t = (\alpha\theta_h^2)_t - \alpha_t\theta_h^2,$$

then

$$\begin{aligned}(\alpha\theta_{h,t},\theta_h) &= \int_\Omega \alpha\theta_{h,t}\theta_h = \frac{1}{2}\int_\Omega \alpha\frac{d}{dt}(\theta_h^2)\\ &= \frac{1}{2}\int_\Omega \frac{d}{dt}(\alpha\theta_h^2) - \frac{1}{2}\int_\Omega \alpha_t\theta_h^2\\ &= \frac{1}{2}\frac{d}{dt}||\theta_h||^2_{A^{-1}} - \frac{1}{2}(\alpha_t\theta_h,\theta_h).\end{aligned}$$

Hence, (3.7) can be rewritten as

$$\begin{aligned}||\rho_{h,t}||_0^2 + \frac{1}{2}\frac{d}{dt}||\theta_h||^2_{A^{-1}} + \frac{1}{2}(\alpha_t\theta_h,\theta_h) &= -\left(M(t,t)\theta_h + \int_0^t M_t(t,s)\theta_h(s)ds,\theta_h\right)\\ &\quad + (c\rho_h,\rho_{h,t}) - (\rho_t,\rho_{h,t}).\end{aligned}$$

Thus, from the ϵ-inequality we derive that

$$||\rho_{h,t}||_0^2 + \frac{d}{dt}||\theta_h||^2_{A^{-1}} \le C\left\{||\theta_h||_0^2 + \int_0^t ||\theta_h(s)||_0^2 ds + ||\rho_h||_0^2 + ||\rho_t||_0^2\right\}$$

and then via integrating from 0 to t, Lemma 2.4, and Gronwall's lemma that

$$||\theta_h||_0^2 \le C\left\{||\theta_h(0)||_0^2 + \int_0^t [||\rho_h(s)||_0^2 + ||\rho_t(s)||_0^2]\right\}. \tag{3.8}$$

It follows from (2.6), Theorem 2.5, and our initial approximation assumption that

$$\begin{aligned}||\theta_h(0)||_0^2 &= ||\bar\sigma_h(0) - \sigma_h(0)||_0^2 \le ||\bar\sigma_h(0) - \sigma(0)||_0^2\\ &\quad + ||\sigma(0) - \Pi_h\sigma(0)||_0^2 + ||\Pi_h\sigma(0) - \sigma_h(0)||_0^2\\ &\le Ch^{2r}||u_0||_{r+1}^2.\end{aligned} \tag{3.9}$$

If (3.1), (3.6), and (3.9) are substituted into (3.8), then we can obtain

$$||\theta_h||_0^2 \le Ch^{2r}\left\{||u_0||_{r+1}^2 + \int_0^t [||u(s)||_r^2 + ||u_t(s)||_r^2]ds\right\}.$$

Then the proofs of Theorem 3.1 are complete via the triangle inequality. ∎

Remark 3.1. The assumption in the above theorem $\|P_hu_0 - u_h(0)\|_0 \le Ch^r||u_0||_r$ and $||\Pi_h\sigma(0) - \sigma_h(0)||_0 \le Ch^r||u_0||_{r+1}$ is available. In fact, from (2.1) and (2.3) we know that

$$(\alpha(0)(\sigma - \sigma_h)(0),\mathbf{v}_h) + ((u - u_h)(0),\nabla\cdot\mathbf{v}_h) = 0, \qquad \mathbf{v}_h \in \mathbf{V}_h. \tag{3.10}$$

When we choose $u_h(0) = P_hu_0$, (3.10) becomes

$$(\alpha(0)(\sigma - \sigma_h)(0),\mathbf{v}_h) = 0, \qquad \mathbf{v}_h \in \mathbf{V}_h,$$

since $(u_0 - P_hu_0,\nabla\cdot\mathbf{v}_h) = 0$ according to (2.5). Thus, we have by virtue of (2.6) that

$$(\sigma(0)(\sigma_h(0) - \Pi_h\sigma(0)),\mathbf{v}_h) = (\alpha(0)(\sigma(0) - \Pi_h\sigma(0)),\mathbf{v}_h) \le Ch^r||u_0||_{r+1}||\mathbf{v}_h||_0$$

which, together with Lemma 2.4, indicates that

$$||\sigma_h(0) - \Pi_h\sigma(0)||_0 \le Ch^r||u_0||_{r+1}.$$

Remark 3.2. Compared with (2.4) the result presented in Theorem 3.1 is sharper, since the regularity requirement in Theorem 3.1 is one order lower for the pressure field than that in (2.4), which demonstrates that the mixed Ritz–Volterra projection is more favorable for the mixed finite element method of (2.1) than the mixed Ritz projection used to obtain (2.4).

4. Local L^2 superconvergence on rectangular elements. In the last decade considerable attention has been given to the analysis of superconvergence of mixed finite element approximations to elliptic [11, 15, 35, 36] and parabolic [4, 5] problems under various norms associated with the Gauss lines for the gradient and the Gauss points for the solution itself. In this section, we will extend these superconvergence results in mixed finite element approximations to our problem of parabolic integro-differential equations.

Following [15] we assume that $\Omega \subset R^2$ is a rectangle and define seminorms on $\mathbf{V}$ and W as follows. Letting $e = [a,b] \times [c,d] \in T_h$, we denote by $(g_1, g_2, \ldots, g_{k+1})$ the Gauss points in $[a,b]$ and $(\hat{g_1}, \hat{g_2}, \ldots, \hat{g}_{k+1})$ the Gauss points in $[c,d]$, and define

$$|||v_1|||_{1,e}^2 := \sum_{j=1}^{k+1} A_j \frac{d-c}{2} \int_a^b |v_1(s, \hat{g}_i)|^2 ds,$$
$$|||v_2|||_{2,e}^2 := \sum_{j=1}^{k+1} A_j \frac{b-a}{2} \int_c^d |v_2(s, g_i)|^2 ds,$$

where $A_j > 0$, $j = 1, 2, \ldots, k+1$, are the coefficients of the Gauss quadrature rule in $[-1, 1]$. Thus, for $\mathbf{v} = (v_1, v_2) \in \mathbf{V}$ and $w \in W$, we define

$$|||\mathbf{v}|||_*^2 := |||v_1|||_1^2 + |||v_2|||_2^2, \ |||v_i|||_i^2 := \sum_{e \in T_h} |||v_i|||_{i,e}^2, \ i = 1, 2,$$
$$|||w|||_*^2 := \frac{1}{4} \sum_{e \in T_h} \sum_{i,j=1}^{k+1} A_i A_j \ \text{area}(e) |w(g_i, \hat{g}_j)|^2.$$

Clearly, these two seminorms are equal to the L^2-norm of functions from $\mathbf{V}_h$ and W_h, respectively [11, 15], where $\mathbf{V}_h \times W_h$ is the Raviart–Thomas finite element space of index k (≥ 0). Moreover, let u^I represent the interpolation function of u of degree k with respect to x and y, respectively, on each element associated with the $(k+1)^2$ Gauss points. First of all, we need the following lemmas.

LEMMA 4.1. *Assume that $\sigma \in \left(H^{k+2}(\Omega)\right)^2 \cap \mathbf{V}$, $u \in H^{k+2}(\Omega)$, and u^I is the interpolation function of u defined by $(k+1)^2$ Gauss points. Then we have for some constant $C > 0$ that*

$$|||\sigma - \Pi_h \sigma|||_* \leq C h^{k+2} ||\sigma||_{k+2},$$
$$||P_h u - u^I||_0 \leq C h^{k+2} ||u||_{k+2}.$$

Proof. The proof can be found in [11, 15]. ∎

LEMMA 4.2. *Assume that $\sigma \in \left(H^{k+2}(\Omega)\right)^2 \cap \mathbf{V}$, $u \in H^{k+1}(\Omega)$, c and β are two $W^{1,\infty}(\Omega)$ functions. Then we have for some constant $C > 0$ that*

$$|(c(P_h u - u), w_h)| \leq C h^{k+2} ||u||_{k+1} ||w_h||_0, \qquad w_h \in W_h,$$
$$|(\beta(\Pi_h \sigma - \sigma), \mathbf{v}_h)| \leq C h^{k+2} ||\sigma||_{k+2} ||\mathbf{v}_h||_0, \qquad \mathbf{v}_h \in \mathbf{V}_h.$$

Proof. Let $\hat{c} := \displaystyle\int_\Omega c/|\Omega| dx$, where $|\Omega|$ is the measure of Ω. Then

$$|c(x,t) - \hat{c}(x,t)| \leq C h ||c||_{1,\infty}$$

which, together with the definition of the L^2-projection operator P_h, yields

$$\begin{aligned}|(c(P_hu-u),w_h)| &= |((c-\hat{c})(P_hu-u),w_h)|\\ &\le Ch||P_hu-u||_0||w_h||_0\\ &\le Ch^{k+2}||u||_{k+1}||w_h||_0.\end{aligned}$$

Thus, we obtain the first estimate in Lemma 4.2.

The proof for the second estimate is referred to in [11]. ∎

THEOREM 4.3. *Let $(\bar{u}_h,\bar{\sigma}_h)$ be the mixed Ritz–Volterra projection of (u,σ) defined by (2.7). Then there exists a positive constant $C>0$, independent of h, such that, for any $0\le t\le T$,*

$$|||u-\bar{u}_h|||_* + |||\sigma-\bar{\sigma}_h|||_* \le Ch^{k+2}\left(||u||_{k+2}+||\sigma||_{k+2}+\int_0^t||\sigma||_{k+2}ds\right).$$

Proof. We first observe by the equality of the norms $|||\cdot|||_*$ and $||\cdot||_0$ for the functions in the finite element spaces W_h and $\mathbf{V}_h$ that

$$\begin{aligned}|||u-\bar{u}_h|||_* &\le ||||u-P_hu|||_* + ||P_hu-\bar{u}_h||_0,\\ |||\sigma-\bar{\sigma}_h|||_* &\le |||\sigma-\Pi_h\sigma|||_* + ||\Pi_h\sigma-\bar{\sigma}_h||_0.\end{aligned}$$

Since $u-u^I=0$ at the $(k+1)^2$ Gauss points in each element e, we have according to Lemma 4.1 that

$$|||P_hu-u|||_* = |||P_hu-u^I|||_* = ||P_hu-u^I||_0 \le Ch^{k+2}||u||_{k+2}.$$

In addition, from Lemma 4.1 we also know

$$|||\sigma-\Pi_h\sigma|||_* \le Ch^{k+2}||\sigma||_{k+2}.$$

Hence, it is sufficient to bound $||P_hu-\bar{u}_h||_0$ and $||\Pi_h\sigma-\bar{\sigma}_h||_0$ to complete the proof of Theorem 4.3.

Let $\xi:=\Pi_h\sigma-\bar{\sigma}_h$ and $\tau:=P_hu-\bar{u}_h$. Then we see from (2.5) and (2.7) that

$$\begin{aligned}&(\alpha\xi,\mathbf{v}_h)+(\nabla\cdot\mathbf{v}_h,\tau)=F_0(\mathbf{v}_h)+F_1(\mathbf{v}_h), && \mathbf{v}_h\in\mathbf{V}_h,\\ &(\nabla\cdot\xi,w_h)+(c\tau,w_h)=G_0(w_h), && w_h\in W_h,\end{aligned}\tag{4.1}$$

where

$$\begin{aligned}F_0(\mathbf{v}_h) &= -\left(\alpha(\sigma-\Pi_h\sigma)+\int_0^t M(t,s)(\sigma-\Pi_h\sigma)(s)ds,\mathbf{v}_h\right), && \mathbf{v}_h\in\mathbf{V}_h,\\ F_1(\mathbf{v}_h) &= -\left(\int_0^t M(t,s)\xi(s)ds,\mathbf{v}_h\right), && \mathbf{v}_h\in\mathbf{V}_h,\\ G_0(w_h) &= -(c(u-P_hu),w_h), && w_h\in W_h.\end{aligned}$$

Since the terms F_0, F_1, and G_0 can be regarded as linear functionals of $\mathbf{v}_h$ and w_h defined on $\mathbf{V}_h$ and W_h, respectively, we then know from the stability result of [1] that for any fixed time $0\le t\le T$

$$||\xi||_{\mathbf{V}}+||\tau||_W \le C\left\{\sup_{\mathbf{v}_h\in\mathbf{V}_h}\frac{|F_0(\mathbf{v}_h)+F_1(\mathbf{v}_h)|}{||\mathbf{v}_h||_{\mathbf{V}}}+\sup_{w_h\in W_h}\frac{|G_0(w_h)|}{||w_h||_W}\right\}.\tag{4.2}$$

Let

$$F_0(t) = \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{|F_0(\mathbf{v}_h)|}{||\mathbf{v}_h||_{\mathbf{V}}} \quad \text{and} \quad G_0(t) = \sup_{w_h \in W_h} \frac{|G_0(w_h)|}{||w_h||_W}$$

and notice that

$$\sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{|F_1(\mathbf{v}_h)|}{||\mathbf{v}_h||_{\mathbf{V}}} = \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{\left|\left(\int_0^t M(t,s)\xi(s)ds, \mathbf{v}_h\right)\right|}{||\mathbf{v}_h||_{\mathbf{V}}} \le C \int_0^t ||\xi(s)||_{\mathbf{V}} ds.$$

Therefore, we find from (4.2) that

$$||\xi||_{\mathbf{V}} + ||\tau||_W \le C\left(F_0(t) + G_0(t) + C\int_0^t ||\xi(s)||_{\mathbf{V}} ds\right)$$

and by Gronwall's inequality that

$$||\xi||_{\mathbf{V}} + ||\tau||_W \le C(F_0(t) + G_0(t)). \tag{4.3}$$

Now we apply Lemma 4.2 to $F_0(t)$ and $G_0(t)$ to obtain

$$F_0(t) \le Ch^{k+2}\left(||\sigma||_{k+2} + \int_0^t ||\sigma(s)||_{k+2} ds\right) \quad \text{and} \quad G_0(t) \le Ch^{k+2}||u||_{k+1}$$

which, together with (4.3), indicates

$$||\xi||_{\mathbf{V}} + ||\tau||_W \le Ch^{k+2}(||u||_{k+1} + |||\sigma|||_{k+2}). \qquad \square$$

COROLLARY 4.4. *Let $(\bar{u}_h, \bar{\sigma}_h)$ be the mixed Ritz–Volterra projection of (u, σ). Then*

$$\begin{aligned}
&|||D_t(u - \bar{u}_h)|||_* + |||D_t(\sigma - \bar{\sigma}_h)|||_* \\
&\le Ch^{k+2}\left\{||u||_{k+1} + ||u_t||_{k+2} + ||\sigma||_{k+2} + ||\sigma_t||_{k+2} + \int_0^t [||u(s)||_{k+1} + ||\sigma(s)||_{k+2}] ds\right\}.
\end{aligned}$$

Proof. Differentiating (4.1) with respect to time t, then we see that ξ_t and τ_t satisfy the same equations with the right-hand sides replaced by

$$\begin{aligned}
F_0'(\mathbf{v}_h) &= -(\alpha(\sigma_t - \Pi_h\sigma_t) + (\alpha_t + M(t,t))(\sigma - \Pi_h\sigma), \mathbf{v}_h) \\
&\quad + \left(\int_0^t M_t(t,s)(\sigma - \Pi_h\sigma)(s)ds, \mathbf{v}_h\right), && \mathbf{v}_h \in \mathbf{V}_h, \\
F_1'(\mathbf{v}_h) &= -\left(M(t,t)\xi + \int_0^t M_t(t,s)\xi(s)ds, \mathbf{v}_h\right), && \mathbf{v}_h \in \mathbf{V}_h, \\
G_0'(w_h) &= -(c_t(u - P_h u + \tau), w_h) - (c(u - P_h u)_t, w_h), && w_h \in W_h.
\end{aligned}$$

Thus, Corollary 4.4 follows from the same argument above. $\square$

In order to obtain superconvergence results for mixed finite element approximations for our parabolic integro-differential equations we choose our initial data approximation $(u_h(0), \sigma_h(0)) \approx (u_0(x), A(0)\nabla u_0(x))$ as the mixed elliptic projection:

$$\begin{aligned}
&(\alpha(0)(\sigma_h(0) - \sigma(0)), \mathbf{v}_h) + (\nabla \cdot \mathbf{v}_h, u_h(0) - u_0) = 0, && \mathbf{v}_h \in \mathbf{V}_h, \\
&(\nabla \cdot (\sigma_h(0) - \sigma(0)), w_h) + (c(0)(u_h(0) - u_0), w_h) = 0, && w_h \in W_h.
\end{aligned} \tag{4.4}$$

THEOREM 4.5. *Let* (u,σ) *and* (u_h,σ_h) *be the solutions of* (2.1) *and* (2.2), *respectively, and* $(u_h(0),\sigma_h(0))$ *is chosen according to* (4.4). *Then there exists a positive constant* $C>0$ *such that, for any* $0\le t\le T$,

$$
\begin{aligned}
&|||u-u_h|||_* + |||\sigma-\sigma_h|||_* \\
&\le Ch^{k+2}\left\{||u||_{k+2}+||\sigma||_{k+2}+\left[\int_0^t (||u||_{k+1}^2+||\sigma||_{k+2}^2+||u_t||_{k+1}^2+||\sigma_t||_{k+2}^2)ds\right]^{1/2}\right\}.
\end{aligned}
$$

Proof. First, the errors are decomposed as

$$
\begin{aligned}
u-u_h &= (u-\bar{u}_h)+(\bar{u}_h-u_h) := \rho+\rho_h,\\
\sigma-\sigma_h &= (\sigma-\bar{\sigma}_h)+(\bar{\sigma}_h-\sigma_h) := \theta+\theta_h,
\end{aligned}
$$

and then by Theorem 4.3 we have that

$$
|||\rho|||_* + |||\theta|||_* \le Ch^{k+2}(||u||_{k+2}+|||\sigma|||_{k+2}).
$$

Moreover, from (2.7) and (4.4) we derive that

$$
\begin{aligned}
&(\alpha(0)\theta_h(0),\mathbf{v}_h)+(\nabla\cdot\mathbf{v}_h,\rho_h(0))=0, \qquad \mathbf{v}_h\in\mathbf{V}_h,\\
&(\nabla\cdot\theta_h(0),w_h)+(c(0)\rho_h(0),w_h)=0, \qquad w_h\in W_h,
\end{aligned}
$$

which, together with the uniqueness of the solution to (2.7), implies

$$
\theta_h(0)=\rho_h(0)=0. \tag{4.5}
$$

Furthermore, from the proof for Corollary 4.4 we know that

$$
||\tau_t||_0 \le Ch^{k+2}\left\{|||u|||_{k+1}+|||\sigma|||_{k+2}+||u_t||_{k+1}+||\sigma_t||_{k+2}\right\}
$$

which, together with the definition of the local L^2-projection operator P_h, demonstrates that

$$
\begin{aligned}
|(\rho_t,\rho_h)| &= |(\tau_t,\rho_h)|\\
&\le Ch^{k+2}\left\{|||u|||_{k+1}+|||\sigma|||_{k+2}+||u_t||_{k+1}+||\sigma_t||_{k+2}\right\}||\rho_h||_0.
\end{aligned}
$$

Noticing that $|||\rho_h|||_* = ||\rho_h||_0$ and $|||\theta_h|||_* = ||\theta_h||_0$ as well as (4.5), we can obtain the desired estimates for ρ_h and θ_h in L^2-norm through the same procedure as that in Theorem 3.1 for ρ_h and θ_h. ∎

5. Global L^2 superconvergence on quadrilaterals. In [20, 25] superconvergence has been obtained in mixed finite element methods on quadrilaterals for elliptic equations. Here we shall extend these results to our parabolic integro-differential equations. The strategy employed here is that we first examine the superclose accuracy between the interpolation function of the exact solution and the mixed finite element solution of (1.1) by means of integral identities, and then we use a suitable interpolation postprocessing method to obtain global superconvergence approximations [25, 26]. As by-products, these superconvergence results can be utilized to form a class of useful a posteriori error estimators to assess the accuracy of the mixed finite element solutions in applications.

Let $\hat{\mathbf{V}}_h(\hat{e}) \times \hat{W}_h(\hat{e})$ be the standard local Raviart–Thomas rectangular space on the reference element $\hat{e} := [-1,1] \times [-1,1]$ of order k (≥ 0); i.e.,

$$\begin{aligned} \hat{\mathbf{V}}_h(\hat{e}) &:= Q_{k+1,k}(\hat{e}) \times Q_{k,k+1}(\hat{e}), \\ \hat{W}_h(\hat{e}) &:= Q_{k,k}(\hat{e}), \end{aligned}$$

where $Q_{m,n}(\hat{e})$ indicates the space of polynomials of degree no more than m and n in x and y on $\hat{e}$, respectively. On arbitrary convex quadrilateral element $e \in T_h$, the local Raviart–Thomas space is defined by

$$\begin{aligned} \mathbf{V}_h(e) &:= \{\mathbf{q} = G\tilde{\mathbf{q}} \circ \hat{F}_e^{-1} : \ \tilde{\mathbf{q}} \in \hat{\mathbf{V}}_h(\hat{e})\}, \\ W_h(e) &:= \{w = \hat{w} \circ \hat{F}_e^{-1} : \ \hat{w} \in \hat{W}_h(\hat{e})\}, \end{aligned}$$

where $\hat{F}_e$ is the affine map which takes $\hat{e}$ onto e and $G := |\det(M_0)|^{-1} M_0$ with M_0 being the Jacobian matrix (derivative) of $\hat{F}_e$. Of course, $\mathbf{V}_h(e) \subset (C^\infty(e))^2$ and $W_h(e) \subset C^\infty(e)$ are no longer of polynomials on e unless e is a parallelogram.

The global Raviart–Thomas finite element space over the partition T_h is defined in the standard way as follows:

$$\begin{aligned} \mathbf{V}_h &:= \{\mathbf{v} \in H(\mathrm{div};\ \Omega) : \ \mathbf{v}|_e \in \mathbf{V}_h(e) \ \forall e \in T_h\}, \\ W_h &:= \{w \in L^2(\Omega) : \ w|_e \in W_h(e) \ \forall e \in T_h\}. \end{aligned}$$

Let $\tilde{\sigma}$ and $\tilde{u}$ be two vector-valued and scalar-valued functions, respectively, on the reference element $\hat{e}$. Recall that the interpolation functions (or the Raviart–Thomas projection) $\hat{\Pi}_h\tilde{\sigma}$ and $\hat{P}_h\tilde{u}$ over $\hat{e}$ are defined by the following linear systems:

$$\begin{aligned} &\int_{\hat{l}_i} (\tilde{\sigma} - \hat{\Pi}_h\tilde{\sigma}) \cdot \mathbf{n} q ds = 0 && \forall q \in P_k(\hat{l}_i), \qquad i = 1,2,3,4, \\ &\int_{\hat{e}} (\tilde{\sigma} - \hat{\Pi}_h\tilde{\sigma}) \cdot \phi = 0 && \forall \phi \in Q_{k-1,k}(\hat{e}) \times Q_{k,k-1}(\hat{e}), \quad \text{and} \\ &\int_{\hat{e}} (\tilde{u} - \hat{P}_h\tilde{u}) q = 0 && \forall q \in Q_{k,k}(\hat{e}), \quad \text{respectively,} \end{aligned} \tag{5.1}$$

where $\hat{l}_i$ ($i = 1,2,3,4$) is one of the four sides of $\hat{e}$, $\mathbf{n}$ is the outward normal vector to $\hat{e}$, and P_r denotes the set of polynomials of total degree no more than r. If $e \in T_h$ is an arbitrary quadrilateral element, and σ and u are two vector-valued and scalar-valued functions defined on e, then their interpolation functions $\Pi_h\sigma$ and P_hu on e are defined by

$$\Pi_h\sigma := G(\hat{\Pi}_h(G^{-1}\hat{\sigma})) \quad \text{and} \quad P_hu := \hat{P}_h\hat{u}, \quad \text{respectively,} \tag{5.2}$$

where $\hat{\sigma} := \sigma \circ \hat{F}_e$ and $\hat{u} := u \circ \hat{F}_e$. Then we have [20]

$$\begin{aligned} &(\nabla \cdot (\sigma - \Pi_h\sigma), w_h) = 0 && \forall w_h \in W_h, \\ &(\nabla \cdot \mathbf{v}_h, u - P_hu) = 0 && \forall \mathbf{v}_h \in \mathbf{V}_h. \end{aligned} \tag{5.3}$$

The semidiscrete mixed finite element method for (1.1) is now defined as follows: Find $(u_h, \sigma_h) \in W_h \times \mathbf{V}_h$ satisfying

$$\begin{aligned} &(u_{h,t}, w_h) - (\nabla \cdot \sigma_h, w_h) - (cu_h, w_h) = (f, w_h), && w_h \in W_h, \\ &(\alpha\sigma_h, \mathbf{v}_h) + \int_0^t (M(t,s)\sigma_h(s), \mathbf{v}_h) ds + (u_h, \nabla \cdot \mathbf{v}_h) = \langle g, \mathbf{n} \cdot \mathbf{v}_h \rangle, && \mathbf{v}_h \in \mathbf{V}_h, \\ &u_h(0) = P_hu_0, \quad \sigma_h(0) = \Pi_h\sigma(0). \end{aligned} \tag{5.4}$$

From (2.1) and (5.4) we derive the following error equation:

$$\begin{aligned}
&(u_t - u_{h,t}, w_h) - (\nabla\cdot(\sigma-\sigma_h), w_h) - (c(u-u_h), w_h) = 0, && w_h \in W_h,\\
&(\alpha(\sigma-\sigma_h), \mathbf{v}_h) + \int_0^t (M(t,s)(\sigma-\sigma_h)(s), \mathbf{v}_h)ds + (u-u_h, \nabla\cdot\mathbf{v}_h) = 0, && \mathbf{v}_h \in \mathbf{V}_h.
\end{aligned} \tag{5.5}$$

From [20, 25] we recall the following lemmas.

LEMMA 5.1. *If $P_h u$ is the interpolation function of u defined as in* (5.2), *and $c \in W^{1,\infty}(\Omega)$, then there exists a constant C such that*

$$|(c(u-P_h u), w_h)| \le Ch^{k+2}||u||_{k+1}||w_h||_0, \quad w_h \in W_h.$$

LEMMA 5.2. *If the finite element partition T_h is h^2-uniform* [20] *or a generalized rectangular mesh* [25], *and $\Pi_h\sigma$ is the interpolation function of σ defined as in* (5.2), *then there exists a constant C such that for sufficiently smooth β*

$$|(\beta(\sigma-\Pi_h\sigma), \mathbf{v}_h)| \le Ch^{k+2}||\sigma||_{k+2}||\mathbf{v}_h||_0, \quad \mathbf{v}_h \in \mathbf{V}_h.$$

We are now ready to get our main theorem in this section.

THEOREM 5.3. *Assume that the finite element partition T_h is h^2-uniform or generalized rectangular and (u_h, σ_h) is the approximate solution of* (1.1) *defined in* (5.4) *by using quadrilateral elements of Raviart–Thomas of order k. If the exact solution u and σ satisfies $u \in H^{k+1}(\Omega)$, and σ, $\sigma_t \in (H^{k+2}(\Omega))^2$, then we have*

$$||u_h - P_h u||_0 + ||\sigma_h - \Pi_h\sigma||_0 \le Ch^{k+2}\left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2 + ||\sigma_t||_{k+2}^2)ds\right]^{1/2}. \tag{5.6}$$

Proof. Let $\rho_h^* := u_h - P_h u$ and $\theta_h^* := \sigma_h - \Pi_h\sigma$. Then it follows from (5.3) and (5.5) that

$$\begin{aligned}
&(\alpha\theta_h^*, \mathbf{v}_h) + \int_0^t (M(t,s)\theta_h^*(s), \mathbf{v}_h)ds + (\rho_h^*, \nabla\cdot\mathbf{v}_h)\\
&\quad = \left(\alpha(\sigma-\Pi_h\sigma) + \int_0^t M(t,s)(\sigma-\Pi_h\sigma)(s)ds, \mathbf{v}_h\right), && \mathbf{v}_h \in \mathbf{V}_h,\\
&(\rho_{h,t}^*, w_h) - (\nabla\cdot\theta_h^*, w_h) - (c\rho_h^*, w_h) = -(c(u-P_h u), w_h), && w_h \in W_h.
\end{aligned} \tag{5.7}$$

Thus, letting $w_h = \rho_h^*$ and $\mathbf{v}_h = \theta_h^*$ in (5.7) we obtain from Lemmas 2.4, 5.1, and 5.2 as well as the ϵ-type inequality that

$$\frac{1}{2}\frac{d}{dt}||\rho_h^*||_0^2 + ||\theta_h^*||_0^2 \le C\left\{\int_0^t ||\theta_h^*||_0^2 ds + ||\rho_h^*||_0^2 + Ch^{2k+4}(||u||_{k+1}^2 + |||\sigma|||_{k+2}^2)\right\}.$$

Integrating from 0 to t and noticing $\rho_h^*(0) = 0$ yield according to Gronwall's lemma that

$$||\rho_h^*||_0^2 + \int_0^t ||\theta_h^*||_0^2 ds \le Ch^{2k+4}\int_0^t (||u||_{k+1}^2 + |||\sigma|||_{k+2}^2)ds$$

or

$$||\rho_h^*||_0 \le Ch^{k+2}\left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2)ds\right]^{1/2}. \tag{5.8}$$

Following the same steps to get the estimate for $\theta_h := \bar{\sigma}_h - \sigma_h$ in Theorem 3.1 we can also obtain

$$||\theta_h^*||_0 \le Ch^{k+2}\left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2 + ||\sigma_t||_{k+2}^2)ds\right]^{1/2}. \tag{5.9}$$

Combining (5.8) with (5.9) implies (5.6). □

As a by-product of (5.6), we immediately gain the following corollary from the inverse property of the finite element space and the approximation property of the local L^2-projection operator P_h.

COROLLARY 5.4. *Assume that T_h is h^2-uniform or a generalized rectangular mesh and the exact solution u and σ satisfies $u \in W^{k+1,\infty}(\Omega)$ and $\sigma \in (H^{k+2}(\Omega))^2$. Then we have for the mixed finite element solution u_h defined by* (5.4) *that*

$$||u - u_h||_\infty \le Ch^{k+1}\left\{||u||_{k+1,\infty} + \left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2)ds\right]^{1/2}\right\}.$$

In order to improve the accuracy of the finite element approximation to the exact solution on a global scale, a reasonable postprocessing method is proposed according to (5.1) and Theorem 5.3 [25, 26]. For this end, we need to define two postprocessing interpolation operators Π_{2h} and P_{2h} to satisfy

$$\begin{aligned}
&\Pi_{2h}\Pi_h = \Pi_{2h},\\
&||\Pi_{2h}\mathbf{v}_h||_0 \le C||\mathbf{v}_h||_0 && \forall \mathbf{v}_h \in \mathbf{V}_h,\\
&||\Pi_{2h}\sigma - \sigma||_0 \le Ch^{k+2}||\sigma||_{k+2} && \forall \sigma \in (H^{k+2}(\Omega))^2,\\
&P_{2h}P_h = P_{2h},\\
&||P_{2h}w_h||_0 \le C||w_h||_0 && \forall w_h \in W_h,\\
&||P_{2h}u - u||_0 \le Ch^{k+2}||u||_{k+2} && \forall u \in H^{k+2}(\Omega).
\end{aligned} \tag{5.10}$$

For easy exposition, we demonstrate our idea mainly for the case of $k = 2$. Thus, we assume that the standard rectangular partition $\hat{T}_h$ has been obtained from $\hat{T}_{2h} = \{\hat{\tau}\}$ with mesh size $2h$ by subdividing each element of $\hat{T}_{2h}$ into four small congruent rectangles. Let $\hat{\tau} := \bigcup_{i=1}^4 \hat{e}_i$ with $\hat{e}_i \in \hat{T}_h$. Thus, we can define two interpolation operators $\hat{\Pi}_{2h}$ and $\hat{P}_{2h}$ associated with $\hat{T}_{2h}$ of degree at most 3 in x and y on $\hat{\tau}$, respectively, according to the following conditions:

$$\begin{aligned}
&\hat{\Pi}_{2h}\tilde{\sigma}|_{\hat{\tau}} \in (Q_{3,3}(\hat{\tau}))^2, && \hat{P}_{2h}\tilde{u}|_{\hat{\tau}} \in Q_{3,3}(\hat{\tau}),\\
&\int_{\hat{l}_i} (\tilde{\sigma} - \hat{\Pi}_{2h}\tilde{\sigma}) \cdot \mathbf{n} q ds = 0 && \forall q \in P_1(\hat{l}_i),\ i = 1, 2, \ldots, 12,\\
&\int_{\hat{e}_i} (\tilde{\sigma} - \Pi_{2h}\tilde{\sigma}) = 0, && i = 1, 2, 3, 4, \quad \text{and}\\
&\int_{\hat{e}_i} (\tilde{u} - \hat{P}_{2h}\tilde{u})q = 0 && \forall q \in Q_{1,1}(\hat{e}_i),\ i = 1, 2, 3, 4, \quad \text{respectively,}
\end{aligned} \tag{5.11}$$

where $\hat{l}_i$ $(i = 1, 2, \ldots, 12)$ is one of the 12 sides of the four small elements $\hat{e}_i$ $(i = 1, 2, 3, 4)$.

Obviously, the following properties can be easily checked by (5.1) for $k = 2$ and (5.11):

$$
\begin{aligned}
&\hat{\Pi}_{2h}\hat{\Pi}_h = \hat{\Pi}_{2h},\\
&||\hat{\Pi}_{2h}\hat{\mathbf{v}}_h||_0 \le C||\hat{\mathbf{v}}_h||_0 \qquad \forall \hat{\mathbf{v}}_h \in \hat{\mathbf{V}}_h,\\
&||\hat{\Pi}_{2h}\tilde{\sigma} - \tilde{\sigma}||_0 \le Ch^4||\tilde{\sigma}||_4 \qquad \forall \tilde{\sigma} \in (H^4(\Omega))^2,\\
&\hat{P}_{2h}\hat{P}_h = \hat{P}_{2h},\\
&||\hat{P}_{2h}\hat{w}_h||_0 \le C||\hat{w}_h||_0 \qquad \forall \hat{w}_h \in \hat{W}_h,\\
&||\hat{P}_{2h}\tilde{u} - \tilde{u}||_0 \le Ch^4||\tilde{u}||_4 \qquad \forall \tilde{u} \in H^4(\Omega).
\end{aligned} \tag{5.12}
$$

Then we can define two interpolation operators Π_{2h} and P_{2h} associated with T_{2h} by

$$
\Pi_{2h}\sigma := G(\hat{\Pi}_{2h}(G^{-1}\sigma \circ \hat{F}_e)) \quad \text{and} \quad P_{2h}u := \hat{P}_{2h}(u \circ \hat{F}_e), \quad \text{respectively,} \tag{5.13}
$$

which satisfy (5.10) by (5.2) and (5.12). Similarly, we can also define Π_{2h} and P_{2h} for the case of $k \neq 2$.

By virtue of the two interpolation operators Π_{2h} and P_{2h} we immediately gain the following global superconvergence theorem.

THEOREM 5.5. *If there is, besides the conditions of Theorem* 5.3, $u \in H^{k+2}(\Omega)$, *then we have*

$$
\begin{aligned}
&||P_{2h}u_h - u||_0 + ||\Pi_{2h}\sigma_h - \sigma||_0\\
&\quad \le Ch^{k+2}\left\{||u||_{k+2} + ||\sigma||_{k+2} + \left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2 + ||\sigma_t||_{k+2}^2)ds\right]^{1/2}\right\}.
\end{aligned}
$$

Proof. From one of the properties of the operator P_{2h} in (5.10) we find that

$$
P_{2h}u_h - u = P_{2h}(u_h - P_h u) + (P_{2h}u - u).
$$

Therefore, it follows from Theorem 5.3 and (5.10) that

$$
\begin{aligned}
||P_{2h}u_h - u||_0 &\le C||u_h - P_h u||_0 + ||P_{2h}u - u||_0\\
&\le Ch^{k+2}\left\{||u||_{k+2} + \left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2)ds\right]^{1/2}\right\}.
\end{aligned}
$$

Analogously, we can obtain

$$
||\Pi_{2h}\sigma_h - \sigma||_0 \le Ch^{k+2}\left\{||\sigma||_{k+2} + \left[\int_0^t (||u||_{k+1}^2 + ||\sigma||_{k+2}^2 + ||\sigma_t||_{k+2}^2)ds\right]^{1/2}\right\}. \qquad \square
$$

It is of great importance for a mixed finite element method to have a computable a posteriori error estimator by which we can assess the accuracy of the mixed finite element solution in applications. One way to construct error estimators is to employ certain superconvergence properties of the finite element solutions. In fact, we have the following theorem.

THEOREM 5.6. *We have under the conditions of Theorem* 5.5 *that*

$$
||u - u_h||_0 = ||P_{2h}u_h - u_h||_0 + O(h^{k+2}), \tag{5.14}
$$

$$||\sigma - \sigma_h||_0 = ||\Pi_{2h}\sigma_h - \sigma_h||_0 + O(h^{k+2}). \tag{5.15}$$

In addition, if there exist positive constants C_1, C_2 *and small* ϵ_1, $\epsilon_2 \in (0,1)$ *such that*

$$||u - u_h||_0 \geq C_1 h^{k+2-\epsilon_1}, \tag{5.16}$$

$$||\sigma - \sigma_h||_0 \geq C_2 h^{k+2-\epsilon_2}, \tag{5.17}$$

then there hold

$$\lim_{h\to 0} \frac{||u - u_h||_0}{||P_{2h}u_h - u_h||_0} = 1, \tag{5.18}$$

$$\lim_{h\to 0} \frac{||\sigma - \sigma_h||_0}{||\Pi_{2h}\sigma_h - \sigma_h||_0} = 1. \tag{5.19}$$

Proof. It follows from Theorem 5.5 and

$$u - u_h = (P_{2h}u_h - u_h) + (u - P_{2h}u_h)$$

that

$$||u - u_h||_0 = ||P_{2h}u_h - u_h||_0 + O(h^{k+2}).$$

Thus, from (5.16) we know

$$\frac{||P_{2h}u_h - u_h||_0}{||u - u_h||_0} + Ch^{\epsilon_1} \geq 1$$

or

$$\lim_{h\to 0} \frac{||P_{2h}u_h - u_h||_0}{||u - u_h||_0} \geq 1. \tag{5.20}$$

Similarly, it follows from (5.16) and

$$||P_{2h}u_h - u_h||_0 = ||u - u_h||_0 + O(h^{k+2})$$

that

$$\overline{\lim_{h\to 0}} \frac{||P_{2h}u_h - u_h||_0}{||u - u_h||_0} \leq 1$$

which, together with (5.20), leads to (5.18).

Analogously, we can obtain (5.15) and (5.19). ∎

We know from (5.14) that the computable error quantity $||P_{2h}u_h - u_h||_0$ is the principal part of the mixed finite element error $||u - u_h||_0$ and can be used as a reliable a posteriori error indicator to assess the accuracy of the mixed finite element solution under the condition (5.16). Also, (5.16) seems to be a reasonable assumption, since $O(h^{k+1})$ is the optimal convergence rate of the mixed finite element solution in L^2-norm. The same comments are also valid for (5.15) and (5.17).

Acknowledgments. The authors express their thanks to the referees whose comments lead to improvements in the final version of the paper.

REFERENCES

[1] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[2] J. R. Cannon and Y. Lin, *Non-classical H^1 projection and Galerkin methods for nonlinear parabolic integro-differential equations*, Calcolo, 25 (1988), pp. 187–201.

[3] J. R. Cannon and Y. Lin, *A priori L^2 error estimates for finite-element methods for nonlinear diffusion equations with memory*, SIAM J. Numer. Anal., 27 (1990), pp. 595–607.

[4] H. Chen, R. E. Ewing, and R. D. Lazarov, *Superconvergence of mixed finite element methods for parabolic problems with nonsmooth initial data*, Numer. Math., 78 (1998), pp. 495–521.

[5] H. Chen, R. E. Ewing, and R. D. Lazarov, *Superconvergence of the mixed finite element approximations to parabolic equations*, in Advances in Numerical Methods and Applications $O(h^3)$, I. T. Dimov, Bl. Sendov, and P. S. Vassilevski, eds., World Scientific, Singapore, 1994, pp. 63–69.

[6] J. H. Cushman, *Nonlocal dispersion in media with continuously evolving scales of heterogeneity*, Transp. Porous Media, 13 (1993), pp. 123–138.

[7] J. H. Cushman, X. Hu, and F. Deng, *Nonlocal reactive transport with physical and chemical heterogeneity: Localization error*, Water Res. Research, 31 (1995), pp. 2219–2237.

[8] J. H. Cushman, X. Hu, and T. R. Ginn, *Nonequilibrium statistical mechanics of preasymptotic dispersion*, J. Statist. Phys., 75 (1994), pp. 859–878.

[9] J. Douglas, Jr., R. E. Ewing, and M. F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media*, RAIRO Anal. Numér., 17 (1983), pp. 249–265.

[10] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[11] J. Douglas, Jr., and J. Wang, *A superconvergence for mixed finite element methods on rectangular domains*, Calcolo, 26 (1989), pp. 121–134.

[12] R. E. Ewing, *Mathematical modeling and simulation for applications of fluid flow in porous media*, in Current and Future Directions in Applied Mathematics, M. Alber, B. Hu, and J. Rosenthal, eds., Birkhauser, Berlin, Germany, 1997, pp. 161–182.

[13] R. E. Ewing, *The need for multidisciplinary involvement in groundwater contaminant simulations*, in Next Generation Environmental Models and Computational Methods, G. Delic and M. F. Wheeler, eds., SIAM, Philadelphia, 1997, pp. 227–245.

[14] R. E. Ewing, *Aspects of upscaling in formulation of flow in porous media*, Adv. Water Res., 20 (1997), pp. 349–358.

[15] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[16] R. E. Ewing, Y. Lin, and R. Lazarov, *Finite volume element approximations of nonlocal reactive flows in porous media*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 285–311.

[17] R. E. Ewing, Y. Lin, and R. Lazarov, *Finite volume element approximations of nonlocal in time one-dimensional flows in porous media*, Computing, 64 (2000), pp. 157–182.

[18] R. E. Ewing, Y. Lin, and J. Wang, *A numerical approximation of nonFickian flows with mixing length growth in porous media*, Acta Math. Univ. Comenian. (N.S.), 70 (2001), pp. 75–84.

[19] R. E. Ewing, Y. Lin, and J. Wang, *A backward Euler method for mixed finite element approximations of nonFickian flows with non-smooth data in porous media*, submitted.

[20] R. E. Ewing, M. M. Liu, and J. Wang, *Superconvergence of mixed finite element approximations over quadrilaterals*, SIAM J. Numer. Anal., 36 (1999), pp. 772–787.

[21] F. Furtado, J. Glimm, W. Lindquist, and L. F. Pereira, *Characterization of mixing length growth for flow in heterogeneous porous media*, in Proceedings of the 11th SPE Symposium on Reservoir Simulation, Anaheim, CA, 1991, pp. 317–322.

[22] J. Glimm, W. Lindquist, F. Pereira, and Q. Zhang, *A theory of macrodispersion for the scale up problem*, Transp. Porous Media, to appear.

[23] X. Hu, F. Deng, and J. H. Cushman, *Nonlocal reactive transport with physical and chemical heterogeneity: Linear nonequilibrium absorption with random K_d*, Water Res. Research, 31 (1995), pp. 2239–2252.

[24] Z. Jiang, *$L^\infty(L^2)$ and $L^\infty(L^\infty)$ error estimates for mixed methods for integro-differential*

equations of parabolic type, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 531–546.

[25] Q. Lin and N. Yan, *The Construction and Analysis of High Efficiency Finite Element Methods*, Hebei University Publishers, Baoding City, People's Republic of China, 1996.

[26] Q. Lin and S. Zhang, *An immediate analysis for global superconvergence for integro-differential equations*, Appl. Math., 42 (1997), pp. 1–21.

[27] Y. Lin, *On maximum norm estimates for Ritz-Volterra projections and applications to some time-dependent problems*, J. Comput. Math., 15 (1997), pp. 159–178.

[28] Y. Lin, V. Thomée, and L. B. Wahlbin, *Ritz–Volterra projections to finite-element spaces and applications to integrodifferential and related equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1047–1070.

[29] S. Neuman and Y.-K. Zhang, *A quasi-linear theory of nonFickian and Fickian subsurface dispersion* I. *Theoretical analysis with application to isotropic media*, Water Res. Research, 26 (1990), pp. 887–902.

[30] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, New York, 1977, pp. 292–315.

[31] I. H. Sloan and V. Thomée, *Time discretization of an integro-differential equation of parabolic type*, SIAM J. Numer. Anal., 23 (1986), pp. 1052–1061.

[32] S. P. K. Sternberg, J. H. Cushman, and R. A. Greenkorn, *Laboratory observation of nonlocal dispersion*, Transp. Porous Media, 23 (1996), pp. 235–251.

[33] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math. 1054, Springer-Verlag, Berlin, 1984.

[34] V. Thomée and N. Y. Zhang, *Error estimates for semidiscrete finite element methods for parabolic integro-differential equations*, Math. Comp., 53 (1989), pp. 121–139.

[35] J. Wang, *Superconvergence and extrapolation for mixed finite element methods on rectangular domains*, Math. Comp., 56 (1991), pp. 477–503.

[36] J. Wang, *Asymptotic expansions and* L^∞*-error estimates for mixed finite element methods for second order elliptic problems*, Numer. Math., 55 (1989), pp. 401–430.

[37] M. F. Wheeler, *A priori* L_2 *error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1561–1582

NEW LOCKING-FREE MIXED METHOD FOR THE REISSNER–MINDLIN THIN PLATE MODEL*

MOHAMED AMARA†, DANIELA CAPATINA-PAPAGHIUC†, AND AMNA CHATTI‡

Abstract. We are interested here in the Reissner–Mindlin model for a bending thin plate with physical boundary conditions. It is well known that this problem depends singularly upon the plate's thickness ε. By decomposing the bending moment and by dualizing its symmetry, we obtain an equivalent mixed formulation of the initial problem whose unknowns now belong to classical Sobolev spaces. We then propose a low-order conforming finite element method for which we obtain optimal error estimates independently upon the small parameter ε. Thus, the discrete method is unconditionally convergent and locking-free. It directly gives an approximation of the bending moment and allows us to recover the two other variables, which are the deflection and the rotation vector.

Key words. mixed formulation, finite element, error estimates, locking-free

AMS subject classifications. 65N12, 65N15, 65N30

PII. S0036142901385222

1. Introduction and notations.

This paper is devoted to the study of the Reissner–Mindlin model for a bending thin plate satisfying physical boundary conditions.

We have already analyzed in [1] the simpler case of the Kirchhoff–Love problem for a bending plate with natural boundary conditions. We proposed there a conforming piecewise linear finite element method which is unconditionally convergent and which gives an optimal convergence rate whenever the exact solution is sufficiently smooth. One thus gets an approximation of the bending moment in the space $(L^2(\Omega))^4$, while the plate's deflection is approximated in the space $H^1(\Omega)$. To do that, the main idea consists of associating with the symmetric bending tensor a unique vector function, which is not a physical variable but which belongs to a classical Sobolev space $(H^1(\Omega))^2$, and then obtaining and discretizing an equivalent formulation in this new unknown.

The aim of the present work is to generalize to the Reissner–Mindlin case the approach that was used in [1] for the Kirchhoff–Love model.

The equations of the Reissner–Mindlin model depend upon a small parameter ε characterizing the plate's thickness. It is well known that this is a singular problem with respect to ε; in the limit case $\varepsilon = 0$ one obtains the Kirchhoff–Love model previously studied in [1]. One of the major difficulties in discretizing the Reissner–Mindlin problem consists of finding a finite element approximation which does not suffer from numerical locking as the plate's thickness becomes very small. For a general presentation of this phenomenon, one may see [3], for instance.

The physical boundary conditions imposed in this paper represent an additional difficulty from both the theoretical and the numerical point of view. Indeed, even if there are many papers in the literature dealing with the approximation of the Reissner–Mindlin model, the great majority of them consider simple Dirichlet bound-

*Received by the editors February 15, 2001; accepted for publication (in revised form) April 2, 2002; published electronically October 23, 2002.
http://www.siam.org/journals/sinum/40-4/38522.html
†Laboratoire de Mathématiques Appliquées, Université de Pau, 64000 Pau, France (mohamed.amara@univ-pau.fr, daniela.capatina-papaghiuc@univ-pau.fr).
‡LAMSIN, Ecole Nationale d'Ingénieurs, 2002 Tunis, Tunisia (amna_chatti@yahoo.fr).

ary conditions for the deflection and the rotation. That means that the plate is supposed to be clamped and in this case one usually eliminates the bending moment from the equations and computes only the deflection and the rotation. This approach is clearly no longer possible when dealing with complex boundary conditions like ours. Moreover, we are interested in obtaining a good approximation of the bending tensor, since in practice it usually represents the quantity of interest for the engineers. Thus, our formulation is new and explicitly takes into account the boundary conditions satisfied by the bending moment.

The approach developed here is based on the same idea as in [1], which is the decomposition of the bending moment by the means of Tartar's lemma (see, for instance, [8]). However, its symmetry is no longer imposed but is dualized by the means of a Lagrange multiplier. Therefore, we now associate with the bending tensor a couple of functions belonging to $(H^1(\Omega))^2 \times H^1(\Omega)$. This finally allows us to obtain an equivalent mixed formulation of the initial problem, whose operator can be written as follows:

$$\begin{pmatrix} A+\varepsilon^2 A_0 & B & C \\ B^T & O & O \\ C^T & O & O \end{pmatrix}, \tag{1.1}$$

where A and B are the same as in the Kirchhoff–Love case. The bilinear form B takes into account the boundary conditions imposed on the bending moment, A_0 takes into account the additional unknown of the Reissner–Mindlin model, that is, the rotation vector, and C dualizes the symmetry of the bending tensor.

Let us remark here that this kind of operator is not typical for such a singularly perturbed problem. Indeed, the mixed formulations usually employed (cf., for instance, [5], and references therein) are obtained by dualizing the constraint imposed in the limit case $\varepsilon = 0$ (that means, in our case, the constraint $\mathbf{r} = \nabla u$ imposed in the Kirchhoff-Love model). Their operator then writes as

$$\begin{pmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{B}^T & \varepsilon^2 \mathcal{C} \end{pmatrix},$$

so our approach in order to avoid the shear locking phenomenon is different.

Before proposing our approximation method, let us first point out the solutions generally proposed in the literature in order to get a locking-free discretization of the Reissner–Mindlin model. However, let us specify that all these methods are introduced for clamped plates and they do not apply to complex boundary conditions. For an exhaustive presentation of the existing results and for recent references on the Reissner–Mindlin model, the reader may see [7].

One of the most commonly used approach consists of writing a mixed formulation as above. But generally it is difficult to find simple and cheap finite element spaces for which the theory of Babuška–Brezzi holds. In practice, this leads to modifying certain operators by the means of reduced integration techniques, or to adding a stabilization term, or to using nonconforming finite elements (eventually enriched with bubble functions). Another solution is to write an equivalent formulation by the means of the Helmholtz decomposition and of two additional unknowns, as in [2] and in [5]. One may also work with the standard variational formulation and employ for the discretization more expensive continuous finite elements (see [11]) or nonconforming elements. Finally, one may employ p or hp methods (see, for instance, [11]), which are known to work well for problems concerned with locking, or use least squares methods, as in [4].

An outline of the paper is as follows. We begin by introducing the boundary value problem and by giving a mathematical framework in which it is well-posed. In section 3 we state a mixed variational formulation whose operator writes as in (1.1) and we show, thanks to the Babuška–Brezzi theory, that this problem is well-posed. Moreover, its main unknown is exactly the bending moment, while the dual unknowns are the displacement's trace and the rotation's normal trace (corresponding to the bilinear form B), respectively, an additional multiplier which dualizes the symmetry of the bending moment (corresponding to C). However, the test-functions corresponding to $A + \varepsilon^2 A_0$ have to satisfy the constraint $\mathrm{div}\mathbf{div}(\cdot) = 0$. In order to avoid its discretization, we associate with the bending moment a unique couple of functions now belonging to classical spaces, and we study an equivalent mixed formulation in these new variables. This is done in section 4, while in section 5 we rigorously establish the link between the two formulations of the problem. Finally, section 6 is devoted to the numerical approximation. We discretize the last saddle point problem by classical finite elements (continuous P_1 and P_2, discontinuous P_0), for which we prove a discrete inf-sup condition uniformly with respect to both h and ε. This insures the unconditional convergence of the method independently of the small parameter ε as well as optimal error estimates. Next, one gets uniform approximations of the initial physical variables in the following spaces: the bending moment in $(H(\mathrm{div};\Omega))^2$ endowed with the weighted-norm $\|\cdot\|_{0,\Omega} + \varepsilon \|\mathbf{div}(\cdot)\|_{0,\Omega}$, the transverse displacement in $H^1(\Omega)$, and the rotation $\mathbf{r}^\varepsilon$ in $L^2(\Omega)^2$, with also an approximation of $curl\mathbf{r}^\varepsilon$ in $L^2(\Omega)$. Let us also notice that the discrete bending moment and rotation are given by simple formulae, while the discrete deflection is obtained by solving a Laplace problem.

As a conclusion, we propose here a well-posed formulation which takes into account the physical boundary conditions imposed in the Reissner–Mindlin model. For any fixed ε, its discretization by simple low-order finite elements is shown to be unconditionally convergent and, moreover, locking-free. If the exact solution is sufficiently smooth, then an optimal convergence rate $O(h)$ is obtained.

2. Physical Reissner–Mindlin model. Let us begin this paragraph by introducing some notations which will be used in what follows. We note that $\mathbf{n} = (n_i)_{1\le i\le 2}$, the unit outward normal vector along Γ, and that $\mathbf{t} = (t_i)_{1\le i\le 2}$, the unit tangent vector to Γ oriented such that $t_1 = n_2$, $t_2 = -n_1$. We also employ in this paper the summation convention of Einstein, and we denote by the letter c any positive constant independent upon both the discretization parameter h and the plate's thickness ε. We agree to write the vectors in bold letters and the tensors in underlined letters. Let us also recall here some classical notation: for any vector function $\mathbf{v}$ and any scalar function v we note that

$$curl\,\mathbf{v} = \partial_1 v_2 - \partial_2 v_1, \qquad \mathrm{div}\,\mathbf{v} = \partial_i v_i,$$

$$\mathbf{curl}\,v = \begin{pmatrix} \partial_2 v \\ -\partial_1 v \end{pmatrix}, \quad \underline{curl}\,\mathbf{v} = \begin{pmatrix} \partial_2 v_1 & -\partial_1 v_1 \\ \partial_2 v_2 & -\partial_1 v_2 \end{pmatrix}, \quad \mathbf{div}\underline{\tau} = \begin{pmatrix} \partial_1\tau_{11} + \partial_2\tau_{12} \\ \partial_1\tau_{21} + \partial_2\tau_{22} \end{pmatrix},$$

and we equally put that

$$\underline{\nabla}\mathbf{v} = \begin{pmatrix} \partial_1 v_1 & \partial_2 v_1 \\ \partial_1 v_2 & \partial_2 v_2 \end{pmatrix}, \quad \underline{\varepsilon}(\mathbf{v}) = \frac{1}{2}(\underline{\nabla}\mathbf{v} + {}^T\underline{\nabla}\mathbf{v}), \quad \underline{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \underline{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

We denote the tangential and the normal derivative on the boundary of a scalar function v by, respectively,

$$\partial_t v = \nabla v \cdot \mathbf{t} \quad \text{and} \quad \partial_n v = \nabla v \cdot \mathbf{n}.$$

As usual, let us denote by Ω the medium surface of the plate and by 2ε the plate's thickness. In view of the finite element discretization, we suppose that Ω is a connected polygonal domain of $\mathbb{R}^2$. The hypothesis of connectivity is not essential but permits an easier presentation of the method. We also suppose that its boundary Γ is decomposed into three disjoint parts $\Gamma = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$, and on each one different boundary conditions are imposed. We consider here the case of linear elasticity and, for the sake of simplicity, the constitutive material is taken as homogeneous and isotropic. For technical reasons, we suppose that Ω has no cuts and that $m(\Gamma_0) > 0$, where $m(\Gamma_0)$ represents the measure of Γ_0.

Let us first recall the equations describing the Kirchhoff–Love problem analyzed in [1], which write as below:

$$\begin{cases} \operatorname{div}(\mathbf{div}\underline{\sigma}) = f & \text{in } \Omega, \\ u = \partial_n u = 0 & \text{on } \Gamma_0, \\ u = 0, \quad \underline{\sigma}\mathbf{n}\cdot\mathbf{n} = 0 & \text{on } \Gamma_1, \\ \underline{\sigma}\mathbf{n}\cdot\mathbf{n} = \partial_t(\underline{\sigma}\mathbf{n}\cdot\mathbf{t}) + \mathbf{div}\underline{\sigma}\cdot\mathbf{n} = 0 & \text{on } \Gamma_2, \\ \sigma_{ij} = (1-\nu)\partial_{ij}u + \nu\Delta u\delta_{ij} & \text{in } \Omega, \end{cases} \tag{2.1}$$

where ν is Poisson's coefficient. The function f represents the force density of the applied transverse loading and belongs to the space $L^2(\Omega)$. The boundary conditions mean that the plate is clamped on Γ_0, simply supported on Γ_1, while Γ_2 is a free boundary.

Then we consider, cf. [6], the following equations for the Reissner–Mindlin problem:

$$\begin{cases} -\dfrac{1}{1-\nu}\mathbf{div}\underline{\sigma}^\varepsilon + \dfrac{1}{\varepsilon^2}(\mathbf{r}^\varepsilon - \nabla u^\varepsilon) = 0 & \text{in } \Omega, \\ \dfrac{1-\nu}{\varepsilon^2}\operatorname{div}(\mathbf{r}^\varepsilon - \nabla u^\varepsilon) = f & \text{in } \Omega, \\ u^\varepsilon = 0, \quad \mathbf{r}^\varepsilon = 0 & \text{on } \Gamma_0, \\ u^\varepsilon = 0, \quad \mathbf{r}^\varepsilon\cdot\mathbf{t} = 0, \quad \underline{\sigma}^\varepsilon\mathbf{n}\cdot\mathbf{n} = 0 & \text{on } \Gamma_1, \\ \mathbf{r}^\varepsilon\cdot\mathbf{t} = \partial_t u^\varepsilon, \quad \underline{\sigma}^\varepsilon\mathbf{n}\cdot\mathbf{n} = \partial_t(\underline{\sigma}^\varepsilon\mathbf{n}\cdot\mathbf{t}) + \mathbf{div}\underline{\sigma}^\varepsilon\cdot\mathbf{n} = 0 & \text{on } \Gamma_2, \\ \sigma_{ij}^\varepsilon = (1-\nu)\,\varepsilon_{ij}(\mathbf{r}^\varepsilon) + \nu(\operatorname{div}\mathbf{r}^\varepsilon)\delta_{ij} & \text{in } \Omega. \end{cases} \tag{2.2}$$

The unknowns of the problem are the transverse displacement of the plate u^ε, the rotation $\mathbf{r}^\varepsilon$ of the unit normal to the medium surface, and the bending moment $\underline{\sigma}^\varepsilon$. Obviously, they depend upon the small parameter ε, which characterizes the plate's thickness.

Let us notice that (2.2) is a singular problem with respect to the small parameter ε. We remark here that, in the simpler case of a clamped plate (i.e., $\Gamma = \Gamma_0$) whose medium surface Ω is a convex domain, one has the following regularity result (see, for instance, [5]):

$$\mathbf{r}^\varepsilon \in (H^2(\Omega))^2, \quad u^\varepsilon \in H^2(\Omega),$$
$$\|\mathbf{r}^\varepsilon\|_{2,\Omega} + \|u^\varepsilon\|_{2,\Omega} + \varepsilon\,\|\mathbf{div}\underline{\sigma}^\varepsilon\|_{1,\Omega} \le c\,\|f\|_{0,\Omega}\,.$$

One cannot improve the above estimate for the term $\mathbf{div}\underline{\sigma}^\varepsilon$, even for a smoother domain and a smoother loading f: indeed, $\mathbf{div}\underline{\sigma}^\varepsilon$ is not uniformly bounded in $(H^1(\Omega))^2$ because of boundary layers in the Reissner–Mindlin model. This represents another source of difficulties in obtaining an approximation method uniformly convergent with respect to ε.

Let us now come back to the above two models of plate, and particularly to the boundary conditions considered in (2.2). With this choice, it is known (cf. Destuynder

and Salaun [6]) that as ε tends towards 0, the Reissner–Mindlin model (2.2) tends towards the Kirchhoff–Love model (2.1) in the following sense:

$$u^\varepsilon \overset{H^1(\Omega)}{\longrightarrow} u, \qquad \mathbf{r}^\varepsilon \overset{H^1(\Omega)}{\longrightarrow} \mathbf{r} = \nabla u,$$
$$\underline{\sigma}^\varepsilon \overset{L^2(\Omega)}{\longrightarrow} \underline{\sigma}, \qquad \mathbf{div}\underline{\sigma}^\varepsilon \overset{H^{-1}(\Omega)}{\longrightarrow} \mathbf{div}\underline{\sigma}.$$

Remark 2.1. One may consider other boundary conditions in the Reissner–Mindlin case, for instance, cf. [6]:

$$\begin{cases} u^\varepsilon = 0, \quad \mathbf{r}^\varepsilon = 0 & \text{on } \Gamma_0, \\ u^\varepsilon = 0, \quad \mathbf{r}^\varepsilon \cdot \mathbf{t} = 0, \quad \underline{\sigma}^\varepsilon \mathbf{n} \cdot \mathbf{n} = 0 & \text{on } \Gamma_1, \\ \underline{\sigma}^\varepsilon \mathbf{n} \cdot \mathbf{n} = \underline{\sigma}^\varepsilon \mathbf{n} \cdot \mathbf{t} = 0 & \text{on } \Gamma_2. \end{cases}$$

In this way, the Reissner–Mindlin model is obtained from the Kirchhoff–Love one by a penalty method. But then, due to the fact that the conditions on the free boundary Γ_2 are different in the two plate models, the convergence of the rotation vector $\mathbf{r}^\varepsilon$ towards ∇u will hold only with respect to the $L^2(\Omega)$-norm.

In order to write the variational formulation of (2.2), we introduce the Hilbert spaces

$$\underline{X} = \left\{ \underline{\tau} \in L^2(\Omega)^4 ;\ D(\underline{\tau}) \in L^2(\Omega) \right\},$$
$$\underline{X}^\varepsilon = \left\{ \underline{\tau} \in L^2(\Omega)^4 ;\ \varepsilon\, \mathbf{div}\underline{\tau} \in (L^2(\Omega))^2,\ D(\underline{\tau}) \in L^2(\Omega) \right\}$$

endowed with their natural norms

$$\| \underline{\tau} \|_{\underline{X}} = (\| \underline{\tau} \|_{0,\Omega}^2 + \| D(\underline{\tau}) \|_{0,\Omega}^2)^{1/2},$$
$$\| \underline{\tau} \|_{\underline{X}^\varepsilon} = (\| \underline{\tau} \|_{0,\Omega}^2 + \varepsilon^2 \| \mathbf{div}\underline{\tau} \|_{0,\Omega}^2 + \| D(\underline{\tau}) \|_{0,\Omega}^2)^{1/2}.$$

We recall that $\underline{X}$ is the space introduced in [1] for the analysis of the Kirchhoff–Love model, and the operator D is defined as follows: $D(\underline{\tau}) = \mathrm{div}(\mathbf{div}\underline{\tau}) = \partial_{ij}\tau_{ij}$. Then one can establish (see [1] for a complete proof) that $(\mathcal{D}(\overline{\Omega}))^4$ is a dense subspace of $\underline{X}$ and the trace operators

$$\gamma_0 : ((\mathcal{D}(\overline{\Omega}))^4, \|\cdot\|_{\underline{X}}) \longrightarrow H^{-1/2}(\Gamma), \qquad \gamma_0(\underline{\tau}) = \underline{\tau}\mathbf{n} \cdot \mathbf{n},$$
$$\gamma_1 : ((\mathcal{D}(\overline{\Omega}))^4, \|\cdot\|_{\underline{X}}) \longrightarrow H^{-3/2}(\Gamma), \qquad \gamma_1(\underline{\tau}) = \partial_t(\underline{\tau}\mathbf{n} \cdot \mathbf{t}) + \mathbf{div}\underline{\tau} \cdot \mathbf{n},$$

are linear and continuous, so they can be extended by continuity on $\underline{X}$. Moreover, for any $v \in H^2(\Omega)$ and any $\underline{\tau} \in \underline{X}$, one has the following Green's formula:

$$\int_\Omega D(\underline{\tau}) v \, d\Omega = \int_\Omega \tau_{ij}\partial_{ij} v \, d\Omega - \langle \gamma_0(\underline{\tau}), \partial_n v \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \gamma_1(\underline{\tau}), v \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}.$$

It is now obvious, since $\underline{X}^\varepsilon \subset \underline{X}$, that the operators γ_0 and γ_1 are well defined on $\underline{X}^\varepsilon$ by simply taking their restriction.

Remark 2.2. One clearly has $\underline{X}^\varepsilon = \underline{X}^{\varepsilon'}$ algebraically when $\varepsilon \neq 0$, $\varepsilon' \neq 0$. We choose this notation in order to preserve a parallelism between the formulations of the Reissner–Mindlin and Kirchhoff–Love models, respectively. We also have

$$\underline{X}^\varepsilon \subset (H(\mathrm{div};\Omega))^2, \qquad \mathbf{div}\underline{X}^\varepsilon \subset H(\mathrm{div};\Omega).$$

For any $f \in L^2(\Omega)$, we introduce

$$\underline{X}^{\varepsilon,f} = \{ \underline{\tau} \in \underline{X}^\varepsilon ;\ D(\underline{\tau}) = f \},$$

and we can easily show that $\underline{X}^{\varepsilon,f} \neq \emptyset$.

3. First formulation with respect to the bending moment. We derive in this section a mixed variational formulation of problem (2.2), whose main unknown will be the bending moment $\underline{\sigma}^\varepsilon$ which is a symmetric second-order tensor.

Let us put

$$M = \left\{ v \in H^{3/2}(\Gamma);\ v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \right\},$$
$$N = \left\{ v \in H^{1/2}(\Gamma);\ v = 0 \text{ on } \Gamma_0 \right\},$$

and let us introduce the bilinear forms $a^\varepsilon(\cdot,\cdot)$, $b(\cdot,\cdot)$, and $c(\cdot,\cdot)$, defined on $\underline{X}^\varepsilon \times \underline{X}^\varepsilon$, on $\underline{X}^\varepsilon \times (M \times N)$, and on $\underline{X}^\varepsilon \times L^2(\Omega)$, respectively, by the following relations:

$$a^\varepsilon(\underline{\sigma},\underline{\tau}) = a(\underline{\sigma},\underline{\tau}) + \varepsilon^2\, a_0(\underline{\sigma},\underline{\tau}),$$
$$b(\underline{\tau},(\zeta,\eta)) = \langle \gamma_1(\underline{\tau}),\zeta\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} - \langle \gamma_0(\underline{\tau}),\eta\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma},$$
$$c(\underline{\tau},\mu) = \int_\Omega (\tau_{12} - \tau_{21})\,\mu\, d\Omega,$$

where

$$a(\underline{\sigma},\underline{\tau}) = \frac{1}{1-\nu}\int_\Omega \underline{\sigma} : \underline{\tau}\, d\Omega - \frac{\nu}{1-\nu^2}\int_\Omega (tr\underline{\sigma})(tr\underline{\tau})\, d\Omega,$$
$$a_0(\underline{\sigma},\underline{\tau}) = \frac{1}{1-\nu}\int_\Omega \mathbf{div}\underline{\sigma}\cdot\mathbf{div}\underline{\tau}\, d\Omega.$$

The forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ as well as the spaces M, N are exactly the same as the ones employed in the Kirchhoff–Love formulation; see [1]. One can immediately notice that if $\underline{\sigma} \in \underline{X}^\varepsilon$ is symmetric, then

$$\forall \mu \in L^2(\Omega), \quad c(\underline{\sigma},\mu) = 0.$$

So, the role of the new term $c(\cdot,\cdot)$ is to dualize the symmetry of the bending tensor, while the bilinear form $\varepsilon^2\, a_0(\cdot,\cdot)$ takes into account the new variable which is the rotation vector.

We now propose the following variational formulation for the Reissner–Mindlin problem:

$$\text{(3.1)} \qquad \begin{cases} \text{find } \underline{\sigma}^\varepsilon \in \underline{X}^{\varepsilon,f},\ (u_0^\varepsilon, r_0^\varepsilon) \in M \times N,\ \lambda^\varepsilon \in L^2(\Omega) \text{ such that} & \\ \forall \underline{\tau} \in \underline{X}^{\varepsilon,0}, & a^\varepsilon(\underline{\sigma}^\varepsilon,\underline{\tau}) + b(\underline{\tau},(u_0^\varepsilon,r_0^\varepsilon)) + c(\underline{\tau},\lambda^\varepsilon) = 0, \\ \forall (\zeta,\eta) \in M \times N, & b(\underline{\sigma}^\varepsilon,(\zeta,\eta)) = 0, \\ \forall \mu \in L^2(\Omega), & c(\underline{\sigma}^\varepsilon,\mu) = 0, \end{cases}$$

which is a generalization of the one introduced in the Kirchhoff–Love case.

In a quite similar way, we show the following theorem.

THEOREM 3.1. *Problem* (3.1) *has a unique solution.*

Proof. It comes from the Babuška–Brezzi theory (see, for instance, [5]). Indeed, the bilinear form $a(\cdot,\cdot)$ is $(L^2(\Omega))^4$-elliptic since $(tr\underline{\tau})^2 \le 2\underline{\tau} : \underline{\tau}$. Therefore, for all $\underline{\tau} \in \underline{X}^{\varepsilon,0}$, one deduces the $\underline{X}^{\varepsilon,0}$-ellipticity of $a^\varepsilon(\cdot,\cdot)$:

$$a^\varepsilon(\underline{\tau},\underline{\tau}) \ge c(\| \underline{\tau} \|_{0,\Omega}^2 + \varepsilon^2 \| \mathbf{div}\underline{\tau} \|_{0,\Omega}^2) = c \| \underline{\tau} \|_{\underline{X}^\varepsilon}^2 .$$

It suffices now to establish the inf-sup condition for the bilinear form $d(\cdot,\cdot)$ obtained by adding the last two equations of (3.1):

$$\forall \underline{\tau} \in \underline{X}^{\varepsilon,0},\ \forall (\zeta,\eta,\mu) \in M \times N \times L^2(\Omega), \quad d(\underline{\tau},(\zeta,\eta,\mu)) = b(\underline{\tau},(\zeta,\eta)) + c(\underline{\tau},\mu).$$

We apply a classical idea, that is, we associate with any $(\zeta,\eta,\mu)\in M\times N\times L^2(\Omega)$ a tensor $\underline{\tau}\in\underline{X}^{\varepsilon,0}$ such that

$$\begin{cases} d(\underline{\tau},(\zeta,\eta,\mu))\geq c(\|\mu\|_{0,\Omega}+\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma})^2 \\ \|\,\underline{\tau}\,\|_{\underline{X}^\varepsilon}\leq c(\|\mu\|_{0,\Omega}+\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma}). \end{cases} \tag{3.2}$$

However, the construction of such a tensor $\underline{\tau}$ is quite technical.

We begin by finding $\underline{\tau}_1$ such that the inf-sup condition for $b(\cdot,\cdot)$ holds. For that, with any couple $(\zeta,\eta)\in M\times N$ we associate the function $\mathbf{q}=(\partial_t\zeta)\,\mathbf{t}+\eta\,\mathbf{n}$. Clearly, we have that $\mathbf{q}\in(H^{1/2}(\Gamma))^2$ and also, since $m(\Gamma_0)\neq 0$, that

$$\|\mathbf{q}\|_{1/2,\Gamma}\leq c(\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma}).$$

Indeed, this comes by noticing that $\mathbf{q}=\nabla w$ on Γ, where w is the unique solution of

$$\begin{cases} \Delta^2 w=0 & \text{in } \Omega,\\ w=\zeta & \text{on } \Gamma,\\ \partial_n w=\eta & \text{on } \Gamma. \end{cases}$$

Let us next consider the following auxiliary problem:

$$\begin{cases} \Delta\omega=0 & \text{in } \Omega,\\ \omega=\mathbf{q} & \text{on } \Gamma, \end{cases}$$

whose solution $\omega\in(H^1(\Omega))^2$ verifies $|\omega|_{1,\Omega}\leq c\,\|\mathbf{q}\|_{\frac12,\Gamma}$. By choosing $\underline{\tau}_1=-\underline{\nabla}\omega$, one obviously has $\mathbf{div}\underline{\tau}_1=0$ as well as $D(\underline{\tau}_1)=0$, so one can write, thanks to an integration by parts, that

$$\begin{aligned} b(\underline{\tau}_1,(\zeta,\eta)) &= \langle\partial_t(\underline{\tau}_1\mathbf{n}\cdot\mathbf{t}),\zeta\rangle_{-\frac32,\frac32,\Gamma}-\langle\underline{\tau}_1\mathbf{n}\cdot\mathbf{n},\eta\rangle_{-\frac12,\frac12,\Gamma}\\ &= -\langle\underline{\tau}_1\mathbf{n}\cdot\mathbf{t},\partial_t\zeta\rangle_{-\frac12,\frac12,\Gamma}-\langle\underline{\tau}_1\mathbf{n}\cdot\mathbf{n},\eta\rangle_{-\frac12,\frac12,\Gamma}\\ &= -\langle\underline{\tau}_1\mathbf{n},\omega\rangle_{-\frac12,\frac12,\Gamma}=|\omega|^2_{1,\Omega}. \end{aligned}$$

So we obtained, for this choice of the tensor $\underline{\tau}_1\in\underline{X}^{\varepsilon,0}$, that

$$b(\underline{\tau}_1,(\zeta,\eta))\geq c(\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma})^2 \quad\text{and}\quad \|\,\underline{\tau}_1\,\|_{\underline{X}^\varepsilon}=|\omega|_{1,\Omega}\leq c(\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma}).$$

Next, we define $\underline{\tau}_2\in\underline{X}^{\varepsilon,0}$ such that $\underline{\tau}=\underline{\tau}_1+\underline{\tau}_2$ satisfies the relations (3.2). The idea is to construct a tensor $\underline{\tau}_2$ with vanishing traces such that $b(\underline{\tau}_2,(\zeta,\eta))=0$, so we will only have to check the inf-sup condition for $c(\cdot,\cdot)$ now. For any $\mu\in L^2(\Omega)$, let us put

$$P(\mu)=\frac{1}{m(\Omega)}\int_\Omega\mu\,d\Omega$$

and consider $\lambda=\mu-P(\mu)-curl\omega$. The fact that

$$\int_\Omega curl\omega\,d\Omega=-\int_\Gamma\omega\cdot\mathbf{t}\,d\Gamma=-\int_\Gamma\partial_t\zeta\,d\Gamma=0$$

implies that λ belongs to $L^2_0(\Omega)=\left\{v\in L^2(\Omega);\ \int_\Omega v\,d\Omega=0\right\}$ and, moreover, that

$$\|\lambda\|_{0,\Omega}\leq c\,\|\mu\|_{0,\Omega}+|\omega|_{1,\Omega}\leq c(\|\mu\|_{0,\Omega}+\|\eta\|_{1/2,\Gamma}+\|\zeta\|_{3/2,\Gamma}).$$

Then there exists, cf. [8], a function $\mathbf{v} \in (H_0^1(\Omega))^2$ such that

$$\operatorname{div} \mathbf{v} = \lambda \text{ in } \Omega \qquad \text{and} \qquad |\mathbf{v}|_{1,\Omega} \leq c \, \|\lambda\|_{0,\Omega} .$$

Therefore, the new function

$$\varphi = \mathbf{v} + \frac{P(\mu)}{2} \begin{pmatrix} x \\ y \end{pmatrix}$$

satisfies the two conditions

$$\begin{aligned} &\operatorname{div} \varphi = \mu - curl\omega \qquad \text{in } \Omega, \\ &|\varphi|_{1,\Omega} \leq |\mathbf{v}|_{1,\Omega} + c \, \|\mu\|_{0,\Omega} \\ &\qquad \leq c(\|\mu\|_{0,\Omega} + \|\eta\|_{1/2,\Gamma} + \|\zeta\|_{3/2,\Gamma}). \end{aligned}$$

The boundary Γ being polygonal, one has that $\partial_t \varphi = c\mathbf{t}$, with $c = \frac{P(\mu)}{2}$, which implies that

$$\partial_t \varphi \cdot \mathbf{n} = 0, \qquad \partial_t(\partial_t \varphi \cdot \mathbf{t}) = 0 \qquad \text{on } \Gamma.$$

By next choosing that $\underline{\tau}_2 = -\underline{curl}\, \varphi \in \underline{X}^{\varepsilon,0}$, one gets

$$\begin{aligned} c(\underline{\tau}_2, \mu) &= \int_\Omega \operatorname{div} \varphi \, \mu \, d\Omega = \|\mu\|_{0,\Omega}^2 - c(\underline{\tau}_1, \mu) \, d\Omega, \\ \gamma_0(\underline{\tau}_2) &= -\partial_t \varphi \cdot \mathbf{n} = 0, \\ \gamma_1(\underline{\tau}_2) &= -\partial_t(\partial_t \varphi \cdot \mathbf{t}) = 0. \end{aligned}$$

Finally, for the tensor $\underline{\tau} = \underline{\tau}_1 + \underline{\tau}_2 \in \underline{X}^{\varepsilon,0}$ with $\mathbf{div}\underline{\tau} = 0$, we get (3.2), which means that the inf-sup condition for $d(\cdot,\cdot)$ uniformly holds with respect to ε. ∎

The next result gives the interpretation of the solution of (3.1) in terms of the Reissner–Mindlin boundary value problem (2.2).

THEOREM 3.2. *Let $(\underline{\sigma}^\varepsilon, (u_0^\varepsilon, r_0^\varepsilon), \lambda^\varepsilon)$ be the solution of* (3.1). *Then $\underline{\sigma}^\varepsilon$ represents the bending moment calculated by the Reissner–Mindlin model* (2.2), *and one equally has that*

$$\begin{cases} r_0^\varepsilon = \mathbf{r}^\varepsilon \cdot \mathbf{n} & \text{on } \Gamma, \\ u_0^\varepsilon = u^\varepsilon & \text{on } \Gamma, \\ \lambda^\varepsilon = -\frac{1}{2} curl \mathbf{r}^\varepsilon & \text{in } \Omega, \end{cases} \tag{3.3}$$

where $(\underline{\sigma}^\varepsilon, \mathbf{r}^\varepsilon, u^\varepsilon)$ satisfies the equations in (2.2).

Proof. The third equation of the variational problem (3.1) gives that $\underline{\sigma}^\varepsilon$ is symmetric, while the second equation implies that $\gamma_1(\underline{\sigma}^\varepsilon) = 0$ and $\gamma_1(\underline{\sigma}^\varepsilon) = 0$.

We introduce, as in the Kirchhoff–Love model, the symmetric tensor

$$\underline{\chi}^\varepsilon = \frac{1}{1-\nu} \left(\underline{\sigma}^\varepsilon - \frac{\nu}{1+\nu} (tr\underline{\sigma}^\varepsilon) \underline{I} \right).$$

We next show, by choosing $\underline{\tau} = \underline{curl}(\mathbf{curl}\, \varphi)$ with $\varphi \in \mathcal{D}(\Omega)$ as a test-function in the first equation of (3.1), that $\underline{\chi}^\varepsilon = \underline{\varepsilon}(\mathbf{r}^\varepsilon)$, which translates into the constitutive law of the plate according to (2.2):

$$\underline{\sigma}^\varepsilon = (1-\nu)\, \underline{\varepsilon}\, (\mathbf{r}^\varepsilon) + \nu (\operatorname{div} \mathbf{r}^\varepsilon) \underline{I}.$$

One clearly obtains, since $\underline{\tau}$ is symmetric, divergence-free, and with null traces, that

$$a(\underline{\sigma}^\varepsilon, \underline{\tau}) = 0 \quad \Leftrightarrow \quad \int_\Omega \underline{\chi}^\varepsilon : \underline{\tau}\, d\Omega = 0.$$

This leads to

$$curl \begin{pmatrix} \partial_1 \chi^\varepsilon_{12} - \partial_2 \chi^\varepsilon_{11} \\ \partial_1 \chi^\varepsilon_{22} - \partial_2 \chi^\varepsilon_{21} \end{pmatrix} = 0,$$

so there exists a function θ^ε such that

$$curl \begin{pmatrix} \chi^\varepsilon_{11} \\ \chi^\varepsilon_{12} - \theta^\varepsilon \end{pmatrix} = curl \begin{pmatrix} \chi^\varepsilon_{21} + \theta^\varepsilon \\ \chi^\varepsilon_{22} \end{pmatrix} = 0.$$

One can still write this as

$$\begin{pmatrix} \chi^\varepsilon_{11} \\ \chi^\varepsilon_{12} - \theta^\varepsilon \end{pmatrix} = \nabla r^\varepsilon_1, \ \begin{pmatrix} \chi^\varepsilon_{21} + \theta^\varepsilon \\ \chi^\varepsilon_{22} \end{pmatrix} = \nabla r^\varepsilon_2$$

or, equivalently,

$$\underline{\chi}^\varepsilon = \underline{\nabla} \mathbf{r}^\varepsilon + \theta^\varepsilon \underline{J}.$$

The symmetry of $\underline{\chi}^\varepsilon$ next gives that $\underline{\chi}^\varepsilon = \underline{\varepsilon}(\mathbf{r}^\varepsilon)$, where the vector $\mathbf{r}^\varepsilon \in (H^1(\Omega))^2$ is unique up to $\underline{\varepsilon}(\mathbf{r}^\varepsilon) = \underline{0}$, i.e., up to a polynomial $\left(\begin{smallmatrix} cy+a \\ -cx+b \end{smallmatrix}\right)$

Let us now take in (3.1) as test-function $\underline{\tau} = \underline{curl}\, \varphi + \frac{1}{2}(\mathrm{div}\varphi)\underline{J}$ with $\varphi \in (\mathcal{D}(\Omega))^2$. Obviously, $\underline{\tau}$ is symmetric and its traces vanish on Γ, so that

$$\begin{aligned} & a(\underline{\sigma}^\varepsilon, \underline{\tau}) + \varepsilon^2 a_0(\underline{\sigma}^\varepsilon, \underline{\tau}) = 0 \\ \Leftrightarrow & \int_\Omega \underline{\nabla} \mathbf{r}^\varepsilon : \underline{\tau}\, d\Omega + \frac{\varepsilon^2}{1-\nu} \int_\Omega \mathbf{div} \underline{\sigma}^\varepsilon \cdot \mathbf{div} \underline{\tau}\, d\Omega = 0 \\ \Leftrightarrow & \left\langle \frac{\varepsilon^2}{1-\nu} \mathbf{div} \underline{\sigma}^\varepsilon - \mathbf{r}^\varepsilon, \mathbf{curl}(\mathrm{div}\varphi) \right\rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} = 0 \\ \Leftrightarrow & \left\langle \nabla \left(curl \left(\frac{\varepsilon^2}{1-\nu} \mathbf{div} \underline{\sigma}^\varepsilon - \mathbf{r}^\varepsilon \right) \right), \varphi \right\rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} = 0 \quad \forall \varphi \in (\mathcal{D}(\Omega))^2. \end{aligned}$$

Since $curl \mathbf{r}^\varepsilon$ is unique up to a constant, we conclude that we can take

$$curl \left(\frac{\varepsilon^2}{1-\nu} \mathbf{div} \underline{\sigma}^\varepsilon - \mathbf{r}^\varepsilon \right) = 0,$$

where $\mathbf{r}^\varepsilon$ now belongs to $(H^1(\Omega)_{|\mathbb{R}})^2$. Thus, one finds a unique $u^\varepsilon \in H^1(\Omega)_{|\mathbb{R}}$ such that

$$\frac{\varepsilon^2}{1-\nu} \mathbf{div} \underline{\sigma}^\varepsilon - \mathbf{r}^\varepsilon = -\nabla u^\varepsilon.$$

We already know that $\underline{\sigma}^\varepsilon \in \underline{X}^{\varepsilon, f}$, which implies that $D(\underline{\sigma}^\varepsilon) = f$. So, we have actually shown that the functions $\underline{\sigma}^\varepsilon, \mathbf{r}^\varepsilon, u^\varepsilon$ satisfy the equations in Ω of the Reissner–Mindlin problem (2.2).

On the other hand, considering as a test-function $\underline{\tau} = \rho \underline{J}$ with $\rho \in \mathcal{D}(\Omega)$ arbitrary gives

$$\begin{aligned}
&\varepsilon^2 a_0(\underline{\sigma}^\varepsilon, \underline{\tau}) + c(\lambda^\varepsilon, \underline{\tau}) = 0 \\
\Leftrightarrow\ & \frac{\varepsilon^2}{1-\nu} \int_\Omega \mathbf{div}\underline{\sigma}^\varepsilon \cdot \mathbf{curl}\, \rho \, d\Omega + 2 \int_\Omega \lambda^\varepsilon \rho \, d\Omega = 0 \\
\Leftrightarrow\ & \left\langle curl \left(\frac{\varepsilon^2}{1-\nu} \mathbf{div}\underline{\sigma}^\varepsilon \right) + 2\lambda^\varepsilon, \rho \right\rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} = 0 \\
&\Leftrightarrow \lambda^\varepsilon = -\frac{1}{2} curl \mathbf{r}^\varepsilon.
\end{aligned}$$

We still have to check the boundary conditions for $\mathbf{r}^\varepsilon$ and u^ε. For that, let us consider an arbitrary tensor $\underline{\tau} \in \underline{X}^{\varepsilon,0}$. By Green's formula,

$$\begin{aligned}
&a^\varepsilon(\underline{\sigma}^\varepsilon, \underline{\tau}) + c(\lambda^\varepsilon, \underline{\tau}) \\
&= \int_\Omega \underline{\varepsilon}\,(\mathbf{r}^\varepsilon) : \underline{\tau} \, d\Omega + \frac{\varepsilon^2}{1-\nu} \int_\Omega \mathbf{div}\underline{\sigma}^\varepsilon \cdot \mathbf{div}\underline{\tau} \, d\Omega - \frac{1}{2} \int_\Omega curl \mathbf{r}^\varepsilon (\tau_{12} - \tau_{21}) \, d\Omega \\
&= \int_\Omega \underline{\nabla} \mathbf{r}^\varepsilon : \underline{\tau} \, d\Omega + \int_\Omega (\mathbf{r}^\varepsilon - \nabla u^\varepsilon) \cdot \mathbf{div}\underline{\tau} \, d\Omega = \langle \underline{\tau}\mathbf{n}, \mathbf{r}^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - \langle \mathbf{div}\underline{\tau} \cdot \mathbf{n}, u^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}.
\end{aligned}$$

In particular, for $\underline{\tau} = \rho \underline{J}$ with $\rho \in \mathcal{D}(\overline{\Omega})$, we have, since $b(\underline{\tau}, (u_0^\varepsilon, r_0^\varepsilon)) = 0$, that

$$\langle \underline{\tau}\mathbf{n}, \mathbf{r}^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - \langle \mathbf{div}\underline{\tau} \cdot \mathbf{n}, u^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0.$$

We decompose $\mathbf{r}^\varepsilon$ in the orthogonal basis $\{\mathbf{n}, \mathbf{t}\}$ and thus

$$\begin{aligned}
&\langle \underline{\tau}\mathbf{n} \cdot \mathbf{t}, \mathbf{r}^\varepsilon \cdot \mathbf{t} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \gamma_0(\underline{\tau}), \mathbf{r}^\varepsilon \cdot \mathbf{n} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} \\
&- \langle \gamma_1(\underline{\tau}), u^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t(\underline{\tau}\mathbf{n} \cdot \mathbf{t}), u^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0 \\
&\Leftrightarrow \langle \rho, \mathbf{r}^\varepsilon \cdot \mathbf{t} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - \langle \partial_t u^\varepsilon, \rho \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0 \\
&\Leftrightarrow \mathbf{r}^\varepsilon \cdot \mathbf{t} = \partial_t u^\varepsilon \quad \text{on } \Gamma.
\end{aligned}$$

The above equality holds in $H^{-1/2}(\Gamma)$, but the fact that $\mathbf{r}^\varepsilon$ belongs to the space $(H^{1/2}(\Gamma))^2$ implies an equality in $H^{1/2}(\Gamma)$; consequently, $u^\varepsilon \in H^{3/2}(\Gamma)$. Next taking $\underline{\tau} = \underline{curl}\, \varphi$ with an arbitrary $\varphi \in (\mathcal{D}(\overline{\Omega}))^2$, we obtain, according to the first equation of (3.1), that

$$\langle \underline{\tau}\mathbf{n}, \mathbf{r}^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - \langle \mathbf{div}\underline{\tau} \cdot \mathbf{n}, u^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \gamma_1(\underline{\tau}), u_0^\varepsilon \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} - \langle \gamma_0(\underline{\tau}), r_0^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0.$$

Since one has $\mathbf{div}\underline{\tau} = 0$ and

$$\begin{aligned}
\langle \underline{\tau}\mathbf{n}, \mathbf{r}^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} &= -\langle \partial_t \varphi \cdot \mathbf{n}, \mathbf{r}^\varepsilon \cdot \mathbf{n} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t(\partial_t \varphi \cdot \mathbf{t}), u^\varepsilon \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}, \\
\langle \gamma_1(\underline{\tau}), u_0^\varepsilon \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} &= -\langle \partial_t(\partial_t \varphi \cdot \mathbf{t}), u_0^\varepsilon \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}, \\
\langle \gamma_0(\underline{\tau}), r_0^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} &= -\langle \partial_t \varphi \cdot \mathbf{n}, r_0^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma},
\end{aligned}$$

one deduces the following relationship:

$$\forall \varphi \in (\mathcal{D}(\overline{\Omega}))^2, \quad -\langle \partial_t \varphi \cdot \mathbf{n}, \mathbf{r}^\varepsilon \cdot \mathbf{n} - r_0^\varepsilon \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t(\partial_t \varphi \cdot \mathbf{t}), u^\varepsilon - u_0^\varepsilon \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} = 0.$$

By a density argument of $\mathcal{D}(\overline{\Omega})$ into $H^1(\Omega)$, it actually turns out that

$$\forall \underline{\tau} \in \underline{X}^0, \quad b(\underline{\tau}, (\mathbf{r}^\varepsilon \cdot \mathbf{n} - r_0^\varepsilon, u^\varepsilon - u_0^\varepsilon)) = 0$$

with $(\mathbf{r}^\varepsilon \cdot \mathbf{n} - r_0^\varepsilon, u^\varepsilon - u_0^\varepsilon) \in H^{1/2}(\Gamma) \times H^{3/2}(\Gamma)$ by now. The same idea as in the Kirchhoff–Love case [1] leads us to $|w|_{2,\Omega} = 0$, where w satisfies the following biharmonic problem:

$$\begin{cases} \Delta^2 w = 0 & \text{in } \Omega, \\ w = u^\varepsilon - u_0^\varepsilon & \text{on } \Gamma, \\ \partial_n w = \mathbf{r}^\varepsilon \cdot \mathbf{n} - r_0^\varepsilon & \text{on } \Gamma. \end{cases}$$

So w is a first-order polynomial. Since $\partial_n w = \mathbf{c} \cdot \mathbf{n}$ with $\mathbf{c} \in \mathbb{R}^2$ and $\mathbf{r}^\varepsilon$ is unique up to a constant, we choose it such that $\mathbf{r}^\varepsilon \cdot \mathbf{n} = r_0^\varepsilon$ on Γ and $\nabla w = 0$. We finally obtain that $u^\varepsilon = u_0^\varepsilon$ on Γ, because u^ε is unique up to a constant, which we take equal to w.

We conclude that $\mathbf{r}^\varepsilon$ and u^ε satisfy the boundary conditions of the Reissner–Mindlin model, and, moreover, we have the equivalence stated in (3.3). ∎

From now on, we write $\underline{\sigma}^\varepsilon \in \underline{X}^{\varepsilon,f}$ as

$$\underline{\sigma}^\varepsilon = \underline{\sigma}^{\varepsilon,0} + \phi^f \underline{I},$$

where ϕ^f is the unique solution of

$$\begin{cases} \Delta\phi = f & \text{in } \Omega, \\ \phi = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ \partial_n \phi = 0 & \text{on } \Gamma_2 \end{cases} \tag{3.4}$$

and where $\underline{\sigma}^{\varepsilon,0} \in \underline{X}^{\varepsilon,0}$. By the means of this decomposition, we get a variational formulation in $\underline{\sigma}^{\varepsilon,0}$ equivalent to (3.1), which we will study in what follows:

$$\begin{cases} \text{find } \underline{\sigma}^{\varepsilon,0} \in \underline{X}^{\varepsilon,0},\ (u_0^\varepsilon, r_0^\varepsilon) \in M \times N,\ \lambda^\varepsilon \in L^2(\Omega) & \text{such that} \\ \forall \underline{\tau} \in \underline{X}^{\varepsilon,0}, & a^\varepsilon(\underline{\sigma}^{\varepsilon,0}, \underline{\tau}) + b(\underline{\tau}, (u_0^\varepsilon, r_0^\varepsilon)) + c(\underline{\tau}, \lambda^\varepsilon) = -a^\varepsilon(\phi^f \underline{I}, \underline{\tau}), \\ \forall (\zeta, \eta) \in M \times N, & b(\underline{\sigma}^{\varepsilon,0}, (\zeta, \eta)) = \langle \phi^f, \eta \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}, \\ \forall \mu \in L^2(\Omega), & c(\underline{\sigma}^{\varepsilon,0}, \mu) = 0. \end{cases} \tag{3.5}$$

Remark 3.1. In order to obtain the transverse displacement u^ε, one can now solve, thanks to (3.3), the second-order elliptic problem

$$\begin{cases} \Delta u^\varepsilon = \frac{1}{1+\nu} tr\underline{\sigma}^\varepsilon - \frac{\varepsilon^2}{1-\nu} f & \text{in } \Omega, \\ u^\varepsilon = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ u^\varepsilon = u_0^\varepsilon & \text{on } \Gamma_2, \end{cases} \tag{3.6}$$

while the rotation vector $\mathbf{r}^\varepsilon$ is given by the relation

$$\mathbf{r}^\varepsilon = \frac{\varepsilon^2}{1-\nu} \mathbf{div}\underline{\sigma}^\varepsilon + \nabla u^\varepsilon. \tag{3.7}$$

4. New mixed variational formulation. The previous formulation (3.5) of the problem is written on the space $\underline{X}^{\varepsilon,0}$, whose tensors satisfy the constraint $D(\underline{\tau}) = \mathrm{div}(\mathbf{div}\underline{\tau}) = 0$. In order to avoid its discretization, we introduce in this section an equivalent formulation, obtained by decomposing the elements of $\underline{X}^{\varepsilon,0}$. This new variational problem has the advantage of using only classical Sobolev spaces (like $H^1(\Omega)$, $H^{1/2}(\Gamma)$, $L^2(\Omega)$) which are easy to approximate by conforming finite elements.

In a similar manner as for the Kirchhoff–Love model [1], one can show (by applying Tartar's lemma twice) that for any $\underline{\tau} \in \underline{X}^{\varepsilon,0}$ there exist unique functions $\rho \in H^1(\Omega)_{|\mathbb{R}}$ and $\varphi \in (H^1(\Omega)_{|\mathbb{R}})^2$ such that $\underline{\tau}$ writes as

$$\underline{\tau} = \underline{curl}\,\varphi + \rho \underline{J}. \tag{4.1}$$

Let us consider the spaces

$$\mathbf{H} = \left\{ \varphi \in (H^1(\Omega))^2;\ \int_\Omega \varphi \, d\Omega = 0 \right\},$$
$$W^\varepsilon = \left\{ \rho \in L^2(\Omega);\ \int_\Omega \rho \, d\Omega = 0,\ \varepsilon\, \mathbf{curl}\rho \in L^2(\Omega)^2 \right\}$$

endowed, respectively, with the norms $|\cdot|_{1,\Omega}$ and $\|\cdot\|_{0,\Omega} + \varepsilon\,|\cdot|_{1,\Omega}$, and let us define the Hilbert space $\mathbf{Y}^\varepsilon = \mathbf{H} \times W^\varepsilon$.

We do not impose here, like in [1], the symmetry of the bending moment because this would lead us to consider a function φ too regular, difficult to approximate by continuous low-order finite elements. For this reason, with any $\underline{\sigma}$, $\underline{\tau} \in \underline{X}^{\varepsilon,0}$ we associate by means of (4.1) the corresponding couples (ψ, ξ), $(\varphi, \rho) \in \mathbf{Y}^\varepsilon$ and we introduce the following bilinear form on $\mathbf{Y}^\varepsilon \times \mathbf{Y}^\varepsilon$:

$$A^\varepsilon((\psi,\xi),(\varphi,\rho)) = A((\psi,\xi),(\varphi,\rho)) + \varepsilon^2 A_0((\psi,\xi),(\varphi,\rho)),$$

where

$$\begin{aligned} A((\psi,\xi),(\varphi,\rho)) = a(\underline{\sigma},\underline{\tau}) = \frac{1}{1-\nu}\int_\Omega &[(\xi - \partial_1\psi_1)(\rho - \partial_1\varphi_1) \\ &+ (\xi - \partial_2\psi_2)(\rho - \partial_2\varphi_2)]\, d\Omega \\ &+ \frac{1}{1-\nu}\int_\Omega (\partial_2\psi_1\partial_2\varphi_1 + \partial_1\psi_2\partial_1\varphi_2)\, d\Omega \\ &- \frac{\nu}{1-\nu^2}\int_\Omega (\partial_2\psi_1 - \partial_1\psi_2)(\partial_2\varphi_1 - \partial_1\varphi_2)\, d\Omega, \\ A_0((\psi,\xi),(\varphi,\rho)) = a_0(\underline{\sigma},\underline{\tau}) = \frac{1}{1-\nu}\int_\Omega &\mathbf{curl}\xi \cdot \mathbf{curl}\rho \, d\Omega. \end{aligned}$$

We also define the bilinear continuous forms $B(\cdot,\cdot)$, $C(\cdot,\cdot)$ on $\mathbf{Y}^\varepsilon \times \mathbf{Z}$ and on $\mathbf{Y}^\varepsilon \times L^2(\Omega)$, respectively, by putting

$$B((\varphi,\rho),\mathbf{q}) = -\langle \partial_t \mathbf{q}, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma},$$
$$C((\varphi,\rho),\lambda) = c(\underline{\tau},\lambda) = \int_\Omega \lambda\,(2\rho - \mathrm{div}\varphi)\, d\Omega,$$

where the space $\mathbf{Z}$ is given by

$$\mathbf{Z} = \left\{ \mathbf{q} \in H^{1/2}(\Gamma)^2;\ \mathbf{q} = 0 \text{ on } \Gamma_0,\ \mathbf{q}\cdot\mathbf{t} = 0 \text{ on } \Gamma_1,\ \int_\Gamma \mathbf{q}\cdot\mathbf{t}\, d\Gamma = 0 \right\}.$$

Remark 4.1. The forms $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ correspond to the equations of the Kirchhoff–Love model, while $C(\cdot,\cdot)$ dualizes the symmetry of the bending tensor and $A_0(\cdot,\cdot)$ takes into account the contribution of the rotation. The space $\mathbf{Z}$ is the same as the one employed in [1]. Let us also notice that with any $\mathbf{q} \in \mathbf{Z}$ one can associate a unique couple $(\zeta,\eta) \in M \times N$ such that $\mathbf{q} = (\partial_t\zeta)\mathbf{t} + \eta\mathbf{n}$; then $B((\varphi,\rho),\mathbf{q}) = b(\underline{\tau},(\zeta,\eta))$; see [1] for more details.

Finally, let us introduce the linear continuous forms $F^\varepsilon(\cdot)$ and $G(\cdot)$ defined on $\mathbf{Y}^\varepsilon$ and on $\mathbf{Z}$, respectively, by the relationship

$$F^\varepsilon((\varphi,\rho)) = -a^\varepsilon(\phi^f \underline{I},\underline{\tau}), \qquad G(\mathbf{q}) = \int_\Gamma \phi^f \mathbf{q}\cdot\mathbf{n}\, d\Gamma.$$

Then we state the following result.

THEOREM 4.1. *The variational problem*

$$\begin{cases} \text{find } (\psi^\varepsilon, \xi^\varepsilon) \in \mathbf{Y}^\varepsilon,\ \mathbf{p}^\varepsilon \in \mathbf{Z},\ \lambda^\varepsilon \in L^2(\Omega) \quad \text{such that} \\ \forall ((\varphi,\rho)) \in \mathbf{Y}^\varepsilon, \quad A^\varepsilon((\psi^\varepsilon,\xi^\varepsilon),(\varphi,\rho)) + B((\varphi,\rho),\mathbf{p}^\varepsilon) + C((\varphi,\rho),\lambda^\varepsilon) = F^\varepsilon((\varphi,\rho)), \\ \forall \mathbf{q} \in \mathbf{Z}, \qquad B((\psi^\varepsilon,\xi^\varepsilon),\mathbf{q}) = G(\mathbf{q}), \\ \forall \mu \in L^2(\Omega), \qquad C((\psi^\varepsilon,\xi^\varepsilon),\mu) = 0 \end{cases} \tag{4.2}$$

has a unique solution.

Proof. The existence and the uniqueness are obtained by the means of the Babuška–Brezzi theory. For that, it is sufficient to check the $\mathbf{Y}^\varepsilon$-ellipticity of the bilinear form $A^\varepsilon(\cdot,\cdot)$ and the inf-sup condition for $D(\cdot,\cdot) = B(\cdot,\cdot) + C(\cdot,\cdot)$.

Concerning the ellipticity, one has for any $(\varphi,\rho) \in \mathbf{Y}^\varepsilon$ that

$$A((\varphi,\rho),(\varphi,\rho)) = a(\underline{\tau},\underline{\tau}) \geq c \| \underline{\tau} \|^2_{0,\Omega},$$

where we have put $\underline{\tau} = \underline{curl}\,\varphi + \rho\underline{J}$. We show next that

$$|\varphi|^2_{1,\Omega} + \|\rho\|^2_{0,\Omega} \leq c \| \underline{\tau} \|^2_{0,\Omega}, \tag{4.3}$$

which leads to the $\mathbf{Y}^\varepsilon$-ellipticity of $A^\varepsilon(\cdot,\cdot)$ uniformly with respect to ε:

$$A^\varepsilon((\varphi,\rho),(\varphi,\rho)) \geq c(|\varphi|^2_{1,\Omega} + \|\rho\|^2_{0,\Omega} + \varepsilon^2 |\rho|^2_{1,\Omega}).$$

One has

$$\| \underline{\tau} \|^2_{0,\Omega} = |\varphi_1|^2_{1,\Omega} + |\varphi_2|^2_{1,\Omega} + 2\|\rho\|^2_{0,\Omega} - 2\int_\Omega \rho\,(\partial_1\varphi_1 + \partial_2\varphi_2)\ d\Omega.$$

According to Girault and Raviart [8], there exists a positive constant k such that the following statement holds, for any $\rho \in L^2_0(\Omega)$:

$$k\|\rho\|_{0,\Omega} \leq \|\nabla\rho\|_{-1,\Omega} = \| \mathbf{div}\underline{\tau} \|_{-1,\Omega} \leq k_1 \| \underline{\tau} \|_{0,\Omega}. \tag{4.4}$$

Next, let δ be a positive real number. Young's inequality implies that

$$2\int_\Omega \rho\,(\partial_1\varphi_1 + \partial_2\varphi_2)\ d\Omega \leq \delta^2 \|\rho\|^2_{0,\Omega} + \frac{1}{\delta^2}|\varphi|^2_{1,\Omega},$$

so

$$(2-\delta^2)\|\rho\|^2_{0,\Omega} + \left(1 - \frac{1}{\delta^2}\right)|\varphi|^2_{1,\Omega} \leq \| \underline{\tau} \|^2_{0,\Omega},$$

and, consequently, by means of (4.4), it holds for $\delta \geq 1$ that

$$|\varphi|^2_{1,\Omega} \leq c \| \underline{\tau} \|^2_{0,\Omega}.$$

So we obtain the estimate (4.3).

We still have to establish the inf-sup condition for the bilinear form $D(\cdot,\cdot)$ on $\mathbf{Y}^\varepsilon \times (\mathbf{Z} \times L^2(\Omega))$, where

$$D((\varphi,\rho),(\mathbf{q},\mu)) = B((\varphi,\rho),\mathbf{q}) + C((\varphi,\rho),\mu).$$

To do that, we fix an arbitrary couple $(\mathbf{q},\mu)\in\mathbf{Z}\times L^2(\Omega)$ and we construct $(\varphi,\rho)\in\mathbf{Y}^\varepsilon$ such that $D((\varphi,\rho),(\mathbf{q},\mu))\geq c(\|\mathbf{q}\|^2_{1/2,\Gamma}+\|\mu\|^2_{0,\Omega})$ and

$$|\varphi|_{1,\Omega}+\|\rho\|_{0,\Omega}+\varepsilon\,|\rho|_{1,\Omega}\leq c(\|\mathbf{q}\|_{1/2,\Gamma}+\|\mu\|_{0,\Omega}).$$

We will actually take $\rho=0$, so we have only to construct $\varphi\in\mathbf{H}$.

Exactly as in the Kirchhoff–Love case [1], with any $\mathbf{q}\in\mathbf{Z}$ we associate the solution $\mathbf{w}$ of the following boundary problem:

$$\begin{cases}\Delta\mathbf{w}=0 & \text{in }\Omega,\\ \mathbf{w}=\mathbf{q} & \text{on }\Gamma,\end{cases}$$

and we take $\varphi_1\in(H^1(\Omega)_{|\mathbb{R}})^2$ such that $\underline{\nabla}\varphi_1=\underline{curl}\,\mathbf{w}$. Then one has, for any $\rho\in W^\varepsilon$, that

$$B((\varphi_1,\rho),\mathbf{q})\geq c\,\|\mathbf{q}\|^2_{1/2,\Gamma}\quad\text{and}\quad|\varphi_1|_{1,\Omega}\leq c\,\|\mathbf{q}\|_{1/2,\Gamma}.$$

We equally have that $\mathrm{div}\varphi_1\in L^2_0(\Omega)$, since

$$\int_\Omega \mathrm{div}\varphi_1\,d\Omega=\int_\Gamma\mathbf{q}\cdot\mathbf{t}\,d\Gamma=0.$$

Next, let us consider an arbitrary $\mu\in L^2(\Omega)$ and put

$$\lambda=\mu-P(\mu)-\mathrm{div}\varphi_1\qquad\text{with }\lambda\in L^2_0(\Omega).$$

According to [8], one knows that there exists $\varphi_2\in(H^1_0(\Omega))^2$ such that

$$\mathrm{div}\varphi_2=\lambda\qquad\text{and}\qquad|\,\varphi_2\,|_{1,\Omega}\leq c\,\|\lambda\|_{0,\Omega}.$$

Finally, we set

$$\varphi=\varphi_1+\varphi_2+\frac{P(\mu)}{2}\begin{pmatrix}x\\ y\end{pmatrix}$$

and choose φ_1 (which is unique up to a constant) such that $\int_\Omega\varphi\,d\Omega=0$. Then we notice that $\varphi\in\mathbf{H}$ and, moreover,

$$\begin{cases}\mathrm{div}\,\varphi=\mu & \text{in }\Omega,\\ |\varphi|_{1,\Omega}\leq c(\|\mathbf{q}\|_{1/2,\Gamma}+\|\lambda\|_{0,\Omega}+\|P(\mu)\|_{0,\Omega})\leq c(\|\mathbf{q}\|_{1/2,\Gamma}+\|\mu\|_{0,\Omega}).\end{cases}$$

The boundary Γ being polygonal, one has that $\partial_t\varphi=\partial_t\varphi_1+c\mathbf{t}$, with $c=\frac{P(\mu)}{2}$, which implies that

$$\partial_t\varphi\cdot\mathbf{n}=\partial_t\varphi_1\cdot\mathbf{n},\qquad\partial_t(\partial_t\varphi\cdot\mathbf{t})=\partial_t(\partial_t\varphi_1\cdot\mathbf{t})\qquad\text{on }\Gamma,$$

so it turns out that

$$D((\varphi,0),\mathbf{q})=B((\varphi_1,0),\mathbf{q})+C((\varphi,0),\mu)\geq c(\|\mathbf{q}\|^2_{1/2,\Gamma}+\|\mu\|^2_{0,\Omega}).$$

Finally, this gives us the desired inf-sup condition, which obviously uniformly holds with respect to ε. □

5. Equivalence with the initial Reissner–Mindlin model. We establish in this paragraph the link between the solution of the previous variational formulation and the solution of (2.2). This is given by the following result.

THEOREM 5.1. *Let* $((\psi^\varepsilon, \xi^\varepsilon), \mathbf{p}^\varepsilon, \lambda^\varepsilon)$ *be the unique solution of* (4.2). *Then we have*

$$\begin{cases} \underline{\sigma}^\varepsilon = \underline{curl}\,\psi^\varepsilon + \xi^\varepsilon \underline{J} + \phi^f \underline{I} & \text{in } \Omega, \\ \mathbf{r}^\varepsilon = \mathbf{p}^\varepsilon & \text{on } \Gamma, \\ -\dfrac{1}{2} curl\, \mathbf{r}^\varepsilon = \lambda^\varepsilon & \text{in } \Omega, \end{cases} \tag{5.1}$$

where $(\underline{\sigma}^\varepsilon, u^\varepsilon, \mathbf{r}^\varepsilon)$ *is the solution of the initial Reissner–Mindlin problem* (2.2).

Proof. We present here the steps of the proof, since it is similar to the one given in [1].

We note $\widetilde{\underline{\sigma}}^\varepsilon = \underline{curl}\,\psi^\varepsilon + \xi^\varepsilon \underline{J} + \phi^f \underline{I}$, for which we obviously have $D(\widetilde{\underline{\sigma}}^\varepsilon) = f$. Moreover, the second equation of (4.2) implies that $\gamma_0(\widetilde{\underline{\sigma}}^\varepsilon) = \gamma_1(\widetilde{\underline{\sigma}}^\varepsilon) = 0$, while the third equation gives that $\widetilde{\underline{\sigma}}^\varepsilon$ is symmetric.

By taking as a test-function a couple $(\mathbf{curl}\,\varphi, 0) \in \mathbf{Y}^\varepsilon$ with an arbitrary $\varphi \in \mathcal{D}(\Omega)$, which corresponds to the tensor $\underline{\tau} = \underline{curl}\mathbf{curl}\,\varphi$, we get according to the proof of Theorem 3.2 the constitutive law of the plate. More precisely, we show that there exists a function $\widetilde{\mathbf{r}}^\varepsilon \in (H^1(\Omega))^2$ unique up to $\underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) = \underline{0}$ such that

$$\widetilde{\underline{\sigma}}^\varepsilon = (1-\nu)\,\underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) + \nu(\mathrm{div}\widetilde{\mathbf{r}}^\varepsilon)\underline{I}.$$

Another choice of the test-function, i.e., $(\varphi - P(\varphi), \frac{1}{2}\mathrm{div}\,\varphi) \in \mathbf{Y}^\varepsilon$ with $\varphi \in (\mathcal{D}(\Omega))^2$ (which means that the associated tensor is $\underline{\tau} = \underline{curl}\varphi + \frac{1}{2}\mathrm{div}\,\varphi$), allows us to obtain, exactly as in Theorem 3.2, the existence of a unique $\widetilde{u}^\varepsilon \in H^1(\Omega)_{|\mathbb{R}}$ satisfying

$$\frac{\varepsilon^2}{1-\nu}\mathbf{div}\widetilde{\underline{\sigma}}^\varepsilon = \widetilde{\mathbf{r}}^\varepsilon - \nabla\widetilde{u}^\varepsilon.$$

We next take as a test-function in (4.2) a couple $(0, \rho^0) \in \mathbf{Y}^\varepsilon$, with $\rho^0 = \rho - P(\rho)$ and $\rho \in \mathcal{D}(\overline{\Omega})$ arbitrary. This leads us to

$$\int_\Omega (curl\,\widetilde{\mathbf{r}}^\varepsilon + 2\lambda^\varepsilon)\rho^0\, d\Omega = 0 \quad \Leftrightarrow \quad curl\,\widetilde{\mathbf{r}}^\varepsilon + 2\lambda^\varepsilon = k, \quad k \in \mathbb{R}.$$

We have thus obtained that $\lambda^\varepsilon = -\frac{1}{2} curl\,\widetilde{\mathbf{r}}^\varepsilon + k$, and we shall next show that $k = 0$.

Let us study for the moment the boundary conditions satisfied by the functions $\widetilde{\mathbf{r}}^\varepsilon$ and $\widetilde{u}^\varepsilon$. The same choice as above, that is, $(0, \rho^0)$, gives us the following relationship, where $\underline{\tau} = \rho^0 \underline{J}$:

$$\int_\Omega \underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) : \underline{\tau}\, d\Omega + \frac{\varepsilon^2}{1-\nu}\int_\Omega \mathbf{div}\widetilde{\underline{\sigma}}^\varepsilon \cdot \mathbf{div}\underline{\tau}\, d\Omega + \int_\Omega \left(k - \frac{1}{2} curl\,\widetilde{\mathbf{r}}^\varepsilon\right)(\tau_{12} - \tau_{21})\, d\Omega = 0$$

$$\Leftrightarrow \int_\Omega (\widetilde{\mathbf{r}}^\varepsilon - \nabla\widetilde{u}^\varepsilon)\cdot\mathbf{curl}\rho^0\, d\Omega - \int_\Omega curl\,\widetilde{\mathbf{r}}^\varepsilon\,\rho^0\, d\Omega + 2k\int_\Omega \rho^0\, d\Omega = 0$$

$$\Leftrightarrow \langle \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} - \partial_t\widetilde{u}^\varepsilon, \rho^0\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0. \tag{5.2}$$

For $\rho \in \mathcal{D}(\Omega)$ with $\int_\Omega \rho\, d\Omega \neq 0$, the above equality implies $\langle \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} - \partial_t\widetilde{u}^\varepsilon, 1\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0$, so (5.2) now writes as

$$\forall \rho \in \mathcal{D}(\overline{\Omega}), \qquad \langle \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} - \partial_t\widetilde{u}^\varepsilon, \rho\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0,$$

which means $\widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} = \partial_t \widetilde{u}^\varepsilon$ on Γ. In order to obtain the other boundary conditions, we consider in the first equation of (4.2) the test-function $(\widehat{\varphi}, 0) \in \mathbf{Y}^\varepsilon$ with

$$\widehat{\varphi} = \varphi' - P(\varphi'), \quad \text{where} \quad \varphi' = \varphi - \frac{P(\mathrm{div}\varphi)}{2} \begin{pmatrix} x \\ y \end{pmatrix},$$

and where $\varphi \in (\mathcal{D}(\overline{\Omega}))^2$ is arbitrary. Then one has

$$\begin{aligned}
&\int_\Omega \underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) : \underline{curl}\, \varphi' \, d\Omega - \langle \partial_t \mathbf{p}^\varepsilon, \varphi' \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \int_\Omega \left(\frac{1}{2} curl\, \widetilde{\mathbf{r}}^\varepsilon - k \right) \mathrm{div}\varphi' \, d\Omega = 0 \\
&\quad \Leftrightarrow \int_\Omega \underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) : \underline{curl}\, \varphi \, d\Omega - \langle \partial_t \mathbf{p}^\varepsilon, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} \\
&\qquad + \frac{1}{2} \int_\Omega curl\, \widetilde{\mathbf{r}}^\varepsilon \, \mathrm{div}\varphi \, d\Omega + \left\langle \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t}, \frac{P(\mathrm{div}\varphi)}{2} \right\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0 \\
&\qquad \Leftrightarrow \int_\Omega \underline{\nabla} \widetilde{\mathbf{r}}^\varepsilon : \underline{curl}\, \varphi \, d\Omega - \langle \partial_t \mathbf{p}^\varepsilon, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0 \\
&\qquad\quad \Leftrightarrow \langle \partial_t (\widetilde{\mathbf{r}}^\varepsilon - \mathbf{p}^\varepsilon), \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0.
\end{aligned}$$

We used here the following relationships, which are true for any $\mathbf{c} \in \mathbb{R}^2$ and any $c \in \mathbb{R}$:

$$\langle \partial_t \mathbf{p}^\varepsilon, \mathbf{c} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0, \qquad \left\langle \partial_t \mathbf{p}^\varepsilon, c \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0, \qquad \langle \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t}, c \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0.$$

Since $\widetilde{\mathbf{r}}^\varepsilon$ is unique up to a constant, we first get that $\widetilde{\mathbf{r}}^\varepsilon = \mathbf{p}^\varepsilon$ on Γ. So we have that

$$\widetilde{\mathbf{r}}^\varepsilon = 0 \quad \text{on } \Gamma_0, \qquad \widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} = 0 \quad \text{on } \Gamma_0 \cup \Gamma_1,$$

and, since $\widetilde{\mathbf{r}}^\varepsilon \cdot \mathbf{t} = \partial_t \widetilde{u}^\varepsilon$ on Γ with $\widetilde{u}^\varepsilon$ unique up to a constant, we obtain $\widetilde{u}^\varepsilon = 0$ on $\Gamma_0 \cup \Gamma_1$.

Now, for the test-function $(\widehat{\varphi}, 0) \in \mathbf{Y}^\varepsilon$ with $\widehat{\varphi} = \varphi - P(\varphi)$ where $\varphi \in (\mathcal{D}(\overline{\Omega}))^2$ is arbitrary, it comes from (4.2) that

$$\begin{aligned}
&\int_\Omega \underline{\varepsilon}(\widetilde{\mathbf{r}}^\varepsilon) : \underline{curl}\, \varphi \, d\Omega - \langle \partial_t \mathbf{p}^\varepsilon, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \int_\Omega \left(\frac{1}{2} curl\, \widetilde{\mathbf{r}}^\varepsilon - k \right) \mathrm{div}\varphi \, d\Omega = 0 \\
&\quad \Leftrightarrow \int_\Omega \underline{\nabla} \widetilde{\mathbf{r}}^\varepsilon : \underline{curl}\, \varphi \, d\Omega - \langle \partial_t \widetilde{\mathbf{r}}^\varepsilon, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - k \int_\Omega \mathrm{div}\varphi \, d\Omega = 0 \\
&\quad\quad \Leftrightarrow k \int_\Omega \mathrm{div}\varphi \, d\Omega = 0 \quad \forall \varphi \in (\mathcal{D}(\overline{\Omega}))^2,
\end{aligned}$$

which means that $k = 0$. Therefore, $(\widetilde{\underline{\sigma}}^\varepsilon, \widetilde{u}^\varepsilon, \widetilde{\mathbf{r}}^\varepsilon)$ satisfies the relationships in (5.1) as well as the Reissner–Mindlin equations (2.2). The uniqueness of the solution of (2.2) ends our proof. ∎

Remark 5.1. We can now calculate the solution $(\underline{\sigma}^\varepsilon, u^\varepsilon, \mathbf{r}^\varepsilon)$ of the Reissner–Mindlin model by means of the solution $((\psi^\varepsilon, \xi^\varepsilon), \mathbf{p}^\varepsilon, \lambda^\varepsilon)$ of problem (4.2). Indeed, the bending moment is given by (5.1), the transverse displacement is obtained as solution of the elliptic problem

$$\begin{cases} \Delta u^\varepsilon = \dfrac{1}{1+\nu} tr \underline{\sigma}^\varepsilon - \dfrac{\varepsilon^2}{1-\nu} f & \text{in } \Omega, \\ u^\varepsilon = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ \partial_t u^\varepsilon = \mathbf{p}^\varepsilon \cdot \mathbf{t} & \text{on } \Gamma_2, \end{cases} \tag{5.3}$$

and the rotation is obtained thanks to the relationship (3.7).

6. Finite element approximation. We are interested here in the conforming discretization of the variational formulation (4.2) which describes the Reissner–Mindlin problem (2.2).

For that, let $(\mathcal{T}_h)_{h>0}$ be a regular family of triangulations of the polygonal domain $\overline{\Omega}$, each $\mathcal{T}_h$ consisting of triangles K—$\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} K$. We employ classical notations: for every triangle K of $\mathcal{T}_h$, we denote by h_K its diameter and by $h = \max_{K\in\mathcal{T}_h} h_K$ the discretization parameter. We also introduce the set $\partial\mathcal{T}_h$ of edges of the triangulation $\mathcal{T}_h$ situated on Γ.

6.1. Discrete variational formulation. In order to approximate $\mathbf{p}^\varepsilon$, we shall use the same finite elements as in the Kirchhoff–Love case, that is,

$$\mathbf{Z}_h = \left\{ \mathbf{q}_h \in \mathbf{Z};\ \ \mathbf{q}_h \in (\mathcal{C}^0(\Gamma))^2 \ \text{ and } \ \forall T \in \partial\mathcal{T}_h,\ \ \mathbf{q}_{h|T} \in (P_1(T))^2 \right\}.$$

The approximation of the additional unknowns $\xi^\varepsilon \in W^\varepsilon$ and $\lambda^\varepsilon \in L^2(\Omega)$ will be achieved in the following finite dimensional spaces:

$$W_h = \left\{ \rho_h \in H^1(\Omega);\ \ \forall K \in \mathcal{T}_h,\ \ \rho_{h|K} \in P_1(K) \right\},$$
$$L_h = \left\{ \lambda_h \in L^2(\Omega);\ \ \forall K \in \mathcal{T}_h,\ \ \lambda_{h|K} \in P_0(K) \right\},$$

while for ψ^ε we employ the finite element space $\mathbf{H}_h = \mathbf{H} \cap (H_h^1)^2$, where

$$H_h^1 = \left\{ \varphi_h \in H^1(\Omega);\ \ \forall K \in \mathcal{T}_h,\ \ \varphi_{h|K} \in P_2(K) \right\}.$$

Remark 6.1. For $\varepsilon \neq 0$, the space W^ε coincides algebraically with $H^1(\Omega)$ so W_h is a subspace of W^ε. The norm considered on W_h is obviously the weighted one previously defined on W^ε.

For the sake of simplicity, we denote

$$\mathbf{Y}_h = \mathbf{H}_h \times W_h \subset \mathbf{Y}^\varepsilon$$

and we put

$$\forall (\varphi_h, \rho_h) \in \mathbf{Y}_h, \quad F_h^\varepsilon((\varphi_h, \rho_h)) = -a^\varepsilon(\phi_h^f \underline{I}, \underline{curl}\,\varphi_h + \rho_h \underline{J})$$
$$\forall \mathbf{q}_h \in \mathbf{Z}_h, \qquad G_h(\mathbf{q}_h) = \int_\Gamma \phi_h^f\, \mathbf{q}_h \cdot \mathbf{n}\, d\Gamma.$$

The discrete function ϕ_h^f is a P_1-continuous finite element approximation of $\phi^f \in V$, the solution of the auxiliary problem (3.4). In order to calculate it explicitly, one can discretize the variational formulation of (3.4) and solve

$$\begin{cases} \text{find } \phi_h^f \in V_h \quad \text{such that} \\ \forall v_h \in V_h, \quad \displaystyle\int_\Omega \nabla\phi_h^f \cdot \nabla v_h\, d\Omega = \int_\Omega f v_h\, d\Omega, \end{cases} \tag{6.1}$$

where

$$V = \left\{ v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \right\}, \qquad V_h = W_h \cap V.$$

It is then obvious that

$$|\,\phi^f - \phi_h^f\,|_{1,\Omega} = \inf_{v_h \in V_h} |\,\phi^f - v_h\,|_{1,\Omega}\,.$$

Then we consider the discrete version of (4.2) written as below:

$$\begin{cases} \text{find } (\psi_h^\varepsilon, \xi_h^\varepsilon) \in \mathbf{Y}_h,\ \mathbf{p}_h^\varepsilon \in \mathbf{Z}_h,\ \lambda_h^\varepsilon \in L_h \quad \text{such that} \\ \forall (\varphi_h, \rho_h) \in \mathbf{Y}_h, \quad A^\varepsilon((\psi_h^\varepsilon, \xi_h^\varepsilon), (\varphi_h, \rho_h)) + B((\varphi_h, \rho_h), \mathbf{p}_h^\varepsilon) + C((\varphi_h, \rho_h), \lambda_h^\varepsilon) \\ \qquad\qquad = F_h^\varepsilon((\varphi_h, \rho_h)), \\ \forall \mathbf{q}_h \in \mathbf{Z}_h, \qquad B((\psi_h^\varepsilon, \xi_h^\varepsilon), \mathbf{q}_h) = G_h(\mathbf{q}_h), \\ \forall \mu_h \in L_h, \qquad C((\psi_h^\varepsilon, \xi_h^\varepsilon), \mu_h) = 0. \end{cases} \tag{6.2}$$

We prove in what follows that the discrete inf-sup condition of Babuška–Brezzi for the above mixed problem uniformly holds with respect to both the discretization parameter h and the plate's thickness ε.

THEOREM 6.1. *There exists a positive constant c, independent of h and ε, such that, for any $(\mathbf{q}_h, \mu_h) \in \mathbf{Z}_h \times L_h$,*

$$\sup_{(\varphi_h, \rho_h) \in \mathbf{Y}_h} \frac{B((\varphi_h, \rho_h), \mathbf{q}_h) + C((\varphi_h, \rho_h), \mu_h)}{|\varphi_h|_{1,\Omega} + \|\rho_h\|_{0,\Omega} + \varepsilon\, |\rho_h|_{1,\Omega}} \geq c(\|\mathbf{q}_h\|_{1/2,\Gamma} + \|\mu_h\|_{0,\Omega}).$$

Proof. We apply Fortin's trick. For that, we will use the continuous inf-sup condition established in Theorem 4.1 and the interpolation operator $P_h : (H^1(\Omega))^2 \cap (\mathcal{C}^0(\overline{\Omega}))^2 \to (H_h^1)^2$ defined hereafter.

Let us note by P_{1h} the classical Lagrange interpolation operator which satisfies, for any $\varphi \in (H^1(K))^2 \cap (\mathcal{C}^0(\overline{K}))^2$,

$$P_{1h}\varphi \in (\mathcal{P}_1(K))^2 \quad \text{and} \quad P_{1h}\varphi(S) = \varphi(S) \quad \forall S \text{ vertex of } K \in \mathcal{T}_h.$$

We introduce the operator P_{2h} defined by $P_{2h}\varphi \in (\mathcal{P}_2(K))^2$, and

$$\begin{aligned} P_{2h}\varphi(S) &= 0 \quad \text{for every vertex } S \text{ of } K, \\ \int_T (\varphi - P_{2h}\varphi)\, d\Gamma &= 0 \quad \text{for every edge } T \text{ of } K. \end{aligned}$$

Then we put on every triangle $K \in \mathcal{T}_h$

$$P_h\varphi = P_{1h}\varphi + P_{2h}(\varphi - P_{1h}\varphi),$$

which clearly satisfies the properties

$$\begin{aligned} &\forall T \in \partial\mathcal{T}_h, \quad \int_T P_h\varphi\, d\Gamma = \int_T \varphi\, d\Gamma, \\ &\forall K \in \mathcal{T}_h, \quad \int_K \operatorname{div}(P_h\varphi)\, d\Omega = \int_K \operatorname{div}\varphi\, d\Omega. \end{aligned} \tag{6.3}$$

If $\varphi \in \mathbf{H} \cap (\mathcal{C}^0(\overline{\Omega}))^2$, then we have only that $P_h\varphi$ belongs to $(H_h^1)^2$ and not to the space $\mathbf{H}$.

Let us now come back to the proof of the uniform inf-sup condition for problem (6.2).

With any $\mathbf{q}_h \in \mathbf{Z}_h$, we associate, exactly as in Theorem 4.1, a function $\varphi_1 \in (H^1(\Omega)_{|\mathbb{R}})^2$ such that

$$B((\varphi_1, 0), \mathbf{q}_h) \geq c\, \|\mathbf{q}_h\|_{1/2,\Gamma}^2 \quad \text{and} \quad |\varphi_1|_{1,\Omega} \leq c\, \|\mathbf{q}_h\|_{1/2,\Gamma}.$$

We recall that $\underline{\nabla}\varphi_1 = \underline{curl}\,\mathbf{w}$, where $\mathbf{w} \in (H^1(\Omega))^2$ is the unique function satisfying $\Delta\mathbf{w} = 0$ in Ω and $\mathbf{w} = \mathbf{q}_h$ on Γ. We also have, by construction, that $\mathrm{div}\varphi_1 \in L_0^2(\Omega)$. Since $\mathbf{q}_h \in (H^1(\Gamma))^2$, we obtain by classical results of regularity of the Laplace operator (see [9], [10]) that $\mathbf{w} \in (H^{1+a}(\Omega))^2$ with $a \in]0, \frac{1}{2}]$. We deduce that $\varphi_1 \in (H^{1+a}(\Omega))^2 \hookrightarrow (\mathcal{C}^0(\overline{\Omega}))^2$, so we can define $P_h\varphi_1$.

Then by considering the discrete function $\varphi_{1h} = P_h\varphi_1 \in (H_h^1)^2$, we have, thanks to (6.3), that

$$B((\varphi_1, 0), \mathbf{q}_h) = B((\varphi_{1h}, 0), \mathbf{q}_h).$$

On the other hand, we obtain, by passing to the reference finite element and using the Bramble–Hilbert lemma, that $|\varphi_1 - P_{2h}\varphi_1|_{1,K} \leq c\,|\varphi_1|_{1,K}$. This implies that

$$\forall K \in \mathcal{T}_h, \quad |\varphi_1 - \varphi_{1h}|_{1,K} \leq c\,|\varphi_1 - P_{1h}\varphi_1|_{1,K} \leq c\,|\varphi_1|_{1,K},$$

so by the triangle inequality, $|\varphi_{1h}|_{1,\Omega} \leq c\,|\varphi_1|_{1,\Omega}$. We now have that

$$\frac{B((\varphi_{1h}, 0), \mathbf{q}_h)}{|\,\varphi_{1h}\,|_{1,\Omega}} \geq c\frac{B((\varphi_1, 0), \mathbf{q}_h)}{|\varphi_1|_{1,\Omega}} \geq c\,\|\mathbf{q}_h\|_{1/2,\Gamma}.$$

Next, following the proof of Theorem 4.1, let us consider an arbitrary $\mu_h \in L_h$ and put

$$\lambda = \mu_h - P(\mu_h) - \mathrm{div}\varphi_{1h} \qquad \text{with } \lambda \in L_0^2(\Omega).$$

According to Girault and Raviart [8], one knows that there exists $\varphi_2 \in H_0^1(\Omega)^2$ such that

$$\mathrm{div}\varphi_2 = \lambda \qquad \text{with} \qquad |\,\varphi_2\,|_{1,\Omega} \leq c\,\|\lambda\|_{0,\Omega}.$$

Finally, we put $\varphi_h' = \varphi_{1h} + P_h\varphi_2 + \frac{P(\mu_h)}{2}\binom{x}{y}$, which belongs to $(H_h^1)^2$, and next consider

$$\varphi_h = \varphi_h' - P(\varphi_h').$$

This last function now belongs to $\mathbf{H}_h$, and it obviously satisfies

$$|\varphi_h|_{1,\Omega} \leq c(\|\mathbf{q}_h\|_{1/2,\Gamma} + \|\lambda\|_{0,\Omega} + \|\mu_h\|_{0,\Omega}) \leq c(\|\mathbf{q}_h\|_{1/2,\Gamma} + \|\mu_h\|_{0,\Omega}).$$

Then we notice that we have, thanks to (6.3),

$$C((\varphi_h, 0), \mu_h) = \|\mu_h\|_{0,\Omega}^2.$$

The boundary Γ being polygonal, one gets $\partial_t\varphi_h = \partial_t\varphi_{1h} + c\mathbf{t}$ with $c = \frac{P(\mu_h)}{2}$, which implies that

$$\partial_t\varphi_h \cdot \mathbf{n} = \partial_t\varphi_{1h} \cdot \mathbf{n}, \quad \partial_t(\partial_t\varphi_h \cdot \mathbf{t}) = \partial_t(\partial_t\varphi_{1h} \cdot \mathbf{t}) \quad \text{on } \Gamma,$$

so we have that

$$B((\varphi_h, 0), \mathbf{q}_h) = B((\varphi_{1h}, 0), \mathbf{q}_h).$$

Finally, this gives us

$$\sup_{(\psi_h,\rho_h)\in\mathbf{Y}_h} \frac{B((\psi_h,\rho_h),\mathbf{q}_h)+C((\psi_h,\rho_h),\mu_h)}{|\psi_h|_{1,\Omega}+\|\rho_h\|_{0,\Omega}+\varepsilon\,|\rho_h|_{1,\Omega}} \geq \frac{B((\varphi_h,0),\mathbf{q}_h)+C((\varphi_h,0),\mu_h)}{|\varphi_h|_{1,\Omega}} \geq c(\|\mathbf{q}_h\|_{1/2,\Gamma}+\|\mu_h\|_{0,\Omega}),$$

which ends the proof. □

Remark 6.2. It is equally possible to approximate ψ^ε by the same finite element as in the Kirchhoff–Love case (whose degrees of freedom are the values at the nodes of the triangulation, to which we add the values at the midpoints of the edges situated on $\Gamma_1\bigcup\Gamma_2$) and λ^ε by piecewise linear elements, and thus one will get a cheaper method. However, the discrete inf-sup condition will not be uniform with respect to h this time, and the convergence of the method will depend upon ε.

The previous result immediately gives, by using the Babuška–Brezzi theory, the following.

THEOREM 6.2. *Problem* (6.2) *admits a unique solution for any positive h and ε. Moreover, the following error bound holds:*

$$\begin{aligned}|\psi^\varepsilon-\psi_h^\varepsilon|_{1,\Omega}+\|\xi^\varepsilon-\xi_h^\varepsilon\|_{0,\Omega}+\varepsilon\,|\xi^\varepsilon-\xi_h^\varepsilon|_{1,\Omega}+\|\mathbf{p}^\varepsilon-\mathbf{p}_h^\varepsilon\|_{1/2,\Gamma}+\|\lambda^\varepsilon-\lambda_h^\varepsilon\|_{0,\Omega}\\ \leq c\{\inf_{\varphi_h\in\mathbf{H}_h}|\psi^\varepsilon-\varphi_h|_{1,\Omega}+\inf_{\rho_h\in W_h}(\|\xi^\varepsilon-\rho_h\|_{0,\Omega}+\varepsilon\,|\xi^\varepsilon-\rho_h|_{1,\Omega})\\ +\inf_{\mathbf{q}_h\in\mathbf{Z}_h}\|\mathbf{p}^\varepsilon-\mathbf{q}_h\|_{1/2,\Gamma}+\inf_{\mu_h\in L_h}\|\lambda^\varepsilon-\mu_h\|_{0,\Omega}+\inf_{v_h\in V_h}|\,\phi^f-v_h\,|_{1,\Omega}\}\end{aligned}$$

with a positive constant c independent of both h and ε. Therefore the proposed approximation method is unconditionally convergent for any fixed ε.

6.2. Approximation of the physical variables. In order to obtain the approximated bending moment, we set (according to Theorem 5.1)

$$\underline{\sigma}_h^\varepsilon=\underline{curl}\,\psi_h^\varepsilon+\xi_h^\varepsilon\underline{J}+\phi_h^f\underline{I}, \tag{6.4}$$

while a P_1-continuous finite element discretization of the boundary value problem (5.3) will give us an approximation u_h^ε of the transverse displacement u^ε. For that, we write that u^ε satisfies the variational formulation

$$\begin{cases}\text{find } u^\varepsilon\in H^1(\Omega) \text{ with } u^\varepsilon=g^\varepsilon \text{ on } \Gamma, \quad \text{such that}\\ \forall v\in H_0^1(\Omega), \quad \displaystyle\int_\Omega\nabla u^\varepsilon\cdot\nabla v\,d\Omega\\ \qquad\qquad =\dfrac{-1}{1+\nu}\displaystyle\int_\Omega(tr\underline{\sigma}^\varepsilon)v\,d\Omega+\dfrac{\varepsilon^2}{1-\nu}\displaystyle\int_\Omega fv\,d\Omega,\end{cases} \tag{6.5}$$

where the function $g^\varepsilon\in H^{3/2}(\Gamma)$ is defined on the boundary Γ by the relationships

$$g^\varepsilon=0 \quad \text{on } \Gamma_0\cup\Gamma_1, \qquad \partial_t g^\varepsilon=\mathbf{p}^\varepsilon\cdot\mathbf{t} \quad \text{on } \Gamma_2.$$

In order to calculate u_h^ε, we consider the next discrete version of (6.5):

$$\begin{cases}\text{find } u_h^\varepsilon\in W_h \text{ with } u_h^\varepsilon=I_h(g_h^\varepsilon) \text{ on } \Gamma, \quad \text{such that}\\ \forall v_h\in W_h^0, \quad \displaystyle\int_\Omega\nabla u_h^\varepsilon\cdot\nabla v_h\,d\Omega=\dfrac{-1}{1+\nu}\displaystyle\int_\Omega(tr\underline{\sigma}_h^\varepsilon)v_h\,d\Omega+\dfrac{\varepsilon^2}{1-\nu}\displaystyle\int_\Omega fv_h\,d\Omega,\end{cases}$$

where $\underline{\sigma}_h^\varepsilon$ is given by (6.4), of course, and where $W_h^0 = W_h \cap H_0^1(\Omega)$. Numerical integration has also been employed on the function g^ε, which has been replaced by an approximation $I_h(g_h^\varepsilon)$. The operator I_h denotes the P_1-continuous interpolation operator on the boundary Γ, and the function g_h^ε is taken as follows:

$$g_h^\varepsilon = 0 \quad \text{on } \Gamma_0 \cup \Gamma_1, \qquad \partial_t g_h^\varepsilon = \mathbf{p}_h^\varepsilon \cdot \mathbf{t} \quad \text{on } \Gamma_2.$$

Since $g_h^\varepsilon \in H^{3/2}(\Gamma) \hookrightarrow \mathcal{C}^0(\Gamma)$, one can clearly define $I_h(g_h^\varepsilon)$.

Then we get the following error bound:

$$|u^\varepsilon - u_h^\varepsilon|_{1,\Omega} \le c\{ \inf_{v_h \in W_h^0} | u^{\varepsilon,0} - v_h |_{1,\Omega} + \| \underline{\sigma}^\varepsilon - \underline{\sigma}_h^\varepsilon \|_{0,\Omega} + \|g^\varepsilon - I_h(g_h^\varepsilon)\|_{1/2,\Gamma}\},$$

where we have put $u^{\varepsilon,0} = u^\varepsilon - u^{\varepsilon,g} \in H_0^1(\Omega)$ and $u^{\varepsilon,g} \in H^1(\Omega)$, a continuous lifting satisfying $u^{\varepsilon,g} = g^\varepsilon$ on Γ and $\Delta u^{\varepsilon,g} = 0$ in Ω.

The discrete rotation vector $\mathbf{r}_h^\varepsilon$ will then be recovered by the means of (3.7):

$$\mathbf{r}_h^\varepsilon = \frac{\varepsilon^2}{1-\nu} \mathbf{div} \underline{\sigma}_h^\varepsilon + \nabla u_h^\varepsilon.$$

We used here the fact that $\mathbf{div}\underline{\sigma}_h^\varepsilon = \mathbf{curl}\xi_h^\varepsilon + \nabla\phi_h^f$ belongs to $L^2(\Omega)$, since $\xi_h^\varepsilon \in H^1(\Omega)$, $\phi_h^f \in H^1(\Omega)$, and $\psi_h^\varepsilon \in (H^1(\Omega))^2$, so the function $\underline{curl}\, \psi_h^\varepsilon$ belongs to $(H(\mathrm{div};\Omega))^2$.

Let us finally notice that the finite element method employed gives us low-order approximations of the physical quantities in the following spaces: $\underline{\sigma}_h^\varepsilon \in (H(\mathrm{div};\Omega))^2$, $u_h^\varepsilon \in H^1(\Omega)$, and $\mathbf{r}_h^\varepsilon \in L^2(\Omega)^2$, with also an approximation of $curl\mathbf{r}^\varepsilon$ in $L^2(\Omega)$ thanks to the multiplier λ^ε. Moreover, the described method uses classical finite element spaces (continuous P_1 and P_2, discontinuous P_0). The preprocessing of ϕ_h^f and post-processing of the displacement u_h^ε are very simple—one has to solve only twice, by P_1-continuous elements, a Laplacian problem. Let us remark that these two discrete problems have the same matrix, which is computed only once.

So, thanks to Theorem 6.2, we are now able to state the main result of the section.

THEOREM 6.3. *The above finite element method for the Reissner–Mindlin problem is unconditionally convergent. The next error estimate holds, with c independent of h and ε:*

$$\begin{aligned}
&\| \underline{\sigma}^\varepsilon - \underline{\sigma}_h^\varepsilon \|_{0,\Omega} \\
&\quad + \varepsilon \| \mathbf{div}(\underline{\sigma}^\varepsilon - \underline{\sigma}_h^\varepsilon) \|_{0,\Omega} + \| D(\underline{\sigma}^\varepsilon) - D(\underline{\sigma}_h^\varepsilon) \|_{-1,\Omega} + |u^\varepsilon - u_h^\varepsilon|_{1,\Omega} + \|\mathbf{r}^\varepsilon - \mathbf{r}_h^\varepsilon\|_{0,\Omega} \\
&\quad \le c\{ \inf_{\varphi_h \in \mathbf{H}_h} |\psi^\varepsilon - \varphi_h|_{1,\Omega} + \inf_{\rho_h \in W_h} (\|\xi^\varepsilon - \rho_h\|_{0,\Omega} + \varepsilon\, |\xi^\varepsilon - \rho_h|_{1,\Omega}) \\
&\quad + \inf_{\mathbf{q}_h \in \mathbf{Z}_h} \|\mathbf{p}^\varepsilon - \mathbf{q}_h\|_{1/2,\Gamma} + \inf_{\mu_h \in L_h} \|\lambda^\varepsilon - \mu_h\|_{0,\Omega} \\
&\quad + \inf_{v_h \in V_h} | \phi^f - v_h |_{1,\Omega} + \inf_{v_h \in W_h^0} | u^{\varepsilon,0} - v_h |_{1,\Omega} + \|g^\varepsilon - I_h(g_h^\varepsilon)\|_{1/2,\Gamma}\}.
\end{aligned}$$

6.3. Convergence rate. In what follows, we are looking for an upper bound of each right-hand side term of the previous inequality.

We first recall that the function $\phi^f \in H^1(\Omega)$ satisfies the boundary value problem (3.4). Then the regularity results for the Laplace operator ensure that there exists $b \in]\frac{1}{2}, 1]$ such that

$$\phi^f \in H^{1+b}(\Omega) \qquad \text{and} \qquad || \phi^f ||_{1+b,\Omega} \le c \|f\|_{0,\Omega},$$

with $b = 1$ if Ω is convex. Then we get

$$\inf_{v_h \in V_h} | \phi^f - v_h |_{1,\Omega} \leq ch^b \|f\|_{0,\Omega} .$$

Next, let us notice that we have by construction of g_h^ε that

$$\| g_h^\varepsilon - I_h(g_h^\varepsilon) \|_{\frac{1}{2},\Gamma} \leq ch \| g_h^\varepsilon \|_{\frac{3}{2},\Gamma} \leq ch \| \mathbf{p}_h^\varepsilon \|_{\frac{1}{2},\Gamma} \leq ch \|f\|_{0,\Omega} .$$

We equally know by classical results on the Laplace operator (see [9], [10]) that there exists $a \in]0,1]$ such that $u^{\varepsilon,g} \in H^{1+a}(\Omega)$. Moreover, we have that

$$\left\|u^{\varepsilon,0}\right\|_{1+a,\Omega} \leq \|u^\varepsilon\|_{1+a,\Omega} + c\|\mathbf{p}^\varepsilon\|_{-1/2+a,\Gamma} \leq c\|f\|_{0,\Omega} .$$

In order to obtain the convergence rate of the discretization method, let us assume the following regularity for the exact solution of (2.2):

$$\begin{gathered} \mathbf{r}^\varepsilon \in H^{1+a}(\Omega)^2, \quad u^\varepsilon \in H^{1+a}(\Omega), \\ \|\mathbf{r}^\varepsilon\|_{1+a,\Omega} + \|u^\varepsilon\|_{1+a,\Omega} + \varepsilon \| \mathbf{div}\underline{\sigma}^\varepsilon \|_{a,\Omega} \leq c\|f\|_{0,\Omega} . \end{gathered} \tag{6.6}$$

This hypothesis is verified in convex domains with $a = 1$, at least for clamped plates (cf., for instance, [5]). Now, the previous inequality implies that

$$|\psi^\varepsilon|_{1+a,\Omega} + \|\xi^\varepsilon\|_{a,\Omega} + \varepsilon\,|\xi^\varepsilon|_{1+a,\Omega} + \|\mathbf{p}^\varepsilon\|_{1/2+a,\Gamma} + \|\lambda^\varepsilon\|_{a,\Omega} \leq c\|f\|_{0,\Omega} ,$$

which, together with Theorem 6.3, allows us to deduce the following theorem.

THEOREM 6.4. *Under the regularity assumption* (6.6), *the discretization method for the Reissner–Mindlin model is convergent to order* $O(h^{min\{a,b\}})$*:*

$$\| \underline{\sigma}^\varepsilon - \underline{\sigma}_h^\varepsilon \|_{0,\Omega} + \varepsilon \| \mathbf{div}(\underline{\sigma}^\varepsilon - \underline{\sigma}_h^\varepsilon) \|_{0,\Omega} + |u^\varepsilon - u_h^\varepsilon|_{1,\Omega} + \|\mathbf{r}^\varepsilon - \mathbf{r}_h^\varepsilon\|_{0,\Omega} \leq ch^{min\{a,b\}} \|f\|_{0,\Omega}$$

independently of the plate's thickness ε*.*

Therefore, our method uses simple conforming finite elements of low degree, is unconditionally convergent and locking-free and, in the case of a convex polygon, for instance (when $a = b = 1$), it is also optimal of order $O(h)$.

REFERENCES

[1] M. AMARA, D. CAPATINA-PAPAGHIUC, AND A. CHATTI, *Bending moment mixed method for the Kirchhoff–Love plate model*, SIAM J. Numer. Anal., to appear.

[2] D.N. ARNOLD AND R.S. FALK, *A uniformly accurate finite element method for the Reissner–Mindlin plate,* SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[3] I. BABUŠKA AND M. SURI, *On locking and robustness in the finite element method*, SIAM J. Numer. Anal., 29 (1992), pp. 1261–1293.

[4] J. BRAMBLE AND T. SUN, *A negative-norm least squares method for Reissner-Mindlin plates*, Math. Comp., 67 (1998), pp. 901–916.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[6] P. DESTUYNDER AND M. SALAUN, *Mathematical Analysis of Thin Plate Models*, Springer-Verlag, Berlin, 1996.

[7] R. FALK AND T. TU, *Locking-free finite elements for the Reissner-Mindlin plate*, Math. Comp., 69 (2000), pp. 911–928.

[8] V. GIRAULT AND P.A. RAVIART, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[9] P. GRISVARD, *Elliptic Problems in Non-Smooth Domains*, Pitman, Boston, 1985.

[10] V.A. KONDRATIEV AND O.A. OLEINIK, *Boundary value problems for partial differential equations in non smooth domains*, Russian Math. Surveys, 38 (1983), pp. 1–86.

[11] M. SURI, I. BABUŠKA, AND C. SCHWAB, *Locking effects in the finite element approximation of plate models*, Math. Comp., 64 (1995), pp. 461–482.

SIAM J. NUMER. ANAL.
Vol. 40, No. 4, pp. 1583–1604

NUMERICAL METHODS FOR STOCHASTIC SYSTEMS PRESERVING SYMPLECTIC STRUCTURE*

G. N. MILSTEIN†‡, YU. M. REPIN‡, AND M. V. TRETYAKOV§

Abstract. Stochastic Hamiltonian systems with multiplicative noise, phase flows of which preserve symplectic structure, are considered. To construct symplectic methods for such systems, sufficiently general fully implicit schemes, i.e., schemes with implicitness both in deterministic and stochastic terms, are needed. A new class of fully implicit methods for stochastic systems is proposed. Increments of Wiener processes in these fully implicit schemes are substituted by some truncated random variables. A number of symplectic integrators is constructed. Special attention is paid to systems with separable Hamiltonians. Some results of numerical experiments are presented. They demonstrate superiority of the proposed symplectic methods over very long times in comparison with nonsymplectic ones.

Key words. stochastic Hamiltonian systems, symplectic integration, implicit methods, mean-square convergence

AMS subject classifications. 60H10, 65C30, 65P10

PII. S0036142901395588

1. Introduction.

Consider the following Cauchy problem for the system of stochastic differential equations (SDEs) in the sense of Stratonovich:

$$dP = f(t,P,Q)dt + \sum_{r=1}^{m} \sigma_r(t,P,Q) \circ dw_r(t), \qquad P(t_0) = p, \tag{1.1}$$

$$dQ = g(t,P,Q)dt + \sum_{r=1}^{m} \gamma_r(t,P,Q) \circ dw_r(t), \qquad Q(t_0) = q,$$

where P, Q, f, g, σ_r, γ_r are n-dimensional column vectors with the components P^i, Q^i, f^i, g^i, σ_r^i, γ_r^i, $i = 1,\dots,n$, and $w_r(t)$, $r = 1,\dots,m$, are independent standard Wiener processes. The diffusion coefficients σ_r, γ_r depend on P, Q (i.e., (1.1) is a system with multiplicative noise), in contrast to [3], where stochastic systems with additive noise are treated.

We suppose that all the coefficients of considered systems are sufficiently smooth functions defined for $(t,p,q) \in [t_0, t_0+T] \times R^d$, $d = 2n$, and they provide the property of extendability of solutions to the interval $[t_0, t_0+T]$. (Additional conditions in connection with considered methods consist of appropriate behavior of partial derivatives of the coefficients on infinity.)

We denote by $X(t;t_0,x) = (P^\top(t;t_0,p,q), Q^\top(t;t_0,p,q))^\top$, $t_0 \le t \le t_0+T$, the solution of problem (1.1). A more detailed notation is $X(t;t_0,x;\omega)$, where ω is an

*Received by the editors September 24, 2001; accepted for publication (in revised form) April 17, 2002; published electronically October 23, 2002. The authors were partially supported by Russian Foundation for Basic Research project 99-01-00134.

http://www.siam.org/journals/sinum/40-4/39558.html

†Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, D-10117 Berlin, Germany (milstein@wias-berlin.de).

‡Department of Mathematics, Ural State University, Lenin Str. 51, 620083 Ekaterinburg, Russia (Grigori.Milstein@usu.ru, Yuri.Repin@usu.ru).

§Department of Mathematics, University of Wales Swansea, Swansea SA2 8PP, UK. Current address: Department of Mathematics and Computer Science, University of Leicester, Leicester LE1 7RH, UK (Michael.Tretyakov@usu.ru, M.Tretiakov@le.ac.uk).

elementary event. It is known that $X(t;t_0,x;\omega)$ is a phase flow (diffeomorphism) for almost every ω. See its properties in, e.g., [1, 2].

If there are functions $H_r(t,p,q)$, $r=0,\dots,m$, such that (see [1] and [3])

$$f^i(t,p,q)=-\partial H_0/\partial q^i,\qquad g^i(t,p,q)=\partial H_0/\partial p^i, \tag{1.2}$$
$$\sigma_r^i(t,p,q)=-\partial H_r/\partial q^i,\quad \gamma_r^i(t,p,q)=\partial H_r/\partial p^i,\quad i=1,\dots,n,\quad r=1,\dots,m,$$

then the phase flow of (1.1) preserves the following symplectic structure:

$$dP\wedge dQ=dp\wedge dq; \tag{1.3}$$

i.e., the sum of the oriented areas of projections onto the coordinate planes $(p^1,q^1),\dots,(p^n,q^n)$ is an integral invariant [4]. To avoid confusion, we note that the differentials in (1.1) and (1.3) have different meanings. In (1.1) P, Q are treated as functions of time and p, q are fixed parameters, while differentiation in (1.3) is made with respect to the initial data p, q.

Let P_k, Q_k, $k=0,\dots,N$, $t_{k+1}-t_k=h_{k+1}$, $t_N=t_0+T$, be a method for (1.1) based on the one-step approximation $\bar P=\bar P(t+h;t,p,q)$, $\bar Q=\bar Q(t+h;t,p,q)$. We say that the method preserves symplectic structure if

$$d\bar P\wedge d\bar Q=dp\wedge dq\,. \tag{1.4}$$

The present paper deals with symplectic integration of the Hamiltonian system with multiplicative noise (1.1), (1.3). It is a continuation of [3], where symplectic methods for Hamiltonian systems with additive noise were proposed. For symplectic integration of deterministic Hamiltonian systems see, e.g., [5, 6, 7, 8, 9] and references therein.

As is known [5], in the case of deterministic general Hamiltonian systems symplectic Runge–Kutta (RK) methods are all implicit. Hence it is natural to expect that to construct symplectic methods for general Hamiltonian systems with multiplicative noise fully implicit methods are needed. The known implicit methods for stochastic systems with multiplicative noise (see [10, 11]) contain implicitness in deterministic terms only. In [12] an implicitness is introduced in stochastic terms as well. However, the methods of [12] are of a very special form. In section 2 a new class of fully implicit methods for general stochastic systems is proposed. Increments of Wiener processes in these implicit schemes are substituted by some truncated random variables. They are important for both theory and practice of numerical integration of SDEs. We use these fully implicit methods in section 3 to construct symplectic methods for general Hamiltonian systems with multiplicative noise. Section 4 is devoted to a special case of separable Hamiltonians. Explicit symplectic integrators are constructed for such systems. In addition, symplectic methods for Hamiltonian systems with small multiplicative noise can be found in the preprint [13]. There one can also find some Liouvillian methods for stochastic systems preserving phase volume. Let us recall that the mean-square methods of higher order contain repeated Ito integrals which are difficult for simulation. In this paper, we prefer to derive methods which are efficient with respect to simulation of the used random variables. That is why orders of the methods derived are not too high. In the last section of the paper we present numerical tests. They clearly demonstrate superiority of the proposed symplectic methods over very long times in comparison with nonsymplectic ones.

2. Fully implicit methods. Construction of implicit methods for stochastic systems with additive noise does not cause any difficulties in principle. However, all

is much more intricate in the case of stochastic systems with multiplicative noise. The known implicit methods for such systems (see [10, 11]) contain implicitness restricted to deterministic terms, e.g., to the drift terms in the implicit Euler scheme. In [12], an implicitness is introduced in stochastic terms as well. However, methods of [12] are of a very special form. In particular, this form does not allow us to construct symplectic methods for stochastic Hamiltonian systems with multiplicative noise. In this section we construct a sufficiently large class of fully implicit methods of mean-square order 1/2 for general stochastic systems. These results are of independent and general interest. That is why in this section we consider SDEs in the Ito sense, following the traditional way of developing numerical methods. At the same time we should note that the Stratonovich form is preferable for SDEs preserving integral invariants.

2.1. The main idea and an example. Let us start with an example. Consider the Ito scalar equation

$$dX = \sigma X dw(t). \tag{2.1}$$

The one-step approximation of the Euler method $\hat{X}$ for (2.1) is

$$\hat{X} = x + \sigma x \Delta w(h). \tag{2.2}$$

We can represent this approximation in the form

$$\hat{X} = x + \sigma \hat{X} \Delta w + \sigma (x - \hat{X}) \Delta w = x - \sigma^2 x (\Delta w)^2 + \sigma \hat{X} \Delta w.$$

As h is small, $(\Delta w)^2 \sim h$, and we obtain the following "natural" implicit method:

$$\tilde{X} = x - \sigma^2 x h + \sigma \tilde{X} \Delta w(h). \tag{2.3}$$

However, this method cannot be realized since $1 - \sigma \Delta w(h)$ can vanish for any small h. Further, for the formal value of $\tilde{X}$ from (2.3)

$$\tilde{X} = \frac{x(1 - \sigma^2 h)}{1 - \sigma \Delta w(h)},$$

we have $E|\tilde{X}| = \infty$. (See [10].) Clearly, the method (2.3) is not suitable. The reason for this is the unboundedness of the random variable $\Delta w(h)$ for any arbitrarily small h.

Our basic idea consists of replacement of $\Delta w(h) = \xi \sqrt{h}$, where ξ is an $\mathcal{N}(0,1)$-distributed random variable, by another random variable $\zeta \sqrt{h} = \zeta_h \sqrt{h}$ such that $\zeta \sqrt{h}$ is bounded and the Euler type method

$$\check{X} = x + \sigma x \zeta \sqrt{h} \tag{2.4}$$

is of the mean-square order 1/2 as well. To achieve this, it is sufficient to require

$$E(\check{X} - \hat{X}) = O(h^{3/2}), \qquad E(\check{X} - \hat{X})^2 = O(h^2). \tag{2.5}$$

We take a symmetric ζ. Then $E(\check{X} - \hat{X}) = 0$. To satisfy the second equation in (2.5), the condition $E(\zeta_h - \xi)^2 = O(h)$ is sufficient.

We shall require a stronger inequality,

$$E(\zeta_h - \xi)^2 \le h^k, \qquad k \ge 1. \tag{2.6}$$

For $A_h > 0$ let

$$\zeta_h = \begin{cases} \xi, \ |\xi| \le A_h, \\ A_h, \ \xi > A_h, \\ -A_h, \ \xi < -A_h. \end{cases} \tag{2.7}$$

Since

$$\begin{aligned} E(\zeta_h - \xi)^2 &= \frac{2}{\sqrt{2\pi}} \int_{A_h}^{\infty} (x - A_h)^2 e^{-x^2/2} dx \\ &= \frac{2}{\sqrt{2\pi}} e^{-A_h^2/2} \int_{A_h}^{\infty} y^2 e^{-y^2/2} e^{-A_h y} dy < e^{-A_h^2/2}, \end{aligned}$$

the inequality (2.6) is fulfilled if $e^{-A_h^2/2} \le h^k$, i.e., $A_h^2 \ge 2k|\ln h|$. Thus, if

$$A_h = \sqrt{2k|\ln h|}, \qquad k \ge 1,$$

then the method based on the one-step approximation (2.4) has the mean-square order 1/2.

LEMMA 2.1. *Let* $A_h = \sqrt{2k|\ln h|}$, $k \ge 1$, *and* ζ_h *be defined by* (2.7). *Then the following inequality holds:*

$$0 \le E(\xi^2 - \zeta_h^2) = 1 - E\zeta_h^2 \le \left(1 + 2\sqrt{2k|\ln h|}\right) h^k. \tag{2.8}$$

Proof. We have

$$\begin{aligned} 1 - E\zeta_h^2 &= \frac{2}{\sqrt{2\pi}} \int_{A_h}^{\infty} (x^2 - A_h^2) e^{-x^2/2} dx \\ &= \frac{2}{\sqrt{2\pi}} \int_{A_h}^{\infty} \left[(x - A_h)^2 + 2A_h(x - A_h)\right] e^{-x^2/2} dx \\ &\le e^{-A_h^2/2} + \frac{4A_h}{\sqrt{2\pi}} \int_{A_h}^{\infty} x e^{-x^2/2} dx = e^{-A_h^2/2} \left(1 + \frac{4A_h}{\sqrt{2\pi}}\right) \le (1 + 2A_h) e^{-A_h^2/2}, \end{aligned}$$

whence (2.8) follows. □

Now consider the following implicit method (for definiteness we put $k = 1$ and $A_h = \sqrt{2|\ln h|}$):

$$\begin{aligned} \bar{X} &= x - \sigma^2 x h + \sigma \bar{X} \zeta_h \sqrt{h}, \\ \bar{X} &= \frac{x(1 - \sigma^2 h)}{1 - \sigma \zeta_h \sqrt{h}}. \end{aligned} \tag{2.9}$$

Since $|\zeta_h| \le \sqrt{2|\ln h|}$, this method is realizable for all h satisfying the inequality

$$2h|\ln h| < \frac{1}{\sigma^2}. \tag{2.10}$$

PROPOSITION 2.2. *The method* (2.9) *is of the mean-square order* 1/2.

Proof. Let us compare method (2.9) with the Euler method (2.2). We get

$$E\bar{X} = x(1 - \sigma^2 h) E \sum_{m=0}^{\infty} \sigma^m \zeta_h^m h^{m/2} = x(1 - \sigma^2 h) E \sum_{m=0}^{\infty} \sigma^{2m} \zeta_h^{2m} h^m.$$

It is obvious from here that the principal term in the expansion of $E(\bar{X}-\hat{X})$ is equal to $x\sigma^2 h(E\zeta_h^2-1)$. Due to Lemma 2.1, we obtain for all sufficiently small h

$$|E(\bar{X}-\hat{X})| \le C|x|\sigma^2\left(1+2\sqrt{2|\ln h|}\right)h^2, \tag{2.11}$$

where C is a positive constant.

Further,

$$\begin{aligned} E(\bar{X}-\hat{X})^2 &= E\left(-\sigma^2 xh+\sigma\bar{X}\zeta_h\sqrt{h}-\sigma x\xi\sqrt{h}\right)^2 \\ &\le 2\sigma^4x^2h^2+2E\left(\sigma\bar{X}\zeta_h\sqrt{h}-\sigma x\xi\sqrt{h}\right)^2 \\ &= 2\sigma^4x^2h^2+2E\left(\sigma\cdot\left(x-\sigma^2xh+\sigma\bar{X}\zeta_h\sqrt{h}\right)\zeta_h\sqrt{h}-\sigma x\xi\sqrt{h}\right)^2 \\ &\le 2\sigma^4x^2h^2+2\sigma^2x^2hE(\zeta_h-\xi)^2+C_1x^2h^2\le C_2x^2h^2 \end{aligned} \tag{2.12}$$

for all sufficiently small h and some positive constants C_1 and C_2. The inequalities (2.11) and (2.12) imply the mean-square convergence of implicit method (2.9) with order $1/2$. □

Introduction of implicitness in the stochastic term leads to the appearance of the compensating term $-\sigma^2xh$ in (2.9). This can be explained in the following way. Since $\bar{X}$ must be close to $x+\sigma x\zeta_h\sqrt{h}$, the expression $x+\sigma\bar{X}\zeta_h\sqrt{h}$ is close to $x+\sigma x\zeta_h\sqrt{h}+\sigma^2x\zeta_h^2h$. Consequently, making use of the compensating term results in $x+\sigma\bar{X}\zeta_h\sqrt{h}-\sigma^2xh=x+\sigma x\zeta_h\sqrt{h}+\sigma^2x(\zeta_h^2-1)h\approx x+\sigma x\zeta_h\sqrt{h}$; i.e., we get the correct result.

Now let us consider the expression $\sigma((1-\beta)x+\beta\bar{X})\zeta_h\sqrt{h}$ which introduces implicitness in the stochastic term with the parameter $0\le\beta\le1$. Clearly, the compensating term in this case is equal to $-\sigma^2\beta xh$. Thus, we derive the following method:

$$\bar{X}=x-\sigma^2\beta xh+\sigma((1-\beta)x+\beta\bar{X})\zeta_h\sqrt{h},\qquad 0\le\beta\le1. \tag{2.13}$$

The following proposition can be proved analogously to Proposition 2.2.

PROPOSITION 2.3. *The method* (2.13)*, as well as the methods*

$$\bar{X}=x-\sigma^2\beta x\zeta_h^2h+\sigma((1-\beta)x+\beta\bar{X})\zeta_h\sqrt{h},\qquad 0\le\beta\le1, \tag{2.14}$$

$$\bar{X}=x-\sigma^2\beta((1-\alpha)x+\alpha\bar{X})h+\sigma((1-\beta)x+\beta\bar{X})\zeta_h\sqrt{h},\qquad 0\le\alpha,\beta\le1, \tag{2.15}$$

are of the mean-square order $1/2$.

2.2. Convergence theorem. Now we are in position to introduce fully implicit methods for general systems of SDEs. For simplicity in writing we deal here with the following scalar Ito SDE:

$$dX=a(t,X)dt+b(t,X)dw(t). \tag{2.16}$$

We suppose that $a(t,x)$, $b(t,x)$, $\frac{\partial b}{\partial x}(t,x)$ are continuous for $t_0\le t\le T$, $x\in\mathbf{R}$, and there exists a positive constant L such that

$$|a(t,y)-a(t,x)|\le L|y-x|,\quad \left|\frac{\partial b}{\partial x}(t,x)\right|\le L,\quad t_0\le t\le T,\ x,y\in\mathbf{R}. \tag{2.17}$$

Note that below the same letter L (or K, or C) is used for various constants.

Consider the implicit one-step approximation (cf. (2.9))

$$\bar{X} = x + a(t,\bar{X})h - b(t,x)\frac{\partial b}{\partial x}(t,x)h + b(t,\bar{X})\zeta_h\sqrt{h}, \tag{2.18}$$

where ζ_h is defined by (2.7) with $A_h = \sqrt{2|\ln h|}$ for definiteness.

LEMMA 2.4. *There exist constants $K > 0$ and $h_0 > 0$ such that for any $h \le h_0$, $t_0 \le t \le T$, $x \in \mathbf{R}$ (2.18) has a unique solution $\bar{X}$ which satisfies the inequality*

$$|\bar{X} - x| \le K(1+|x|)\left(|\zeta_h|\sqrt{h} + h\right). \tag{2.19}$$

The solution $\bar{X}$ of (2.18) can be found by the method of simple iteration with x as the initial approximation.

Proof. For any fixed t, x, and h, let us introduce the function

$$\varphi(z) = x + a(t,z)h - b(t,x)\frac{\partial b}{\partial x}(t,x)h + b(t,z)\zeta_h\sqrt{h}.$$

Then (2.18) can be written as

$$\bar{X} = \varphi(\bar{X}).$$

There is a positive constant C such that for any $z \in \mathbf{R}$

$$|\varphi(z) - x| \le |a(t,x)|h + |a(t,z) - a(t,x)|h + |b(t,x)||\zeta_h|\sqrt{h} + |b(t,z) - b(t,x)||\zeta_h|\sqrt{h}$$
$$+ \left|b(t,x)\frac{\partial b}{\partial x}(t,x)\right| h \le C(1+|x|)\left(|\zeta_h|\sqrt{h} + h\right) + L|z-x|\left(|\zeta_h|\sqrt{h} + h\right).$$

Further, for any z_1, $z_2 \in \mathbf{R}$

$$|\varphi(z_2) - \varphi(z_1)| \le L|z_2 - z_1|\left(|\zeta_h|\sqrt{h} + h\right).$$

Clearly, there exist positive constants K and h_0 such that for any $h \le h_0$, $x \in \mathbf{R}$

$$L\left(|\zeta_h|\sqrt{h} + h\right) < 1,$$

and if

$$|z - x| \le K(1+|x|)\left(|\zeta_h|\sqrt{h} + h\right),$$

then

$$|\varphi(z) - x| \le K(1+|x|)\left(|\zeta_h|\sqrt{h} + h\right).$$

Let us note that the constants K in the last two inequalities are the same. Now the lemma follows from the contraction mapping principle. □

In addition to (2.17) suppose that there exist continuous $\partial a/\partial t$, $\partial b/\partial t$, and $\partial^2 b/\partial x^2$ and the inequalities

$$\left|\frac{\partial a}{\partial t}(t,x)\right| \le L(1+|x|), \quad \left|\frac{\partial b}{\partial t}(t,x)\right| \le L(1+|x|), \quad t_0 \le t \le T, \quad x \in \mathbf{R} \tag{2.20}$$

hold.

THEOREM 2.5. *Assume* (2.17) *and* (2.20). *Let there exist* $\delta > 0$ *such that if* $|y - x| \le \delta(1 + |x|)$, *then the inequality*

$$\left| b(t,x)\frac{\partial^2 b}{\partial x^2}(t,y) \right| \le L, \quad t_0 \le t \le T \tag{2.21}$$

holds.

Then the implicit method based on the one-step approximation (2.18) *converges in mean-square with the order* $1/2$.

Proof. Let $\hat{X}$ be the Euler approximation for (2.16):

$$\hat{X} = x + a(t,x)h + b(t,x)\Delta w(h).$$

Using the condition (2.17) only, we get

$$\begin{aligned} E|\bar{X} - \hat{X}|^2 &\le E\left| a(t,\bar{X})h - a(t,x)h + b(t,\bar{X})\zeta_h\sqrt{h} - b(t,x)\Delta w(h) - b(t,x)\frac{\partial b}{\partial x}(t,x)h \right|^2 \\ &\le LE|a(t,\bar{X}) - a(t,x)|^2h^2 + LE|b(t,\bar{X}) - b(t,x)|^2\zeta_h^2 h \\ &\quad + Lb^2(t,x)E(\zeta_h - \xi)^2 h + L\left| b(t,x)\frac{\partial b}{\partial x}(t,x) \right|^2 h^2 \\ &\le LE|\bar{X} - x|^2h^2 + LE|\bar{X} - x|^2\zeta_h^2 h + L(1 + |x|)^2E(\zeta_h - \xi)^2 h + L(1 + |x|)^2h^2. \end{aligned}$$

It follows from here, Lemma 2.4, the inequality $E\zeta^4 < E\xi^4 = 3$, and (2.6) that

$$E|\bar{X} - \hat{X}|^2 \le L(1 + |x|)^2h^2. \tag{2.22}$$

Now let us proceed to the evaluation of $E(\bar{X} - \hat{X})$. We have

(2.23)
$$|E(\bar{X} - \hat{X})| \le |Ea(t,\bar{X}) - a(t,x)|h + \left| E(b(t,\bar{X}) - b(t,x))\zeta_h\sqrt{h} - b(t,x)\frac{\partial b}{\partial x}(t,x)h \right|.$$

Due to Lemma 2.4, $E|\bar{X} - x| \le K(1 + |x|)(E|\zeta_h|\sqrt{h} + h)$. Then

$$|Ea(t,\bar{X}) - a(t,x)|h \le C(1 + |x|)h^{3/2}. \tag{2.24}$$

We have

$$\begin{aligned} &(b(t,\bar{X}) - b(t,x))\zeta_h\sqrt{h} - b(t,x)\frac{\partial b}{\partial x}(t,x)h \\ &\quad = \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot (\bar{X} - x)\zeta_h\sqrt{h} - b(t,x)\frac{\partial b}{\partial x}(t,x)h \\ &\quad = \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot \left(a(t,\bar{X})h + b(t,\bar{X})\zeta_h\sqrt{h} - b(t,x)\frac{\partial b}{\partial x}(t,x)h \right)\zeta_h\sqrt{h} \\ &\qquad - b(t,x)\frac{\partial b}{\partial x}(t,x)h \\ &\quad = \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot \left(a(t,\bar{X}) - b(t,x)\frac{\partial b}{\partial x}(t,x)h \right)\zeta_h h^{3/2} \\ &\qquad + \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot b(t,\bar{X})\zeta_h^2 h - b(t,x)\frac{\partial b}{\partial x}(t,x)h, \end{aligned} \tag{2.25}$$

where $0 \le \theta \le 1$.

Since $|\bar{X} - x| \le \rho(1 + |x|)$, where $\rho \to 0$ as $h \to 0$, we get $|\bar{X}| \le |x| + |\bar{X} - x| \le K(1 + |x|)$ for all sufficiently small h. Therefore

$$\left| E \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot a(t, \bar{X}) \zeta_h h^{3/2} \right| \le KE|a(t, \bar{X})\zeta_h| h^{3/2} \le KE(1 + |\bar{X}|)|\zeta_h| h^{3/2} \le K(1 + |x|)h^{3/2}. \tag{2.26}$$

Clearly,

$$\left| E \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot b(t, x) \frac{\partial b}{\partial x}(t, x) \zeta_h h^{3/2} \right| \le K(1 + |x|)h^{3/2}.$$

Let us estimate the last two terms in (2.25). We obtain

$$\begin{aligned} &\frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot b(t, \bar{X}) \zeta_h^2 h - b(t, x) \frac{\partial b}{\partial x}(t, x) h \\ &= \left(\frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) - \frac{\partial b}{\partial x}(t, x) \right) b(t, \bar{X}) \zeta_h^2 h \\ &+ \frac{\partial b}{\partial x}(t, x)(b(t, \bar{X}) - b(t, x)) \zeta_h^2 h + \frac{\partial b}{\partial x}(t, x) b(t, x)(\zeta_h^2 - 1) h \\ &= \frac{\partial^2 b}{\partial x^2}(t, x + \theta_1(\bar{X} - x)) \cdot \theta(\bar{X} - x) \cdot b(t, \bar{X}) \zeta_h^2 h \\ &+ \frac{\partial b}{\partial x}(t, x) \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot (\bar{X} - x) \zeta_h^2 h + \frac{\partial b}{\partial x}(t, x) b(t, x)(\zeta_h^2 - 1) h, \end{aligned}$$

where $0 \le \theta, \theta_1 \le 1$. Due to Lemma 2.4, we get $|x + \theta_1(\bar{X} - x) - \bar{X}| \le |\bar{X} - x| \le K(|\zeta_h|\sqrt{h} + h)(1 + |x|)$. For all sufficiently small h we have $K(|\zeta_h|\sqrt{h} + h) < \delta$ and consequently due to (2.21)

$$\left| \frac{\partial^2 b}{\partial x^2}(t, x + \theta_1(\bar{X} - x)) \cdot b(t, \bar{X}) \right| \le L. \tag{2.27}$$

Using (2.27), the conditions (2.17), and Lemmas 2.1 and 2.4, we obtain for the last two terms in (2.25)

$$\left| E \frac{\partial b}{\partial x}(t, x + \theta(\bar{X} - x)) \cdot b(t, \bar{X}) \zeta_h^2 h - b(t, x) \frac{\partial b}{\partial x}(t, x) h \right| \le K(1 + |x|)h^{3/2}. \tag{2.28}$$

Thus, (2.23)–(2.28) give

$$|E(\bar{X} - \hat{X})| \le K(1 + |x|)h^{3/2}. \tag{2.29}$$

It follows from (2.22) and (2.29) (see [10]) that the method based on (2.18) is of the mean-square order 1/2. □

REMARK 2.1. *The condition* (2.21) *is satisfied if, for instance,*

$$|b(t, x)| \le L, \quad \left| \frac{\partial^2 b}{\partial x^2}(t, x) \right| \le L, \quad t_0 \le t \le T, \quad x \in \mathbf{R} \tag{2.30}$$

or

$$\left| \frac{\partial^2 b}{\partial x^2}(t, x) \right| \le \frac{L}{1 + |x|}, \quad t_0 \le t \le T, \quad x \in \mathbf{R} \tag{2.31}$$

holds.

Let us underline that the conditions of Theorem 2.5 *are not necessary and the method is applicable more widely. This is true for other methods proposed in the paper as well.*

REMARK 2.2. *Let the function* $c(t,x) := b(t,x)\frac{\partial b}{\partial x}(t,x)$ *satisfy the condition*

$$|c(t,y) - c(t,x)| \le L|y-x|. \tag{2.32}$$

Consider the implicit one-step approximation

$$\bar{X} = x + a(t,\bar{X})h - b(t,\bar{X})\frac{\partial b}{\partial x}(t,\bar{X})h + b(t,\bar{X})\zeta_h\sqrt{h}. \tag{2.33}$$

It is not difficult to prove that Theorem 2.5 *is true for the implicit method based on* (2.33) *provided* (2.32) *is fulfilled.*

2.3. General construction. Let

$$dX^i = a^i(t,X)dt + \sum_{r=1}^{m} b_r^i(t,X)dw_r(t),\ i = 1,\dots,d. \tag{2.34}$$

Introduce the following one-step approximation:

(2.35)

$$\begin{aligned}\bar{X}^i = x^i &+ \sum_{k=1}^{l}\lambda_k^i a^i(t+\nu_k^i h, (1-\alpha_{k1}^i)x^1 + \alpha_{k1}^i\bar{X}^1,\dots,(1-\alpha_{kd}^i)x^d + \alpha_{kd}^i\bar{X}^d)h\\ &+ \sum_{r=1}^{m}\sum_{k=1}^{l}\mu_{rk}^i b_r^i(t+\nu_{rk}^i h, (1-\beta_{rk1}^i)x^1 + \beta_{rk1}^i\bar{X}^1,\dots,(1-\beta_{rkd}^i)x^d + \beta_{rkd}^i\bar{X}^d)\zeta_{rh}\sqrt{h}\\ &+ A^i,\end{aligned}$$

where $0 \le \nu,\alpha,\beta \le 1,\ \lambda,\mu \ge 0,\ \sum_{k=1}^{l}\lambda_k^i = 1,\ \sum_{k=1}^{l}\mu_{rk}^i = 1,\ i = 1,\dots,d,\ l$ is a positive integer, and A^i are some expressions to be found. Substituting the Euler-type approximation

$$\hat{X}^j = x^j + a^j(t,x)h + \sum_{s=1}^{m} b_s^j(t,x)\zeta_{sh}\sqrt{h}$$

instead of $\bar{X}^j,\ j = 1,\dots,d$, in b_r^i, we obtain

$$\begin{aligned}&b_r^i(t+\nu_{rk}^i h, (1-\beta_{rk1}^i)x^1 + \beta_{rk1}^i\bar{X}^1,\dots,(1-\beta_{rkd}^i)x^d + \beta_{rkd}^i\bar{X}^d)\\ &\quad\approx b_r^i(t,x) + \sum_{j=1}^{d}\frac{\partial b_r^i}{\partial x^j}(t,x)\beta_{rkj}^i\sum_{s=1}^{m} b_s^j(t,x)\zeta_{sh}\sqrt{h}.\end{aligned}$$

It is clear from here that either

$$A^i = -\sum_{r=1}^{m}\sum_{k=1}^{l}\mu_{rk}^i\sum_{j=1}^{d}\frac{\partial b_r^i}{\partial x^j}(t,x)\beta_{rkj}^i\sum_{s=1}^{m} b_s^j(t,x)\zeta_{sh}\sqrt{h}\zeta_{rh}\sqrt{h} \tag{2.36}$$

or

$$A^i = -\sum_{r=1}^{m}\sum_{k=1}^{l}\mu_{rk}^i\sum_{j=1}^{d}\frac{\partial b_r^i}{\partial x^j}(t,x)\beta_{rkj}^i b_r^j(t,x)h \tag{2.37}$$

can be put in (2.35).

Substituting one of these expressions in (2.35), we obtain a multiparameter family of implicit methods. It is also possible to introduce implicitness in A^i by changing t, x as it was done in the terms connecting with a^i. Moreover, the family can be extended if some a^i or b^i_r are represented as sums of terms. In this case the coefficients λ, ν, α, μ, β can differ for different terms.

It can be proved that under appropriate conditions of smoothness and boundedness on the coefficients of (2.34) the method based on the one-step approximation (2.35) with A^i as in (2.36) or (2.37) is of the mean-square order 1/2. The proof is analogous to the proof of Theorem 2.5.

Here and below we will not precisely indicate conditions on the coefficients a and b_r assuming that appropriate conditions on the coefficients hold. These conditions can be restored using the general theory [10] and Theorem 2.5. (See also Remarks 2.1 and 2.2.)

Let us give an example of fully implicit methods:

$$\bar{X} = x + a(t, \bar{X})h - \sum_{r=1}^{m}\sum_{j=1}^{d} \frac{\partial b_r}{\partial x^j}(t, \bar{X}) b_r^j(t, \bar{X})h + \sum_{r=1}^{m} b_r(t, \bar{X})\zeta_{rh}\sqrt{h}.$$

Further, in the case of SDEs in the sense of Stratonovich,

$$dX = a(t, X)dt + \sum_{r=1}^{m} b_r(t, X) \circ dw_r(t), \tag{2.38}$$

we construct the following derivative-free fully implicit method (midpoint method):

$$X_{k+1} = X_k + a\left(t_k + \frac{h}{2}, \frac{X_k + X_{k+1}}{2}\right)h + \sum_{r=1}^{m} b_r\left(t_k, \frac{X_k + X_{k+1}}{2}\right)(\zeta_{rh})_k\sqrt{h}. \tag{2.39}$$

For $b^i_r = 0$, this method coincides with the well-known deterministic midpoint scheme, which has the second order of convergence.

In the general case the method (2.39) is of the mean-square order 1/2. In the commutative case, i.e., when $\Lambda_i b_r = \Lambda_r b_i$ (here the operator $\Lambda_r := (b_r, \partial/\partial x)$), or in the case of a system with one noise (i.e., $m = 1$), the midpoint method (2.39) has the first mean-square order of convergence which is stated in the next theorem. (Its proof is available in the preprint [13].)

THEOREM 2.6. *Suppose that the commutative conditions* $\Lambda_i b_r = \Lambda_r b_i$, $i, r = 1, \dots, m$, *are fulfilled. Let* ζ_{rh} *be defined by* (2.7) *with* $A_h = \sqrt{4|\ln h|}$. *Then the method* (2.39) *for the system* (2.38) *has the first mean-square order of convergence.*

3. Symplectic methods for the general Hamiltonian system. Here, using the results of the previous section, we construct symplectic methods for the general Hamiltonian system with multiplicative noise (1.1), (1.3). Its Ito form reads

$$\begin{aligned} dP &= f dt + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n} \frac{\partial \sigma_r}{\partial p^j}\sigma_r^j dt + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n} \frac{\partial \sigma_r}{\partial q^j}\gamma_r^j dt + \sum_{r=1}^{m}\sigma_r dw_r(t), \\ dQ &= g dt + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n} \frac{\partial \gamma_r}{\partial p^j}\sigma_r^j dt + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n} \frac{\partial \gamma_r}{\partial q^j}\gamma_r^j dt + \sum_{r=1}^{m}\gamma_r dw_r(t). \end{aligned} \tag{3.1}$$

Introduce the following implicit method:

$$P_{k+1} = P_k + fh - \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\left(\frac{\partial\sigma_r}{\partial p^j}\sigma_r^j - \frac{\partial\sigma_r}{\partial q^j}\gamma_r^j\right)h + \sum_{r=1}^{m}\sigma_r\cdot(\zeta_{rh})_k\sqrt{h},$$
$$Q_{k+1} = Q_k + gh - \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\left(\frac{\partial\gamma_r}{\partial p^j}\sigma_r^j - \frac{\partial\gamma_r}{\partial q^j}\gamma_r^j\right)h + \sum_{r=1}^{m}\gamma_r\cdot(\zeta_{rh})_k\sqrt{h}, \tag{3.2}$$

where all the functions have t, P_{k+1}, Q_k as their arguments.

THEOREM 3.1. *The implicit method* (3.2) *for the system* (3.1), (1.3) (*or for system* (1.1), (1.3)) *is symplectic and of the mean-square order* 1/2.

Proof. The method (3.2) belongs to the family (2.35) and, consequently, the assertion about its order of convergence follows from the previous section. Let us prove symplecticness of the method. It is convenient to write the one-step approximation corresponding to (3.2) in the form

$$\bar{P}^i = p^i - \frac{\partial H_0}{\partial q^i}h - \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial^2 H_r}{\partial q^i\partial p^j}\frac{\partial H_r}{\partial q^j}h - \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial^2 H_r}{\partial q^i\partial q^j}\frac{\partial H_r}{\partial p^j}h - \sum_{r=1}^{m}\frac{\partial H_r}{\partial q^i}\zeta_{rh}\sqrt{h},$$
$$\bar{Q}^i = q^i + \frac{\partial H_0}{\partial p^i}h + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial^2 H_r}{\partial p^i\partial p^j}\frac{\partial H_r}{\partial q^j}h + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial^2 H_r}{\partial p^i\partial q^j}\frac{\partial H_r}{\partial p^j}h + \sum_{r=1}^{m}\frac{\partial H_r}{\partial p^i}\zeta_{rh}\sqrt{h}, \tag{3.3}$$

where $i = 1, \dots, n$ and all the functions have $t, \bar{P}, q$ as their arguments.

Introduce the following function $F(t,p,q)$ (h, ζ_{rh} are fixed here):

$$F(t,p,q) = H_0(t,p,q)h + \frac{1}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial H_r}{\partial q^j}(t,p,q)\frac{\partial H_r}{\partial p^j}(t,p,q)h + \sum_{r=1}^{m}H_r(t,p,q)\zeta_{rh}\sqrt{h}.$$

Then (3.3) can be written as

$$\bar{P}^i = p^i - \frac{\partial F}{\partial q^i}(t,\bar{P},q),$$
$$\bar{Q}^i = q^i + \frac{\partial F}{\partial p^i}(t,\bar{P},q). \tag{3.4}$$

We have (the arguments everywhere are t, $\bar{P}$, q)

$$\sum_{i=1}^{n}d\bar{P}^i\wedge d\bar{Q}^i = \sum_{i=1}^{n}d\bar{P}^i\wedge\left(dq^i + \sum_{j=1}^{n}F''_{p^ip^j}d\bar{P}^j + \sum_{j=1}^{n}F''_{p^iq^j}dq^j\right)$$
$$= \sum_{i=1}^{n}d\bar{P}^i\wedge dq^i + \sum_{i=1}^{n}\sum_{j=1}^{n}F''_{p^ip^j}d\bar{P}^i\wedge d\bar{P}^j + \sum_{i=1}^{n}\sum_{j=1}^{n}F''_{p^iq^j}d\bar{P}^i\wedge dq^j.$$

Since $d\bar{P}^i\wedge d\bar{P}^j = -d\bar{P}^j\wedge d\bar{P}^i$, we get

$$\sum_{i=1}^{n}d\bar{P}^i\wedge d\bar{Q}^i = \sum_{i=1}^{n}d\bar{P}^i\wedge dq^i + \sum_{i=1}^{n}\sum_{j=1}^{n}F''_{p^iq^j}d\bar{P}^i\wedge dq^j$$
$$= \sum_{i=1}^{n}d\bar{P}^i\wedge dq^i + \sum_{i=1}^{n}\sum_{j=1}^{n}F''_{q^ip^j}d\bar{P}^j\wedge dq^i. \tag{3.5}$$

Further,

$$d\bar{P}^i = dp^i - \sum_{j=1}^{n} F''_{q^i p^j} d\bar{P}^j - \sum_{j=1}^{n} F''_{q^i q^j} dq^j.$$

Substituting $\sum_{j=1}^{n} F''_{q^i p^j} d\bar{P}^j$ from here in (3.5), we obtain

$$\sum_{i=1}^{n} d\bar{P}^i \wedge d\bar{Q}^i = \sum_{i=1}^{n} d\bar{P}^i \wedge dq^i + \sum_{i=1}^{n} \left(dp^i - d\bar{P}^i - \sum_{j=1}^{n} F''_{q^i q^j} dq^j \right) \wedge dq^i$$
$$= \sum_{i=1}^{n} dp^i \wedge dq^i - \sum_{i=1}^{n} \sum_{j=1}^{n} F''_{q^i q^j} dq^j \wedge dq^i = \sum_{i=1}^{n} dp^i \wedge dq^i. \qquad \square$$

A more general symplectic method for the Hamiltonian system (1.1), (1.3) has the form

$$\begin{aligned} P_{k+1} &= P_k + f(t_k + \beta h, \alpha P_{k+1} + (1-\alpha)P_k, (1-\alpha)Q_{k+1} + \alpha Q_k)h \\ &\quad + \left(\frac{1}{2} - \alpha\right) \sum_{r=1}^{m} \sum_{j=1}^{n} \left(\frac{\partial \sigma_r}{\partial p^j} \sigma_r^j - \frac{\partial \sigma_r}{\partial q^j} \gamma_r^j \right) h + \sum_{r=1}^{m} \sigma_r \cdot (\zeta_{rh})_k \sqrt{h}, \\ Q_{k+1} &= Q_k + g(t_k + \beta h, \alpha P_{k+1} + (1-\alpha)P_k, (1-\alpha)Q_{k+1} + \alpha Q_k)h \\ &\quad + \left(\frac{1}{2} - \alpha\right) \sum_{r=1}^{m} \sum_{j=1}^{n} \left(\frac{\partial \gamma_r}{\partial p^j} \sigma_r^j - \frac{\partial \gamma_r}{\partial q^j} \gamma_r^j \right) h + \sum_{r=1}^{m} \gamma_r \cdot (\zeta_{rh})_k \sqrt{h}, \end{aligned} \tag{3.6}$$

where σ_r, γ_r, $r = 1, \dots, m$, and their derivatives are calculated at $(t_k, \alpha P_{k+1} + (1-\alpha)P_k, (1-\alpha)Q_{k+1} + \alpha Q_k)$, and α, $\beta \in [0,1]$ are parameters.

Using arguments similar to ones in the proof of Theorem 3.1, we obtain the following theorem.

THEOREM 3.2. *The implicit method* (3.6) *for the system* (1.1), (1.3) (*or for system* (3.1), (1.3)) *is symplectic and of the mean-square order* 1/2.

The method (3.2) is a particular case of (3.6) when $\alpha = 1$, $\beta = 0$. If $\alpha = \beta = 1/2$ the method (3.6) becomes the midpoint method (cf. (2.39)):

$$\begin{aligned} P_{k+1} &= P_k + f\left(t_k + \frac{h}{2}, \frac{P_k + P_{k+1}}{2}, \frac{Q_k + Q_{k+1}}{2}\right) h \\ &\quad + \sum_{r=1}^{m} \sigma_r \left(t_k, \frac{P_k + P_{k+1}}{2}, \frac{Q_k + Q_{k+1}}{2}\right) (\zeta_{rh})_k \sqrt{h}, \\ Q_{k+1} &= Q_k + g\left(t_k + \frac{h}{2}, \frac{P_k + P_{k+1}}{2}, \frac{Q_k + Q_{k+1}}{2}\right) h \\ &\quad + \sum_{r=1}^{m} \gamma_r \left(t_k, \frac{P_k + P_{k+1}}{2}, \frac{Q_k + Q_{k+1}}{2}\right) (\zeta_{rh})_k \sqrt{h}. \end{aligned} \tag{3.7}$$

REMARK 3.1. *In the commutative case, i.e., when* $\Lambda_i b_r = \Lambda_r b_i$, *or in the case of a system with one noise* (*i.e.*, $m = 1$), *the symplectic method* (3.7) *for* (1.1), (1.3) *has the first mean-square order of convergence.*

REMARK 3.2. *In the case of Hamiltonians that are separable in the noise part, i.e., when* $H_r(t,p,q) = U_r(t,q) + V_r(t,p)$, $r = 1, \dots, m$, *we can obtain symplectic*

methods for (1.1), (1.3) *which are explicit in stochastic terms and do not need truncated random variables. For instance,* (3.2) *acquires the form*

$$
\begin{aligned}
P_{k+1} &= P_k + f(t_k, P_{k+1}, Q_k)h \\
&\quad + \frac{h}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial \sigma_r}{\partial q^j}(t_k, Q_k)\cdot\gamma_r^j(P_{k+1}) + \sum_{r=1}^{m}\sigma_r(t_k, Q_k)\Delta_k w_r, \\
Q_{k+1} &= Q_k + g(t_k, P_{k+1}, Q_k)h \\
&\quad - \frac{h}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial \gamma_r}{\partial p^j}(P_{k+1})\cdot\sigma_r^j(t_k, Q_k) + \sum_{r=1}^{m}\gamma_r(t_k, P_{k+1})\Delta_k w_r.
\end{aligned} \tag{3.8}
$$

Of course, if it is necessary, fully implicit methods which require truncated random variables can be used in the case of separable Hamiltonians as well.

REMARK 3.3. *It is possible to construct fully explicit symplectic methods for the following partitioned system:*

$$
\begin{aligned}
dP &= f(t,Q)dt + \sum_{r=1}^{m}\sigma_r(t,Q)\circ dw_r(t),\ P(t_0) = p, \\
dQ &= g(P)dt + \sum_{r=1}^{m}\gamma_r(t)dw_r(t),\ Q(t_0) = q,
\end{aligned} \tag{3.9}
$$

with $f^i = -\partial U_0/\partial q^i$, $g^i = \partial V_0/\partial p^i$, $\sigma_r^i = -\partial U_r/\partial q^i$, $r = 1,\dots,m$, $i = 1,\dots,n$.

For instance, the explicit partitioned Runge–Kutta (*PRK*) *method* (*cf.* (4.5)–(4.6))

$$
\begin{aligned}
\mathcal{Q}_1 &= Q_k + \alpha h g(P_k), \\
\mathcal{P}_1 &= P_k + hf(t_k + \alpha h, \mathcal{Q}_1) + \frac{h}{2}\sum_{r=1}^{m}\sum_{j=1}^{n}\frac{\partial \sigma_r}{\partial q^j}(t_k, \mathcal{Q}_1)\cdot\gamma_r^j(t_k), \\
\mathcal{Q}_2 &= \mathcal{Q}_1 + (1-\alpha)hg(\mathcal{P}_1),
\end{aligned} \tag{3.10}
$$

$$
\begin{aligned}
P_{k+1} &= \mathcal{P}_1 + \sum_{r=1}^{m}\sigma_r(t_k, \mathcal{Q}_2)\Delta_k w_r, \\
Q_{k+1} &= \mathcal{Q}_2 + \sum_{r=1}^{m}\gamma_r(t_k)\Delta_k w_r,\ k = 0,\dots,N-1,
\end{aligned} \tag{3.11}
$$

with the parameter $0 \le \alpha \le 1$, *is symplectic and of the mean-square order* $1/2$.

A particular case of the system (3.9) *is considered in the next section, where explicit symplectic methods of a higher order are proposed.*

4. Explicit symplectic methods in the case of separable Hamiltonians. Consider a special case of the Hamiltonian system (1.1), (1.3) such that

$$
H_0(t,p,q) = V_0(p) + U_0(t,q), \quad H_r(t,p,q) = U_r(t,q), \quad r = 1,\dots,m. \tag{4.1}
$$

In this case we get the following system in the sense of Stratonovich:

$$
\begin{aligned}
dP &= f(t,Q)dt + \sum_{r=1}^{m}\sigma_r(t,Q)\circ dw_r(t), \qquad P(t_0) = p, \\
dQ &= g(P)dt,\ Q(t_0) = q,
\end{aligned} \tag{4.2}
$$

with

(4.3)

$$f^i = -\partial U_0/\partial q^i, \quad g^i = \partial V_0/\partial p^i, \quad \sigma_r^i = -\partial U_r/\partial q^i, \quad r = 1, \dots m, \quad i = 1, \dots, n.$$

We note that it is not difficult to consider a slightly more general separable Hamiltonian $H_0(t,p,q) = V_0(t,p) + U_0(t,q)$, but we restrict ourselves to H_0 from (4.1).

It is obvious that the system (4.2) has the same form in the sense of Ito.

For $V_0(p) = \frac{1}{2}(M^{-1}p, p)$ with M a constant, symmetric, invertible matrix, the system (4.2) takes the form

$$dP = f(t,Q)dt + \sum_{r=1}^{m} \sigma_r(t,Q)dw_r(t), \qquad P(t_0) = p, \tag{4.4}$$
$$dQ = M^{-1}Pdt, \qquad Q(t_0) = q.$$

This system can be written as a second-order differential equation with multiplicative noise. Some physical applications of stochastic symplectic integration for such systems are discussed in [14].

Due to specific features of the system (4.2), (4.3) we have succeeded in construction of explicit partitioned Runge-Kutta (PRK) methods of a higher order.

4.1. First-order methods. A PRK method for (4.2) has the form (cf. (3.10)–(3.11)):

$$\mathcal{Q}_1 = Q_k + \alpha h g(P_k), \qquad \mathcal{P}_1 = P_k + hf(t_k + \alpha h, \mathcal{Q}_1), \tag{4.5}$$
$$\mathcal{Q}_2 = \mathcal{Q}_1 + (1-\alpha)hg(\mathcal{P}_1),$$

$$P_{k+1} = \mathcal{P}_1 + \sum_{r=1}^{m} \sigma_r(t_k, \mathcal{Q}_2)\Delta_k w_r, \quad Q_{k+1} = \mathcal{Q}_2, \quad k = 0, \dots, N-1, \tag{4.6}$$

where $0 \le \alpha \le 1$ is a parameter.

THEOREM 4.1. *The explicit method* (4.5)–(4.6) *for the system* (4.2) *with* (4.3) *is symplectic and of the first mean-square order.*

Proof. In the case of the system (4.2) the operators Λ_r take the form $\Lambda_r = (\sigma_r, \partial/\partial p)$. Since σ_r do not depend on p, we get $\Lambda_i \sigma_j = 0$. It is known [10] that in such a case the Euler method has the first mean-square order of accuracy. Comparing the method (4.5)–(4.6) with the Euler method, it is not difficult to get that the method (4.5)–(4.6) is of the first mean-square order as well.

Due to (4.3), $\partial \sigma_r^i/\partial q^j = \partial \sigma_r^j/\partial q^i$. Using this, we obtain $dP_{k+1} \wedge dQ_{k+1} = d\mathcal{P}_1 \wedge d\mathcal{Q}_2$. It is easy to prove that $d\mathcal{P}_1 \wedge d\mathcal{Q}_2 = d\mathcal{P}_1 \wedge d\mathcal{Q}_1 = dP_k \wedge dQ_k$. Therefore the method (4.5)–(4.6) is symplectic. □

REMARK 4.1. *By swapping the roles of p and q, we can propose the following symplectic method of the first mean-square order for the system* (4.2) *with* (4.3)*:*

$$\mathcal{P} = P_k + \alpha h f(t_k, Q_k), \qquad \mathcal{Q} = Q_k + hg(\mathcal{P}), \tag{4.7}$$

$$P_{k+1} = \mathcal{P} + (1-\alpha)hf(t_{k+1}, \mathcal{Q}) + \sum_{r=1}^{m} \sigma_r(t_k, \mathcal{Q})\Delta_k w_r, \qquad Q_{k+1} = \mathcal{Q}. \tag{4.8}$$

4.2. Methods of order 3/2. Consider the relations

$$\mathcal{P}_i = p + h\sum_{j=1}^{s}\alpha_{ij}f(t+c_jh,\mathcal{Q}_j) + \sum_{j=1}^{s}\sum_{r=1}^{m}\sigma_r(t+d_jh,\mathcal{Q}_j)\left(\lambda_{ij}\varphi_r+\mu_{ij}\psi_r\right), \tag{4.9}$$

$$\mathcal{Q}_i = q + h\sum_{j=1}^{s}\hat{\alpha}_{ij}g(\mathcal{P}_j),\quad i=1,\dots,s,$$

$$\bar{P} = p + h\sum_{i=1}^{s}\beta_i f(t+c_ih,\mathcal{Q}_i) + \sum_{i=1}^{s}\sum_{r=1}^{m}\sigma_r(t+d_ih,\mathcal{Q}_i)\left(\nu_i\varphi_r+\varkappa_i\psi_r\right), \tag{4.10}$$

$$\bar{Q} = q + h\sum_{i=1}^{s}\hat{\beta}_i g(\mathcal{P}_i),$$

where φ_r, ψ_r do not depend on p and q, the parameters α_{ij}, $\hat{\alpha}_{ij}$, β_i, $\hat{\beta}_i$, λ_{ij}, μ_{ij}, ν_i, $\varkappa_i$ satisfy the conditions

$$\beta_i\hat{\alpha}_{ij} + \hat{\beta}_j\alpha_{ji} - \beta_i\hat{\beta}_j = 0, \tag{4.11}$$

$$\nu_i\hat{\alpha}_{ij} + \hat{\beta}_j\lambda_{ji} - \nu_i\hat{\beta}_j = 0,\quad \varkappa_i\hat{\alpha}_{ij} + \hat{\beta}_j\mu_{ji} - \varkappa_i\hat{\beta}_j = 0,\quad i,j=1,\dots,s,$$

and c_i, d_i are arbitrary parameters.

If $\sigma_r \equiv 0$ the relations (4.9)–(4.10) coincide with a general form of s-stage PRK methods for deterministic differential equations. (See, e.g., [5, p. 34].) It is known [9, 5] that the symplectic condition holds for $\bar{P}$, $\bar{Q}$ from (4.9)–(4.10) with (4.11) in the case of $\sigma_r \equiv 0$. By a generalization of the proof of Theorem 6.2 from [5], we prove the following lemma. (Another generalization is given in [3].)

LEMMA 4.2. *The relations* (4.9)–(4.10) *with conditions* (4.11) *preserve symplectic structure, i.e.,* $d\bar{P}\wedge d\bar{Q} = dp\wedge dq$.

Proof. Denote for awhile: $f_i = f(t+c_ih,\mathcal{Q}_i)$, $g_i = g(\mathcal{P}_i)$, $\sigma_{ri} = \sigma_r(t+d_ih,\mathcal{Q}_i)$. We get

$$\begin{aligned} d\bar{P}\wedge d\bar{Q} = dp\wedge dq + h\sum_{j=1}^{s}\hat{\beta}_j dp\wedge dg_j + h\sum_{i=1}^{s}\beta_i df_i\wedge dq + h^2\sum_{i=1}^{s}\sum_{j=1}^{s}\beta_i\hat{\beta}_j df_i\wedge dg_j \\ + \sum_{i=1}^{s}\sum_{r=1}^{m}\left(\nu_i\varphi_r+\varkappa_i\psi_r\right)d\sigma_{ri}\wedge dq + h\sum_{i=1}^{s}\sum_{j=1}^{s}\sum_{r=1}^{m}\left(\nu_i\varphi_r+\varkappa_i\psi_r\right)\hat{\beta}_j d\sigma_{ri}\wedge dg_j. \end{aligned} \tag{4.12}$$

Then we express $dp\wedge dg_j$ from

$$d\mathcal{P}_j\wedge dg_j = dp\wedge dg_j + h\sum_{i=1}^{s}\alpha_{ji}df_i\wedge dg_j + \sum_{i=1}^{s}\sum_{r=1}^{m}\left(\lambda_{ji}\varphi_r+\mu_{ji}\psi_r\right)d\sigma_{ri}\wedge dg_j$$

and substitute it in (4.12). Analogously, we act with $df_i\wedge dq$ and $d\sigma_{ri}\wedge dq$ finding them from the expressions for $df_i\wedge d\mathcal{Q}_i$ and $d\sigma_{ri}\wedge d\mathcal{Q}_i$. As a result, using (4.11), we obtain

$$\begin{aligned} d\bar{P}\wedge d\bar{Q} = dp\wedge dq + h\sum_{i=1}^{s}\hat{\beta}_i d\mathcal{P}_i\wedge dg_i + h\sum_{i=1}^{s}\beta_i df_i\wedge d\mathcal{Q}_i \\ + \sum_{i=1}^{s}\sum_{r=1}^{m}\left(\nu_i\varphi_r+\varkappa_i\psi_r\right)d\sigma_{ri}\wedge d\mathcal{Q}_i. \end{aligned}$$

Taking into account that the wedge product is skew-symmetric, the vector functions f, g, σ_r are gradients, f, σ_r do not depend on p, and g does not depend on q, it is not difficult to see that each of the terms $d\mathcal{P}_i \wedge dg_i$, $df_i \wedge d\mathcal{Q}_i$, $d\sigma_{ri} \wedge d\mathcal{Q}_i$ vanishes. Therefore $d\bar{P} \wedge d\bar{Q} = dp \wedge dq$. □

Introduce the following 2-stage explicit PRK method for the system (4.2), (4.3):

$$\mathcal{Q}_1 = Q_k, \qquad \mathcal{P}_1 = P_k + \frac{h}{4} f(t_k, \mathcal{Q}_1) + \frac{1}{2}\sum_{r=1}^{m} \sigma_r(t_k, \mathcal{Q}_1)\,(3(J_{r0})_k - \Delta_k w_r)\,, \tag{4.13}$$

$$\mathcal{Q}_2 = \mathcal{Q}_1 + \frac{2}{3} h g(\mathcal{P}_1),$$

$$\mathcal{P}_2 = \mathcal{P}_1 + \frac{3}{4} h f\left(t_k + \frac{2}{3}h, \mathcal{Q}_2\right) + \frac{3}{2}\sum_{r=1}^{m} \sigma_r\left(t_k + \frac{2}{3}h, \mathcal{Q}_2\right)(-(J_{r0})_k + \Delta_k w_r)\,,$$

$$P_{k+1} = \mathcal{P}_2, \quad Q_{k+1} = \mathcal{Q}_2 + \frac{h}{3} g(\mathcal{P}_2), \quad k = 0, \dots, N-1, \tag{4.14}$$

where

$$J_{r0} := \frac{1}{h}\int_t^{t+h} (w_r(\vartheta) - w_r(t))\, d\vartheta. \tag{4.15}$$

THEOREM 4.3. *The explicit PRK method* (4.13)–(4.14) *for system* (4.2), (4.3) *preserves symplectic structure and has the mean-square order* 3/2.

Proof. The method (4.13)–(4.14) has the form of (4.9)–(4.10), and its parameters satisfy the conditions (4.11). Then, Lemma 4.2 implies that this method preserves symplectic structure.

Now let us prove mean-square order of convergence of (4.13)–(4.14). To this end, introduce the one-step approximation for (4.2):

$$\tilde{P} = p + \sum_{r=1}^{m} \sigma_r \Delta w_r + hf + \sum_{r=1}^{m}\left[\frac{\partial \sigma_r}{\partial t} + \sum_{i=1}^{n} g^i \frac{\partial \sigma_r}{\partial q^i}\right] I_{0r} + \frac{h^2}{2}\left[\frac{\partial f}{\partial t} + \sum_{i=1}^{n} g^i \frac{\partial f}{\partial q^i}\right], \tag{4.16}$$

$$\tilde{Q} = q + hg + \sum_{r=1}^{m}\sum_{i=1}^{n} \sigma_r^i \frac{\partial g}{\partial p^i} I_{r0} + \frac{h^2}{2}\left[\sum_{i=1}^{n} f^i \frac{\partial g}{\partial p^i} + \frac{1}{2}\sum_{r=1}^{m}\sum_{i,j=1}^{n} \sigma_r^i \sigma_r^j \frac{\partial^2 g}{\partial p^i \partial p^j}\right],$$

where

$$I_{0r} = \int_t^{t+h} (\vartheta - t)\, dw_r(\vartheta), \qquad I_{r0} = \int_t^{t+h} (w_r(\vartheta) - w_r(t))\, d\vartheta = hJ_{r0}, \tag{4.17}$$

and all the coefficients are calculated at (t, p, q). We note that

$$(\Delta w_r - J_{r0})\, h = I_{0r}.$$

Using the general theory of numerical integration of SDEs [10], it is not difficult to show that the method based on (4.16) is of the mean-square order 3/2. Our nearest aim is to prove that the one-step approximation $\bar{P}$, $\bar{Q}$ corresponding to the method (4.13)–(4.14) is such that

$$\left| E\begin{bmatrix} \bar{P} - \tilde{P} \\ \bar{Q} - \tilde{Q} \end{bmatrix} \right| = O(h^3), \qquad \left(E\begin{bmatrix} \bar{P} - \tilde{P} \\ \bar{Q} - \tilde{Q} \end{bmatrix}^2 \right)^{1/2} = O(h^2). \tag{4.18}$$

Expanding the right-hand sides of the approximation $\bar{P}$, $\bar{Q}$ about (t,p,q), we obtain

$$\bar{P} = p + hf + \frac{h^2}{2}\frac{\partial f}{\partial t} + \frac{3}{4}h\sum_{i=1}^{n}\Delta\mathcal{Q}_2^i\frac{\partial f}{\partial q^i} + \sum_{r=1}^{m}\sigma_r\Delta w_r + \frac{3}{2}\sum_{r=1}^{m}\sum_{i=1}^{n}\Delta\mathcal{Q}_2^i\frac{\partial\sigma_r}{\partial q^i}\left(\Delta w_r - J_{r0}\right) + h\sum_{r=1}^{m}\frac{\partial\sigma_r}{\partial t}\left(\Delta w_r - J_{r0}\right) + \rho_1, \tag{4.19}$$

$$\bar{Q} = q + hg + \frac{h}{3}\sum_{i=1}^{n}\left(2\Delta\mathcal{P}_1^i + \Delta\mathcal{P}_2^i\right)\frac{\partial g}{\partial p^i} + \frac{h}{6}\sum_{i,j=1}^{n}(2\Delta\mathcal{P}_1^i\Delta\mathcal{P}_1^j + \Delta\mathcal{P}_2^i\Delta\mathcal{P}_2^j)\frac{\partial^2 g}{\partial p^i\partial p^j} + \rho_2,$$

$$\Delta\mathcal{P}_1 := \mathcal{P}_1 - p = \frac{h}{4}f + \frac{1}{2}\sum_{r=1}^{m}\sigma_r\left(3J_{r0} - \Delta w_r\right),$$

$$\Delta\mathcal{Q}_2 := \mathcal{Q}_2 - q = \frac{2}{3}hg + \frac{2}{3}h\sum_{i=1}^{n}\Delta\mathcal{P}_1^i\frac{\partial g}{\partial p^i} + \frac{h}{3}\sum_{i,j=1}^{n}\Delta\mathcal{P}_1^i\Delta\mathcal{P}_1^j\frac{\partial^2 g}{\partial p^i\partial p^j} + r_1,$$

$$\Delta\mathcal{P}_2 := \mathcal{P}_2 - p = hf + \sum_{r=1}^{m}\sigma_r\Delta w_r + r_2,$$

where all the coefficients are calculated at (t,p,q).

Due to properties of the Wiener process and Ito integrals, we have

$$E\Delta w_i = 0, \quad E\Delta w_i\Delta w_j = \delta_{ij}h, \quad E\Delta w_i\Delta w_j\Delta w_k = 0, \quad E\left(\Delta w_i\right)^4 = 3h^2,$$
$$EJ_{i0} = 0, \quad EJ_{i0}J_{j0} = \delta_{ij}\frac{h}{3}, \quad EJ_{i0}J_{j0}J_{k0} = 0, \quad E\left(J_{i0}\right)^4 = \frac{h^2}{3},$$
$$E\Delta w_iJ_{j0} = \delta_{ij}\frac{h}{2}, \quad E\Delta w_i\Delta w_jJ_{k0} = 0, \quad E\Delta w_iJ_{j0}J_{k0} = 0. \tag{4.20}$$

Then, under appropriate conditions on smoothness and boundedness of the coefficients of (4.2), we get

$$|E\rho_i| = O(h^3), \quad E\left(\rho_i\right)^2 = O(h^5), \quad i = 1,2,$$
$$|Er_1| = O(h^3), \quad E\left(r_1\right)^2 = O(h^5), \quad |Er_2| = O(h^2), \quad E\left(r_2\right)^2 = O(h^3). \tag{4.21}$$

In addition to (4.20) we note that

$$E\left(\Delta w_r - J_{r0}\right)\left(3J_{r0} - \Delta w_r\right) = 0, \quad E\left(3J_{r0} - \Delta w_r\right)^2 = h. \tag{4.22}$$

Using (4.20)–(4.22), we obtain from (4.19)

$$\bar{P} = p + \sum_{r=1}^{m}\sigma_r\Delta w_r + hf + \sum_{r=1}^{m}\left[\frac{\partial\sigma_r}{\partial t} + \sum_{i=1}^{n}g^i\frac{\partial\sigma_r}{\partial q^i}\right]I_{0r} + \frac{h^2}{2}\left[\frac{\partial f}{\partial t} + \sum_{i=1}^{n}g^i\frac{\partial f}{\partial q^i}\right] + R_1,$$

$$\bar{Q} = q + hg + \sum_{r=1}^{m}\sum_{i=1}^{n}\sigma_r^i\frac{\partial g}{\partial p^i}I_{r0} + \frac{h^2}{2}\left[\sum_{i=1}^{n}f^i\frac{\partial g}{\partial p^i} + \frac{1}{2}\sum_{r=1}^{m}\sum_{i,j=1}^{n}\sigma_r^i\sigma_r^j\frac{\partial^2 g}{\partial p^i\partial p^j}\right] + R_2$$

with R_i, $i = 1, 2$, such that

$$|ER_i| = O(h^3), \quad E\left(R_i\right)^2 = O(h^4), \quad i = 1, 2.$$

This implies (4.18). It follows from (4.18) that the method (4.13)–(4.14) is of the mean-square order 3/2. □

REMARK 4.2. *The random variables* $\Delta_k w_r(h)$, $(J_{r0})_k$ *have a Gaussian joint distribution, and they can be simulated at each step by* $2m$ *independent* $\mathcal{N}(0,1)$*-distributed random variables* ξ_{rk} *and* η_{rk}, $r = 0, \ldots, m$:

$$\Delta_k w_r(h) = \xi_{rk}\sqrt{h}, \qquad (J_{r0})_k = \left(\xi_{rk}/2 + \eta_{rk}/\sqrt{12}\right)\sqrt{h}\,.$$

As a result, the method (4.13)–(4.14) *takes the constructive form.*

REMARK 4.3. *It is very unusual that the direct expansion of* (4.13)–(4.14) *does not contain the habitual term* $\frac{h^2}{4}\sum_{r=1}^{m}\sum_{i,j=1}^{n}\frac{\partial^2 g}{\partial p^i \partial p^j}\sigma_r^i\sigma_r^j$. *The similar term in the expansion contains some combinations of* Δw_r *and* J_{r0} *instead of* h. (*See a similar remark in* [3].)

REMARK 4.4. *In the case of* $\sigma_r = 0$, $r = 1, \ldots, m$, *the method* (4.13)–(4.14) *coincides with the well-known deterministic symplectic PRK method of the second order. Adapting other explicit deterministic second-order PRK methods from* [5, 9], *it is possible to construct other explicit symplectic methods of the order* 3/2 *for the system* (4.2), (4.3).

REMARK 4.5. *In the case of a more general system than* (4.2) *methods of the order* 3/2 *require simulation of repeated Ito integrals which is a laborious problem from the computational point of view. We do not consider such methods in the paper.* (*See also the introduction.*)

Lemma 4.2 *can be generalized for the general separable case, i.e., for the system* (1.1), (1.3) *with* $H_r = V_r(p) + U_r(t,q)$, $r = 0, 1, \ldots, m$, *and it can also be generalized for the general stochastic Hamiltonian system* (1.1), (1.3). *In the case of systems with one noise repeated Ito integrals can effectively be simulated, and generalizations of Lemma* 4.2 *can be used for constructing high-order symplectic methods for Hamiltonian systems with one noise* (*i.e., when* $m = 1$).

5. Numerical tests.

5.1. Kubo oscillator. The system of SDEs in the sense of Stratonovich (Kubo oscillator)

$$\begin{aligned} dX^1 &= -aX^2 dt - \sigma X^2 \circ dw(t), \qquad X^1(0) = x^1, \\ dX^2 &= aX^1 dt + \sigma X^1 \circ dw(t), \qquad X^2(0) = x^2, \end{aligned} \tag{5.1}$$

is often used for testing numerical methods. (See, e.g., [15], where some nonsymplectic stochastic methods based on deterministic symplectic methods are used.) Here a and σ are constants, and $w(t)$ is a one-dimensional standard Wiener process.

The phase flow of this system preserves symplectic structure. Moreover, the quantity $\mathcal{H}(x^1, x^2) = \left(x^1\right)^2 + \left(x^2\right)^2$ is conservative for this system; i.e.,

$$\mathcal{H}(X^1(t), X^2(t)) = \mathcal{H}(x^1, x^2) \text{ for } t \geq 0.$$

This means that a phase trajectory of (5.1) belongs to the circle with the center at the origin and with the radius $\sqrt{\mathcal{H}(x^1, x^2)}$.

We test here three methods. In application to (5.1) the symplectic PRK method (3.8) takes the following form:

$$X^1_{k+1} = X^1_k - aX^2_k h - \frac{\sigma^2}{2} X^1_{k+1} h - \sigma X^2_k \Delta_k w, \tag{5.2}$$
$$X^2_{k+1} = X^2_k + aX^1_{k+1} h + \frac{\sigma^2}{2} X^2_k h + \sigma X^1_{k+1} \Delta_k w.$$

This method is implicit in the deterministic part only.

The midpoint method (3.7) applied to the system with one noise (5.1) reads

$$X^1_{k+1} = X^1_k - a\frac{X^2_k + X^2_{k+1}}{2} h - \sigma \frac{X^2_k + X^2_{k+1}}{2} (\zeta_h)_k \sqrt{h}, \tag{5.3}$$
$$X^2_{k+1} = X^2_k + a\frac{X^1_k + X^1_{k+1}}{2} h + \sigma \frac{X^1_k + X^1_{k+1}}{2} (\zeta_h)_k \sqrt{h}.$$

This is a fully implicit method. Note that due to specific features of the system (5.1), the formula (5.3) is valid (solvable) not only in the case of the truncated random variable ζ_h but also if we put $\Delta_k w$ instead of $(\zeta_h)_k \sqrt{h}$.

The method (5.3) is of the first mean-square order. The method (5.2) is of the mean-square order $1/2$ as well as the Euler method:

$$X^1_{k+1} = X^1_k - aX^2_k h - \frac{\sigma^2}{2} X^1_k h - \sigma X^2_k \Delta_k w, \tag{5.4}$$
$$X^2_{k+1} = X^2_k + aX^1_k h - \frac{\sigma^2}{2} X^2_k h + \sigma X^1_k \Delta_k w,$$

which, of course, is not symplectic.

Figure 1 gives approximations of a sample phase trajectory of (5.1) simulated by the symplectic methods (5.2) and (5.3) and by the Euler method (5.4). The initial condition is $x^1 = 1$, $x^2 = 0$. The corresponding exact phase trajectory belongs to the circle with the center at the origin and with the unit radius.

We see that the Euler method is not appropriate for simulation of the oscillator (5.1) on long time intervals, while the symplectic methods preserve conservative properties of the Kubo oscillator.

These experiments also demonstrate that the midpoint method is much more accurate than the other methods applied. It is not difficult to check that $\mathcal{H}(x^1, x^2)$ is conserved by the midpoint method (5.3), but it is not conserved by the symplectic PRK method (5.2). This is similar to the deterministic case. Indeed, it is known [8, 5] that symplectic deterministic RK methods (e.g., the midpoint scheme) conserve all quadratic functions that are conserved by the Hamiltonian system being integrated, while deterministic PRK methods do not possess this property.

5.2. A model for synchrotron oscillations of particles in storage rings. In [14] a model describing synchrotron oscillations of particles in storage rings under the influence of external fluctuating electromagnetic fields was considered. This model can be written in the following form:

$$dP = -\omega^2 \sin(Q) dt - \sigma_1 \cos(Q) dw_1 - \sigma_2 \sin(Q) dw_2, \tag{5.5}$$
$$dQ = P dt.$$

P and Q are scalars here. The system (5.5) is of the form (4.2), and therefore its phase flow preserves symplectic structure.

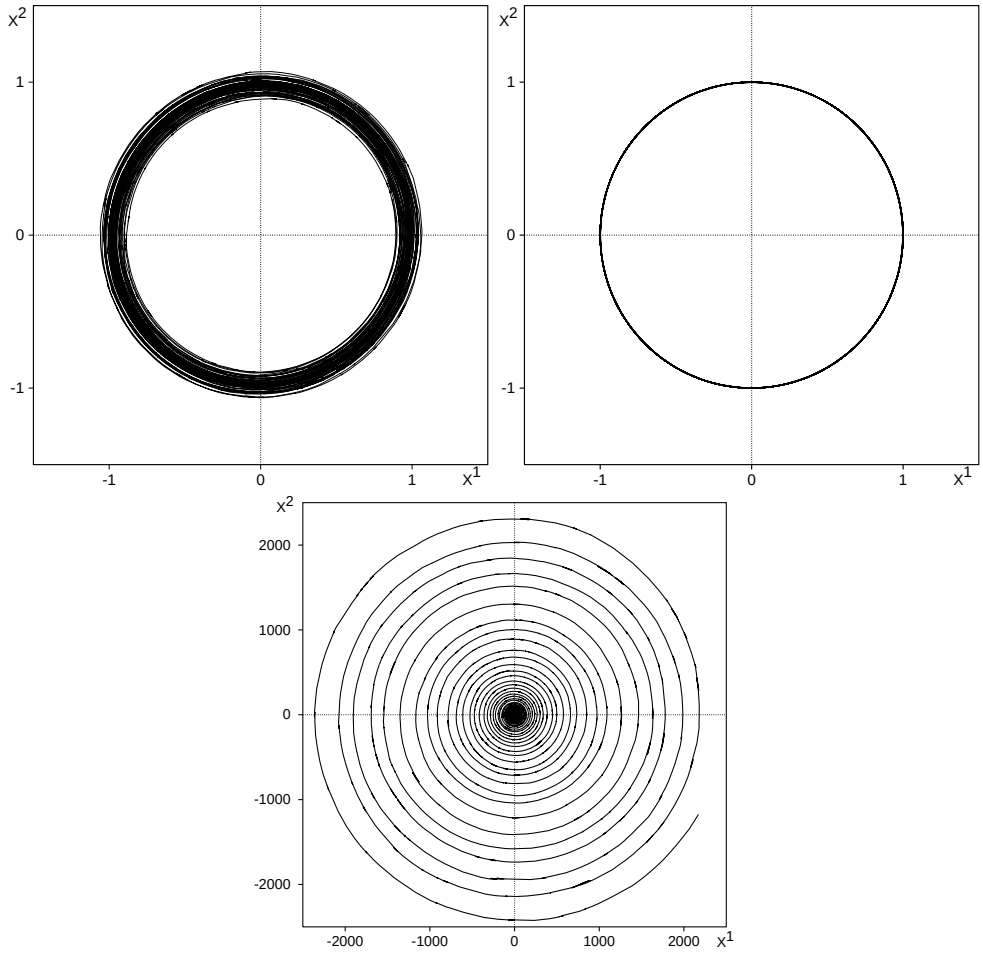


FIG. 1. *A sample phase trajectory of* (5.1) *with* $X^1(0) = 1$, $X^2(0) = 0$ *obtained by the symplectic method* (5.2) *(top left), the midpoint method* (5.3) *(top right), and by the Euler method* (5.4) *(bottom) for* $a = 2$, $\sigma = 0.3$, $h = 0.02$ *on the time interval* $t \le 200$.

The Euler method for (5.5) takes the form

$$P_{k+1} = P_k - h\omega^2 \sin(Q_k) - h^{1/2}(\sigma_1 \cos(Q_k)\Delta_k w_1 + \sigma_2 \sin(Q_k)\Delta_k w_2), \tag{5.6}$$
$$Q_{k+1} = Q_k + hP_k.$$

In application to (5.5) the explicit symplectic method (4.5)–(4.6) with $\alpha = 1$ is written as

$$\mathcal{Q} = Q_k + hP_k, \tag{5.7}$$
$$P_{k+1} = P_k - h\omega^2 \sin(\mathcal{Q}) - h^{1/2}(\sigma_1 \cos(\mathcal{Q})\Delta_k w_1 + \sigma_2 \sin(\mathcal{Q})\Delta_k w_2), \qquad Q_{k+1} = \mathcal{Q}.$$

Both methods are of the first mean-square order.

Approximations of a sample trajectory of (5.5) simulated by the symplectic method (5.7) and the Euler method (5.6) are plotted on Figure 2. The trajectory obtained by the symplectic method with $h = 0.02$ (solid line) visually coincides with

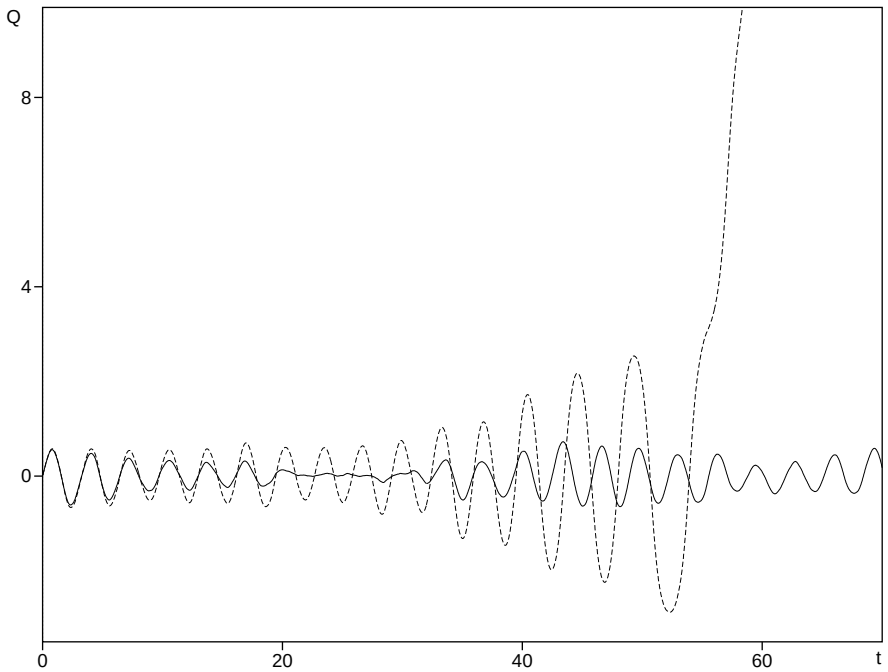


FIG. 2. *A sample trajectory of* (5.5) *for* $\omega = 2$, $\sigma_1 = 0.2$, $\sigma_2 = 0.1$, $h = 0.02$. *Solid line—the symplectic method* (5.7), *dashed line—the Euler method* (5.6).

the one obtained for a smaller step, e.g., for $h = 0.002$, using the same sample paths for the Wiener processes; i.e., this trajectory visually coincides with the exact solution of (5.5). This figure clearly demonstrates that the Euler method (dashed line) is unacceptable for simulation of the solution to (5.5) on a long time interval, while the symplectic method (5.7) produces quite accurate results despite both methods having the same mean-square order of accuracy.

REFERENCES

[1] J.-M. BISMUT, *Mécanique Aléatoire,* Lecture Notes in Math. 866, Springer, Berlin, 1981.
[2] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
[3] G.N. MILSTEIN, YU. M. REPIN, AND M.V. TRETYAKOV, *Symplectic integration of Hamiltonian systems with additive noise*, SIAM J. Numer. Anal., 39 (2002), pp. 2066–2088.
[4] V.I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer, Berlin, 1989.
[5] J.M. SANZ-SERNA AND M.P. CALVO, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.
[6] E. HAIRER, S.P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations.* I. *Non-stiff Problems*, Springer, Berlin, 1993.
[7] P.J. CHANNEL AND C. SCOVEL, *Symplectic integration of Hamiltonian systems*, Nonlinearity, 3 (1990), pp. 231–259.
[8] J.M. SANZ-SERNA, *Symplectic integrators for Hamiltonian problems: An overview*, Acta Numer. 1, Cambridge University Press, Cambridge, UK, 1992, pp. 243–286.
[9] YU. B. SURIS, *Hamiltonian methods of Runge-Kutta type and their variational interpretation*, Mat. Model., 2 (1990), pp. 78–87.
[10] G.N. MILSTEIN, *Numerical Integration of Stochastic Differential Equations*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.
[11] P.E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*,

Springer, Berlin, 1992.

[12] G.N. Milstein, E. Platen, and H. Schurz, *Balanced implicit methods for stiff stochastic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1010–1019.

[13] G.N. Milstein, Yu. M. Repin, and M.V. Tretyakov, *Mean-Square Symplectic Methods for Hamiltonian Systems with Multiplicative Noise*, Preprint 670, Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin, Germany, 2001.

[14] M. Seeßelberg, H.P. Breuer, H. Mais, F. Petruccione, and J. Honerkamp, *Simulation of one-dimensional noisy Hamiltonian systems and their application to particle storage rings*, Z. Phys. C, 62 (1994), pp. 62–73.

[15] M.V. Tretyakov and S.V. Tret'jakov, *Numerical integration of Hamiltonian systems with external noise*, Phys. Lett. A, 194 (1994), pp. 371–374.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1605–1631

A ROBUST FINITE ELEMENT METHOD FOR DARCY–STOKES FLOW*

KENT ANDRE MARDAL†, XUE-CHENG TAI‡, AND RAGNAR WINTHER§

Abstract. Finite element methods for a family of systems of singular perturbation problems of a saddle point structure are discussed. The system is approximately a linear Stokes problem when the perturbation parameter is large, while it degenerates to a mixed formulation of Poisson's equation as the perturbation parameter tends to zero. It is established, basically by numerical experiments, that most of the proposed finite element methods for Stokes problem or the mixed Poisson's system are not well behaved uniformly in the perturbation parameter. This is used as the motivation for introducing a new "robust" finite element which exhibits this property.

Key words. singular perturbation problems, Darcy–Stokes flow, nonconforming finite elements, uniform error estimates

AMS subject classifications. 65N12, 65N15, 65N30

PII. S0036142901383910

1. Introduction.

Let $\Omega \subset \mathbb{R}^2$ be a bounded and connected polygonal domain with boundary $\partial\Omega$. In this paper we shall consider finite element methods for the following singular perturbation problem:

$$\begin{aligned} (\boldsymbol{I} - \varepsilon^2 \boldsymbol{\Delta})\boldsymbol{u} - \mathbf{grad}\, p &= \boldsymbol{f} \quad \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u} &= g \quad \text{in } \Omega, \\ \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Here $\varepsilon \in (0, 1]$ is a parameter, while $\boldsymbol{\Delta} = \mathbf{diag}(\Delta, \Delta)$ is the Laplace operator on vector fields. The vector field $\boldsymbol{f}$ and scalar field g represent the data. The problem (1.1) admits only a solution if the function g has mean value zero on Ω and "the pressure" p is determined only up to addition of a constant.

We note that when ε is not too small, and $g = 0$, this problem is simply a standard Stokes problem but with an additional nonharmful lower order term. However, if $\boldsymbol{f} = 0$ and ε approaches zero, then the model problem formally tends to a mixed formulation of the Poisson equation with homogeneous Neumann boundary conditions.

When $\varepsilon = 0$ the first equation in (1.1) has the form of Darcy's law for flow in a homogeneous porous medium, where $\boldsymbol{u}$ is a volume averaged velocity. In fact, the system (1.1) can be regarded as a macroscopic model for flow in an "almost porous media," where $\boldsymbol{u}$ and p represent volume averaged velocity and pressure, respectively. The zero order velocity term in the first equation of (1.1) then typically represents a *Stokes drag*. An attempt to derive Darcy's law from volume averaged Stokes flow is, for example, discussed in [16]. Generalizations of the system (1.1) have also been proposed in the modeling of macrosegregation formation in binary alloy solidification;

*Received by the editors January 22, 2001; accepted for publication (in revised form) April 23, 2002; published electronically October 31, 2002. This research was supported in part by the Research Council of Norway under grants 128224/431, 133755/441, and 135420/431.
http://www.siam.org/journals/sinum/40-5/38391.html
†Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway (kent-and@ifi.uio.no).
‡Department of Mathematics, University of Bergen, Johannes Brunsgt. 12, 5007 Bergen, Norway (Xue-Cheng.Tai@mi.uib.no).
§Department of Informatics and Department of Mathematics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway (rwinther@ifi.uio.no).

cf. [13]. Systems of the form (1.1) may also arise from time discretizations of the Navier–Stokes equation, where the parameter ε corresponds to the square root of the time step; cf. [4]. However, the study of such time discretizations is not the motivation for the present paper.

The purpose of the present paper is to discuss a finite element method for the model problem (1.1) with convergence properties that are uniform with respect to the perturbation parameter ε. In section 2 we will introduce some notations and discuss various properties of the model (1.1). Discretizations of the model problem by the finite element method is described in section 3. In particular, we will state stability conditions which are uniform with respect to the parameter ε and show, by numerical experiments, that the standard discretizations, proposed either for $\varepsilon = 1$ or $\varepsilon = 0$, do not satisfy these stability conditions. A new nonconforming finite element discretization is then proposed in section 4. We show that this new discretization is uniformly stable, and, as a consequence, we establish in section 5 error estimates which are uniform in ε under the assumption that proper regularity estimates hold for the solution. In section 6 we then study the asymptotic smoothness of the solution of (1.1) as ε tends to zero. Based on these regularity results we show that, for fixed data $\boldsymbol{f}$ and g, a uniform $O(h^{1/2})$ error estimate in a suitable energy norm can be derived.

In the final section of this paper we study an elliptic system which formally is a generalization of (1.1). This system is given by

$$\begin{aligned} (\boldsymbol{I} - \varepsilon^2 \boldsymbol{\Delta})\boldsymbol{u} - \delta^{-2}\,\mathbf{grad}(\operatorname{div} \boldsymbol{u} - g) &= \boldsymbol{f} \quad && \text{in } \Omega, \\ \boldsymbol{u} &= 0 \quad && \text{on } \partial\Omega, \end{aligned} \tag{1.2}$$

where $\varepsilon, \delta \in (0, 1]$. By introducing $p = \delta^{-2}(\operatorname{div} \boldsymbol{u} - g)$ this system can be alternatively written on the mixed form

$$\begin{aligned} (\boldsymbol{I} - \varepsilon^2 \boldsymbol{\Delta})\boldsymbol{u} - \mathbf{grad}\, p &= \boldsymbol{f} \quad && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u} - \delta^2 p &= g \quad && \text{in } \Omega, \\ \boldsymbol{u} &= 0 \quad && \text{on } \partial\Omega. \end{aligned} \tag{1.3}$$

Note that this system also has meaning when $\delta = 0$, and in this case the system reduces to (1.1).

The symmetric and positive definite system (1.2) is discretized by a straightforward finite element approach, utilizing the new nonconforming velocity space constructed earlier in this paper; i.e., the mixed system (1.3) is not introduced in the discretization. We show, by numerical experiments and theory, that under the assumption of sufficiently regular solutions we obtain error estimates which are uniform both in ε and δ.

2. Preliminaries. We will use $H^m = H^m(\Omega)$ to denote the Sobolev space of scalar functions on Ω with m derivatives in $L^2 = L^2(\Omega)$, with norm $\|\cdot\|_m$. Furthermore, the notation $\|\cdot\|_{m,K}$ is used to indicate that the norm is defined with respect to a domain K different from Ω. The seminorm derived from the partial derivatives of order equal m is denoted $|\cdot|_m$, i.e., $|\cdot|_m^2 = \|\cdot\|_m^2 - \|\cdot\|_{m-1}^2$. The space $H_0^m = H_0^m(\Omega)$ will denote the closure in H^m of $C_0^\infty(\Omega)$. The dual space of H_0^m with respect to the L^2 inner product will be denoted by H^{-m}. Furthermore, L_0^2 will denote the space of L^2 functions with mean value zero. A space written in boldface denotes a 2-vector valued analogue of the corresponding scalar space. The notation $(\cdot, \cdot)$ is used to denote the L^2 inner product on scalar, vector, and matrix valued functions.

Below we shall encounter the intersection and sum of Hilbert spaces. We therefore recall the basic definitions of these concepts. If X and Y are Hilbert spaces, both continuously contained in some larger Hilbert spaces, then the intersection $X \cap Y$ and the sum $X + Y$ are themselves Hilbert spaces with the norms

$$\|z\|_{X\cap Y} = (\|z\|_X^2 + \|z\|_Y^2)^{1/2}$$

and

$$\|z\|_{X+Y} = \inf_{\substack{z=x+y \\ x\in X,\, y\in Y}} (\|x\|_X^2 + \|y\|_Y^2)^{1/2}.$$

Furthermore, if $X \cap Y$ is dense in both X and Y, then $(X \cap Y)^* = X^* + Y^*$. We refer the reader to [3, Chapter 2] for these results.

If q is a scalar field, then $\mathbf{grad}\, q$ will denote the gradient of q, while $\operatorname{div} \boldsymbol{v}$ denotes the divergence of a vector field $\boldsymbol{v}$. We shall also use the differential operators

$$\mathbf{curl}\, q = \begin{pmatrix} -\partial q/\partial x_2 \\ \partial q/\partial x_1 \end{pmatrix} \quad \text{and} \quad \operatorname{rot} \boldsymbol{v} = \partial \boldsymbol{v}_1/\partial x_2 - \partial \boldsymbol{v}_2/\partial x_1.$$

Note that, due to Green's theorem, these definitions lead to the following "integration by parts formula":

$$\int_\Omega \mathbf{curl}\, q \cdot \boldsymbol{v}\, dx = \int_\Omega q \operatorname{rot} \boldsymbol{v}\, dx + \int_{\partial\Omega} q(\boldsymbol{v} \cdot \boldsymbol{t})\, d\tau, \tag{2.1}$$

where $\boldsymbol{t}$ is the unit tangent vector in the counterclockwise direction on $\partial\Omega$, and τ is the arclength.

The gradient of a vector field $\boldsymbol{v}$ is denoted $\boldsymbol{D}\boldsymbol{v}$; i.e., $\boldsymbol{D}\boldsymbol{v}$ is the 2×2 matrix with elements

$$(\boldsymbol{D}\boldsymbol{v})_{i,j} = \partial \boldsymbol{v}_i/\partial x_j, \qquad 1 \le i, j \le 2.$$

Hence, for any $\boldsymbol{u} \in \boldsymbol{H}^2$ and $\boldsymbol{v} \in \boldsymbol{H}_0^1$ we have

$$-(\boldsymbol{\Delta}\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{D}\boldsymbol{u}, \boldsymbol{D}\boldsymbol{v}) \equiv \int_\Omega \boldsymbol{D}\boldsymbol{u} : \boldsymbol{D}\boldsymbol{v}\, dx,$$

where the colon denotes the scalar product of matrix fields. Recall also the identity

$$\boldsymbol{\Delta} = \mathbf{grad} \operatorname{div} - \mathbf{curl} \operatorname{rot}, \tag{2.2}$$

which can be verified by a direct computation. As a consequence, we obtain the identity

$$(\boldsymbol{D}\boldsymbol{u}, \boldsymbol{D}\boldsymbol{v}) = (\operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{v}) + (\operatorname{rot} \boldsymbol{u}, \operatorname{rot} \boldsymbol{v}) \quad \forall \boldsymbol{u} \in \boldsymbol{H}^1,\ \boldsymbol{v} \in \boldsymbol{H}_0^1. \tag{2.3}$$

In addition to the function spaces introduced above we will also use the space $\boldsymbol{H}(\operatorname{div}) = \boldsymbol{H}(\operatorname{div}; \Omega)$ consisting of all vector fields in $\boldsymbol{L}^2$ with divergence in L^2, i.e.,

$$\boldsymbol{H}(\operatorname{div}) = \{\boldsymbol{v} \in \boldsymbol{L}^2 : \operatorname{div} \boldsymbol{v} \in L^2\}.$$

Similarly,

$$\boldsymbol{H}(\operatorname{rot}) = \{\boldsymbol{v} \in \boldsymbol{L}^2 : \operatorname{rot} \boldsymbol{v} \in L^2\},$$

and the norms of these spaces are denoted by $\|\cdot\|_{\mathrm{div}}$ and $\|\cdot\|_{\mathrm{rot}}$, respectively. Furthermore, $\boldsymbol{H}_0(\mathrm{div})$ is the closed subspace of $\boldsymbol{H}(\mathrm{div})$ consisting of functions with vanishing normal components on the boundary; i.e.,

$$\boldsymbol{H}_0(\mathrm{div}) = \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}) : \boldsymbol{v} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega\},$$

where $\boldsymbol{n}$ is the unit outward normal vector.

Throughout this paper $a_\varepsilon(\cdot,\cdot) : \boldsymbol{H}^1 \times \boldsymbol{H}^1 \mapsto \mathbb{R}$ will denote the bilinear form

$$a_\varepsilon(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v}) + \varepsilon^2(\boldsymbol{D}\boldsymbol{u},\boldsymbol{D}\boldsymbol{v}).$$

A weak formulation of problem (1.1) is given by the following:

Find $(\boldsymbol{u},p) \in \boldsymbol{H}_0^1 \times L_0^2$ such that

$$\begin{aligned} a_\varepsilon(\boldsymbol{u},\boldsymbol{v}) + (p, \operatorname{div}\boldsymbol{v}) &= (\boldsymbol{f},\boldsymbol{v}) \quad && \forall \boldsymbol{v} \in \boldsymbol{H}_0^1, \\ (\operatorname{div}\boldsymbol{u}, q) &= (g,q) && \forall q \in L_0^2. \end{aligned} \tag{2.4}$$

Here we assume that data $(\boldsymbol{f},g)$ is given in $\boldsymbol{H}^{-1} \times L_0^2$.

The problem (2.4) has a unique solution $(\boldsymbol{u},p) \in \boldsymbol{H}_0^1 \times L_0^2$. This follows from standard results for Stokes problem; cf., for example, [11]. However, the bound on $(\boldsymbol{u},p) \in \boldsymbol{H}_0^1 \times L_0^2$ will degenerate as ε tends to zero. In fact, for the reduced problem (2.4) with $\varepsilon = 0$ the space $\boldsymbol{H}_0^1 \times L_0^2$ is not a proper function space for the solution. However, the theory developed in [6] can be applied in this case if we seek $(\boldsymbol{u},p)$ either in $\boldsymbol{H}_0(\mathrm{div}) \times L_0^2$ or in $\boldsymbol{L}^2 \times (H^1 \cap L_0^2)$, and with data $(\boldsymbol{f},g)$ in the proper dual spaces. These results are, in fact, consequences of standard results for the Poisson equation.

The fact that the regularity of the solution is changed when ε becomes zero strongly suggests that ε-dependent norms and function spaces are required in order to obtain stability estimates independent of ε. Furthermore, since the reduced problem is well posed for two completely different choices of function spaces, this indicates that there are at least two different choices of ε-dependent norms. In the present paper we will study the problem (1.1) with respect to an ε-dependent norm which reduces to the norm in $\boldsymbol{H}_0(\mathrm{div}) \times L_0^2$ when $\varepsilon = 0$. Our goal is to derive discretizations which are uniformly stable with respect to ε in this norm. This appears to be the proper choice if we want to study discretizations which also can be generalized to nonmixed approximations of elliptic problems of the form (1.2).

Remark. When we refer to the *reduced system* corresponding to (1.1) we refer to the system (1.1) with $\varepsilon = 0$ and the boundary condition $\boldsymbol{u} = 0$ replaced by $\boldsymbol{u} \cdot \boldsymbol{n} = 0$. This system has a weak formulation given by (2.4) but with the solution space $\boldsymbol{H}_0^1$ replaced by $\boldsymbol{H}_0(\mathrm{div})$. ∎

The space $\boldsymbol{H}_0(\mathrm{div}) \cap \varepsilon \cdot \boldsymbol{H}_0^1$, with norm $|\!|\!| \cdot |\!|\!|_\varepsilon$ given by

$$|\!|\!|\boldsymbol{v}|\!|\!|_\varepsilon^2 = \|\boldsymbol{v}\|_0^2 + \|\operatorname{div}\boldsymbol{v}\|_0^2 + \varepsilon^2\|\boldsymbol{D}\boldsymbol{v}\|_0^2,$$

is equal to $\boldsymbol{H}_0^1$ as a set for $\varepsilon > 0$ but equal to $\boldsymbol{H}_0(\mathrm{div})$ for $\varepsilon = 0$. The system (2.4) can alternatively be written as the system

$$\mathcal{A}_\varepsilon \begin{pmatrix} \boldsymbol{u} \\ p \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ g \end{pmatrix},$$

where the coefficient operator $\mathcal{A}_\varepsilon$ is given by

$$\mathcal{A}_\varepsilon = \begin{pmatrix} \boldsymbol{I} - \varepsilon^2\boldsymbol{\Delta} & -\mathbf{grad} \\ \operatorname{div} & 0 \end{pmatrix}. \tag{2.5}$$

Let $\boldsymbol{X}_\varepsilon$ be the product space $(\boldsymbol{H}_0(\mathrm{div}) \cap \varepsilon \cdot \boldsymbol{H}_0^1) \times L_0^2$ and $\boldsymbol{X}_\varepsilon^*$ the corresponding dual space with respect to the L^2 inner product. This space can also be expressed as

$$\boldsymbol{X}_\varepsilon^* = (\boldsymbol{H}^{-1}(\mathrm{rot}) + \varepsilon^{-1}\boldsymbol{H}^{-1}) \times L_0^2.$$

Here the + sign has the interpretation as the sum of Hilbert spaces, and the space $\boldsymbol{H}^{-1}(\mathrm{rot})$ is given by

$$\boldsymbol{H}^{-1}(\mathrm{rot}) = \{\boldsymbol{v} \in \boldsymbol{H}^{-1} : \mathrm{rot}\, \boldsymbol{v} \in H^{-1}\}.$$

The operator $\mathcal{A}_\varepsilon$ can be seen to be an isomorphism mapping $\boldsymbol{X}_\varepsilon$ into $\boldsymbol{X}_\varepsilon^*$. Furthermore, the corresponding operator norms

$$||\mathcal{A}_\varepsilon||_{\mathcal{L}(X_\varepsilon, X_\varepsilon^*)} \quad \text{and} \quad ||\mathcal{A}_\varepsilon^{-1}||_{\mathcal{L}(X_\varepsilon^*, X_\varepsilon)}$$

are independent of ε. In fact, with the definitions above, this is also true for $\varepsilon \in [0, 1]$; i.e., the endpoint $\varepsilon = 0$ can be included.

The uniform boundedness of $\mathcal{A}_\varepsilon$ is straightforward to check from the definitions above, while the uniform boundedness of the inverse can be verified from the two Brezzi conditions; cf. [6]. For the present problem these conditions read as follows:

There are constants $\alpha_0, \beta_0 > 0$, independent of ε, such that

$$\sup_{\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div}) \cap \varepsilon \cdot \boldsymbol{H}_0^1} \frac{(q, \mathrm{div}\, \boldsymbol{v})}{|\!|\!| \boldsymbol{v} |\!|\!|_\varepsilon} \geq \alpha_0 \|q\|_0 \quad \forall q \in L_0^2 \tag{2.6}$$

and

$$a_\varepsilon(\boldsymbol{v}, \boldsymbol{v}) \geq \beta_0 |\!|\!| \boldsymbol{v} |\!|\!|_\varepsilon^2 \quad \forall \boldsymbol{v} \in \boldsymbol{Z}, \tag{2.7}$$

where $\boldsymbol{Z} = \{\boldsymbol{v} \in \boldsymbol{H}_0^1 : \mathrm{div}\, \boldsymbol{v} = 0\}$.

Since it is well known (cf., for example, [11, Chapter 1, Corollary 2.4]) that condition (2.6) holds for $\varepsilon = 1$, it also holds for all $\varepsilon \in [0, 1]$ with the same constant α_0. Furthermore, condition (2.7) holds trivially with $\beta_0 = 1$ for $\varepsilon \in [0, 1]$.

3. Uniformly stable discretizations. The purpose of this section is to discuss finite element discretizations of the system (1.1). In particular, we shall be interested in discretizations which are stable uniformly in the parameter $\varepsilon \in (0, 1]$.

Let $\boldsymbol{V}_h \subset \boldsymbol{H}_0^1$ and $Q_h \subset L_0^2$ be finite element spaces, where $h \in (0, 1]$ is a discretization parameter. The weak formulation (2.4) leads to the following corresponding finite element discretization:

Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned} a_\varepsilon(\boldsymbol{u}_h, \boldsymbol{v}) + (p_h, \mathrm{div}\, \boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}) \quad && \forall \boldsymbol{v} \in \boldsymbol{V}_h, \\ (\mathrm{div}\, \boldsymbol{u}_h, q) &= (g, q) \quad && \forall q \in Q_h. \end{aligned} \tag{3.1}$$

Remark. Below we shall also encounter several examples of nonconforming approximations of (2.4), i.e., the space $\boldsymbol{V}_h \not\subseteq \boldsymbol{H}_0^1$. In all these examples the bilinear form $a_\varepsilon(\cdot, \cdot)$ is understood to be the sum of the corresponding integrals over each element. No extra jump terms are added. The same remark applies to the energy norm, $|\!|\!| \cdot |\!|\!|_\varepsilon$. □

The discretization (3.1) is stable in the sense of [6] if proper discrete analogues of the conditions (2.6) and (2.7) hold. These conditions are the following.

Stability conditions. The discretization (3.1) is said to be uniformly stable if there exist constants $\alpha, \beta > 0$, independent of ε and h, such that

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{(q, \operatorname{div} \boldsymbol{v})}{|\!|\!|\boldsymbol{v}|\!|\!|_\varepsilon} \geq \alpha \|q\|_0 \quad \forall q \in Q_h \tag{3.2}$$

and

$$a_\varepsilon(\boldsymbol{v}, \boldsymbol{v}) \geq \beta |\!|\!|\boldsymbol{v}|\!|\!|_\varepsilon^2 \quad \forall \boldsymbol{v} \in \boldsymbol{Z}_h, \tag{3.3}$$

where $\boldsymbol{Z}_h = \{\boldsymbol{v} \in \boldsymbol{V}_h : (\operatorname{div} \boldsymbol{v}, q) = 0 \quad \forall q \in Q_h\}$.

For the case $\varepsilon = 1$, or more precisely for ε bounded away from zero, the second condition is obvious. In this case there are several choices of pairs of finite element spaces which satisfy (3.2) with α independent of h. We mention, for example, the Mini element proposed in [1] or the $P_2 - P_0$ element; i.e., we choose continuous quadratic velocities for $\boldsymbol{V}_h$ and the corresponding space of piecewise constants for Q_h; cf. [10]. For a general review of stable Stokes elements we refer the reader to [8].

However, most of these spaces do *not* lead to discretizations which are stable uniformly in ε. The main reason for this is that when ε approaches zero the second condition is no longer obvious. In fact, for the reduced problem with $\varepsilon = 0$ the condition (3.3) requires

$$\|\boldsymbol{v}\|_0^2 \geq \beta \|\boldsymbol{v}\|_{\operatorname{div}}^2 \quad \forall \boldsymbol{v} \in \boldsymbol{Z}_h.$$

Hence, we must have

$$\|\operatorname{div} \boldsymbol{v}\|_0 \leq c \|\boldsymbol{v}\|_0 \quad \forall \boldsymbol{v} \in \boldsymbol{Z}_h \tag{3.4}$$

for a suitable constant c independent of h, and this condition does not hold for the common conforming stable Stokes elements.

Example 3.1. We consider the problem (1.1) with Ω taken as the unit square. The domain is triangulated by first dividing it into $h \times h$ squares. Then, each square is divided into two triangles by the diagonal with a negative slope. The system is then discretized using the $P_2 - P_0$ element with respect to this triangulation; i.e., $\boldsymbol{V}_h \subset \boldsymbol{H}_0^1$ consists of piecewise quadratic functions, while $Q_h \subset L_0^2$ is the space of discontinuous piecewise constants. This discretization is known to be stable when $\varepsilon > 0$ is fixed; cf. [10]. However, our purpose here is to investigate how the convergence behaves as ε becomes small.

We consider the system (1.1) with the function g chosen to be identical zero, while $\boldsymbol{f} = \boldsymbol{u} - \varepsilon^2 \boldsymbol{\Delta} \boldsymbol{u} - \mathbf{grad}\, p$, where $\boldsymbol{u} = \mathbf{curl} \sin^2(\pi x_1) \sin^2(\pi x_2)$ and $p = \sin(\pi x_1)$. Hence, in this example the solution is independent of ε.

In Table 3.1 we have computed the relative L^2 error in the velocity $\boldsymbol{u}$; i.e., $e(h) = \|\boldsymbol{u} - \boldsymbol{u}_h\|_0 / \|\boldsymbol{u}\|_0$ for different values of ε and h. A third order Gauss–Legendre rule (cf. [17]) was used here, and in all the other examples of this section, to perform the necessary integrations. For each fixed ε the convergence rate with respect to h, γ is estimated by assuming $e(h) = ch^\gamma$ and by computing a least squares fit to this log-linear relation.

When $\varepsilon = 1$ the convergence seems to be at least quadratic with respect to h in this case. However, the convergence deteriorates as ε becomes smaller, and for $\varepsilon = 0$ there is no convergence.

Table 3.2 is based on the corresponding relative errors in the energy norm, i.e., the norm $|\!|\!| \cdot |\!|\!|_\varepsilon$ for velocity and the L^2 norm for pressure. For simplicity only the estimated convergence rates are given.

TABLE 3.1
The relative L^2 error in velocity obtained by the $P_2 - P_0$ element.

$\varepsilon \backslash$ h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1	3.84e-2	4.75e-3	6.41e-4	1.04e-4	2.11e-5	2.72
2^{-2}	6.15e-2	1.73e-2	4.65e-3	1.20e-3	3.05e-4	1.92
2^{-4}	4.55e-1	2.10e-1	6.78e-2	1.86e-2	4.79e-3	1.67
2^{-8}	9.31e-1	9.68e-1	9.43e-1	8.14e-1	5.32e-1	0.19
0	9.35e-1	9.84e-1	1.00	1.01	1.02	-0.03

TABLE 3.2
Estimated convergence rates for the velocity and pressure, measured in the energy norm, for the $P_2 - P_0$ element.

ε	1	2^{-2}	2^{-4}	2^{-8}	0
rate, velocity	1.84	1.01	0.70	-0.79	-1.03
rate, pressure	1.06	1.01	1.09	0.13	-0.20

These results indicate a similar degenerate behavior with respect to ε. In fact, when $\varepsilon = 0$ the norm, $\|\boldsymbol{u}_h\|_\varepsilon$, seems to grow like h^{-1} as h approaches zero. This must be due to the fact that only the projection of $\operatorname{div} \boldsymbol{u}_h$ into piecewise constants is controlled by the method in this case. □

Example 3.2. We repeat the experiment above but with the difference that we use the nonconforming Crouzeix–Raviart element instead of the $P_2 - P_0$ element; i.e., $\boldsymbol{V}_h$ consists of piecewise linear vector fields which are continuous at the midpoint of each edge of the triangulation, while $Q_h \subset L_0^2$ is the space of piecewise constants. It is well known that for any fixed $\varepsilon > 0$ this element leads to a stable discretization; cf. [10].

In Table 3.3 we have again computed the relative L^2 error in the velocity $\boldsymbol{u}$ for different values of ε and h.

TABLE 3.3
The relative L^2 error in velocity obtained by the nonconforming Crouzeix–Raviart element.

$\varepsilon \backslash$ h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1	1.83e-1	4.89e-2	1.26e-2	3.19e-3	8.02e-4	1.96
2^{-2}	2.19e-1	6.89e-2	1.91e-2	4.96e-3	1.26e-3	1.87
2^{-4}	6.42e-1	3.86e-1	1.53e-1	4.58-2	1.21-3	1.45
2^{-8}	9.51e-1	1.00	1.01	9.43e-1	7.44e-1	0.08
0	9.53e-1	1.01	1.04	1.05	1.06	-0.04

The L^2 convergence appears to be quadratic when ε is large. However, also in this case the convergence deteriorates as ε decreases, and for the reduced problem, with $\varepsilon = 0$, the observed values for the relative error is monotonically increasing.

The corresponding estimates of the convergence rates in the energy norm decreases from approximately linear convergence to no convergence as is shown by Table 3.4.

In fact, the divergence of the Crouzeix–Raviart element in the case $\varepsilon = 0$ is not surprising. Since the divergence-free vector fields in this case can be realized as the curl operator applied to the corresponding Morley space, this behavior of the Crouzeix–Raviart element is closely tied to the divergence of the Morley element for the Poisson equation; cf. [14]. □

TABLE 3.4
Estimated convergence rates for the velocity and pressure, measured in the energy norm, for the Crouzeix–Raviart element.

ε	1	2^{-2}	2^{-4}	2^{-8}	0
rate, velocity	0.98	0.97	0.74	0.03	-0.03
rate, pressure	1.00	0.93	0.98	0.12	-0.03

The two examples above show that the $P_2 - P_0$ element and the nonconforming Crouzeix–Raviart element, which both are known to be stable for $\varepsilon = 1$, fail to give methods which converge uniformly in ε. The divergence of the $P_2 - P_0$ element for $\varepsilon = 0$ is basically due to the fact that the estimate (3.4) does not hold, and therefore the method is unstable, while the divergence of the Crouzeix–Raviart method is caused by the inconsistency of the method.

Example 3.3. We repeat the experiment above once more, but this time the system (1.1) is discretized by using the Mini element; i.e., $\boldsymbol{V}_h \subset \boldsymbol{H}_0^1$ consists of linear combinations of piecewise linear functions and cubic bubble functions with support on a single triangle, while $Q_h \subset L_0^2$ is the space of continuous piecewise linear functions.

In Table 3.5 we have computed the relative error in the velocity, with respect to the energy norm $|||\cdot|||_\varepsilon$, for different values of ε and h.

TABLE 3.5
The relative error in velocity, measured in the energy norm, for the Mini element.

$\varepsilon\backslash$ h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1	3.01	1.65	8.42e-1	4.22e-1	2.11e-1	0.96
2^{-2}	2.70	1.55	7.80e-1	3.90e-1	1.95e-1	0.96
2^{-4}	3.71	1.67	7.89e-1	3.87e-1	1.92e-1	1.07
2^{-8}	7.32	4.28	2.79	1.64	6.51e-1	0.84
0	7.44	4.76	3.70	3.39	3.30	0.28

When $\varepsilon = 1$ the convergence seems to be linear with respect to h. This agrees with the theoretical results given in [1]. The convergence deteriorates as ε becomes smaller, and for $\varepsilon = 0$ there seems to be essentially no convergence in the energy norm.

An interesting observation can be made for the Mini element if we consider the corresponding errors for the pressure p. In Table 3.6 we study the relative error given by $\|p - p_h\|_0 / \|p\|_0$.

TABLE 3.6
The relative L^2 error in the pressure obtained by the Mini element.

$\varepsilon\backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1	8.78	2.81	8.85e-1	2.95e-1	1.02e-1	1.61
2^{-2}	6.09e-1	1.84e-1	5.62e-2	1.85e-2	6.40e-3	1.64
2^{-4}	6.08e-2	1.51e-2	3.88e-3	1.21e-3	4.07e-4	1.81
2^{-8}	3.58e-2	9.93e-3	2.34e-3	4.10e-4	6.00e-5	2.30
0	3.59e-2	1.02e-2	2.75e-3	7.23e-4	1.87e-4	1.90

The surprising observation is that for the pressure the convergence seems to be uniform with respect to ε. In fact, the convergence rate seems to improve as ε tends

to zero, and for ε small the convergence with respect to h appears to be quadratic. This is a striking difference to what we observed in Examples 3.1 and 3.2. In both these cases the error in the pressure diverges as ε tend to zero; cf. Tables 3.2 and 3.4.

What we have observed here is not special to the present example. The Mini element leads to a discretization which is uniformly stable with respect to ε in a proper ε-dependent norm different from $|||\cdot|||_\varepsilon$. If we define the solution space $\boldsymbol{X}_\varepsilon$ by

$$\boldsymbol{X}_\varepsilon = (\boldsymbol{L}^2 \cap \varepsilon \cdot \boldsymbol{H}_0^1) \times ((\boldsymbol{H}^1 \cap L_0^2) + \varepsilon^{-1} \cdot L^2), \tag{3.5}$$

then it can be shown that the Mini element will in fact produce a uniformly stable discretization in the corresponding energy norm. This norm degenerates to the norm of $\boldsymbol{L}^2 \times H^1$ as ε tends to zero; cf. the discussion in section 2. In order to confirm this behavior we computed the relative error in velocity once more, but this time we used the L^2 norm instead of $|||\cdot|||_\varepsilon$. The results are given in Table 3.7.

TABLE 3.7
The relative L^2 error in velocity obtained by the Mini element.

$\epsilon \backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1	3.54e-1	1.03e-1	2.64e-2	6.60e-3	1.65e-3	1.95
2^{-2}	3.16e-1	8.79e-2	2.20e-2	5.48e-3	1.37e-3	1.97
2^{-4}	1.90e-1	4.60e-2	1.07e-2	2.59e-3	6.42e-4	2.06
2^{-8}	1.81e-1	7.23e-2	2.87e-2	8.70e-3	1.74e-3	1.64
0	1.82e-1	7.66e-2	3.59e-2	1.76e-2	8.75e-3	1.09

We observe that as ε decreases from one to zero the corresponding convergence rate decreases from approximately two to one. However, there is no sign which indicates that the behavior will deteriorate below linear convergence. To complete the picture we have also computed the estimated convergence rates for the pressure in H^1. The results are given in Table 3.8.

TABLE 3.8
Estimated convergence rates for the H^1 error of the pressure obtained by the Mini element.

ε	1	2^{-2}	2^{-4}	2^{-8}	0
rate	0.61	0.64	0.86	0.99	0.99

The estimated convergence rate is clearly below one when $\varepsilon = 1$, while it improves towards one as ε is decreased. This is consistent with the fact that the norm of the pressure component of the product space (3.5) is weaker than the H^1 norm for each $\varepsilon > 0$ but approaches the H^1 norm as ε approaches zero.

The results above seem to confirm that the Mini element leads to a uniformly convergent discretization as long as the error is properly measured. However, as motivated in section 2, in the present paper we are interested in a discretization of the system (1.1) which has a uniform behavior when the error is measured in $(\boldsymbol{H}_0(\mathrm{div}) \cap \varepsilon \cdot \boldsymbol{H}_0^1) \times L_0^2$. Therefore, for our purpose here, the Mini element should not be regarded as a uniformly stable element. ∎

Let us recall that if a standard conforming Stokes element is not uniformly stable with respect to ε, then this instability must be caused by the failure of the second stability condition (3.3) or, equivalently, (3.4). Note that the stability condition (3.4) will be trivially satisfied if the spaces $\boldsymbol{V}_h \times Q_h$ are constructed such that all elements of $\boldsymbol{Z}_h$ are divergence-free, i.e., $\boldsymbol{Z}_h \subset \boldsymbol{Z}$. In fact, nearly all proposed finite element

methods for the reduced problem will have this property. This is, for example, true for the Raviart–Thomas spaces (cf. [15]) and for the Brezzi–Douglas–Marini spaces of [7]. However, in all these cases the spaces $\boldsymbol{V}_h$ will be only a subspace of $\boldsymbol{H}_0(\mathrm{div})$ and not of $\boldsymbol{H}_0^1$, due to the fact that only the normal components of the elements of $\boldsymbol{V}_h$ are required to be continuous across element edges. It is therefore not clear that these spaces will be useful for problems of the form (1.1) with $\varepsilon > 0$.

Example 3.4. We repeat the calculation done in the three examples above, but now we use the lowest order Raviart–Thomas space for the discretization. Hence, for $\varepsilon = 0$ we will expect to obtain linear convergence with respect to h. On the other hand, for $\varepsilon > 0$ the method is nonconforming and there seems to be no reason to expect that the method is convergent in this case. In Table 3.9 we have computed the estimated convergence rates with respect to h for the relative L^2 errors of the velocity $\boldsymbol{u}$ and the pressure p for different values of ε.

TABLE 3.9
Estimated convergence rates for the L^2 errors of the velocity and pressure for the Raviart–Thomas element.

ε	1	2^{-2}	2^{-4}	2^{-8}	0
rate, velocity	-0.07	-0.07	0.28	0.97	0.97
rate, pressure	-0.04	0.08	0.86	1.01	1.01

As expected, the method appears to be divergent for $\varepsilon > 0$. ∎

4. A robust nonconforming finite element space. The four examples presented above illustrate that none of the standard elements proposed for the case $\varepsilon = 1$ or $\varepsilon = 0$ will lead to a discretization of the problem (1.1) with uniform convergence properties with respect to ε, when the error is measured in the norm of the space $(\boldsymbol{H}_0(\mathrm{div}) \cap \varepsilon \cdot \boldsymbol{H}_0^1) \times L_0^2$. The purpose of the rest of this paper is therefore to construct and analyze a new finite element space which has this property.

4.1. The finite element space. In order to describe the new finite element space we will first define the proper polynomial space, or shape functions, on a given triangle. Let $T \subset \mathbb{R}^2$ be a triangle and consider the polynomial space of vector fields on T given by

$$\boldsymbol{V}(T) = \{\boldsymbol{v} \in \mathbb{P}_3^2 : \operatorname{div} \boldsymbol{v} \in \mathbb{P}_0, \quad (\boldsymbol{v} \cdot \boldsymbol{n})|_e \in \mathbb{P}_1 \quad \forall e \in \mathcal{E}(T)\}.$$

Here $\mathbb{P}_k$ denotes the set of polynomials of degree k and $\mathcal{E}(T)$ denotes the set of the edges of T. Furthermore, $\boldsymbol{n}$ is the unit normal vector on the edge e. Below we will also use $\boldsymbol{t}$ to denote the unit tangent vector on e, while τ denotes the arc length along e.

The space $\mathbb{P}_3^2$ is a vector space of dimension 20. Furthermore, the conditions

$$\operatorname{div} \boldsymbol{v} \in \mathbb{P}_0 \quad \text{and} \quad (\boldsymbol{v} \cdot \boldsymbol{n})|_e \in \mathbb{P}_1 \quad \forall e \in \mathcal{E}(T)$$

represent at most 11 linearly independent constraints on this space. Therefore we must have

$$\dim \boldsymbol{V}(T) \geq 9.$$

In fact, we shall show that $\dim \boldsymbol{V}(T) = 9$.

LEMMA 4.1. *The space $\boldsymbol{V}(T)$ is a linear space of dimension nine. Furthermore, an element $\boldsymbol{v} \in \boldsymbol{V}(T)$ is uniquely determined by the following degrees of freedom:*

- $\int_e (\boldsymbol{v}\cdot\boldsymbol{n})\tau^k\,d\tau$, $k=0,1$, *for all* $e\in\mathcal{E}(T)$.
- $\int_e (\boldsymbol{v}\cdot\boldsymbol{t})\,d\tau$ *for all* $e\in\mathcal{E}(T)$.

Proof. Since $\boldsymbol{V}(T)$ is a vector space of dimension ≥ 9 it is enough to show that elements of $\boldsymbol{V}(T)$ are uniquely determined by the given nine degrees of freedom. Assume that $\boldsymbol{v}\in\boldsymbol{V}(T)$ with all the degrees of freedom equal zero. In particular, this implies that

$$(\boldsymbol{v}\cdot\boldsymbol{n})|_{\partial T}\equiv 0.$$

As a consequence of this

$$\int_T \operatorname{div}\boldsymbol{v}\,dx=\int_{\partial T}\boldsymbol{v}\cdot\boldsymbol{n}\,d\tau=0.$$

Hence, since $\operatorname{div}\boldsymbol{v}\in\mathbb{P}_0$, we conclude that $\boldsymbol{v}$ is divergence-free.

However, since $\boldsymbol{v}\in\mathbb{P}_3^2$ is divergence-free we must have $\boldsymbol{v}=\mathbf{curl}\,w$ for a suitable scalar function $w\in\mathbb{P}_4$. Furthermore, since

$$(\mathbf{grad}\,w\cdot\boldsymbol{t})|_e=(\boldsymbol{v}\cdot\boldsymbol{n})|_e=0$$

for each edge e, we conclude that $\mathbf{grad}\,w\cdot\boldsymbol{t}\equiv 0$ on ∂T. Since w is uniquely determined only up to a constant, we can therefore assume that $w\equiv 0$ on ∂T.

Hence, w is of the form $w=pb$, where $p\in\mathbb{P}_1$ and b is the cubic bubble function with respect to T; i.e., $b=\lambda_1\lambda_2\lambda_3$, where $\lambda_i(x)$ are the barycentric coordinates of x with respect to the three corners of T. In particular, $\frac{\partial b}{\partial\boldsymbol{n}}|_e$ does not change sign on e. Furthermore,

$$\left.\frac{\partial w}{\partial\boldsymbol{n}}\right|_{\partial T}=p\left.\frac{\partial b}{\partial\boldsymbol{n}}\right|_{\partial T}$$

and

$$\int_e p\frac{\partial b}{\partial\boldsymbol{n}}\,d\tau=\int_e\frac{\partial w}{\partial\boldsymbol{n}}\,d\tau=\int_e\boldsymbol{v}\cdot\boldsymbol{t}\,d\tau=0\quad\forall e\in\mathcal{E}(T).$$

We can therefore conclude that p has a root in the interior of e. However, if $p\in\mathbb{P}_1$ with a root in the interior of each edge of T, then $p\equiv w\equiv 0$. ∎

Let $\{\mathcal{T}_h\}$ be a shape regular family of triangulations of Ω, where h is the maximal diameter. Furthermore, let $\mathcal{E}_h$ be the set of edges of $\mathcal{T}_h$. Define a finite element space of vector fields $\boldsymbol{V}_h$, associated with the triangulation $\mathcal{T}_h$, as all functions $\boldsymbol{v}\in\boldsymbol{V}_h$ such that

- $\boldsymbol{v}|_T\in\boldsymbol{V}(T)$ for all $T\in\mathcal{T}_h$,
- $\int_e(\boldsymbol{v}\cdot\boldsymbol{n})\tau^k\,d\tau$ is continuous for $k=0,1$ for all $e\in\mathcal{E}_h$,
- $\int_e(\boldsymbol{v}\cdot\boldsymbol{t})\,d\tau$ is continuous for all $e\in\mathcal{E}_h$.

Here we assume that $\boldsymbol{v}$ is extended to be zero outside Ω; i.e., if e is an edge on the boundary of Ω, then we require

$$\int_e(\boldsymbol{v}\cdot\boldsymbol{n})\tau^k\,d\tau=0,\quad k=0,1,\quad\text{and}\quad\int_e(\boldsymbol{v}\cdot\boldsymbol{t})\,d\tau=0.$$

It follows from Lemma 4.1 that any function $\boldsymbol{v}\in\boldsymbol{V}_h$ is uniquely determined by the two lowest order moments of $\boldsymbol{v}\cdot\boldsymbol{n}$ and by the mean value of $\boldsymbol{v}\cdot\boldsymbol{t}$ for all interior edges; cf. Figure 4.1.

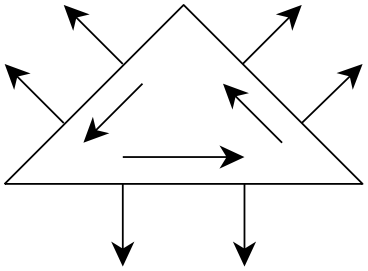


FIG. 4.1. *The degrees of freedom of the new nonconforming element.*

If $\boldsymbol{v} \in \boldsymbol{V}_h$, then the normal component $\boldsymbol{v} \cdot \boldsymbol{n}$ is continuous for all interior edges. Therefore, $\boldsymbol{V}_h \subset \boldsymbol{H}_0(\mathrm{div})$. However, the tangential component of $\boldsymbol{v}$ is not continuous; only the mean value with respect to each edge is continuous. Therefore, $\boldsymbol{V}_h \not\subseteq \boldsymbol{H}_0^1$. In addition to the space $\boldsymbol{V}_h$ we let $Q_h \subset L_0^2$ denote the space of scalar piecewise constants with respect to the triangulation $\mathcal{T}_h$.

In the rest of this paper $\boldsymbol{V}_h$ and Q_h will always refer to the finite element spaces just introduced. The corresponding nonconforming finite element approximation of the system (1.1) is defined by the system (3.1).

4.2. Properties of the new finite element space. It follows from the definition of $\boldsymbol{V}_h$ that $\operatorname{div} \boldsymbol{V}_h \subset Q_h$. Hence, if we define $\boldsymbol{Z}_h \subset \boldsymbol{V}_h$ as the weakly divergence-free elements of $\boldsymbol{V}_h$, i.e.,

$$\boldsymbol{Z}_h = \{\boldsymbol{v} \in \boldsymbol{V}_h : (\operatorname{div} \boldsymbol{v}, q) = 0 \quad \forall q \in Q_h\},$$

then these elements are in fact divergence-free.

Remark. It can be seen that

$$\boldsymbol{Z}_h = \mathbf{curl}\, W_h, \tag{4.1}$$

where W_h is an associated nonconforming H^2-element. Locally, on each triangle, W_h consists of all $\mathbb{P}_4$ polynomials which reduce to a quadratic on each edge. In addition, $W_h \subset H_0^1$ and the average of the normal derivatives of functions in W_h are continuous on each edge. The finite element space W_h is precisely described and analyzed in [14]. The identity (4.1) was actually the main motivation for the construction of the space $\boldsymbol{V}_h$. More precisely, the spaces W_h, $\boldsymbol{V}_h$, and Q_h are related such that the sequence

$$0 \longrightarrow W_h/\mathbb{R} \xrightarrow{\mathbf{curl}} \boldsymbol{V}_h \xrightarrow{\operatorname{div}} Q_h \longrightarrow 0$$

is exact. In particular, $\operatorname{div} \boldsymbol{V}_h = Q_h$. ∎

Define an interpolation operator $\boldsymbol{\Pi}_h : \boldsymbol{H}_0^1 \mapsto \boldsymbol{V}_h$ by

$$\int_e (\boldsymbol{\Pi}_h \boldsymbol{v} \cdot \boldsymbol{n}) \tau^k \, d\tau = \int_e (\boldsymbol{v} \cdot \boldsymbol{n}) \tau^k \, d\tau, \qquad k = 0, 1,$$
$$\int_e (\boldsymbol{\Pi}_h \boldsymbol{v} \cdot \boldsymbol{t}) \, d\tau = \int_e (\boldsymbol{v} \cdot \boldsymbol{t}) \, d\tau$$

for all $e \in \mathcal{E}_h$. In addition, let $P_h : L_0^2 \mapsto Q_h$ be the L^2 projection. From the definition of the operator $\boldsymbol{\Pi}_h$ we easily verify the commutativity property

$$\operatorname{div} \boldsymbol{\Pi}_h \boldsymbol{v} = P_h \operatorname{div} \boldsymbol{v} \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1. \tag{4.2}$$

In fact, for all $T \in \mathcal{T}_h$

$$\int_T \operatorname{div} \boldsymbol{\Pi}_h \boldsymbol{v}\, dx = \int_{\partial T} (\boldsymbol{\Pi}_h \boldsymbol{v} \cdot \boldsymbol{n})\, d\tau = \int_{\partial T} (\boldsymbol{v} \cdot \boldsymbol{n})\, d\tau = \int_T \operatorname{div} \boldsymbol{v}\, dx,$$

and hence (4.2) follows.

Since Q_h is the space of piecewise constants the L^2 projection P_h onto Q_h satisfies

$$\|w - P_h w\|_0 \le ch\|w\|_1 \tag{4.3}$$

for all $w \in H^1 \cap L_0^2$, where $c > 0$ is independent of h and w. The operator $\boldsymbol{\Pi}_h$ is well defined on $\boldsymbol{H}_0^1$, it is locally defined on each triangle, and it preserves linear functions locally. Furthermore, the polynomial space $\boldsymbol{V}(T)$ is invariant under affine Piola transformations. More precisely, let $T \in \mathcal{T}_h$ and let $\phi(x) = Bx + c$ be an affine map of T onto a reference triangle $\hat{T}$. Then the Piola transform, $\boldsymbol{v} \mapsto \hat{\boldsymbol{v}}$, where

$$\hat{\boldsymbol{v}}(\hat{x}) = (\det B)^{-1} B v(x), \qquad \hat{x} = \phi(x),$$

maps $\boldsymbol{V}(T)$ onto $\boldsymbol{V}(\hat{T})$. Therefore, approximation estimates for the operator $\boldsymbol{\Pi}_h$ can be derived from standard scaling arguments utilizing the shape regularity of $\{\mathcal{T}_h\}$. In particular, there exists a constant $c > 0$, independent of h, such that

$$\|\boldsymbol{\Pi}_h \boldsymbol{v}\|_{\operatorname{div}} \le \|\boldsymbol{\Pi}_h \boldsymbol{v}\|_{1,h} \le c\|\boldsymbol{v}\|_1. \tag{4.4}$$

In addition, from the Bramble–Hilbert lemma, using the fact that $\boldsymbol{\Pi}_h$ preserves linears locally, we can further conclude that

$$\|\boldsymbol{\Pi}_h \boldsymbol{v} - \boldsymbol{v}\|_{j,h} \le ch^{k-j}|\boldsymbol{v}|_k \quad \text{for } 0 \le j \le 1 \le k \le 2 \tag{4.5}$$

and for all $\boldsymbol{v} \in \boldsymbol{H}_0^1 \cap \boldsymbol{H}^k$. Here $\|\cdot\|_{j,h}$ denotes the piecewise $\boldsymbol{H}^j$-norm

$$\|\boldsymbol{v}\|_{j,h}^2 = \sum_{T \in \mathcal{T}_h} \|\boldsymbol{v}\|_{j,T}^2.$$

In fact, if $\hat{T}$ is a reference triangle, and $\hat{\boldsymbol{\Pi}} : \boldsymbol{H}^1(\hat{T}) \mapsto \boldsymbol{V}(\hat{T})$ the corresponding interpolation operator, then for all $\boldsymbol{v} \in H^1(\hat{T})$

$$\|\hat{\boldsymbol{\Pi}}\boldsymbol{v}\|_{0,\hat{T}} \le c_1\|\boldsymbol{v}\|_{0,\partial\hat{T}} \le c_2\|\boldsymbol{v}\|_{0,\hat{T}}^{1/2}\|\boldsymbol{v}\|_{1,\hat{T}}^{1/2},$$

where c_1 and c_2 depend only on $\hat{T}$. Hence, from a scaling argument we also obtain the low order estimate

$$\|\boldsymbol{\Pi}_h \boldsymbol{v} - \boldsymbol{v}\|_0 \le ch^{1/2}\|\boldsymbol{v}\|_0^{1/2}\|\boldsymbol{v}\|_1^{1/2} \tag{4.6}$$

for all $\boldsymbol{v} \in \boldsymbol{H}_0^1$.

Next we will verify the stability conditions (3.2) and (3.3) for the product space $\boldsymbol{V}_h \times Q_h$. However, due to the fact that we are considering a nonconforming finite element approximation of the system (1.1), where $\boldsymbol{V}_h \not\subseteq \boldsymbol{H}_0^1$, the norm $|||\cdot|||_\varepsilon$ has to be properly modified. For each $\boldsymbol{v} \in \boldsymbol{V}_h$ we define

$$|||\boldsymbol{v}|||_{\varepsilon,h}^2 = \|\boldsymbol{v}\|_{\operatorname{div}}^2 + \varepsilon^2 \sum_{T \in \mathcal{T}_h} \|\boldsymbol{D}\boldsymbol{v}\|_{0,T}^2.$$

Note that for $\varepsilon = 0$ this norm is simply equal to $\|\cdot\|_{\mathrm{div}}$, while for $\varepsilon = 1$ it is equivalent, uniformly in h, to the piecewise $\boldsymbol{H}^1$-norm $\|\cdot\|_{1,h}$.

LEMMA 4.2. *There exists a constant $\alpha_1 > 0$, independent of h, such that*

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{(q, \operatorname{div} \boldsymbol{v})}{\|\boldsymbol{v}\|_{1,h}} \geq \alpha_1 \|q\|_0 \quad \forall q \in Q_h.$$

Proof. This follows by a standard argument from the properties of the interpolation operator $\boldsymbol{\Pi}_h$ and the corresponding continuous result (2.6). In fact, since for any $\boldsymbol{v} \in \boldsymbol{H}_0^1$ and $q \in Q_h$ we have

$$(q, \operatorname{div} \boldsymbol{\Pi}_h \boldsymbol{v}) = (q, \operatorname{div} \boldsymbol{v})$$

and

$$\|\boldsymbol{\Pi}_h \boldsymbol{v}\|_{1,h} \leq c_1 \|\boldsymbol{v}\|_1,$$

we can take $\alpha_1 = \alpha_0/c_1$. ∎

The following uniform stability result is an immediate consequence of the previous lemma.

THEOREM 4.1. *The pair of spaces $(\boldsymbol{V}_h, Q_h)$ satisfies the uniform stability conditions* (3.2) *and* (3.3) *but with the norm $|||\cdot|||_\varepsilon$ replaced by $|||\cdot|||_{\varepsilon,h}$.*

Proof. The norms $|||\cdot|||_{1,h}$ and $\|\cdot\|_{1,h}$ are equivalent on $\boldsymbol{V}_h$, and $|||\cdot|||_{\varepsilon,h}$ decreases as ε decreases. It follows from Lemma 4.2 that condition (3.2) holds. Since $\boldsymbol{Z}_h \subset \boldsymbol{Z}$ the second condition (3.3) holds with $\beta = 1$. ∎

5. Error estimates for smooth solutions. Since our new finite element space $(\boldsymbol{V}_h, Q_h)$ satisfies the proper stability conditions (3.2) and (3.3), uniformly with respect to ε, it seems probable that the corresponding finite element method will in fact have uniform convergence properties. In the present section we shall investigate this question under the assumption that the solution (u, p) of the continuous problem is sufficiently smooth, while the effect of the ε-dependent boundary layers will be taken into account in the next section.

We will start the discussion here with a numerical example which is completely similar to Examples 3.1–3.3.

Example 5.1. We redo the computations done in Examples 3.1–3.3, but this time we use the finite element spaces constructed above. In all the numerical examples with the new element we used a fifth order Gauss–Legendre method (cf. [17]) as integration rule.

In Table 5.1 we have computed the estimated convergence rates with respect to h for the velocity and the pressure.

TABLE 5.1
Estimated convergence rates for the velocity and the pressure for the new nonconforming element.

ε	1	2^{-2}	2^{-4}	2^{-8}	0
rate, velocity in L^2	1.93	1.94	1.94	1.90	1.92
rate, velocity in $\|\|\|\cdot\|\|\|_\varepsilon$	0.98	0.99	1.05	1.72	1.92
rate, pressure in L^2	0.98	1.00	1.00	1.00	1.00

We observe that the convergence rates in L^2 appear to be close to quadratic in velocity and linear in pressure uniformly with respect to $\varepsilon \in [0, 1]$, while the

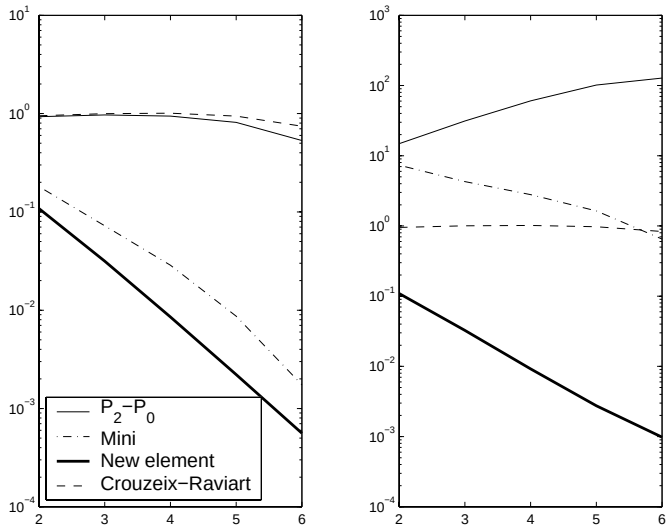


FIG. 5.1. *The errors in velocity, measured in the L^2 norm and the energy norm, as functions of $\sigma = -\log(h)/\log(2)$.*

convergence in the energy norm appears to be at least linear for each $\varepsilon > 0$. In fact, as ε approaches zero the convergence rate tends to two. This improved convergence is partly due to the fact that the exact solution $\boldsymbol{u}$ is divergence-free in this case.

To make a direct comparison between the $P_2 - P_0$ element, the Crouzeix–Raviart element, the Mini element, and the new element when ε is small compared to h, we have plotted the errors in velocity for the different methods as functions of σ, where $h = 2^{-\sigma}$. Here we have chosen $\varepsilon = 2^{-8}$. The errors are plotted, in a logarithmic scale, in Figure 5.1.

To the left the L^2 errors are plotted, while the errors in the energy norm are depicted to the right. We observe that the Mini element and the new element behave comparably with respect to the L^2 norm, while the new element clearly is superior to all the other methods with respect to the energy norm. ∎

The rest of this section will be devoted to establishing error estimates for the new nonconforming finite element method. Throughout this section we will assume that $\boldsymbol{u} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$, where $(\boldsymbol{u}, p)$ is the weak solution of (2.4). For convenience we also introduce the notation $\|\cdot\|_a$ for the norm on $\boldsymbol{V}_h$ associated with the bilinear form a_ε, i.e.,

$$\|\boldsymbol{v}\|_a^2 = \|\boldsymbol{v}\|_0^2 + \sum_{T \in \mathcal{T}_h} \|\boldsymbol{D}\boldsymbol{v}\|_{0,T}^2.$$

For any $\boldsymbol{v} \in \boldsymbol{V}_h$, we define the consistency error $E_{\varepsilon,h}(\boldsymbol{u}, \boldsymbol{v})$ by

$$E_{\varepsilon,h}(\boldsymbol{u}, \boldsymbol{v}) = \varepsilon^2 \sum_{e \in \mathcal{E}_h} \int_e (\operatorname{rot} \boldsymbol{u}) \ [\boldsymbol{v} \cdot \boldsymbol{t}] \, d\tau.$$

Here, if T_- and T_+ are two triangles, sharing an edge e, then $[w] = [w]_e = w|_{T_+} - w|_{T_-}$ denotes the jump of w across e, while $\boldsymbol{t}$ is the unit tangent vector along e corresponding to the clockwise direction on T_+. Since $[\boldsymbol{v} \cdot \boldsymbol{n}]_e = 0$ for any $\boldsymbol{v} \in \boldsymbol{V}_h$ it follows from

(2.2) and Green's theorem, in particular from (2.1), that

$$\begin{aligned} a_\varepsilon(\boldsymbol{u},\boldsymbol{v}) + (p, \operatorname{div}\boldsymbol{v}) &= (\boldsymbol{f},\boldsymbol{v}) + E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v}) \quad && \forall \boldsymbol{v} \in \boldsymbol{V}_h, \\ (\operatorname{div}\boldsymbol{u}, q) &= (g, q) && \forall q \in L^2_0, \end{aligned} \tag{5.1}$$

where the term $E_{\varepsilon,h}$ appears due to the fact that $\boldsymbol{V}_h \not\subseteq \boldsymbol{H}^1_0$.

In the error analysis below we will need proper estimates on the consistency error $E_{\varepsilon,h}$. The following bounds are therefore useful.

LEMMA 5.1. *If $\boldsymbol{u} \in \boldsymbol{H}^2 \cap \boldsymbol{H}^1_0$, then*

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})|}{\|\boldsymbol{v}\|_a} \le c\,\varepsilon \begin{cases} h\|\operatorname{rot}\boldsymbol{u}\|_1, \\ h^{1/2}\|\operatorname{rot}\boldsymbol{u}\|_1^{1/2}\|\operatorname{rot}\boldsymbol{u}\|_0^{1/2}, \end{cases}$$

where $c > 0$ is independent of ε and h.

Proof. Let $e \in \mathcal{E}_h$ and $\boldsymbol{v} \in \boldsymbol{H}^1_0 + \boldsymbol{V}_h$. Since the mean value with respect to e of $\boldsymbol{v}\cdot\boldsymbol{t}$ is zero, it follows from a standard scaling argument (cf., for example, [5, section 8.3] or [14, section 4] for similar arguments) that for any $\phi \in H^1$

$$\begin{aligned} \textstyle\int_e \phi[\boldsymbol{v}\cdot\boldsymbol{t}]d\tau &\le \inf_{\lambda,\mu\in\mathbb{R}} \|\phi-\lambda\|_{0,e}\|[\boldsymbol{v}\cdot\boldsymbol{t}-\mu]\|_{0,e} \\ &\le \begin{cases} ch|\phi|_{1,\Omega_e}(|\boldsymbol{v}|_{1,T_-}+|\boldsymbol{v}|_{1,T_+}), \\ ch^{1/2}|\phi|^{1/2}_{1,\Omega_e}\|\phi\|^{1/2}_{0,\Omega_e}(|\boldsymbol{v}|_{1,T_-}+|\boldsymbol{v}|_{1,T_+}). \end{cases} \end{aligned} \tag{5.2}$$

Here T_- and T_+ denote the two triangles meeting the edge e and $\Omega_e = T_- \cup T_+$. Since

$$|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})| \le \varepsilon^2 \sum_{e\in\mathcal{E}_h} \left| \int_e (\operatorname{rot}\boldsymbol{u})\ [\boldsymbol{v}\cdot\boldsymbol{t}]\, d\tau \right|,$$

the desired estimate follows by applying the estimate (5.2) with $\phi = \operatorname{rot}\boldsymbol{u}$, summing over all edges, and using the fact that

$$\sum_{e\in\mathcal{E}_h} |\boldsymbol{v}|^2_{1,T} \le \varepsilon^{-2} a_\varepsilon(\boldsymbol{v},\boldsymbol{v}). \qquad \square$$

Let $(\boldsymbol{u}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ be the approximation of $(\boldsymbol{u}, p)$ derived from the discrete system (3.1). From (3.1) and (5.1) we obtain

$$a_\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{v}) + (p-p_h, \operatorname{div}\boldsymbol{v}) = E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v}) \tag{5.3}$$

for all $\boldsymbol{v} \in \boldsymbol{V}_h$. Furthermore,

$$\operatorname{div}\boldsymbol{u}_h = P_h \operatorname{div}\boldsymbol{u} = \operatorname{div}\boldsymbol{\Pi}_h\boldsymbol{u}.$$

Therefore, taking $\boldsymbol{v} = \boldsymbol{\Pi}_h\boldsymbol{u} - \boldsymbol{u}_h$ in (5.3) we obtain

$$a_\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{\Pi}_h\boldsymbol{u}-\boldsymbol{u}_h) = E_{\varepsilon,h}(\boldsymbol{u}, \boldsymbol{\Pi}_h\boldsymbol{u}-\boldsymbol{u}_h).$$

Since a_ε is an inner product we further have

$$\begin{aligned} \|\boldsymbol{\Pi}_h\boldsymbol{u}-\boldsymbol{u}_h\|^2_a &\le \|\boldsymbol{u}-\boldsymbol{\Pi}_h\boldsymbol{u}\|^2_a + 2a_\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{\Pi}_h\boldsymbol{u}-\boldsymbol{u}_h) \\ &\le \|\boldsymbol{u}-\boldsymbol{\Pi}_h\boldsymbol{u}\|^2_a + 2E_{\varepsilon,h}(\boldsymbol{u}, \boldsymbol{\Pi}_h\boldsymbol{u}-\boldsymbol{u}_h). \end{aligned}$$

Hence, we conclude that

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_a \le 2\left(\|\boldsymbol{u}-\boldsymbol{\Pi}_h\boldsymbol{u}\|_a + \sup_{\boldsymbol{v}\in\boldsymbol{V}_h}\frac{|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})|}{\|\boldsymbol{v}\|_a}\right). \tag{5.4}$$

From this basic bound we easily derive the following error estimate.

THEOREM 5.1. *If $\boldsymbol{u}\in\boldsymbol{H}^2\cap\boldsymbol{H}_0^1$ and $p\in H^1\cap L_0^2$, then the following estimates hold:*

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}_h\|_0+\varepsilon\|\operatorname{rot}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 &\le c(h^2+\varepsilon h)\|\boldsymbol{u}\|_2,\\ \|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 &\le ch\|\operatorname{div}\boldsymbol{u}\|_1,\\ \|p-p_h\|_0 &\le ch(\|p\|_1+(\varepsilon+h)\|\boldsymbol{u}\|_2).\end{aligned}$$

Here $c>0$ is a constant independent of ε and h.

Remark. Here, and below, the differential operators $\boldsymbol{D}$ and rot, applied to vector fields in $\boldsymbol{V}_h$, are defined locally on each triangle of the triangulation $\mathcal{T}_h$. ∎

Proof. The first estimate is a direct consequence of (4.5), (5.4), and Lemma 5.1. The second estimate follows from the bound (4.3) and the fact that $\operatorname{div}\boldsymbol{u}_h = P_h\operatorname{div}\boldsymbol{u}$.

In order to establish the third estimate we first observe that (4.3) implies that

$$\|p-P_hp\|_0 \le ch\|p\|_1. \tag{5.5}$$

Hence, it remains only to estimate P_hp-p_h. However, from the modified inf-sup condition (3.2) (cf. Theorem 4.1) we obtain

$$\|P_hp-p_h\|_0 \le \alpha^{-1}\sup_{\boldsymbol{v}\in\boldsymbol{V}_h}\frac{(P_hp-p_h,\operatorname{div}\boldsymbol{v})}{|\!|\!|\boldsymbol{v}|\!|\!|_{\varepsilon,h}}.$$

Furthermore, for any $\boldsymbol{v}\in\boldsymbol{V}_h$ we have

$$\begin{aligned}(P_hp-p_h,\operatorname{div}\boldsymbol{v}) &= (p-p_h,\operatorname{div}\boldsymbol{v})\\ &= -a_\varepsilon(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{v})+E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v}),\end{aligned}$$

which implies that

$$|(P_hp-p_h,\operatorname{div}\boldsymbol{v})| \le \left(\|\boldsymbol{u}-\boldsymbol{u}_h\|_a+\sup_{\boldsymbol{v}\in\boldsymbol{V}_h}\frac{|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})|}{\|\boldsymbol{v}\|_a}\right)|\!|\!|\boldsymbol{v}|\!|\!|_{\varepsilon,h}$$

or

$$\|P_hp-p_h\|_0 \le \alpha^{-1}\left(\|\boldsymbol{u}-\boldsymbol{u}_h\|_a+\sup_{\boldsymbol{v}\in\boldsymbol{V}_h}\frac{|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})|}{\|\boldsymbol{v}\|_a}\right). \tag{5.6}$$

From the previous estimates we therefore obtain

$$\|P_hp-p_h\|_0 \le c(h^2+\varepsilon h)\|\boldsymbol{u}\|_2,$$

and together with (5.5) this establishes the desired estimate on the error $\|p-p_h\|_0$. ∎

Remark. As an alternative to the estimates given in Theorem 5.1 we can also obtain

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_0+\varepsilon\|\operatorname{rot}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 \le ch(\|\boldsymbol{u}\|_1+\varepsilon\|\boldsymbol{u}\|_2) \tag{5.7}$$

and

$$\|p - p_h\|_0 \leq ch(\|p\|_1 + \|\boldsymbol{u}\|_1 + \varepsilon\|\boldsymbol{u}\|_2). \tag{5.8}$$

These modifications are obtained if we use the estimate

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u}\|_0 \leq ch\|\boldsymbol{u}\|_1,$$

obtained from (4.5), in (5.4) instead of the corresponding quadratic estimate. Even if the modified estimates are weaker for uniformly smooth solutions, they are sometimes preferable for more singular solutions. □

6. Boundary layers and uniform error estimates. In general, we cannot expect that the norm $\|\boldsymbol{u}\|_2$ of the solution of (1.1) is bounded independently of ε. In fact, as ε approaches zero even $\|\operatorname{rot} \boldsymbol{u}\|_0$ should be expected to blow up. Hence, the convergence estimates given in Theorem 5.1 will deteriorate as ε becomes small. The following example shows that this behavior of the error is in fact real.

Example 6.1. In this example we study the convergence for an ε-dependent solution. Let $\boldsymbol{u} = \varepsilon \operatorname{\mathbf{curl}} e^{-x_1 x_2/\varepsilon}$, $p = \varepsilon e^{-x_1/\varepsilon}$, $\boldsymbol{f} = \boldsymbol{u} - \varepsilon^2 \boldsymbol{\Delta} \boldsymbol{u} - \operatorname{\mathbf{grad}} p$, and let g be identical zero. In fact, $\boldsymbol{u}$ is not the solution of the corresponding system (1.1), since the boundary conditions are not satisfied. However, the adaption of the new method to nonhomogeneous boundary conditions is straightforward.

The significance of the solution $\boldsymbol{u}$ just given is related to the fact that the quantities $\|\operatorname{rot} \boldsymbol{u}\|_0$ and $\varepsilon\|\operatorname{rot} \boldsymbol{u}\|_1$ both are of order $\varepsilon^{-1/2}$ as ε tends to zero. As we will see in Lemma 6.1, this behavior is typical for solutions of the singular perturbation problem (1.1). For solutions with this singular behavior the estimates (5.7) and (5.8) lead to error bounds of the form

$$|||\boldsymbol{u} - \boldsymbol{u}_h|||_\varepsilon, \|p - p_h\|_0 \leq ch\varepsilon^{-1/2}, \tag{6.1}$$

where c is a constant independent of ε and h. In Table 6.1 we have computed the absolute error $|||\boldsymbol{u} - \boldsymbol{u}_h|||_\varepsilon$ for different values of ε and h. For each fixed ε the convergence rate with respect to h is estimated.

TABLE 6.1
The absolute error in velocity, measured in the energy norm, obtained by the new nonconforming element.

$\varepsilon \backslash$ h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
2^{-2}	7.29e-2	3.60e-2	1.77e-2	8.75e-3	4.36e-3	0.98
2^{-6}	8.89e-2	5.88e-2	3.71e-2	2.06e-2	1.05e-2	0.77
2^{-8}	1.12e-1	6.89e-2	4.07e-2	2.66e-2	1.73e-2	0.67
2^{-10}	1.17e-1	8.16e-2	5.48e-2	3.34e-2	1.93e-2	0.65
2^{-12}	1.17e-1	8.20e-2	5.74e-2	4.02e-2	2.71e-2	0.52

We observe that for ε sufficiently large the convergence rate is approximately one, but the estimated rate decreases when ε approaches zero. These results seem to confirm the claim that the convergence is linear with respect to h for each fixed ε. However, when h is sufficiently large compared to ε we do not observe this linear rate.

In Table 6.2 we give the corresponding absolute L^2 errors for the pressure.

Again the estimated convergence rate is approximately one for ε large. Then it starts to decrease with ε as in Table 6.1. However, in this case the convergence rate

TABLE 6.2
The absolute L^2 error in the pressure obtained by the new nonconforming element.

$\varepsilon\backslash$ h	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
2^{-2}	2.32e-2	1.11e-2	5.36e-3	2.64e-3	1.31e-3	1.04
2^{-6}	9.00e-3	5.33e-3	2.62e-3	1.15e-3	4.61e-4	1.07
2^{-8}	5.28e-3	3.24e-3	2.18e-3	1.23e-3	5.97e-4	0.77
2^{-10}	4.93e-3	2.54e-3	1.33e-3	7.93e-4	5.32e-4	0.81
2^{-12}	4.92e-3	2.51e-3	1.24e-3	6.22e-4	3.27e-4	0.98

increases roughly back to one when ε is superclose to zero. We will comment on this phenomenon for the error of the pressure at the end of this section.

The estimate (6.1) does not imply uniform convergence with respect to ε for our new finite element method. However, as a consequence of the theory below, we will obtain an improved estimate of the form

$$|||\boldsymbol{u}-\boldsymbol{u}_h|||_\varepsilon, \|p-p_h\|_0 \le c\min(h^{1/2}, h\varepsilon^{-1/2}) \tag{6.2}$$

for solutions with a singular behavior similar to the solution $\boldsymbol{u}$ studied here. Note that this is in fact consistent with the results of Tables 6.1 and 6.2, where we never observe a convergence rate below a half. □

The main purpose of this section is to establish error estimates which are uniform with respect to the perturbation parameter ε. We shall show a uniform $O(h^{1/2})$ error estimate in the energy norm. We observe that if $g \in H^1 \cap L_0^2$, then it follows directly from Theorem 5.1 that

$$\|\operatorname{div}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 \le c\,h\|g\|_1, \tag{6.3}$$

where the constant c is independent of ε and h. Hence, we have uniform linear convergence for the error of the divergence. In contrast to this, the remaining part of the error will be affected by boundary layers as ε becomes small. However, the following uniform convergence estimate will be derived.

THEOREM 6.1. *If $\boldsymbol{f} \in \boldsymbol{H}(\mathrm{rot})$ and $g \in H_+^1$, then there is a constant c, independent of $\boldsymbol{f}$, g, ε, and h, such that*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_0 + \varepsilon\|\operatorname{rot}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 + \|p-p_h\|_0 \le c\,h^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}).$$

Here the Sobolev space H_+^1 is a space contained in H^1, with an associated norm, $\|\cdot\|_{1,+}$, slightly stronger than $\|\cdot\|_1$. This space will be precisely defined below.

The derivation of the uniform error estimate above will depend heavily on certain regularity estimates for the solution of the system (1.1). For example, we shall estimate the blowup of $\|\operatorname{rot}\boldsymbol{u}\|_1$ as ε approaches zero. We shall therefore first derive these regularity estimates.

For convenience of the reader we repeat the system (1.1):

$$\begin{aligned} (\boldsymbol{I}-\varepsilon^2\boldsymbol{\Delta})\boldsymbol{u} - \mathbf{grad}\,p &= \boldsymbol{f} \quad \text{in } \Omega,\\ \operatorname{div}\boldsymbol{u} &= g \quad \text{in } \Omega,\\ \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{6.4}$$

We also repeat that the domain Ω is a polygonal domain in $\mathbb{R}^2$. In fact, in the discussion of this section we shall assume that Ω in addition is *convex*. If $\varepsilon \in (0,1]$,

$\boldsymbol{f} \in \boldsymbol{L}^2$, and $g = 0$, then the corresponding weak solution admits the additional regularity that $(\boldsymbol{u}, p) \in (\boldsymbol{H}_0^1 \times L_0^2) \cap (\boldsymbol{H}^2 \times H^1)$. This regularity result follows directly from the result for the corresponding Stokes problem on a convex domain which can be found in [12, Corollary 7.3.3.5]. In fact, the same regularity holds for $g \neq 0$ if we restrict the data g to the space H_+^1.

In order to define this space let $x_1, x_2, \ldots, x_N \in \partial\Omega$ denote the vertices of Ω. The space H_+^1 is given by

$$H_+^1 = \left\{ g \in H^1 \cap L_0^2 : \int_\Omega \frac{|g(x)|^2}{|x - x_j|^2}\, dx < \infty,\ j = 1, 2, \ldots, N \right\},$$

with associated norm

$$\|g\|_{1,+}^2 = \|g\|_1^2 + \sum_{j=1}^{N} \int_\Omega \frac{|g(x)|^2}{|x - x_j|^2}\, dx.$$

Hence, functions in H_+^1 vanish weakly at each vertex of Ω.

It is established in [2] that

$$\operatorname{div}(\boldsymbol{H}^2 \cap \boldsymbol{H}_0^1) = H_+^1.$$

Furthermore, the divergence operator has a bounded right inverse, $\boldsymbol{R} : H_+^1 \mapsto \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$; i.e., $\operatorname{div} \boldsymbol{R}g = g$ for all $g \in H_+^1$ and

$$\|\boldsymbol{R}g\|_2 \leq c\|g\|_{1,+}.$$

Note that if $(\boldsymbol{u}, p)$ solves (6.4), then $(\boldsymbol{u} - \boldsymbol{R}g, p)$ solves a corresponding problem with $g = 0$. From the result in the case $g = 0$ we can therefore conclude that $(\boldsymbol{u}, p) \in (\boldsymbol{H}_0^1 \times L_0^2) \cap (\boldsymbol{H}^2 \times H^1)$ for any $(\boldsymbol{f}, g) \in \boldsymbol{L}^2 \times H_+^1$.

The following result gives an upper bound for the blowup of the norm $\|\operatorname{rot} \boldsymbol{u}\|_1$ as ε tends to zero.

LEMMA 6.1. *Assume that $\boldsymbol{f} \in \boldsymbol{H}(\mathrm{rot})$, $g \in H_+^1$, and let $(\boldsymbol{u}, p)$ be the corresponding solution of* (6.4). *There exists a constant $c > 0$, independent of ε, $\boldsymbol{f}$ and g, such that*

$$\varepsilon^{1/2} \|\operatorname{rot} \boldsymbol{u}\|_0 + \varepsilon^{3/2} \|\operatorname{rot} \boldsymbol{u}\|_1 \leq c\, (\|\operatorname{rot} \boldsymbol{f}\|_0 + \|g\|_{1,+}). \tag{6.5}$$

Proof. We first construct a function $\hat{\boldsymbol{u}} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$ such that

$$\operatorname{div} \hat{\boldsymbol{u}} = g \quad \text{and} \quad \operatorname{rot} \boldsymbol{\Delta} \hat{\boldsymbol{u}} = 0. \tag{6.6}$$

In fact, the function $\hat{\boldsymbol{u}}$ can be constructed by defining

$$\hat{\boldsymbol{u}} = \boldsymbol{R}g + \mathbf{curl}\, \psi,$$

with $\psi \in H_0^2$ being the weak solution of the biharmonic equation

$$\begin{aligned} \Delta^2 \psi &= \operatorname{rot} \boldsymbol{\Delta} \boldsymbol{R} g \qquad \text{in } \Omega, \\ \psi &= \frac{\partial \psi}{\partial \boldsymbol{n}} = 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

We observe that, since $\boldsymbol{R}g \in \boldsymbol{H}^2$, the right-hand side is in H^{-1}. Therefore, from the regularity of solutions of the biharmonic equation on convex domains (cf. [12,

Theorem 7.2.2.3]), we have that $\psi \in H^3$, and $\|\psi\|_3 \le c\|\operatorname{rot} \boldsymbol{\Delta R} g\|_{-1}$. Hence, $\hat{\boldsymbol{u}} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$, and

$$\|\hat{\boldsymbol{u}}\|_2 \le c\,\|g\|_{1,+}. \tag{6.7}$$

Furthermore, clearly $\operatorname{div} \hat{\boldsymbol{u}} = \operatorname{div} \boldsymbol{R} g = g$, and for any $\mu \in C_0^\infty$ we have

$$(\boldsymbol{\Delta}\hat{\boldsymbol{u}}, \mathbf{curl}\,\mu) = (\boldsymbol{\Delta R} g, \mathbf{curl}\,\mu) - (\Delta\psi, \Delta\mu) = 0.$$

Hence, the second property in (6.6) also holds.

Define $\boldsymbol{v} = \boldsymbol{u} - \hat{\boldsymbol{u}}$. Then $(\boldsymbol{v}, p) \in (\boldsymbol{H}_0^1 \times L_0^2) \cap (\boldsymbol{H}^2 \times H^1)$ is the weak solution of the problem

$$\begin{aligned} (\boldsymbol{I} - \varepsilon^2\boldsymbol{\Delta})\boldsymbol{v} - \mathbf{grad}\, p &= \hat{\boldsymbol{f}} && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{v} &= 0 && \text{in } \Omega, \\ \boldsymbol{v} &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{6.8}$$

where $\hat{\boldsymbol{f}} = \boldsymbol{f} + \varepsilon^2\boldsymbol{\Delta}\hat{\boldsymbol{u}} - \hat{\boldsymbol{u}}$. Clearly, $\hat{\boldsymbol{f}} \in \boldsymbol{L}^2$. In fact, $\hat{\boldsymbol{f}} \in \boldsymbol{H}(\mathrm{rot})$, since

$$\operatorname{rot} \hat{\boldsymbol{f}} = \operatorname{rot} \boldsymbol{f} - \operatorname{rot} \hat{\boldsymbol{u}}.$$

Furthermore, there is a constant c, independent of ε, $\boldsymbol{f}$ and g, such that

$$\|\operatorname{rot} \hat{\boldsymbol{f}}\|_0 \le c\,(\|\operatorname{rot} \boldsymbol{f}\|_0 + \|g\|_{1,+}). \tag{6.9}$$

Since $\boldsymbol{v} \in \boldsymbol{L}^2$ and $\operatorname{div} \boldsymbol{v} = 0$ there exists $\phi \in H^1$, uniquely determined up to a constant, such that $\boldsymbol{v} = \mathbf{curl}\,\phi$ [11, Theorem I.3.1]. Hence, since $\boldsymbol{v} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$, we can choose $\phi \in H^3 \cap H_0^2$. In fact, by applying the rot operator, as a map from $\boldsymbol{L}^2$ to H^{-1}, to the first equation of (6.8) we obtain

$$\begin{aligned} -\Delta\phi + \varepsilon^2\Delta^2\phi &= \operatorname{rot} \hat{\boldsymbol{f}} && \text{in } \Omega, \\ \phi = \frac{\partial\phi}{\partial \boldsymbol{n}} &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The function ϕ is uniquely determined by this problem. This singular perturbation problem was in fact studied in [14], where it was established that [14, Lemma 5.1]

$$\varepsilon^{1/2}\|\phi\|_2 + \varepsilon^{3/2}\|\phi\|_3 \le c\|\operatorname{rot} \hat{\boldsymbol{f}}\|_0,$$

and as a consequence

$$\varepsilon^{1/2}\|\operatorname{rot} \boldsymbol{v}\|_0 + \varepsilon^{3/2}\|\operatorname{rot} \boldsymbol{v}\|_1 \le c\|\operatorname{rot} \hat{\boldsymbol{f}}\|_0.$$

Therefore, since $\boldsymbol{u} = \boldsymbol{v} + \hat{\boldsymbol{u}}$, (6.7) and (6.9) imply

$$\begin{aligned} \varepsilon^{1/2}\|\operatorname{rot} \boldsymbol{u}\|_0 + \varepsilon^{3/2}\|\operatorname{rot} \boldsymbol{u}\|_1 &\le c\|\operatorname{rot} \hat{\boldsymbol{f}}\|_0 + \varepsilon^{1/2}(\|\operatorname{rot} \hat{\boldsymbol{u}}\|_0 + \varepsilon\|\operatorname{rot} \hat{\boldsymbol{u}}\|_1) \\ &\le c(\|\operatorname{rot} \boldsymbol{f}\|_0 + \|g\|_{1,+}). \end{aligned}$$

This completes the proof. ∎

In addition to the ε-dependent bound on the solution $(\boldsymbol{u}, p)$ of (6.4) derived above, we shall also need convergence estimates on how fast these solutions converge to the solution of the reduced system.

The reduced system corresponding to (6.4) is of the form

$$\begin{aligned} \boldsymbol{u}^0 - \mathbf{grad}\, p^0 &= \boldsymbol{f} && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u}^0 &= g && \text{in } \Omega, \\ \boldsymbol{u}^0 \cdot \boldsymbol{n} &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{6.10}$$

A precise weak formulation of this system is given by the following:

Find $(\boldsymbol{u}^0, p^0) \in \boldsymbol{H}_0(\operatorname{div}) \times L^2_0$ such that

$$\begin{aligned} (\boldsymbol{u}^0, \boldsymbol{v}) + (p^0, \operatorname{div} \boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}) && \forall \boldsymbol{v} \in \boldsymbol{H}_0(\operatorname{div}), \\ (\operatorname{div} \boldsymbol{u}^0, q) &= (g, q) && \forall q \in L^2_0. \end{aligned} \tag{6.11}$$

If $(\boldsymbol{f}, g) \in \boldsymbol{H}^{-1}(\operatorname{rot}) \times L^2_0$, then this system admits a unique solution. In fact, if $\boldsymbol{f} \in \boldsymbol{H}(\operatorname{rot})$, then $\boldsymbol{u}^0 \in \boldsymbol{H}(\operatorname{rot})$ with $\operatorname{rot} \boldsymbol{u}^0 = \operatorname{rot} \boldsymbol{f}$. Therefore,

$$\boldsymbol{u}^0 \in \boldsymbol{H}_0(\operatorname{div}) \cap \boldsymbol{H}(\operatorname{rot}),$$

and hence (cf. [11, Proposition 3.1, Chapter 1]) $\boldsymbol{u}^0 \in \boldsymbol{H}^1$. As a consequence, $p^0 \in H^1$. Furthermore, the corresponding solution map is continuous; i.e., there exists a constant c, independent of $\boldsymbol{f}$ and g, such that

$$\|\boldsymbol{u}^0\|_1 + \|p^0\|_1 \le c\,(\|\boldsymbol{f}\|_{\operatorname{rot}} + \|g\|_0). \tag{6.12}$$

LEMMA 6.2. *Assume that $\boldsymbol{f} \in \boldsymbol{H}(\operatorname{rot})$, $g \in H^1_+$, and let $(\boldsymbol{u}, p)$ be the corresponding solution of* (6.4). *There exists a constant $c > 0$, independent of ε, $\boldsymbol{f}$ and g, such that*

$$\|\boldsymbol{u} - \boldsymbol{u}^0\|_0 + \|p - p^0\|_1 \le c\,\varepsilon^{1/2}(\|\boldsymbol{f}\|_{\operatorname{rot}} + \|g\|_{1,+}).$$

Proof. It follows from (2.2), the weak formulation of (6.4), and Green's theorem that for any $\boldsymbol{v} \in \boldsymbol{H}^1 \cap \boldsymbol{H}_0(\operatorname{div})$ the solution $(\boldsymbol{u}, p)$ satisfies

$$\begin{aligned} (\boldsymbol{u}, \boldsymbol{v}) + \varepsilon^2(\operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{v}) + \varepsilon^2(\operatorname{rot} \boldsymbol{u}, \operatorname{rot} \boldsymbol{v}) + \varepsilon^2 \int_{\partial\Omega} (\operatorname{rot} \boldsymbol{u})(\boldsymbol{v} \cdot \boldsymbol{t}) d\tau \\ + (p, \operatorname{div} \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}). \end{aligned}$$

By subtracting from this the first equation of (6.11), we obtain

$$(\boldsymbol{u} - \boldsymbol{u}^0, \boldsymbol{v}) + \varepsilon^2(\operatorname{rot} \boldsymbol{u}, \operatorname{rot} \boldsymbol{v}) + \varepsilon^2 \int_{\partial\Omega} (\operatorname{rot} \boldsymbol{u})(\boldsymbol{v} \cdot \boldsymbol{t}) d\tau = 0$$

for any $\boldsymbol{v} \in \boldsymbol{H}^1 \cap \boldsymbol{H}_0(\operatorname{div})$ with $\operatorname{div} \boldsymbol{v} = 0$. Hence, if we take $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}^0$, and observe that $\operatorname{rot} \boldsymbol{u}^0 = \operatorname{rot} \boldsymbol{f}$ and $\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}^0) = 0$, we derive the identity

$$\|\boldsymbol{u} - \boldsymbol{u}^0\|_0^2 + \varepsilon^2 \|\operatorname{rot} \boldsymbol{u}\|_0^2 = \varepsilon^2 \int_{\partial\Omega} (\operatorname{rot} \boldsymbol{u})(\boldsymbol{u}^0 \cdot \boldsymbol{t}) d\tau + \varepsilon^2(\operatorname{rot} \boldsymbol{u}, \operatorname{rot} \boldsymbol{f}),$$

which immediately leads to the bound

$$\|\boldsymbol{u} - \boldsymbol{u}^0\|_0^2 + \frac{\varepsilon^2}{2} \|\operatorname{rot} \boldsymbol{u}\|_0^2 \le \varepsilon^2 \|\operatorname{rot} \boldsymbol{f}\|_0^2 + \varepsilon^2 \int_{\partial\Omega} (\operatorname{rot} \boldsymbol{u})(\boldsymbol{u}^0 \cdot \boldsymbol{t}) d\tau. \tag{6.13}$$

In order to estimate the boundary integral we note that it follows from Lemma 6.1 and [12, Theorem 1.5.1.10] that

$$\|\operatorname{rot} \boldsymbol{u}\|_{0,\partial\Omega} \le c \|\operatorname{rot} \boldsymbol{u}\|_0^{1/2} \|\operatorname{rot} \boldsymbol{u}\|_1^{1/2} \le c\varepsilon^{-1} (\|\operatorname{rot} \boldsymbol{f}\|_0 + \|g\|_{1,+}).$$

Together with the estimate (6.12) this leads to

$$\varepsilon^2 \int_{\partial\Omega} (\operatorname{rot} \boldsymbol{u})(\boldsymbol{u}^0 \cdot \boldsymbol{t}) d\tau \le \varepsilon^2 \| \operatorname{rot} \boldsymbol{u}\|_{0,\partial\Omega} \|\boldsymbol{u}^0\|_1 \le c\,\varepsilon(\|\boldsymbol{f}\|^2_{\mathrm{rot}} + \|g\|^2_{1,+}).$$

Hence, the estimate

$$\|\boldsymbol{u} - \boldsymbol{u}^0\|_0 + \varepsilon^2 \| \operatorname{rot} \boldsymbol{u}\|_0^2 \le c\,\varepsilon^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}) \tag{6.14}$$

follows.

The estimate for $\|p - p^0\|_1$ is now a direct consequence of the identity

$$\begin{aligned} \mathbf{grad}(p - p^0) &= \boldsymbol{u} - \boldsymbol{u}^0 - \varepsilon^2 \boldsymbol{\Delta} \boldsymbol{u} \\ &= \boldsymbol{u} - \boldsymbol{u}^0 + \varepsilon^2(\mathbf{curl} \operatorname{rot} \boldsymbol{u} - \mathbf{grad}\, g) \end{aligned}$$

and the previously established bounds. In fact, it follows from Lemma 6.1 and (6.14) that

$$\begin{aligned} \| \mathbf{grad}(p - p^0)\|_0 &\le \|\boldsymbol{u} - \boldsymbol{u}^0\|_0 + \varepsilon^2(\| \operatorname{rot} \boldsymbol{u}\|_1 + \|g\|_1) \\ &\le c\,\varepsilon^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}). \end{aligned}$$

Since $p - p^0 \in L^2_0$, an application of the Poincaré inequality completes the proof. ∎

The regularity bounds derived above will now be used to prove the uniform convergence estimates.

Proof of Theorem 6.1. Recall that since $\boldsymbol{u} \in \boldsymbol{H}^1_0$ it follows from [11, Proposition 3.1, Chapter 1] that

$$\|\boldsymbol{u}\|_1 \le c(\| \operatorname{div} \boldsymbol{u}\|_0 + \| \operatorname{rot} \boldsymbol{u}\|_0).$$

Furthermore, by the standard H^2-regularity for solutions of the Poisson equation on convex domains, and (2.2), we obtain

$$\|\boldsymbol{u}\|_2 \le c\|\boldsymbol{\Delta} \boldsymbol{u}\|_0 \le c(\| \operatorname{div} \boldsymbol{u}\|_1 + \| \operatorname{rot} \boldsymbol{u}\|_1).$$

Hence, from the estimates given in Lemmas 6.1 and 6.2 we conclude that

$$\varepsilon^2 \|\boldsymbol{u}\|_2 + \varepsilon \|\boldsymbol{u}\|_1 + \|\boldsymbol{u} - \boldsymbol{u}^0\|_0 + \|p - p^0\|_1 \le c\varepsilon^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}). \tag{6.15}$$

The desired estimate on the velocity error will be derived from (5.4). We will first establish the interpolation estimate

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u}\|_0 + \varepsilon \|\boldsymbol{D}(\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u})\|_1 \le c\, h^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}). \tag{6.16}$$

From (4.6), (6.12), and (6.15) we have

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u}\|_0 &\le \|(\boldsymbol{I} - \boldsymbol{\Pi}_h)(\boldsymbol{u} - \boldsymbol{u}^0)\|_0 + \|\boldsymbol{u}^0 - \boldsymbol{\Pi}_h \boldsymbol{u}^0\|_0 \\ &\le c h^{1/2} \left(\|\boldsymbol{u} - \boldsymbol{u}^0\|_0^{1/2} \|\boldsymbol{u} - \boldsymbol{u}^0\|_1^{1/2} + h^{1/2} \|\boldsymbol{u}^0\|_1\right) \\ &\le c h^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}} + \|g\|_{1,+}). \end{aligned}$$

Furthermore, from (4.4), (4.5), and (6.15),

$$\varepsilon\|\boldsymbol{D}(\boldsymbol{u}-\boldsymbol{\Pi}_h\boldsymbol{u})\|_0 \le c\varepsilon\|\boldsymbol{u}\|_1^{1/2}\|\boldsymbol{u}-\boldsymbol{\Pi}_h\boldsymbol{u}\|_1^{1/2} \le c\varepsilon h^{1/2}\|\boldsymbol{u}\|_1^{1/2}\|\boldsymbol{u}\|_2^{1/2} \le ch^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+}).$$

The estimate (6.16) is therefore verified.

Similarly, since $\|\boldsymbol{u}\|_1^{1/2}\|\boldsymbol{u}\|_2^{1/2} \le c\varepsilon^{-1}(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+})$, we obtain from Lemma 5.1 that

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h}\frac{|E_{\varepsilon,h}(\boldsymbol{u},\boldsymbol{v})|}{\|\boldsymbol{v}\|_a} \le ch^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+}). \tag{6.17}$$

However, by combining (5.4), (6.3), (6.16), and (6.17), this implies

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_0+\varepsilon\|\operatorname{rot}(\boldsymbol{u}-\boldsymbol{u}_h)\|_0 \le ch^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+}). \tag{6.18}$$

In order to establish the estimate for the $\|p-p_h\|_0$ note that (4.3) and (6.15) imply

$$\|P_hp-p\|_0 \le ch\|p\|_1 \le ch(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+}).$$

Finally, by (5.6), (6.17), and (6.18),

$$\|P_hp-p_h\|_0 \le ch^{1/2}(\|\boldsymbol{f}\|_{\mathrm{rot}}+\|g\|_{1,+}).$$

This completes the proof of Theorem 6.1. ∎

Remark. Even if Lemma 6.2 states that $\|p\|_1$ is uniformly bounded with respect to ε, we are not able to prove that $\|p-p_h\|_0$ converges linearly in h uniformly in ε. The convergence rate is polluted by the blowup of $\boldsymbol{u}$. This seems to agree with what we observed in Example 6.1; cf. Table 6.2. ∎

7. An associated elliptic system. In this section we shall study the elliptic system (1.2) given by

$$\begin{aligned}(\boldsymbol{I}-\varepsilon^2\boldsymbol{\Delta})\boldsymbol{u}-\delta^{-2}\,\mathbf{grad}(\operatorname{div}\boldsymbol{u}-g) &= \boldsymbol{f} \quad \text{in } \Omega,\\ \boldsymbol{u} &= 0 \quad \text{on } \partial\Omega,\end{aligned} \tag{7.1}$$

where $\varepsilon,\delta \in (0,1]$. Recall that by introducing $p=\delta^{-2}\operatorname{div}\boldsymbol{u}$ this system can be alternatively written on the mixed form (1.3). Hence, as δ approaches zero the system formally reduces to (1.1).

The system (7.1) will be discretized by a standard finite element approach; i.e., the mixed system (1.3) is not introduced in the discretization. Let the bilinear form $b_{\varepsilon,\delta}(\cdot,\cdot)$ be defined by

$$\begin{aligned}b_{\varepsilon,\delta}(\boldsymbol{u},\boldsymbol{v}) &= a_\varepsilon(\boldsymbol{u},\boldsymbol{v})+\delta^{-2}(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{v})\\ &= (\boldsymbol{u},\boldsymbol{v})+\varepsilon^2(\boldsymbol{D}\boldsymbol{u},\boldsymbol{D}\boldsymbol{v})+\delta^{-2}(\operatorname{div}\boldsymbol{u},\operatorname{div}\boldsymbol{v}).\end{aligned}$$

For a given finite element space $\boldsymbol{V}_h$, the corresponding standard finite element discretization of (7.1) is given by the following:

Find a $\boldsymbol{u}_h \in \boldsymbol{V}_h$ such that

$$b_{\varepsilon,\delta}(\boldsymbol{u}_h,\boldsymbol{v}) = (\boldsymbol{f},\boldsymbol{v})+\delta^{-2}(g,\operatorname{div}\boldsymbol{v}) \qquad \forall\boldsymbol{v}\in\boldsymbol{V}_h. \tag{7.2}$$

Our purpose here is to discuss this discretization when the finite element space $\boldsymbol{V}_h$ is the space introduced in section 4. Since this space is not a subspace of $\boldsymbol{H}_0^1$ this will lead to a nonconforming discretization of the system (7.1). However, before we analyze this discretization, we will present some numerical experiments based on the system (7.1).

Example 7.1. In all the examples presented in this section we consider the system (7.1) with $\boldsymbol{u} = \mathbf{curl}\sin^2(\pi x_1)\sin^2(\pi x_2)$, $g = 0$, and $\boldsymbol{f} = \boldsymbol{u} - \varepsilon^2\boldsymbol{\Delta u}$. Hence, the solution is independent of ε and δ.

We consider the problem (7.1) with Ω taken as the unit square. The domain is triangulated as described in Example 3.1. The system is then discretized by solving the system (7.2), where the space $\boldsymbol{V}_h$ is the standard space of continuous piecewise linear functions with respect to this triangulation.

In the present example we have used $\varepsilon = 1$, while δ and h vary. In Table 7.1 we have computed the relative error in the L^2 norm for different values of δ and h.

TABLE 7.1
The relative L^2 error using piecewise linear elements, $\varepsilon = 1$.

$\delta\backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1.00	3.87e-1	1.32e-1	3.69e-2	9.52e-3	2.39e-3	1.85
0.10	9.19e-1	7.28e-1	4.34e-1	1.88e-1	6.20e-2	0.97
0.01	1.00	9.96e-1	9.82e-1	9.32e-1	7.88e-1	0.08

As expected we observe approximately quadratic convergence with respect to h for $\delta = 1$. However, the convergence clearly deteriorates as δ tends to zero. □

Example 7.2. We repeat the experiment above, but we extend the finite element space and use the corresponding velocity space of the Mini element instead of the piecewise linear space. It is interesting to note that the L^2 convergence deteriorates, as δ gets small, also in this case, in contrast to what we have observed in Table 3.7. The relative L^2 error is given in Table 7.2.

TABLE 7.2
The relative L^2 error using the Mini element, $\varepsilon = 1$.

$\delta\backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1.00	3.80e-1	1.30e-1	3.62e-2	9.34e-3	2.35e-3	1.85
0.10	9.19e-1	7.28e-1	4.34e-1	1.88e-1	6.20e-2	0.97
0.01	9.99e-1	9.96e-1	9.82e-1	9.33e-1	7.88e-1	0.08

We observe that the results are almost identical to the ones we obtained in the piecewise linear case. Hence, the extra bubble functions have almost no effect. Of course, the main reason for the difference between the results given here, for δ small, and the results given in Example 3.3, where $\delta = 0$, is that the second equation of the mixed method used previously implicitly introduces a reduced integration in the divergence term. □

Example 7.3. We repeat the experiment above once more, but this time we use the new nonconforming element. In Table 7.3 we have computed the relative error in the energy norm, i.e., the norm generated by the form $b_{\varepsilon,\delta}$, for different values of δ and h.

In contrast to the other examples above, in this case the convergence seems to be linear with respect to h, uniformly in δ. We also observe that the errors are almost

TABLE 7.3
The relative error in energy norm for the new nonconforming element, $\varepsilon = 1$.

$\delta \backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1.00	1.84	9.83e-1	4.98e-1	2.50e-1	1.25e-1	0.97
0.10	1.83	9.66e-1	4.87e-1	2.44e-1	1.22e-1	0.98
0.01	1.83	9.66e-1	4.87e-1	2.44e-1	1.22e-1	0.98

TABLE 7.4
The relative error in energy norm for the new nonconforming element, $\varepsilon = 0.01$.

$\delta \backslash h$	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	Rate
1.00	1.04e-1	3.23e-2	8.94e-3	2.21e-3	5.29e-4	1.91
0.10	1.04e-1	3.23e-2	8.94e-3	2.21e-3	5.29e-4	1.91
0.01	1.04e-1	3.23e-2	8.94e-3	2.21e-3	5.29e-4	1.91

independent of δ.

Next, we reduce ε and take $\varepsilon = 0.01$ and redo the experiment. The results are given in Table 7.4.

We observe that to the given accuracy the numerical solution is independent of δ, clearly indicating that the numerical solutions are close to a pure curl field independent of δ, which is precisely the form of the exact solution in this case. A similar observation is done if we take $\varepsilon = 0$. □

The numerical experiments just presented indicate that the nonconforming space $\boldsymbol{V}_h$, introduced in section 4 above, is well suited for problem (7.1). We will give a partial theoretical justification for this claim by deriving a generalization of Theorem 5.1.

We assume throughout this section that $\boldsymbol{u} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$. Let $\|\cdot\|_b$ be the energy norm associated with the system (7.1), i.e.,

$$\|\boldsymbol{v}\|_b^2 = b_{\varepsilon,\delta}(\boldsymbol{v}, \boldsymbol{v}).$$

It is a straightforward consequence of the second Strang lemma (cf. [9, Theorem 4.2.2]) that there exists a $c > 0$, independent of ε, h and $\boldsymbol{u}$, such that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_b^2 \leq \|\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u}\|_b^2 + c \sup_{\boldsymbol{v} \in \boldsymbol{V}_h} \frac{|E_{\varepsilon,h}(\boldsymbol{u}, \boldsymbol{v})|^2}{\|\boldsymbol{v}\|_b^2}, \tag{7.3}$$

where the inconsistency error $E_{\varepsilon,h}$ is introduced in section 5. However, since $\|\boldsymbol{v}\|_b \geq \|\boldsymbol{v}\|_a$, the inconsistency term can be bounded as in Lemma 5.1. Furthermore, (4.5) implies

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u}\|_a \leq c(h^2 + \varepsilon h)\|\boldsymbol{u}\|_2.$$

As a consequence of the fact that $\operatorname{div} \boldsymbol{\Pi}_h \boldsymbol{u} = P_h \operatorname{div} \boldsymbol{u}$, it is also true that

$$\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_0^2 = \|\operatorname{div}(\boldsymbol{u} - \boldsymbol{\Pi}_h \boldsymbol{u})\|_0^2 + \|\operatorname{div}(\boldsymbol{\Pi}_h \boldsymbol{u} - \boldsymbol{u}_h)\|_0^2.$$

Thus, we can conclude from (7.3) that

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|_a^2 + \delta^{-2}\|(I - P_h)\operatorname{div} \boldsymbol{u}\|_0^2 + \delta^{-2}\|\operatorname{div}(\boldsymbol{\Pi}_h \boldsymbol{u} - \boldsymbol{u}_h)\|_0^2 \\ \leq c(h^2 + \varepsilon h)^2 \|\boldsymbol{u}\|_2^2 + \delta^{-2}\|(I - P_h)\operatorname{div} \boldsymbol{u}\|_0^2. \end{aligned}$$

We therefore have established the following convergence result.

THEOREM 7.1. *If* $\boldsymbol{u} \in \boldsymbol{H}^2 \cap \boldsymbol{H}_0^1$, *then*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 + \varepsilon \|\operatorname{rot}(\boldsymbol{u} - \boldsymbol{u}_h)\|_0 + \delta^{-1} \|\operatorname{div}(\boldsymbol{\Pi}_h \boldsymbol{u} - \boldsymbol{u}_h)\|_0 \leq c(h^2 + \varepsilon h)\|\boldsymbol{u}\|_2.$$

Here $c > 0$ *is a constant independent of* ε, δ, *and* h.

Note that from this result we can conclude that if ε and h are fixed, and δ approaches zero, then $\operatorname{div} \boldsymbol{u}_h$ converges in L^2 to $P_h \operatorname{div} \boldsymbol{u}$. Furthermore, the divergence of the error can be controlled by this estimate since

$$\begin{aligned} \|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|_0 &\leq \|(I - P_h) \operatorname{div} \boldsymbol{u}\|_0 + \|\operatorname{div}(\boldsymbol{\Pi}_h \boldsymbol{u} - \boldsymbol{u}_h)\|_0 \\ &\leq ch \|\operatorname{div} \boldsymbol{u}\|_1 + c\delta(\varepsilon^2 + h\varepsilon)\|\boldsymbol{u}\|_2. \end{aligned}$$

Of course, exactly as for the problem (1.1) we can argue that, in general cases, the norm $\|\boldsymbol{u}\|_2$ will not remain bounded as ε and δ approach zero. Hence, ideally we would like to generalize the results of section 6 to the problem (7.1). However, this discussion is outside the scope of this paper.

Acknowledgments. The authors are grateful to Professors D.N. Arnold, R.S. Falk, and Z. Cai for many useful discussions.

REFERENCES

[1] D.N. ARNOLD, F. BREZZI, AND M. FORTIN, *A stable finite element method for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[2] D.N. ARNOLD, L.R. SCOTT, AND M. VOGELIUS, *Regular inversion of the divergence operator with Dirichlet boundary conditions on a polygon*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 15 (1988), pp. 169–192.

[3] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces*, Springer-Verlag, Berlin, New York, 1976.

[4] J.H. BRAMBLE AND J.E. PASCIAK, *Iterative techniques for time dependent Stokes problem*, Comput. Math. Appl., 33 (1997), pp. 13–30.

[5] S.C. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[6] F. BREZZI, *On the existence, uniqueness and approximation of saddle–point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationelle Sér. Rouge Anal. Numér., 8 (1974), pp. 129–151.

[7] F. BREZZI, J. DOUGLAS, AND L.D. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[8] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[9] P.G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] M. CROUZEIX AND P.A. RAVIART, *Conforming and non–conforming finite element methods for solving the stationary Stokes equations*, Rev. Française Automat. Informat. Recherche Opérationelle Sér. Rouge Anal. Numér., 7 (1973), pp. 33–76.

[11] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.

[12] P. GRISVARD, *Elliptic Problems on Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[13] E. HAUG, T. RUSTEN, AND H. THEVIK, *A mathematical model for macrosegregation formation in binary alloy solidification*, in Numerical Methods and Software Tools in Industrial Mathematics, M. Dæhlen and A. Tveito, eds., Birkhäuser, Boston, 1997.

[14] T.K. NILSSEN, X-C. TAI, AND R. WINTHER, *A robust nonconforming* H^2*–element*, Math. Comp., 70 (2000), pp. 489–505.

[15] P.A. RAVIART AND J.M. THOMAS, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977.

[16] S. WHITAKER, *Flow in porous media* I: *A theoretical derivation of Darcy's law*, Transp. Porous Media, 1 (1986), pp. 3–25.

[17] O.C. ZIENKIEWICZ AND R.L. TAYLOR, *The Finite Element Method*, Butterworth-Heinemann, Oxford, UK, 2000.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1632–1649

BENDING MOMENT MIXED METHOD FOR THE KIRCHHOFF–LOVE PLATE MODEL*

MOHAMED AMARA†, DANIELA CAPATINA-PAPAGHIUC†, AND AMNA CHATTI‡

Abstract. We deal with the Kirchhoff–Love model for a bending thin plate with physical boundary conditions. We propose here a new mixed formulation, based on a decomposition of the bending moment. For its discretization, we employ classical low-order conforming finite elements. Then the discrete formulation allows us to obtain directly an approximation of the bending moment, while the deflection is recovered by solving an additional second order elliptic problem. We establish optimal error estimates which prove that the method is unconditionally convergent. Moreover, its convergence rate is optimal whenever the exact solution is sufficiently smooth.

Key words. Kirchhoff–Love model, mixed formulation, finite element, error estimates

AMS subject classifications. 65N12, 65N15, 65N30

PII. S0036142900379680

1. Introduction.

We are interested in this paper in the analysis, from the theoretical and the numerical point of view, of the Kirchhoff–Love model for a bending thin plate satisfying physical boundary conditions; cf. [4]. The unknowns of the problem are the plate's deflection and the bending moment, which is a second order symmetric tensor. The usual approach for solving the Kirchhoff–Love model consists of eliminating the stress tensor and calculating next the displacement, which satisfies a fourth order elliptic problem. However, due to the boundary conditions considered here, this approach is not possible, and, moreover, our goal is to obtain a good approximation for the stress tensor, which represents in practice the quantity of physical interest.

Therefore, we propose a new mixed formulation whose main unknown is the bending moment, while the multiplier is the gradient of the displacement on the lateral boundary. In order to avoid the discretization of the constraint imposed on the bending moment and on the test-functions, $div\,\mathbf{div} = 0$, the main idea is to decompose them by applying twice the Tartar lemma [9] and also by using the symmetry of the above tensors. A similar idea is used by Destuynder and Salaun [4], but they apply only once Tartar's lemma, so they get a completely different problem.

We thus obtain an equivalent mixed formulation of the problem, whose main unknown is not a physical variable but whose spaces are very easy to approximate. Indeed, we are now dealing with classical Sobolev spaces, $H^1(\Omega)$ and $H^{1/2}(\Gamma)$, respectively. We employ for the approximation conforming low-order finite elements, for which we prove a uniform inf-sup condition with respect to the discretization parameter. So, thanks to the Babuška–Brezzi theory (cf., for instance, [2]), we establish the unconditional convergence of the proposed method. Moreover, the convergence rate is proved to be optimal $O(h)$ whenever the continuous solution is sufficiently smooth. It is then obvious how to obtain an approximate bending tensor, while the deflection of the plate is recovered by solving an additional Laplacian problem.

*Received by the editors October 16, 2000; accepted for publication (in revised form) March 12, 2002; published electronically October 31, 2002.
http://www.siam.org/journals/sinum/40-5/37968.html
†Laboratoire de Mathématiques Appliquées, Université de Pau, 64000 Pau, France (mohamed.amara@univ-pau.fr, daniela.capatina-papaghiuc@univ-pau.fr).
‡LAMSIN, Ecole Nationale d'Ingénieurs, 2002 Tunis, Tunisia (amna_chatti@yahoo.fr).

The same idea of decomposing the bending moment applies to the Reissner–Mindlin thin plate model, with natural boundary conditions. The variational formulations presented here can then be generalized to the Reissner–Mindlin case, and they lead to a discrete conforming method whose convergence is uniform with respect to the plate's thickness, which means that no locking phenomenon occurs.

The paper is organized as follows. In section 2 we introduce the continuous Kirchhoff–Love problem, and in section 3 we state and analyze a first mixed variational formulation with respect to the bending moment. Section 4 deals with an equivalent variational problem, obtained by decomposing the above tensor. Finally, in section 5 we study its discretization by means of classical finite elements; we establish optimal error estimates and deduce the unconditional convergence of the finite element method. We conclude with the approximation of the quantities which interested us at the beginning of the paper, that is, the bending moment and the deflection of the plate.

2. Presentation of the continuous problem. Let $\Omega \subset \mathbb{R}^2$ denote the medium surface of the thin plate and Γ its lateral boundary. We suppose that Γ is decomposed into three disjoint open parts,

$$\Gamma = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2.$$

On each part we impose different boundary conditions. More precisely, we suppose that the plate is clamped on Γ_0, simply supported on Γ_1, while Γ_2 represents the free boundary. In view of the finite element approximation of the problem, we also suppose that Ω is a polygonal and connected bounded domain. The hypothesis of connectivity is not essential but permits an easier presentation of the method.

The mechanical framework considered here is linear elasticity, and, in order to simplify the writing, the constitutive material is taken to be homogeneous and isotropic. Then the equations describing the Kirchhoff–Love model may be written as below (see, for instance, [4] for more details):

$$\left\{\begin{array}{ll} \Delta^2 u = f & \text{in } \Omega, \\ u = \partial_n u = 0 & \text{on } \Gamma_0, \\ u = 0, \quad \sigma_{ij} n_i n_j = 0 & \text{on } \Gamma_1, \\ \sigma_{ij} n_i n_j = \partial_t(\sigma_{ij} t_i n_j) + \partial_j \sigma_{ij} n_i = 0 & \text{on } \Gamma_2, \\ \sigma_{ij} = (1-\nu)\partial_{ij} u + \nu \Delta u \delta_{ij} & \text{in } \Omega, \end{array}\right. \tag{2.1}$$

where ν is the Poisson coefficient. The unknowns of the problem are the deflection of the plate u and the bending moment $\underline{\sigma} = (\sigma_{ij})_{1\le i,j\le 2}$, which is a symmetric second order tensor. A transverse loading is applied, of force density denoted (after scaling) by f. In what follows, we will take $f \in L^2(\Omega)$. We will suppose for technical reasons that Ω has no cuts, Γ_0 is not empty, and, moreover, $\Gamma_0 \cup \Gamma_1$ is connected.

We denote $\mathbf{n} = (n_i)_{1\le i\le 2}$ the unit outward normal vector along Γ and $\mathbf{t} = (t_i)_{1\le i\le 2}$ the unit tangent vector to Γ oriented such that $t_1 = n_2$, $t_2 = -n_1$. We also employ in this paper the summation convention of Einstein, and we denote by the letter c any positive constant independent of the discretization. We agree to write the vectors in bold letters and the tensors in underlined letters.

Let us also recall here some classical notation which will be used in what follows; for any vector function $\mathbf{v}$ and any scalar function v we denote

$$curl\, \mathbf{v} = \partial_1 v_2 - \partial_2 v_1, \qquad div\, \mathbf{v} = \partial_i v_i,$$
$$\mathbf{curl}\, v = \begin{pmatrix} \partial_2 v \\ -\partial_1 v \end{pmatrix}, \qquad \underline{curl}\, \mathbf{v} = \begin{pmatrix} \partial_2 v_1 & -\partial_1 v_1 \\ \partial_2 v_2 & -\partial_1 v_2 \end{pmatrix},$$

and we also put

$$\mathbf{div}\underline{\tau} = \begin{pmatrix} \partial_1\tau_{11} + \partial_2\tau_{12} \\ \partial_1\tau_{21} + \partial_2\tau_{22} \end{pmatrix}, \quad \underline{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \underline{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

We also introduce the tangential, respectively, the normal, derivative on the boundary for a scalar function v:

$$\partial_t v = \nabla v \cdot \mathbf{t} \quad \text{and} \quad \partial_n v = \nabla v \cdot \mathbf{n}.$$

We will denote by $\langle \cdot, \cdot \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}$ and $\langle \cdot, \cdot \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}$ the duality pairings between the spaces $H^{1/2}(\Gamma)$ and its topological dual $H^{-1/2}(\Gamma)$, respectively, $H^{3/2}(\Gamma)$ and its dual $H^{-3/2}(\Gamma)$. Let us also recall that $H^{1/2}(\Gamma)$ and $H^{3/2}(\Gamma)$ are the spaces of traces of functions belonging to $H^1(\Omega)$, respectively, to $H^2(\Omega)$. The Kirchhoff–Love model (2.1) is a rather general one. It includes the case of a clamped plate (obtained for $\Gamma_1 = \Gamma_2 = \emptyset$ and modelled by a classical biharmonic problem), as well as the case of a simply supported plate (obtained when $\Gamma_0 = \Gamma_2 = \emptyset$).

One of the first questions which arise in the analysis of problem (2.1) is how to give a mathematical framework in which the above problem is well-posed. For that, we consider the following spaces:

$$V = \left\{ v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \right\},$$
$$\underline{X} = \left\{ \underline{\tau} = (\tau_{ij})_{1 \le i,j \le 2};\ \tau_{ij} \in L^2(\Omega),\ \ D(\underline{\tau}) \in L^2(\Omega) \right\},$$

endowed with the natural norms

$$\|v\|_V = |v|_{1,\Omega},$$
$$\| \underline{\tau} \|_{\underline{X}} = (\| \underline{\tau} \|_{0,\Omega}^2 + \| D(\underline{\tau}) \|_{0,\Omega}^2)^{1/2},$$

where $D(\underline{\tau}) = div(\mathbf{div}\underline{\tau}) = \partial_{ij}\tau_{ij}$. It is obvious that V and $\underline{X}$ are Hilbert spaces with respect to these norms, and, moreover, we have the following result.

PROPOSITION 2.1. *The space $(\mathcal{D}(\overline{\Omega}))^4$ is a dense subspace of $\underline{X}$.*

Proof. Let $\underline{\sigma} \in \underline{X}$ belong to the orthogonal complement of $(\mathcal{D}(\overline{\Omega}))^4$ in $\underline{X}$; that is,

$$\forall \underline{\tau} \in (\mathcal{D}(\overline{\Omega}))^4, \quad \langle \underline{\sigma}, \underline{\tau} \rangle_{\underline{X}} = \int_\Omega \underline{\sigma} : \underline{\tau}\, d\Omega + \int_\Omega D(\underline{\sigma}) D(\underline{\tau})\, d\Omega = 0.$$

By taking $\underline{\tau} \in (\mathcal{D}(\Omega))^4 \subset (\mathcal{D}(\overline{\Omega}))^4$ one obtains, for all $i, j \in \{1, 2\}$, the following equality in the sense of distributions:

$$\sigma_{ij} + \partial_{ij} D(\underline{\sigma}) = 0 \quad \text{in } \mathcal{D}'(\Omega).$$

This immediately implies that $D(\underline{\sigma}) \in H^2(\Omega)$. By taking next as test-function $\underline{\tau} = \alpha \underline{I}$ with $\alpha \in \mathcal{D}(\overline{\Omega})$, one obtains that

$$D(\underline{\sigma}) = \partial_n D(\underline{\sigma}) = 0 \ \text{ on } \Gamma.$$

So, $D(\underline{\sigma})$ satisfies the following boundary value problem:

$$\begin{cases} D(\underline{\sigma}) \in H_0^2(\Omega), \\ D(\underline{\sigma}) + \Delta^2 D(\underline{\sigma}) = 0 \quad \text{in } \Omega, \end{cases}$$

which admits a unique solution $D(\underline{\sigma}) = 0$. Therefore, we obtain that $\underline{\sigma} = \underline{0}$. We have thus established that $(\mathcal{D}(\overline{\Omega})^4)^\perp = \{\underline{0}\}$ for the scalar product of $\underline{X}$, and the Hahn–Banach theorem completes the proof. ∎

This preliminary result allows us to define rigorously the trace operators for the bending moment $\underline{\sigma}$. We proceed in the usual way; that is, we first consider the operators γ_0 and γ_1 defined by

$$\gamma_0 : (\mathcal{D}(\overline{\Omega}))^4 \longrightarrow L^2(\Gamma), \qquad \gamma_0(\underline{\tau}) = \underline{\tau}\mathbf{n}\cdot\mathbf{n} = \tau_{ij}n_in_j$$

and, respectively,

$$\gamma_1 : (\mathcal{D}(\overline{\Omega}))^4 \longrightarrow L^2(\Gamma), \qquad \gamma_1(\underline{\tau}) = \partial_t(\underline{\tau}\mathbf{n}\cdot\mathbf{t}) + \mathbf{div}\underline{\tau}\cdot\mathbf{n} = \partial_t(\tau_{ij}t_in_j) + \partial_j\tau_{ij}n_i,$$

and we establish the following result.

THEOREM 2.2. *Operators* $\gamma_0 : ((\mathcal{D}(\overline{\Omega}))^4, \|\cdot\|_{\underline{X}}) \longrightarrow (L^2(\Gamma), \|\cdot\|_{-\frac{1}{2},\Gamma})$ *and* $\gamma_1 : ((\mathcal{D}(\overline{\Omega}))^4, \|\cdot\|_{\underline{X}}) \longrightarrow (L^2(\Gamma), \|\cdot\|_{-\frac{3}{2},\Gamma})$ *are linear and continuous. They can be extended by continuity to the whole space* $\underline{X}$*, with values in* $H^{-1/2}(\Gamma)$ *and* $H^{-3/2}(\Gamma)$*, respectively. Moreover, for any* $v \in H^2(\Omega)$ *one has the following Green-type formula:*

$$\int_\Omega D(\underline{\tau})v\, d\Omega = \int_\Omega \tau_{ij}\partial_{ij}v\, d\Omega - \langle\gamma_0(\underline{\tau}), \partial_n v\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle\gamma_1(\underline{\tau}), v\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}. \tag{2.2}$$

Proof. Let us consider any $\underline{\tau} \in (\mathcal{D}(\overline{\Omega}))^4$. Then by classical Green's formula one has, for any $v \in H^2(\Omega)$, that

$$\begin{aligned}\int_\Omega D(\underline{\tau})v\, d\Omega &= -\int_\Omega \mathbf{div}\underline{\tau}\cdot\nabla v\, d\Omega + \int_\Gamma (\mathbf{div}\underline{\tau}\cdot\mathbf{n})v\, d\Gamma \\ &= \int_\Omega \tau_{ij}\partial_{ij}v\, d\Omega + \int_\Gamma (\mathbf{div}\underline{\tau}\cdot\mathbf{n})v\, d\Gamma - \int_\Gamma (\underline{\tau}\mathbf{n})\cdot\nabla v\, d\Gamma.\end{aligned}$$

One can next write

$$\begin{aligned}\int_\Gamma (\underline{\tau}\mathbf{n})\cdot\nabla v\, d\Gamma &= \int_\Gamma (\underline{\tau}\mathbf{n}\cdot\mathbf{n})\partial_n v\, d\Gamma + \int_\Gamma (\underline{\tau}\mathbf{n}\cdot\mathbf{t})\partial_t v\, d\Gamma \\ &= \int_\Gamma (\underline{\tau}\mathbf{n}\cdot\mathbf{n})\partial_n v\, d\Gamma - \int_\Gamma \partial_t(\underline{\tau}\mathbf{n}\cdot\mathbf{t})v\, d\Gamma\end{aligned}$$

and finally obtain, for any $\underline{\tau} \in (\mathcal{D}(\overline{\Omega}))^4$,

$$\int_\Omega D(\underline{\tau})v\, d\Omega = \int_\Omega \tau_{ij}\partial_{ij}v\, d\Omega - \langle\gamma_0(\underline{\tau}), \partial_n v\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle\gamma_1(\underline{\tau}), v\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}.$$

We next establish the continuity of the linear operators γ_0 and γ_1 defined on the space $(\mathcal{D}(\overline{\Omega}))^4$. For that, it is sufficient to write that

$$\begin{aligned}(\|\gamma_0(\underline{\tau})\|^2_{-1/2,\Gamma} + \|\gamma_1(\underline{\tau})\|^2_{-3/2,\Gamma})^{1/2} &\le c \sup_{\substack{\lambda\in H^{1/2}(\Gamma)\\ \mu\in H^{3/2}(\Gamma)}} \frac{\langle\gamma_0(\underline{\tau}),\lambda\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle\gamma_1(\underline{\tau}),\mu\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}}{(\|\lambda\|^2_{1/2,\Gamma} + \|\mu\|^2_{3/2,\Gamma})^{1/2}} \\ &\le c' \sup_{v\in H^2(\Omega)} \frac{\langle\gamma_0(\underline{\tau}),\partial_n v\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} - \langle\gamma_1(\underline{\tau}),v\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}}{\|v\|_{2,\Omega}}.\end{aligned}$$

The last inequality comes by associating, to any $(\lambda,\mu) \in H^{1/2}(\Gamma)\times H^{3/2}(\Gamma)$, a function $v \in H^2(\Omega)$ defined as the unique solution of the biharmonic problem

$$\begin{cases}\Delta^2 v = 0 & \text{in } \Omega,\\ v = -\mu & \text{on } \Gamma,\\ \partial_n v = \lambda & \text{on } \Gamma\end{cases}$$

and by noticing that $\|v\|_{2,\Omega} \leq c(\|\lambda\|^2_{1/2,\Gamma} + \|\mu\|^2_{3/2,\Gamma})^{1/2}$. Then one obviously has

$$\| \gamma_0(\underline{\tau}) \|_{-1/2,\Gamma} + \| \gamma_1(\underline{\tau}) \|_{-3/2,\Gamma} \leq c \sup_{v \in H^2(\Omega)} \frac{\displaystyle\int_\Omega D(\underline{\tau})v \, d\Omega - \int_\Omega \tau_{ij}\partial_{ij}v \, d\Omega}{\|v\|_{2,\Omega}} \leq c \, \| \underline{\tau} \|_{\underline{X}}$$

and by density (cf. Proposition 2.1) it comes that γ_0 and γ_1 can be continuously extended to $\underline{X}$. Green-type formula (2.2) also holds, due to the density of $(\mathcal{D}(\overline{\Omega}))^4$ in $\underline{X}$. ∎

3. Formulation with respect to the bending moment. In this section we propose a first variational formulation of the initial plate problem, whose main unknown will be the bending moment $\underline{\sigma}$. For that, let us begin by introducing the Hilbert spaces

$$\begin{aligned} M &= \left\{ v \in H^{3/2}(\Gamma) ; \ v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \right\}, \\ N &= \left\{ v \in H^{1/2}(\Gamma) ; \ v = 0 \text{ on } \Gamma_0 \right\}, \\ \underline{X}^0 &= \{ \underline{\tau} \in \underline{X}; \ D(\underline{\tau}) = 0 \}, \end{aligned}$$

as well as the subset of $\underline{X}$:

$$\underline{X}^f = \{ \underline{\tau} \in \underline{X}; \ D(\underline{\tau}) = f \}.$$

Let us also define the following continuous bilinear forms:

$$\forall (\underline{\sigma}, \underline{\tau}) \in \underline{X} \times \underline{X}, \qquad a(\underline{\sigma}, \underline{\tau}) = \frac{1}{1-\nu} \int_\Omega \underline{\sigma} : \underline{\tau} \, d\Omega - \frac{\nu}{1-\nu^2} \int_\Omega (tr\underline{\sigma})(tr\underline{\tau}) \, d\Omega;$$

$$\forall \underline{\tau} \in \underline{X}, \ \forall (\mu, \lambda) \in M \times N, \quad b(\underline{\tau}, (\mu, \lambda)) = \langle \gamma_1(\underline{\tau}), \mu \rangle_{-\frac{3}{2}, \frac{3}{2}, \Gamma} - \langle \gamma_0(\underline{\tau}), \lambda \rangle_{-\frac{1}{2}, \frac{1}{2}, \Gamma}.$$

It is also useful to introduce the boundary value problem

$$\left\{ \begin{array}{ll} \Delta\phi = f & \text{in } \Omega, \\ \phi = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ \partial_n \phi = 0 & \text{on } \Gamma_2, \end{array} \right. \tag{3.1}$$

which admits a unique solution $\phi^f \in V$. Thus, the set $\underline{X}^f$ is not empty since $\underline{\tau} = \phi^f \underline{I}$ belongs to $\underline{X}^f$.

Next, we consider the mixed formulation of the Kirchhoff–Love model given by

$$\left\{ \begin{array}{ll} \text{find } \underline{\sigma} \in \underline{X}^f, \ (u_0, u_1) \in M \times N \quad \text{such that,} & \\ \quad \forall \underline{\tau} \in \underline{X}^0, & a(\underline{\sigma}, \underline{\tau}) + b(\underline{\tau}, (u_0, u_1)) = 0; \\ \quad \forall (v_0, v_1) \in M \times N, & b(\underline{\sigma}, (v_0, v_1)) = 0. \end{array} \right.$$

Obviously, one can write the solution $\underline{\sigma} \in \underline{X}^f$ in the following way:

$$\underline{\sigma} = \underline{\sigma}^0 + \phi^f \underline{I},$$

with ϕ^f verifying (3.1) and $\underline{\sigma}^0 \in \underline{X}^0$. By the means of this decomposition, we can work from now on with the new unknown $\underline{\sigma}^0$. Since one has

$$\gamma_0(\phi^f \underline{I}) = \phi^f, \qquad \gamma_1(\phi^f \underline{I}) = 0,$$

the previous variational problem now writes under the equivalent form

$$(3.2)\quad \left\{\begin{array}{ll} \text{find } \underline{\sigma}^0 \in \underline{X}^0,\ (u_0,u_1)\in M\times N \quad \text{such that,} & \\ \forall \underline{\tau}\in \underline{X}^0, & a(\underline{\sigma}^0,\underline{\tau})+b(\underline{\tau},(u_0,u_1)) = -a(\phi^f \underline{I},\underline{\tau});\\ \forall (v_0,v_1)\in M\times N, & b(\underline{\sigma}^0,(v_0,v_1)) = \langle \phi^f, v_1\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}. \end{array}\right.$$

We establish in the next theorem the well-posedness of this problem, as well as its link with the initial Kirchhoff–Love model.

THEOREM 3.1. *Problem* (3.2) *has a unique solution* $(\underline{\sigma}^0,u_0,u_1)$. *Moreover, one has that*

$$\left\{\begin{array}{ll} \underline{\sigma}^0+\phi^f\underline{I}=\underline{\sigma} & \text{in } \Omega\\ u_0=u & \text{on } \Gamma\\ u_1=\partial_n u & \text{on } \Gamma, \end{array}\right.$$

where $(\underline{\sigma},u)$ *satisfies the Kirchhoff–Love equations* (2.1).

Proof. We make use of the classical Babuška–Brezzi theory for mixed formulations in order to obtain the existence and the uniqueness of the solution of (3.2).

First of all, let us note that the bilinear form $a(\cdot,\cdot)$ is elliptic on $\underline{X}^0\times\underline{X}^0$; indeed, the inequality

$$(tr\underline{\tau})^2 \le 2\underline{\tau}:\underline{\tau}$$

implies that,

$$\forall \underline{\tau}\in\underline{X}^0, \qquad a(\underline{\tau},\underline{\tau}) \ge \frac{1}{1+\nu}\parallel \underline{\tau}\parallel^2_{0,\Omega}.$$

Second, we verify the inf-sup condition for the bilinear form $b(\cdot,\cdot)$. For that, let us consider an arbitrary couple $(v_0,v_1)\in M\times N$, to which we associate the unique solution $w\in H^2(\Omega)$ of the following boundary value problem:

$$(3.3)\quad \left\{\begin{array}{ll} \Delta^2 w=0 & \text{in } \Omega,\\ w=-v_0 & \text{on } \Gamma,\\ \partial_n w=-v_1 & \text{on } \Gamma. \end{array}\right.$$

We define next $\tau_{ij}=\partial_{ij}w$ for $1\le i,j\le 2$; then we have that $\underline{\tau}=(\tau_{ij})_{1\le i,j\le 2}\in\underline{X}^0$ and also that $\parallel\underline{\tau}\parallel_{\underline{X}}=|w|_{2,\Omega}$. Green's formula (2.2) gives

$$\langle\gamma_1(\underline{\tau}),v_0\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}-\langle\gamma_0(\underline{\tau}),v_1\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}=\int_\Omega \tau_{ij}\partial_{ij}w\,d\Omega=|w|^2_{2,\Omega},$$

so, using the continuity of the trace operators for w, we get that

$$\sup_{\underline{\tau}\in\underline{X}^0}\frac{\langle\gamma_1(\underline{\tau}),v_0\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}-\langle\gamma_0(\underline{\tau}),v_1\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}}{\parallel\underline{\tau}\parallel_{\underline{X}}}\ge |w|_{2,\Omega}\ge c(\|v_0\|_{\frac{3}{2},\Gamma}+\|v_1\|_{\frac{1}{2},\Gamma}).$$

The hypotheses of the Babuška–Brezzi theorem are now satisfied, so problem (3.2) admits a unique solution.

In order to interpret this variational problem in the sense of distributions, let us begin by taking as test-function in the first equation of (3.2) a tensor $\underline{\tau}\in(\mathcal{D}(\Omega))^4$ with $D(\underline{\tau})=0$. Then we have

$$(3.4)\qquad \int_\Omega \underline{\chi}:\underline{\tau}\,d\Omega=0,$$

where we have put, for the simplicity of the writing,

$$\underline{\chi} = \frac{1}{1-\nu}\left(\underline{\sigma} - \frac{\nu}{1+\nu}(tr\underline{\sigma})\underline{I}\right), \qquad \underline{\sigma} = \underline{\sigma}^0 + \phi^f \underline{I}.$$

By taking next in the relation (3.4) $\underline{\tau} = \rho \underline{J}$ with $\rho \in \mathcal{D}(\Omega)$, we obtain that the tensor $\underline{\chi}$ is symmetric. For the test-function $\underline{\tau} = \underline{curl}\,\varphi$ with $\varphi \in (\mathcal{D}(\Omega))^2$, we get

$$curl \begin{pmatrix} \chi_{11} \\ \chi_{12} \end{pmatrix} = curl \begin{pmatrix} \chi_{21} \\ \chi_{22} \end{pmatrix} = 0,$$

so there exists $\theta \in (H^1(\Omega)/\mathbb{R})^2$ such that $\underline{\chi} = \underline{\nabla}\theta$. The symmetry of $\underline{\chi}$ then implies $\theta = \nabla u$, where $u \in H^2(\Omega)$. Therefore, we finally obtain that $\chi_{ij} = \partial_{ij} u$ or, equivalently,

$$\sigma_{ij} = (1-\nu)\partial_{ij} u + \nu \Delta u \delta_{ij},$$

with $u \in H^2(\Omega)$ unique up to a first order polynomial. The Green-type formula (2.2) and the first equation of (3.2) next give, for all $\underline{\tau} \in \underline{X}^0$, that

$$\langle \gamma_1(\underline{\tau}), u_0 - u\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} - \langle \gamma_0(\underline{\tau}), u_1 - \partial_n u\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0.$$

Consequently, by using the same idea as in the proof of the inf-sup condition for $b(\cdot,\cdot)$, we get that $|w|_{2,\Omega} = 0$, where w now satisfies the following biharmonic problem:

$$\begin{cases} \Delta^2 w = 0 & \text{in } \Omega, \\ w = u_0 - u & \text{on } \Gamma, \\ \partial_n w = u_1 - \partial_n u & \text{on } \Gamma. \end{cases}$$

So w is a first order polynomial, which we can take to be null, and thus we obtain that

$$u_{|\Gamma} = u_0 \in M, \qquad \partial_n u_{|\Gamma} = u_1 \in N.$$

Finally, since $D(\underline{\sigma}) = f$ and since $\underline{\sigma}$ satisfies, from the second equation of (3.2), the boundary conditions $\gamma_1(\underline{\sigma}) = 0$ on Γ_2, $\gamma_0(\underline{\sigma}) = 0$ on $\Gamma_1 \cup \Gamma_2$, we now get that $(\underline{\sigma}, u)$ satisfies the equations of the Kirchhoff–Love model (2.1), which ends the proof. ∎

REMARK 1. *This result allows us to recover the deflection of the plate, too. Indeed, u is the unique solution of the biharmonic problem*

$$\begin{cases} \Delta^2 u = f & in\ \Omega, \\ u = u_0 & on\ \Gamma, \\ \partial_n u = u_1 & on\ \Gamma. \end{cases}$$

However, in what follows we will calculate the displacement u more directly, as the solution of a second order elliptic problem.

4. Equivalent mixed formulation. We introduce in this section another variational formulation for (2.1), which is equivalent to (3.2) and whose unknowns now belong to classical Sobolev spaces. Then we will approximate this new problem, and thus we will avoid the discretization of the constraint $div(\mathbf{div}\underline{\tau}) = 0$ imposed on the test-functions of (3.2). This equivalent formulation is based on the decomposition of the elements of $\underline{X}^0$, which is presented in the next paragraph.

4.1. Characterization of the subspace $\underline{\boldsymbol{X}}^{\mathbf{0}}$. For any $\underline{\tau} \in \underline{X}^0$, one has $D(\underline{\tau}) = 0$ that is $div(\mathbf{div}\underline{\tau}) = 0$. Applying Tartar's lemma (see [5] or [9] for instance), one gets the existence of a unique $\rho \in L_0^2(\Omega)$ such that $\mathbf{div}\,\underline{\tau} = \mathbf{curl}\,\rho$. This translates into the relations

$$\begin{cases} \partial_1\tau_{11} + \partial_2\tau_{12} = \partial_2\rho, \\ \partial_1\tau_{21} + \partial_2\tau_{22} = -\partial_1\rho \end{cases}$$

or, equivalently,

$$\begin{cases} \partial_1\tau_{11} + \partial_2(\tau_{12} - \rho) = 0, \\ \partial_1(\tau_{21} + \rho) + \partial_2\tau_{22} = 0. \end{cases}$$

We have employed above the usual notation

$$L_0^2(\Omega) = \left\{ \rho \in L^2(\Omega); \quad \int_\Omega \rho \, d\Omega = 0 \right\}.$$

By setting $\mathbf{q}_1 = \begin{pmatrix} \tau_{11} \\ \tau_{12} - \rho \end{pmatrix}$ and $\mathbf{q}_2 = \begin{pmatrix} \tau_{21} + \rho \\ \tau_{22} \end{pmatrix}$, the previous relations write $div\,\mathbf{q}_i = 0$ for $1 \le i \le 2$. One more application of the Tartar lemma gives the existence of a unique function $\varphi = (\varphi_1, \varphi_2) \in (H^1(\Omega))^2$, with $\int_\Omega \varphi \, d\Omega = 0$, such that

$$\begin{cases} \tau_{11} = \partial_2\varphi_1, \\ \tau_{12} = \rho - \partial_1\varphi_1, \\ \tau_{21} = -\rho + \partial_2\varphi_2, \\ \tau_{22} = -\partial_1\varphi_2. \end{cases}$$

So, $\underline{\tau}$ belongs to $\underline{X}^0$ if and only if there exist two functions $\varphi \in (H^1(\Omega))^2$ and $\rho \in L^2(\Omega)$, both unique up to a constant, such that

$$\underline{\tau} = \underline{curl}\,\varphi + \rho\underline{J}.$$

It follows that if $\underline{\tau} \in \underline{X}^f$, then there exist unique functions $\widetilde{\varphi} \in (H^1(\Omega))^2$ with $\int_\Omega \widetilde{\varphi} \, d\Omega = 0$ and $\widetilde{\rho} \in L_0^2(\Omega)$ such that

$$\underline{\tau} = \underline{curl}\,\widetilde{\varphi} + \widetilde{\rho}\underline{J} + \phi^f \underline{I},$$

with $\phi^f \in V$ the solution of problem (3.1).

Concerning the trace operators γ_0 and γ_1, we can express them for every $\underline{\tau} \in \underline{X}^0$ in the following way:

$$\begin{cases} \gamma_0(\underline{\tau}) = (\underline{curl}\,\varphi\,\mathbf{n}) \cdot \mathbf{n} = -\partial_t\varphi \cdot \mathbf{n}, \\ \gamma_1(\underline{\tau}) = \partial_t(\underline{\tau}\mathbf{n} \cdot \mathbf{t}) + \mathbf{div}\underline{\tau} \cdot \mathbf{n} = -\partial_t(\partial_t\varphi \cdot \mathbf{t}). \end{cases} \tag{4.1}$$

If, moreover, the tensor $\underline{\tau} \in \underline{X}^0$ is symmetric, then

$$2\rho = div\varphi.$$

In this case we write $\underline{\tau}$ as below:

$$\underline{\tau} = \underline{curl}\,\varphi + \frac{1}{2}(div\varphi)\underline{J} = \begin{pmatrix} \partial_2\varphi_1 & (\partial_2\varphi_2 - \partial_1\varphi_1)/2 \\ (\partial_2\varphi_2 - \partial_1\varphi_1)/2 & -\partial_1\varphi_2 \end{pmatrix}, \tag{4.2}$$

with a unique function φ now belonging to

$$\mathbf{H} = \left\{ \varphi \in (H^1(\Omega))^2; \ \int_\Omega \varphi \, d\Omega = 0, \ \int_\Omega div\varphi \, d\Omega = 0 \right\}.$$

Obviously, $\mathbf{H}$ is a Hilbert space with respect to the norm

$$[\varphi]_{\mathbf{H}} = \left(\|\partial_2\varphi_1\|^2_{0,\Omega} + \frac{1}{2}\|\partial_2\varphi_2 - \partial_1\varphi_1\|^2_{0,\Omega} + \|\partial_1\varphi_2\|^2_{0,\Omega} \right)^{1/2}.$$

By the means of the operator $T : (H^1(\Omega))^2 \to (H^1(\Omega))^2$ defined by $T(\varphi) = (\varphi_2, -\varphi_1)$, one can immediately see that $[\varphi]_{\mathbf{H}} = \| \underline{\varepsilon}(T(\varphi)) \|_{0,\Omega}$ and $div\varphi = curl\, T(\varphi)$. Here, $\underline{\varepsilon}(\mathbf{v})$ is the strain tensor associated to a function $\mathbf{v}$ with $\varepsilon_{ij}(v) = \frac{1}{2}(\partial_i v_j + \partial_j v_i)$ for $1 \le i, j \le 2$, and the L^2-norm of a tensor is the square root of the sum of the squares of the L^2-norms of the tensor elements. So, by Korn's inequality we have that $[\cdot]_{\mathbf{H}}$ and $\|\cdot\|_{1,\Omega}$ are equivalent norms on $\mathbf{H}$.

From now on, since we know that $\|\cdot\|_{1,\Omega}$ and $|\cdot|_{1,\Omega}$ are also equivalent norms on $\mathbf{H}$, we shall consider the space $\mathbf{H}$ endowed with the norm $|\cdot|_{1,\Omega}$.

REMARK 2. *Let us point out here that to any $\varphi \in (H^1(\Omega))^2$ we can associate a function $\widetilde{\varphi} \in \mathbf{H}$ such that*

$$\underline{curl}\, \varphi + \frac{1}{2}(div\varphi)\underline{J} = \underline{curl}\, \widetilde{\varphi} + \frac{1}{2}(div\widetilde{\varphi})\underline{J},$$
$$\partial_t\varphi \cdot \mathbf{n} = \partial_t\widetilde{\varphi} \cdot \mathbf{n}, \quad \partial_t(\partial_t\varphi \cdot \mathbf{t}) = \partial_t(\partial_t\widetilde{\varphi} \cdot \mathbf{t}), \quad [\varphi]_{\mathbf{H}} = [\widetilde{\varphi}]_{\mathbf{H}}.$$

Indeed, let us put $a = \int_\Omega div\varphi \, d\Omega$ and consider the function $\varphi' = \frac{a}{2m(\Omega)}\binom{x}{y} + \mathbf{c}$, with $\mathbf{c} \in \mathbb{R}^2$ chosen such that $\int_\Omega \varphi \, d\Omega = \int_\Omega \varphi' \, d\Omega$. Clearly, one has $\partial_t\varphi' = \frac{a}{2m(\Omega)}\mathbf{t}$ on Γ which leads to $\partial_t\varphi' \cdot \mathbf{n} = 0$, $\partial_t(\partial_t\varphi' \cdot \mathbf{t}) = 0$. Thus, we can take $\widetilde{\varphi} = \varphi - \varphi'$ which satisfies the above conditions.

4.2. New variational formulation of the Kirchhoff–Love model. In what follows, since we know that the bending moment $\underline{\sigma}$ is symmetric, we will use the decomposition (4.2) for the symmetric elements of $\underline{X}^0$ in order to write down and to study a new equivalent formulation of problem (3.2), whose main unknown will now be a function ψ belonging to $\mathbf{H}$.

To do that, let us define a bilinear continuous form $A(\cdot,\cdot)$ on $\mathbf{H} \times \mathbf{H}$ by putting

$$A(\psi,\varphi) = a\left(\underline{curl}\, \psi + \frac{1}{2}(div\psi)\underline{J}, \underline{curl}\, \varphi + \frac{1}{2}(div\varphi)\underline{J} \right)$$
$$= \frac{1}{1-\nu}\int_\Omega \left[\partial_2\psi_1\partial_2\varphi_1 + \partial_1\psi_2\partial_1\varphi_2 + \frac{1}{2}(\partial_2\psi_2 - \partial_1\psi_1)(\partial_2\varphi_2 - \partial_1\varphi_1) \right] d\Omega$$
$$- \frac{\nu}{1-\nu^2}\int_\Omega (\partial_2\psi_1 - \partial_1\psi_2)(\partial_2\varphi_1 - \partial_1\varphi_2) \, d\Omega.$$

Let us also calculate from (4.1), for any $(v_0, v_1) \in M \times N$,

$$b\left(\underline{curl}\, \varphi + \frac{1}{2}(div\varphi)\underline{J}, (v_0, v_1) \right) = -\langle \partial_t(\partial_t\varphi \cdot \mathbf{t}), v_0 \rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} + \langle \partial_t\varphi \cdot \mathbf{n}, v_1 \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}.$$

Considering a lifting $w \in H^2(\Omega)$ which satisfies $w = v_0$ on Γ and $\partial_n w = v_1$ on Γ, one can now write

$$b\left(\underline{curl}\, \varphi + \frac{1}{2}(div\varphi)\underline{J}, (v_0, v_1) \right) = \langle \partial_t\varphi \cdot \mathbf{t}, \partial_t v_0 \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t\varphi \cdot \mathbf{n}, v_1 \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}$$
$$= \langle \partial_t\varphi \cdot \mathbf{t}, \nabla w \cdot \mathbf{t} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t\varphi \cdot \mathbf{n}, \nabla w \cdot \mathbf{n} \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}$$
$$= \langle \partial_t\varphi, \nabla w \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = -\langle \partial_t \nabla w, \varphi \rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}.$$

REMARK 3. *In order to establish in a rigorous manner the previous relations, one may use the continuity of the trace operators* $\gamma_0,\ \gamma_1$ *and the density of* $\mathcal{D}(\overline{\Omega})$ *in* $H^1(\Omega)$.

This leads us to introduce a new bilinear form $B(\cdot,\cdot)$ on $\mathbf{H}\times\mathbf{Z}$ by setting

$$B(\varphi,\mathbf{q}) = -\langle \partial_t \mathbf{q}, \varphi\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma},$$

where

$$\mathbf{Z} = \left\{ \mathbf{q}\in (H^{1/2}(\Gamma))^2;\ \mathbf{q}=0 \text{ on } \Gamma_0,\ \mathbf{q}\cdot\mathbf{t}=0 \text{ on } \Gamma_1,\ \int_\Gamma \mathbf{q}\cdot\mathbf{t}\, d\Gamma = 0 \right\}$$

is endowed with the usual norm $\|\cdot\|_{1/2,\Gamma}$. It is obvious that to any $\mathbf{q}\in\mathbf{Z}$, one can now associate a unique couple $(v_0,v_1)\in M\times N$ by putting

$$\mathbf{q} = (\partial_t v_0)\mathbf{t} + v_1\mathbf{n}.$$

We also introduce the linear continuous forms $F(\cdot)$ and $G(\cdot)$ defined on $\mathbf{H}$, respectively, on $\mathbf{Z}$, by

$$F(\varphi) = -a\left(\phi^f \underline{I}, \underline{curl}\,\varphi + \frac{1}{2}(div\varphi)\underline{J}\right) = -\frac{1}{1+\nu}\int_\Omega \phi^f\,(\partial_2\varphi_1 - \partial_1\varphi_2)\ d\Omega,$$

$$G(\mathbf{q}) = \int_\Gamma \phi^f \mathbf{q}\cdot\mathbf{n}\, d\Gamma.$$

We are now able to write the following mixed variational problem:

$$\left\{\begin{array}{lll} \text{find } \psi\in\mathbf{H},\ \mathbf{p}\in\mathbf{Z} & \text{such that,} & \\ \forall\varphi\in\mathbf{H}, & A(\psi,\varphi)+B(\varphi,\mathbf{p}) & = F(\varphi); \\ \forall\mathbf{q}\in\mathbf{Z}, & B(\psi,\mathbf{q}) & = G(\mathbf{q}), \end{array}\right. \tag{4.3}$$

and we are able to prove in the next result that it is well-posed.

THEOREM 4.1. *Problem* (4.3) *has a unique solution.*

Proof. We use, once more, the Babuška–Brezzi theory for mixed formulations. We begin by establishing that the bilinear form $A(\cdot,\cdot)$ is $\mathbf{H}$-elliptic. Indeed, let φ be an arbitrary element of $\mathbf{H}$. We put $\underline{\tau} = \underline{curl}\,\varphi + \frac{1}{2}(div\varphi)\underline{J}$; then we clearly have $\underline{\tau}\in\underline{X}^0$, and we already know that there exists a constant $c>0$ depending only on ν such that

$$A(\varphi,\varphi) = a(\underline{\tau},\underline{\tau}) \geq c\,\|\,\underline{\tau}\,\|^2_{0,\Omega}\,.$$

On the other hand, one has

$$\|\,\underline{\tau}\,\|_{0,\Omega} = [\varphi]_{\mathbf{H}} \geq c\,|\varphi|_{1,\Omega}\,,$$

so we deduce the ellipticity of $A(\cdot,\cdot)$ on $\mathbf{H}\times\mathbf{H}$.

The continuity of the bilinear form $B(\cdot,\cdot)$ on $\mathbf{H}\times\mathbf{Z}$ may be established by classical means:

$$|\,B(\varphi,\mathbf{q})\,| \leq \|\partial_t\mathbf{q}\|_{-1/2,\Gamma}\cdot\|\varphi\|_{1/2,\Gamma} \leq c\,\|\mathbf{q}\|_{1/2,\Gamma}\cdot|\varphi|_{1,\Omega}\,.$$

As to the inf-sup condition for $B(\cdot,\cdot)$, let us consider for any $\mathbf{q}\in\mathbf{Z}$ the boundary value problem

$$\left\{\begin{array}{ll} \Delta\mathbf{w} = 0 & \text{in } \Omega, \\ \mathbf{w} = \mathbf{q} & \text{on } \Gamma. \end{array}\right.$$

Since $\partial_t \mathbf{q} = -(\underline{curl}\,\mathbf{w})\mathbf{n} \in (H^{-1/2}(\Gamma))^2$, we can write that

$$\sup_{\varphi\in\mathbf{H}} \frac{B(\varphi,\mathbf{q})}{|\varphi|_{1,\Omega}} = \sup_{\varphi\in\mathbf{H}} \frac{\displaystyle\int_\Omega \underline{curl}\,\mathbf{w} : \underline{\nabla}\varphi \, d\Omega}{|\varphi|_{1,\Omega}}.$$

On the other hand, since $\Delta w_i = -curl(\mathbf{curl}\, w_i) = 0$, for $1 \le i \le 2$, one gets that there exists $\mathbf{z} \in (H^1(\Omega))^2$, unique up to a constant, such that $\underline{curl}\,\mathbf{w} = \underline{\nabla}\mathbf{z}$. We can take $\mathbf{z}$ such that $\int_\Omega \mathbf{z}\, d\Omega = 0$. Moreover, we have

$$\int_\Omega div\mathbf{z}\, d\Omega = \int_\Omega (\partial_2 w_1 - \partial_1 w_2)\, d\Omega = \int_\Gamma \mathbf{w}\cdot\mathbf{t}\, d\Gamma = \int_\Gamma \mathbf{q}\cdot\mathbf{t}\, d\Gamma = 0,$$

so $\mathbf{z}$ belongs to $\mathbf{H}$. This implies

$$\sup_{\varphi\in\mathbf{H}} \frac{\displaystyle\int_\Omega \underline{curl}\,\mathbf{w} : \underline{\nabla}\varphi \, d\Omega}{|\varphi|_{1,\Omega}} \ge \frac{\displaystyle\int_\Omega \underline{curl}\,\mathbf{w} : \underline{\nabla}\mathbf{z} \, d\Omega}{|\mathbf{z}|_{1,\Omega}} = |\mathbf{z}|_{1,\Omega} = \| \underline{curl}\,\mathbf{w} \|_{0,\Omega}\,.$$

So, we have now established that

$$\sup_{\varphi\in\mathbf{H}} \frac{B(\varphi,\mathbf{q})}{|\varphi|_{1,\Omega}} \ge \| \underline{curl}\,\mathbf{w} \|_{0,\Omega} = |\mathbf{w}|_{1,\Omega}\,.$$

However, $\mathbf{q} = 0$ on Γ_0 and, since we supposed that $m(\Gamma_0) > 0$, by Poincaré's inequality and by trace theorem we obtain that

$$|\mathbf{w}|_{1,\Omega} \ge c' \|\mathbf{w}\|_{1,\Omega} \ge c \|\mathbf{q}\|_{\frac{1}{2},\Gamma}\,,$$

which allows us to conclude that problem (4.3) admits a unique solution. ∎

4.3. Equivalence with the initial problem. We present here the link between the solution of the previous variational problem (4.3) and the solution of the initial Kirchhoff–Love model. This is stated in the next theorem.

THEOREM 4.2. *Let* $(\psi, \mathbf{p}) \in \mathbf{H}\times\mathbf{Z}$ *be the unique solution of* (4.3). *Then one has*

$$\begin{cases} \underline{\sigma\prime} = \underline{curl}\,\psi + \dfrac{1}{2}(div\psi)\underline{J} + \phi^f \underline{I} & \text{in } \Omega, \\ \nabla u = \mathbf{p} \quad \text{on} \quad \Gamma, \end{cases} \tag{4.4}$$

where $(\underline{\sigma\prime}, u)$ *is the solution of* (2.1) *and* ϕ^f *the solution of* (3.1).

Proof. Let us consider the tensor $\underline{\sigma}'$ associated to the function ψ by means of the relation

$$\underline{\sigma}' = \underline{curl}\,\psi + \frac{1}{2}(div\psi)\underline{J} + \phi^f \underline{I}.$$

Then obviously $D(\underline{\sigma}') = f$ and $\underline{\sigma}'$ is symmetric. Moreover, to any $(v_0, v_1) \in M \times N$ we associate, as in Theorem 3.1, the auxiliary problem (3.3), and we set $\mathbf{q} = \nabla w$. Then one has $\mathbf{q} \in \mathbf{Z}$ and the second equation of (4.3) implies that

$$\langle \partial_t\psi, \nabla w\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = \int_\Gamma \phi^f \partial_n w \, d\Gamma$$

$$\Leftrightarrow \langle \partial_t\psi\cdot\mathbf{t}, \partial_t w\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} + \langle \partial_t\psi\cdot\mathbf{n} - \phi^f, \partial_n w\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0$$

$$\Leftrightarrow \langle \partial_t(\partial_t\psi\cdot\mathbf{t}), v_0\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma} - \langle \partial_t\psi\cdot\mathbf{n} - \phi^f, v_1\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma} = 0,$$

which gives

$$\left\{\begin{array}{ll}\partial_t(\partial_t\psi\cdot\mathbf{t})=0 & \text{in } H^{-3/2}(\Gamma_2),\\ \partial_t\psi\cdot\mathbf{n}-\phi^f=0 & \text{in } H^{-1/2}(\Gamma_1\cup\Gamma_2),\end{array}\right. \Longleftrightarrow \left\{\begin{array}{ll}\gamma_1(\underline{\sigma}')=0 & \text{on } \Gamma_2,\\ \gamma_0(\underline{\sigma}')=0 & \text{on } \Gamma_1\cup\Gamma_2.\end{array}\right.$$

Next, to any $\varphi\in\mathbf{H}$ we associate $\underline{\tau}\in\underline{X}^0$ by putting $\underline{\tau}=\underline{curl}\,\varphi+\frac{1}{2}(div\varphi)\underline{J}$. We take (cf. Remark 2) in the first equation of (4.3) a test-function $\varphi=\varphi^*-\binom{ax+c_1}{ay+c_2}\in\mathbf{H}$ with $\varphi^*\in(\mathcal{D}(\Omega))^2$ arbitrary, and we thus get that

$$a(\underline{\sigma}',\underline{\tau})=0.$$

By introducing the symmetric tensor

$$\underline{\chi}'=\frac{1}{1-\nu}\left(\underline{\sigma}'-\frac{\nu}{1+\nu}(tr\underline{\sigma}')\,\underline{I}\right),$$

we may write this last relation as

$$\int_\Omega \underline{\chi}':\underline{\tau}\,d\Omega=0 \iff \int_\Omega(\chi'_{11}\partial_2\varphi_1^*-\chi'_{12}\partial_1\varphi_1^*)+(\chi'_{12}\partial_2\varphi_2^*-\chi'_{22}\partial_1\varphi_2^*)\,d\Omega=0$$

for every $\varphi^*\in(\mathcal{D}(\Omega))^2$. We deduce, exactly as in the proof of Theorem 3.1, the existence of a function $u'\in H^2(\Omega)$, unique up to a first order polynomial, such that

$$\chi'_{ij}=\partial_{ij}u' \quad \text{for} \quad 1\le i,j\le 2;$$

that is,

$$\underline{\sigma}'=(1-\nu)\begin{pmatrix}\partial_{11}u' & \partial_{12}u'\\ \partial_{21}u' & \partial_{22}u'\end{pmatrix}+\nu(\Delta u')\underline{I}.$$

Now, for $\varphi^*\in(H^1(\Omega))^2$, the Green-type formula (2.2) gives that

$$\langle\gamma_1(\underline{\tau}),u'\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}-\langle\gamma_0(\underline{\tau}),\partial_n u'\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}=-a(\underline{\sigma}',\underline{\tau})=-\langle\partial_t\mathbf{p},\varphi^*\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}.$$

The above relation can be written as below:

$$-\langle\partial_t(\partial_t\varphi^*\cdot\mathbf{t}),u'\rangle_{-\frac{3}{2},\frac{3}{2},\Gamma}+\langle\partial_t\varphi^*\cdot\mathbf{n},\partial_n u'\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}=-\langle\partial_t\mathbf{p},\varphi^*\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}$$

$$\Leftrightarrow\langle\partial_t(\nabla u'-\mathbf{p}),\varphi^*\rangle_{-\frac{1}{2},\frac{1}{2},\Gamma}=0$$

for all $\varphi^*\in(H^1(\Omega))^2$. So we see that $\mathbf{p}=\nabla u'+\mathbf{c}$ on Γ with $\mathbf{c}\in\mathbb{R}^2$, and since $\nabla u'\in(H^1(\Omega)/\mathbb{R})^2$ we can choose $\mathbf{c}=0$. So we obtain, because $\mathbf{p}\in\mathbf{Z}$, that

$$\left\{\begin{array}{ll}\nabla u'=0 & \text{on } \Gamma_0,\\ \partial_t u'=0 & \text{on } \Gamma_1,\end{array}\right. \Longleftrightarrow \left\{\begin{array}{ll}\partial_n u'=0 & \text{on } \Gamma_0,\\ u'=c & \text{on } \Gamma_0\cup\Gamma_1.\end{array}\right.$$

However, now u' belongs to $H^2(\Omega)/\mathbb{R}$, so we can fix the constant c by taking $u'=0$ on $\Gamma_0\cup\Gamma_1$. We have thus proved that $(\underline{\sigma}',u')$ satisfies the Kirchhoff–Love problem (2.1), and the theorem is established. □

As a conclusion, it is sufficient to solve the mixed problem (4.3) and use relation (4.4) to see that $\underline{\sigma}^0=\underline{curl}\,\psi+\frac{1}{2}(div\psi)\underline{J}$ and, implicitly, the tensor $\underline{\sigma}=\underline{\sigma}^0+\phi^f\underline{I}$ is the solution of the initial Kirchhoff–Love model.

REMARK 4. *Let us notice that we can also calculate the displacement u as the unique solution of the following second order elliptic problem:*

$$\begin{cases} \Delta u = \dfrac{1}{1+\nu}(tr\underline{\sigma}) & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \\ \partial_n u = \mathbf{p}\cdot\mathbf{n} & \text{on } \Gamma_2. \end{cases} \tag{4.5}$$

The equation in Ω is obtained by taking the trace in the constitutive law of the plate, and it can also be written in terms of the solution ψ of (4.3) *as* $\Delta u = \frac{1}{1+\nu}(-curl\,\psi + 2\phi^f)$.

So, from now on, we shall study the variational problem (4.3). More precisely, we are interested in its finite element approximation, which will be presented in the next section.

5. Finite element approximation.

5.1. Discrete variational formulation. Let $(\mathcal{T}_h)_{h>0}$ be a regular family of triangulations of the polygonal domain $\overline{\Omega}$, each $\mathcal{T}_h$ consisting of triangles K : $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} K$. For every triangle K of $\mathcal{T}_h$, we denote by h_K its diameter, and we define the discretization parameter $h = \max_{K\in\mathcal{T}_h} h_K$. We also introduce the set of edges of the triangulation $\mathcal{T}_h$ situated on $\Gamma_1 \cup \Gamma_2$,

$$\partial\mathcal{T}_h^1 = \{T;\ T \text{ edge of } K \in \mathcal{T}_h, T \subset (\Gamma_1 \cup \Gamma_2)\},$$

and we agree to denote by $\mathcal{T}_h^*$ the set of triangles $K \in \mathcal{T}_h$ which have at least an edge situated on $\Gamma_1 \cup \Gamma_2$. We consider two finite dimensional spaces $\mathbf{H}_h \subset \mathbf{H}$ and $\mathbf{Z}_h \subset \mathbf{Z}$ which we take as below:

$$\begin{aligned} H_h^1 &= \{\varphi_h \in H^1(\Omega);\ \forall K \in \mathcal{T}_h,\ \varphi_{h|K} \in P_1(K) \text{ if } K \notin \mathcal{T}_h^* \\ &\qquad\qquad \text{and } \varphi_{h|K} \in P_2(K) \text{ if } K \in \mathcal{T}_h^*\}, \\ \mathbf{H}_h &= \mathbf{H} \cap (H_h^1)^2, \\ \mathbf{Z}_h &= \left\{\mathbf{q}_h \in \mathbf{Z};\ \mathbf{q}_h \in (\mathcal{C}^0(\Gamma))^2 \text{ and } \forall T \in \partial\mathcal{T}_h^1,\ \mathbf{q}_{h|T} \in (P_1(T))^2\right\}. \end{aligned}$$

The degrees of freedom of $\varphi_h \in \mathbf{H}_h$ are the values of φ_h at the nodes of the triangulation $\mathcal{T}_h$, to which we add the values at the midpoints of the edges situated on $\Gamma_1 \cup \Gamma_2$ (i.e., the bubble-functions on the boundary $\Gamma_1 \cup \Gamma_2$).

Let us now write down the discrete version of the continuous problem (4.3) as follows:

$$\begin{cases} \text{find } \psi_h \in \mathbf{H}_h,\ \mathbf{p}_h \in \mathbf{Z}_h \quad \text{such that,} \\ \quad \forall \varphi_h \in \mathbf{H}_h, \quad A(\psi_h, \varphi_h) + B(\varphi_h, \mathbf{p}_h) = F_h(\varphi_h); \\ \quad \forall \mathbf{q}_h \in \mathbf{Z}_h, \quad B(\psi_h, \mathbf{q}_h) = G_h(\mathbf{q}_h). \end{cases} \tag{5.1}$$

The linear forms $F(\cdot)$ and $G(\cdot)$ are replaced in the discrete case by

$$\begin{aligned} F_h(\varphi_h) &= -\frac{1}{1+\nu}\int_\Omega \phi_h^f\,(\partial_2\varphi_{1h} - \partial_1\varphi_{2h})\ d\Omega, \\ G_h(\mathbf{q}_h) &= \int_\Gamma \phi_h^f\,\mathbf{q}_h\cdot\mathbf{n}\ d\Gamma. \end{aligned}$$

The discrete function ϕ_h^f is a P_1-continuous finite element approximation of ϕ^f, the solution of the auxiliary problem (3.1). In order to calculate it explicitly, one can

discretize the variational formulation of (3.1) and solve

$$\left\{ \begin{array}{l} \text{find } \phi_h^f \in V_h \quad \text{such that,} \\ \forall v_h \in V_h, \quad \displaystyle\int_\Omega \nabla \phi_h^f \cdot \nabla v_h \, d\Omega = \int_\Omega f v_h \, d\Omega, \end{array} \right. \tag{5.2}$$

where

$$V_h = \left\{ v_h \in V;\ v_{h|K} \in P_1(K)\ \forall K \in \mathcal{T}_h \right\}.$$

It is then obvious that

$$| \phi^f - \phi_h^f |_{1,\Omega} = \inf_{v_h \in V_h} | \phi^f - v_h |_{1,\Omega} .$$

5.2. Error estimates. With the above choice for the finite element spaces, we can show that the variational problem (5.1) has a unique solution. Moreover, the discrete inf-sup condition of Babuška–Brezzi holds uniformly with respect to the discretization parameter h, a result which is stated in the next lemma.

LEMMA 5.1. *There exists a positive constant c independent of h such that,*

$$\forall \mathbf{q}_h \in \mathbf{Z}_h, \quad \sup_{\varphi_h \in \mathbf{H}_h} \frac{B(\varphi_h, \mathbf{q}_h)}{|\varphi_h|_{1,\Omega}} \geq c \parallel \mathbf{q}_h \parallel_{1/2,\Gamma} .$$

Proof. We apply Fortin's trick (see [2], [8]). For that, we will use the continuous inf-sup condition, established in Theorem 4.1, and the interpolation operator $P_h : (H^1(\Omega))^2 \to (H_h^1)^2$ defined hereafter.

Let us denote by P_{1h} the classical Lagrange interpolation operator which satisfies, for any $\varphi \in (H^1(K))^2 \cap (\mathcal{C}^0(\overline{K}))^2$,

$$P_{1h}\varphi \in (\mathcal{P}_1(K))^2 \quad \text{and} \quad P_{1h}\varphi(S) = \varphi(S)$$

for every vertex S of $K \in \mathcal{T}_h$. For the triangles satisfying $K \in \mathcal{T}_h^*$ we also introduce the operator P_{2h} defined by $P_{2h}\varphi \in (\mathcal{P}_2(K))^2$ and

$$\begin{gathered} P_{2h}\varphi(S) = 0 \text{ for every vertex } S \text{ of } K, \\ \int_T (\varphi - P_{2h}\varphi) \, d\Gamma = 0, \end{gathered}$$

where T represents the edge of K situated on $\Gamma_1 \cup \Gamma_2$. Then we put (see also [2]) on every triangle $K \in \mathcal{T}_h$

$$P_h\varphi = \left\{ \begin{array}{ll} P_{1h}\varphi & \text{if} \quad K \notin \mathcal{T}_h^*, \\ P_{1h}\varphi + P_{2h}(\varphi - P_{1h}\varphi) & \text{if} \quad K \in \mathcal{T}_h^*, \end{array} \right.$$

which clearly has the property

$$\forall T \subset \Gamma_1 \cup \Gamma_2, \quad \int_T P_h\varphi \, d\Gamma = \int_T \varphi \, d\Gamma. \tag{5.3}$$

If $\varphi \in \mathbf{H} \cap (\mathcal{C}^0(\overline{\Omega}))^2$, then we have only $P_h\varphi \in (H_h^1)^2$. We construct, according to Remark 2,

$$\widetilde{P_h\varphi} = P_h\varphi - \left(\begin{array}{c} ax + c_1 \\ ay + c_2 \end{array} \right) \in \mathbf{H}_h.$$

Now let us come back to the proof of the uniform inf-sup condition for problem (5.1). To any $\mathbf{q}_h \in \mathbf{Z}_h$, we associate exactly as in Theorem 4.1 a function $\mathbf{z} \in \mathbf{H}$ such that

$$\frac{B(\mathbf{z}, \mathbf{q}_h)}{|\mathbf{z}|_{1,\Omega}} \geq c \, \|\mathbf{q}_h\|_{\frac{1}{2},\Gamma} \geq |\mathbf{z}|_{1,\Omega} \, .$$

We note that $\underline{\nabla}\mathbf{z} = \underline{curl}\, \mathbf{w}$ with $\mathbf{w} \in (H^1(\Omega))^2$ and $\Delta \mathbf{w} = 0$ in Ω and $\mathbf{w} = \mathbf{q}_h$ on Γ. We have $\mathbf{q}_h \in (H^1(\Gamma))^2$; then by classical results of regularity of the Laplace operator (see [6], [7]) we have that $\mathbf{w} \in (H^{1+a}(\Omega))^2$ with $a \in]0, \frac{3}{2}]$. We deduce that $\mathbf{z} \in (H^{1+a}(\Omega))^2 \hookrightarrow (\mathcal{C}^0(\overline{\Omega}))^2$, so we can define $P_h\mathbf{z}$.

Then by considering the discrete function $\mathbf{z}_h = \widetilde{P_h\mathbf{z}} \in \mathbf{H}_h$ and using (5.3) we obtain

$$B(\mathbf{z}, \mathbf{q}_h) = B(P_h\mathbf{z}, \mathbf{q}_h) = B\left(\widetilde{P_h\mathbf{z}}, \mathbf{q}_h\right).$$

The last equality comes from the fact that

$$B\left(\begin{pmatrix} ax + c_1 \\ ay + c_2 \end{pmatrix}, \mathbf{q}_h\right) = a \int_\Gamma \mathbf{q}_h \cdot \mathbf{t} \, d\Gamma = 0.$$

On the other hand, we obtain by passing to the reference finite element and using the Bramble–Hilbert lemma (see [1], [3]), for any K such that $K \in \mathcal{T}_h^*$, that $|\mathbf{z} - P_{2h}\mathbf{z}|_{1,K} \leq c \, |\mathbf{z}|_{1,K}$. This implies

$$\forall K \in \mathcal{T}_h, \quad |\mathbf{z} - P_h\mathbf{z}|_{1,K} \leq c \, |\mathbf{z} - P_{1h}\mathbf{z}|_{1,K} \leq c \, |\mathbf{z}|_{1,K} \, ,$$

so, by the triangle inequality,

$$|P_h\mathbf{z}|_{1,\Omega} \leq c \, |\mathbf{z}|_{1,\Omega} \, .$$

By taking into account the fact that

$$c' \left|\widetilde{P_h\mathbf{z}}\right|_{1,\Omega} \leq \left[\widetilde{P_h\mathbf{z}}\right]_{\mathbf{H}} = [P_h\mathbf{z}]_{\mathbf{H}} \leq c \, |P_h\mathbf{z}|_{1,\Omega} \, ,$$

we now obtain that

$$\sup_{\varphi_h \in \mathbf{H}_h} \frac{B(\varphi_h, \mathbf{q}_h)}{|\varphi_h|_{1,\Omega}} \geq \frac{B\left(\widetilde{P_h\mathbf{z}}, \mathbf{q}_h\right)}{\left|\widetilde{P_h\mathbf{z}}\right|_{1,\Omega}} \geq c \frac{B(\mathbf{z}, \mathbf{q}_h)}{|\mathbf{z}|_{1,\Omega}} \geq c \, \|\mathbf{q}_h\|_{\frac{1}{2},\Gamma} \, ,$$

which completes the proof of the lemma. □

This allows us to get the following error estimate (cf. [2], [8]):

$$\begin{aligned} |\psi - \psi_h|_{1,\Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{\frac{1}{2},\Gamma} \leq c \Bigg\{ & \inf_{\varphi_h \in \mathbf{H}_h} |\psi - \varphi_h|_{1,\Omega} + \inf_{\mathbf{q}_h \in \mathbf{Z}_h} \|\mathbf{p} - \mathbf{q}_h\|_{\frac{1}{2},\Gamma} \\ & + \sup_{\varphi_h \in \mathbf{H}_h} \frac{F_h(\varphi_h) - F(\varphi_h)}{|\varphi_h|_{1,\Omega}} + \sup_{\mathbf{q}_h \in \mathbf{Z}_h} \frac{G_h(\mathbf{q}_h) - G(\mathbf{q}_h)}{\|\mathbf{q}_h\|_{\frac{1}{2},\Gamma}} \Bigg\} , \end{aligned} \tag{5.4}$$

with c independent of the triangulation. It is obvious that

$$\sup_{\varphi_h \in \mathbf{H}_h} \frac{F_h(\varphi_h) - F(\varphi_h)}{|\varphi_h|_{1,\Omega}} \leq c \mid \phi^f - \phi_h^f \mid_{1,\Omega} \leq c \inf_{v_h \in V_h} \mid \phi^f - v_h \mid_{1,\Omega},$$

while the last term can be bounded as below:

$$\sup_{\mathbf{q}_h\in\mathbf{Z}_h}\frac{G_h(\mathbf{q}_h)-G(\mathbf{q}_h)}{\|\mathbf{q}_h\|_{\frac12,\Gamma}}=\sup_{\mathbf{q}_h\in\mathbf{Z}_h}\frac{\displaystyle\int_\Gamma(\phi_h^f-\phi^f)\,\mathbf{q}_h\cdot\mathbf{n}\,d\Gamma}{\|\mathbf{q}_h\|_{\frac12,\Gamma}}\le c\parallel(\phi_h^f-\phi^f)\mathbf{n}\parallel_{-\frac12,\Gamma}$$
$$\le c\parallel\phi_h^f-\phi^f\parallel_{0,\Gamma}\le c\parallel\phi_h^f-\phi^f\parallel_{1,\Omega}\le c\inf_{v_h\in V_h}\mid\phi^f-v_h\mid_{1,\Omega}.$$

So, we have now established the following result.

THEOREM 5.2. *The approximation method* (5.1) *of the variational problem* (4.3) *is unconditionally convergent and satisfies the following error estimate:*

$$|\psi-\psi_h|_{1,\Omega}+\|\mathbf{p}-\mathbf{p}_h\|_{\frac12,\Gamma}$$
$$\le c\left\{\inf_{\varphi_h\in\mathbf{H}_h}|\psi-\varphi_h|_{1,\Omega}+\inf_{\mathbf{q}_h\in\mathbf{Z}_h}\|\mathbf{p}-\mathbf{q}_h\|_{\frac12,\Gamma}+\inf_{v_h\in V_h}\mid\phi^f-v_h\mid_{1,\Omega}\right\},$$

with a constant c independent of the discretization.

5.3. Convergence rate. First of all, let us recall that the function ϕ^f belonging to $H^1(\Omega)$ satisfies the boundary value problem (3.1):

$$\begin{cases}\Delta\phi^f=f & \in L^2(\Omega),\\ \phi^f=0 & \text{on }\Gamma_0\cup\Gamma_1,\\ \partial_n\phi^f=0 & \text{on }\Gamma_2.\end{cases}$$

The regularity results for the Laplace operator ensure (see [6], [7]) that there exists $b\in]\frac12,1]$ such that

$$\phi^f\in H^{1+b}(\Omega),\qquad||\phi^f||_{1+b,\Omega}\le c\parallel f\parallel_{0,\Omega}.$$

Then we get

$$\inf_{v_h\in V_h}\mid\phi^f-v_h\mid_{1,\Omega}\le ch^b\parallel f\parallel_{0,\Omega}.$$

If Ω is convex, we have $b=1$ and then we obtain an optimal behavior of the approximation method given in (5.2). If the domain Ω is not convex but has no cuts, then we have $b\in]\frac12,1[$.

In order to give the convergence rate of the discretization method proposed in (5.1), we assume that the solution $(\underline{\sigma},u)$ of the initial Kirchhoff–Love model (2.1) has the following smoothness:

$$\underline{\sigma}\in(H^a(\Omega))^4,\qquad u\in H^{2+a}(\Omega),\qquad 0<a\le1,\qquad \parallel\underline{\sigma}\parallel_{a,\Omega}+\parallel u\parallel_{2+a,\Omega}\le c\parallel f\parallel_{0,\Omega}. \tag{5.5}$$

Then we immediately obtain from Theorem 5.2 and standard interpolation results that

$$|\psi-\psi_h|_{1,\Omega}+\|\mathbf{p}-\mathbf{p}_h\|_{\frac12,\Gamma}\le ch^{\min\{a,b\}}\parallel f\parallel_{0,\Omega}. \tag{5.6}$$

5.4. Approximate bending tensor and displacement. It is now easy to come back to the calculus of the quantities which interested us at the beginning of this paper, that is, the bending moment $\underline{\sigma}$ and the deflection of the plate u. Concerning the tensor $\underline{\sigma}$, we already know from Theorem 4.2 that

$$\underline{\sigma} = \underline{curl}\,\psi + \frac{1}{2}(div\psi)\underline{J} + \phi^f \underline{I},$$

where ψ satisfies the equations (4.3). Therefore we set

$$\underline{\sigma}_h = \underline{curl}\,\psi_h + \frac{1}{2}(div\psi_h)\underline{J} + \phi_h^f \underline{I}, \tag{5.7}$$

with ψ_h the solution of (5.1) and with ϕ_h^f given by (5.2). Then it is obvious that

$$\| \underline{\sigma} - \underline{\sigma}_h \|_{0,\Omega} \leq c\,|\psi - \psi_h|_{1,\Omega} + || \phi^f - \phi_h^f ||_{0,\Omega} \leq ch^{\min\{a,b\}} \| f \|_{0,\Omega}\ . \tag{5.8}$$

One equally obtains the following estimate, with respect to the norm of $H^{-1}(\Omega)$:

$$\| D(\underline{\sigma}) - D(\underline{\sigma}_h) \|_{-1,\Omega} = \| \Delta(\phi^f - \phi_h^f) \|_{-1,\Omega} \leq | \phi^f - \phi_h^f |_{1,\Omega} \leq ch^b \| f \|_{0,\Omega}\ .$$

Now, concerning the plate's deflection, we know that the continuous solution u satisfies the boundary value problem (4.5), which may be written in variational form

$$\begin{cases} \text{find } u \in V \quad \text{such that,} \\ \forall v \in V, \quad \displaystyle\int_\Omega \nabla u \cdot \nabla v\, d\Omega = \frac{-1}{1+\nu} \int_\Omega (tr\underline{\sigma}) v\, d\Omega + \int_{\Gamma_2} \mathbf{p} \cdot \mathbf{n}\, v\, d\Gamma. \end{cases}$$

Then we calculate the approximation u_h of u as the solution of the next discrete problem:

$$\begin{cases} \text{find } u_h \in V_h \quad \text{such that,} \\ \forall v_h \in V_h, \quad \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h\, d\Omega = \frac{-1}{1+\nu} \int_\Omega (tr\underline{\sigma}_h) v_h\, d\Omega + \int_{\Gamma_2} \mathbf{p}_h \cdot \mathbf{n}\, v_h\, d\Gamma, \end{cases} \tag{5.9}$$

where $\underline{\sigma}_h$ is of course given by (5.7) and $\mathbf{p}_h$ by the problem (5.1). We obtain the same matrix in the relations (5.9) and (5.2), so we have to compute it only once. This leads us to the following error bound:

$$\begin{aligned} |u - u_h|_{1,\Omega} &\leq c \left\{ \inf_{v_h \in V_h} |u - v_h|_{1,\Omega} + \| \underline{\sigma} - \underline{\sigma}_h \|_{0,\Omega} + \| (\mathbf{p} - \mathbf{p}_h) \cdot \mathbf{n} \|_{-\frac{1}{2},\Gamma_2} \right\} \\ &\leq c \left\{ \inf_{v_h \in V_h} |u - v_h|_{1,\Omega} + \| \underline{\sigma} - \underline{\sigma}_h \|_{0,\Omega} + \| \mathbf{p} - \mathbf{p}_h \|_{\frac{1}{2},\Gamma_2} \right\}. \end{aligned} \tag{5.10}$$

Thanks to Theorem 5.2 and to estimates (5.8) and (5.10), we immediately obtain the convergence rate of the approximation of the bending moment $\underline{\sigma}$, as well as of the deflection u. As a conclusion, we finally state the following result.

THEOREM 5.3. *Under the previous smoothness hypotheses on the solution* $(\underline{\sigma}, u)$*, one has*

$$\| \underline{\sigma} - \underline{\sigma}_h \|_{0,\Omega} + \| u - u_h \|_{1,\Omega} + \| D(\underline{\sigma}) - D(\underline{\sigma}_h) \|_{-1,\Omega} \leq ch^{\min\{a,b\}} \| f \|_{0,\Omega},$$

where the constant c *is independent of the discretization.*

So, the approximation method of the Kirchhoff–Love model described in this paper is unconditionally convergent and is optimal whenever the solution $(\underline{\sigma}, u)$ of (2.1), as well as the solution ϕ^f of the Laplacian (3.1), are sufficiently smooth. More precisely, if $\underline{\sigma} \in (H^1(\Omega))^4$ and $\phi^f \in H^2(\Omega)$ (which is the case, for instance, when Ω is convex since $a = b = 1$), then the convergence rate is $O(h)$.

REFERENCES

[1] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[2] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[3] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[4] P. Destuynder and M. Salaun, *Mathematical Analysis of Thin Plate Models*, Springer-Verlag, Berlin, 1996.
[5] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[6] P. Grisvard, *Elliptic Problems in Non-Smooth Domains*, Pitman, Boston, 1985.
[7] V.A. Kondratiev and O.A. Oleinik, *Boundary value problems for partial differential equations in nonsmooth domains*, Russian Math. Surveys, 38 (1983), pp. 1–86.
[8] J.E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II: Finite Element Methods, P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 523–639.
[9] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, 1979.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1650–1665

CONSTRUCTING RANDOMLY SHIFTED LATTICE RULES IN WEIGHTED SOBOLEV SPACES*

I. H. SLOAN†, F. Y. KUO‡, AND S. JOE‡

Abstract. Shifted rank-1 lattice rules, a special class of quasi-Monte Carlo methods, have recently been proposed by the present authors for the integration of functions belonging to certain "weighted" Sobolev spaces. The shifts in these rules were generated in a deterministic manner. In contrast, in this paper we generate these shifts randomly. This allows probabilistic estimates for the error in a given integral. It also reduces the number of operations required to find the generating vectors for the underlying lattice rules component-by-component. The rules thus constructed achieve a worst-case strong tractability error bound in an average or probabilistic sense.

Key words. randomized quasi-Monte Carlo methods, shifted lattice rules, worst-case error, tractability

AMS subject classifications. 65D30, 65D32, 68Q25

PII. S0036142901393942

1. Introduction.

In the recent paper [8] by the present authors, quasi-Monte Carlo (QMC) rules of the form

$$R_{n,d}(f,\boldsymbol{\Delta}) = \frac{1}{n}\sum_{i=1}^{n} f\left(\left\{\frac{i\boldsymbol{z}}{n}+\boldsymbol{\Delta}\right\}\right) \tag{1.1}$$

were constructed for the approximation of integrals over the d-dimensional unit cube. Here n is assumed to be prime, and $\boldsymbol{z}\in\{1,2,\ldots,n-1\}^d$ is an integer vector, with the braces around a vector indicating that we take the fractional part of each component of the vector. In the given construction the components of $\boldsymbol{z}$ were constructed one component at a time, as were the components of $\boldsymbol{\Delta}$, the "shift". Rules of the form (1.1) are called "shifted rank-1 lattice rules", the rank-1 lattice rules being rules of the form

$$R_{n,d}(f) := R_{n,d}(f,\mathbf{0}) = \frac{1}{n}\sum_{i=1}^{n} f\left(\left\{\frac{i\boldsymbol{z}}{n}\right\}\right). \tag{1.2}$$

Further information about lattice rules may be found in [7]. The shifted rank-1 lattice rules constructed in [8] had the advantage that they achieved the worst-case strong tractability error bounds in certain weighted Sobolev spaces.

The construction of the rules in [8] for fixed n and all dimensions up to d required $O(n^3d^2)$ operations, making the construction computationally expensive for large n. Moreover, like other QMC rules, the application of those rules did not yield a practical error estimate.

*Received by the editors August 17, 2001; accepted for publication (in revised form) April 1, 2002; published electronically October 31, 2002. This research was supported by the Australian Research Council.

http://www.siam.org/journals/sinum/40-5/39394.html

†School of Mathematics, The University of New South Wales, Sydney NSW 2052, Australia (sloan@maths.unsw.edu.au).

‡Department of Mathematics, The University of Waikato, Private Bag 3105, Hamilton, New Zealand (fkuo@math.waikato.ac.nz, stephenj@math.waikato.ac.nz).

In [8], the components of the shift $\boldsymbol{\Delta}$ were taken from the set $\{1/(2n), 3/(2n), \ldots, (2n-1)/(2n)\}$, with the successive components of the shift being obtained in a deterministic manner through minimizing a certain functional over the finite set. Here we propose the alternative of allowing the components of $\boldsymbol{\Delta}$ to be random numbers in $[0,1]$, opening the possibility of repeating the calculation with a number of independent shifts so as to allow error estimation. The underlying idea is not new. For example, as early as 1976 Cranley and Patterson [1] pointed out the benefits of using random shifts with rank-1 lattice rules; later this idea was generalized to other lattice rules by Joe [5]. Later still, Owen in [6] introduced a similar randomization idea ("scrambled (t,m,s)-nets") into nets. By now the idea of combining randomization (or "Monte Carlo") ideas with deterministic QMC ideas is commonplace. The key underlying concepts behind such randomized QMC methods are discussed in [3]. The novel element of the present paper is that the lattice rules that we construct here, one component at a time, are specifically designed for use with random shifts, and are in a certain sense optimal for this purpose. An important advantage is that, as we shall see, the cost of constructing rules with n points and all dimensions up to d is reduced to $O(n^2d^2)$, allowing calculations with much larger values of n.

Let q be a positive integer. In this paper we approximate the integral

$$I_d(f) = \int_{[0,1]^d} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{1.3}$$

where f is at least continuous, by

$$\bar{R}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q) = \frac{1}{q}\sum_{m=1}^{q} R_{n,d}(f, \boldsymbol{\Delta}_m) = \frac{1}{qn}\sum_{m=1}^{q}\sum_{i=1}^{n} f\left(\left\{\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}_m\right\}\right), \tag{1.4}$$

where $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q$ are q independent random shifts drawn from a uniform distribution on $[0,1]^d$. In section 2 we will show that $\bar{R}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$ is an unbiased estimator of $I_d(f)$.

In this work the quality of the approximation (1.4) to the integral (1.3) is measured by the root-mean-square value of the worst-case error for f in the unit ball for a particular Hilbert space H_d, that is, by the square root of

$$E\left[D^2(\bar{R}_{n,d})\right] := \int_{[0,1]^{qd}} D^2\left(\bar{R}_{n,d}(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)\right)\, \mathrm{d}\boldsymbol{\Delta}_1 \cdots \mathrm{d}\boldsymbol{\Delta}_q,$$

where

$$\begin{aligned} &D\left(\bar{R}_{n,d}(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)\right) \\ &:= \sup\left\{|\bar{R}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q) - I_d(f)| : \|f\|_{H_d} \le 1, f \in H_d\right\}. \end{aligned} \tag{1.5}$$

The Hilbert space H_d in which our functions f lie is taken here, as in [8], to be a slight generalization of the weighted Sobolev spaces introduced by Sloan and Woźniakowski in [9]. In that work the successive coordinate directions were associated with a nonincreasing sequence of "weights" $\gamma_1, \gamma_2, \ldots$ to express the common reality that successive coordinate directions become less important. The integration problem was found to be "strongly QMC tractable" if and only if

$$\sum_{j=1}^{\infty} \gamma_j < \infty.$$

(The integration problem is said to be strongly QMC tractable if the minimal number of function evaluations n in a QMC rule

$$Q_{n,d}(f) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{t}_i) \tag{1.6}$$

needed to reduce the initial error $I_d(f)$ by a factor of $\varepsilon > 0$ is bounded by a polynomial in ε^{-1} independently of d.) In the present generalization, there are two additional sequences $\{\beta_j\}$ and $\{a_j\}$ of positive real numbers, and H_d is a tensor product of 1-dimensional Hilbert spaces,

$$H_d = H_1^{(1)} \otimes H_1^{(2)} \otimes \cdots \otimes H_1^{(d)}$$

with the inner product of 1-dimensional functions f and g in the jth 1-dimensional space $H_1^{(j)}$ being

$$\beta_j^{-1} f(a_j) g(a_j) + \gamma_j^{-1} \int_0^1 f'(x) g'(x)\,\mathrm{d}x;$$

the function spaces of [9] are recovered by setting $a_j = 1$ and $\beta_j = 1$. The necessary and sufficient condition for strong tractability now becomes (see [8])

$$\sum_{j=1}^{\infty} \frac{\gamma_j}{\beta_j} < \infty. \tag{1.7}$$

Under this condition the precise result of [9, Lemma 8], appropriately generalized, is that for each $n \ge 1$ and $d \ge 1$ there exist QMC rules of the form (1.6) such that

$$D(Q_{n,d}) \le \frac{C_d}{n^{1/2}}, \tag{1.8}$$

where $D(Q_{n,d})$ is defined analogously to (1.5), and

$$C_d := \left(\prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \tfrac{1}{2}\right)\right] - \prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \tfrac{1}{3}\right)\right] \right)^{\frac{1}{2}}$$

which is bounded independently of d by

$$C_\infty := \left(\prod_{j=1}^{\infty} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \tfrac{1}{2}\right)\right] - \prod_{j=1}^{\infty} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \tfrac{1}{3}\right)\right] \right)^{\frac{1}{2}},$$

a finite number if and only if (1.7) holds. In [10] the stronger result was established that (for prime n) there even exist shifted lattice rules (i.e., QMC rules of the special form (1.1)) that satisfy (1.8). However, the results in both [9] and [10] were not constructive.

A crucial aspect of the arguments in this paper, as in [8], [9] and [10], is that H_d is a reproducing kernel Hilbert space with a simple kernel. In section 3 we study the worst-case error $D(Q_{n,d})$ in a reproducing kernel Hilbert space, specialize to our particular Hilbert space H_d, and discuss tractability. In section 4 we explain the

component-by-component construction of a generating vector $\boldsymbol{z}$ such that at every step the strong tractability bound (1.8) is preserved. The final algorithm (Algorithm 4.2) can be written in a few lines, and for a given (prime) n yields all components of $\boldsymbol{z}$ up to the dth in a time of at most $O(n^2d^2)$, and gives $\sqrt{E\left[D^2(\bar{R}_{n,d})\right]}$ satisfying (1.8). Section 5 gives the results of some searches, and some comparisons of computed worst-case errors with the bound (1.8), for n up to approximately 32000 and d up to 100.

2. An unbiased estimator. First we show that the approximation $\bar{R}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$ is an unbiased estimator of $I_d(f)$. Since this result is true not only for rank-1 lattice rules but also for all other integration rules, we state the result in its general form.

Let $Q_{n,d}(f)$ be an n-point general quadrature rule with $n \geq 1$,

$$Q_{n,d}(f) := \sum_{i=1}^{n} w_i f(\boldsymbol{t}_i), \tag{2.1}$$

where $\boldsymbol{t}_i \in [0,1]^d$, $w_i \in \mathbb{R}$, and $\sum_{i=1}^{n} w_i = 1$. For $\boldsymbol{\Delta} \in [0,1]^d$, let $Q_{n,d}(f, \boldsymbol{\Delta})$ denote the $\boldsymbol{\Delta}$-shifted rule by

$$Q_{n,d}(f, \boldsymbol{\Delta}) := \sum_{i=1}^{n} w_i f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}).$$

For q a positive integer and $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q \in [0,1]^d$, let $\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$ denote the approximation obtained by taking an average over q random shifts; that is,

$$\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q) := \frac{1}{q} \sum_{m=1}^{q} Q_{n,d}(f, \boldsymbol{\Delta}_m) = \frac{1}{q} \sum_{m=1}^{q} \sum_{i=1}^{n} w_i f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_m\}),$$

where $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q \in [0,1]^d$ are independent random vectors having a uniform distribution on $[0,1]^d$.

The proof that (1.4) is an unbiased estimator of the integral rests on the following easily established property.

LEMMA 2.1. *Let $f \in L_1([0,1]^d)$. For all $\boldsymbol{t} \in \mathbb{R}^d$,*

$$\int_{[0,1]^d} f(\{\boldsymbol{t} + \boldsymbol{x}\}) \, \mathrm{d}\boldsymbol{x} = \int_{[0,1]^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = I_d(f).$$

The theorem below then follows easily.

THEOREM 2.2. *The family of shifted rules $Q_{n,d}(f, \boldsymbol{\Delta})$ is an unbiased estimator of the integral $I_d(f)$, in the sense that*

$$E\left[Q_{n,d}(f, \cdot)\right] := \int_{[0,1]^d} Q_{n,d}(f, \boldsymbol{\Delta}) \, \mathrm{d}\boldsymbol{\Delta} = I_d(f).$$

Proof. We have

$$\begin{aligned} E\left[Q_{n,d}(f,\cdot)\right] &= \int_{[0,1]^d} \sum_{i=1}^{n} w_i f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} \\ &= \sum_{i=1}^{n} w_i \int_{[0,1]^d} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} \\ &= \sum_{i=1}^{n} w_i I_d(f) = I_d(f), \end{aligned}$$

which completes the proof. □

COROLLARY 2.3 (cf. Corollary 3 of [5]). *The mean* $\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$ *is an unbiased estimate of* $I_d(f)$ *and has variance*

$$\sigma^2 = \frac{1}{q} E\left[\left(Q_{n,d}(f,\cdot) - I_d(f)\right)^2\right] = \frac{1}{q} \int_{[0,1]^d} \left(Q_{n,d}(f,\boldsymbol{\Delta}) - I_d(f)\right)^2 \mathrm{d}\boldsymbol{\Delta}.$$

Remark 1. Since the theorem and the corollary hold for any general quadrature rule $Q_{n,d}$, then it certainly holds for rank-1 lattice rules, thus the shifted rank-1 lattice rules given in (1.1) and (1.4) are unbiased estimates of the integral $I_d(f)$.

Remark 2. It is well known that an unbiased estimate of σ, the standard error of the mean $\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$, is

$$\tilde{\sigma} = \left(\frac{1}{q(q-1)} \sum_{m=1}^{q} \left(Q_{n,d}(f,\boldsymbol{\Delta}_m) - \bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)\right)^2 \right)^{1/2}.$$

By using the well-known Chebyshev inequality,

$$\text{probability}\left(\left|\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q) - I_d(f)\right| < k\sigma\right) \ge 1 - \frac{1}{k^2},$$

this estimate of σ allows us to calculate confidence intervals for the error, that is, an interval in which the true error must lie with a fixed probability.

3. Hilbert spaces and tractability. We recall that a d-dimensional reproducing kernel Hilbert space consists of a Hilbert space H_d along with a reproducing kernel $K(\boldsymbol{x}, \boldsymbol{y})$ which has the property that $f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x})\rangle_{H_d}$ for $f \in H_d$ and $\boldsymbol{x} \in [0,1]^d$, where $\langle \cdot, \cdot \rangle_{H_d}$ is the inner product in H_d. The corresponding norm in H_d is $\|f\|_{H_d} = \langle f, f\rangle_{H_d}^{1/2}$.

Let $D(Q_{n,d}, K)$ denote the corresponding "discrepancy" for the rule $Q_{n,d}$ given by (2.1), which we recall is the same as the worst-case error for all f in the unit ball in H_d. Thus,

$$D(Q_{n,d}, K) = \sup\{|Q_{n,d}(f) - I_d(f)| : \|f\|_{H_d} \le 1,\ f \in H_d\}.$$

One may show (for example, see [9]) that

$$\begin{aligned} D^2(Q_{n,d}, K) = &\int_{[0,1]^{2d}} K(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - 2\sum_{i=1}^{n} w_i \int_{[0,1]^d} K(\boldsymbol{t}_i, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k K(\boldsymbol{t}_i, \boldsymbol{t}_k). \end{aligned} \tag{3.1}$$

Since the approximation we are interested in is $\bar{Q}_{n,d}(f, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_q)$, we shall follow the usual analysis of QMC methods and consider the mean square discrepancy given by

$$E\left[D^2(\bar{Q}_{n,d}, K)\right] = \int_{[0,1]^{qd}} D^2\left(\frac{1}{q}\sum_{m=1}^{q}\sum_{i=1}^{n} w_i f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_m\}), K\right) \mathrm{d}\boldsymbol{\Delta}_1 \cdots \mathrm{d}\boldsymbol{\Delta}_q.$$

THEOREM 3.1. *The mean square discrepancy is given by*

$$E\left[D^2(\bar{Q}_{n,d}, K)\right] = \frac{1}{q}\left[\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} - \int_{[0,1]^{2d}} K(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\right].$$

Proof. It follows from (3.1) that

$$\begin{aligned} D^2(\bar{Q}_{n,d}, K) = {} & \int_{[0,1]^{2d}} K(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - \frac{2}{q}\sum_{m=1}^{q}\sum_{i=1}^{n} w_i \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_m\}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ & + \frac{1}{q^2}\sum_{m=1}^{q}\sum_{\ell=1}^{q}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_m\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}_\ell\}). \end{aligned} \tag{3.2}$$

It is clear that the expectation of the first term on the right-hand side of this expression is just itself and that the expectation value of the second term is

$$\begin{aligned} -2\sum_{i=1}^{n} w_i \int_{[0,1]^{2d}} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{y} &= -2\sum_{i=1}^{n} w_i \int_{[0,1]^{2d}} K(\boldsymbol{\Delta}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{y} \\ &= -2\int_{[0,1]^{2d}} K(\boldsymbol{\Delta}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{y}, \end{aligned} \tag{3.3}$$

where we have made use of Lemma 2.1 and $\sum_{i=1}^{n} w_i = 1$. For the third term, we note that the $m \neq \ell$ case is different from the $m = \ell$ case. It is then not too hard to see that the expectation of the third term is given by

$$\begin{aligned} & \frac{q^2 - q}{q^2}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^{2d}} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}'\})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}' \\ & \qquad + \frac{q}{q^2}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} \\ & = \frac{q-1}{q}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^{2d}} K(\boldsymbol{\Delta}, \boldsymbol{\Delta}')\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}' \\ & \qquad + \frac{1}{q}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} \\ & = \frac{q-1}{q}\int_{[0,1]^{2d}} K(\boldsymbol{\Delta}, \boldsymbol{\Delta}')\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}' \\ & \qquad + \frac{1}{q}\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta}, \end{aligned} \tag{3.4}$$

where again we have made use of Lemma 2.1 and $\sum_{i=1}^n w_i = 1$. The result now follows from (3.2), (3.3), and (3.4). □

To relate the result given in Theorem 3.1 to past results, we recall that associated with any reproducing kernel $K(\boldsymbol{x}, \boldsymbol{y})$ is the "shift-invariant" kernel $K^*(\boldsymbol{x}, \boldsymbol{y})$ defined by

$$K^*(\boldsymbol{x}, \boldsymbol{y}) := \int_{[0,1]^d} K(\{\boldsymbol{x} + \boldsymbol{\Delta}\}, \{\boldsymbol{y} + \boldsymbol{\Delta}\}) \, \mathrm{d}\boldsymbol{\Delta}. \tag{3.5}$$

By shift-invariant, we mean that for arbitrary $\boldsymbol{\Delta} \in [0,1]^d$

$$K^*(\boldsymbol{x}, \boldsymbol{y}) = K^*(\{\boldsymbol{x} + \boldsymbol{\Delta}\}, \{\boldsymbol{y} + \boldsymbol{\Delta}\}),$$

so by taking $\boldsymbol{\Delta} = -\boldsymbol{y}$ we can write

$$K^*(\boldsymbol{x}, \boldsymbol{y}) = K^*(\{\boldsymbol{x} - \boldsymbol{y}\}, \boldsymbol{0}), \qquad \boldsymbol{x}, \boldsymbol{y} \in [0,1]^d. \tag{3.6}$$

We then have from Hickernell and Woźniakowski [4] that

$$E\left[D^2(Q_{n,d}, K)\right] = D^2(Q_{n,d}, K^*);$$

that is, the expected value of the squared discrepancy for the shifted rule in the Hilbert space with reproducing kernel K is exactly the same as the squared discrepancy for the original unshifted rule in the Hilbert space with reproducing kernel K^*. Thus we would expect the expression for $E\left[D^2(\bar{Q}_{n,d}, K)\right]$ given in Theorem 3.1 to match this result in the case when $q = 1$. In fact, we have the following result.

THEOREM 3.2.

$$E\left[D^2(\bar{Q}_{n,d}, K)\right] = \frac{1}{q} D^2(Q_{n,d}, K^*),$$

with

$$D^2(Q_{n,d}, K^*) = \sum_{i=1}^n \sum_{k=1}^n w_i w_k K^*(\{\boldsymbol{t}_i - \boldsymbol{t}_k\}, \boldsymbol{0}) - \int_{[0,1]^d} K^*(\boldsymbol{x}, \boldsymbol{0}) \, \mathrm{d}\boldsymbol{x}. \tag{3.7}$$

Proof. The expression for $D^2(Q_{n,d}, K^*)$ in (3.7) (given in section 4 of [4]) follows easily from (3.1) by using (3.6) and Lemma 2.1.

We see from Theorem 3.1 that $E\left[D^2(\bar{Q}_{n,d}, K)\right]$ depends on the expressions

$$\sum_{i=1}^n \sum_{k=1}^n w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\}) \, \mathrm{d}\boldsymbol{\Delta} \quad \text{and} \quad \int_{[0,1]^{2d}} K(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}.$$

By making use of (3.5) and then (3.6), we see that

$$\begin{aligned} &\sum_{i=1}^n \sum_{k=1}^n w_i w_k \int_{[0,1]^d} K(\{\boldsymbol{t}_i + \boldsymbol{\Delta}\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}\}) \, \mathrm{d}\boldsymbol{\Delta} \\ &= \sum_{i=1}^n \sum_{k=1}^n w_i w_k K^*(\boldsymbol{t}_i, \boldsymbol{t}_k) = \sum_{i=1}^n \sum_{k=1}^n w_i w_k K^*(\{\boldsymbol{t}_i - \boldsymbol{t}_k\}, \boldsymbol{0}). \end{aligned} \tag{3.8}$$

Application of Lemma 2.1 shows that

$$\int_{[0,1]^{2d}} K(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} = \int_{[0,1]^{2d}} K(\{\boldsymbol{x} + \boldsymbol{y}\}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}.$$

Upon changing the order of integration in this last integral as well as replacing $\boldsymbol{y}$ by $\boldsymbol{\Delta}$, we obtain

$$\begin{aligned}\int_{[0,1]^{2d}} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} &= \int_{[0,1]^{2d}} K(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\boldsymbol{\Delta})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{x}\\ &= \int_{[0,1]^{2d}} K(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\{\boldsymbol{0}+\boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{x} = \int_{[0,1]^d} K^*(\boldsymbol{x},\boldsymbol{0})\,\mathrm{d}\boldsymbol{x},\end{aligned}$$

where the final step follows from (3.5). The result then follows from Theorem 3.1, together with this expression and (3.8). ∎

We then see that $E\left[D^2(\bar{Q}_{n,d},K)\right]$ may be calculated by making use of (3.7). In the case when the quadrature points of $Q_{n,d}$ are the points of a rank-1 lattice rule $R_{n,d}$, namely, the points $\{i\boldsymbol{z}/n\}$, $i=1,\ldots,n$ (cf. (1.2)), the double sum in (3.7) may be reduced to a single sum.

COROLLARY 3.3. *With $\boldsymbol{t}_i=\{i\boldsymbol{z}/n\}$, we have*

$$D^2(R_{n,d},K^*) = \frac{1}{n}\sum_{i=1}^{n} K^*(\boldsymbol{t}_i,\boldsymbol{0}) - \int_{[0,1]^d} K^*(\boldsymbol{x},\boldsymbol{0})\,\mathrm{d}\boldsymbol{x}.$$

Proof. In this case $w_i=1/n$ and by the properties of a rank-1 lattice rule, we have

$$\{\{\boldsymbol{t}_i-\boldsymbol{t}_k\}:1\le i,k\le n\} = \{\boldsymbol{t}_i:1\le i\le n\},$$

from which the result follows. ∎

So far we have looked at the general theory and have not specified the reproducing kernel Hilbert space. To specify the space, let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be two sequences of positive numbers. We shall consider the d-dimensional weighted Sobolev spaces that have reproducing kernels of the form

$$K(\boldsymbol{x},\boldsymbol{y}) = K_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{d}\left(\beta_j+\gamma_j\mu_{a_j}(x_j,y_j)\right), \tag{3.9}$$

where

$$\mu_a(x,y) = \begin{cases} \min(|x-a|,|y-a|) & \text{if } (x-a)(y-a)>0,\\ 0 & \text{if } (x-a)(y-a)\le 0.\end{cases}$$

These and similar Sobolev spaces have been considered previously in works such as [2], [4], [8], and [9]. Common choices of the parameters are $a_j=1$ or $a_j=1/2$ for $1\le j\le d$. Then it may be shown (see [8]) that the associated shift-invariant kernel $K^*(\boldsymbol{x},\boldsymbol{y})$ defined by (3.5) is given by

$$K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{d}\left[\beta_j+\gamma_j\left(|x_j-y_j|^2-|x_j-y_j|+a_j^2-a_j+\tfrac{1}{2}\right)\right],$$

which may be written as

$$K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{d}\left[\beta_j+\gamma_j\left(B_2\left(|x_j-y_j|\right)+a_j^2-a_j+\tfrac{1}{3}\right)\right],$$

where $B_2(x) = x^2 - x + 1/6$ is the second-degree Bernoulli polynomial. Upon making use of Corollary 3.3, we find that

$$D^2(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(B_2\left(\left\{\frac{iz_j}{n}\right\}\right) + a_j^2 - a_j + \frac{1}{3}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]. \tag{3.10}$$

The integration problem is said to be "QMC tractable" if the minimal number $n_0(\varepsilon, d)$ of quadrature points $\boldsymbol{t}_i$ needed in (2.1) with $w_i = 1/n$ to reduce the worst-case error from its initial value $\|I_d\|$ by a factor $\varepsilon > 0$ is bounded by a polynomial in ε^{-1} and d. The problem is said to be "strongly QMC tractable" if that bound is independent of d.

To show that there exist rules of the form (1.4) which achieve strongly tractability error bounds in a probabilistic sense, we need to show that there exist rules for which $D^2(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}})$ given in (3.10) satisfies a corresponding bound. It follows from Lemma 8 in [9] that for $a_j = 1$, $j = 1, \ldots, d$, there exists a generating vector $\boldsymbol{z}$ such that

$$D^2(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}) \le \frac{1}{n}\left(\prod_{j=1}^{d}\left(\beta_j + \frac{\gamma_j}{2}\right) - \prod_{j=1}^{d}\left(\beta_j + \frac{\gamma_j}{3}\right)\right), \tag{3.11}$$

whereas $\|I_d\|$ for this case is given by (see [8])

$$\|I_d\| = \prod_{j=1}^{d}\left(\beta_j + \frac{\gamma_j}{3}\right). \tag{3.12}$$

Moreover, arguments similar to those in the proof of Theorem 5.2 in that paper show that if $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ satisfy

$$\sum_{j=1}^{\infty}\frac{\gamma_j}{\beta_j} < \infty,$$

then the ratio of the bound in (3.11) to $\|I_d\|$ is bounded independently of d, demonstrating the strong QMC tractability of multivariate integration under this condition.

4. Construction of the lattice rule $\boldsymbol{R_{n,d}}$. From the preceding discussion, it is clear that if we wish to use random shifts of rank-1 lattice rules to estimate the integral $I_d(f)$, then we should choose rank-1 lattice rules which give the best value of $D(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}})$. This means finding a $\boldsymbol{z}$ for the rule in (1.2) in some class so as to obtain the best value of $D(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}})$.

Here we propose finding such a $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)$ by doing a search one component at a time. Thus we fix $z_1 = 1$ and find $z_2 \in \{1, 2, \ldots, n-1\}$ to minimize $D(R_{n,2}, K^*_{2,\boldsymbol{\beta},\boldsymbol{\gamma}})$. With z_1 and z_2 fixed, we then find $z_3 \in \{1, 2, \ldots, n-1\}$ to minimize $D(R_{n,3}, K^*_{3,\boldsymbol{\beta},\boldsymbol{\gamma}})$, and so on.

We first give the theoretical foundation which ensures that the resulting rank-1 lattice rule achieves a bound (3.11) that corresponds to strong tractability. In fact, the bound that we obtain in Theorem 4.1 below (for the case $a_j = 1$) is exactly the

bound given in (3.11). Throughout the rest of the paper, we will use the shorter notation

$$D^2(\boldsymbol{z}) := D^2(R_{n,d}, K^*_{d,\boldsymbol{\beta},\boldsymbol{\gamma}}),$$

where $\boldsymbol{z}$ is the generating vector for the rank-1 lattice rule $R_{n,d}$.

THEOREM 4.1. *Let n be a prime number, $n > 1$. Suppose there exists an integer vector $\hat{\boldsymbol{z}} \in \{1, 2, \ldots, n-1\}^d$ such that*

$$D^2(\hat{\boldsymbol{z}}) \le \frac{1}{n}\left(\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{2}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{3}\right)\right]\right). \tag{4.1}$$

Then there exists $z_{d+1} \in \{1, 2, \ldots, n-1\}$ such that

$$D^2\left(\hat{\boldsymbol{z}}, z_{d+1}\right) \le \frac{1}{n}\left(\prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{2}\right)\right] - \prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{3}\right)\right]\right).$$

Moreover, the bound (4.1) *holds for $d = 1$.*
(Here $(\hat{\boldsymbol{z}}, z_{d+1}) \in \{1, 2, \ldots, n-1\}^{d+1}$ is just $\hat{\boldsymbol{z}}$ with the one additional component z_{d+1}.)

Proof. It follows from (3.10) that, for any $\boldsymbol{z} \in \{1, 2, \ldots, n-1\}^d$ and $z_{d+1} \in \{1, 2, \ldots, n-1\}$, we have

$$\begin{aligned}
&D^2\left(\boldsymbol{z}, z_{d+1}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(B_2\left(\left\{\frac{iz_j}{n}\right\}\right) + a_j^2 - a_j + \frac{1}{3}\right)\right] \\
&\qquad \times \left[\beta_{d+1} + \gamma_{d+1}\left(B_2\left(\left\{\frac{iz_{d+1}}{n}\right\}\right) + a_{d+1}^2 - a_{d+1} + \frac{1}{3}\right)\right] \\
&\quad - \prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right] \\
&= \left[\beta_{d+1} + \gamma_{d+1}\left(a_{d+1}^2 - a_{d+1} + \frac{1}{3}\right)\right] D^2(\boldsymbol{z}) \\
&\quad + \frac{\gamma_{d+1}}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(B_2\left(\left\{\frac{iz_j}{n}\right\}\right) + a_j^2 - a_j + \frac{1}{3}\right)\right] B_2\left(\left\{\frac{iz_{d+1}}{n}\right\}\right).
\end{aligned}$$

Upon separating out the $i = n$ term and using $B_2(0) = 1/6$, this expression for $D^2\left(\boldsymbol{z}, z_{d+1}\right)$ becomes

$$\begin{aligned}
&\left[\beta_{d+1} + \gamma_{d+1}\left(a_{d+1}^2 - a_{d+1} + \frac{1}{3}\right)\right] D^2(\boldsymbol{z}) + \frac{\gamma_{d+1}}{6n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] \\
&+ \frac{\gamma_{d+1}}{n}\sum_{i=1}^{n-1}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(B_2\left(\left\{\frac{iz_j}{n}\right\}\right) + a_j^2 - a_j + \frac{1}{3}\right)\right] B_2\left(\left\{\frac{iz_{d+1}}{n}\right\}\right).
\end{aligned} \tag{4.2}$$

In this expression, the only z_{d+1} dependence is in the last term. Suppose we average this last term over all the possible values of z_{d+1} to form

$$\Theta = \frac{\gamma_{d+1}}{n} \sum_{i=1}^{n-1} \prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(B_2 \left(\left\{ \frac{i z_j}{n} \right\} \right) + a_j^2 - a_j + \frac{1}{3} \right) \right] \times \frac{1}{n-1} \sum_{z_{d+1}=1}^{n-1} B_2 \left(\left\{ \frac{i z_{d+1}}{n} \right\} \right).$$

When n is prime, for fixed i satisfying $1 \le i \le n-1$ the values of $\{iz_{d+1}/n\}$ as z_{d+1} runs from 1 to $n-1$ are just $1/n, 2/n, \ldots, (n-1)/n$ in some order, and hence we have

$$\frac{1}{n-1} \sum_{z_{d+1}=1}^{n-1} B_2 \left(\left\{ \frac{i z_{d+1}}{n} \right\} \right) = \frac{1}{n-1} \left(\sum_{z_{d+1}=1}^{n} B_2 \left(\frac{z_{d+1}}{n} \right) - B_2(1) \right).$$

By recalling that $B_2(x) = x^2 - x + 1/6$ and using the well-known sums for the first n positive integers and the squares of the first n positive integers, we have

$$\begin{aligned} \sum_{z_{d+1}=1}^{n} B_2 \left(\frac{z_{d+1}}{n} \right) &= \sum_{z_{d+1}=1}^{n} \left[\left(\frac{z_{d+1}}{n} \right)^2 - \left(\frac{z_{d+1}}{n} \right) + \frac{1}{6} \right] \\ &= \frac{(n+1)(2n+1)}{6n} - \frac{n+1}{2} + \frac{n}{6} = \frac{1}{6n}. \end{aligned} \tag{4.3}$$

It then follows that

$$\frac{1}{n-1} \sum_{z_{d+1}=1}^{n-1} B_2 \left(\left\{ \frac{i z_{d+1}}{n} \right\} \right) = \frac{1}{n-1} \left(\frac{1}{6n} - \frac{1}{6} \right) = -\frac{1}{6n},$$

and hence

$$\begin{aligned} \Theta &= -\frac{\gamma_{d+1}}{6n^2} \sum_{i=1}^{n-1} \prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(B_2 \left(\left\{ \frac{i z_j}{n} \right\} \right) + a_j^2 - a_j + \frac{1}{3} \right) \right] \\ &= -\frac{\gamma_{d+1}}{6n} \left(D^2(\boldsymbol{z}) - \frac{1}{n} \prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \frac{1}{2} \right) \right] + \prod_{j=1}^{d} \left[\beta_j + \gamma_j \left(a_j^2 - a_j + \frac{1}{3} \right) \right] \right), \end{aligned}$$

where we have made use of (3.10). Since Θ involves an average over z_{d+1}, it follows

that there exists $z_{d+1} \in \{1, 2, \ldots, n-1\}$ for which

$$
\begin{aligned}
&D^2(\boldsymbol{z}, z_{d+1}) \\
&\le \left[\beta_{d+1} + \gamma_{d+1}\left(a_{d+1}^2 - a_{d+1} + \frac{1}{3}\right)\right] D^2(\boldsymbol{z}) + \frac{\gamma_{d+1}}{6n} \prod_{j=1}^{d} \left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] \\
&\quad - \frac{\gamma_{d+1}}{6n}\left(D^2(\boldsymbol{z}) - \frac{1}{n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right]\right. \\
&\qquad\qquad \left. + \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right) \\
&= \left[\beta_{d+1} + \gamma_{d+1}\left(a_{d+1}^2 - a_{d+1} + \frac{1}{3} - \frac{1}{6n}\right)\right] D^2(\boldsymbol{z}) \\
&\quad + \frac{\gamma_{d+1}}{6n}\left(\left(1 + \frac{1}{n}\right)\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right).
\end{aligned}
$$

Using the hypothesis in the theorem, we now obtain

$$
\begin{aligned}
&D^2(\hat{\boldsymbol{z}}, z_{d+1}) \\
&\le \left[\beta_{d+1} + \gamma_{d+1}\left(a_{d+1}^2 - a_{d+1} + \frac{1}{3} - \frac{1}{6n}\right)\right] \\
&\qquad \times \frac{1}{n}\left(\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right) \\
&\quad + \frac{\gamma_{d+1}}{6n}\left(\left(1 + \frac{1}{n}\right)\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right) \\
&= \frac{1}{n}\left(\prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] - \prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right) \\
&\quad - \gamma_{d+1}\left(\frac{1}{6n} - \frac{1}{6n^2}\right)\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right] \\
&\le \frac{1}{n}\left(\prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{2}\right)\right] - \prod_{j=1}^{d+1}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]\right),
\end{aligned}
$$

which is the desired result.

In one dimension, it is known that there is only one n-point lattice rule, namely, the n-point rectangle rule. Thus we may take $z_1 = 1$ and obtain from (3.10) that

$$
\begin{aligned}
D^2(1) &= \frac{1}{n}\sum_{i=1}^{n}\left[\beta_1 + \gamma_1\left(B_2\left(\left\{\frac{i}{n}\right\}\right) + a_1^2 - a_1 + \frac{1}{3}\right)\right] - \left[\beta_1 + \gamma_1\left(a_1^2 - a_1 + \frac{1}{3}\right)\right] \\
&= \frac{\gamma_1}{n}\sum_{i=1}^{n} B_2\left(\left\{\frac{i}{n}\right\}\right) = \frac{\gamma_1}{6n^2} \le \frac{\gamma_1}{6n},
\end{aligned}
$$

where we have made use of (4.3). Thus (4.1) holds for $d = 1$. □

Given n (a prime number) and d, Theorem 4.1 leads to an algorithm for finding a generating vector $\boldsymbol{z} = (z_1, z_2, \dots)$ such that

$$D^2(z_1, z_2, \dots, z_d) \le \frac{1}{n}\left(\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{2}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \tfrac{1}{3}\right)\right]\right)$$

for arbitrarily large values of d. We note that this bound is the average of $D^2(Q_{n,d}, K_{d,\boldsymbol{\beta},\boldsymbol{\gamma}})$ over all the points $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n$ of an n-point QMC rule $Q_{n,d}$.

ALGORITHM 4.2.

1. *Set z_1, the first component of $\boldsymbol{z}$, to* 1.
2. *For $d = 2, 3, \dots, d_{\max} - 1, d_{\max}$, find $z_d \in \{1, 2, \dots, n-1\}$ such that*

$$D^2(z_1, z_2, \dots, z_d) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(B_2\left(\left\{\frac{iz_j}{n}\right\}\right) + a_j^2 - a_j + \frac{1}{3}\right)\right] - \prod_{j=1}^{d}\left[\beta_j + \gamma_j\left(a_j^2 - a_j + \frac{1}{3}\right)\right]$$

is minimized.

We see from the algorithm that the cost of constructing a rank-1 lattice rule for all dimensions up to $d_{\max}$ is $O(n^2 d_{\max}^2)$. This can be reduced to $O(n^2 d_{\max})$ if we store (at a cost of $O(n)$ storage) the products during the search. Also, if at a later time components of $\boldsymbol{z}$ for further dimensions are needed, the algorithm can be restarted in an obvious way.

5. Numerical searches. Here we present some results from implementing the algorithm given in the previous section for dimensions d up to $d_{\max} = 100$. We consider only values of z_d satisfying $1 \le z_d \le (n-1)/2$, since

$$B_2\left(\left\{\frac{iz_d}{n}\right\}\right) = B_2\left(1 - \left\{\frac{i(n-z_d)}{n}\right\}\right) = B_2\left(\left\{\frac{i(n-z_d)}{n}\right\}\right).$$

All the runs had $\beta_j = 1$ and $a_j = 1$ for $1 \le j \le d$. The values of n used were the prime numbers 2003, 8009, and 32003, so $n^{1/2}$ approximately doubles from one value of n to the next. The results are mainly presented in graphical form. However, in Table 5.1, we present d, z_d, and the corresponding value of $D(\boldsymbol{z})$ for the case when $\gamma_j = 0.9^j$ and $n = 2003$ to allow readers to check their own implementation of Algorithm 4.2.

Figure 5.1 shows a graph of $D(\boldsymbol{z})$ against d, for d increasing in steps of 5, in the case when $\gamma_j = 0.9^j$. Figure 5.2 shows the same information for the case $\gamma_j = 1/j^2$. It seems, on comparing the results with $n = 2003$, 8009, and 32003, that the convergence of $D(\boldsymbol{z})$ to zero as n increases is faster than the theoretically predicted $O(n^{-1/2})$ (since the latter would predict that successive errors would halve from one value of n to the next).

From the discussion preceding Algorithm 4.2, we see that the $\boldsymbol{z}$ constructed using the algorithm is such that $D(\boldsymbol{z}) \le C_d n^{-1/2}$, where (since $\beta_j = 1$ and $a_j = 1$)

$$C_d = \left(\prod_{j=1}^{d}\left(1 + \frac{\gamma_j}{2}\right) - \prod_{j=1}^{d}\left(\beta_j + \frac{\gamma_j}{3}\right)\right)^{1/2}.$$

TABLE 5.1
$n = 2003$ *and* $\gamma_j = 0.9^j$.

d	z_d	$D(\boldsymbol{z})$	d	z_d	$D(\boldsymbol{z})$	d	z_d	$D(\boldsymbol{z})$
1	1	1.9336E-04	35	781	4.5241E-02	69	512	5.0287E-02
2	765	4.3028E-04	36	739	4.5735E-02	70	455	5.0303E-02
3	343	8.6132E-04	37	761	4.6183E-02	71	58	5.0317E-02
4	849	1.4677E-03	38	723	4.6589E-02	72	802	5.0330E-02
5	702	2.4072E-03	39	571	4.6957E-02	73	284	5.0341E-02
6	880	3.5417E-03	40	809	4.7291E-02	74	680	5.0352E-02
7	416	4.8706E-03	41	458	4.7593E-02	75	442	5.0361E-02
8	449	6.4552E-03	42	168	4.7867E-02	76	372	5.0369E-02
9	581	8.2240E-03	43	470	4.8115E-02	77	932	5.0377E-02
10	989	1.0138E-02	44	612	4.8340E-02	78	943	5.0384E-02
11	735	1.2055E-02	45	968	4.8543E-02	79	14	5.0390E-02
12	378	1.4090E-02	46	350	4.8727E-02	80	247	5.0396E-02
13	326	1.6169E-02	47	857	4.8893E-02	81	133	5.0401E-02
14	465	1.8189E-02	48	550	4.9043E-02	82	204	5.0405E-02
15	892	2.0215E-02	49	956	4.9179E-02	83	749	5.0409E-02
16	591	2.2222E-02	50	89	4.9303E-02	84	655	5.0413E-02
17	354	2.4215E-02	51	195	4.9415E-02	85	429	5.0416E-02
18	927	2.6125E-02	52	900	4.9516E-02	86	263	5.0419E-02
19	743	2.7972E-02	53	644	4.9608E-02	87	372	5.0422E-02
20	461	2.9721E-02	54	435	4.9690E-02	88	943	5.0424E-02
21	217	3.1356E-02	55	873	4.9765E-02	89	442	5.0426E-02
22	628	3.2902E-02	56	160	4.9832E-02	90	932	5.0428E-02
23	488	3.4353E-02	57	296	4.9893E-02	91	680	5.0430E-02
24	82	3.5704E-02	58	44	4.9947E-02	92	14	5.0432E-02
25	725	3.6967E-02	59	205	4.9996E-02	93	204	5.0433E-02
26	853	3.8133E-02	60	595	5.0041E-02	94	888	5.0434E-02
27	564	3.9210E-02	61	672	5.0081E-02	95	284	5.0436E-02
28	837	4.0212E-02	62	198	5.0117E-02	96	802	5.0437E-02
29	395	4.1133E-02	63	614	5.0149E-02	97	247	5.0438E-02
30	64	4.1977E-02	64	273	5.0178E-02	98	749	5.0438E-02
31	366	4.2749E-02	65	36	5.0205E-02	99	655	5.0439E-02
32	60	4.3459E-02	66	859	5.0229E-02	100	429	5.0440E-02
33	34	4.4108E-02	67	712	5.0250E-02			
34	753	4.4699E-02	68	266	5.0269E-02			

To get an idea of how much smaller the actual values of $D(\boldsymbol{z})$ are compared to the bound of $C_d n^{-1/2}$, we take $d = 100$ and for the three values of n calculate the ratio $D(\boldsymbol{z})/(C_d n^{-1/2})$. The three values of this ratio in the two cases $\gamma_j = 0.9^j$ and $\gamma_j = 1/j^2$ are plotted against n (on a log scale) and displayed in Figure 5.3. The straight lines in the figure are intended to make it easier to view the data.

Figure 5.3 also provides some information about how good the rules are compared to the classical Monte Carlo algorithm. As is well known, this is a randomized algorithm in which the quadrature points are chosen randomly from a uniform distribution on $[0,1]^d$. For any particular sampling of n points, we can compute a worst-case error in our space H_d by using (3.7) with $w_i = 1/n$. It is shown in [9] that the expected value of the squared worst-case error is given by C_d^2/n. In this sense Figure 5.3 indicates that the rules constructed here on average make better point selections than classical Monte Carlo, with the advantage increasing with n.

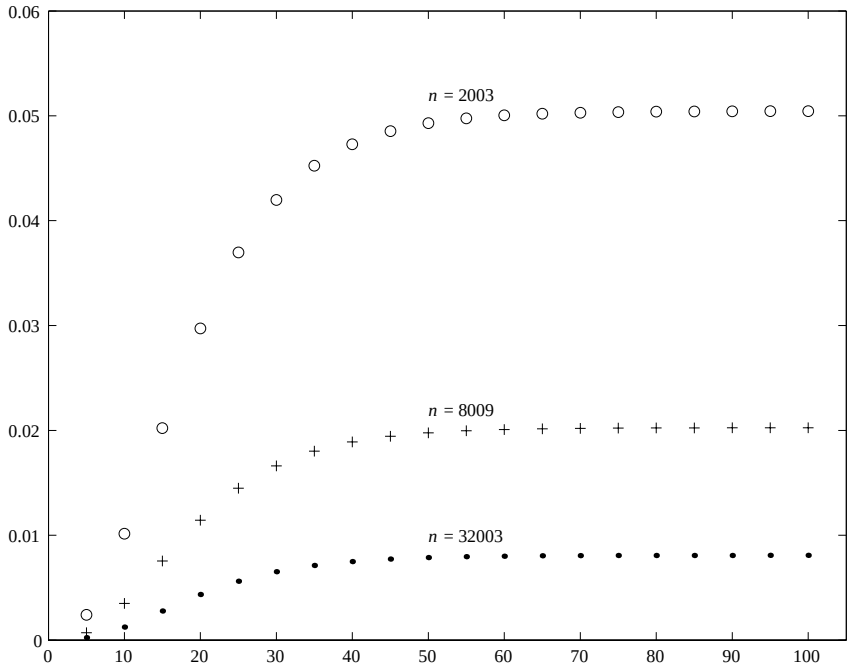


FIG. 5.1. $D(\boldsymbol{z})$ *against* d *for* $\gamma_j = 0.9^j$ *with* $n = 2003$, 8009, *and* 32003.

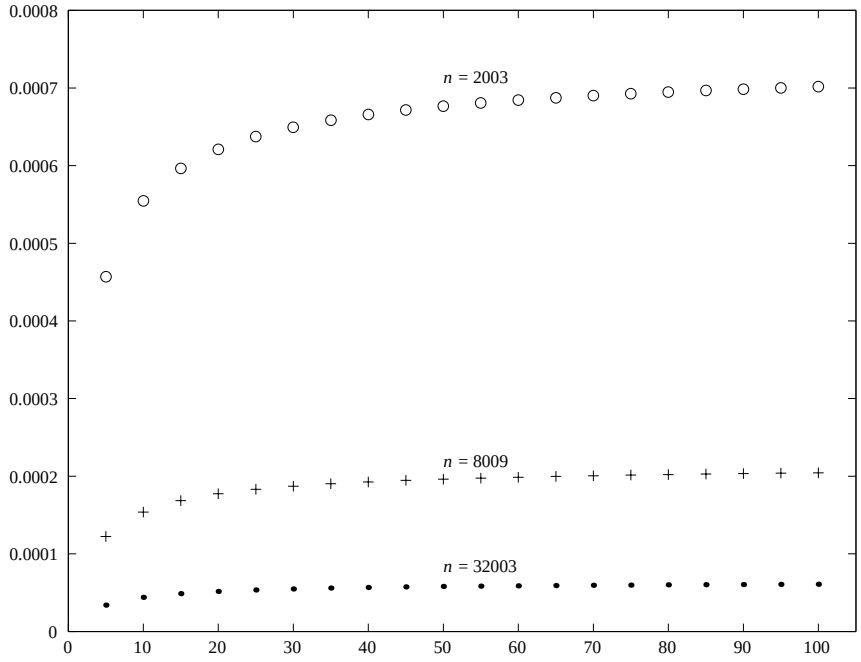


FIG. 5.2. $D(\boldsymbol{z})$ *against* d *for* $\gamma_j = 1/j^2$ *with* $n = 2003$, 8009, *and* 32003.

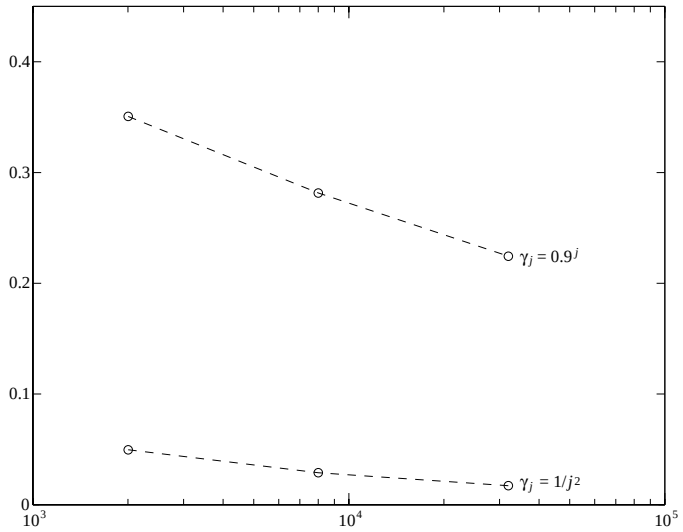


FIG. 5.3. *The ratio* $D(\boldsymbol{z})/(C_d n^{-1/2})$ *against* n *for* $n = 2003$, 8009, *and* 32003 *with* $\gamma_j = 0.9^j$ *and* $\gamma_j = 1/j^2$.

REFERENCES

[1] R. CRANLEY AND T. N. L. PATTERSON, *Randomization of number theoretic methods for multiple integration*, SIAM J. Numer. Anal., 13 (1976), pp. 904–914.

[2] F. J. HICKERNELL, *Lattice rules: How well do they measure up?*, in Random and Quasi-Random Point Sets, Lecture Notes in Statist. 138, P. Hellekalek and G. Larcher, eds., Springer-Verlag, New York, 1998, pp. 109–166.

[3] F. J. HICKERNELL AND H. S. HONG, *Quasi-Monte Carlo methods and their randomizations*, in Proceedings of the IMS Workshop on Applied Probability, 2002, to appear.

[4] F. J. HICKERNELL AND H. WOŹNIAKOWSKI, *Integration and approximation in arbitrary dimensions*, Adv. Comput. Math., 12 (2000), pp. 25–58.

[5] S. JOE, *Randomization of lattice rules for numerical multiple integration*, J. Comput. Appl. Math., 31 (1990), pp. 299–304.

[6] A. B. OWEN, *Randomly permuted* (t, m, s)*-nets and* (t, s)*-sequences*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Lecture Notes in Statist. 106, H. Niederreiter and P. J.-S. Shiue, eds., Springer-Verlag, New York, 1995, pp. 299–317.

[7] I. H. SLOAN AND S. JOE, *Lattice Rules for Multiple Integration*, Clarendon Press, Oxford, UK, 1994.

[8] I. H. SLOAN, F. Y. KUO, AND S. JOE, *On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces*, Math. Comp., 71 (2002), pp. 1609–1640.

[9] I. H. SLOAN AND H. WOŹNIAKOWSKI, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.

[10] I. H. SLOAN AND H. WOŹNIAKOWSKI, *Tractability of multivariate integration for weighted Korobov classes*, J. Complexity, 17 (2001), pp. 697–721.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1666–1682

COMPUTING ZEROS ON A REAL INTERVAL THROUGH CHEBYSHEV EXPANSION AND POLYNOMIAL ROOTFINDING*

JOHN P. BOYD†

Abstract. Robust polynomial rootfinders can be exploited to compute the roots on a real interval of a *nonpolynomial* function $f(x)$ by the following: (i) expand f as a Chebyshev polynomial series, (ii) convert to a polynomial in ordinary, series-of-powers form, and (iii) apply the polynomial rootfinder. (Complex-valued roots and real roots outside the target interval are discarded.) The expansion is most efficiently done by adaptive Chebyshev interpolation with N equal to a power of two, where N is the degree of the truncated Chebyshev series. All previous evaluations of f can then be reused when N is increased; adaption stops when N is sufficiently large so that further increases produce no significant change in the interpolant. We describe two conversion strategies. The "convert-to-powers" method uses multiplication by mildly ill-conditioned matrices to create a polynomial of degree N. The "degree-doubling" strategy defines a polynomial of larger degree $2N$ but is very well-conditioned. The "convert-to-powers" method, although faster, restricts N to moderate values; this can always be accomplished by subdividing the target interval. Both these strategies allow simultaneous approximation of many roots on an interval, whether simple or multiple, for nonpolynomial $f(x)$.

Key words. rootfinding, single transcendental equation, Chebyshev series

AMS subject classifications. 65H05, 42C10, 65E05

PII. S0036142901398325

1. Introduction.

One irony of the history of mathematics is that the problem of finding the roots of a polynomial, which taxed the brains of mathematicians from the Babylonians to Diophantus to Omar Khayyam to Cardano and Tartaglia in the Renaissance to Lagrange, Abel, Gauss, and Galois around the turn of the 19th century, is now largely uninteresting. Reliable polynomial rootfinding software, which requires no a priori estimates for the zeros, is now a part of almost all language packages (Matlab, Maple, Mathematica) and Fortran libraries (NAG, IMSL, etc.). The undergraduate who casually executes the one-line Matlab command, **roots(p)**, where **p** is a vector containing the coefficients of the polynomial, is blissfully ignorant of the three centuries of struggle to move from Ferrari's literal solution of the quartic, published in 1545, to Hermite's solution of the quintic 320 years later.

Nevertheless, the problem of finding the roots of a single transcendental equation in a single unknown is still a staple of numerical analysis courses. The reason is that, until recently, there was no black box for computing the zeros of a nonpolynomial $f(x)$. Bisection and Brent's algorithm will reliably find *some* roots, but this is not the same as finding *all roots* on an interval. It is particularly easy to miss zeros that are closely spaced or multiple.

Kavvadias and Vrahatis [10], Kavvadias, Makri, and Vrahatis [11], and Smiley and Chun [13] have developed bisection-like but more sophisticated subdivision strategies for reliable transcendental rootfinding. However, these algorithms are relatively slow. Later, we shall explain how these subdivision strategies can in principle be accelerated

*Received by the editors November 15, 2001; accepted for publication (in revised form) March 28, 2002; published electronically October 31, 2002. This work was supported by National Science Foundation grant OCE9986368.

http://www.siam.org/journals/sinum/40-5/39832.html

†Department of Atmospheric, Oceanic, and Space Sciences and Laboratory for Scientific Computation, University of Michigan, 2455 Hayward Avenue, Ann Arbor, MI 48109 (jpboyd@engin.umich.edu, http://www-personal.engin.umich.edu:/∼ jpboyd/).

(or replaced!) by the ideas introduced here.

We show that it is easy to extend the existing library software for *polynomial* f to *general* f merely by a simple intervention: expanding f as a Chebyshev series and then converting the Chebyshev approximation to an ordinary polynomial. As seen through the lens of Chebyshev polynomials, there is no such thing as a "transcendental" function: all rootfinding problems are polynomial rootfinding problems.

In his book *Applied Analysis* (1956) [12], Lanczos published the first example of the "Chebyshevization" of rootfinding. At a time when general polynomial solvers did not exist, he collapsed a cubic equation (hard) to a quadratic (easy!) by expanding the cubic as a Chebyshev series and then neglecting the third degree term.

In an earlier paper [5], we extended Lanczos's strategy to complicated $f(x)$. However, our earlier work was criticized because it did not provide estimates for the condition number of the conversion-to-powers step. Since Wilkinson's famous example of a very badly conditioned polynomial (well illustrated on pp. 330–331 of [1]), all right-thinking numerical analysts have cringed at working with a polynomial expressed as a sum of powers of x. In this work, we show that, although there is some ill-conditioning, the convert-to-powers strategy is robust and reliable if the degree of the Chebyshev expansion is restricted to moderate N (i.e., $N < 18$ or so). By subdividing an interval with many roots into subintervals, and applying a separate Chebyshev expansion to each one, the Chebyshev-to-powers strategy can be applied to almost all functions which are analytic on a desired target interval.

Furthermore, there is an alternative strategy, discussed here for the first time, which allows extraction of roots from a polynomial $h_{2N}(z)$ whose coefficients are simply those of the Chebyshev series. The ill-conditioning is completely eliminated. However, the degree of $h_{2N}(z)$ is twice that of the truncated Chebyshev series from whence it came.

Figure 1 schematically summarizes our algorithm.

The first issue is, How is the Chebyshev series computed? The answer is that $f(x)$ must be evaluated at a set of discrete points on the target interval; the Chebyshev coefficients are then given by a matrix-vector multiply where the vector holds the set of grid-point values of $f(x)$ and the elements of the matrix are trigonometric functions. The complete procedure is described in the appendix.

The second issue is, How does one determine when the truncation N is large enough? There is a well-established theory for doing this as reviewed in our book [6] and previous article [5]. The most systematic strategy mimics that of the Clenshaw–Curtis quadrature: the number of points is doubled until the approximation ceases to change; all previously used values of $f(x)$ are reused by finer approximations so that nothing is wasted. We shall briefly review stopping criteria below.

The third issue is, How does one convert a truncated Chebyshev series to an ordinary polynomial? We offer two ways. In the "convert-to-powers" strategy, the coefficients of the powers of x are the product of an upper triangular matrix with the vector of Chebyshev coefficients; the matrix elements are integers computed by a simple recurrence.

The "degree-doubling" algorithm defines an associated polynomial whose degree is twice that of the truncation of the Chebyshev series. However, the real part of the roots of this polynomial which lie on the unit circle in the complex plane are the roots of $f(x)$ on the real interval $x \in [a, b]$.

The fourth issue is, Given that the roots of a polynomial are notoriously sensitive to small perturbations to the coefficients of the powers of x, how ill-conditioned is the

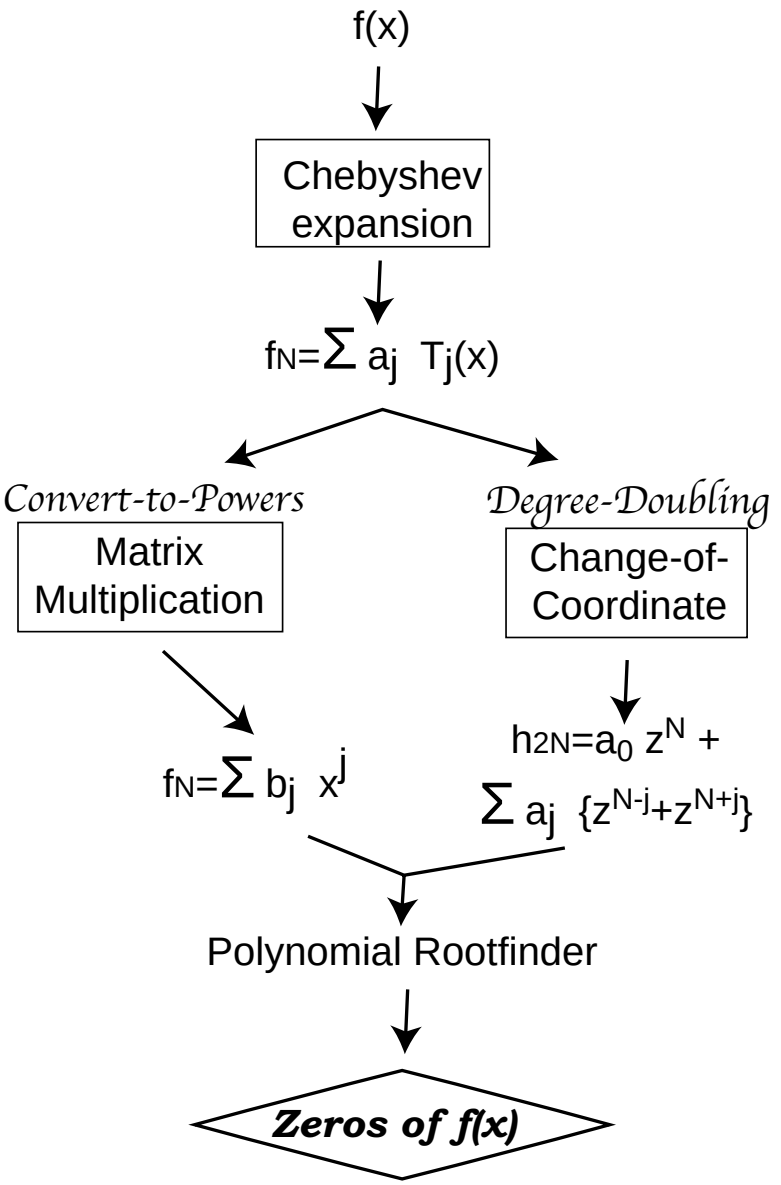


FIG. 1. *Schematic of rootfinding for a nonpolynomial* $f(x)$.

convert-to-powers algorithm? The answer is that if the Chebyshev degree is restricted, the roots of the polynomial *on the target interval* will be very good approximations to those of the truncated Chebyshev series, whose zeros are in turn very good approximations to those of the original $f(x)$. Small errors can easily be corrected by one or two secant or Newton iterations with $f(x)$. If the interval is large and has many roots, it may be necessary to subdivide the interval into subdomains and compute a different moderate degree Chebyshev series on each. With these precautions of degree restriction and interval subdivision followed by iteration with $f(x)$ itself, the convert-to-powers algorithm can yield roots to full machine precision.

In this article, we offer four novelties (beyond the earlier work of Lanczos and [5]):

1. Condition number estimates for the triangular matrices that convert Chebyshev coefficients to the coefficients of the same polynomial expressed as a sum of powers of x.
2. A second strategy for converting a truncated Chebyshev polynomial to a different polynomial of twice the degree but with (effectively) the same roots; the coefficients of the new polynomial are the same as the Chebyshev coefficients.
3. The Chebyshev-to-polynomial method is extended to *unbounded* intervals, either infinite or semi-infinite, so long as $f(x)$ asymptotes to a constant at infinity and has only a finite number of real roots.
4. A different, grid-point-value-based "stopping" criterion for assessing when f is approximated to sufficient accuracy by a Chebyshev series truncated after the Nth term.

The strength of the algorithm is that no a priori knowledge of the roots is needed. The reliability of existing polynomial solvers is extended to nonpolynomial $f(x)$.

The sections of the article are as follows:

Sec. 2: first two issues (computing the Chebyshev expansion).
Sec. 3: convert-to-powers and its condition number.
Sec. 4: bounds on errors in roots and their application.
Sec. 5: the degree-doubling theorem.
Sec. 6: numerical example: roots of $J_0(x)$.
Sec. 7: searching a region in the complex plane instead of a real interval.
Sec. 8: rootfinding on an infinite interval.
Sec. 9: summary and open problems.

2. Adaptive computation of the Chebyshev coefficients. Our Chebyshev approximation is a finite series of Chebyshev polynomials which interpolates $f(x)$ at a set of $(N+1)$ points known as the Chebyshev–Lobatto grid. This differs little from the truncation of the infinite Chebyshev polynomial series of f [6]. One must evaluate $f(x)$, the function whose roots are sought, at $(N+1)$ points on the target interval $x \in [a,b]$. The Chebyshev coefficients are then the vector which is the product of a square matrix with the column vector of grid-point values of f. (The grid points and the matrix elements are given in the appendix for arbitrary N.)

If $f(x)$ is expensive to evaluate, the best strategy is to restrict N to be a power of two. In this case, all previously computed grid-point values of $f(x)$ can be reused when N is doubled so that the maximum number of evaluations of f is never more than the smallest (power-of-two) N for which the "stopping criteria," is met. A similar strategy is employed in the adaptive, spectrally accurate Clenshaw–Curtis quadrature scheme [6].

As explained in [6], Chebyshev series for a function f which is *analytic* on the interval $x \in [a,b]$ converge *geometrically* fast; that is, the jth term (and also the absolute value of the jth coefficient) are bounded by ρ^j for some $\rho < 1$. The error in the $(N+1)$-point interpolation is typically the same order of magnitude as the last computed Chebyshev coefficient a_N [6]. [5] proposed a cautious "stopping criterion": increase N until $\sum_{j=[(2/3)N]}^{N} |a_j| < \epsilon$, where $[(2/3)N]$ denotes the integer closest to $2N/3$.

One can also use a grid-point value criterion which is given here for the first time:

$$\max |f_{2N}(x_j) - f_N(x_j)| < \epsilon, \tag{1}$$

where the difference is calculated for all the points on the grid of $(2N+1)$ points which are not on the coarser grid of $(N+1)$ points. (At points common to both grids, both interpolants equal f and therefore each other.) Since the error of f_N tends to be a maximum roughly halfway between the points of the coarse grid, the difference at these intermediate points is likely to be quite close to the true maximum pointwise error of f_N.

Both these stopping criteria are *very conservative*; f_{2N} will usually have an error much smaller than ϵ with the grid-point criterion, which forces the error of the lower order approximation f_N to be less than ϵ. Reliability is built on conservative strategies, however.

2.1. Subdivision. As explained in the next section, conversion-to-powers is a well-conditioned process only if N is restricted to moderate degree. What if large N is needed to obtain an accurate Chebyshev approximation?

The answer is that one can divide the interval into subintervals. Our recommendation is to expand f on the entire interval first, even if this requires using large N. If the maximum degree that allows satisfactory conversion-to-powers is $N_{\max}$, the asymptotic theory of Chebyshev expansions [6] suggests that one should subdivide into $[N/N_{\max}]$ subintervals where the square brackets denote the integer closest to the ratio of $N/N_{\max}$. Cautious arithmurgists are encouraged to use a somewhat larger number of subdivisions.

Once the split into subdomains has been made, the algorithm can be applied on each subinterval without modification.

2.2. Scaling. Chebyshev expansions are highly uniform in the sense that the maximum pointwise error (absolute error) oscillates with peaks and troughs of similar amplitude over the entire expansion interval, $x \in [a,b]$. If $f(x)$ is itself highly *nonuniform*, such as $\exp(10x)\sin(x)$, then the Chebyshev series will have large *relative* errors where $f(x)$ is very small.

There are two remedies. The first is to subdivide into subintervals sufficiently small so that $f(x)$ varies only mildly over each subdomain. The second is to multiply f by a smooth scaling function that eliminates the large fluctuations in magnitude. For our example, $\tilde{f} \equiv \exp(-10x)f(x) = \sin(x)$ has the same roots as f, but, because it is much more uniform, the Chebyshev expansion of $\tilde{f}$ will have much smaller relative error and yield much more accurate approximations to the roots. Devising a smooth scaling function may be difficult, however, as illustrated in [5].

2.3. Nonanalytic/nonsmooth $f(x)$. Chebyshev expansions converge, but at a very slow rate, if $f(x)$ has poles, branch points, discontinuities, or other singularities on the expansion interval, $x \in [a,b]$, including singularities at the endpoints. We therefore caution the reader that Chebyshev rootfinding methods are useful only when $f(x)$ is analytic everywhere on the expansion interval including the endpoints. (Singularities off the expansion interval, whether at real or complex locations, however, are powerless to destroy the good properties of the algorithm and are thus largely irrelevant.)

3. Converting Chebyshev series to polynomials, I: Convert-to-powers. The Chebyshev expansion of $f(x)$ on $x \in [a,b]$ is

$$f_N \equiv \sum_{j=0}^{N} a_j T_j(y) = \sum_{j=}^{N} b_j\, y^j, \qquad y(x) \in [-1,1], \tag{2}$$

where

$$y \equiv \frac{2x-(b+a)}{b-a}, \qquad y \in [-1,1] \;\leftrightarrow\; x \in [a,b], \tag{3}$$

is the stretched-and-translated argument of the Chebyshev polynomials. The cost of transforming from $\{a_j\}$ to $\{b_j\}$ can be halved by splitting f_N into its even and odd parts: let $\vec{a}^{even}$ and $\vec{b}^{even}$ be vectors containing the even degree coefficients $\{a_0, a_2, a_4 \ldots\}$ and $\{b_0, b_2, b_4 \ldots\}$, respectively. Then

$$\vec{b}^{even} = \vec{\vec{Q}}^{even}\, \vec{a}^{even}. \tag{4}$$

Explicitly, the upper left block is given by

$$\vec{\vec{Q}}^{even,block} = \begin{vmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & 2 & -8 & 18 & -32 & 50 & -72 & 98 & -128 \\ 0 & 0 & 8 & -48 & 160 & -400 & 840 & -1568 & 2688 \\ 0 & 0 & 0 & 32 & -256 & 1120 & -3584 & 9408 & -21504 \\ 0 & 0 & 0 & 0 & 128 & -1280 & 6912 & -26880 & 84480 \\ 0 & 0 & 0 & 0 & 0 & 512 & -6144 & 39424 & -180224 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2048 & -28672 & 212992 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8192 & -131072 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32768 \end{vmatrix}. \tag{5}$$

The jth column is composed of the coefficients of T_{2j-2} in powers of x.

Similarly,

$$\vec{\vec{Q}}^{odd,block} = \begin{vmatrix} 1 & -3 & 5 & -7 & 9 & -11 & 13 & -15 & 17 \\ 0 & 4 & -20 & 56 & -120 & 220 & -364 & 560 & -816 \\ 0 & 0 & 16 & -112 & 432 & -1232 & 2912 & -6048 & 11424 \\ 0 & 0 & 0 & 64 & -576 & 2816 & -9984 & 28800 & -71808 \\ 0 & 0 & 0 & 0 & 256 & -2186 & 16640 & -70400 & 239360 \\ 0 & 0 & 0 & 0 & & 1024 & -13312 & 92160 & -452608 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4096 & -61440 & 487424 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 16384 & -278528 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 65536 \end{vmatrix}. \tag{6}$$

The elements can be computed by recurrence relation:

$$Q_{11}^{even} = 1, \qquad Q_{jj}^{even} = 2^{2j-3}, \qquad j = 2, 3, \ldots. \tag{7}$$

The recurrence, vertically up the j-column from the diagonal, is then

$$Q_{j-K,j}^{even} = \text{round}\left\{ -\frac{(2j-2K)(2j-2K-1)}{2K\,(4j-2K-4)} Q_{j-K+1,j}^{even} \right\}, \tag{8}$$

$$Q_{jj}^{odd} = 2^{2j-2}, \qquad j = 1, 2, 3, \ldots, \tag{9}$$

$$Q_{j-K,j}^{odd} = \text{round}\left\{ -\frac{(2j-2K+1)(j-K)}{K\,(4j-2K-2)} Q_{j-K+1,j}^{odd} \right\}. \tag{10}$$

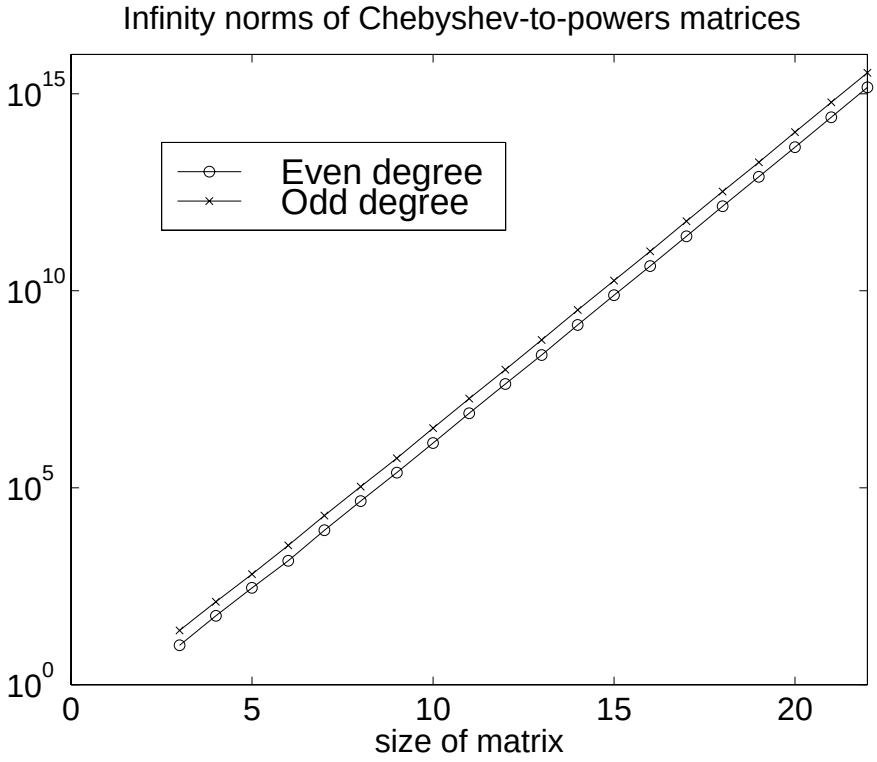


FIG. 2. *L_∞ matrix norms of $\vec{\vec{Q}}^{even}$ and $\vec{\vec{Q}}^{odd}$ versus the dimension of the matrix j.*

Because the elements are *integers*, we can eliminate roundoff error in the recurrences by rounding to the nearest integer.

Figure 2 shows how the norms of the transformed matrices grow exponentially, roughly as

$$||\vec{\vec{Q}}^{even}||_\infty \sim 0.016\,(5.8)^j, \qquad ||\vec{\vec{Q}}^{odd}||_\infty \sim 0.039\,(5.8)^j. \tag{11}$$

If we wish to avoid sacrificing more than six decimal places of accuracy or, more precisely, to guarantee that the errors in the coefficients of the power form are no more than a million times the floating point and truncation errors in the Chebyshev coefficients, we must restrict the size of $\vec{\vec{Q}}^{even}$ and $\vec{\vec{Q}}^{odd}$ to nine or less since

$$||\vec{\vec{Q}}^{even}||_\infty(9\times 9) = 243,712\,, \qquad ||\vec{\vec{Q}}^{odd}||_\infty(9\times 9) = 559,104. \tag{12}$$

If we put the even and odd polynomials together, we obtain a general nonsymmetric polynomial of degree 17.

4. Condition number of polynomial roots. Gautschi [8] and Winkler [14] give theorems on the condition number of polynomial roots. To give the flavor of these ideas without undue complexity, we shall state a simpler result applicable only in the limit of an arbitrarily small perturbation. Recall that the argument of the Chebyshev polynomials (and of the sum-of-powers into which it is transformed) is $y \in [-1, 1]$,

which is the image of the interval $x \in [a, b]$ through the linear change-of-coordinate $y \equiv (2x - (b + a))(b - a)$.

THEOREM 4.1 (sensitivity of polynomial roots). *Let y_* denote a real-valued root of multiplicity m on the interval $y \in [-1, 1]$ for a function f which is written terms of a basis as*

$$f(y) = \sum_{j=0}^{N} a_j \phi_j(y), \qquad y \in [-1, 1], \tag{13}$$

where the basis functions are either Chebyshev polynomials or powers of y and in either event satisfy

$$|\phi_j(y)| \le 1 \qquad \forall\, y \in [-1, 1]. \tag{14}$$

Let $\tilde{y}$ denote the root of the perturbed function $\tilde{f}$ which is equal to f except the kth coefficient has been altered by an amount ϵ:

$$\tilde{f}(y) = \sum_{j=0, j\neq k}^{N} a_j \phi_j(y) \; + (a_k + \epsilon)\, \phi_k(y), \qquad y \in [-1, 1]. \tag{15}$$

Then

$$|\tilde{y} - y_*| \le \epsilon^{1/m} \left| \frac{1}{m!} \frac{d^m f}{dy^m}(y_*) \right|^{-1/m} + O(\epsilon^{(m+1)/m}), \qquad |\epsilon| \ll 1. \tag{16}$$

Proof. Taylor's theorem at $y = y_*$ is

$$\tilde{f}(y) \approx \epsilon \phi_k(y) + \frac{1}{m!} \frac{d^m f}{dy^m}(y_*)(y - y_*)^m + O((y - y_*)^{m+1}), \tag{17}$$

since, at an mth order zero of f, the function itself and its first $(m - 1)$ derivatives are zero by definition. Solving the Taylor approximation for the root of the perturbed function, $\tilde{y}$, and then invoking the inequality that all basis functions are bounded in magnitude by one on the interval, is sufficient to prove the theorem.

The theorem implies that Wilkinson's famous example of an ill-conditioned polynomial, which has haunted the dreams of numerical analysts for a generation, is dreaded unduly. When the largest root of interest has magnitude $y_{\max} \gg 1$, then y^k can be become as large as $(y_{\max})^k$ at that root, and thus tiny changes in the coefficient of y^k can produce huge changes in the largest root. However, when $|y| \le 1$, all the powers of y are bounded by one, and a simple root will be altered only an $O(\epsilon)$ amount by an $O(\epsilon)$ perturbation of the coefficients.

Thus, for our purposes, the powers-of-x form is *not* ill-conditioned. The only difficulty is that the Chebyshev-to-powers matrix multiplication greatly magnifies small errors in the Chebyshev coefficients. Thus, when N is large, an alteration of ϵ in the kth Chebyshev coefficient will produce changes of millions or billions of ϵ in the coefficients of the powers-of-x form. This in turn will produce a comparable change in a simple root.

We conclude that if N is restricted to moderate values, such as $N < 18$, by subdividing the original interval, then the Chebyshev-to-powers algorithm will be reasonably well-conditioned, where "reasonably" means that we lose no more than six decimal places to the ugly condition numbers of the transformation matrices $\vec{Q}^{even}$, $\vec{Q}^{odd}$, and are still able to compute the roots to nine or ten decimal places.

5. Conversion to a polynomial, II: Degree-doubling. An alternative method for deriving a polynomial from a truncated Chebyshev series is given by the following.

THEOREM 5.1 (associated double-degree polynomial). *Let $f_N(x)$ be a polynomial:*

$$f_N(x) \equiv \sum_{j=0}^{N} a_j\, T_j(x). \tag{18}$$

Define a polynomial h of twice the degree through

$$h_{2N}(z) \equiv \sum_{j=0}^{2N} b_j z^j, \tag{19}$$

where

$$b_j = \begin{cases} a_{j-N}, & j > N, \\ 2\, a_0, & j = N, \\ a_{N-j}, & j < N. \end{cases} \tag{20}$$

Then the roots x_k of f_N on the real interval $x \in [-1, 1]$ are related to the roots z_k of $h_N(z)$ on the unit disk in the complex z-plane through

$$x_k = \Re(z_k). \tag{21}$$

Proof. The identity $T_j(x) = \cos(jt)$ when $x = \cos(t)$ plus $\cos(t) \equiv (\exp(it) + \exp(it))/2$ implies that

$$f_N(\cos(t)) \equiv \sum_{j=0}^{N} a_j\, \{\exp(it) + \exp(-it)\}\, /2. \tag{22}$$

Define

$$h_{2N}\left(\exp(it)\right) \equiv 2\, \exp(iNt) f_N(\cos(t)). \tag{23}$$

Because $\exp(iNt)$ never vanishes, the roots of the product are identical with those of $f_N(\cos(t))$. Defining $z \equiv \exp(it)$ and recalling $\exp(ijt) = [\exp(it)]^j$ proves the theorem.

6. Numerical example. Figure 3 shows the success of the Chebyshev algorithm using the two different strategies for polynomial creation: convert-to-powers on the left and degree-doubling on the right.

The left panel shows that, for some $f(x)$ at least, the restriction to $N \leq 17$ for convert-to-powers is very conservative. With $N = 40$, the maximum relative error is less than 1 part in 3800 in all of the first 19 roots of the J_0 Bessel function. By increasing N, the error can be reduced to $O(10^{-12})$ for some of the roots. However, there are signs of ill-conditioning: the roots at the ends of chosen expansion interval do not converge with increasing N.

The degree-doubling method requires more computation because the "black box" polynomial rootfinder is asked to solve a polynomial of degree $2N$ instead of N. (The cost of the triangular matrix multiplications of the convert-to-powers scheme is only $(N^2/2)$ multiplications and the same number of additions.) However, the right panel shows that the degree-doubling method is completely free of the ill-conditioning that dogs the convert-to-powers method for large N.

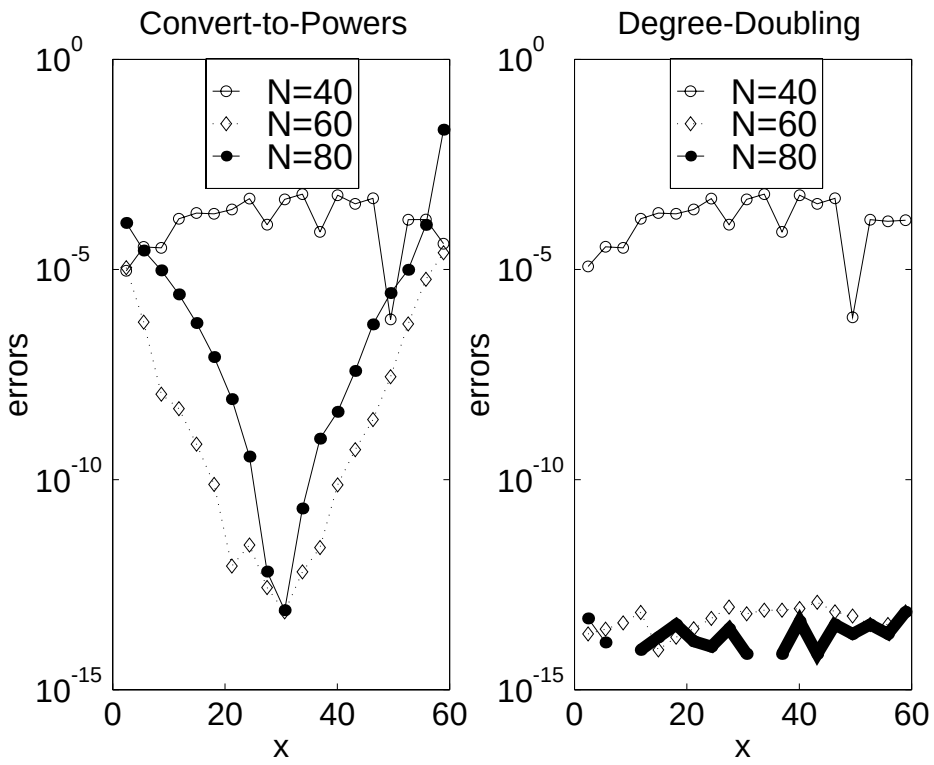


FIG. 3. *Comparison of the "convert-to-powers" method (left) and the "degree-doubling" algorithm (right) for computing the first* 19 *roots of* $J_0(x)$ *by Chebyshev expansions truncated after the Nth term on the interval* $x \in [0, 60]$.

Nevertheless, the convert-to-powers strategy is very effective even for N far larger than the restrictions suggested in earlier sections. Why is the ill-conditioning completely not ruinous for $N = 80$? Figure 4 (left) shows that the coefficients asymptote to a plateau of $O(10^{-16})$ for $N > 60$, which is a magnitude controlled by roundoff error. The very small size of the high degree Chebyshev coefficients keeps these coefficients from causing major problems when the series is converted to an ordinary polynomial of degree 80. It is important to the note from (5) and (6) that the size of the matrix elements increases rapidly with the column so that the elements that dominate the condition number of these matrices are the multipliers of very tiny coefficients in the Bessel–Chebyshev series.

Even with the mild ill-conditioning, it is remarkable that 19 roots can be captured so accurately with no more than two to four evaluations of $f(x)$ per root. Another robust interval rootfinding such as bisection would surely require far more evaluations of f.

When N is restricted to smaller values (and the expansion interval proportionately reduced), the ill-conditioning disappears as predicted. Figure 5 shows that, for a smaller interval, the errors are nearly uniform over the interval and decrease uniformly as N increases. (No comparison with the degree-doubling scheme is shown because for this case, where N is small and the convert-to-powers method is well-conditioned, there is no graphically discernible difference between the two algorithms.) With $N =$

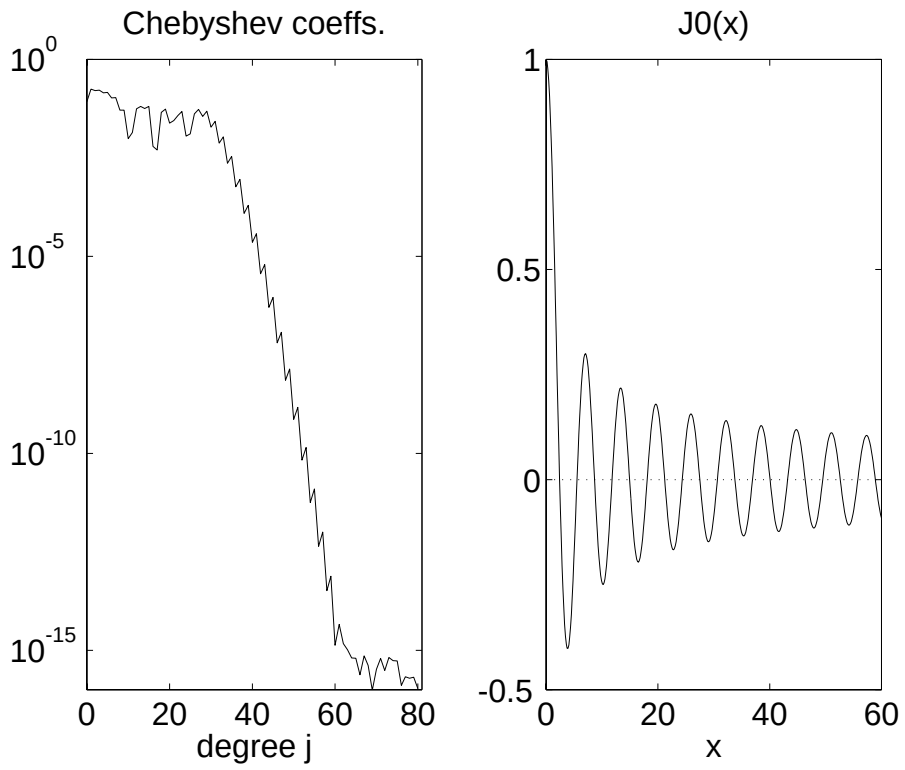


FIG. 4. *Left: absolute values of the Chebyshev coefficients of* $J_0(x)$ *for the interval* $x \in [0, 60]$. *Right: A plot of* J_0 *on this same interval, which contains* 19 *roots of* J_0.

28, somewhat larger than our ultraconservative recommended maximum of $N = 17$, the first six roots of J_0 are all approximated to within a relative error of less than 1 part in a hundred billion!

7. Generalization and alternative: Interpolation in powers around a circle in the complex plane. To generalize our method to the complex plane, the crucial fact is that a power series is optimal for interpolation in a *disk* in the complex plane in the same way that Chebyshev polynomials are optimal for interpolation on a real interval [9]. Interpolation by a series of Chebyshev polynomials on a real interval is replaced by interpolation of a series of powers of z on a circle in the complex z-plane. By applying interpolation-on-a-circle to many overlapping circles, polynomial rootsolvers can thus be applied to find roots of nontranscendental functions in arbitrary regions of the complex plane. Although they employ a different strategy to find roots within a circle, Delves and Lyness give a good discussion of such regional rootfinding methods [7].

8. Rootfinding on the whole real axis. If a function f has an *infinite* numbers of roots on the real axis, it is obviously impractical to find them all numerically. However, it is often possible to find an *asymptotic* approximation to the roots of large $|x|$ and then numerically compute the finite number of roots for which $|x|$ is too small for the asymptotic formula to be accurate. For the J_0 Bessel function used as an

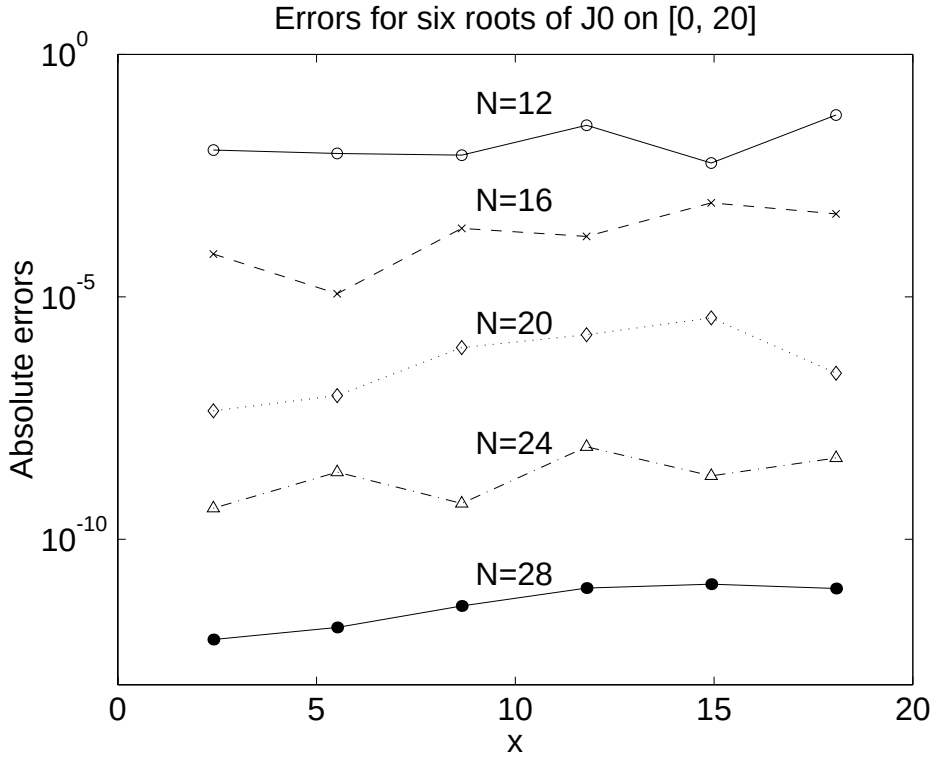


FIG. 5. *Absolute errors for the first six roots of the J_0 Bessel function. N is the degree of the Chebyshev approximation used to generate the polynomial whose roots were computed to approximate those of the Bessel function.*

example, for instance, the kth root, conventionally denoted by $j_{0,k}$, is asymptotically

$$j_{0,k} \sim (k - 1/4)\pi + \frac{1}{8(k-1/4)\pi} - \frac{31}{6(4k-1)^3\pi^3} + O(k^{-5}). \tag{24}$$

This approximation has an absolute error of only 0.0018 even for the first root and an error of just 2.7×10^{-10} for the 20th root.

When f has only a finite number of roots on an unbounded interval, it is possible to find them directly by using a change-of-coordinate that maps the infinite interval into the canonical interval for Chebyshev polynomials, $x \in [-1, 1]$. One can then apply the Chebyshev-to-polynomial algorithms, either convert-to-powers or degree-doubling, without modification.

If the coordinate on the infinite interval is denoted by y, then a good mapping [2] is

$$y = \frac{L\,x}{\sqrt{1-x^2}}; \qquad \leftrightarrow \qquad x = \frac{y}{\sqrt{L^2+y^2}}, \qquad x \in [-1, 1],\ \ y \in [-\infty, \infty], \tag{25}$$

where L is a constant, the user-choosable map parameter. Although the optimum L is problem-dependent [2, 6], the Chebyshev rate of convergence is not very sensitive to L, and $L = 1$ is a good choice in most applications. Other mappings can work, too, as discussed in [4] and [6]; a similar transformation for the semi-infinite interval is explained in [3] and [6].

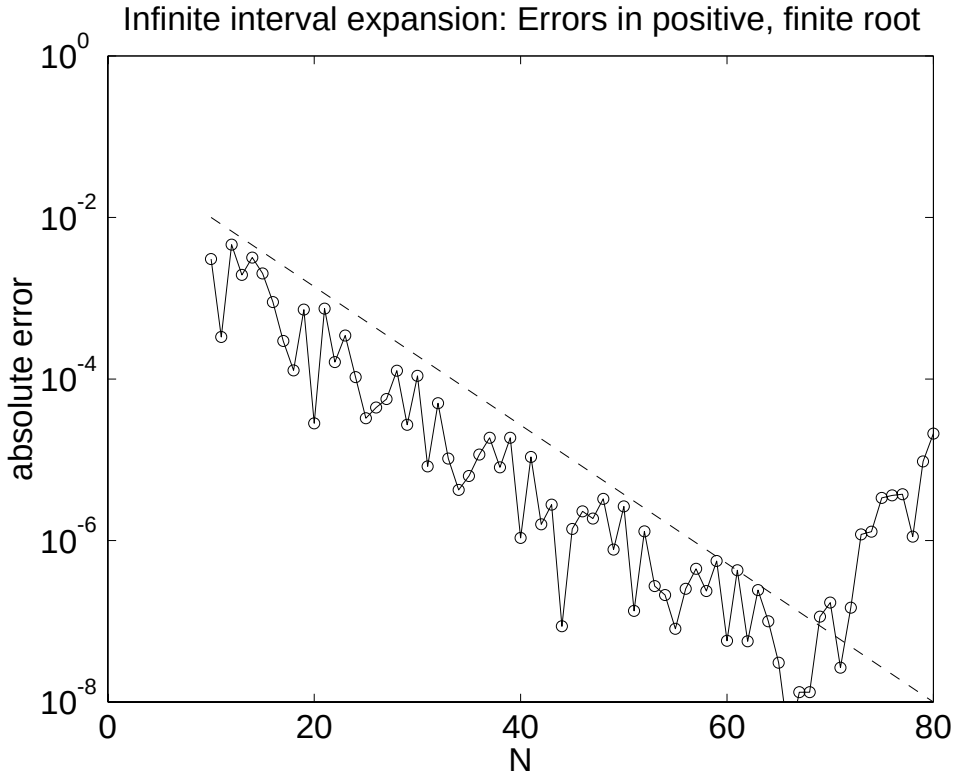


FIG. 6. *Infinite interval example,* $f(y) \equiv (2y^2 - 1)\exp(-[1/2]y^2)$*, with symmetry exploited by using Chebyshev expansion only on* $x \in [0, 1]$*, which is the image of the positive real axis. The dashed line is a guideline,* $0.072\exp(-0.197N)$*, to show that the trend of the error is exponential decrease with* N*. The rise in error for large* N *is due to the use of the "convert-to-powers" method of computing an ordinary polynomial from the Chebyshev series; the degree-doubling method is free of this problem.*

For example, the function

$$f(y) \equiv (2y^2 - 1)\exp(-[1/2]y^2) \tag{26}$$

has only two roots on the infinite interval, $y_* = \pm 1/\sqrt{2}$. Under the mapping, this becomes

$$f(y) = \left(2\frac{L^2x^2}{1-x^2} - 1\right)\exp\left(-\frac{1}{2}\frac{L^2x^2}{1-x^2}\right). \tag{27}$$

Because this is *symmetric* with respect to the origin, we can improve efficiency by computing a Chebyshev expansion only on $x \in [0, 1]$. The function has only a single finite root on the whole positive real axis at $y = 1/\sqrt{2}$, which is equivalent to $x = 1/\sqrt{3}$ in the transformed coordinate.

Figure 6 shows that the error decreases exponentially with N, the degree of the Chebyshev approximation, with oscillations in N superimposed. At $N = 70$, the roundoff error in the Chebyshev-to-powers conversion finally asserts itself so that further increases in N worsen the error compared to $N = 70$ where the root is approximated to a relative error of about 1 part in six million. This roundoff problem

can be suppressed, as in earlier examples, by using the "degree-doubling" conversion instead.

The accuracy is not as good for the finite interval example, which is hardly surprising. Nevertheless, with the added device of a mapping of the infinite interval into a finite interval, the Chebyshev-to-powers rootfinding method is successful.

9. Chebyshev methods and other "black box" rootfinders. The rootfinders of Kavvadias and Vrahatis [10], Kavvadias, Makri, and Vrahatis [11], and Smiley and Chun [13] both claim to be "black boxes," that is, to find the roots of a transcendental $f(x)$ without (i) requiring user input except for a subroutine to evaluate $f(x)$ and (ii) requiring the user to understand their algorithms—unnecessary because of their reliability. Both methods are a kind of "bisection-with-macho"; after the interval is subdivided, tests are applied to exclude subintervals which are root-free. Subdivision-and-test is then recursively applied until all the roots are isolated in sufficiently narrow intervals. Kavvadias and Vrahatis test for roots by numerically evaluating the Kronecker–Picard integral, whose value is the number of roots; Smiley and Chun use the cheaper but less precise criterion that if the Lipschitz constant L, defined by $L = \max(|f(x) - f(y)|/|x - y|)$ for all x, y on the interval is such that $|f(a)|, |f(b)| > L|b - a|$, then the interval $x \in [a, b]$ must be root-free. Refinements such as Newton–Ralphson iteration in the "end game" (when a root has been isolated within a small interval) and local Lipschitz constants are used to accelerate convergence for both methods.

We shall not attempt detailed comparisons between our algorithm and theirs. The efficiency of both Kronecker–Picard and Lipschitz-test algorithms strongly depend upon subdivision strategies, numerical quadrature tactics, Lipschitz constant approximation schemes, and so on. These black boxes will significantly improve as further experience allows better "tuning."

Instead, we will merely note that these algorithms require a large number of evaluations of $f(x)$ because of the repeated subdivisions and also the numerical quadratures or Lipschitz constant approximations. In principle, the cost of these evaluations could be dramatically reduced by replacing $f(x)$ by its Chebyshev interpolant. It follows that our ideas are perhaps complementary rather than competitive with subdivide-and-test methods.

However, our algorithm is completely self-contained. Kronecker–Picard and Lipschitz-test algorithms are useful in an $f \to Chebyshev$ approach only if these algorithms are superior to polynomial rootfinders. Are they? Alternatively, subdivide-and-test methods will fail to benefit from replacement of f by its Chebyshev proxy only for problems where f is cheap to evaluate and the subdivide-and-test methods converge faster than polynomial rootfinders. Such comparisons are highly problem-dependent and also implementation-dependent. We must leave these as open research questions.

10. Summary and open problems. Our main conclusion is that by "Chebyshevizing" a function $f(x)$, that is, by replacing $f(x)$ by its Chebyshev interpolant, the availability of robust polynomial rootfinders can be leveraged into reliable software for finding the roots of a *smooth, analytic* but otherwise *arbitrary* function $f(x)$ on a given *real interval*. Our Matlab subfunction that computes the roots using the convert-to-powers method has only 45 executable statements, and the degree-doubling algorithm is even shorter. Automation of subdivision would require a few additional lines, but the overall algorithms are commendably simple: all the complexity is in the polynomial rootfinder, which the user borrows from a software library.

By using an unbounded-interval-to-finite-interval mapping, the method easily generalizes to finding the real-valued roots of a function over the entire real axis if these roots are finite in number and the function is sufficiently smooth as $|x| \to \infty$ so that the transformed function is C^∞.

The "degree-doubling" method converts the truncated Chebyshev series of degree N into an ordinary polynomial of degree $2N$ whose coefficients are the same as the Chebyshev coefficients. This completely eliminates the ill-conditioning problem. However, because the degree is doubled, the computational cost is greater than for the convert-to-powers algorithm.

The convert-to-powers method has the flaw that it is mildly ill-conditioned. This difficulty can be cured by restricting N to moderate degree (less than 18) and subdividing the original target interval into as many subintervals as needed so that $f(x)$ is well-approximated by a Chebyshev series of restricted degree on each subinterval. However, degree restriction and subdivision are annoying complications. The reward is that the polynomial rootfinder is only asked to solve a polynomial of degree N rather than $2N$.

The numerical examples show that the convert-to-powers method is not as ill-conditioned as the norms of the conversion matrices would indicate. The reason is that, for a given problem, the true condition number depends upon the rate of convergence of the Chebyshev series as well as upon the matrix norms. The most extreme example is when $f(x)$ is a polynomial of finite degree k so that all Chebyshev coefficients a_j are zero for $j > k$. In this case, only the upper left $(k+1)/2 \times (k+1)/2$ blocks of the transformation matrices have anything to operate on. The effective condition number is determined by these blocks and not by the actual size of N, which may be much larger. For the Bessel function example, the Chebyshev series of J_0 is not identically zero for large degree, but the exponentially fast decrease of the Chebyshev coefficients does drastically reduce the effective condition number. An open problem is to develop a refined f-dependent condition number that takes the rate of Chebyshev convergence into account.

Because of the competing virtues and flaws of well-conditioned versus mildly ill-conditioned, fast versus slow, it is not possible to anoint either the degree-doubling or convert-to-powers algorithm as the "best" choice. What can be said is that both work well.

A minor unsolved problem, discussed at length in [5] but not here, is to find a good multiplicative scaling function when $f(x)$ varies by many orders of magnitude on the search interval $x \in [a, b]$. Because Chebyshev expansions are highly uniform in *absolute* error, there may be annoyingly large *relative* errors when $f(x)$ is badly scaled in the sense of having huge maxima on some parts of the interval but only tiny peaks and valleys on other subintervals. In theory, this difficulty can always be solved by subdividing the interval into subintervals and applying the algorithm on each subdomain.

The major unsolved problem is to find a good *direct* way to find all the roots of a polynomial on a real interval when the polynomial is defined by its Chebyshev coefficients without prior conversion to powers-of-x form. If such an algorithm could be found, then both the convert-to-powers and degree-doubling procedures become unnecessary.

Seen through the lens of Chebyshev polynomial series, there is no such thing as a nonpolynomial function. Every $f(x)$ is a truncated Chebyshev series in disguise. From the Chebyshev perspective, it is as easy to simultaneously find all roots of a

function $f(x)$ on a real interval, whether simple zeros or multiple roots, as it is for a polynomial.

Appendix A. Chebyshev interpolation of a function $f(x)$.

Goal. Compute a Chebyshev series, including terms up to and including T_N, on the interval $x \in [a, b]$.

Step 1. Create the interpolation points (Lobatto grid):

$$x_k \equiv \frac{b-a}{2} \cos\left(\pi \frac{k}{N}\right) + \frac{b+a}{2}, \qquad k = 0, 1, 2, \ldots, N. \tag{28}$$

Step 2. Compute the elements of the $(N+1) \times (N+1)$ interpolation matrix.

Define $p_j = 2$ if $j = 0$ or $j = N$ and $p_j = 1, j \in [1, N-1]$. Then the elements of the interpolation matrix are

$$I_{jk} = \frac{2}{p_j\, p_k\, N} \cos\left(j\pi \frac{k}{N}\right). \tag{29}$$

Step 3. Compute the grid-point values of $f(x)$, the function to be approximated:

$$f_k \equiv f(x_k), \qquad k = 0, 1, \ldots, N. \tag{30}$$

Step 4. Compute the coefficients through a vector-matrix multiply:

$$a_j = \sum_{k=0}^{N} I_{jk}\, f_k, \qquad j = 0, 1, 2, \ldots, N. \tag{31}$$

The approximation is

$$f \approx \sum_{j=0}^{N} a_j T_j\left(\frac{2x-(b+a)}{b-a}\right) = \sum_{j=0}^{N} a_j \cos\left\{j \arccos\left(\frac{2x-(b+a)}{b-a}\right)\right\}. \tag{32}$$

REFERENCES

[1] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[2] J. P. Boyd, *Exponentially convergent Fourier/Chebyshev quadrature schemes on bounded and infinite intervals*, J. Sci. Comput., 2 (1987), pp. 99–109.

[3] J. P. Boyd, *Orthogonal rational functions on a semi-infinite interval*, J. Comput. Phys., 70 (1987), pp. 63–88.

[4] J. P. Boyd, *The rate of convergence of Fourier coefficients for entire functions of infinite order with application to the Weideman-Cloot sinh-mapping for pseudospectral computations on an infinite interval*, J. Comput. Phys., 110 (1994), pp. 360–372.

[5] J. P. Boyd, *A Chebyshev polynomial interval-searching method ("Lanczos economization") for solving a nonlinear equation with application to the nonlinear eigenvalue problem*, J. Comput. Phys., 118 (1995), pp. 1–8.

[6] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover, New York, 2001.

[7] L. M. Delves and J. N. Lyness, *On numerical integration round a closed contour*, Math. Comp., 21 (1967), pp. 561–577.

[8] W. Gautschi, *On the condition number of algebraic equations*, Numer. Math., 21 (1973), pp. 405–424.

[9] K. O. Geddes and J. C. Mason, *Polynomial approximation by projections on the unit circle*, SIAM J. Numer. Anal., 12 (1975), pp. 111–120.

[10] D. J. Kavvadias and M. N. Vrahatis, *Locating and computing all the simple roots and extrema of a function*, SIAM J. Sci. Comput., 17 (1996), pp. 1232–1248.

[11] D. J. Kavvadias, F. S. Makri, and M. N. Vrahatis, *Locating and computing arbitrarily distributed zeros*, SIAM J. Sci. Comput., 21 (1999), pp. 954–969.
[12] C. Lanczos, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1956. Reprinted by Dover (1988).
[13] M. W. Smiley and C. Chun, *An algorithm for finding all solutions of a nonlinear system*, J. Comput. Appl. Math., 137 (2001), pp. 293–315.
[14] J. R. Winkler, *Condition numbers of a root*, Appl. Numer. Math., 38 (2001), pp. 275–285.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1683–1697

NONCONFORMING FINITE ELEMENT ANALYSIS FOR A PLATE CONTACT PROBLEM*

WEIMIN HAN† AND LIE-HENG WANG‡

Abstract. In this paper we analyze nonconforming finite element methods for solving a fourth order elliptic variational inequality of the second kind arising in a plate frictional contact problem. The variational inequality involves a nondifferentiable term due to the frictional contact. Optimal order error estimates are derived for both continuous and discontinuous nonconforming finite elements.

Key words. nonconforming finite element method, elliptic variational inequality of fourth order, plate frictional contact problem, optimal order error estimate

AMS subject classifications. 65N30, 74S05

PII. S0036142901390731

1. Introduction.

Variational inequalities form an important family of nonlinear boundary value or initial-boundary value problems. Interest in variational inequalities originates in applications from mechanics and physics. A partial list of the applications that lead to variational inequalities include the following: contact mechanics, non-Newtonian fluid flows such as Bingham fluids, obstacle problems, optimal control, plasticity, Stefan problems, unilateral problems, and so on. An early comprehensive reference on the topic is [8], where many problems in mechanics and physics are formulated and studied in the framework of variational inequalities. More recent references on the mathematical analysis of variational inequalities include [1, 10, 20, 22, 23]. Comprehensive references concerning the numerical analysis of variational inequalities, especially those arising in mechanical problems, include [12, 13, 14, 15, 16, 17, 19]. These references focus on numerical analysis for variational inequalities involving second order differential operators. Numerical study of fourth order variational inequalities is also available from some of these references; e.g., a duality approach based on conforming finite elements is analyzed in [16].

Nonconforming finite element methods are a natural choice in employing finite element methods for solving fourth order boundary value problems since the smoothness requirement on finite element functions is weakened. An early reference on the mathematical analysis of nonconforming finite element methods for the plate bending problem is [21]. Application of nonconforming finite element methods is not limited to fourth order problems; they offer more efficient solution algorithms for numerous other problems (cf. [3, p. 208]). Convergence and error estimation of nonconforming finite element methods are more involved compared to that of conforming finite element methods. A patch test was proposed and is widely used by engineers for convergence analysis of nonconforming finite element methods (cf. [2, 18]). However, it

*Received by the editors June 12, 2001; accepted for publication (in revised form) April 16, 2002; published electronically November 14, 2002.

http://www.siam.org/journals/sinum/40-5/39073.html

†Department of Mathematics, University of Iowa, Iowa City, IA 52242 (whan@math.uiowa.edu). The work of this author was supported by K. C. Wong Education Foundation, and the National Science Foundation under grants DMS-9874015 and DMS-0106781.

‡State Key Laboratory of Scientific and Engineering Computing, ICMSEC, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (wlh@lsec.cc.ac.cn). The work of this author was supported by the National Natural Science Foundation of China.

is shown in [26] that the patch test is neither a necessary nor a sufficient condition for convergence. One finds in [26] a rigorous necessary and sufficient condition for convergence of nonconforming finite element solutions to variational equations of some boundary value problems. Some further developments along this line can be found in [25, 30], where convergence conditions are studied which are easier to examine. A summary account of nonconforming finite element methods can be found in [5] or, more recently, in [6]. In particular, in these latter references, one can find some discussions of the four nonconforming finite elements mentioned later in this paper: the continuous nonconforming elements of the Zienkiewicz triangle and Adini's rectangle, and the discontinuous nonconforming elements of Morley's triangle and the Fraeijs De Veubeke triangle.

In this paper, we derive error estimates for continuous and discontinuous nonconforming finite elements in solving a fourth order elliptic variational inequality of the second kind. A variational inequality of the second kind is featured by the presence of nondifferentiable terms in the formulation. Variational inequalities of the second kind are commonly seen in frictional contact problems. In this paper we adopt a plate frictional contact problem as our model fourth order variational inequality of the second kind for error analysis of nonconforming finite element methods; the ideas and results reported here can be extended to nonconforming finite element methods for other fourth order elliptic variational inequalities of the second kind. Literature on nonconforming finite element methods for fourth order variational inequalities is rather small at the moment. The only papers on this topic we know of are [27, 28, 29]. Note that in these papers the variational inequalities being approximated are of the first kind; i.e., they are imposed over convex sets, and no nondifferentiable terms are involved. To analyze nonconforming finite elements for fourth order variational inequalities of the second kind, we need to employ new techniques.

The paper is organized as follows. In section 2, we introduce the plate contact problem and show some properties for the solution of the problem. In section 3, we present an abstract result for nonconforming methods that will be used in deriving error estimates later in the paper. Sections 4 and 5 are devoted to error estimation of continuous and discontinuous nonconforming finite element methods for the plate contact problem, respectively.

2. The plate contact problem. Consider a thin flat plate $\Omega \times (-d/2, d/2)$, where $\Omega \subset \mathbb{R}^2$, $d > 0$ is the thickness of the plate and is assumed to be small. Assume the three-dimensional material is isotropic, linearly elastic with Poisson's ratio $\nu \in (0, 1/2)$ and Young's modulus $E > 0$. The plate is subject to a normal force of density $D_0 f(\boldsymbol{x})$ with the stiffness coefficient of the plate

$$D_0 = \frac{E\, d^3}{12\,(1-\nu^2)}.$$

Denote by $u = u(\boldsymbol{x})$, $\boldsymbol{x} \in \overline{\Omega}$, the vertical deflection of the plate. Let the boundary $\Gamma = \partial\Omega$ of the plate be decomposed into three mutually disjoint parts: $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2 \cup \overline{\Gamma}_3$ such that Γ_1, Γ_2, and Γ_3 are relatively open, $\overline{\Gamma}_1 \cap \overline{\Gamma}_3 = \emptyset$, and meas $(\Gamma_1) > 0$. The boundary is assumed to be Lipschitz continuous, and the unit outward normal vector is denoted by $\boldsymbol{n} = (n_1, n_2)^T$. The tangential vector is $\boldsymbol{\tau} = (\tau_1, \tau_2)^T$ with $\tau_1 = -n_2$, $\tau_2 = n_1$. Both $\boldsymbol{n}$ and $\boldsymbol{\tau}$ exist a.e. on Γ. Assume the plate is clamped on Γ_1:

$$u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_1,$$

is free on Γ_2:

$$M(u) = N(u) = 0 \quad \text{on } \Gamma_2,$$

and is in frictional contact with a rigid foundation on Γ_3. Here,

$$M(u) = -\Delta u + (1-\nu)\,\partial_{\tau\tau} u,$$
$$N(u) = \partial_n \Delta u + (1-\nu)\,\partial_\tau(\partial_{n\tau} u).$$

Notice that for a smooth function u, $M(u)$ and $N(u)$ are defined a.e. on Γ. The quantity $M(u)$ can be interpreted as the tangential moment, while $-N(u)$ represents a force. Here and throughout the paper, we use the following notations:

$$\partial_{11} u = \frac{\partial^2 u}{\partial x_1^2}, \quad \partial_{12} u = \frac{\partial^2 u}{\partial x_1 \partial x_2}, \quad \partial_n u = \frac{\partial u}{\partial n}, \quad \partial_\tau u = \frac{\partial u}{\partial \tau}, \quad \cdots .$$

Introduce the function space

$$V = \{\, v \in H^2(\Omega) \mid v = \partial_n v = 0 \text{ on } \Gamma_1 \,\}. \tag{2.1}$$

Over the space V, we define a bilinear form

$$a(u,v) = \int_\Omega \left[\Delta u\, \Delta v + (1-\nu)\,(2\,\partial_{12} u\, \partial_{12} v - \partial_{11} u\, \partial_{22} v - \partial_{22} u\, \partial_{11} v)\right] dx, \tag{2.2}$$

and a functional

$$j(v) = \int_{\Gamma_3} g\,|v|\, ds, \tag{2.3}$$

where g is given. For the data of the problem, we assume

$$f \in L^2(\Omega), \quad g \in L^2(\Gamma_3),\ g > 0 \text{ a.e. on } \Gamma_3. \tag{2.4}$$

We will use the notation

$$(f,v) = \int_\Omega f\, v\, dx.$$

The plate frictional contact problem is defined through a minimal energy principle:

$$u \in V,\ J(u) = \inf_{v \in V} J(v),$$

where

$$J(v) = \frac{1}{2}\, a(v,v) + j(v) - (f,v).$$

The quantity $D_0 J(v)$ is the total energy, and $j(v)$ is the contribution from the frictional contact. It is easy to show that the minimization problem is equivalent to the following variational inequality.

PROBLEM 2.1. *Find $u \in V$ such that*

$$a(u, v-u) + j(v) - j(u) \geq (f, v-u) \quad \forall\, v \in V. \tag{2.5}$$

Wellposedness of Problem 2.1 follows from a standard argument.

THEOREM 2.2. *Problem* 2.1 *has a unique solution.*

Proof. Since meas $(\Gamma_1) > 0$, the bilinear form is coercive on V:

$$a(v,v) \geq \alpha \|v\|_V^2 \quad \forall\, v \in V.$$

We also observe that over the space V, $a(\cdot,\cdot)$ is continuous, $j(\cdot)$ is continuous and convex, and f defines a linear continuous functional. Thus Problem 2.1 is an elliptic variational inequality of the second kind and has a unique solution (cf. [12]). ∎

To obtain the corresponding strong form of the boundary value problem, we assume the solution u is smooth (say, $u \in C^4(\overline{\Omega})$). For any $v \in H^2(\Omega)$, we have

$$\int_\Omega \Delta^2 u\, v\, dx = \int_\Omega \Delta u\, \Delta v\, dx + \int_\Gamma \partial_n \Delta u\, v\, ds - \int_\Gamma \Delta u\, \partial_n v\, ds. \tag{2.6}$$

It is easy to verify the equality

$$\int_\Omega (2\,\partial_{12} u\,\partial_{12} v - \partial_{11} u\,\partial_{22} v - \partial_{22} u\,\partial_{11} v)\, dx = \int_\Gamma (-\partial_{\tau\tau} u\,\partial_n v + \partial_{n\tau} u\,\partial_\tau v)\; ds.$$

If the boundary Γ is smooth, we further have

$$\int_\Omega (2\,\partial_{12} u\,\partial_{12} v - \partial_{11} u\,\partial_{22} v - \partial_{22} u\,\partial_{11} v)\, dx = -\int_\Gamma (\partial_{\tau\tau} u\,\partial_n v + \partial_\tau(\partial_{n\tau} u)\, v)\; ds. \tag{2.7}$$

Then from (2.6) we have

$$a(u,v) = \int_\Omega \Delta^2 u\, v\, dx - \int_\Gamma N(u)\, v\, ds - \int_\Gamma M(u)\, \partial_n v\, ds \quad \forall\, v \in H^2(\Omega). \tag{2.8}$$

Using this relation in the variational inequality (2.5), we can follow a standard argument (cf., e.g., [8, 19]) to conclude that u satisfies the relations

$$\Delta^2 u = f \quad \text{in } \Omega, \tag{2.9}$$

$$u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_1, \tag{2.10}$$

$$M(u) = N(u) = 0 \quad \text{on } \Gamma_2, \tag{2.11}$$

$$\left.\begin{array}{l} M(u) = 0, \quad |N(u)| \leq g, \\ |N(u)| < g \Longrightarrow u = 0, \\ |N(u)| = g \Longrightarrow u = \lambda\, N(u) \text{ for some } \lambda \geq 0 \end{array}\right\} \text{on } \Gamma_3. \tag{2.12}$$

This is the strong form of the plate contact problem studied in [8]. We comment that $g > 0$ can be interpreted as the frictional bound.

When the boundary Γ is only piecewise smooth, the right-hand side of the relation (2.7) needs to be replaced by

$$\sum_P \big(\partial_{n\tau} u(P_-) - \partial_{n\tau} u(P_+)\big)\, v(P) - \int_\Gamma (\partial_{\tau\tau} u\,\partial_n v + \partial_\tau(\partial_{n\tau} u)\, v)\; ds,$$

where P is any corner point on the boundary and $\partial_{n\tau} u(P_-)$ and $\partial_{n\tau} u(P_+)$ are the left and right limiting values of $\partial_{n\tau} u$ at P along Γ directed counterclockwise. Then the relations (2.9)–(2.12) are to be supplemented with continuity conditions of the form

$$\partial_{n\tau} u(P_-) = \partial_{n\tau} u(P_+).$$

Such conditions can be interpreted as "corner force conditions" (cf. [7]).

The main purpose of the paper is to analyze nonconforming finite element methods for Problem 2.1. For this, we need a characterization of the solution of Problem 2.1, following an idea found in [13]. Let

$$\Lambda = \{\, \mu \in L^\infty(\Gamma_3) \mid |\mu| \le 1 \text{ a.e. on } \Gamma_3 \,\}.$$

THEOREM 2.3. *A function u is a solution of Problem* 2.1 *if and only if there exists $\lambda \in \Lambda$ such that*

$$a(u,v) + \int_{\Gamma_3} g\,\lambda\, v\, ds = (f,v) \quad \forall\, v \in V, \tag{2.13}$$

$$\lambda\, u = |u| \quad \text{a.e. on } \Gamma_3. \tag{2.14}$$

Proof. Suppose u is a solution of Problem 2.1. By taking $v = 0$ and $2\,u$ in (2.5), we obtain

$$a(u,u) + j(u) = (f,u). \tag{2.15}$$

Then, from (2.5),

$$a(u,v) + j(v) \ge (f,v) \quad \forall\, v \in V. \tag{2.16}$$

It is easy to see that (2.5) is equivalent to (2.15) and (2.16). From (2.16), we get

$$(f,v) - a(u,v) \le j(v) \quad \forall\, v \in V.$$

Replacing v by $-v$ in this inequality, we obtain

$$(f,v) - a(u,v) \ge -j(v) \quad \forall\, v \in V.$$

Therefore,

$$|(f,v) - a(u,v)| \le j(v) \quad \forall\, v \in V. \tag{2.17}$$

Let γ be the trace operator defined on V and denote $H_{\Gamma_3} = \gamma(V)|_{\Gamma_3}$ with the norm

$$\|v\|_{H_{\Gamma_3}} = \inf\{\, \|w\|_V \mid w \in V,\, w|_{\Gamma_3} = v \,\}.$$

Then, from (2.17), we see that $H_{\Gamma_3} \ni v \mapsto (f,v) - a(u,v)$ defines a linear mapping on H_{Γ_3}; here, we use the same symbol v for a function from V with the trace v on Γ_3. Thus $\ell(v) = (f,v) - a(u,v)$ is a linear mapping on H_{Γ_3} and, from (2.17),

$$|\ell(v)| \le \int_{\Gamma_3} g\,|v|\, ds \quad \forall\, v \in H_{\Gamma_3}.$$

Obviously, $H_{\Gamma_3} \subset L^1(\Gamma_3)$. By the Hahn–Banach theorem, the linear functional ℓ can be the extended to the space $L^1(\Gamma_3)$, and we have the existence of $\lambda \in \Lambda$ such that

$$\ell(v) = \int_{\Gamma_3} g\,\lambda\, v\, ds \quad \forall\, v \in L^1(\Gamma_3).$$

Therefore, (2.13) holds. Using (2.15), we then have

$$\int_{\Gamma_3} g\,(\lambda\, u - |u|)\, ds = 0.$$

Since $|\lambda| \leq 1$ a.e. on Γ_3, we have the relation (2.14).

Proof of the converse statement is easy and is hence omitted. □

By comparing (2.8) with (2.13) and using the equality boundary conditions of the solution u, we find that

$$g\,\lambda = N(u) \quad \text{on } \Gamma_3.$$

Thus, the "Lagrange multiplier" λ can be interpreted as a scaled shearing force.

Note that the relations (2.9)–(2.12) are valid only if the solution $u \in V$ of Problem 2.1 is smooth, e.g., $u \in C^4(\overline{\Omega})$. Consequently, these relations cannot be used in error analysis. In finite element error analysis, we need a reasonable solution regularity stronger than $u \in V$. Such a regularity result for Problem 2.1 does not seem to be available in the current literature. In this paper, we will assume

$$u \in H^3(\Omega). \tag{2.18}$$

Now let us derive some relations for the solution $u \in V$ of Problem 2.1. In (2.13), we let $v \in C_0^\infty(\Omega)$ to obtain

$$\Delta^2 u = f \quad \text{in the sense of distributions.}$$

Since $f \in L^2(\Omega)$, we actually have

$$\Delta^2 u = f \quad \text{in } L^2(\Omega), \tag{2.19}$$

and then also

$$\Delta^2 u = f \quad \text{a.e. in } \Omega. \tag{2.20}$$

Since $\Delta^2 u \in L^2(\Omega)$ and $u \in H^3(\Omega)$, we can define $\partial_n \Delta u \in H^{-1/2}(\Gamma)$ by the relation (cf., e.g., [11])

$$\langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} = \int_\Omega \left[\Delta^2 u\, v + \nabla(\Delta u)\cdot\nabla v\right] dx \quad \forall\, v \in H^1(\Omega). \tag{2.21}$$

For the bilinear form (2.2), we then have

$$\begin{aligned} a(u,v) &= \int_\Omega \Delta^2 u\, v\, dx - \langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} + \int_\Gamma \Delta u\, \partial_n v\, ds \\ &\quad + (1-\nu)\int_\Gamma (-\partial_{\tau\tau} u\, \partial_n v + \partial_{n\tau} u\, \partial_\tau v)\, ds \\ &= (f,v) - \int_\Gamma M(u)\, \partial_n v\, ds - \langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} + (1-\nu)\int_\Gamma \partial_{n\tau} u\, \partial_\tau v\, ds. \end{aligned}$$

Thus by (2.13) we have

$$-\int_\Gamma M(u)\, \partial_n v\, ds - \langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} + (1-\nu)\int_\Gamma \partial_{n\tau} u\, \partial_\tau v\, ds + \int_{\Gamma_3} g\,\lambda\, v\, ds = 0 \quad \forall\, v \in V. \tag{2.22}$$

By a standard procedure (cf. [8]), it can then be established that

$$M(u) = 0 \quad \text{a.e. on } \Gamma_2 \cup \Gamma_3. \tag{2.23}$$

Then from (2.22) we obtain

$$-\langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} + (1-\nu)\int_\Gamma \partial_{n\tau} u\, \partial_\tau v\, ds + \int_{\Gamma_3} g\,\lambda\, v\, ds = 0 \quad \forall\, v \in V. \tag{2.24}$$

Now the closure of V in $H^1(\Omega)$ is

$$H^1_{\Gamma_1}(\Omega) = \{\, v \in H^1(\Omega) \mid v = 0 \text{ a.e. on } \Gamma_1 \,\}.$$

Denote

$$\tilde{H}^1_{\Gamma_1}(\Omega) = \{\, v \in H^1_{\Gamma_1}(\Omega) \mid \partial_\tau v \in L^2(\Gamma) \,\}.$$

Then from (2.24) we conclude that

$$-\langle \partial_n \Delta u, v\rangle_{1/2,\Gamma} + (1-\nu)\int_\Gamma \partial_{n\tau} u\, \partial_\tau v\, ds + \int_{\Gamma_3} g\,\lambda\, v\, ds = 0 \quad \forall\, v \in \tilde{H}^1_{\Gamma_1}(\Omega). \tag{2.25}$$

3. An abstract error estimate. Let $\{\mathcal{T}_h\}_h$ be a family of finite element partitions of the domain $\overline{\Omega}$. Here $h \to 0+$ is a discretization parameter. A typical element in $\mathcal{T}_h$ is denoted by T. Let $\{V_h\}_h$ be a family of corresponding finite element spaces used to approximate the space V. We consider the case of nonconforming approximation. Thus, in general, $V_h \not\subset V$. Then the discrete approximation problem is the following.

PROBLEM 3.1. *Find $u \in V_h$ such that*

$$a_h(u_h, v_h - u_h) + j(v_h) - j(u_h) \geq (f, v_h - u_h) \quad \forall\, v_h \in V_h, \tag{3.1}$$

where the discrete bilinear form is

$$a_h(u,v) = \sum_{T\in\mathcal{T}_h} \int_T \left[\Delta u\, \Delta v + (1-\nu)\,(2\,\partial_{12}u\,\partial_{12}v - \partial_{11}u\,\partial_{22}v - \partial_{22}u\,\partial_{11}v)\right] dx. \tag{3.2}$$

Assume

$$\|v_h\|_h = \left\{ \sum_{T\in\mathcal{T}_h} |v_h|^2_{2,T} \right\}^{1/2}, \quad v_h \in V_h,$$

is a norm on V_h. Then the bilinear form (3.2) is coercive on V_h. Obviously, $a_h(\cdot,\cdot)$ is continuous:

$$|a_h(u,v)| \leq M\, \|u_h\|_h \|v_h\|_h \quad \forall\, u_h, v_h \in V + V_h.$$

We also observe that $j(\cdot)$ is continuous and convex on V_h, and f defines a linear continuous functional on V_h. Therefore, Problem 3.1 has a unique solution.

The following abstract error estimate is inspired by Falk's work [9] and the work of Brezzi, Hager, and Raviart [4]. It plays an important role in error analysis for the approximation of the variational inequality and can be viewed as an extension of the Strang lemma (cf. [5]) for variational equations to variational inequalities.

THEOREM 3.2. *For the solutions of the Problems* 2.1 *and* 3.1, *we have the inequality*

$$\|u - u_h\|^2_h \leq c \inf_{v_h \in V_h} \left\{ \|u - v_h\|^2_h + R_h(v_h, u_h) \right\}, \tag{3.3}$$

where

$$R_h(v_h, u_h) = a_h(u, v_h - u_h) + j(v_h) - j(u_h) - (f, v_h - u_h) \tag{3.4}$$

is a discrete residual.

Proof. For any $v_h \in V_h$, we have

$$\begin{aligned}\alpha \|u_h - v_h\|_h^2 &\le a_h(u_h - v_h, u_h - v_h)\\ &= a_h(u - v_h, u_h - v_h) + a_h(u_h - u, u_h - v_h)\\ &\le M \|u - v_h\|_h \|u_h - v_h\|_h + a_h(u_h, u_h - v_h) - a_h(u, u_h - v_h)\\ &\le M \|u - v_h\|_h \|u_h - v_h\|_h + R_h(v_h, u_h),\end{aligned}$$

where in the last step we used the defining inequality (3.1). Using the inequality

$$M \|u - v_h\|_h \|u_h - v_h\|_h \le \frac{\alpha}{2} \|u_h - v_h\|_h^2 + \frac{M^2}{2\,\alpha} \|u - v_h\|_h^2$$

we obtain

$$\|u_h - v_h\|_h^2 \le c \{\|u - v_h\|_h^2 + R_h(v_h, u_h)\}.$$

Now the relation (3.3) follows from

$$\|u - u_h\|_h \le \|u - v_h\|_h + \|u - v_h\|_h$$

and the arbitrariness of $v_h \in V_h$. □

4. Continuous nonconforming finite element approximation. We consider some continuous nonconforming plate elements in this section. Let $\{\mathcal{T}_h\}_h$ be a family of regular triangulation of $\overline{\Omega}$, and let $\{V_h\}_h \subset C^0(\overline{\Omega})$ be a corresponding family of nonconforming finite element subspaces of V. We assume

$$\Big| \sum_T \int_{\partial T} w\, \partial_n v_h \, ds \Big| \le c\, h \, \|w\|_{1,\Omega} \|v_h\|_h \quad \forall\, v_h \in V_h, \tag{4.1}$$

and the finite element interpolation error estimate

$$\|w - \Pi_h w\|_h \le c\, h \, \|w\|_{3,\Omega} \quad \forall\, w \in V \cap H^3(\Omega). \tag{4.2}$$

Here $\Pi_h w \in V_h$ denotes the finite element interpolant of w.

THEOREM 4.1. *Assume* (2.18), (4.1), *and* (4.2). *Then we have the error estimate*

$$\|u - u_h\|_h \le c\, h \, (\|u\|_{3,\Omega} + h^{1/4} \|g\|_{0,\Gamma_3}). \tag{4.3}$$

Proof. Let us first estimate the terms involved in the residual $R_h(v_h, u_h)$. We have

(4.4)

$$\begin{aligned}a_h(u, u_h - v_h) &= \sum_T \int_T \big\{\Delta u\, \Delta(u_h - v_h) + (1 - \nu)\,(2\, \partial_{12} u\, \partial_{12}(u_h - v_h)\\ &\qquad - \partial_{11} u\, \partial_{22}(u_h - v_h) - \partial_{22} u\, \partial_{11}(u_h - v_h))\big\}\, ds\\ &= -\sum_T \int_T \nabla(\Delta u) \cdot \nabla(u_h - v_h)\, dx + \sum_T \int_{\partial T} \Delta u\, \partial_n(u_h - v_h)\, ds\\ &\quad + (1 - \nu) \sum_T \int_{\partial T} \big\{ - \partial_{\tau\tau} u\, \partial_n(u_h - v_h) + \partial_{n\tau} u\, \partial_\tau(u_h - v_h)\big\}\, ds.\end{aligned}$$

Since $V_h \subset C(\overline{\Omega})$, we have $u_h, v_h \in H^1(\Omega)$ and

(4.5)
$$-\sum_T \int_T \nabla(\Delta u)\cdot\nabla(u_h - v_h)\,dx = -\int_\Omega \nabla(\Delta u)\cdot\nabla(u_h - v_h)\,dx$$
$$= \int_\Omega \Delta^2 u\,(u_h - v_h)\,dx - \langle \partial_n \Delta u, u_h - v_h\rangle_{1/2,\Gamma}.$$

Then

$$\begin{aligned}
&(f, u_h - v_h) - a_h(u, u_h - v_h)\\
&\quad = -\sum_T \int_{\partial T} \left[\Delta u - (1-\nu)\,\partial_{\tau\tau} u\right] \partial_n(u_h - v_h)\,ds + \langle \partial_n \Delta u, u_h - v_h\rangle_{1/2,\Gamma}\\
&\qquad - (1-\nu)\sum_T \int_{\partial T} \partial_{n\tau} u\,\partial_\tau(u_h - v_h)\,ds\\
&\quad = -\sum_T \int_{\partial T} M(u)\,\partial_n(u_h - v_h)\,ds - (1-\nu)\sum_T \sum_{\substack{\gamma\subset\partial T\\ \gamma\not\subset\Gamma}} \int_\gamma \partial_{n\tau} u\,\partial_\tau(u_h - v_h)\,ds\\
&\qquad - (1-\nu)\sum_T \sum_{\substack{\gamma\subset\partial T\\ \gamma\subset\Gamma}} \int_\gamma \partial_{n\tau} u\,\partial_\tau(u_h - v_h)\,ds + \langle \partial_n \Delta u, u_h - v_h\rangle_{1/2,\Gamma}.
\end{aligned}$$

Since $u_h, v_h \in C(\overline{\Omega})$, we have

$$\sum_T \sum_{\substack{\gamma\subset\partial T\\ \gamma\not\subset\Gamma}} \int_\gamma \partial_{n\tau} u\,\partial_\tau(u_h - v_h)\,ds = 0.$$

Thus,

$$\begin{aligned}
(f, u_h - v_h) - a_h(u, u_h - v_h) = &-\sum_T \int_{\partial T} M(u)\,\partial_n(u_h - v_h)\,ds + \langle \partial_n \Delta u, u_h - v_h\rangle_{1/2,\Gamma}\\
&- (1-\nu)\int_\Gamma \partial_{n\tau} u\,\partial_\tau(u_h - v_h)\,ds.
\end{aligned}$$

Using the relation (2.25), we obtain

(4.6)
$$(f, u_h - v_h) - a_h(u, u_h - v_h) = -\sum_T \int_{\partial T} M(u)\,\partial_n(u_h - v_h)\,ds - \int_{\Gamma_3} g\,\lambda\,(u_h - v_h)\,ds.$$

Then

(4.7)
$$R_h(v_h, u_h) = \int_{\Gamma_3} g\,(|v_h| - \lambda\,v_h - |u_h| + \lambda\,u_h)\,ds - \sum_T \int_{\partial T} M(u)\,\partial_n(u_h - v_h)\,ds.$$

The last term on the right-hand side of (4.7) is estimated by (4.1):

$$\left|\sum_T \int_{\partial T} M(u)\,\partial_n(v_h - u_h)\,ds\right| \le c\,h\,\|u\|_{3,\Omega}\|v_h - u_h\|_h.$$

We now estimate the first term on the right-hand side of (4.7). Since $|\lambda| \le 1$ a.e. on Γ_3,

$$\int_{\Gamma_3} g\,(|v_h| - \lambda\, v_h - |u_h| + \lambda\, u_h)\, ds \le \int_{\Gamma_3} g\,(|v_h| - \lambda\, v_h)\, ds.$$

In the following, we choose $v_h = \Pi_h u$. We have

$$\begin{aligned}\int_{\Gamma_3} g\,(|\Pi_h u| - \lambda\,\Pi_h u)\, ds &= \int_{\Gamma_3} g\ (|\Pi_h u| - |u| - \lambda\,(\Pi_h u - u))\ ds \\ &\le 2\int_{\Gamma_3} g\,|u - \Pi_h u|\, ds \\ &\le 2\,\|g\|_{0,\Gamma_3}\|u - \Pi_h u\|_{0,\Gamma_3}.\end{aligned}$$

From [26], for any element side $\gamma \subset \Gamma_3$, denoting T the element that has the side γ, we have

$$\begin{aligned}\|u - \Pi_h u\|_{0,\gamma} &\le c\left(h^{-1}\|u - \Pi_h u\|_{0,T}^2 + h\,|u - \Pi_h u|_{1,T}^2\right)^{1/2} \\ &\le c\left(h^{-1}h^6|u|_{3,T}^2 + h\,h^4|u|_{3,T}^2\right)^{1/2} \\ &\le c\,h^{5/2}|u|_{3,T}.\end{aligned}$$

Thus,

$$\begin{aligned}\|u - \Pi_h u\|_{0,\Gamma_3} &= \left(\sum_{\gamma\subset\Gamma_3}\|u - \Pi_h u\|_{0,\gamma}^2\right)^{1/2} \\ &\le c\,h^{5/2}\left(\sum_{T:\partial T\cap\Gamma_3\ne\emptyset}|u|_{3,T}^2\right)^{1/2} \\ &\le c\,h^{5/2}|u|_{3,\Omega}.\end{aligned}$$

Summarizing, we have the bound

$$R_h(\Pi_h u, u_h) \le c\,h\,\|u\|_{3,\Omega}\|\Pi_h u - u_h\|_h + c\,h^{5/2}\|g\|_{0,\Gamma_3}\|u\|_{3,\Omega}.$$

So, from (3.3),

$$\|u - u_h\|_h^2 \le c\,\{\|u - \Pi_h u\|_h^2 + h\,\|u\|_{3,\Omega}\|\Pi_h u - u_h\|_h + h^{5/2}\|g\|_{0,\Gamma_3}\|u\|_{3,\Omega}\}.$$

The term $\|\Pi_h u - u_h\|_h$ is bounded by $\|\Pi_h u - u\|_h + \|u - u_h\|_h$. Using the interpolation error estimate (4.2), we then obtain the error estimate (4.3). □

One example of a continuous nonconforming finite element is the Zienkiewicz triangle. Assume $\overline{\Omega}$ is such that it is possible to split it into triangles with all sides parallel to three fixed directions. This property is valid if Ω is the union of rectangles with sides parallel to two fixed directions and right triangles with two sides parallel to the two fixed directions. Let $\{\mathcal{T}_h\}$ be a regular family of partitions of $\overline{\Omega}$ into such triangles. Then the Zienkiewicz triangle consists of piecewise incomplete polynomials of degree less than or equal to 3. On each triangle, the polynomial is determined by its values and the values of its two first order derivatives at the three vertices; for details, cf. [5]. For this element, we have (4.1) and (4.2) (cf. [24]).

Another example is Adini's rectangle. Assume $\overline{\Omega} \subset \mathbb{R}^2$ can be partitioned into rectangles (e.g., if Ω is the union of rectangles with sides parallel to two fixed directions). Let $\{\mathcal{T}_h\}_h$ be a regular family of partitions of $\overline{\Omega}$ into rectangles with sides parallel to the coordinate axes. Then Adini's rectangle is defined as a piecewise polynomial corresponding to the partition $\mathcal{T}_h$ such that, on each element, it is a polynomial from the space $\mathcal{P}_3(\mathbb{R}^2) + [x_1^3 x_2, x_1 x_2^3]$, with the values of function and of the two first partial derivatives with respect to x_1 and x_2 at the four vertices of the element as the finite element parameters. For the vertices on $\overline{\Gamma}_1$, the parameters are taken to be zero for V_h. Then, from [26], we have (4.1) and (4.2).

We conclude that for both the Zienkiewicz triangle and Adini's rectangle, the optimal order error estimate (4.3) holds.

5. Discontinuous nonconforming finite element approximation. In this section, we consider discontinuous nonconforming finite element approximations of the plate contact problem. Let $\{V_h\}_h \not\subset C^0(\overline{\Omega})$ be a family of nonconforming finite element subspaces of V corresponding to a regular family $\{\mathcal{T}_h\}_h$ of triangulations of $\overline{\Omega}$ such that the finite element functions are continuous at the vertices of the corresponding triangulation. We still assume (4.1) and (4.2).

THEOREM 5.1. *Assume* (2.18), (4.1), *and* (4.2). *Also assume the finite element functions are continuous at the vertices of the corresponding triangulation. Then we have the error estimate*

$$\|u - u_h\|_h \le c\, h \left\{ \|u\|_{3,\Omega} + h^{1/4} \|g\|_{0,\Gamma_3} + h\, \|f\|_{0,\Omega} \right\}. \tag{5.1}$$

Proof. Since $V_h \not\subset C(\overline{\Omega})$ implies $V_h \not\subset H^1(\Omega)$, we must modify the expression (4.5) as follows. Denote $w_h = u_h - v_h$ and let w_h^I be the continuous piecewise linear interpolant of w_h. Since $w_h^I \in C(\overline{\Omega})$, $w_h^I \in H^1(\Omega)$. First we write

$$\begin{aligned}
-\sum_T \int_T \nabla(\Delta u) \cdot \nabla(u_h - v_h)\, dx &= -\sum_T \int_T \nabla(\Delta u) \cdot \nabla w_h\, dx \\
&= -\sum_T \int_T \nabla(\Delta u) \cdot \nabla w_h^I\, dx \\
&\quad -\sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I)\, dx \\
&= \int_\Omega \Delta^2 u\, w_h^I\, dx - \langle \partial_n \Delta u, w_h^I \rangle_{1/2,\Gamma} \\
&\quad -\sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I)\, dx.
\end{aligned}$$

Then

$$\begin{aligned}
a_h(u, u_h - v_h) = \int_\Omega \Delta^2 u\, w_h^I\, dx - \langle \partial_n \Delta u, w_h^I \rangle_{1/2,\Gamma} - \sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I)\, dx \\
+ \sum_T \int_{\partial T} \Delta u\, \partial_n w_h ds + (1-\nu) \sum_T \int_{\partial T} (-\partial_{\tau\tau} u\, \partial_n w_h + \partial_{n\tau} u\, \partial_\tau w_h)\, ds.
\end{aligned}$$

Use the relation (2.20),

$$a_h(u, u_h - v_h) = (f, w_h^I) - \langle \partial_n \Delta u, w_h^I \rangle_{1/2,\Gamma} + \sum_T \int_{\partial T} \left\{ \Delta u - (1-\nu)\, \partial_{\tau\tau} u \right\} \partial_n w_h \, ds$$
$$+ (1-\nu) \sum_T \int_{\partial T} \partial_{n\tau} u \partial_\tau w_h \, ds - \sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I) \, dx.$$

Hence,

$$\begin{aligned}(f, u_h - v_h) - a_h(u, u_h - v_h) &= (f, w_h - w_h^I) + \sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I) \, dx + \langle \partial_n \Delta u, w_h^I \rangle_{1/2,\Gamma} \\ &\quad - \sum_T \int_{\partial T} M(u)\, \partial_n w_h \, ds - (1-\nu) \sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau w_h \, ds.\end{aligned}$$

Now

$$\sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau w_h \, ds = \sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau w_h^I \, ds + \sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau (w_h - w_h^I) \, ds.$$

For each side γ of the elements, define a piecewise constant projection operator $P_0^\gamma : L^1(\gamma) \to \mathbb{R}$ by

$$P_0^\gamma(v) = \frac{1}{|\gamma|} \int_\gamma v \, ds.$$

Since

$$\int_\gamma \partial_\tau (w_h - w_h^I) \, ds = 0,$$

we have

$$\begin{aligned}\sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau (w_h - w_h^I) \, ds &= \sum_T \sum_{\gamma \subset \partial T} \int_\gamma \partial_{n\tau} u \, \partial_\tau (w_h - w_h^I) \, ds \\ &= \sum_T \sum_{\gamma \subset \partial T} \int_\gamma (\partial_{n\tau} u - P_0^\gamma(\partial_{n\tau} u)) \, \partial_\tau (w_h - w_h^I) \, ds \\ &\le c\, h\, |u|_{3,\Omega} \|w_h\|_h.\end{aligned}$$

Using the fact $w_h^I \in C(\overline{\Omega})$ we have

$$\sum_T \int_{\partial T} \partial_{n\tau} u \, \partial_\tau w_h^I \, ds = \sum_{\gamma \subset \Gamma} \int_\gamma \partial_{n\tau} u \, \partial_\tau w_h^I \, ds.$$

Also,

$$(f, w_h - w_h^I) \le \|f\|_{0,\Omega} \|w_h - w_h^I\|_{0,\Omega} \le c\, h^2 \|f\|_{0,\Omega} \|w_h\|_h$$

and

$$\sum_T \int_T \nabla(\Delta u) \cdot \nabla(w_h - w_h^I) \le \sum_T |u|_{3,T} |w_h - w_h^I|_{1,T} \le c\, h\, |u|_{3,\Omega} \|w_h\|_h.$$

Using the estimate (4.1), we have

$$(f, u_h - v_h) - a_h(u, u_h - v_h) \le c\,h\,(|u|_{3,\Omega} + h\,\|f\|_{0,\Omega})\,\|w_h\|_h + \langle \partial_n \Delta u, w_h^I\rangle_{1/2,\Gamma} - (1-\nu)\int_\Gamma \partial_{n\tau} u\,\partial_\tau w_h^I\,ds.$$

By (2.25), we then obtain

$$(f, u_h - v_h) - a_h(u, u_h - v_h) \le c\,h\,(|u|_{3,\Omega} + h\,\|f\|_{0,\Omega})\,\|w_h\|_h + \int_{\Gamma_3} g\,\lambda\,w_h^I\,ds.$$

Thus, for the residual term defined in (3.4), we have

$$\begin{aligned} R_h(v_h, u_h) &\le \int_{\Gamma_3} g\,(|v_h| - |u_h| + \lambda\,w_h^I)\,ds + c\,h\,(|u|_{3,\Omega} + h\,\|f\|_{0,\Omega})\,\|w_h\|_h \\ &= \int_{\Gamma_3} g\,(|v_h| - |u_h| + \lambda\,w_h)\,ds + \int_{\Gamma_3} g\,\lambda\,(w_h^I - w_h)\,ds \\ &\quad + c\,h\,(|u|_{3,\Omega} + h\,\|f\|_{0,\Omega})\,\|w_h\|_h. \end{aligned} \tag{5.2}$$

The second term on the right is bounded as follows:

$$\begin{aligned} \int_{\Gamma_3} g\,\lambda\,(w_h^I - w_h)\,ds &\le \sum_{\gamma\subset\Gamma_3}\int_\gamma g\,|w_h - w_h^I|\,ds \\ &\le \|g\|_{0,\Gamma_3}\left(\sum_{\gamma\subset\Gamma_3}\|w_h - w_h^I\|_{0,\gamma}^2\right)^{1/2} \\ &\le c\,\|g\|_{0,\Gamma_3}\left(\sum_{\partial T\cap\Gamma_3\neq\emptyset}\left[h^{-1}\|w_h - w_h^I\|_{0,T}^2 + h\,|w_h - w_h^I|_{1,T}^2\right]\right)^{1/2} \\ &\le c\,\|g\|_{0,\Gamma_3} h^{3/2}\left(\sum_{T:\partial T\cap\Gamma_3\neq\emptyset}|w_h|_{2,T}^2\right)^{1/2} \\ &\le c\,h^{3/2}\|g\|_{0,\Gamma_3}\|w_h\|_h. \end{aligned}$$

The first term $\int_{\Gamma_3} g\,(|v_h| - |u_h| + \lambda\,w_h)\,ds$ can be handled similarly as in the proof of Theorem 4.1. So, with $v_h = \Pi_h u$ in (5.2), we have the bound

$$\begin{aligned} R_h(\Pi_h u, u_h) &\le c\,(h^{3/2}\|g\|_{0,\Gamma_3} + h\,|u|_{3,\Omega} + h^2\|f\|_{0,\Omega})\,\|\Pi_h u - u_h\|_h \\ &\quad + c\,h^{5/2}\|g\|_{0,\Gamma_3}\|u\|_{3,\Omega}. \end{aligned} \tag{5.3}$$

Now, by (3.3), we have

$$\begin{aligned} \|u - u_h\|_h^2 \le c\,(&\|u - \Pi_h u\|_h^2 + (h^{3/2}\|g\|_{0,\Gamma_3} + h\,|u|_{3,\Omega} + h^2\|f\|_{0,\Omega})\,\|\Pi_h u - u_h\|_h \\ &+ h^{5/2}\|g\|_{0,\Gamma_3}\|u\|_{3,\Omega}), \end{aligned}$$

from which we can derive the error estimate (5.1) as in the proof of Theorem 4.1. □

As examples of discontinuous nonconforming finite element spaces for the plate contact problem, we mention Morley's triangle and the Fraeijs De Veubeke triangle.

Assume Ω is a polygonal domain and let $\{\mathcal{T}_h\}$ be a regular family of partitions of $\overline{\Omega}$ into triangles. For Morley's triangle, on each element, the finite element function is quadratic and is uniquely determined by the function values at the three vertices and the normal derivative at the three midside nodes. For the Fraeijs De Veubeke triangle, on each element, the finite element function is cubic and is uniquely determined by the function values at the three vertices and at the center and the normal derivative at the Gaussian points of second order on each side. From their constructions, we see that for both Morley's triangle and the Fraeijs De Veubeke triangle, the finite element functions are continuous at the vertices of the corresponding triangulation. For both elements, (4.1) and (4.2) are valid (cf. [26]).

We conclude that for both Morley's triangle and the Fraeijs De Veubeke triangle, the optimal order error estimate (5.1) holds.

Acknowledgment. We thank the two referees whose suggestions led to an improvement of this paper.

REFERENCES

[1] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities: Applications to Free-Boundary Problems*, John Wiley, New York, 1984.

[2] G.P. Bazeley, Y.K. Cheung, B.M. Irons, and O.C. Zienkiewicz, *Triangular elements in bending—conforming and non-conforming solutions*, in Proceedings of the Conference on Matrix Methods in Structural Mechanics, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, 1965.

[3] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[4] F. Brezzi, W.W. Hager, and P.A. Raviart, *Error estimates for the finite element solution of variational inequalities, Part* I. *Primal theory*, Numer. Math., 28 (1977), pp. 431–443.

[5] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[6] P.G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 17–351.

[7] I. Doghri, *Mechanics of Deformable Solids*, Springer-Verlag, Berlin, 2000.

[8] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[9] R. Falk, *Error estimates for the approximation of a class of variational inequalities*, Math. Comput., 28 (1974), pp. 963–971.

[10] A. Friedman, *Variational Principles and Free-boundary Problems*, John Wiley, New York, 1982.

[11] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, New York, 1986.

[12] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[13] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[14] W. Han and B.D. Reddy, *Plasticity: Mathematical Theory and Numerical Analysis*, Springer-Verlag, New York, 1999.

[15] W. Han and M. Sofonea, *Quasistatic Contact Problems in Viscoelasticity and Viscoplasticity*, AMS and International Press, 2002.

[16] J. Haslinger, I. Hlaváček, and J. Nečas, *Numerical methods for unilateral problems in solid mechanics*, in Handbook of Numerical Analysis, Vol. IV, P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1996, pp. 313–485.

[17] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Solution of Variational Inequalities in Mechanics*, Springer-Verlag, New York, 1988.

[18] B.M. Irons and A. Razzaque, *Experience with the patch test for convergence of finite element methods*, in Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A.R. Aziz, ed., Academic Press, New York, 1972, pp. 557–587.

[19] N. Kikuchi and J.T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.
[20] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
[21] P. Lascaux and P. Lesaint, *Some nonconforming finite elements for the plate bending problem*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 9 (1975), pp. 9–53.
[22] P.D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*, Birkhäuser, Boston, 1985.
[23] J.F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland, Amsterdam, 1987.
[24] Z.C. Shi, *The generalized patch test for Zienkiewicz's triangles*, J. Comput. Math., 2 (1984), pp. 279–286.
[25] Z.C. Shi, *The F-E-M-Test for nonconforming finite elements*, Math. Comp., 49 (1987), pp. 391–405.
[26] F. Stummel, *The generalized patch test*, SIAM J. Numer. Anal., 16 (1979), pp. 449–471.
[27] L.H. Wang, *Morley's element approximation to a fourth order variational inequality with curvature obstacle*, Math. Numer. Sinica, 12 (1990), pp. 279–284.
[28] L.H. Wang, *Some nonconforming finite element approximations of a fourth order variational inequality with displacement obstacle*, Math. Numer. Sinica, 12 (1990), pp. 352–356.
[29] L.H. Wang, *Some strongly discontinuous nonconforming finite element approximations for a fourth order variational inequality with displacement obstacle*, Math. Numer. Sinica, 14 (1992), pp. 98–101.
[30] M. Wang, *On the necessity and sufficiency of the patch test for convergence of nonconforming finite elements*, SIAM J. Numer. Anal., 39 (2001), pp. 363–384.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1698–1715

A PRIORI ERROR ESTIMATES FOR MIXED FINITE ELEMENT APPROXIMATIONS OF THE ACOUSTIC WAVE EQUATION*

ELEANOR W. JENKINS†, BÉATRICE RIVIÈRE†, AND MARY F. WHEELER†

Abstract. In this paper we derive optimal a priori $L^\infty(L^2)$ error estimates for mixed finite element displacement formulations of the acoustic wave equation. The computational complexity of this approach is equivalent to the traditional mixed finite element formulations of the second order hyperbolic equations in which the primary unknowns are pressure and the gradient of pressure. However, the displacement formulations with the physical variables of interest, displacement and pressure, requires less regularity on the displacement.

Key words. acoustic wave equation, mixed finite elements, error estimation

AMS subject classifications. 35L05, 65M12, 65M60, 65M15

PII. S0036142901388068

1. Introduction.

There is significant interest in simulating the effects of wave propagation in heterogeneous media to aid in the interpretation of field data and to predict the damage patterns due to earthquakes. Simulated waveform data (seismograms) computed for an assumed earth model are compared against the recorded data. If the match is unacceptable, the model is perturbed, and the simulation is redone and compared again. This procedure is implemented formally by global optimization techniques [19] resulting in a description of an earth model (with its associated uncertainties) that explains the observations. Thus there is a need for a fast and accurate simulation technique that can be used for real time analysis of seismograms.

In the past, wave simulation has been successfully modeled using finite difference methods [10, 13], but these solutions have been expensive to compute. The use of structured finite differences in simulating earthquake responses in the Los Angeles Basin requires 35 billion grid points [3], which emphasizes the need for unstructured meshes. The staggered grid approach described in [10] may be used to solve problems on the order of millions using a workstation, but memory optimization routines must be used and it is unclear that the method is easily parallelized.

Finite element discretization methods have the advantage of handling complex geometries and straightforward local discretization techniques using error indicators. It is also easy to incorporate free surface boundary conditions and nonmatching grids. In [13], Marfurt concludes that finite element methods may be the most cost-effective way to simulate wave fields.

Previous attempts at wave simulation using finite elements have used continuous Galerkin methods [2, 3, 9, 13, 18], discontinuous Galerkin methods [11, 17], and mixed finite element methods [7, 8]. We propose mixed finite element displacement formulations for solving the acoustic wave equation, which are described below. These approximations can be defined for general meshes.

*Received by the editors April 18, 2001; accepted for publication (in revised form) April 25, 2002; published electronically November 14, 2002. This research was supported by the NSF through NPACI and the National Partnership for Advanced Computational Infrastructure, grant 10152711.
http://www.siam.org/journals/sinum/40-5/38806.html

†Center for Subsurface Modeling, Texas Institute for Computational and Applied Mathematics, University of Texas at Austin, ACES 5.340, 201 E. 24th Street, Austin, TX 78712 (lea@ticam.utexas.edu, riviere@ticam.utexas.edu, mfw@ticam.utexas.edu).

Let Ω be a bounded domain in $\mathbb{R}^n, n = 1, 2, 3$, with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$. The general form of the wave equation is

$$\rho \boldsymbol{u}_{tt} - \nabla \cdot \tilde{\boldsymbol{\tau}} = \boldsymbol{f} \qquad \text{in } \Omega \times (0,T)\,, \tag{1.1}$$

$$\nabla \cdot \boldsymbol{u} = 0 \qquad \text{on } \Gamma_D \times (0,T)\,, \tag{1.2}$$

$$\boldsymbol{u} \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \Gamma_N \times (0,T)\,, \tag{1.3}$$

$$\boldsymbol{u}\,(\cdot,0) = \boldsymbol{u}_0 \qquad \text{in } \Omega, \tag{1.4}$$

$$\boldsymbol{u}_t\,(\cdot,0) = \boldsymbol{u}_1 \qquad \text{in } \Omega, \tag{1.5}$$

where $\boldsymbol{u}$ is the displacement, ρ is the density, and $\tilde{\boldsymbol{\tau}}$ is the stress tensor given by the generalized Hooke's law $\tilde{\boldsymbol{\tau}} = \lambda(\nabla \cdot \boldsymbol{u})\tilde{\boldsymbol{I}} + \mu(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T)$. Here $\lambda > 0$ and μ are the Lamé coefficients characterizing the material. We let $\boldsymbol{f}$ represent a general source term, and let $\boldsymbol{u}_0$ and $\boldsymbol{u}_1$ be the initial conditions on displacements and velocities, and we assume that $\boldsymbol{f}$, $\boldsymbol{u}_0$, and $\boldsymbol{u}_1$ are smooth enough so that there is a unique solution $\boldsymbol{u} \in \mathcal{C}^2\,((0,T) \times \Omega)$ to (1.1)–(1.5) [12].

The acoustic problem is the limiting case with $\mu = 0$. In this case, (1.1) becomes

$$\rho \boldsymbol{u}_{tt} - \nabla \cdot (\lambda(\nabla \cdot \boldsymbol{u})\tilde{\boldsymbol{I}}) = \boldsymbol{f}. \tag{1.6}$$

We assume that ρ and λ are bound above and below by the positive constants ρ_0, ρ_1, λ_0, and λ_1, respectively.

A standard approach in geophysics modeling is to solve a scalar wave equation. Here, since $p = \lambda \nabla \cdot \boldsymbol{u}$, we have $p_{tt} = \lambda \nabla \cdot \boldsymbol{u}_{tt}$. We substitute this expression into (1.6) to obtain

$$p_{tt} - \lambda \nabla \cdot \frac{1}{\rho}\,(\nabla p) = \tilde{f}, \tag{1.7}$$

where $\tilde{f} = \nabla \cdot \boldsymbol{f}$.

A priori error estimates for solving (1.7) with a constant λ were obtained by Cowsar, Dupont, and Wheeler [7, 8]. We propose an alternative mixed finite elements displacement formulation that requires less regularity on the displacement solution than the approach in [7, 8]. We describe this method in section 3. We derive the error estimates for the continuous-in-time problem in section 4. For the discrete-in-time problem, stability results and error estimates are obtained in section 5. In section 6 we present conclusions.

2. Preliminaries. In this section we describe the notation used in this paper, we introduce the functional spaces, two projection operators and their approximation properties, and we recall Gronwall's inequality, which is needed in the error analysis.

2.1. Inner products and norms. The L^2 inner product over Ω is defined as

$$(u, v) = \int_\Omega uv\, d\Omega,$$

and we denote by $\|\cdot\|_{L^2}$ the L^2 norm over Ω, i.e., $\|u\|_{L^2(\Omega)} = (u,u)^{\frac{1}{2}}$. The inner product over the boundary $\partial\Omega$ is denoted

$$\langle u, v \rangle = \int_{\partial\Omega} uv\, d\Omega$$

for $u, v \in H^{\frac{1}{2}+\epsilon}(\Omega)$, with $\epsilon > 0$. The time-space norm $\|\cdot\|_{L^2(0,T;L^2(\Omega))}$ is defined as

$$\|u\|_{L^2(0,T;L^2(\Omega))} = \|u\|_{L^2(L^2)} = \left(\int_0^T \|u\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}.$$

The time-space norm $\|\cdot\|_{L^\infty(L^2)}$ is similarly defined.

2.2. Gronwall's inequality. We recall Gronwall's inequality, which states that if $y \geq 0$ satisfies $y_t \leq ky(t) + h(t)$ for $0 \leq t \leq \tau$, where $k \geq 0$ is a constant and $h(t) \geq 0$, $h \in L^1((0,\tau))$, then

$$y(t) \leq e^{k\tau} \left[y(0) + \int_0^\tau h(s)\, ds \right]$$

for all $t \in [0,\tau]$ [14].

2.3. Functional spaces and projections. We define the standard Sobolev spaces for mixed methods:

$$\begin{aligned} \boldsymbol{H}(\Omega, \mathrm{div}) &= \{\boldsymbol{v} : \boldsymbol{v} \in (\boldsymbol{L}^2(\Omega))^n, \nabla \cdot \boldsymbol{v} \in \boldsymbol{L}^2(\Omega)\}, \\ \boldsymbol{V} &= \{\boldsymbol{v} \in \boldsymbol{H}(\Omega, \mathrm{div}) : \boldsymbol{v} \cdot \boldsymbol{\nu}|_{\Gamma_N} = 0\}, \\ W &= H^{\frac{1}{2}+\epsilon}(\Omega) \text{ for any } \epsilon > 0. \end{aligned}$$

Let $\{\mathcal{E}_h\}_{h>0}$ be a quasi-uniform family of finite element partitions of Ω, where h is the maximal element diameter. Let $\boldsymbol{V}_h \times W_h$ be any of the usual mixed finite element approximating subspaces of $\boldsymbol{V} \times W$, that is, the Raviart–Thomas–Nedelec spaces [15, 16], Brezzi–Douglas–Marini spaces [5], or Brezzi–Douglas–Fortin–Marini spaces [4].

Each of these mixed spaces has a projection operator $\Pi_h : \boldsymbol{H}(\Omega, \mathrm{div}) \to \boldsymbol{V}_h$ such that for any $\boldsymbol{z} \in \boldsymbol{H}(\Omega, \mathrm{div})$

$$(\nabla \cdot \Pi_h \boldsymbol{z}, w) = (\nabla \cdot \boldsymbol{z}, w) \quad \forall w \in W_h. \tag{2.1}$$

In addition, if $\boldsymbol{z} \in \boldsymbol{H}(\Omega, \mathrm{div}) \cap \boldsymbol{H}^k(\Omega)$, we also have

$$\|\Pi_h \boldsymbol{z} - \boldsymbol{z}\|_0 \leq Ch^j \|\boldsymbol{z}\|_j, \quad 1 \leq j \leq k, \tag{2.2}$$

where k is associated with the degree of the polynomial, and $\|\cdot\|_s$ is the standard Sobolev norm on $(\boldsymbol{H}^s(\Omega))^n$ [1].

Let $\mathcal{P}_h$ be the L^2 projection of W onto W_h such that

$$(\mathcal{P}_h \phi, w) = (\phi, w) \quad \forall \phi \in W, \quad \forall w \in W_h. \tag{2.3}$$

In addition, if $\phi \in W \cap H^k(\Omega)$, then we also have

$$\|\mathcal{P}_h \phi - \phi\|_s \leq Ch^{j-s} \|\phi\|_j, \quad 0 \leq s \leq k, \quad 0 \leq j \leq k. \tag{2.4}$$

3. Model problem and scheme. We observe that $\nabla \cdot (\lambda(\nabla \cdot \boldsymbol{u})\tilde{\boldsymbol{I}}) = \nabla(\lambda \nabla \cdot \boldsymbol{u})$ so that (1.6) may be rewritten as

$$\rho \boldsymbol{u}_{tt} - \nabla(\lambda \nabla \cdot \boldsymbol{u}) = f. \tag{3.1}$$

By introducing $p = \lambda \nabla \cdot \boldsymbol{u}$, we present (3.1) in a mixed finite element form:

$$\rho \boldsymbol{u}_{tt} - \nabla p = \boldsymbol{f} \quad \text{in } \Omega \times (0,T)\,, \tag{3.2}$$

$$\lambda^{-1} p = \nabla \cdot \boldsymbol{u} \quad \text{in } \Omega \times (0,T)\,. \tag{3.3}$$

A similar approach for the linear elasticity problem has been presented by Brezzi and Fortin in [6].

Let $\boldsymbol{v} \in \boldsymbol{V}$ and $w \in W$. If we multiply (3.2) and (3.3) by $\boldsymbol{v}$ and w, respectively, and integrate over Ω we get the weak formulation

$$(\rho \boldsymbol{u}_{tt}, \boldsymbol{v}) - (\nabla p, \boldsymbol{v}) = (\boldsymbol{f}(t), \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \tag{3.4}$$

$$\left(\lambda^{-1} p, w\right) - (\nabla \cdot \boldsymbol{u}, w) = 0 \quad \forall\, w \in W. \tag{3.5}$$

Clearly, if $\boldsymbol{u}$, p satisfy (3.2) and (3.3), then $\boldsymbol{u}$, p satisfy (3.4) and (3.5).

If we integrate by parts in (3.4) we get, for any $\boldsymbol{v} \in \boldsymbol{V}$,

$$(\nabla p, \boldsymbol{v}) = \langle p, \boldsymbol{v} \cdot \nu \rangle - (p, \nabla \cdot \boldsymbol{v}) = -\,(p, \nabla \cdot \boldsymbol{v})\,,$$

and the weak formulation becomes the following:

For any $t \geq 0$, find $(\boldsymbol{u}(t), p(t)) \in \boldsymbol{V} \times W$ such that

$$(\boldsymbol{u}(0), \boldsymbol{v}) = (\boldsymbol{u}_0, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \tag{3.6}$$

$$(\boldsymbol{u}_t(0), \boldsymbol{v}) = (\boldsymbol{u}_1, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \tag{3.7}$$

$$\left(\lambda^{-1} p(0), w\right) = (\nabla \cdot \boldsymbol{u}_0, w) \quad \forall\, w \in W, \tag{3.8}$$

$$(\rho \boldsymbol{u}_{tt}(t), \boldsymbol{v}) + (p(t), \nabla \cdot \boldsymbol{v}) = (\boldsymbol{f}(t), \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \quad \forall\, t > 0, \tag{3.9}$$

$$\left(\lambda^{-1} p(t), w\right) - (\nabla \cdot \boldsymbol{u}(t), w) = 0 \quad \forall\, w \in W, \quad \forall\, t > 0. \tag{3.10}$$

Note that in [7, 8] it is necessary that $\nabla p \in \boldsymbol{H}(\Omega, \mathrm{div})$ so that $\nabla \cdot \boldsymbol{u} \in H^2(\Omega)$. Here, we require only that $\nabla \cdot \boldsymbol{u} \in H^{\frac{1}{2}}(\Omega)$. It is clear that the solution $\boldsymbol{u} \in \mathcal{C}^2((0,T) \times \Omega)$ of problem (1.1)–(1.5) with $p = \lambda \nabla \cdot \boldsymbol{u}$ is a solution to (3.6)–(3.10). The uniqueness is provided by the following lemma.

LEMMA 3.1. *Let* $(\boldsymbol{u}_a, p_a)$ *and* $(\boldsymbol{u}_b, p_b)$ *be two solutions of* (3.6)–(3.10). *Then* $\boldsymbol{u}_a = \boldsymbol{u}_b$ *and* $p_a = p_b$.

Proof. Let $\boldsymbol{\chi} = \boldsymbol{u}_a - \boldsymbol{u}_b$ and $\psi = p_a - p_b$. Then by subtracting the equations satisfied by these solutions we have

$$(\boldsymbol{\chi}(0), \boldsymbol{v}) = 0 \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \tag{3.11}$$

$$(\boldsymbol{\chi}_t(0), \boldsymbol{v}) = 0 \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \tag{3.12}$$

$$\left(\lambda^{-1} \psi(0), w\right) = 0 \quad \forall\, w \in W, \tag{3.13}$$

$$(\rho \boldsymbol{\chi}_{tt}(t), \boldsymbol{v}) + (\psi(t), \nabla \cdot \boldsymbol{v}) = 0 \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}, \quad \forall\, t > 0, \tag{3.14}$$

$$\left(\lambda^{-1} \psi(t), w\right) - (\nabla \cdot \boldsymbol{\chi}(t), w) = 0 \quad \forall\, w \in W, \quad \forall\, t > 0. \tag{3.15}$$

We differentiate (3.15) with respect to time:

$$\left(\lambda^{-1} \psi_t, w\right) - (\nabla \cdot \boldsymbol{\chi}_t, w) = 0 \quad \forall w \in W, \quad \forall t > 0. \tag{3.16}$$

We choose $\boldsymbol{v} = \boldsymbol{\chi}_t$ and $w = p$. Thus (3.14) and (3.16) become

$$(\rho \boldsymbol{\chi}_{tt}, \boldsymbol{\chi}_t) + (\psi, \nabla \cdot \boldsymbol{\chi}_t) = 0, \tag{3.17}$$

$$\left(\lambda^{-1} \psi_t, \psi\right) - (\nabla \cdot \boldsymbol{\chi}_t, \psi) = 0. \tag{3.18}$$

We add (3.17) and (3.16) so that

$$(\rho \boldsymbol{\chi}_{tt}, \boldsymbol{\chi}_t) + \left(\lambda^{-1}\psi_t, \psi\right) = 0 \tag{3.19}$$

and thus

$$\frac{1}{2}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\psi\right\|^2_{L^2(\Omega)} = 0. \tag{3.20}$$

We integrate (3.20) with respect to time to obtain

$$\frac{1}{2}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\left\|\lambda^{-\frac{1}{2}}\psi\right\|^2_{L^2(\Omega)} = C_1, \tag{3.21}$$

where C_1 is a constant independent of time. As (3.21) holds for any t, it holds in particular for $t = 0$. We have from the initial data that $\|\boldsymbol{\chi}_t(0)\|^2_{L^2(\Omega)} = \|\psi(0)\|^2_{L^2(\Omega)} = 0$, so $C_1 = 0$. Hence

$$\frac{1}{2}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\left\|\lambda^{-\frac{1}{2}}\psi\right\|^2_{L^2(\Omega)} = 0 \tag{3.22}$$

and therefore

$$\frac{1}{2}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} = \frac{1}{2}\left\|\lambda^{-\frac{1}{2}}\psi\right\|^2_{L^2(\Omega)} = 0. \tag{3.23}$$

However, we assume that both λ and ρ are bounded above and below away from 0. Hence we must have that

$$\boldsymbol{\chi}_t = \psi = 0.$$

Now we consider $\boldsymbol{\chi}_t = 0$. We integrate with respect to time and obtain

$$\boldsymbol{\chi} = C_2,$$

where C_2 is a constant independent of time. This holds for any t; in particular, it holds for $t = 0$. Again, we use the initial conditions and see that $\boldsymbol{\chi}(0) = 0$, and hence $C_2 = 0$. Thus $\boldsymbol{\chi} = 0$, which concludes the proof. ∎

The mixed finite element approximation to $(\boldsymbol{u}(t), p(t))$ for any $t \geq 0$ is given by the functions $(\boldsymbol{U}(t), P(t)) \in \boldsymbol{V}_h \times W_h$ satisfying

$$(\boldsymbol{U}(0), \boldsymbol{v}) = (\Pi_h \boldsymbol{u}_0, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{3.24}$$

$$(\boldsymbol{U}_t(0), \boldsymbol{v}) = (\Pi_h \boldsymbol{u}_1, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{3.25}$$

$$(P(0), w) = (p(0), w) \quad \forall\, w \in W_h, \tag{3.26}$$

$$(\rho \boldsymbol{U}_{tt}(t), \boldsymbol{v}) + (P(t), \nabla \cdot \boldsymbol{v}) = (\boldsymbol{f}(t), \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \quad \forall\, t > 0, \tag{3.27}$$

$$\left(\lambda^{-1} P(t), w\right) - (\nabla \cdot \boldsymbol{U}(t), w) = 0 \quad \forall\, w \in W_h, \quad \forall\, t > 0. \tag{3.28}$$

The existence and uniqueness of a solution $(\boldsymbol{U}(t), P(t))$ to (3.24)–(3.28) is shown in the following lemma.

LEMMA 3.2. *A solution* $(\boldsymbol{U}(t), P(t))$ *to* (3.24)–(3.28) *exists and is unique.*

Proof. Because we are operating in a finite dimensional space, it suffices to show uniqueness. Uniqueness follows directly from the proof of the previous lemma. ∎

4. Continuous in time a priori error estimates. In this section, we prove the convergence of the scheme (3.24)–(3.28) in the $L^\infty(L^2)$ norm. We first derive an estimate for the error in the pressure and the velocity, then for the error in the displacement.

THEOREM 4.1. *For $t \geq 0$, let $(\boldsymbol{u}(t), p(t))$ be the solution of the problem (3.6)–(3.10). Assume that $\boldsymbol{u}_t \in L^\infty\left(\boldsymbol{L}^2(\Omega)\right)$, $\boldsymbol{u}_{tt} \in L^2(\boldsymbol{H}^k(\Omega))$, $p \in L^\infty\left(L^2(\Omega)\right)$, and $p_t \in L^2\left(H^k(\Omega)\right)$. Then there exists a constant C independent of h such that*

$$\left\|\rho^{\frac{1}{2}}(\boldsymbol{u}_t - \boldsymbol{U}_t)\right\|_{L^\infty(L^2)} + \left\|\lambda^{-\frac{1}{2}}(p - P)\right\|_{L^\infty(L^2)} \leq Ch^k\left(\|\boldsymbol{u}_{tt}\|_{L^2(H^k)} + \|p_t\|_{L^2(H^k)}\right), \tag{4.1}$$

where k is associated with the degree of the finite element polynomial.

Proof. For simplification, we denote $\boldsymbol{\chi} = \boldsymbol{U} - \Pi_h\boldsymbol{u}$, $\xi = P - \mathcal{P}_h p$, $\boldsymbol{\eta} = \boldsymbol{u} - \Pi_h\boldsymbol{u}$, and $\zeta = p - \mathcal{P}_h p$, where Π_h and $\mathcal{P}_h$ have been defined in section 2.3. These definitions hold throughout the paper. If we subtract $\Pi_h\boldsymbol{u}$ and $\mathcal{P}_h p$ from (3.9), (3.10), (3.27), and (3.28) we obtain

$$(\rho\boldsymbol{\chi}_{tt}, \boldsymbol{v}) + (\xi, \nabla\cdot\boldsymbol{v}) = (\rho\boldsymbol{\eta}_{tt}, \boldsymbol{v}) + (\zeta, \nabla\cdot\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \tag{4.2}$$

$$\left(\lambda^{-1}\xi, w\right) - (\nabla\cdot\boldsymbol{\chi}, w) = \left(\lambda^{-1}\zeta, w\right) - (\nabla\cdot\boldsymbol{\eta}, w) \quad \forall w \in W_h. \tag{4.3}$$

Since $\boldsymbol{\chi} \in \boldsymbol{V}_h$, we can set $\boldsymbol{v} = \boldsymbol{\chi}_t$ in (4.2):

$$(\rho\boldsymbol{\chi}_{tt}, \boldsymbol{\chi}_t) + (\xi, \nabla\cdot\boldsymbol{\chi}_t) = (\rho\boldsymbol{\eta}_{tt}, \boldsymbol{\chi}_t) + (\zeta, \nabla\cdot\boldsymbol{\chi}_t). \tag{4.4}$$

We then differentiate (4.3) with respect to time to obtain

$$\left(\lambda^{-1}\xi_t, w\right) - (\nabla\cdot\boldsymbol{\chi}_t, w) = \left(\lambda^{-1}\zeta_t, w\right) - (\nabla\cdot\boldsymbol{\eta}_t, w) \quad \forall w \in W_h. \tag{4.5}$$

As $\xi \in W_h$, we can set $w = \xi$ in (4.5), which gives

$$\left(\lambda^{-1}\xi_t, \xi\right) - (\nabla\cdot\boldsymbol{\chi}_t, \xi) = \left(\lambda^{-1}\zeta_t, \xi\right) - (\nabla\cdot\boldsymbol{\eta}_t, \xi). \tag{4.6}$$

Adding (4.4) and (4.6) gives

$$\frac{1}{2}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)} = (\rho\boldsymbol{\eta}_{tt}, \boldsymbol{\chi}_t) + (\zeta, \nabla\cdot\boldsymbol{\chi}_t) + \left(\lambda^{-1}\zeta_t, \xi\right) - (\nabla\cdot\boldsymbol{\eta}_t, \xi). \tag{4.7}$$

If we use the definitions (2.1) and (2.3) of the Π_h and $\mathcal{P}_h$ projections, we obtain

$$\frac{1}{2}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)} = (\rho\boldsymbol{\eta}_{tt}, \boldsymbol{\chi}_t) + \left(\lambda^{-1}\zeta_t, \xi\right). \tag{4.8}$$

We use the Cauchy–Schwarz inequality to bound the right-hand side of (4.8) so that

$$\frac{1}{2}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)} \leq \left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|_{L^2(\Omega)}\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_{tt}\right\|_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\zeta_t\right\|_{L^2(\Omega)}\left\|\lambda^{-\frac{1}{2}}\xi\right\|_{L^2(\Omega)},$$

and use the fact that $2ab \le a^2 + b^2$ to get

$$\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)} \le \left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)} + \left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_{tt}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\zeta_t\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)}.$$

Now we can apply Gronwall's inequality and

$$\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^2(\Omega)}(t) + \left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^2(\Omega)}(t) \le \left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t(0)\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\xi(0)\right\|_{L^2(\Omega)} + \int_0^t \left(\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_{tt}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\zeta_t\right\|^2_{L^2(\Omega)}\right).$$

If we take the supremum over all t and use the initial conditions we obtain

$$\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}_t\right\|^2_{L^\infty(L^2)} + \left\|\lambda^{-\frac{1}{2}}\xi\right\|^2_{L^\infty(L^2)} \le \left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_{tt}\right\|^2_{L^2(L^2)} + \left\|\lambda^{-\frac{1}{2}}\zeta_t\right\|^2_{L^2(L^2)}.$$

We complete the proof by using the approximation properties (2.2) and (2.4) of the projections. ∎

We now derive an estimate of the error in the displacement in the $L^\infty(L^2)$ norm.

THEOREM 4.2. *For $t \ge 0$, let $(\boldsymbol{u}(t), p(t))$ be the solution of problem (3.6)–(3.10). Assume that $\boldsymbol{u} \in L^\infty\left(L^2(\Omega)\right)$, $\boldsymbol{u}_t \in L^2(\boldsymbol{H}^k(\Omega))$, $\boldsymbol{u}_t(0) \in \boldsymbol{H}^k(\Omega)$, and $p \in L^2\left(H^k(\Omega)\right)$. Then there exists a constant C independent of h such that*

$$\left\|\rho^{\frac{1}{2}}(\boldsymbol{u} - \boldsymbol{U})\right\|_{L^\infty(L^2)} \le Ch^k\left(\|\boldsymbol{u}_t\|_{L^2(H^k)} + \|\boldsymbol{u}_t(0)\|_{H^k} + \|p\|_{L^2(H^k)}\right), \tag{4.9}$$

where k is associated with the degree of the finite element polynomial.

Proof. We first obtain the same equations (4.2) and (4.3) that arrive after taking into account the definition of the Π_h and the $\mathcal{P}_h$ projections

$$(\rho\boldsymbol{\chi}_{tt}, \boldsymbol{v}) + (\xi, \nabla\cdot\boldsymbol{v}) = (\rho\boldsymbol{\eta}_{tt}, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \tag{4.10}$$

$$\left(\lambda^{-1}\xi, w\right) - (\nabla\cdot\boldsymbol{\chi}, w) = (\lambda^{-1}\zeta, w) \quad \forall w \in W_h. \tag{4.11}$$

If we integrate (4.10) from 0 to t, noting that $\boldsymbol{\chi}_t(0) = 0$, we obtain

$$(\rho\boldsymbol{\chi}_t, \boldsymbol{v}) + \left(\int_0^t \xi, \nabla\cdot\boldsymbol{v}\right) = (\rho\boldsymbol{\eta}_t, \boldsymbol{v}) - (\rho\boldsymbol{\eta}_t(0), \boldsymbol{v}). \tag{4.12}$$

We set $\boldsymbol{v} = \boldsymbol{\chi}$, $\phi = \int_0^t \xi(s)\,ds$, and $w = \phi$. Then (4.12) and (4.11) become

$$(\rho\boldsymbol{\chi}_t, \boldsymbol{\chi}) + (\phi, \nabla\cdot\boldsymbol{\chi}) = (\rho\boldsymbol{\eta}_t, \boldsymbol{\chi}) - (\rho\boldsymbol{\eta}_t(0), \boldsymbol{\chi}), \tag{4.13}$$

$$\left(\lambda^{-1}\xi, \phi\right) - (\nabla\cdot\boldsymbol{\chi}, \phi) = (\lambda^{-1}\zeta, \phi). \tag{4.14}$$

Adding (4.13) and (4.14) gives

$$(\rho\boldsymbol{\chi}_t, \boldsymbol{\chi}) + \left(\lambda^{-1}\xi, \phi\right) = (\rho\boldsymbol{\eta}_t, \boldsymbol{\chi}) - (\rho\boldsymbol{\eta}_t(0), \boldsymbol{\chi}) + (\lambda^{-1}\zeta, \phi).$$

Therefore

$$(\rho\boldsymbol{\chi}_t, \boldsymbol{\chi}) + \left(\lambda^{-1}\phi_t, \phi\right) = (\rho\boldsymbol{\eta}_t, \boldsymbol{\chi}) - (\rho\boldsymbol{\eta}_t(0), \boldsymbol{\chi}) + (\lambda^{-1}\zeta, \phi)$$

so that

$$\frac{1}{2}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}+\frac{1}{2}\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}=(\rho\boldsymbol{\eta}_t,\boldsymbol{\chi})-(\rho\boldsymbol{\eta}_t(0),\boldsymbol{\chi})+(\lambda^{-1}\zeta,\phi).$$

Multiplying through by 2 and using the Cauchy–Schwarz inequality gives

$$\begin{aligned}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}+\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}&\le 2\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|_{L^2(\Omega)}\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t\right\|_{L^2(\Omega)}\\&\quad+2\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|_{L^2(\Omega)}\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t(0)\right\|_{L^2(\Omega)}+2\left\|\lambda^{-\frac{1}{2}}\zeta\right\|_{L^2(\Omega)}\left\|\lambda^{-\frac{1}{2}}\phi\right\|_{L^2(\Omega)}.\end{aligned}$$

Hence

$$\begin{aligned}\frac{d}{dt}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}+\frac{d}{dt}\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}&\le 2\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t\right\|^2_{L^2(\Omega)}\\&\quad+\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t(0)\right\|^2_{L^2(\Omega)}+\left\|\lambda^{-\frac{1}{2}}\zeta\right\|^2_{L^2(\Omega)}+\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}.\end{aligned}\tag{4.15}$$

We apply Gronwall's inequality to (4.15) and note that $\phi(0)=0$ and $\boldsymbol{\chi}(0)=0$ from (3.24). Then

$$\begin{aligned}&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}(t)+\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}(t)\\&\qquad\le\int_0^t\left(\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_s\right\|^2_{L^2(\Omega)}+\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_s(0)\right\|^2_{L^2(\Omega)}+\left\|\lambda^{-\frac{1}{2}}\zeta\right\|^2_{L^2(\Omega)}\right)ds\end{aligned}$$

and thus

$$\begin{aligned}&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^2(\Omega)}(t)+\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{L^2(\Omega)}(t)\\&\qquad\le Ct\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t(0)\right\|^2_{L^2(\Omega)}+C\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t\right\|^2_{L^2(L^2)}+C\left\|\lambda^{-\frac{1}{2}}\zeta\right\|^2_{L^2(L^2)}.\end{aligned}$$

We take the supremum over all t to get

$$\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{L^\infty(L^2)}\le C\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t\right\|^2_{L^2(L^2)}+C\left\|\rho^{\frac{1}{2}}\boldsymbol{\eta}_t(0)\right\|^2_{L^2(\Omega)}+C\left\|\lambda^{-\frac{1}{2}}\zeta\right\|^2_{L^2(L^2)}.$$

The final result is obtained by using the approximation properties (2.2) and (2.4). ∎

5. Explicit method. In this section, we define further notation, we formulate the fully discrete mixed finite element scheme, and we analyze the stability and the convergence of the discrete method.

5.1. Notation and scheme. Let $\Delta t>0$ be the time step size and define $t_i=i\Delta t$ with $t_N=T$. For any function ϕ of time, let ϕ^n denote $\phi(t_n)$. We denote $\phi^{n+\frac{1}{2}}=(\phi^n+\phi^{n+1})/2$, and we define the following terms for the discrete temporal derivatives:

$$\partial_t\phi^n=\frac{\phi^{n+1}-\phi^{n-1}}{2\Delta t},\quad \partial_t\phi^{n+\frac{1}{2}}=\frac{\phi^{n+1}-\phi^n}{\Delta t},\quad \partial_t^2\phi^n=\frac{\phi^{n+1}-2\phi^n+\phi^{n-1}}{\Delta t^2}.$$

We easily see that we have

$$\partial_t^2 \phi^n = \frac{\partial_t \phi^{n+\frac{1}{2}} - \partial_t \phi^{n-\frac{1}{2}}}{\Delta t} \quad \text{and} \quad \partial_t \phi^n = \frac{\partial_t \phi^{n+\frac{1}{2}} + \partial_t \phi^{n-\frac{1}{2}}}{2}. \tag{5.1}$$

The fully discrete mixed finite element scheme is as follows: find $\left(\boldsymbol{U}^{n+1}, P^{n+1}\right)$ in $V_h \times W_h$ such that

$$\left(\boldsymbol{U}^0, \boldsymbol{v}\right) = \left(\Pi_h \boldsymbol{u}_0, \boldsymbol{v}\right) \qquad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h \tag{5.2}$$

$$\left(P^0, w\right) = \left(p^0, w\right) \qquad \forall\, w \in W_h \tag{5.3}$$

$$\left(\rho \frac{2}{\Delta t} \partial_t \boldsymbol{U}^{\frac{1}{2}}, \boldsymbol{v}\right) + \left(P^0, \nabla \cdot \boldsymbol{v}\right) = \left(\boldsymbol{f}^0 + \rho \frac{2}{\Delta t} \Pi_h \boldsymbol{u}_1, \boldsymbol{v}\right) \qquad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{5.4}$$

$$\left(\rho \partial_t^2 \boldsymbol{U}^n, \boldsymbol{v}\right) + \left(P^n, \nabla \cdot \boldsymbol{v}\right) = \left(\boldsymbol{f}^n, \boldsymbol{v}\right) \qquad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{5.5}$$

$$\left(\lambda^{-1} P^{n+1}, w\right) - \left(\nabla \cdot \boldsymbol{U}^{n+1}, w\right) = 0 \qquad \forall\, w \in W_h. \tag{5.6}$$

5.2. Stability condition for the discrete problem. We show that the scheme is stable for the Dirichlet (homogeneous) problem, and in particular show that the temporal iterates are bound by the initial data. We consider (5.5) and (5.6) for the homogeneous case,

$$\left(\rho \partial_t^2 \boldsymbol{U}^n, \boldsymbol{v}\right) + \left(P^n, \nabla \cdot \boldsymbol{v}\right) = 0 \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \tag{5.7}$$

$$\left(\lambda^{-1} P^{n+1}, w\right) - \left(\nabla \cdot \boldsymbol{U}^{n+1}, w\right) = 0 \qquad \forall w \in W_h. \tag{5.8}$$

As in [7], we use the "inverse assumption," which states that there exists a constant C_0, independent of h, such that

$$\|\nabla \cdot \phi\|_{L^2(\Omega)} \leq C_0 h^{-1} \|\phi\|_{L^2(\Omega)} \tag{5.9}$$

for $\phi \in W_h$.

THEOREM 5.1. *The explicit scheme defined by* (5.2)–(5.6) *is stable if* $\Delta t < \frac{2h\rho_0^{\frac{1}{2}}}{C_0 \lambda_1^{\frac{1}{2}}}$. *That is,*

$$\left(1 - \frac{\Delta t^2 C_0^2 \lambda_1}{4h^2 \rho_0}\right) \left\|\partial_t \boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2 + \left\|P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2 \tag{5.10}$$

is bound by initial data.

Proof. If we subtract (5.8) from itself, with $n+1$ replaced by $n-1$, we get

$$\left(\lambda^{-1}\left(P^{n+1} - P^{n-1}\right), w\right) - \left(\nabla \cdot \left(\boldsymbol{U}^{n+1} - \boldsymbol{U}^{n-1}\right), w\right) = 0, \quad w \in W_h. \tag{5.11}$$

As (5.7) holds for all $\boldsymbol{v} \in \boldsymbol{V}_h$ and (5.11) holds for all $w \in W_h$, we set $\boldsymbol{v} = \partial_t \boldsymbol{U}^n$ and $w = \frac{P^n}{2\Delta t}$. This gives

$$\left(\rho \partial_t^2 \boldsymbol{U}^n, \partial_t \boldsymbol{U}^n\right) + \left(P^n, \nabla \cdot \partial_t \boldsymbol{U}^n\right) = 0, \tag{5.12}$$

$$\left(\lambda^{-1}\left(P^{n+1} - P^{n-1}\right), \frac{P^n}{2\Delta t}\right) - \left(\nabla \cdot \left(\boldsymbol{U}^{n+1} - \boldsymbol{U}^{n-1}\right), \frac{P^n}{2\Delta t}\right) = 0. \tag{5.13}$$

If we add (5.12) and (5.13), we obtain

$$\left(\rho \partial_t^2 \boldsymbol{U}^n, \frac{\boldsymbol{U}^{n+1} - \boldsymbol{U}^{n-1}}{2\Delta t}\right) + \left(\lambda^{-1}\left(P^{n+1} - P^{n-1}\right), \frac{P^n}{2\Delta t}\right) = 0. \tag{5.14}$$

Substituting (5.1) into (5.14) above yields

$$\begin{aligned}(5.15)\quad &\frac{1}{2\Delta t}\left(\rho\left(\partial_t \boldsymbol{U}^{n+\frac{1}{2}}-\partial_t \boldsymbol{U}^{n-\frac{1}{2}}\right),\partial_t \boldsymbol{U}^{n+\frac{1}{2}}+\partial_t \boldsymbol{U}^{n-\frac{1}{2}}\right)+\left(\lambda^{-1}\partial_t P^n,P^n\right)\\ &=\frac{1}{2\Delta t}\left(\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\right)+\left(\lambda^{-1}\partial_t P^n,P^n\right)=0.\end{aligned}$$

We now examine $(\partial_t P^n, P^n)$. We have

$$(5.16)\qquad \partial_t P^n=\frac{P^{n+1}+P^n-P^n-P^{n-1}}{2\Delta t}=\frac{\left(P^{n+\frac{1}{2}}-P^{n-\frac{1}{2}}\right)}{\Delta t}.$$

We can rewrite P^n as

$$\begin{aligned}P^n&=\frac{P^n+P^{n+1}}{4}+\frac{P^{n-1}+P^n}{4}-\frac{\Delta t^2}{4}\left(\frac{P^{n+1}-2P^n+P^{n-1}}{\Delta t^2}\right)\\ &=\frac{P^n+P^{n+1}}{4}+\frac{P^{n-1}+P^n}{4}-\frac{\Delta t^2}{4}\partial_t^2 P^n\\ (5.17)\qquad &=\frac{P^{n+\frac{1}{2}}}{2}+\frac{P^{n-\frac{1}{2}}}{2}-\frac{\Delta t^2}{4}\partial_t^2 P^n.\end{aligned}$$

If we use (5.16) and (5.17) in $(\partial_t P^n, P^n)$, we have

$$\begin{aligned}\left(\lambda^{-1}\partial_t P^n,P^n\right)&=\frac{1}{2\Delta t}\left(\lambda^{-1}\left(P^{n+\frac{1}{2}}-P^{n-\frac{1}{2}}\right),P^{n+\frac{1}{2}}+P^{n-\frac{1}{2}}\right)\\ &\quad-\left(\lambda^{-1}\partial_t P^n,\frac{\Delta t^2}{4}\partial_t^2 P^n\right)\\ &=\frac{1}{2\Delta t}\left(\lambda^{-1}\left(P^{n+\frac{1}{2}}-P^{n-\frac{1}{2}}\right),P^{n+\frac{1}{2}}+P^{n-\frac{1}{2}}\right)\\ &\quad-\left(\frac{\lambda^{-1}}{2}\left(\partial_t P^{n+\frac{1}{2}}+\partial_t P^{n-\frac{1}{2}}\right),\frac{\Delta t^2}{4}\frac{\left(\partial_t P^{n+\frac{1}{2}}-\partial_t P^{n-\frac{1}{2}}\right)}{\Delta t}\right)\\ &=\frac{1}{2\Delta t}\left[\left\|\lambda^{-\frac{1}{2}}P^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\lambda^{-\frac{1}{2}}P^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\right.\\ &\qquad\left.-\frac{\Delta t^2}{4}\left(\left\|\lambda^{-\frac{1}{2}}\partial_t P^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\lambda^{-\frac{1}{2}}\partial_t P^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\right)\right].\end{aligned}$$

Thus (5.15) becomes

$$\begin{aligned}(5.18)\quad &\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\\ &\qquad+\left\|\lambda^{-\frac{1}{2}}P^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\lambda^{-\frac{1}{2}}P^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\\ &\qquad-\frac{\Delta t^2}{4}\left(\left\|\lambda^{-\frac{1}{2}}\partial_t P^{n+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\left\|\lambda^{-\frac{1}{2}}\partial_t P^{n-\frac{1}{2}}\right\|^2_{L^2(\Omega)}\right)=0.\end{aligned}$$

If we sum (5.18) from $n=1,\ldots,N$, we get

$$\begin{aligned}&\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{N+\frac{1}{2}}\right\|^2_{L^2(\Omega)}+\left\|\lambda^{-\frac{1}{2}}P^{N+\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\frac{\Delta t^2}{4}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}}\right\|^2_{L^2(\Omega)}\\ &\qquad=\left\|\rho^{\frac{1}{2}}\partial_t \boldsymbol{U}^{\frac{1}{2}}\right\|^2_{L^2(\Omega)}+\left\|\lambda^{-\frac{1}{2}}P^{\frac{1}{2}}\right\|^2_{L^2(\Omega)}-\frac{\Delta t^2}{4}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{\frac{1}{2}}\right\|^2_{L^2(\Omega)}.\end{aligned}$$

We recall from (5.11) that

$$\left(\lambda^{-1}\left(P^{n+1}-P^{n}\right),w\right)-\left(\nabla\cdot\left(\boldsymbol{U}^{n+1}-\boldsymbol{U}^{n}\right),w\right)=0.$$

Then, by the Cauchy–Schwarz inequality and (5.9),

$$\begin{aligned}\left(\lambda^{-\frac{1}{2}}\left(P^{N+1}-P^{N}\right),w\right)&=\left(\nabla\cdot\left(\boldsymbol{U}^{N+1}-\boldsymbol{U}^{N}\right),w\right)\\&\leq\left\|\nabla\cdot\left(\boldsymbol{U}^{N+1}-\boldsymbol{U}^{N}\right)\right\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}\\&\leq\frac{C_0}{h}\left\|\boldsymbol{U}^{N+1}-\boldsymbol{U}^{N}\right\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}\\&\leq\frac{C_0}{h}\left(\Delta t\right)\left\|\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}.\end{aligned}$$

Thus

$$\left(\Delta t\right)\left(\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}},w\right)\leq\frac{C_0}{h}\left(\Delta t\right)\left\|\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}.$$

We choose $w=\partial_t P^{N+\frac{1}{2}}$, so

$$\begin{aligned}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2&\leq\frac{C_0}{h}\left\|\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\left\|\partial_t P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\\&\leq\frac{C_0\lambda_1^{\frac{1}{2}}}{h\rho_0^{\frac{1}{2}}}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\end{aligned}$$

or

$$\left\|\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}\leq\frac{C_0\lambda_1^{\frac{1}{2}}}{h\rho_0^{\frac{1}{2}}}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}.$$

Hence

$$\begin{aligned}&\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{\frac{1}{2}}\right\|_{L^2(\Omega)}^2+\left\|\lambda^{-\frac{1}{2}}P^{\frac{1}{2}}\right\|_{L^2(\Omega)}^2-\frac{\Delta t^2}{4}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\quad=\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2+\left\|\lambda^{-\frac{1}{2}}P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\qquad-\frac{\Delta t^2}{4}\left\|\lambda^{-\frac{1}{2}}\partial_t P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\quad\geq\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2+\left\|\lambda^{-\frac{1}{2}}P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\qquad-\frac{\Delta t^2C_0^2\lambda_1}{4h^2\rho_0}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\quad=\left(1-\frac{\Delta t^2C_0^2\lambda_1}{4h^2\rho_0}\right)\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{U}^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2\\&\qquad+\left\|\lambda^{-\frac{1}{2}}P^{N+\frac{1}{2}}\right\|_{L^2(\Omega)}^2.\end{aligned}$$

Thus the temporal iterates are bound by the initial data and the discrete in time scheme is stable if $\Delta t<\frac{2h\rho_0^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}}C_0}$. ∎

5.3. Discrete in time a priori error estimates. We now can derive the estimates for the fully discrete scheme.

THEOREM 5.2. *If* $\boldsymbol{u} \in L^\infty(\boldsymbol{H}(\Omega;\mathrm{div}))$, $\frac{\partial^3 \boldsymbol{u}}{\partial t^3} \in L^1\left(\boldsymbol{L}^2(\Omega)\right)$ $\frac{\partial^4 \boldsymbol{u}}{\partial t^4} \in L^\infty\left(\boldsymbol{L}^2(\Omega)\right)$, *and* $p \in L^\infty(L^2(\Omega))$, *then for* $\{\boldsymbol{U}^n, P^n\}$ *defined by* (5.3)–(5.6) *there exists a constant* C *independent of* h *and* Δt *such that if* $\Delta t < \frac{2h\rho_0^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}} C_0}$, *then*

$$\begin{aligned}(5.19)\quad &\left\|\rho^{\frac{1}{2}}(\boldsymbol{u}-\boldsymbol{U})\right\|_{l^\infty(L^2)} + \left\|\lambda^{\frac{1}{2}}(p-P)\right\|_{l^\infty(L^2)} \\ &\qquad \le C\left(h^k + \Delta t^2\right)\left(\|\boldsymbol{u}\|_{L^\infty(H^k)} + \left\|\frac{\partial^3 \boldsymbol{u}}{\partial t^3}\right\|_{L^\infty(L^2)} + \|p\|_{L^\infty(L^2)}\right),\end{aligned}$$

where k *is associated with the degree of the finite element polynomial.*

Proof. From (3.9)–(3.10) and (5.5)–(5.6), and by the properties of the L^2 and Π_h projections, we can write

$$\left(\rho\partial_t^2 \boldsymbol{\chi}^n, \boldsymbol{v}\right) + \left(\xi^n, \nabla\cdot\boldsymbol{v}\right) = \left(\rho\partial_t^2\boldsymbol{\eta}^n, \boldsymbol{v}\right) + \left(\boldsymbol{r}^n, \boldsymbol{v}\right), \tag{5.20}$$

$$\left(\lambda^{-1}\xi^{n+1}, w\right) - \left(\nabla\cdot\boldsymbol{\chi}^{n+1}, w\right) = \left(\lambda^{-1}\zeta^{n+1}, w\right), \tag{5.21}$$

where $\boldsymbol{r}^n = \rho(\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^n) - \partial_t^2 \boldsymbol{u}^n)$. We introduce

$$\phi^0 = \frac{\Delta t}{2}\xi^0, \quad \phi^n = \frac{\Delta t}{2}\xi^0 + \Delta t\sum_{i=1}^n \xi^i.$$

Using (5.1) in (5.20) gives

$$\left(\rho\frac{\partial_t\boldsymbol{\chi}^{n+\frac{1}{2}} - \partial_t\boldsymbol{\chi}^{n-\frac{1}{2}}}{\Delta t}, \boldsymbol{v}\right) + \left(\xi^n, \nabla\cdot\boldsymbol{v}\right) = \left(\rho\frac{\partial_t\boldsymbol{\eta}^{n+\frac{1}{2}} - \partial_t\boldsymbol{\eta}^{n-\frac{1}{2}}}{\Delta t}, \boldsymbol{v}\right) + \left(\boldsymbol{r}^n, \boldsymbol{v}\right).$$

Summing over time levels and multiplying through by Δt gives

$$\begin{aligned}(5.22)\quad &\left(\rho\left(\partial_t\boldsymbol{\chi}^{n+\frac{1}{2}} - \partial_t\boldsymbol{\chi}^{\frac{1}{2}}\right), \boldsymbol{v}\right) + \left(\phi^n - \phi^0, \nabla\cdot\boldsymbol{v}\right) \\ &\qquad = \left(\rho\left(\partial_t\boldsymbol{\eta}^{n+\frac{1}{2}} - \partial_t\boldsymbol{\eta}^{\frac{1}{2}}\right), \boldsymbol{v}\right) + \left(\Delta t\sum_{i=1}^n \boldsymbol{r}^i, \boldsymbol{v}\right),\end{aligned}$$

since $\Delta t\sum_{i=1}^n \xi^i = \phi^n - \phi^0$.

Using (5.4), we obtain

$$\begin{aligned}&\left(\rho\partial_t\boldsymbol{\chi}^{\frac{1}{2}}, \boldsymbol{v}\right) + \frac{\Delta t}{2}\left(\xi^0, \nabla\cdot\boldsymbol{v}\right) \\ &\qquad = \left(\rho\partial_t\boldsymbol{\eta}^{\frac{1}{2}}, \boldsymbol{v}\right) + \left(\rho(\Pi\boldsymbol{u}_1 - \boldsymbol{u}_1), \boldsymbol{v}\right) - \frac{1}{2\Delta t}\int_0^{\Delta t}\rho\left(\Delta t - t\right)^2\left(\frac{\partial^3\boldsymbol{u}}{\partial t^3}, \boldsymbol{v}\right)dt,\end{aligned}$$

and thus (5.22) reduces to

$$\left(\rho\partial_t\boldsymbol{\chi}^{n+\frac{1}{2}}, \boldsymbol{v}\right) + \left(\phi^n, \nabla\cdot\boldsymbol{v}\right) = \left(\rho\partial_t\boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{v}\right) + \left(\boldsymbol{R}^n, \boldsymbol{v}\right), \tag{5.23}$$

where $\boldsymbol{R}^n$ is defined as

$$\boldsymbol{R}^n = \Delta t \sum_{i=1}^{n} \boldsymbol{r}^i + \rho(\Pi_h \boldsymbol{u}_1 - \boldsymbol{u}_1) - \frac{1}{2\Delta t} \int_0^{\Delta t} \rho \left(\Delta t - t\right)^2 \frac{\partial^3 \boldsymbol{u}}{\partial t^3}(t)\, dt.$$

We now rewrite (5.21) by noting that $\xi^{n+1} = \partial_t \phi^{n+\frac{1}{2}}$, so that

$$\left(\lambda^{-1} \partial_t \phi^{n+\frac{1}{2}}, w\right) - \left(\nabla \cdot \boldsymbol{\chi}^{n+1}, w\right) = (\lambda^{-1} \zeta^{n+1}, w). \tag{5.24}$$

We choose $\boldsymbol{v} = \boldsymbol{\chi}^{n+\frac{1}{2}}$ and $w = \phi^{n+\frac{1}{2}}$, which, when substituted into (5.23) and (5.24), gives

$$\left(\rho \partial_t \boldsymbol{\chi}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) + \left(\phi^n, \nabla \cdot \boldsymbol{\chi}^{n+\frac{1}{2}}\right) = \left(\rho \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) + \left(\boldsymbol{R}^n, \boldsymbol{\chi}^{n+\frac{1}{2}}\right), \tag{5.25}$$

$$\left(\lambda^{-1} \partial_t \phi^{n+\frac{1}{2}}, \phi^{n+\frac{1}{2}}\right) - \left(\nabla \cdot \boldsymbol{\chi}^{n+1}, \phi^{n+\frac{1}{2}}\right) = \left(\lambda^{-1} \zeta^{n+1}, \phi^{n+\frac{1}{2}}\right). \tag{5.26}$$

We expand (5.25) and (5.26) to get

$$\begin{aligned}
&\left(\rho \left(\frac{\boldsymbol{\chi}^{n+1} - \boldsymbol{\chi}^n}{\Delta t}\right), \frac{\boldsymbol{\chi}^{n+1} + \boldsymbol{\chi}^n}{2}\right) + \left(\phi^n, \frac{\nabla \cdot \left(\boldsymbol{\chi}^{n+1} + \boldsymbol{\chi}^n\right)}{2}\right) \\
&\qquad = \left(\rho \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) + \left(\boldsymbol{R}^n, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) \\
&\left(\lambda^{-1} \left(\frac{\phi^{n+1} - \phi^n}{\Delta t}\right), \frac{\phi^{n+1} + \phi^n}{2}\right) - \left(\nabla \cdot \boldsymbol{\chi}^{n+1}, \frac{\phi^{n+1} + \phi^n}{2}\right) = \left(\lambda^{-1} \zeta^{n+1}, \phi^{n+\frac{1}{2}}\right),
\end{aligned}$$

so that

$$\begin{aligned}
&\frac{1}{2\Delta t} \left(\left\|\rho^{\frac{1}{2}} \boldsymbol{\chi}^{n+1}\right\|_{L^2(\Omega)}^2 - \left\|\rho^{\frac{1}{2}} \boldsymbol{\chi}^n\right\|_{L^2(\Omega)}^2\right) + \frac{1}{2}\left(\phi^n, \nabla \cdot \boldsymbol{\chi}^{n+1}\right) + \frac{1}{2}\left(\phi^n, \nabla \cdot \boldsymbol{\chi}^n\right) \\
&\qquad = \left(\rho \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) + \left(\boldsymbol{R}^n, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) \qquad (5.27) \\
&\frac{1}{2\Delta t} \left(\left\|\lambda^{-\frac{1}{2}} \phi^{n+1}\right\|_{L^2(\Omega)}^2 - \left\|\lambda^{-\frac{1}{2}} \phi^n\right\|_{L^2(\Omega)}^2\right) = \frac{1}{2}\left(\nabla \cdot \boldsymbol{\chi}^{n+1}, \phi^n\right) + \frac{1}{2}\left(\nabla \cdot \boldsymbol{\chi}^{n+1}, \phi^{n+1}\right) \\
&\qquad + \left(\lambda^{-1} \zeta^{n+1}, \phi^{n+\frac{1}{2}}\right). \qquad (5.28)
\end{aligned}$$

Adding (5.27) and (5.28) and multiplying by $2\Delta t$ gives

$$\begin{aligned}
&\left\|\rho^{\frac{1}{2}} \boldsymbol{\chi}^{n+1}\right\|_{L^2(\Omega)}^2 - \left\|\rho^{\frac{1}{2}} \boldsymbol{\chi}^n\right\|_{L^2(\Omega)}^2 + \left\|\lambda^{-\frac{1}{2}} \phi^{n+1}\right\|_{L^2(\Omega)}^2 - \left\|\lambda^{-\frac{1}{2}} \phi^n\right\|_{L^2(\Omega)}^2 \\
&\qquad + \Delta t \left[\left(\phi^n, \nabla \cdot \boldsymbol{\chi}^n\right) - \left(\nabla \cdot \boldsymbol{\chi}^{n+1}, \phi^{n+1}\right)\right] \\
&\qquad = 2\Delta t \left[\left(\rho \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) + \left(\boldsymbol{R}^n, \boldsymbol{\chi}^{n+\frac{1}{2}}\right)\right] + 2\Delta t \left(\lambda^{-1} \zeta^{n+1}, \phi^{n+\frac{1}{2}}\right).
\end{aligned} \tag{5.29}$$

The terms on the right-hand side are bound using the Cauchy–Schwarz inequality as

$$\begin{aligned}
\left(\rho \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) &\le \left\|\rho^{\frac{1}{2}} \partial_t \boldsymbol{\eta}^{n+\frac{1}{2}}\right\|_{L^2(\Omega)} \left\|\boldsymbol{\chi}^{n+\frac{1}{2}}\right\|_{L^2(\Omega)} \\
\left(\boldsymbol{R}^n, \boldsymbol{\chi}^{n+\frac{1}{2}}\right) &\le \left\|\boldsymbol{R}^n\right\|_{L^2(\Omega)} \left\|\boldsymbol{\chi}^{n+\frac{1}{2}}\right\|_{L^2(\Omega)} \\
\left(\lambda^{-1} \zeta^{n+1}, \phi^{n+\frac{1}{2}}\right) &\le \left\|\lambda^{-\frac{1}{2}} \zeta^{n+1}\right\|_{L^2(\Omega)} \left\|\lambda^{-\frac{1}{2}} \phi^{n+\frac{1}{2}}\right\|_{L^2(\Omega)}.
\end{aligned}$$

Sum (5.29) over time levels to get

$$\begin{aligned}
&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} - \left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{0}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)} \\
&\quad - \left\|\lambda^{-\frac{1}{2}}\phi^{0}\right\|^2_{L^2(\Omega)} - \Delta t\left[\left(\nabla\cdot\boldsymbol{\chi}^{n+1},\phi^{n+1}\right) - \left(\phi^0,\nabla\cdot\boldsymbol{\chi}^0\right)\right] \\
&\le 2\Delta t\sum_{i=0}^{n}\left(\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} + \left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\right) \\
&\quad +\Delta t\sum_{i=0}^{n}\left(\left\|\lambda^{-\frac{1}{2}}\zeta^{i+1}\right\|_{L^2(\Omega)}\left(\left\|\lambda^{-\frac{1}{2}}\phi^{i+1}\right\|_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{i}\right\|_{L^2(\Omega)}\right)\right). \qquad (5.30)
\end{aligned}$$

After imposing the initial conditions (5.2) and (5.3) in (5.30) we have

$$\begin{aligned}
&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)} - \Delta t\left(\nabla\cdot\boldsymbol{\chi}^{n+1},\phi^{n+1}\right) \\
&\le 2\Delta t\sum_{i=0}^{n}\left(\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} + \left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\right) \\
&\quad + \Delta t\sum_{i=0}^{n}\left(\left\|\lambda^{-\frac{1}{2}}\zeta^{i+1}\right\|_{L^2(\Omega)}\left(\left\|\lambda^{-\frac{1}{2}}\phi^{i+1}\right\|_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{i}\right\|_{L^2(\Omega)}\right)\right).
\end{aligned}$$

Using the Cauchy–Schwarz inequality and the inverse assumption (5.9) and choosing h and Δt such that $\Delta t < \frac{2h\rho_0^{\frac{1}{2}}}{C_0\lambda_1^{\frac{1}{2}}}$, we have that

$$\begin{aligned}
\Delta t\left(\nabla\cdot\boldsymbol{\chi}^{n+1},\phi^{n+1}\right) &\le \Delta t\left\|\nabla\cdot\boldsymbol{\chi}^{n+1}\right\|_{L^2(\Omega)}\left\|\phi^{n+1}\right\|_{L^2(\Omega)} \\
&\le \Delta t C_0 h^{-1}\left\|\boldsymbol{\chi}^{n+1}\right\|_{L^2(\Omega)}\left\|\phi^{n+1}\right\|_{L^2(\Omega)} \\
&\le \frac{\Delta t C_0\lambda_1^{\frac{1}{2}}}{2h\rho_0^{\frac{1}{2}}}\left(\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)}\right) \\
&< \left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)} \\
&\le C\Delta t\sum_{i=0}^{n}\left(\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} + \left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\right) \\
&\qquad + 2\left\|\lambda^{-\frac{1}{2}}\phi\right\|_{l^\infty(L^2)}\left(\Delta t\sum_{i=0}^{n}\left\|\lambda^{-\frac{1}{2}}\zeta^{i+1}\right\|_{L^2(\Omega)}\right).
\end{aligned}$$

Since $\left\|\boldsymbol{\chi}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} \leq \|\boldsymbol{\chi}\|_{l^\infty(L^2)}$, then

$$\begin{aligned}
&\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}^{n+1}\right\|^2_{L^2(\Omega)} + \left\|\lambda^{-\frac{1}{2}}\phi^{n+1}\right\|^2_{L^2(\Omega)} \\
&\quad\leq \frac{C2\Delta t}{\rho_0^{\frac{1}{2}}}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|_{l^\infty(L^2)}\left(\sum_{i=0}^{n}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} + \sum_{i=0}^{n}\left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\right) \\
&\qquad\quad + \frac{1}{4}\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{l^\infty(L^2)} + C\left(\Delta t\sum_{i=0}^{n}\left\|\lambda^{-\frac{1}{2}}\zeta^{i+1}\right\|_{L^2(\Omega)}\right)^2 \\
&\quad\leq \frac{1}{4}\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{l^\infty(L^2)} + C\Delta t^2\left(\sum_{i=0}^{N}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\right)^2 \\
&\qquad\quad + C\Delta t^2\left(\sum_{i=0}^{N}\left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\right)^2 + \frac{1}{4}\left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{l^\infty(L^2)} \\
&\qquad\quad + C\left(\Delta t\sum_{i=0}^{n}\left\|\lambda^{-\frac{1}{2}}\zeta^{i+1}\right\|_{L^2(\Omega)}\right)^2.
\end{aligned}$$

If we take the supremum on n on the left-hand side we get

$$\begin{aligned}
\left\|\rho^{\frac{1}{2}}\boldsymbol{\chi}\right\|^2_{l^\infty(L^2)} + \left\|\lambda^{-\frac{1}{2}}\phi\right\|^2_{l^\infty(L^2)} &\leq C\left\|\lambda^{-\frac{1}{2}}\zeta\right\|^2_{l^\infty(L^2)} \\
&\quad + C\Delta t^2\left(\sum_{i=0}^{N}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)}\right)^2 + C\Delta t^2\left(\sum_{i=0}^{N}\left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)}\right)^2.
\end{aligned}$$

The first two terms in the right-hand side of the previous inequality can be bound using the approximation properties and the bound

$$\Delta t\sum_{i=0}^{N}\left\|\rho^{\frac{1}{2}}\partial_t\boldsymbol{\eta}^{i+\frac{1}{2}}\right\|_{L^2(\Omega)} \leq C\left(h^k\left\|\boldsymbol{u}\right\|_{L^\infty(H^k(\Omega))} + \Delta t^2\left\|\frac{\partial^3\boldsymbol{u}}{\partial t^3}\right\|_{L^1(0,T;L^2(\Omega))}\right).$$

We bound the last term by

$$\begin{aligned}
\Delta t\sum_{i=0}^{N}\left\|\boldsymbol{R}^i\right\|_{L^2(\Omega)} &\leq C\left\|\boldsymbol{R}\right\|_{l^\infty(L^2)} \\
&\leq C\Delta t\sum_{i=1}^{N}\left\|\boldsymbol{r}^i\right\|_{L^2(\Omega)} + C\left\|\rho\left(\Pi_h\boldsymbol{u}_1 - \boldsymbol{u}_1\right)\right\|_{L^2(\Omega)} \\
&\quad + C\left\|\frac{1}{2\Delta t}\int_0^{\Delta t}\rho(\Delta t - t)^2\frac{\partial^3\boldsymbol{u}}{\partial t^3}(t)\,dt\right\|_{L^2(\Omega)}.
\end{aligned}$$

We obtain a bound on $\|\boldsymbol{r}^i\|_{L^2(\Omega)}$ by first expanding $\boldsymbol{r}^i$ as

$$\boldsymbol{r}^i = \rho(\boldsymbol{u}^i_{tt} - \partial_t^2\boldsymbol{u}^i). \tag{5.31}$$

We note that

$$\partial_t^2\boldsymbol{u}^i = \frac{\boldsymbol{u}^{i+1} - 2\boldsymbol{u}^i + \boldsymbol{u}^{i-1}}{\Delta t^2} \tag{5.32}$$

and use Taylor series to expand $\boldsymbol{u}^{i+1}$ and $\boldsymbol{u}^{i-1}$. We have

$$\begin{aligned}\boldsymbol{u}^{i-1} &= \boldsymbol{u}\left(t_i - \Delta t\right)\\ &= \boldsymbol{u}(t^i) - \Delta t\frac{\partial \boldsymbol{u}}{\partial t}(t^i) + \frac{\Delta t^2}{2}\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) - \frac{\Delta t^3}{6}\frac{\partial^3 \boldsymbol{u}}{\partial t^3}(t^i)\\ &\quad + \frac{1}{6}\int_{t^i-\Delta t}^{t^i}\left(t^i - \Delta t - t\right)^3\frac{\partial^4 \boldsymbol{u}}{\partial t^4}(t)\,dt\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{u}^{i+1} &= \boldsymbol{u}(t_i + \Delta t)\\ &= \boldsymbol{u}(t^i) + \Delta t\frac{\partial \boldsymbol{u}}{\partial t}(t^i) + \frac{\Delta t^2}{2}\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) + \frac{\Delta t^3}{6}\frac{\partial^3 \boldsymbol{u}}{\partial t^3}(t^i)\\ &\quad + \frac{1}{6}\int_{t^i}^{t^i+\Delta t}\left(t^i + \Delta t - t\right)^3\frac{\partial^4 \boldsymbol{u}}{\partial t^4}(t)\,dt\end{aligned}$$

so that

$$\begin{aligned}\boldsymbol{u}^{i+1} + \boldsymbol{u}^{i-1} &= 2\boldsymbol{u}(t^i) + \Delta t^2\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) + \frac{1}{6}\left[\int_{t^i}^{t^i+\Delta t}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}(t)\left(t^i + \Delta t - t\right)^3\,dt\right.\\ &\qquad \left. + \int_{t^i-\Delta t}^{t^i}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}(t)\left(t^i - \Delta t - t\right)^2\,dt\right]\\ &= 2\boldsymbol{u}(t^i) + \Delta t^2\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) + \frac{1}{6}\int_{-\Delta t}^{\Delta t}\left(|t| - \Delta t\right)^3\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\left(t^i + t\right)\,dt. \qquad (5.33)\end{aligned}$$

When we use (5.33) and (5.32) in (5.31) we get

$$\begin{aligned}\boldsymbol{r}^i &= \rho\left(\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) - \partial_t^2\boldsymbol{u}^i\right)\\ &= \frac{\rho}{6\Delta t^2}\int_{-\Delta t}^{\Delta t}\left(|t| - \Delta t\right)^3\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\left(t^i + t\right)\,dt\end{aligned}$$

and thus

$$\boldsymbol{r}^i = \rho\left(\frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t^i) - \partial_t^2\boldsymbol{u}^i\right) = \frac{\rho}{6\Delta t^2}\int_{-\Delta t}^{\Delta t}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\left(t^i + \Delta t\right)\left(|t| - \Delta t\right)^3\,dt.$$

Therefore

$$\begin{aligned}\left\|\boldsymbol{r}^i\right\|_{L^2(\Omega)}^2 &\le C\Delta t^2\int_\Omega\left[\int_{t^i-\Delta t}^{t^i+\Delta t}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}(t)\right]^2\,dt\\ &\le C\Delta t^3\int_{t^i-\Delta t}^{t^i+\Delta t}\left\|\rho^{\frac{1}{2}}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\right\|_{L^2(\Omega)}^2\\ &\le C\Delta t^4\left\|\rho^{\frac{1}{2}}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\right\|_{L^\infty(L^2)}^2,\end{aligned}$$

so that

$$\Delta t\sum_{i=1}^n\left\|\boldsymbol{r}^i\right\|_{L^2(\Omega)} \le C\Delta t^2\left\|\rho^{\frac{1}{2}}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\right\|_{L^\infty(L^2)}\sum_{i=1}^n\Delta t \le C\Delta t^2\left\|\rho^{\frac{1}{2}}\frac{\partial^4 \boldsymbol{u}}{\partial t^4}\right\|_{L^\infty(L^2)}.$$

Similarly,

$$\begin{aligned}\left\| \frac{1}{2\Delta t} \int_0^{\Delta t} \rho(\Delta t - t)^2 u^{(3)}(t) \right\|_{L^2(\Omega)} &\leq C\Delta t^2 \int_\Omega \left(\int_0^{\Delta t} \rho \frac{\partial^3 \boldsymbol{u}}{\partial t^3}(t)\, dt \right)^2 \\ &\leq C\Delta t^3 \int_0^{\Delta t} \left\| \rho^{\frac{1}{2}} \frac{\partial^3 \boldsymbol{u}}{\partial t^3} \right\|^2_{L^2(\Omega)} dt \\ &\leq C\Delta t^4 \left\| \rho^{\frac{1}{2}} \frac{\partial^3 \boldsymbol{u}}{\partial t^3} \right\|^2_{L^\infty(L^2)}.\end{aligned}$$

Finally, using the approximation result (2.2) and combining all the bounds, we get

$$\Delta t \sum_{i=0}^{N} \left\| \boldsymbol{R}^i \right\|_{L^2(\Omega)} \leq C(h^k + \Delta t^2),$$

which concludes the proof of the discrete estimate. ∎

Remark. The convergence rate in space can be at most quadratic, as $\Delta t < Ch$.

6. Conclusions. We have developed a priori error estimates for mixed finite element displacement formulations of the acoustic wave equation. Our scheme maintains the same computational complexity as earlier mixed finite element formulations for second order hyperbolic equations, as we have not introduced any additional unknowns. Our formulations require less regularity on the displacement than standard approaches.

We have shown convergence of the scheme via our continuous-in-time estimates, and we have shown that in the temporally discrete case we expect a quadratic convergence rate.

Acknowledgment. The authors would like to thank Dr. Mrinal Sen of the Institute for Geophysics at the University of Texas at Austin for his insights into this problem.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure and Appl. Math. 65, Academic Press, New York, 1975.
[2] G. A. Baker, *Error estimates for finite element methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 13 (1976), pp. 564–576.
[3] H. Bao, J. Bielak, O. Ghattas, L. F. Kallivokas, D. R. O'Hallaron, J. R. Shewchuk, and J. Xu, *Large-scale simulation of elastic wave propagation in heterogeneous media on parallel computers*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 85–102.
[4] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.
[5] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.
[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[7] L. C. Cowsar, T. F. Dupont, and M. F. Wheeler, *A priori estimates for mixed finite element methods for the wave equation*, Comput. Methods App. Mech. Engrg. 82 (1990), pp. 205–222.
[8] L. C. Cowsar, T. F. Dupont, and M. F. Wheeler, *A priori estimates for mixed finite element approximations of second-order hyperbolic equations with absorbing boundary conditions*, SIAM J. Numer. Anal., 33 (1996), pp. 492–504.

[9] T. Dupont, *L^2-estimates for Galerkin methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 880–889.
[10] R. W. Graves, *Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences*, Bull. Seismol. Soc. Amer., 86 (1996), pp. 1091–1106.
[11] C. Johnson, *Discontinuous Galerkin finite element methods for second order hyperbolic problems*, Comput. Methods App. Mech. Engrg. 107 (1993), pp. 117–129.
[12] R. J. Knops and L. E. Payne, *Uniqueness Theorems in Linear Elasticity*, Springer Tracts Nat. Philos. 19, Springer-Verlag, Berlin, 1971.
[13] K. J. Marfurt, *Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations*, Geophysics, 49 (1984), pp. 533–549.
[14] R. McOwen, *Partial Differential Equations: Methods and Applications*, Prentice-Hall, Upper Saddle River, NJ, 1995.
[15] J. C. Nedelec, *Mixed finite elements in $\mathcal{R}^3$*, Numer. Math., 35 (1980), pp. 315–341.
[16] R. A. Raviart and J. M. Thomas, *Mathematical Aspects of the Finite Element Method,* Lecture Notes in Math., 106, Springer-Verlag, Berlin, 1997, pp. 292–315.
[17] B. Rivière and M. Wheeler, *Discontinuous Finite Element Methods for Acoustic and Elastic Wave Problems. Part* I: *Semidiscrete Error Estimates*, TICAM report 01-02, University of Texas, Austin, TX, 2001.
[18] F. J. Sabadell, F. J. Serón, and J. Badal, *A parallel laboratory for simulation and visualization of seismic wavefields*, Geophys. Prospecting, 48 (2000), pp. 377–398.
[19] M. K. Sen and P. L. Stoffa, *Global Optimization Methods in Geophysical Inversion*, Elsevier Science, Amsterdam, The Netherlands, 1995.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1716–1762

SIMULATION OF THE SOLUTION OF A VISCOUS POROUS MEDIUM EQUATION BY A PARTICLE METHOD*

KARL OELSCHLÄGER†

Abstract. To solve a particular partial differential equation, namely a viscous porous medium equation, we discuss a particle method, which is based on the concept of moderately interacting many-particle systems. Our approach may be classified as a combination of a smoothed particle hydrodynamics method and a particle-mesh method. Quantitatively, it is assessed in terms of estimates on the approximation error.

Key words. particle methods, many-particle systems, smoothed particle hydrodynamics

AMS subject classifications. 65C35, 60K35, 65M12, 68U20

PII. S0036142900363377

1. Introduction.

In this paper we consider the numerical simulation of a *viscous porous medium equation*,

$$\partial_t \rho = \frac{1}{2}\Delta\rho + \frac{1}{2}\Delta\rho^2 = \frac{1}{2}\Delta\rho + \nabla\cdot(\rho\nabla\rho), \qquad \rho(.,0) = \rho_0, \tag{1.1}$$

by a *particle method*, which is based on the concept of *moderately interacting many-particle systems*.

The approximation of the solution ρ of (1.1) by the *empirical processes* of such many-particle systems has been studied analytically in [12] and [13]. In those papers we have for any $N \in \mathbb{N}$ a family of N particles in $\mathbb{R}^d$, whose positions at time $t \geq 0$ are denoted by $X_N^k(t)$, $k = 1, \ldots, N$. It is assumed that they evolve according to the system

$$dX_N^k(t) = -\frac{1}{N}\sum_{\substack{m=1,\ldots,N\\ m\neq k}} \nabla\phi_N\big(X_N^k(t) - X_N^m(t)\big)dt + dW^k(t), \qquad k = 1, \ldots, N,\ t \geq 0, \tag{1.2}$$

of coupled *stochastic differential equations*. In (1.2) the processes $W^1, W^2, \ldots$ are independent, standard Brownian motions in $\mathbb{R}^d$. Moreover, the *interaction potential* ϕ_N is obtained from a fixed function ϕ_1 by the scaling

$$\phi_N(x) = \theta_N^d \phi_1(\theta_N x), \quad x \in \mathbb{R}^d,\ N \in \mathbb{N}, \tag{1.3}$$

where

$$\theta_N = N^{\beta/d} \quad \text{for some} \quad \beta \in \Big(0, \frac{d}{d+2}\Big), \tag{1.4}$$

and ϕ_1 is a differentiable, symmetric probability density, i.e.,

$$\phi_1 \in C_b^1(\mathbb{R}^d), \quad \phi_1 \geq 0, \quad \int_{\mathbb{R}^d} dx\ \phi_1(x) = 1, \quad \phi_1(x) = \phi_1(-x),\ x \in \mathbb{R}^d. \tag{1.5}$$

*Received by the editors May 3, 2000; accepted for publication (in revised form) February 28, 2002; published electronically November 14, 2002. This work was supported by the Deutsche Forschungsgemeinschaft.

http://www.siam.org/journals/sinum/40-5/36337.html

†Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-69120, Heidelberg, Germany (karl.oelschlaeger@urz.uni-heidelberg.de).

The scaling parameter θ_N represents the *inverse interaction range* in the system (1.2). In particular, β determines the decrease of the interaction range θ_N^{-1} in comparison to the *typical distance between neighboring particles*, which as explained below in this section is $O(N^{-1/d})$ as $N \to \infty$.

In [12] and [13] we discuss in particular the empirical processes

$$\mathbb{X}_N(t) = \frac{1}{N} \sum_{m=1}^{N} \delta_{X_N^m(t)}, \quad t \geq 0,\ N \in \mathbb{N}, \tag{1.6}$$

of the many-particle systems (1.2), where δ_a denotes the Dirac measure at $a \in \mathbb{R}^d$. If $\mathbb{X}_N(0)$, $N \in \mathbb{N}$, and ϕ_1 are sufficiently *regular*, the convergence

$$\lim_{N \to \infty} \mathbb{X}_N = \rho \tag{1.7}$$

of the empirical processes to the solution ρ of (1.1) can be demonstrated; cf. [12]. In particular, (1.7) is proved as L^2-convergence of a regularized version of $\mathbb{X}_N$ to ρ as $N \to \infty$. As an extension in [13] the rate of the convergence (1.7) is specified. More precisely, we deduce an expansion

$$\mathbb{X}_N(t) \sim \rho(.,t) + \sum_{r=1}^{\lfloor d/4\beta \rfloor} \theta_N^{-2r} \rho_r(.,t) + \frac{1}{\sqrt{N}} \zeta(t), \quad t \geq 0,\ N \to \infty,$$

where the functions ρ_r, $r = 1, \ldots, \lfloor d/4\beta \rfloor$, are solutions of suitable linear diffusion equations and ζ is some $\mathcal{S}'(\mathbb{R}^d)$-valued Gaussian process. Here, $\mathcal{S}'(\mathbb{R}^d)$ is the space of *tempered distributions* on $\mathbb{R}^d$ and $\lfloor z \rfloor = \max\{n \in \mathbb{Z} : n \leq z\}$, $z \in \mathbb{R}$.

The convergence result (1.7) suggests determining a numerical solution of the partial differential equation (1.1) by simulating the many-particle system (1.2) for sufficiently large N. Then, the associated empirical process can be considered as an approximation to the analytical solution of (1.1). Such an approach to solve an evolution equation by using a many-particle system as an auxiliary tool is called a *particle method*. Physically motivated applications can be found, e.g., in numerical studies in astro-, hydro-, or plasmaphysics, where they have a long tradition but very often a formal or heuristic basis; cf. [4], [8]. Mathematical investigations of particle methods usually refer to simpler evolution problems, which are considered to exhibit typical features and difficulties of more complex, realistic problems, and nevertheless are accessible to a detailed analysis. For example, particle simulations of the *Boltzmann equation*, cf., e.g., [11], or *Vlasov–Poisson–Fokker–Planck equations*, cf., e.g., [6], have been studied in detail. Furthermore, applications of particle methods to solve hydrodynamical equations, cf., e.g., [1], [10], or specific reaction-diffusion equations, cf., e.g., [2], [18], have been investigated. Also in the present paper we deal with a simple partial differential equation, namely (1.1), and utilize this context to propose a particular particle method together with an analytical investigation. We emphasize that our considerations should also be applicable in more general situations. In particular, it seems that they can be extended to such evolution equations, which like those in [14], [15], [16] can be approximated by suitable *moderately interacting many-particle systems*. Indeed, at least the computer program employed to perform the simulations, which will be described in section 6, has already been modified such that parabolic systems of partial differential equations as those discussed in [14] may be handled.

Moderately interacting many-particle systems are essentially characterized by the feature that the range of the different types of interactions between the particles is both large in comparison to the typical distance between neighboring particles and small in comparison to the spatial size of the whole system. For example, by (1.3)–(1.5) the range of the interaction in the system (1.2) is $O(\theta_N^{-1}) = O(N^{-\beta/d})$. On the other hand, the fact that to any particle the mass $1/N$ is attached, cf. (1.6), and the convergence relation (1.7) imply that in the situations studied in [12] and [13] the typical distance between neighboring particles is $O(N^{-1/d})$ as $N \to \infty$, and furthermore that the spatial size of the whole system is $O(1)$. In particular, since $\beta \in (0, d/(d+2))$, the above-mentioned characteristic of moderate interaction can be observed in this case. As an extension of [12] and [13] we have described and investigated in [14], [15], [16] various types of moderately interacting many-particle systems, which also may involve different species of particles. More precisely, we derive in these papers in the limit as the particle number tends to ∞ systems of reaction-diffusion equations, cf. [14], and particular systems of hydrodynamical equations, cf. [15] and [16], as limit dynamics for empirical processes. Although the dynamics of the many-particle systems in [14], [15], [16] may be much more complex than in (1.2), the crucial ideas to determine analytically the respective limit behavior are quite similar. In all these cases we consider vectors $(\mathbb{X}_{N,1}(t) * \kappa_N, \ldots, \mathbb{X}_{N,R}(t) * \kappa_N)$ of regularized versions of the empirical processes $\mathbb{X}_{N,1}(t), \ldots, \mathbb{X}_{N,R}(t)$, $t \geq 0$, $N \in \mathbb{N}$, for the various species, where R is the number of species and κ_N, $N \in \mathbb{N}$, a sequence of suitable convolution kernels converging to δ_0 as $N \to \infty$. As a result an L^2-convergence like $\lim_{N\to\infty} \sum_{r=1}^{R} \|\mathbb{X}_{N,r}(t) * \kappa_N - \rho_r(.,t)\|_2 = 0$, $t \geq 0$, to the solution $(\rho_1, \ldots, \rho_r)$ of the limit dynamics is deduced.

Of course, systems like (1.2), which are continuous in space and time, cannot be simulated directly on a computer. For that aim we have to deal with problems related to discretization. In particular, we have to replace the many-particle dynamics (1.2) by some modified version, where to a certain extent for *continuous features* discrete counterparts are employed. Consequently, we shall introduce in section 2 a family of many-particle systems, whose members differ in the number N of particles, a spatial mesh size δ, the time step h, and the scaling parameter θ for the interaction potential. In an extension, we shall also present a family of many-particle systems, whose members additionally differ in a parameter τ determining the adaption of the mesh to local spatial irregularities. More precisely, instead of (1.2) we shall define

- for any $N \in \mathbb{N}$ a family of systems of N coupled evolution equations
- in discrete time with time step $h \in (0,1)$ and
- an interaction potential ϕ_θ obtained by the scaling

$$\phi_\theta(x) = \theta^d \phi_1(\theta x), \quad x \in \mathbb{R}^d,\ \theta > 1, \tag{1.8}$$

such that

- the *force field* corresponding to $(x,t) \to (1/N) \sum_{m=1}^{N} \nabla\phi_N(x - X_N^m(t))$ in (1.2) is determined by the values of the *particle density*, which is defined in analogy to $(x,t) \to (1/N) \sum_{m=1}^{N} \phi_N(x - X_N^m(t))$, in a discrete spatial mesh with mesh size $\delta \in (0,1)$. Furthermore,
- depending on $\tau > 0$ this mesh may be adapted, i.e. refined, in those regions of space, where the particle density behaves irregularly.

Quite similarly as in [12], [13], [14], [15], [16] we shall study regularized versions of the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ of the discretized modifications of (1.2), namely functions like $(x,s) \to (\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(s) * \phi_\theta)(x)$. As a main result of the present paper,

which is formulated precisely in section 3.2, we shall demonstrate the convergence in an L^2-sense of these functions to the solution ρ of (1.1) as $N, \theta \to \infty$ and $\delta, h(,\tau) \to 0$. In particular, an estimate for the rate of that convergence will be provided. The crucial parts of the proofs of our results are contained in section 4, whereas some auxiliary ingredients are deferred to section 5. In section 6 we shall present the outcomes of some numerical simulations.

Regarding the technical basis of the present work, we have indicated in the previous paragraph that the interaction kernels ϕ_θ, $\theta > 1$, and also certain related functions are employed to regularize the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$. As a consequence, several technical properties of these kernels have to be assumed. These suppositions are collected in section 3.1. We note that a few estimates used in sections 4 and 5 in some possibly modified form also appear in [12], [13], [14], [15], [16]. A certain amount of the technical difficulties encountered in the present work is associated with the fact that in the many-particle system (1.2) and also in its discretized modifications the range of the interaction is strictly positive, whereas in the dynamics (1.1) of the limit ρ this range vanishes. To handle that problem a family ρ_θ, $\theta > 1$, of functions solving particular integro-differential equations, cf. (3.16), will be utilized as an auxiliary tool. The dynamics of both these functions and the discretized modifications of (1.2) is governed by the interaction kernels ϕ_θ, $\theta > 1$. Therefore, it will become possible to study the *distance* between $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ and ρ_θ as $N, \theta \to \infty$ and $\delta, h(,\tau) \to 0$ in detail. On the other hand, in an accompanying paper [17] it is demonstrated that $\lim_{\theta\to\infty} \rho_\theta = \rho$, where explicit relations for the rate of this convergence are also given. As a final consequence, the desired estimate on the rate of convergence of $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ to ρ as $N, \theta \to \infty$ and $\delta, h(,\tau) \to 0$ will follow.

The study in the present paper is essentially a theoretical investigation of the so-called *smoothed particle hydrodynamics*- (SPH-) method, which has been described, e.g., in [1], [5]. The crucial feature of that particular particle method is the use of regularizations of the particle configurations, which are obtained by convolutions with smooth kernels approaching δ_0 as the particle number tends to infinity, to determine estimates for densities and also their derivatives. In particular, we employ the close connection to moderately interacting many-particle systems, which is established by this property.

As mentioned above, to determine the force field describing the interaction between the particles we shall utilize the values of the particle density on a discrete spatial mesh. For this reason our approach here may also be classified as a *particle-mesh method*; cf., e.g., [8]. As an important consequence, we do not need to consider explicitly the interaction between individual pairs of particles, and therefore it will become less expensive to simulate systems with large particle numbers.

Sometimes in applications of particle methods to the solution of partial differential equations the particles represent point concentrations of some *derivative* of the solution. This happens, for example, in connection with the *vortex method* to solve equations for incompressible fluid motion, cf., e.g., [10], and also in related studies of particular reaction-diffusion equations in one dimension; cf., e.g., [3], [18]. In contrast, as usual in SPH-models the particles in the present paper describe concentrations of the solution itself.

To conclude this first section we mention some notation, which will be used in the remaining parts of this paper.

By

$$\widetilde{f}(\lambda) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} dx \ \exp(i\lambda x) f(x), \quad \lambda \in \mathbb{R}^d, \ f \in L^1(\mathbb{R}^d),$$

we denote the *Fourier transform*. For its natural extensions to finite measures μ or square integrable functions g the notation $\widetilde{\mu}$ or $\widetilde{g}$ is also employed.

The set $\mathcal{M}(\mathbb{R}^d)$ contains the *finite measures* on $\mathbb{R}^d$. The *variation* of any $\mu \in \mathcal{M}(\mathbb{R}^d)$ is denoted by $|\mu|$.

To quantify regularity properties in some L^2-sense the *Sobolev norms*

$$\|f\|_{(m)} = \left(\sum_{k=0}^{m} \|\nabla^{\otimes k} f\|_2^2\right)^{1/2}, \tag{1.9}$$

$$\|f\|_{(m,1)} = \left(\sum_{k=0}^{m} \int_{\mathbb{R}^d} dy\ (1+|y|)|\nabla^{\otimes k} f(y)|^2\right)^{1/2}, \quad m = 0,1,2,\ldots,$$

for sufficiently regular real-valued functions f on $\mathbb{R}^d$ are utilized. Here, $\nabla^{\otimes k} f$ denotes the tensor of all partial derivatives of order k of a function f. The *Sobolev spaces* associated with the norms $\|.\|_{(m)}$, $m = 0,1,2,\ldots$, are

$$H_m^2(\mathbb{R}^d) = \big\{f \in L^2(\mathbb{R}^d) : \|f\|_{(m)} < \infty\big\}, \quad m = 0,1,2,\ldots.$$

As an abbreviation for integrals we shall apply the notations

$$\langle \mu, f\rangle = \int_{\mathbb{R}^d} \mu(dx) f(x), \qquad \langle f, g\rangle = \int_{\mathbb{R}^d} dx\ f(x)g(x),$$

whenever the right sides are well-defined for a measure μ and functions f and g.

The discrete sets

$$\mathcal{T}_h = \{0, h, 2h, \ldots\}, \quad h > 0,$$

represent collections of equidistant points on the time axis. For $s \geq 0$ we define

$$\lfloor s \rfloor_h = h\lfloor s/h \rfloor = \max\{ph : p \in \mathbb{Z},\ ph \leq s\};$$

i.e., $\lfloor s \rfloor_h$ is the largest element of $\mathcal{T}_h$, which is not larger than s. In addition to $\lfloor . \rfloor$ we will also use the notation

$$\lceil z \rceil = \min\{n \in \mathbb{Z} : n \geq z\}, \quad z \in \mathbb{R}.$$

For the cardinality of a finite set $\mathcal{N}$ the notation $|\mathcal{N}|$ will be employed.

Centered Gaussian densities in $\mathbb{R}^d$ are denoted as

$$\sigma_{d;\alpha}(x) = \frac{1}{(2\pi\alpha)^{d/2}} \exp\Big(-\frac{x^2}{2\alpha}\Big), \quad x \in \mathbb{R}^d,\ \alpha > 0.$$

To describe mathematically the randomness of the initial positions of the particles and the Brownian motions, which model the random part of their dynamics, we suppose the existence of some *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. Any random variable or random process appearing in this paper should be measurable with respect to $(\Omega, \mathcal{F}, \mathbb{P})$. The expectation operator in $(\Omega, \mathcal{F}, \mathbb{P})$ is denoted by $\mathbb{E}[.]$. Moreover, we assume that $(\mathcal{F}_s)_{s\geq 0}$ is a filtration of σ-algebras $\mathcal{F}_s \subseteq \mathcal{F}$ such that the initial positions of the particles are measurable with respect to $\mathcal{F}_0$ and any Brownian motion in this paper is *adapted* to $(\mathcal{F}_s)_{s\geq 0}$.

We denote by $C, C', C'', \ldots$ positive, finite constants, which may vary from place to place. In general, these constants are independent of the particle number N, the

discretization parameters δ and h, the scaling variable θ, the adaption parameter τ, or other variables with respect to those uniform estimates have to be deduced. If, however, the dependence on particular parameters $\alpha_1, \dots, \alpha_M$ is to be emphasized, the notation $C(\alpha_1, \dots, \alpha_M)$ is employed.

For the subsequent sections we choose some fixed $T \in (0, \infty)$ and then restrict our studies to the time interval $[0, T]$. As a simplification we also formulate our considerations only for $d = 2$. No essential new problems would be encountered for $d = 1$ or $d > 2$. Furthermore, our computer simulations described in section 6 are also performed for the two-dimensional case.

2. A discretized version of the many-particle system (1.2). In this section we describe how the many-particle system (1.2), which is continuous in space and time, may be discretized such that its simulation on a computer becomes possible. In other words, we construct a particle method, which can be used to solve (1.1) numerically.

First, we note that as a consequence of the symmetry of ϕ_1, cf. (1.5), the restriction "$m \neq k$" in the sum in (1.2) has no effect, and therefore to simplify the notation it will be omitted from now on.

To obtain a modification of the system (1.2), which is suitable for a numerical simulation,

(i) we discretize the time axis and utilize an *Euler scheme* based on this discretization to determine the evolution in time of the particle positions. Additionally,

(ii) we employ some *spatial mesh* and for any discrete time point we calculate at its vertices the *local population density*. Quite similarly as in moderately interacting many-particle systems this density is defined as the convolution of the empirical process of the particle positions with ϕ_θ; cf. (1.8). Depending on the spatial variations of the population density

(ii$_a$) the mesh size can be adapted locally. Furthermore,

(iii) we use the discrete set of values of the population density, which has been computed in (ii) (and (ii$_a$)), to determine by some interpolation procedure the *force field* at the particle positions. Together with some Gaussian random variables, which model the increments of the Brownian motions, this force field is employed to accomplish the evolution in time.

We note that the population density determined in (ii) (and (ii$_a$)) may also be used to obtain a graphical representation of the simulation results.

As a consequence of the rough outline (i)–(iii) our modification of the many-particle system (1.2) in the first case, when no adaption is performed, depends on four parameters, namely the particle number N, the mesh size δ of the nonadapted initial mesh, the time step h, and additionally on the scaling variable θ of the interaction potential ϕ_θ. Hence, we then typically need four indices N, δ, h, and θ to characterize the various mathematical objects appearing in the investigation of our particle method. Next, in the second case, when adaption is also included, the additional parameter τ has to be taken into account.

Subsequently in this section, we shall supplement (i)–(iii) by missing details. Especially, we have to describe the adaption procedure in (ii$_a$) and the interpolation step in (iii).

First, let us fix a particle number N, a mesh size δ, a time step h, a scaling parameter θ, and also for use in the second case an adaption parameter τ. We now will define a system of N particles, whose positions in $\mathbb{R}^2$ at time ph, $p = 0, 1, 2, \dots$, will be denoted by $Y^k_{N,\delta,h,\theta(,\tau)}(ph)$, $k = 1, \dots, N$. The relevant information about

these particle positions is summarized in the associated empirical process

$$\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(ph) = \frac{1}{N}\sum_{m=1}^{N} \delta_{Y^m_{N,\delta,h,\theta(,\tau)}(ph)}, \quad p = 0,1,2,\ldots; \tag{2.1}$$

i.e., $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(ph)$ is a finite measure having mass $1/N$ at the particle positions $Y^k_{N,\delta,h,\theta(,\tau)}(ph)$, $k = 1,\ldots,N$.

To characterize the time evolution of the particles, first their drift has to be described. Similarly as above in (1.2) in the original many-particle system this drift is determined by the interaction kernel ϕ_θ, which is defined in (1.8). A look at (1.2) suggests evaluating for any $p = 0,1,2,\ldots$ the sum $(1/N)\sum_{m=1}^N \nabla\phi_\theta\big(y - Y^m_{N,\delta,h,\theta(,\tau)}(ph)\big)$ at the particle positions $y \in \{Y^1_{N,\delta,h,\theta(,\tau)}(ph),\ldots,Y^N_{N,\delta,h,\theta(,\tau)}(ph)\}$. Evidently, for any time step the expense for computing these sums exactly would be proportional to N^2. This quadratic dependence on N would restrict severely the study of systems with large particle numbers. Hence, as indicated in (ii)–(iii) we shall provide an alternative expression for the drift on the particles. Ultimately, our analytical calculations summarized in Theorem 2 will imply that this expression is a good approximation as $N,\theta \to \infty$ and $\delta, h(,\tau) \to 0$.

Our starting point is the observation that for a smooth initial state ρ_0 the solution ρ of (1.1) also remains smooth for $t > 0$. Then, for any $t \geq 0$ the empirical measures $\mathbb{X}_N(t)$ of the system (1.2), which converge to $\rho(.,t)$ as $N \to \infty$, cf. [12], [13], should also be smooth in some generalized sense. Continuing this line of reasoning one may suspect that the empirical measures $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)$ associated with the simulation possess some smoothness properties too, at least, if they constitute the desired approximation to $\rho(.,t)$ at all. For this reason, to determine the drift acting on the particles and furthermore the particle density needed for a graphical representation of the simulation results, it should suffice to evaluate the function $y \to (1/N)\sum_{m=1}^N \phi_\theta\big(y - Y^m_{N,\delta,h,\theta(,\tau)}(t)\big)$ in some sufficiently dense discrete subset of $\mathbb{R}^2$.

More precisely, to implement (ii)–(iii) we proceed as follows:

(A) We choose some region $Q \subseteq \mathbb{R}^2$. This region should contain the positions of *most* of the particles during the time interval $[0,T]$ of our considerations. For simplicity we select a rectangle Q. Both Q and also the rectangles $R^1, R^2, \ldots$, which are constructed below in (B) and (C), are half open; i.e., they have the form $[a,b) \times [c,d)$ for suitable $a,b,c,d \in \mathbb{R}$. However, occasionally when considering a single R^m it will be convenient to assume that rectangle to be closed. In particular, the corners A^m, B^m, C^m and D^m, cf. Figure 1, are then contained in R^m. We denote the width of Q in the x_1- (x_2-) direction by w_1 (w_2). The introduction of the bounded set Q is necessary, since the lattices constructed below in (B) and (C) have to be finite.

(B) Within Q a regular, rectangular lattice $\mathcal{L}_0$, which consists of a set of rectangles $R^1,\ldots,R^{|\mathcal{L}_0|}$, whose sides have length $w_1/\lceil w_1/\delta\rceil$ ($w_2/\lceil w_2/\delta\rceil$) in the x_1- (x_2-) direction, is constructed. At any time point ph, when the drift of the particles has to be computed, the *density* $y \to d_{N,\delta,h,\theta(,\tau)}(y,ph) = (1/N)\sum_{m=1}^N \phi_\theta\big(y - Y^m_{N,\delta,h,\theta(,\tau)}(ph)\big)$ is determined both at the vertices of the lattice and the centers of the rectangles R^1, $\ldots, R^{|\mathcal{L}_0|}$. As indicated in Figure 1 the corners and the center of some rectangle R^m are denoted by A^m, B^m, C^m, D^m and M^m, respectively. We suppose that

$$\delta < \delta_0 = \min\{w_1, w_2\}/2, \tag{2.2}$$

in order that $\mathcal{L}_0$ contains a 3×3 submesh.

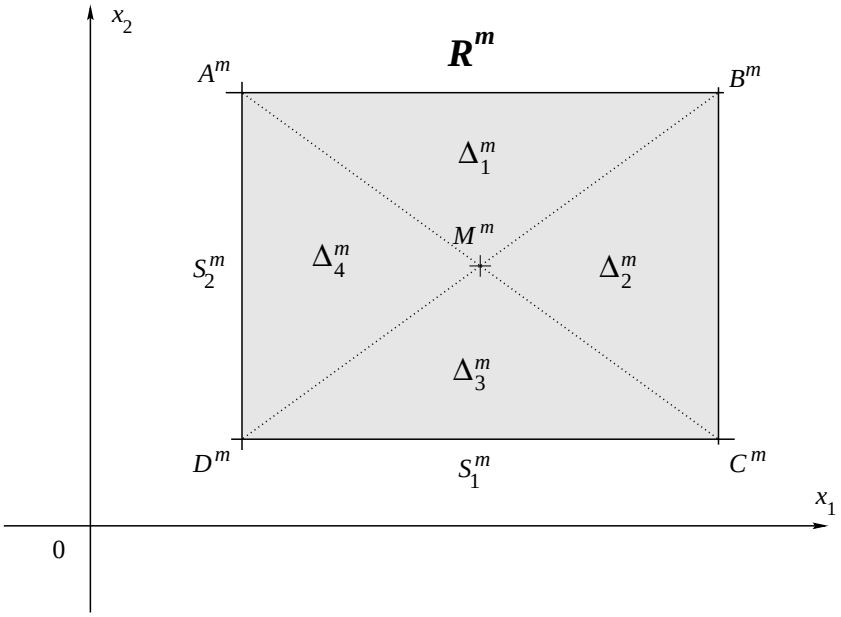


FIG. 1. *A typical rectangle within the mesh in Q.*

(C) For some time ph it may happen that the discretization of the function $y \to d_{N,\delta,h,\theta(,\tau)}(y, ph)$, which is associated with the lattice $\mathcal{L}_0$ as described in (B), is too coarse. In such a situation the above-mentioned local adaption of the mesh size, cf. (ii_a), may be performed. More precisely, the adaption procedure should be started, if in some rectangles the density $d_{N,\delta,h,\theta,\tau}(., ph)$ is not *flat* enough, where *flatness* is measured in terms of $|\nabla^{\otimes 2} d_{N,\delta,h,\theta,\tau}(., ph)|$.

In our situation we have given only the values of the function $d_{N,\delta,h,\theta,\tau}(., ph)$ in some discrete set of points in Q. These values can be utilized to compute for any R^m a quantity $\kappa^m_{N,\delta,h,\theta,\tau}(ph)$ related to $\text{diam}(R^m) \sup_{x \in R^m} |\nabla^{\otimes 2} d_{N,\delta,h,\theta,\tau}(x, ph)|$. For clarity in the present paragraph we defer the detailed description of the determination of $\kappa^m_{N,\delta,h,\theta,\tau}(ph)$ to (E). Employing now the adaption parameter $\tau > 0$ we consider $d_{N,\delta,h,\theta,\tau}(., ph)$ as not flat enough in R^m if

$$|\kappa^m_{N,\delta,h,\theta,\tau}(ph)| > \tau. \tag{2.3}$$

In this way in any rectangle the flatness of $d_{N,\delta,h,\theta,\tau}(., ph)$ is examined, and finally any R^m, where (2.3) holds, is divided into four smaller rectangles having equal scales and the point M^m as a common corner. Next, the density $d_{N,\delta,h,\theta,\tau}(., ph)$ is computed at the centers and those corners of the new rectangles, where it has not yet been calculated during previous calculations referring to time ph. If necessary, i.e., if as a consequence of divisions new rectangles have been created, the discretization of Q is checked again. More precisely, using (2.3) the flatness of $d_{N,\delta,h,\theta,\tau}(., ph)$ within the various rectangles is examined once more and perhaps further rectangles are divided. Continuing with this procedure we finally arrive at some possibly irregular lattice $\mathcal{L}(ph)$, which consists of rectangles $R^1, \ldots, R^{|\mathcal{L}(ph)|}$ and locally may have an appearance as in Figure 2. $\mathcal{L}(ph)$ is *adapted* to the density $d_{N,\delta,h,\theta,\tau}(., ph)$; i.e., the rectangles R^m, $m = 1, \ldots, |\mathcal{L}(ph)|$, constructed by the procedure described above are small enough such that within any R^m the function $y \to d_{N,\delta,h,\theta,\tau}(y, ph)$ is sufficiently

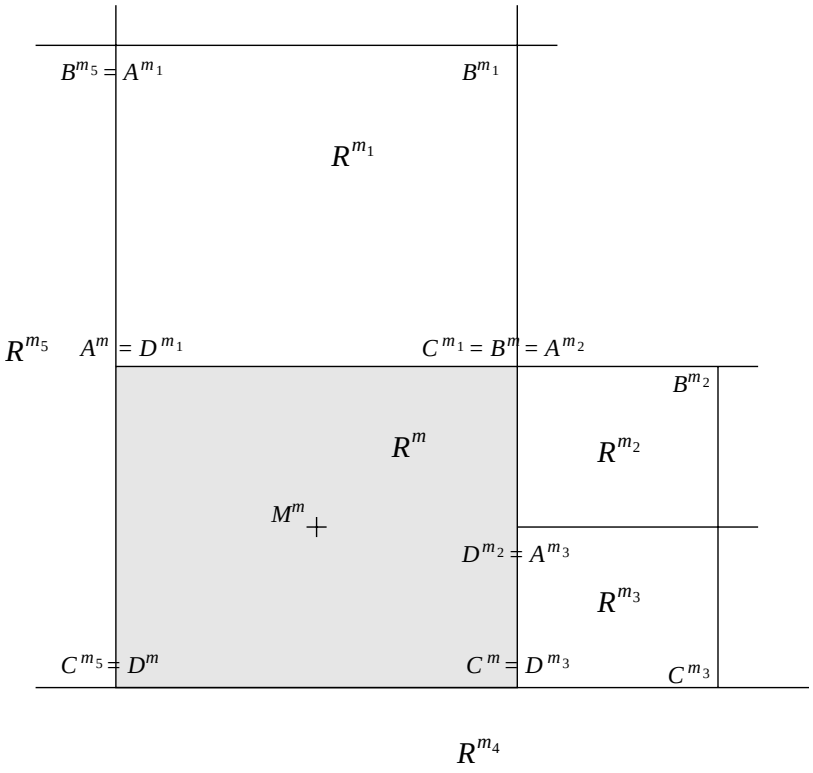


FIG. 2. *Details of a locally refined mesh.*

flat according to the criterion specified by (2.3).

Let

$$S_1^m = B_1^m - A_1^m = C_1^m - D_1^m,$$
$$S_2^m = A_2^m - D_2^m = B_2^m - C_2^m, \quad m = 1, \dots, |\mathcal{L}(ph)|, \tag{2.4}$$

with A_1^m (A_2^m), ... denoting the x_1- (x_2-) coordinate of A^m, ..., be the width of R^m in the x_1- (x_2-) direction. Obviously, the ratios S_1^m/S_2^m, $m = 1, 2, \dots$, coincide for $\mathcal{L}_0$ and equal $w_1\lceil w_2/\delta\rceil/(w_2\lceil w_1/\delta\rceil)$. Furthermore, they are preserved by the subsequent divisions of rectangles described above, and therefore we observe

$$\underline{C}(Q) \le \frac{S_1^m}{S_2^m} \le \overline{C}(Q),\ 0 \le S_1^m,\ S_2^m \le \delta, \quad m = 1, \dots, |\mathcal{L}(ph)|, \tag{2.5}$$

where $\underline{C}(Q), \overline{C}(Q) \in (0, \infty)$ are constants, which depend only on the rectangle Q, i.e., on w_1 and w_2.

(D) For the graphical representation of the *particle density* of the many-particle system defining the particle method the density values at the corners and the centers of the rectangles of the lattice $\mathcal{L}(ph)$, which are calculated by performing (B) and (C), may be used immediately. However, as indicated above in (iii) for the determination of the drift acting on the particles an *interpolation procedure* has to be utilized. To explain the details we fix one particle at position $y_P = Y^k_{N,\delta,h,\theta(,\tau)}(ph)$ at time ph.

If $y_P \notin Q$, where Q is the region introduced in (A), i.e., if y_P is not covered by some rectangle in the lattice $\mathcal{L}(ph)$, then in accordance with (1.2) we define

$$\widehat{D}_{N,\delta,h,\theta(,\tau)}(y_P, ph) = \nabla d_{N,\delta,h,\theta(,\tau)}(y_P, ph) = \frac{1}{N}\sum_{m=1}^{N} \nabla\phi_\theta\big(y_P - Y^m_{N,\delta,h,\theta(,\tau)}(ph)\big). \tag{2.6}$$

If $y_P \in Q$, we first determine the unique rectangle R^m such that $y_P \in R^m$. As shown in Figure 1, R^m is divided into four triangles Δ_1^m, Δ_2^m, Δ_3^m, and Δ_4^m, where these triangles have the corners A^m, B^m, M^m (for Δ_1^m), B^m, C^m, M^m (for Δ_2^m), C^m, D^m, M^m (for Δ_3^m), and D^m, A^m, M^m (for Δ_4^m). We suppose that the points on the diagonals of R^m belong to the respective adjacent triangle on the left and that $M^m \in \Delta_4^m$. For any triangle Δ_r^m the density values at its corners, which are calculated in (B) and (C), are employed to determine the drift acting on particles contained in its interior. For example, if $y_P \in \Delta_1^m$, we may define the vector

$$\widehat{D}_{N,\delta,h,\theta(,\tau)}(y_P, ph) = \begin{pmatrix} \dfrac{d_{N,\delta,h,\theta(,\tau)}(B^m, ph) - d_{N,\delta,h,\theta(,\tau)}(A^m, ph)}{S_1^m} \\ \dfrac{d_{N,\delta,h,\theta(,\tau)}(A^m, ph) + d_{N,\delta,h,\theta(,\tau)}(B^m, ph) - 2d_{N,\delta,h,\theta(,\tau)}(M^m, ph)}{S_2^m} \end{pmatrix}, \tag{2.7}$$

where S_1^m and S_2^m are introduced in (2.4). Obviously, (2.7) characterizes a particular discretization of $\nabla d_{N,\delta,h,\theta(,\tau)}(y_P, ph)$. If $y_P \in \Delta_r^m$, $r = 2, 3, 4$, immediate modifications of (2.7) may be used. For details we refer the reader to (5.1).

For convenience, we do not utilize $\widehat{D}_{N,\delta,h,\theta(,\tau)}$ as defined in (2.6) and (2.7) directly for the determination of the drift acting on the particles but employ a bounded modification. More precisely, we define

$$D_{N,\delta,h,\theta(,\tau)}(y, ph) = \begin{cases} \dfrac{\widehat{D}_{N,\delta,h,\theta(,\tau)}(y, ph)}{|\widehat{D}_{N,\delta,h,\theta(,\tau)}(y, ph)|} \min\{K, |\widehat{D}_{N,\delta,h,\theta(,\tau)}(y, ph)|\}, & \text{if } |\widehat{D}_{N,\delta,h,\theta(,\tau)}(y, ph)| \neq 0, \\ 0 & \text{if } |\widehat{D}_{N,\delta,h,\theta(,\tau)}(y, ph)| = 0, \end{cases} \quad y \in \mathbb{R}^2,\ p = 0, 1, 2, \ldots, \tag{2.8}$$

where the constant $K > 0$ is a suitable upper bound to the drift acting on the particles asymptotically as $N, \theta \to \infty$ and $\delta, h(, \tau) \to 0$; cf. (3.15).

(E) We still have to describe the determination of the quantity $\kappa_{\cdot N,\delta,h,\theta,\tau}(.)$ to complete the discussion in (C) of the adaption of the lattice $\mathcal{L}(.)$ to the local fluctuations of the density $d_{N,\delta,h,\theta,\tau}$. For that aim we suppose that for some fixed time ph after the division of several rectangles in an intermediate step the lattice $\mathcal{L}'(ph)$ has been obtained. Then a fixed rectangle R^m and its environment may look similarly as in Figure 2.

Within R^m the function $d_{N,\delta,h,\theta,\tau}(., ph)$ is given at least in A^m, B^m, C^m, D^m, and M^m, and therefore in any case four independent differences like $d_{N,\delta,h,\theta,\tau}(B^m, ph) - d_{N,\delta,h,\theta,\tau}(A^m, ph)$ or $d_{N,\delta,h,\theta,\tau}(M^m, ph) - d_{N,\delta,h,\theta,\tau}(D^m, ph)$ are available to estimate the five components of $\nabla d_{N,\delta,h,\theta,\tau}(x, ph)$ and $\nabla^{\otimes 2} d_{N,\delta,h,\theta,\tau}(x, ph)$ for $x \in R^m$. Consequently, if not by previous steps in the construction of $\mathcal{L}'(ph)$ sides of R^m have been

divided and thus the values of $d_{N,\delta,h,\theta,\tau}(.,ph)$ at some intermediate points are known, we possibly will also have to take into account some values of $d_{N,\delta,h,\theta,\tau}(.,ph)$ in points of neighboring rectangles to estimate those partial derivatives.

As far as $\nabla d_{N,\delta,h,\theta,\tau}(x,ph)$, $x \in R^m$, is concerned, our approximations are described in (D); cf. (2.7) and also (5.1). Obviously, for those approximations only the values of $d_{N,\delta,h,\theta,\tau}(.,ph)$ in $\{A^m, B^m, C^m, D^m, M^m\}$ are used. They are also sufficient to estimate some particular linear combinations of partial derivatives of second order. More precisely, we obtain

$$
\begin{aligned}
(2.9)\ & (S_1^m)^2\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(x,ph) + (S_2^m)^2\nabla_2\nabla_2 d_{N,\delta,h,\theta,\tau}(x,ph)\\
&\quad \simeq 2\big(d_{N,\delta,h,\theta,\tau}(A^m,ph) + d_{N,\delta,h,\theta,\tau}(B^m,ph)\\
&\qquad + d_{N,\delta,h,\theta,\tau}(C^m,ph) + d_{N,\delta,h,\theta,\tau}(D^m,ph) - 4d_{N,\delta,h,\theta,\tau}(M^m,ph)\big),\\
&S_1^m S_2^m \nabla_1\nabla_2 d_{N,\delta,h,\theta,\tau}(x,ph)\\
&\quad \simeq d_{N,\delta,h,\theta,\tau}(B^m,ph) + d_{N,\delta,h,\theta,\tau}(D^m,ph)\\
&\qquad - d_{N,\delta,h,\theta,\tau}(A^m,ph) - d_{N,\delta,h,\theta,\tau}(C^m,ph), \quad x \in R^m.
\end{aligned}
$$

To justify (2.9) the right sides may be expanded in a Taylor series of second order. Some hints can be found in the proof of Lemma 3.

Contrary to (2.9), in order to approximate the derivatives $\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(.,ph)$ and $\nabla_2\nabla_2 d_{N,\delta,h,\theta,\tau}(.,ph)$ separately $d_{N,\delta,h,\theta,\tau}(.,ph)$-values of some points belonging to neighboring rectangles are possibly needed. Let us consider first the upper boundary of R^m, i.e., the interval $[A^m, B^m]$ between A^m and B^m. It may happen that by previous divisions in $\mathcal{L}'(ph)$ the rectangle above R^m has been divided. Then as a particular consequence, $d_{N,\delta,h,\theta,\tau}((A^m + B^m)/2, ph)$ has been computed. In this case we define

$$
\begin{aligned}
(2.10)\quad K_{1,u}^m = 4\big(&d_{N,\delta,h,\theta,\tau}(A^m,ph)\\
&+ d_{N,\delta,h,\theta,\tau}(B^m,ph) - 2d_{N,\delta,h,\theta,\tau}((A^m + B^m)/2,ph)\big)/(S_1^m)^2
\end{aligned}
$$

as an approximation to $\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(x,ph)$, $x \in [A^m, B^m]$. Otherwise, i.e., if $d_{N,\delta,h,\theta,\tau}((A^m + B^m)/2, ph)$ is not known, we examine the structure of $\mathcal{L}'(ph)$ in the linear continuations of the interval $[A^m, B^m]$ beyond A^m and B^m, respectively. Since the lattice $\mathcal{L}'(ph)$, which represents the present state of refinement of the original lattice $\mathcal{L}_0$, is obtained by several divisions of rectangles as described in (C), either one or both continuations of $[A^m, B^m]$ belong to the respective boundary of another neighboring rectangle in $\mathcal{L}'(ph)$. With $\mathcal{E}_u^m \subseteq \{l, r\}$ we introduce an enumeration of those continuations. We suppose that $l \in \mathcal{E}_u^m$ ($r \in \mathcal{E}_u^m$) if and only if $[A^m, B^m]$ can be extended within $\mathcal{L}'(ph)$ on its left (right) side beyond A^m (B^m). For example, in Figure 2 we have $\mathcal{E}_u^m = \{r\}$ and $\mathcal{E}_u^{m_2} = \{r, l\}$. If $l \in \mathcal{E}_u^m$ ($r \in \mathcal{E}_u^m$), we define $E_{u,l}^m$ ($E_{u,r}^m$) as that vertex in $\mathcal{L}'(ph)$ on the extension of $[A^m, B^m]$ beyond A^m (B^m), whose distance to A^m (B^m) is minimal. In particular, in Figure 2 we get $E_{u,r}^m = B^{m_2}$ and $E_{u,l}^{m_2} = A^m$. We note that $d_{N,\delta,h,\theta,\tau}(E_{u,q}^m, ph)$, $q \in \mathcal{E}_u^m$, has been determined during the previous steps in the construction of $\mathcal{L}'(ph)$. Therefore, we may now introduce

$$K^m_{1,u,l} = \frac{2}{(A^m_1 - E^m_{u,l,1})(B^m_1 - A^m_1)}\Bigg(\frac{B^m_1 - A^m_1}{B^m_1 - E^m_{u,l,1}} d_{N,\delta,h,\theta,\tau}(E^m_{u,l}, ph) + \frac{A^m_1 - E^m_{u,l,1}}{B^m_1 - E^m_{u,l,1}} d_{N,\delta,h,\theta,\tau}(B^m, ph) - d_{N,\delta,h,\theta,\tau}(A^m, ph)\Bigg), \quad \text{if } l \in \mathcal{E}^m_u, \tag{2.11}$$

and

$$K^m_{1,u,r} = \frac{2}{(B^m_1 - A^m_1)(E^m_{u,r,1} - B^m_1)}\Bigg(\frac{E^m_{u,r,1} - B^m_1}{E^m_{u,r,1} - A^m_1} d_{N,\delta,h,\theta,\tau}(A^m, ph) + \frac{B^m_1 - A^m_1}{E^m_{u,r,1} - A^m_1} d_{N,\delta,h,\theta,\tau}(E^m_{u,r}, ph) - d_{N,\delta,h,\theta,\tau}(B^m, ph)\Bigg), \quad \text{if } r \in \mathcal{E}^m_u, \tag{2.12}$$

where $E^m_{u,q,k}$, $k = 1, 2$, denotes the x_k-coordinate of $E^m_{u,q}$, $q \in \mathcal{E}^m_u$. By employing Taylor expansions of second order for $d_{N,\delta,h,\theta,\tau}(., ph)$ at A^m and B^m it may be shown that $K^m_{1,u,l} \simeq \nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(A^m, ph)$ and $K^m_{1,u,r} \simeq \nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(B^m, ph)$, respectively. For details we refer to Lemma 3 and its proof. Consequently, in the case where $d_{N,\delta,h,\theta,\tau}((A^m + B^m)/2, ph)$ is not known we may use $K^m_{1,u,l}$ and $K^m_{1,u,r}$ to introduce an alternative to (2.10), namely

$$K^m_{1,u} = \frac{1}{|\mathcal{E}^m_u|} \sum_{q \in \mathcal{E}^m_u} K^m_{1,u,q}, \tag{2.13}$$

as an approximation to $\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in [A^m, B^m]$.

For the lower boundary of R^m, i.e., the interval $[C^m, D^m]$ between C^m and D^m, we can define a supplement $K^m_{1,l}$ to $K^m_{1,u}$. First, if $d_{N,\delta,h,\theta,\tau}((C^m + D^m)/2, ph)$ has been determined during previous calculations, an obvious modification of (2.10) may be used. Otherwise, the considerations leading to (2.13) have to be transferred from $[A^m, B^m]$ to $[C^m, D^m]$. More precisely, in a study of the continuations of $[C^m, D^m]$ beyond C^m and D^m we have to introduce the set $\mathcal{E}^m_l$, and then, similarly as in (2.11) and (2.12), we have to define $K^m_{1,l,q}$, $q \in \mathcal{E}^m_l$, such that finally an analogue of (2.13) for an approximation $K^m_{1,l}$ for $\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in [C^m, D^m]$, is obtained.

Now, with

$$K^m_1 = \frac{1}{2}(K^m_{1,u} + K^m_{1,l}) \tag{2.14}$$

an approximation to $\nabla_1\nabla_1 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in R^m$, can be defined.

Next, the considerations leading to (2.14) may be "rotated by $\pi/2$" to deduce quite analogously an approximation K^m_2 to $\nabla_2\nabla_2 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in R^m$.

For the definition of $\kappa^m_{N,\delta,h,\theta,\tau}(ph)$, which is employed in (C) to quantify *flatness* of $d_{N,\delta,h,\theta,\tau}(., ph)$ in R^m, we do not utilize approximations like K^m_1 or K^m_2 for spatial derivatives of second order of $d_{N,\delta,h,\theta,\tau}(., ph)$ in R^m directly but instead use a suitable upper bound for such approximations. Taking into account (2.10)–(2.14) and also their modifications, which refer to the intervals $[B^m, C^m]$, $[C^m, D^m]$, $[D^m, A^m]$ and have not been written down explicitly, we therefore introduce further definitions.

First, by $\mathcal{X}_1^m$ we denote the set of those intervals $\eta \in \mathcal{I} = \{[A^m, B^m], [B^m, C^m], [C^m, D^m], [D^m, A^m]\}$ such that during previous divisions of some rectangles in the neighborhood of R^m we have determined $d_{N,\delta,h,\theta,\tau}(M_\eta, ph)$ with M_η denoting the center of η. Let P_η and E_η be the endpoints of η. Next, by $\mathcal{X}_2^m$ we denote the set of all linear continuations of intervals in $\mathcal{I} \setminus \mathcal{X}_1^m$ beyond their respective endpoints, which belong to the boundary of a neighboring rectangle. Associated with any $\eta \in \mathcal{X}_2^m$ is a set $\Theta_\eta = \{P_\eta, M_\eta, E_\eta\}$ of three points such that P_η and M_η are neighboring corners of R^m, η is an extension of $[P_\eta, M_\eta]$ beyond M_η, and E_η is lying outside R^m but is contained both in η and $\mathcal{L}'(ph)$ with minimal distance to R^m. In the end we obtain for any R^m a set $\mathcal{X}^m = \mathcal{X}_1^m \cup \mathcal{X}_2^m$ with a set $\Theta_\eta = \{P_\eta, M_\eta, E_\eta\}$ of vertices in $\mathcal{L}'(ph)$ for any $\eta \in \mathcal{X}^m$.

Following (2.10)–(2.12) we then define

$$K_\eta = \frac{1}{|P_\eta - M_\eta||M_\eta - E_\eta|} \left| \frac{|M_\eta - E_\eta|}{|P_\eta - E_\eta|} d_{N,\delta,h,\theta,\tau}(P_\eta, ph) + \frac{|P_\eta - M_\eta|}{|P_\eta - E_\eta|} d_{N,\delta,h,\theta,\tau}(E_\eta, ph) - d_{N,\delta,h,\theta,\tau}(M_\eta, ph) \right|, \quad \eta \in \mathcal{X}^m. \tag{2.15}$$

These quantities K_η, $\eta \in \mathcal{X}^m$, may be combined with an analogous estimate for the approximation to $\nabla_1 \nabla_2 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in R^m$, given by (2.9). Consequently, we are directed to the definition

$$\kappa^m_{N,\delta,h,\theta,\tau}(ph) = \left(\frac{1}{S_1^m S_2^m} \big| d_{N,\delta,h,\theta,\tau}(A^m, ph) + d_{N,\delta,h,\theta,\tau}(C^m, ph) - d_{N,\delta,h,\theta,\tau}(B^m, ph) - d_{N,\delta,h,\theta,\tau}(D^m, ph) \big| + \sum_{\eta \in \mathcal{X}^m} K_\eta \right) \sqrt{S_1^m S_2^m}, \quad m = 1, \ldots, |\mathcal{L}'(ph)|. \tag{2.16}$$

With $\kappa^m_{N,\delta,h,\theta,\tau}(ph)$ we have provided the desired quantity to estimate flatness of $d_{N,\delta,h,\theta,\tau}(., ph)$ within R^m. Its use in a procedure to obtain a mesh $\mathcal{L}(ph)$, which is adapted to the local density fluctuations of $d_{N,\delta,h,\theta,\tau}(., ph)$ in Q, is explained in (C).

Remark 1. (i) The definition (2.16) of $\kappa^m_{N,\delta,h,\theta,\tau}(.)$ is symmetric; i.e., any corner of R^m and any direction is considered in the same way. However, to obtain the rate of convergence given by the right side of (3.24) in Theorem 2(b) with a possibly different $C(T)$ it would suffice to include in (2.16) instead of $\sum_{\eta \in \mathcal{X}^m} K_\eta$ only two terms accounting for the x_1- and x_2-direction, respectively. In particular, these terms would provide suitable upper bounds to approximations of $\nabla_1 \nabla_1 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in R^m$, and $\nabla_2 \nabla_2 d_{N,\delta,h,\theta,\tau}(x, ph)$, $x \in R^m$, which are needed in connection with the second part of Lemma 3.

(ii) Obviously, the choice of the region Q, which is that part of $\mathbb{R}^2$ where our simulation procedure may be characterized as a *particle-mesh method*, has a considerable influence on the expenses for the computations. In particular, if Q is large, it is left by only a few particles, and therefore for the determination of their drift only a small number of extensive summations as in (2.6) has to be performed. On the other hand, for any time ph the number of mesh points, where $d_{N,\delta,h,\theta(,\tau)}(., ph)$ has to be determined, being $\sim |\mathcal{L}(ph)|$ gets large with Q. For the determination of a convenient region Q we have not performed rigorous considerations but rely on some heuristic arguments. For details we refer to Remark 3(vi) in section 3.

(iii) The restriction (2.2) on the spatial discretization parameter δ ensures that $\mathcal{L}_0$ contains at least three mesh points in each direction. This is a necessary require-

ment in order that the adaption step described by (C) and (E) can be performed, in particular since only in this case sets like $\mathcal{E}_u^m$, which are introduced in the paragraph after (2.10), are always nonempty.

Having specified at some fixed time ph the negative drift acting on the respective particles according to (2.6)–(2.8) we may finally determine the particle positions at time $ph+h$ by following an Euler scheme; cf. (i). In particular, we may define

$$Y^k_{N,\delta,h,\theta(,\tau)}(ph+h) = Y^k_{N,\delta,h,\theta(,\tau)}(ph) - hD_{N,\delta,h,\theta(,\tau)}\big(Y^k_{N,\delta,h,\theta(,\tau)}(ph), ph\big) + \sqrt{h}Z^{k,p}, \quad k=1,\dots,N, \tag{2.17}$$

where $Z^{k,p}$, $k=1,\dots,N$, $p=0,1,2,\dots$, are independent $\mathbb{R}^2$-valued *Gaussian* random variables with mean 0 and variance 1. To justify (2.17) we remark that the random vector $(\sqrt{h}Z^{1,p},\dots,\sqrt{h}Z^{N,p})$ has the same distribution as $\big(W^1(ph+h) - W^1(ph),\dots,W^N(ph+h) - W^N(ph)\big)$, which describes the increments in the time interval $[ph, ph+h)$ of the Brownian motions $W^1,\dots,W^N$ appearing in (1.2).

The crucial ideas to discretize (1.2) are formulated in (i)–(iii), where the details of (ii)–(iii) can be found in (A)–(E). In short form the resulting procedure may be summarized as follows:

(0) First, as a domain, which may be regarded as the essential living space of the population of particles under consideration, some rectangle Q is chosen. Then, as discretization parameters the number N of particles, the mesh size $\delta \in (0,\delta_0)$ and the time step $h>0$ are selected. Moreover, $\theta>1$ is used as scaling parameter for the interaction, and some $\tau>0$ is introduced, if the adaption of the mesh should be performed.

(1) We determine suitable initial positions $Y^k_{N,\delta,h,\theta(,\tau)}(0)$, $k=1,\dots,N$, such that the initial population density $\rho(.,0)=\rho_0$ of the limit system (1.1) is *approximated* by the empirical measure $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0) = (1/N)\sum_{k=1}^N \delta_{Y^k_{N,\delta,h,\theta(,\tau)}(0)}$. Some hints on the choice of these initial positions can be found in Remark 3(iii) in section 3.

(2) For any time point ph, $p=0,1,2,\dots$, we first determine the population density at the points of some sufficiently fine mesh $\mathcal{L}(ph)$ in Q, cf. (B), (C), in the course of which the adaption procedure (C), (E) may also be performed. Next, following (D), cf. (2.6)–(2.8), we compute the negative drift acting on the respective particles. Finally, the new particle positions at time $ph+h$ are fixed according to (2.17).

(3) For sufficiently many time points the values of the population density at the corners and centers of the rectangles of the respective lattice $\mathcal{L}(ph)$ are saved in some file. These values can be used for a graphical representation of the simulation results.

Of course, for the computer simulations the particle positions $Y^k_{N,\delta,h,\theta(,\tau)}(t)$, $k=1,\dots,N$, need to be defined only for time points t in the discrete set $\mathcal{T}_h$. However, for our subsequent considerations, where the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ are compared analytically with processes in continuous time, it will be convenient to define $Y^k_{N,\delta,h,\theta(,\tau)}(t)$ for any $t\ge 0$. For that purpose, we extend the system (2.17) of *stochastic difference equations* to some system of *stochastic differential equations*, namely

$$Y^k_{N,\delta,h,\theta(,\tau)}(t) = Y^k_{N,\delta,h,\theta(,\tau)}(0) - \int_0^t ds\ D_{N,\delta,h,\theta(,\tau)}\big(Y^k_{N,\delta,h,\theta(,\tau)}(\lfloor s\rfloor_h), \lfloor s\rfloor_h\big) + W^k(t), \tag{2.18}$$

$$k = 1, \dots, N,\ t \geq 0,\ \delta \in (0, \delta_0),\ h > 0,\ \theta > 1,\ (\tau > 0,)\ N \in \mathbb{N}.$$

As we have indicated by our notation, the respective solutions $(Y^1_{N,\delta,h,\theta(,\tau)}(t), \dots, Y^N_{N,\delta,h,\theta(,\tau)}(t))$ of (2.17) and (2.18) coincide in distribution for $t \in \mathcal{T}_h$. This is fairly obvious, since in (2.18) the drift $-D_{N,\delta,h,\theta(,\tau)}$ acting on the particles remains constant between two successive time points $ph, ph + h \in \mathcal{T}_h$ and equals the "drift" at time ph in the system (2.17), and since the increments of the Brownian motions $W^k(ph + h) - W^k(ph)$ in (2.18) have the same law as the random variables $\sqrt{h}Z^{k,p}$ appearing in (2.17).

Now, with (2.18) definition (2.1) of the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)$ can be extended from $t \in \mathcal{T}_h$ to any $t \geq 0$. In our analytical investigation of the procedure sketched by (0)–(3) we shall study the convergence of those empirical processes in continuous time to the solution ρ of (1.1) in the limit as $N, \theta \to \infty$ and $\delta, h(, \tau) \to 0$. As a consequence of several similarities between (1.2) and (2.18) we can rely in these considerations on some ideas from [13]. Of course, the fact that in (2.18) the drift is determined by discretizing the gradient $(y, t) \to (1/N) \sum_{m=1}^N \nabla\phi_\theta(y - Y^m_{N,\delta,h,\theta(,\tau)}(t))$ of the population density both in space and time will lead to new problems, which did not appear in [13].

3. Results. In this section we first collect some assumptions, in particular on the kernel ϕ_1 and the initial state ρ_0 of the solution ρ of (1.1). Next, we shall describe the family ρ_θ, $\theta > 1$, of auxiliary functions, which had been mentioned in section 1. Finally, our main results and some supplementary remarks are presented.

3.1. Technical preliminaries. According to (1.8) the function ϕ_1 can be considered as the basis of the family of interaction kernels ϕ_θ, $\theta > 1$. Moreover, as indicated in section 1 these interaction kernels are used to regularize the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$, $N \in \mathbb{N}$, $\delta \in (0, \delta_0)$, $h > 0$, $\theta > 1(, \tau > 0)$. In view of those applications (1.5) has to be supplemented by various other assumptions on ϕ_1.

First, ϕ_1 is supposed to be a convolution product

$$\phi_1 = \phi_1^r * \phi_1^r, \tag{3.1}$$

where ϕ_1^r is a smooth, symmetric probability density, i.e.,

$$\phi_1^r \in C_b^1(\mathbb{R}^2), \quad \phi_1^r \geq 0, \quad \int_{\mathbb{R}^2} dx\ \phi_1^r(x) = 1, \quad \phi_1^r(x) = \phi_1^r(-x),\ x \in \mathbb{R}^2. \tag{3.2}$$

As far as smoothness properties of ϕ_1^r are concerned, we also need

$$\widetilde{\phi_1^r} \in C_b^2(\mathbb{R}^2), \qquad \int_{\mathbb{R}^2} dx\ |\nabla\phi_1^r(x)|^2 < \infty. \tag{3.3}$$

Functions related to ϕ_1^r are defined by

$$U_{k_1,k_2}(x) = (-1)^{k_1+k_2} \frac{x_1^{k_1} x_2^{k_2}}{k_1! k_2!} \phi_1^r(x), \qquad k_1, k_2 = 0, 1, \dots, \overline{k},\ k_1 + k_2 \leq \overline{k},\ x \in \mathbb{R}^2, \tag{3.4}$$

where

$$\overline{k} = 2. \tag{3.5}$$

They are supposed to satisfy

$$\left|\widetilde{U_{k_1,k_2}}(\lambda)\right| \le C\left|\widetilde{\phi_1^r}(\lambda)\right|, \quad k_1,k_2=0,1,\dots,\overline{k}-1,\ 1\le k_1+k_2\le\overline{k}-1,\ \lambda\in\mathbb{R}^2, \tag{3.6}$$

and

$$|U_{k_1,k_2}(x)| \le C\left(\frac{1}{1+|x|^3}\right)^{1/2}, \quad k_1,k_2=0,1,\dots,\overline{k},\ k_1+k_2=\overline{k},\ x\in\mathbb{R}^2. \tag{3.7}$$

Furthermore, we assume that ϕ_1 has bounded moments of all orders, i.e.,

$$\int_{\mathbb{R}^2} dx\ |x|^m\phi_1(x)<\infty, \quad m=1,2,\dots. \tag{3.8}$$

To regularize the empirical processes we shall use functions which are obtained from ϕ_1^r by the scaling (1.8), namely

$$\phi_\theta^r(x)=\theta^2\phi_1^r(\theta x), \quad x\in\mathbb{R}^2,\ \theta>1. \tag{3.9}$$

Obviously, modifications of (3.1) and (3.2) also hold for these functions, i.e.,

$$\phi_\theta=\phi_\theta^r*\phi_\theta^r, \tag{3.10}$$
$$\phi_\theta^r\in C_b^1(\mathbb{R}^2),\ \phi_\theta^r\ge 0,\ \int_{\mathbb{R}^2} dx\ \phi_\theta^r(x)=1,\ \phi_\theta^r(x)=\phi_\theta^r(-x),\ x\in\mathbb{R}^2,\quad \theta>1.$$

Moreover, by (1.5), (1.8), (3.2), and (3.9) the kernels ϕ_θ and ϕ_θ^r satisfy

$$\lim_{\theta\to\infty}\phi_\theta=\lim_{\theta\to\infty}\phi_\theta^r=\delta_0 \quad \text{in } \mathcal{S}'(\mathbb{R}^2). \tag{3.11}$$

In particular, by (3.11) the study of the limit behavior of the function $(x,t)\to(\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)*\phi_\theta^r)(x)$, $x\in\mathbb{R}^2$, $0\le t\le T$, in order to obtain information on the asymptotics of $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ as $N,\theta\to\infty$, $\delta,h(,\tau)\to 0$ is justified.

As a consequence of (3.1), (3.3), (3.9), and (3.10) we furthermore deduce with

$$\phi_1\in C_b^2(\mathbb{R}^2), \qquad \phi_\theta\in C_b^2(\mathbb{R}^2),\ \theta>1, \tag{3.12}$$

a strengthening of the first relation in (1.5). The remaining parts of (1.5) are obvious consequences of (3.1) and (3.2).

Remark 2. Similarly as in our analytical investigations of moderately interacting many-particle systems [12], [13], [14], [15], [16] the *factorization property* (3.1) will be needed to transfer crucial parts of the study of the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ to an L^2-context, where the attention is drawn to the functions $(x,t)\to(\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)*\phi_\theta^r)(x)$. For that purpose the assumption that ϕ_1^r and therefore also ϕ_θ^r is a symmetric probability density, cf. (3.2), (3.10), is essential. The remaining assumptions about ϕ_1 and ϕ_1^r are technical and needed only to deduce various estimates. We note that (3.2)–(3.8) are satisfied, in particular, if $\phi_1^r=C(-\Delta+1)^{-n}$, $n=2,3,\dots$, where C is a suitable constant.

Next, we suppose that the initial state ρ_0 of the limit process ρ is a smooth probability density, i.e.,

$$\rho_0\ge 0, \qquad \int_{\mathbb{R}^2} dx\ \rho_0(x)=1, \qquad \|\rho_0\|_{(m,1)}<\infty,\ m=0,1,2,\dots, \tag{3.13}$$

where the norms $\|.\|_{(m,1)}$, $m = 0, 1, 2, \ldots$, are introduced in (1.9).

We note that the limit dynamics (1.1) represents a simple nonlinear parabolic problem, which may be handled quite well by classical methods; cf. Remark (i) in [17, section 2]. In particular, (3.13) implies

$$\rho(.,t) \geq 0, \quad \int_{\mathbb{R}^2} dx\ \rho(x,t) = 1, \qquad t \geq 0, \tag{3.14}$$
$$\rho \in C_b^\infty\big(\mathbb{R}^2 \times [0,T]\big), \qquad \sup_{t \leq T} \|\rho(.,t)\|_{(m)} < \infty,\ m = 0, 1, 2, \ldots.$$

In (2.8) the constant K is introduced as upper bound for the absolute value of the drift acting on the particles in the many-particle system (2.17) or (2.18). The convergence $\lim_{N,\theta \to \infty,\ \delta,h(,\tau) \to 0} \mathbb{Y}_{N,\delta,h,\theta(,\tau)} = \rho$ of course can only be expected if K is also in this limit an upper bound to the drift. This is guaranteed, for example, if K satisfies

$$\sup_{t \leq T}\ \|\nabla \rho(.,t)\|_\infty + 1 < K. \tag{3.15}$$

We note that the left side of (3.15) is finite as a consequence of (3.14). We also mention that the interpretation of $-\nabla\rho(.,t)$ as *drift* is an immediate consequence of the fact that (1.1) may be considered as a *Fokker–Planck equation*.

As indicated in section 1 we have to take care of the range of the interaction in the various dynamical systems appearing in this paper. In particular, many-particle systems like (1.2) or its analogues conceived for computer simulations like (2.17) or (2.18) and the partial differential equation (1.1) differ in their range of interaction. By (1.3) the range in the system (1.2) is $O(\theta_N^{-1})$, and by (1.8) and the construction of $D_{N,\delta,h,\theta(,\tau)}$ in section 2 the systems (2.17) and (2.18) have a range of order $O(\theta^{-1})$. Hence, in (1.2) and (2.17) or (2.18) the interaction has strictly positive range for any $N \in \mathbb{N}$, $\delta \in (0, \delta_0)$, $h > 0$, $\theta > 1$ (and $\tau > 0$). On the other hand, in the partial differential equation (1.1) the interaction is strictly local; i.e., its range is 0. This essential difference in the structure of the many-particle systems and their limit, respectively, will lead to difficulties when the *distance* between $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ and ρ is studied analytically. For this reason we introduce as an auxiliary tool a family of deterministic functions ρ_θ, $\theta > 1$, whose time evolutions are governed by an interaction similar as in the many-particle systems (1.2), (2.17), or (2.18), and which furthermore are approximating the limit ρ as $\theta \to \infty$. In particular, we define ρ_θ as the solution of the integro-differential equation

$$\partial_t \rho_\theta = \frac{1}{2}\Delta\rho_\theta + \nabla \cdot \big(\rho_\theta \nabla(\rho_\theta * \phi_\theta)\big), \qquad \rho_\theta(.,0) = \rho_0, \tag{3.16}$$

where ϕ_θ is the interaction kernel employed in (2.17) or (2.18), and ρ_0 coincides with the initial state of ρ; cf. (1.1).

By (3.11) the *formal* convergence of ρ_θ to the solution ρ of (1.1) is quite obvious. The convergence can be stated rigorously, also quantitatively, and additional regularity properties, which hold uniformly in θ, can be deduced; cf. Theorem 3 in [17]. In particular, for any $\theta > \theta_0$, where $\theta_0 = \theta_0(T)$ is sufficiently large, there exists a unique solution $\rho_\theta \in C_b^\infty(\mathbb{R}^2 \times [0,T])$ of (3.16), which satisfies

$$\rho_\theta(.,t) \geq 0, \quad \int_{\mathbb{R}^2} dx\ \rho_\theta(x,t) = 1, \qquad t \leq T,\ \theta > \theta_0, \tag{3.17}$$

$$\sup_{t\le T,\theta>\theta_0} \|\rho_\theta(.,t)\|_{(m,1)} < \infty, \quad m = 0,1,2,\dots,$$
$$\sup_{x\in\mathbb{R}^2,t\le T,\theta>\theta_0} |\nabla^{\otimes m}\partial_t^k \rho_\theta(x,t)| < \infty, \quad m,k = 0,1,2,\dots,$$

and

$$\sup_{x\in\mathbb{R}^2,t\le T,\theta>\theta_0} \theta^2|\nabla^{\otimes m}\partial_t^k(\rho_\theta(x,t)-\rho(x,t))| < \infty, \quad m,k = 0,1,2,\dots. \tag{3.18}$$

We note that the conditions (3.1), (3.2), (3.8), and (3.13) are employed in [17, Theorem 3], to verify (3.17) and (3.18).

In particular, possibly after increasing θ_0, the relations (3.15) and (3.18) imply

$$\sup_{t\le T,\theta>\theta_0} \|\nabla\rho_\theta(.,t)\|_\infty < K. \tag{3.19}$$

So far the parameters N, δ, h, θ (and τ) have been considered as independent. However, the convergence $\lim_{N,\theta\to\infty,\ \delta,h\to 0} \mathbb{Y}_{N,\delta,h,\theta} = \rho$ or $\lim_{N,\theta\to\infty,\ \delta,h,\tau\to 0} \mathbb{Y}_{N,\delta,h,\theta,\tau} = \rho$ can only be expected when certain relations for these parameters hold. For example, the mesh size δ has to be much smaller than the interaction range θ^{-1}. In this context, we introduce some particular sets

$$\begin{aligned}\mathcal{P}_1 &= \big\{(N,\delta,h,\theta)\in\mathbb{N}\times(0,\delta_0)\times(0,\infty)\times(\theta_0,\infty) : (\delta^2+h)\theta^8 < \overline{C}\big\},\\ \mathcal{P}_2 &= \big\{(N,\delta,h,\theta,\tau)\in\mathbb{N}\times(0,\delta_0)\times(0,\infty)\times(\theta_0,\infty)\times(0,\infty) :\\ &\qquad\qquad \tau^2 + h\theta^8 + \delta^4\theta^{10} < \overline{C}\big\},\end{aligned} \tag{3.20}$$

where $\overline{C}$ is some arbitrary but fixed constant. Obviously, in order that the parameters N, δ, h, θ (and τ) stay in these sets, θ may grow only slowly as $\delta, h \to 0$.

3.2. Results and supplementary remarks. By (3.18) the convergence of ρ_θ to ρ as $\theta\to\infty$ is specified quite precisely. Hence, as a crucial part of our considerations we have to study the distance between $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ and ρ_θ. A result sufficient for our purposes is given by the following proposition.

PROPOSITION 1. *For any $N\in\mathbb{N}$, $\delta\in(0,\delta_0)$, $h>0$, $\theta>\theta_0$ (and $\tau>0$) we consider the solution ρ_θ of* (3.16) *and also the many-particle system as described in section* 2. *We emphasize that the positions $Y^k_{N,\delta,h,\theta(,\tau)}(t)$, $0\le t\le T$, $k=1,\dots,N$, of the particles of that many-particle system solve* (2.17) *or* (2.18)*, where the negative drift $D_{N,\delta,h,\theta(,\tau)}$ is determined according to the particular discretization procedure summarized essentially by* (2.6)–(2.8). *The associated empirical process $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ is introduced in* (2.1). *Moreover, the constants δ_0, θ_0, and K are characterized in* (2.2) *and the last part of section* 3.1*, respectively.*

We assume that the interaction kernels ϕ_θ, $\theta>\theta_0$, which determine the time evolution of the many-particle system and the solution of (3.16)*, satisfy* (1.8) *and* (3.1)–(3.8). *Furthermore, for the initial state ρ_0 of ρ and ρ_θ, $\theta>\theta_0$, the regularity properties* (3.13) *are supposed to hold. Then, the distance between $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ and ρ_θ may be estimated as follows.*

(a) *If the adaption step* (C) *in section* 2 *is not performed, we get*

$$\begin{aligned}\mathbb{E}\bigg[\sup_{t\le T}\|(\mathbb{Y}_{N,\delta,h,\theta}(t)-\rho_\theta(.,t))*\phi_\theta^r\|_2^2 + \int_0^T dt\ \|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(t)-\rho_\theta(.,t))*\phi_\theta^r)\|_2^2\bigg]\\ \le C(T)\left(\mathbb{E}\big[\|(\mathbb{Y}_{N,\delta,h,\theta}(0)-\rho_0)*\phi_\theta^r\|_2^2\big] + \frac{\theta^4}{N} + (\delta^2+h)\theta^8 + \theta^{2(1-\overline{k})}\right),\\ (N,\delta,h,\theta)\in\mathcal{P}_1.\end{aligned} \tag{3.21}$$

(b) *If the adaption step* (C) *in section* 2 *is included, we deduce*

$$\mathbb{E}\Big[\sup_{t\le T}\|(\mathbb{Y}_{N,\delta,h,\theta,\tau}(t)-\rho_\theta(.,t))*\phi_\theta^r\|_2^2+\int_0^T dt\ \|\nabla((\mathbb{Y}_{N,\delta,h,\theta,\tau}(t)-\rho_\theta(.,t))*\phi_\theta^r)\|_2^2\Big]$$
$$\le C(T)\Big(\mathbb{E}\big[\|(\mathbb{Y}_{N,\delta,h,\theta,\tau}(0)-\rho_0)*\phi_\theta^r\|_2^2\big]+\frac{\theta^4}{N}+\delta^4\theta^{10}+\tau^2+h\theta^8+\theta^{2(1-\overline{k})}\Big),$$
$$(N,\delta,h,\theta,\tau)\in\mathcal{P}_2. \tag{3.22}$$

By Proposition 1 and estimates for $\|\rho(.,t)-\rho_\theta(.,t)\|_2^2$ and $\|\rho_\theta(.,t)-\rho_\theta(.,t)*\phi_\theta^r\|_2^2$, which are provided by (3.17), (3.18) and Lemma 4, we may now quite easily deduce our main estimates.

THEOREM 2. *Let us consider the same situation as in Proposition* 1. *Then, the distance between* $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ *and* ρ *may be estimated as follows.*

(a) *If the adaption step* (C) *in section* 2 *is not performed, we get*

$$\mathbb{E}\Big[\sup_{t\le T}\|\mathbb{Y}_{N,\delta,h,\theta}(t)*\phi_\theta^r-\rho(.,t)\|_2^2+\int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta}(t)*\phi_\theta^r-\rho(.,t))\|_2^2\Big] \tag{3.23}$$
$$\le C(T)\left(\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta}(0)*\phi_\theta^r-\rho_0\|_2^2\big]+\frac{\theta^4}{N}+(\delta^2+h)\theta^8+\theta^{-2}\right),$$
$$(N,\delta,h,\theta)\in\mathcal{P}_1.$$

(b) *If the adaption step* (C) *in section* 2 *is included, we deduce*

$$\mathbb{E}\Big[\sup_{t\le T}\|\mathbb{Y}_{N,\delta,h,\theta,\tau}(t)*\phi_\theta^r-\rho(.,t)\|_2^2+\int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta,\tau}(t)*\phi_\theta^r-\rho(.,t))\|_2^2\Big] \tag{3.24}$$
$$\le C(T)\left(\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta,\tau}(0)*\phi_\theta^r-\rho_0\|_2^2\big]+\frac{\theta^4}{N}+\delta^4\theta^{10}+\tau^2+h\theta^8+\theta^{-2}\right),$$
$$(N,\delta,h,\theta,\tau)\in\mathcal{P}_2.$$

Remark 3. (i) The right sides of related estimates in Proposition 1 and Theorem 2 essentially coincide, since for any $t\ge 0$ the convergence $\lim_{\theta\to\infty}\rho_\theta(.,t)*\phi_\theta^r=\rho(.,t)$ is fast in comparison to the convergence $\lim_{N,\theta\to\infty,\ \delta,h(,\tau)\to 0}(\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)-\rho_\theta(.,t))*\phi_\theta^r=0$.

The factor $1/N$ on the right sides of (3.21)–(3.24) corresponds to the usual $1/\sqrt{N}$-dependence of the *distance* between some deterministic object and its *Monte-Carlo simulation* involving N quantities, which are stochastically independent.

(ii) As a consequence of the scaling relations (1.8) and (3.9) the diameters of the respective *supports* of the kernels ϕ_θ and ϕ_θ^r decrease with the same order $O(\theta^{-1})$ as $\theta\to\infty$. Here, we have to define "support" in a generalized sense, e.g., as the set of those points in $\mathbb{R}^2$, whose distance to the mean 0 of the probability densities ϕ_θ or ϕ_θ^r is less than the respective *standard error*. Hence, since the function ρ is smooth, cf. (3.14), Theorem 2 demonstrates that in our situation on the scale of its range of interaction, which obviously may be identified with the diameter of the support of ϕ_θ, the many-particle system (2.17) or (2.18) behaves regularly; i.e., the *interaction force* acting on the particles does not exhibit extreme *fluctuations*. In retrospect, this observation is a justification of the discretization step (B) and the subsequent interpolation step (D) in section 2.

(iii) In a simple method the initial positions $Y^k_{N,\delta,h,\theta(,\tau)}(0)$, $k=1,\dots,N$, of the particles may be chosen as independently and identically distributed (i.i.d.) random

variables in $\mathbb{R}^2$ with density ρ_0. By (3.10) we then obtain

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0) * \phi_\theta^r - \rho_0\|_2^2\big] \\
&\quad = \mathbb{E}\big[\langle \mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0), \mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0) * \phi_\theta\rangle\big] - 2\mathbb{E}\big[\langle \mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0), \rho_0 * \phi_\theta^r\rangle\big] + \|\rho_0\|_2^2 \\
&\quad = \frac{1}{N^2}\sum_{k,l=1}^{N} \mathbb{E}\big[\phi_\theta\big(Y^k_{N,\delta,h,\theta(,\tau)}(0) - Y^l_{N,\delta,h,\theta(,\tau)}(0)\big)\big] \\
&\qquad - \frac{2}{N}\sum_{k=1}^{N} \mathbb{E}\big[(\rho_0 * \phi_\theta^r)\big(Y^k_{N,\delta,h,\theta(,\tau)}(0)\big)\big] + \|\rho_0\|_2^2 \\
&\quad = \frac{N(N-1)}{N^2}\int_{\mathbb{R}^2} dx \int_{\mathbb{R}^2} dy\ \rho_0(x)\rho_0(y)\phi_\theta(x-y) + \frac{1}{N}\phi_\theta(0) \\
&\qquad - 2\int_{\mathbb{R}^2} dx\ \rho_0(x)(\rho_0 * \phi_\theta^r)(x) + \|\rho_0\|_2^2 \\
&\quad = \frac{N-1}{N}\|\rho_0 * \phi_\theta^r\|_2^2 + \frac{1}{N}\phi_\theta(0) - 2\langle \rho_0, \rho_0 * \phi_\theta^r\rangle + \|\rho_0\|_2^2 \\
&\quad = \|\rho_0 * \phi_\theta^r - \rho_0\|_2^2 - \frac{1}{N}\|\rho_0 * \phi_\theta^r\|_2^2 + \frac{1}{N}\phi_\theta(0).
\end{aligned}$$

Consequently, (3.2), (3.3), (3.9), (3.13), and Lemma 4 imply

$$\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(0) * \phi_\theta^r - \rho_0\|_2^2\big] \le C\left(\theta^{-4} + \frac{\theta^2}{N}\right). \tag{3.25}$$

Since $N, \theta \ge 1$, we observe that for this particular choice of the initial positions of the particles (3.23) may be replaced by

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{t\le T}\|\mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r - \rho(.,t)\|_2^2 + \int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r - \rho(.,t))\|_2^2\bigg] \\
\le C(T)\Big(\frac{\theta^4}{N} + (\delta^2 + h)\theta^8 + \theta^{-2}\Big), \quad (N,\delta,h,\theta) \in \mathcal{P}_1,
\end{aligned} \tag{3.26}$$

and that (3.24) turns into

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{t\le T}\|\mathbb{Y}_{N,\delta,h,\theta,\tau}(t) * \phi_\theta^r - \rho(.,t)\|_2^2 + \int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta,\tau}(t) * \phi_\theta^r - \rho(.,t))\|_2^2\bigg] \\
\le C(T)\left(\frac{\theta^4}{N} + \delta^4\theta^{10} + \tau^2 + h\theta^8 + \theta^{-2}\right), \quad (N,\delta,h,\theta,\tau) \in \mathcal{P}_2.
\end{aligned} \tag{3.27}$$

These considerations demonstrate that in the present situation an optimized choice of the initial particle positions, e.g., by employing so-called *low-discrepancy methods*, cf. [11], does not lead to an improvement of the final bound for $\mathbb{E}\big[\sup_{t\le T}\|\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t) * \phi_\theta^r - \rho(.,t)\|_2^2 + \int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t) * \phi_\theta^r - \rho(.,t))\|_2^2\big]$.

(iv) Obviously, the CPU-time for the simulation of our many-particle system increases with decreasing δ, h, and τ and increasing N. Moreover, the various terms on the right sides of (3.23), (3.24), (3.26), and (3.27) exhibit opposing monotonicity properties as functions of θ. Hence, to achieve in the limit as $N, \theta \to \infty$ and $\delta, h(, \tau) \to 0$ with minimal efforts a fixed order of accuracy it is suggested to determine N, δ, h, θ (and τ) in such a way that the terms on the right sides of our estimates have equal

weights. Then, any modification of a parameter would lead to either an increasing approximation error or to an increasing CPU-time without improving the accuracy. In particular, first θ has to be chosen such that θ^{-2} corresponds to the desired accuracy. Next, in the case where the adaption step is not performed, N, δ, and h should satisfy

$$N \sim \theta^6, \quad h \sim \theta^{-10}, \quad \delta \sim \theta^{-5}. \tag{3.28}$$

On the other hand, if the adaption step is included, N, δ, h, and τ should be determined according to

$$N \sim \theta^6, \quad h \sim \theta^{-10}, \quad \delta \sim \theta^{-3}, \quad \tau \sim \theta^{-1}. \tag{3.29}$$

(v) Of course, in the derivation of the estimates (3.21)–(3.24) rigorous mathematical arguments will be used. As a supplement we shall present in section 4.3 some formal considerations, which lead to nonrigorous improvements of several intermediate results. Those considerations are essentially based on a hypothesis about the uniform regularity of the functions $\mathbb{R}^2 \times [0,T] \ni (x,t) \to (\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t) * \phi_\theta)(x)$, $N \in \mathbb{N}$, $\delta < \delta_0$, $h < T$, $\theta > \theta_0(, \tau > 0)$, and they finally amount to discard the θ-dependence of exactly those contributions to the right sides of (3.23) and (3.24), which emerge as a consequence of the space-time discretization in our simulation method; cf. (i)–(iii) in section 2.

In particular, if the adaption step (C) in section 2 is not performed, and if (4.8) and (4.11) are modified, the estimate (3.23) may formally be replaced by

$$\begin{aligned} \mathbb{E}\Big[\sup_{t\le T} \|\mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r - \rho(.,t)\|_2^2 + \int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r - \rho(.,t))\|_2^2\Big] \\ \lesssim C(T)\Big(\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta}(0) * \phi_\theta^r - \rho_0\|_2^2\big] + \frac{\theta^4}{N} + \delta^2 + h + \theta^{-2}\Big), \\ (N,\delta,h,\theta) \in \mathcal{P}_1,\ \theta \ll N^{1/2}. \end{aligned} \tag{3.30}$$

Now taking into account (3.25) we observe that for fixed θ instead of (3.28) the quantities N, δ, and h should satisfy

$$N \sim \theta^6, \quad h \sim \theta^{-2}, \quad \delta \sim \theta^{-1} \tag{3.31}$$

to guarantee an "optimal" performance of our simulation procedure.

In the case where the adaption step (C) in section 2 is included, the relation (4.8) has to be replaced by (4.18). As indicated in section 4.3, too, a formal improvement of that estimate is also possible and first leads to

$$\begin{aligned} \mathbb{E}\Big[\sup_{t\le T} \|\mathbb{Y}_{N,\delta,h,\theta,\tau}(t) * \phi_\theta^r - \rho(.,t)\|_2^2 + \int_0^T dt\ \|\nabla(\mathbb{Y}_{N,\delta,h,\theta,\tau}(t) * \phi_\theta^r - \rho(.,t))\|_2^2\Big] \\ \lesssim C(T)\Big(\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta,\tau}(0) * \phi_\theta^r - \rho_0\|_2^2\big] + \frac{\theta^4}{N} + \delta^4 + \tau^2 + h + \theta^{-2}\Big), \\ (N,\delta,h,\theta,\tau) \in \mathcal{P}_2,\ \theta \ll N^{1/2}, \end{aligned} \tag{3.32}$$

instead of (3.24), and subsequently to

$$N \sim \theta^6, \quad h \sim \theta^{-2}, \quad \delta \sim \theta^{-1/2}, \quad \tau \sim \theta^{-1}, \tag{3.33}$$

as a substitute for (3.29).

If in a computer simulation of $Y^k_{N,\delta,h,\theta(,\tau)}(.)$, $k = 1,\dots,N$, the choice of the parameters N, h, δ (and τ) is based on (3.31) or (3.33) instead of (3.28) or (3.29), for both discretization parameters δ and h much larger values may be chosen. In particular, a considerable reduction of the expenses may be expected.

Finally, we mention that those contributions to the derivations of (3.23) and (3.24), which are not related to the space-time discretization in our simulation procedure, have close analogues in the corresponding calculations in the study [13] of the asymptotics of the associated many-particle system (1.2) as $N \to \infty$. It therefore seems that also formally our estimates for those terms cannot be improved.

(vi) In step (A) in section 2 the region $Q \subseteq \mathbb{R}^2$ is characterized as a rectangle "containing most of the particles" during the simulation interval $[0,T]$. To determine Q heuristically we may first introduce some rectangle Q_0 satisfying $\int_{Q_0} dx\ \rho_0(x) \geq 1-\varepsilon$, where $\varepsilon > 0$ and ρ_0, which by (3.13) is a probability density, is the initial state of the solution ρ of (1.1). Next, we use the fact that (1.1) can be considered as a *Fokker–Planck equation* with diffusion coefficient $1/2$ and drift vector $-\nabla\rho(.,t)$, $t \geq 0$. In particular, together with Q_0 the diffusion coefficient and the drift vector can be employed to guess some rectangle $Q = Q_T$ satisfying $\int_Q dx\ \rho(x,t) \gtrsim 1-\varepsilon$, $0 \leq t \leq T$. Of course, in general $\nabla\rho(.,t)$, $t > 0$, is not known in advance. However, the maximum principle implies that the solution ρ of (1.1) in $\mathbb{R}^1$ satisfies $\|\rho'(.,t)\|_\infty \leq \|\rho_0'\|_\infty$, $t \geq 0$. Also in $\mathbb{R}^d$, $d > 1$, as a consequence of its nonsingular diffusion the dynamics (1.1) is smoothing. Hence, it seems to be justified to use $\|\nabla\rho_0\|_\infty$ as an estimate for $\sup_{t\leq T}\|\nabla\rho(.,t)\|_\infty$. When taking into account that a mass density subject to diffusion with coefficient $1/2$ spreads by $\sqrt{T}$ during $[0,T]$ we are led to define $Q = Q_T$ as that rectangle, whose boundaries have the distance $\sqrt{T} + \|\nabla\rho_0\|_\infty T$ to their counterparts in Q_0.

In the simulations described in section 6 the initial state ρ_0 is always a weighted superposition of Gaussian densities. In these cases it is particularly easy to determine Q_0 and $\|\nabla\rho_0\|_\infty$ in terms of the weights, the centers, and the standard deviations of the various Gaussians, and then to guess Q in the way sketched above.

We note that in some sense the quantity $1-\varepsilon$ may be considered as a measure for the fraction of the particle-mesh method within our procedure; cf. Remark 1(ii) in section 2.

(vii) The cut-off step (2.8) for the negative drift $D_{N,\delta,h,\theta(,\tau)}$ acting on the particles appears as a technical tool in the proofs of Proposition 1 and Theorem 2. In particular, for the estimation of the terms on the right side of (4.4) it will be useful that uniformly in $N \in \mathbb{N}$, $\delta < \delta_0$, $h > 0$, $\theta > \theta_0$ (and $\tau > 0$) the drift is bounded. Later on, in the simulations described in section 6 the vector $-\widehat{D}_{N,\delta,h,\theta(,\tau)}$ introduced in (2.6), (2.7), and the associated paragraph is employed as drift; i.e., (2.8) is omitted.

4. Proofs. In this section we present the main parts of the proofs. Several auxiliary results will be provided in section 5.

4.1. Proof of Proposition 1. To simplify our considerations we temporarily concentrate on the proof of (a); i.e., we suppose that in the determination of the negative drift $D_{N,\delta,h,\theta}$ acting on the particles the adaption step (C) in section 2 is not performed.

Essentially, we now proceed in a similar way as in [13] or [14]. In particular, we shall employ the integro-differential equation (3.16) for ρ_θ and the stochastic differential equations (2.18) for the positions $Y^k_{N,\delta,h,\theta}(.)$, $k = 1,\dots,N$, of the particles together with *Itô's formula*, cf. [9], to write down an equation describing the time

evolution of the stochastic process $t \to \|(\mathbb{Y}_{N,\delta,h,\theta}(t) - \rho_\theta(.,t)) * \phi_\theta^r\|_2^2$. As a result, with (2.1) and (3.10) we get

$$
\begin{aligned}
(4.1)\ & \|(\mathbb{Y}_{N,\delta,h,\theta}(t)-\rho_\theta(.,t)) * \phi_\theta^r\|_2^2 \\
&= \big\langle \mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r, \mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r \big\rangle \\
&\quad - 2\big\langle \mathbb{Y}_{N,\delta,h,\theta}(t) * \phi_\theta^r, \rho_\theta(.,t) * \phi_\theta^r \big\rangle + \big\langle \rho_\theta(.,t) * \phi_\theta^r, \rho_\theta(.,t) * \phi_\theta^r \big\rangle \\
&= \frac{1}{N^2} \sum_{k,l=1}^N \phi_\theta\big(Y^k_{N,\delta,h,\theta}(t) - Y^l_{N,\delta,h,\theta}(t)\big) \\
&\quad - \frac{2}{N} \sum_{k=1}^N \big(\rho_\theta(.,t) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(t)) + \big\langle \rho_\theta(.,t), \rho_\theta(.,t) * \phi_\theta \big\rangle \\
&= \|(\mathbb{Y}_{N,\delta,h,\theta}(0) - \rho_0) * \phi_\theta^r\|_2^2 \\
&\quad - \int_0^t ds\ \frac{2}{N^2} \sum_{k,l=1}^N \nabla\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s) - Y^l_{N,\delta,h,\theta}(s)\big) \cdot D_{N,\delta,h,\theta}\big(Y^k_{N,\delta,h,\theta}(\lfloor s \rfloor_h), \lfloor s \rfloor_h\big) \\
&\quad + \int_0^t ds\ \frac{1}{N^2} \sum_{\substack{k,l=1,\dots,N \\ k \neq l}} \Delta\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s) - Y^l_{N,\delta,h,\theta}(s)\big) \\
&\quad + \int_0^t \frac{2}{N^2} \sum_{k,l=1}^N \nabla\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s) - Y^l_{N,\delta,h,\theta}(s)\big) \cdot dW^k(s) \\
&\quad + \int_0^t ds\ \frac{2}{N} \sum_{k=1}^N \nabla\big(\rho_\theta(.,s) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s)) \cdot D_{N,\delta,h,\theta}\big(Y^k_{N,\delta,h,\theta}(\lfloor s \rfloor_h), \lfloor s \rfloor_h\big) \\
&\quad - \int_0^t ds\ \frac{1}{N} \sum_{k=1}^N \Delta\big(\rho_\theta(.,s) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s)) \\
&\quad - \int_0^t \frac{2}{N} \sum_{k=1}^N \nabla\big(\rho_\theta(.,s) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s)) \cdot dW^k(s) \\
&\quad + \int_0^t ds\ 2\big\langle \rho_\theta(.,s), \nabla(\mathbb{Y}_{N,\delta,h,\theta}(s) * \phi_\theta) \cdot \nabla(\rho_\theta(.,s) * \phi_\theta) \big\rangle \\
&\quad - \int_0^t ds\ \big\langle \rho_\theta(.,s), \Delta(\mathbb{Y}_{N,\delta,h,\theta}(s) * \phi_\theta) \big\rangle \\
&\quad - \int_0^t ds\ 2\big\langle \rho_\theta(.,s), |\nabla(\rho_\theta(.,s) * \phi_\theta)|^2 \big\rangle + \int_0^t ds\ \big\langle \rho_\theta(.,s), \Delta(\rho_\theta(.,s) * \phi_\theta) \big\rangle, \\
&\quad 0 \le t \le T,\ N \in \mathbb{N},\ \delta \in (0, \delta_0),\ h > 0,\ \theta > \theta_0.
\end{aligned}
$$

To determine an upper bound for the right side of (4.1) we collect the various terms in several groups.

First, we deduce for the terms which contain the Laplace operator and therefore represent the deterministic contributions of the Brownian motions the relation

$$
\begin{aligned}
(4.2)\ & \frac{1}{N^2} \sum_{\substack{k,l=1,\dots,N \\ k \neq l}} \Delta\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s) - Y^l_{N,\delta,h,\theta}(s)\big) - \frac{1}{N} \sum_{k=1}^N \Delta\big(\rho_\theta(.,s) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s)) \\
&\quad - \big\langle \rho_\theta(.,s), \Delta(\mathbb{Y}_{N,\delta,h,\theta}(s) * \phi_\theta) \big\rangle + \big\langle \rho_\theta(.,s), \Delta(\rho_\theta(.,s) * \phi_\theta) \big\rangle
\end{aligned}
$$

$$
\begin{aligned}
&= \big\langle \mathbb{Y}_{N,\delta,h,\theta}(s), \Delta(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta)\big\rangle - \frac{1}{N}\Delta\phi_\theta(0) - \big\langle \mathbb{Y}_{N,\delta,h,\theta}(s), \Delta(\rho_\theta(.,s)*\phi_\theta)\big\rangle \\
&\quad - \big\langle \rho_\theta(.,s), \Delta(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta)\big\rangle + \big\langle \rho_\theta(.,s), \Delta(\rho_\theta(.,s)*\phi_\theta)\big\rangle \\
&= -\big\langle \nabla(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta^r), \nabla(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta^r)\big\rangle - \frac{\theta^4}{N}\Delta\phi_1(0) \\
&\quad + 2\big\langle \nabla(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta^r), \nabla(\rho_\theta(.,s)*\phi_\theta^r)\big\rangle - \big\langle \nabla(\rho_\theta(.,s)*\phi_\theta^r), \nabla(\rho_\theta(.,s)*\phi_\theta^r)\big\rangle \\
&= -\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s))*\phi_\theta^r)\|_2^2 + \frac{\theta^4}{N}\int_{\mathbb{R}^2} dz\ |\nabla\phi_1^r(z)|^2, \\
&\quad 0 \le s \le T,\ N \in \mathbb{N},\ \delta \in (0,\delta_0),\ h > 0,\ \theta > \theta_0,
\end{aligned}
$$

where we have utilized in particular (1.8), (3.1), (3.2), (3.10), and (3.12).

Next, following (2.8) we introduce the notation

$$
\nabla_{(K)} f(y) = \begin{cases} \dfrac{\nabla f(y)}{|\nabla f(y)|}\min\{K, |\nabla f(y)|\} & \text{if } |\nabla f(y)| \neq 0, \\ 0 & \text{if } |\nabla f(y)| = 0, \end{cases} \qquad y \in \mathbb{R}^2,\ f \in C_b^1(\mathbb{R}^2), \tag{4.3}
$$

where the constant K is characterized by (3.15).

Now, collecting those terms on the right side of (4.1), which are related to the interaction, we get

$$
\begin{aligned}
&-\frac{2}{N^2}\sum_{k,l=1}^N \nabla\phi_\theta\big(Y_{N,\delta,h,\theta}^k(s) - Y_{N,\delta,h,\theta}^l(s)\big)\cdot D_{N,\delta,h,\theta}\big(Y_{N,\delta,h,\theta}^k(\lfloor s\rfloor_h), \lfloor s\rfloor_h\big) \\
&\quad + \frac{2}{N}\sum_{k=1}^N \nabla\big(\rho_\theta(.,s)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s))\cdot D_{N,\delta,h,\theta}\big(Y_{N,\delta,h,\theta}^k(\lfloor s\rfloor_h), \lfloor s\rfloor_h\big) \\
&\quad + 2\big\langle \rho_\theta(.,s), \nabla(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta)\cdot\nabla(\rho_\theta(.,s)*\phi_\theta)\big\rangle - 2\big\langle \rho_\theta(.,s), |\nabla(\rho_\theta(.,s)*\phi_\theta)|^2\big\rangle \\
&= \frac{2}{N}\sum_{k=1}^N \nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\qquad \cdot\big(D_{N,\delta,h,\theta}\big(Y_{N,\delta,h,\theta}^k(\lfloor s\rfloor_h), \lfloor s\rfloor_h\big) - \nabla_{(K)}\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(\lfloor s\rfloor_h))\big) \\
&\quad + \frac{2}{N}\sum_{k=1}^N \nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\qquad\qquad \cdot\big(\nabla_{(K)}\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(\lfloor s\rfloor_h)) \\
&\qquad\qquad\qquad - \nabla_{(K)}\big(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s))\big) \\
&\quad + \frac{2}{N}\sum_{k=1}^N \nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\qquad\qquad \cdot\nabla_{(K)}\big(\big(\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s)\big)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\quad + \frac{2}{N}\sum_{k=1}^N \nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\qquad\qquad \cdot\big(\nabla_{(K)}(\rho_\theta(.,s)*\phi_\theta) - \nabla(\rho_\theta(.,s)*\phi_\theta)\big)(Y_{N,\delta,h,\theta}^k(s)) \\
&\quad + 2\big\langle \mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s), \nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)\cdot\nabla(\rho_\theta(.,s)*\phi_\theta)\big\rangle
\end{aligned} \tag{4.4}
$$

$$= \quad \sum_{j=1}^{5} A^j_{N,\delta,h,\theta}(s), \quad 0 \le s \le T,\ N \in \mathbb{N},\ \delta \in (0,\delta_0),\ h > 0,\ \theta > \theta_0.$$

In the subsequent calculations we shall determine upper bounds for $A^j_{N,\delta,h,\theta}$, $j = 1,\dots,5$.

For $A^1_{N,\delta,h,\theta}$ we obtain

$$\begin{aligned}(4.5)\quad |A^1_{N,\delta,h,\theta}(s)| &\le \frac{2}{N}\sum_{k=1}^{N}\big|\nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\ &\qquad \big\|D_{N,\delta,h,\theta}(.,\lfloor s\rfloor_h) - \nabla_{(K)}(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta)\big\|_\infty \\ &= 2\big\langle \mathbb{Y}_{N,\delta,h,\theta}(s), |\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)) * \phi_\theta)|\big\rangle \\ &\qquad \big\|D_{N,\delta,h,\theta}(.,\lfloor s\rfloor_h) - \nabla_{(K)}(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta)\big\|_\infty.\end{aligned}$$

By (3.10) and (3.17) the first factor on the right side of (4.5) can be estimated as

$$\begin{aligned}(4.6)\quad &\big\langle \mathbb{Y}_{N,\delta,h,\theta}(s), |\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)) * \phi_\theta)|\big\rangle \\ &\quad \le \big\langle \mathbb{Y}_{N,\delta,h,\theta}(s), \phi^r_\theta * |\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)) * \phi^r_\theta)|\big\rangle \\ &\quad \le \|\mathbb{Y}_{N,\delta,h,\theta}(s) * \phi^r_\theta\|_2 \|\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)) * \phi^r_\theta)\|_2 \\ &\quad \le \big(\|(\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s)) * \phi^r_\theta\|_2 + 1\big)\|\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)) * \phi^r_\theta)\|_2.\end{aligned}$$

On the other hand, by (1.8), (2.1), (2.6)–(2.8), (3.1), (3.2), (3.12), (4.3), and Lemma 3 we also obtain

$$\begin{aligned}(4.7)\quad &\big\|D_{N,\delta,h,\theta}(.,\lfloor s\rfloor_h) - \nabla_{(K)}(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta)\big\|_\infty \\ &\qquad \le \big\|\widehat{D}_{N,\delta,h,\theta}(.,\lfloor s\rfloor_h) - \nabla(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta)\big\|_\infty \\ &\qquad \le C\delta\|\nabla^{\otimes 2}(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta)\|_\infty \\ &\qquad \le C\delta\theta^4.\end{aligned}$$

Hence, as a summary of (4.5)–(4.7) we deduce

$$\begin{aligned}(4.8)\quad |A^1_{N,\delta,h,\theta}(s)| &\le C\delta^2\theta^8\big(\|(\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s)) * \phi^r_\theta\|_2^2 + 1\big) \\ &\qquad + \frac{1}{8}\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s)) * \phi^r_\theta)\|_2^2, \\ &\qquad 0 \le s \le T,\ N \in \mathbb{N},\ \delta \in (0,\delta_0),\ h > 0,\ \theta > \theta_0.\end{aligned}$$

We note that (4.7) is the only estimate which has to be modified when the adaption step (C) in section 2 is included in our algorithm, i.e., when part (b) of Proposition 1 is proved. In that case we will have to take into account in particular the second part of (5.2).

Next, we turn to the estimation of $|A^2_{N,\delta,h,\theta}|$. By (1.8), (3.10), (3.12), and (4.3)we obtain

$$\begin{aligned}(4.9)\quad &|A^2_{N,\delta,h,\theta}(s)| \\ &\quad \le \frac{2}{N}\sum_{k=1}^{N}\big|\nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\ &\qquad \big|\nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)) \\ &\qquad\qquad -\nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(s) * \phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big|\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{N}\sum_{k=1}^{N}\big|\nabla\big(\big(\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)\big(Y^k_{N,\delta,h,\theta}(s)\big)\big| \\
&\qquad \bigg|\frac{1}{N}\sum_{l=1}^{N}\Big(\nabla\phi_\theta\big(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)\big) \\
&\qquad\qquad -\nabla\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s)-Y^l_{N,\delta,h,\theta}(s)\big)\Big)\bigg| \\
&\le C\theta^4\frac{1}{N^2}\sum_{k,l=1}^{N}\big|\nabla\big(\big(\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)\big(Y^k_{N,\delta,h,\theta}(s)\big)\big| \\
&\qquad \big(|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^k_{N,\delta,h,\theta}(s)|+|Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(s)|\big) \\
&\le C\theta^4\int_{\mathbb{R}^2}dx\,\big|\nabla\big(\big(\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi^r_\theta\big)(x)\big||\gamma_{N,\delta,h,\theta}(x,s)| \\
&\le C\theta^4\|\nabla((\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s))*\phi^r_\theta)\|_2\|\gamma_{N,\delta,h,\theta}(.,s)\|_2,
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma_{N,\delta,h,\theta}(x,s) = \frac{1}{N^2}\sum_{k,l=1}^{N}\phi^r_\theta\big(x-Y^k_{N,\delta,h,\theta}(s)\big)\big(&|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^k_{N,\delta,h,\theta}(s)| \\
&+|Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(s)|\big).
\end{aligned}
$$

Equation (3.10) yields

$$
\begin{aligned}
&\mathbb{E}\big[\|\gamma_{N,\delta,h,\theta}(.,s)\|_2^2\big] \\
&= \mathbb{E}\bigg[\frac{1}{N^4}\sum_{k,k',l,l'=1}^{N}\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s)-Y^{k'}_{N,\delta,h,\theta}(s)\big) \\
&\qquad \big(|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^k_{N,\delta,h,\theta}(s)|+|Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(s)|\big) \\
&\qquad \big(|Y^{k'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^{k'}_{N,\delta,h,\theta}(s)|+|Y^{l'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^{l'}_{N,\delta,h,\theta}(s)|\big)\bigg] \\
&= \frac{1}{N^4}\sum_{k,k',l,l'=1}^{N}\mathbb{E}\Big[\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s)-Y^{k'}_{N,\delta,h,\theta}(s)\big) \\
&\qquad \big(|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^k_{N,\delta,h,\theta}(s)|+|Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(s)|\big) \\
&\qquad \big(|Y^{k'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^{k'}_{N,\delta,h,\theta}(s)|+|Y^{l'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^{l'}_{N,\delta,h,\theta}(s)|\big)\Big] \\
&\le \frac{2}{N^3}\sum_{k,k',l=1}^{N}\mathbb{E}\bigg[\mathbb{E}\Big[\phi_\theta\big(Y^k_{N,\delta,h,\theta}(s)-Y^{k'}_{N,\delta,h,\theta}(s)\big) \\
&\qquad \big(|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^k_{N,\delta,h,\theta}(s)|^2 \\
&\qquad\qquad +|Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^l_{N,\delta,h,\theta}(s)|^2\big)\Big|\mathcal{F}_{\lfloor s\rfloor_h}\Big]\bigg].
\end{aligned}
\tag{4.10}
$$

According to the definition of their dynamics, cf. (2.17) or (2.18), for any triple (k,k',l) with $k\neq k'\neq l\neq k$ the processes $Y^k_{N,\delta,h,\theta}$, $Y^{k'}_{N,\delta,h,\theta}$, and $Y^l_{N,\delta,h,\theta}$ have the form (5.23) in the time interval $\big[\lfloor s\rfloor_h,\lfloor s\rfloor_h+h\big)$, where their respective negative drift, namely $D_{N,\delta,h,\theta}(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h),\lfloor s\rfloor_h)$, $D_{N,\delta,h,\theta}(Y^{k'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h),\lfloor s\rfloor_h)$, and

$D_{N,\delta,h,\theta}(Y^l_{N,\delta,h,\theta}(\lfloor s\rfloor_h),\lfloor s\rfloor_h)$, is constant. Consequently, (1.8), (2.8), (3.10), (3.12), (3.17), (4.10), and Lemma 6 imply

$$\begin{aligned}
&\mathbb{E}\big[\|\gamma_{N,\delta,h,\theta}(.,s)\|_2^2\big]\\
&\quad\le Ch\bigg(\frac{\exp(2K^2T)}{N^3}\sum_{\substack{k,k',l=1,\dots,N\\ k\neq k'}}\mathbb{E}\Big[\big(\phi_\theta*\sigma_{2;8(s-\lfloor s\rfloor_h)}\big)\big(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-Y^{k'}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)\big)\Big]\\
&\qquad\qquad+\frac{\phi_\theta(0)}{N}\bigg)\\
&\quad\le Ch\big(\mathbb{E}\big[\|\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi^r_\theta*\sigma_{2;4(s-\lfloor s\rfloor_h)}\|_2^2\big]+\theta^2/N\big)\\
&\quad\le Ch\big(\mathbb{E}\big[\|(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-\rho_\theta(.,\lfloor s\rfloor_h))*\phi^r_\theta\|_2^2\big]+1+\theta^2/N\big),
\end{aligned}$$

where in the first line of the right side the terms with $k=k'$ contribute $Ch\phi_\theta(0)/N$. Therefore, (4.9) leads to

$$\begin{aligned}
\mathbb{E}\big[|A^2_{N,\delta,h,\theta}(s)|\big]&\le Ch\theta^8\big(\mathbb{E}\big[\|(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)-\rho_\theta(.,\lfloor s\rfloor_h))*\phi^r_\theta\|_2^2\big]+1+\theta^2/N\big)\\
&\quad+\frac{1}{8}\mathbb{E}\big[\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2^2\big],\\
&\quad 0\le s\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0.
\end{aligned}\tag{4.11}$$

Since for any $f\in C^1_b(\mathbb{R}^2)$ and any $x\in\mathbb{R}^2$ the vectors $\nabla f(x)$ and $\nabla_{(K)}f(x)$ are parallel, cf. (4.3), we immediately observe

$$A^3_{N,\delta,h,\theta}(s)\le 0,\quad 0\le s\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0.\tag{4.12}$$

Moreover, by (3.10) and (3.19), which is a consequence of the particular choice of the constants K and θ_0, we conclude

$$A^4_{N,\delta,h,\theta}(s)=0,\quad 0\le s\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0.\tag{4.13}$$

For the estimation of $|A^5_{N,\delta,h,\theta}(s)|$ we may employ Lemma 5 in the case $d=2$, $L=\overline{k}$, $\kappa=\phi^r_1$, and $\kappa_\theta=\phi^r_\theta$. In this situation the assumptions (5.9), (5.11), (5.13), and (5.14) are satisfied by (3.2), (3.6), (3.7), and (3.9). Additionally, we also choose $\mu(dx)=\mathbb{Y}_{N,\delta,h,\theta}(s)(dx)-\rho_\theta(x,s)dx$, $h=\nabla((\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s))*\phi^r_\theta)$, and $g=\nabla\rho_\theta(.,s)*\phi_\theta$. Although both g and h are $\mathbb{R}^2$-valued now, Lemma 5 may still be applied. In particular, with (2.1), (3.10), and (3.17) we obtain

$$\begin{aligned}
&|A^5_{N,\delta,h,\theta}(s)|\\
&\quad\le C\bigg(\|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta\|_2\sum_{l=0}^{\overline{k}-1}\theta^{-l}\|\nabla^{\otimes(l+1)}\rho_\theta(.,s)*\phi_\theta\|_\infty\\
&\qquad\qquad+\theta^{1-\overline{k}}\int_{\mathbb{R}^2}\big\{\mathbb{Y}_{N,\delta,h,\theta}(s)(dx)+\rho_\theta(x,s)dx\big\}\|\nabla^{\otimes(\overline{k}+1)}\rho_\theta(.,s)*\phi_\theta\|_\infty\bigg)\\
&\qquad\|\nabla((\rho_\theta(.,s)-\mathbb{Y}_{N,\delta,h,\theta}(s))*\phi^r_\theta)\|_2\\
&\quad\le C\big(\|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta\|_2+\theta^{1-\overline{k}}\big)\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2\\
&\quad\le C\big(\|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta\|_2^2+\theta^{2(1-\overline{k})}\big)\\
&\qquad+\frac{1}{8}\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2^2,\\
&\qquad 0\le s\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0.
\end{aligned}\tag{4.14}$$

Now, since we have completed the derivation of an upper bound for the right side of (4.4), we turn to those terms on the right side of (4.1) which explicitly contain the Brownian motions $W^1, W^2, \ldots$. For those contributions *Doob's inequality*, cf. [9], (2.1), (3.2), (3.9), and (3.10) imply

$$
\begin{aligned}
(4.15)\quad & \mathbb{E}\Big[\sup_{s\le t}\Big|\int_0^s \frac{2}{N^2}\sum_{k,l=1}^{N}\nabla\phi_\theta\big(Y^k_{N,\delta,h,\theta}(u)-Y^l_{N,\delta,h,\theta}(u)\big)\cdot dW^k(u)\\
&\qquad -\int_0^s \frac{2}{N}\sum_{k=1}^{N}\nabla(\rho_\theta(.,u)*\phi_\theta)(Y^k_{N,\delta,h,\theta}(u))\cdot dW^k(u)\Big|\Big]\\
&= \mathbb{E}\Big[\sup_{s\le t}\Big|\int_0^s \frac{2}{N}\sum_{k=1}^{N}\nabla\big((\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(u))\cdot dW^k(u)\Big|\Big]\\
&\le \frac{C}{\sqrt{N}}\Big(\mathbb{E}\Big[\int_0^t ds\frac{1}{N}\sum_{k=1}^{N}\big|\nabla\big((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big|^2\Big]\Big)^{1/2}\\
&= \frac{C}{\sqrt{N}}\Big(\int_0^t ds\ \mathbb{E}\big[\big\langle\mathbb{Y}_{N,\delta,h,\theta}(s),|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi_\theta)|^2\big\rangle\big]\Big)^{1/2}\\
&\le \frac{C}{\sqrt{N}}\Big(\int_0^t ds\ \mathbb{E}\big[\big\langle\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi^r_\theta,|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)|^2\big\rangle\big]\Big)^{1/2}\\
&\le \frac{C}{N}\|\phi^r_\theta\|_\infty+\frac{1}{8}\int_0^t ds\ \mathbb{E}\big[\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2^2\big]\\
&\le C\frac{\theta^2}{N}+\frac{1}{8}\int_0^t ds\ \mathbb{E}\big[\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2^2\big],\\
&\qquad 0\le t\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0.
\end{aligned}
$$

By (4.2), (4.4), (4.8), and (4.11)–(4.15) the estimates for the terms on the right side of (4.1) are now finished. If the parameters N, δ, h, and θ are restricted to the set $\mathcal{P}_1$, cf. (3.20), and if (3.3) is taken into account, we obtain as a summary

$$
\begin{aligned}
&\|(\mathbb{Y}_{N,\delta,h,\theta}(t)-\rho_\theta(.,t))*\phi^r_\theta\|_2^2\\
&\quad\le \|(\mathbb{Y}_{N,\delta,h,\theta}(0)-\rho_0)*\phi^r_\theta\|_2^2+C\left(\frac{\theta^4}{N}+\delta^2\theta^8+\theta^{2(1-\overline{k})}\right)\\
&\quad -\frac{3}{4}\int_0^t ds\ \|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta)\|_2^2\\
&\quad +C\int_0^t ds\ \|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta\|_2^2+\int_0^t ds\ |A^2_{N,\delta,h,\theta}(s)|\\
&\quad +\sup_{s\le t}\Big|\int_0^s\frac{2}{N}\sum_{k=1}^{N}\nabla\big((\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(u))\cdot dW^k(u)\Big|,\\
&\quad 0\le t\le T,\ (N,\delta,h,\theta)\in\mathcal{P}_1,
\end{aligned}
$$

and therefore

$$
\begin{aligned}
(4.16)\quad &\sup_{s\le t}\Big(\|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi^r_\theta\|_2^2\\
&\qquad +\frac{3}{4}\int_0^s du\ \|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi^r_\theta)\|_2^2\Big)
\end{aligned}
$$

$$\begin{aligned}
&\le \|(\mathbb{Y}_{N,\delta,h,\theta}(0)-\rho_0)*\phi_\theta^r\|_2^2 + C\left(\frac{\theta^4}{N}+\delta^2\theta^8+\theta^{2(1-\overline{k})}\right)\\
&\quad + C\int_0^t ds\ \|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi_\theta^r\|_2^2 + \int_0^t ds\ |A^2_{N,\delta,h,\theta}(s)|\\
&\quad + \sup_{s\le t}\left|\int_0^s \frac{2}{N}\sum_{k=1}^N \nabla\big((\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(u))\cdot dW^k(u)\right|,\\
&\quad 0\le t\le T,\ (N,\delta,h,\theta)\in\mathcal{P}_1.
\end{aligned}$$

Next, by (3.20), (4.11), (4.15), and (4.16) we deduce

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\le t}\left(\|(\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi_\theta^r\|_2^2\right.\right.\\
&\qquad\left.\left. + \frac{3}{4}\int_0^s du\ \|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi_\theta^r)\|_2^2\right)\right]\\
&\quad\le \mathbb{E}\big[\|(\mathbb{Y}_{N,\delta,h,\theta}(0)-\rho_0)*\phi_\theta^r\|_2^2\big] + C\left(\frac{\theta^4}{N}+(\delta^2+h)\theta^8+\theta^{2(1-\overline{k})}\right)\\
&\quad + \frac{1}{4}\int_0^t ds\ \mathbb{E}\big[\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s)-\rho_\theta(.,s))*\phi_\theta^r)\|_2^2\big]\\
&\quad + C\int_0^t ds\ \mathbb{E}\left[\sup_{u\le s}\|(\mathbb{Y}_{N,\delta,h,\theta}(u)-\rho_\theta(.,u))*\phi_\theta^r\|_2^2\right],\\
&\quad 0\le t\le T,\ (N,\delta,h,\theta)\in\mathcal{P}_1.
\end{aligned}$$

Now, by an application of Gronwall's lemma the relation (3.21) follows.

To prove (b) we have only to check the estimate (4.7) another time, since all remaining parts of the proof of (a) are independent of the adaption step (C) in section 2.

In particular, by Lemma 3 and the adaption rule introduced in step (C) in section 2 we obtain

$$\begin{aligned}
&\big\|\widehat{D}_{N,\delta,h,\theta,\tau}(.,\lfloor s\rfloor_h)-\nabla(\mathbb{Y}_{N,\delta,h,\theta,\tau}(\lfloor s\rfloor_h)*\phi_\theta)\big\|_\infty\\
&\quad\le C\left(\sup_{m=1,\dots,|\mathcal{L}(\lfloor s\rfloor_h)|}|\kappa^m_{N,\delta,h,\theta,\tau}(\lfloor s\rfloor_h)|+\delta^2\|\nabla^{\otimes 3}(\mathbb{Y}_{N,\delta,h,\theta,\tau}(\lfloor s\rfloor_h)*\phi_\theta)\|_\infty\right)\\
&\quad\le C(\tau+\delta^2\theta^5).
\end{aligned}\tag{4.17}$$

Consequently, with (4.5), (4.6), and the first part of (4.7) we deduce

$$\begin{aligned}
|A^1_{N,\delta,h,\theta,\tau}(s)| &\le C(\tau^2+\delta^4\theta^{10})\big(\|(\mathbb{Y}_{N,\delta,h,\theta,\tau}(s)-\rho_\theta(.,s))*\phi_\theta^r\|_2^2+1\big)\\
&\quad + \frac{1}{8}\|\nabla((\mathbb{Y}_{N,\delta,h,\theta,\tau}(s)-\rho_\theta(.,s))*\phi_\theta^r)\|_2^2,\\
&\quad 0\le s\le T,\ N\in\mathbb{N},\ \delta\in(0,\delta_0),\ h>0,\ \theta>\theta_0,\ \tau>0,
\end{aligned}\tag{4.18}$$

instead of (4.8).

Now, the proof of (b) may be finished in just the same way as that of (a).

4.2. Proof of Theorem 2. As consequence of (3.14), (3.17), and (3.18) we obtain

$$\begin{aligned}(4.19)\quad &\sup_{t\le T,\theta>\theta_0}\theta^2\|\nabla(\rho_\theta(.,t)-\rho(.,t))\|_2^2\\ &\quad=\sup_{t\le T,\theta>\theta_0}(-\theta^2)\big\langle\Delta(\rho_\theta(.,t)-\rho(.,t)),\rho_\theta(.,t)-\rho(.,t)\big\rangle\\ &\quad\le\sup_{t\le T,\theta>\theta_0}\theta^2\|\Delta(\rho_\theta(.,t)-\rho(.,t))\|_\infty\big\langle\rho_\theta(.,t)+\rho(.,t),1\big\rangle<\infty,\end{aligned}$$

and quite similarly

$$(4.20)\qquad \sup_{t\le T,\theta>\theta_0}\theta^2\|\rho_\theta(.,t)-\rho(.,t)\|_2^2<\infty.$$

By (1.9), (3.17), and Lemma 4 we also observe that

$$(4.21)\qquad \sup_{t\le T,\theta>\theta_0}\theta^4\big(\|\rho_\theta(.,t)-\rho_\theta(.,t)*\phi_\theta^r\|_2^2+\|\nabla(\rho_\theta(.,t)-\rho_\theta(.,t)*\phi_\theta^r)\|_2^2\big)<\infty.$$

We note here that the assumptions (5.9) and (5.10), which are needed for an application of Lemma 4 in the case $\kappa=\phi_1^r$, are satisfied by (3.2) and (3.3).

Now, (3.23) follows from (3.5), (3.21), and (4.19)–(4.21), whereas to obtain (3.24) we have to take into account (3.22) instead of (3.21).

4.3. Some formal improvements. In this subsection we present those formal considerations, which are mentioned in Remark 3(v) in section 3.

To support our arguments we introduce in $\mathbb{R}^2$ a many-particle system, whose dynamics is determined by a system of coupled stochastic differential equations like (1.2), where, however, the interaction is defined in terms of the kernel ϕ_θ; cf. (1.8), instead of ϕ_N; cf. (1.3), (1.4). Obviously, this many-particle system may be described by empirical processes $\mathbb{X}_{N,\theta}$, $N\in\mathbb{N}$, $\theta>1$, which are given in a similar way as in (1.6) or (2.1). As in (1.7) the convergence $\lim_{N,\theta\to\infty}\mathbb{X}_{N,\theta}=\rho$ may be expected, at least, if it is ensured that N and θ converge to their common limit ∞ in such a way that the range of the interaction ($=O(\theta^{-1})$) is much larger than the typical distance between neighboring particles ($=O(N^{-1/2})$). Moreover, $\lim_{N,\theta\to\infty}\partial_t^n\nabla^{\otimes m}(\mathbb{X}_{N,\theta}(.)*\phi_\theta)=\partial_t^n\nabla^{\otimes m}\rho$, $m=0,1,2,3$, $n=0,1$, should also hold. In particular, for $m=0,1,2,3$ and $n=0,1$ the functions $\mathbb{R}^2\times[0,T]\ni(x,t)\to\partial_t^n\nabla^{\otimes m}(\mathbb{X}_{N,\theta}(t)*\phi_\theta)(x)$ are supposed to be "bounded" uniformly in $N\in\mathbb{N}$ and $\theta\ll N^{1/2}$.

Next, we assume that the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}$ and $\mathbb{X}_{N,\theta}$ exhibit similar regularity properties if the discretization parameters δ and h are sufficiently small.

As a summary of the formal assumptions collected so far we are led to the following hypothesis:

(HS) The functions $\mathbb{R}^2\times[0,T]\ni(x,t)\to\partial_t^n\nabla^{\otimes m}(\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t)*\phi_\theta)(x)$, $m=0,1,2,3$, $n=0,1$, are bounded uniformly in $(N,\delta,h,\theta)\in\mathcal{P}_1$ (or $(N,\delta,h,\theta,\tau)\in\mathcal{P}_2$) and $\theta\ll N^{1/2}$.

Since the Brownian motions $W^1,W^2,\ldots$, which are not differentiable, and the negative drift $D_{N,\delta,h,\theta(,\tau)}$, which as a consequence of the space-time discretization is even discontinuous, contribute in the particular form determined by (2.1) and (2.18) to the functions considered in (HS), the regularity properties stated there cannot hold in a strong, classical sense. However, at least some weak version of (HS) is supposed to be valid. That should be sufficient for our purposes, since the calculations sketched below in the remaining part of this subsection are performed in a context, where an integration over $\mathbb{R}^2\times[0,t]$, $0\le t\le T$, is involved.

As a first consequence of (HS) we may omit θ^4 on the right side of (4.7). Subsequently, on the right side of (4.8) the term $\delta^2\theta^8$ may be replaced by δ^2.

Next, for the second contribution to the sum constituting the first estimate for $|A^2_{N,\delta,h,\theta}(s)|$ in (4.9) we deduce

$$\begin{aligned}
&\big|\nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)) - \nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\
&\quad\le \big|\nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h)) - \nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\
&\qquad + \big|\nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(\lfloor s\rfloor_h)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s)) - \nabla\big(\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\
&\quad= Q^{k,1}_{N,\delta,h,\theta}(s) + Q^{k,2}_{N,\delta,h,\theta}(s).
\end{aligned}$$

First, the hypothesis (HS) yields $Q^{k,2}_{N,\delta,h,\theta}(s) \lesssim Ch$. Furthermore, when taking into account additionally the fact that the processes $[v, v+h) \ni u \to Y^k_{N,\delta,h,\theta}(u)$, $v \in \mathcal{T}_h \cap [0,T]$, $k = 1,\dots,N$, are diffusion processes with constant drift bounded by K and diffusion coefficient $1/2$, we observe $Q^{k,1}_{N,\delta,h,\theta}(s) \le C|Y^k_{N,\delta,h,\theta}(\lfloor s\rfloor_h) - Y^k_{N,\delta,h,\theta}(s)| \lesssim C(\sqrt{h}+h)$. As a consequence, we may formally replace (4.9) by

$$\begin{aligned}
|A^2_{N,\delta,h,\theta}(s)| &\lesssim \frac{C\sqrt{h}}{N}\sum_{k=1}^N \big|\nabla\big(\big(\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s)\big)*\phi_\theta\big)(Y^k_{N,\delta,h,\theta}(s))\big| \\
&\le C\sqrt{h}\big\langle \mathbb{Y}_{N,\delta,h,\theta}(s)*\phi^r_\theta, |\nabla((\rho_\theta(.,s) - \mathbb{Y}_{N,\delta,h,\theta}(s))*\phi^r_\theta)|\big\rangle \\
&\le Ch\|\mathbb{Y}_{N,\delta,h,\theta}(s)*\phi^r_\theta\|_2^2 + \frac{1}{8}\|\nabla((\mathbb{Y}_{N,\delta,h,\theta}(s) - \rho_\theta(.,s))*\phi^r_\theta)\|_2^2.
\end{aligned}$$

Hence, on a formal level in (4.11) the factor θ^8 may be omitted.

By the modifications of (4.8) and (4.11) discussed so far we now can conclude that (3.30) may formally be used instead of (3.23).

In the case, where the adaption step (C) in section 2 is performed, we may utilize (HS) to omit on the right sides of (4.17) and (4.18) the factors θ^5 and θ^{10}, respectively. Since the remaining estimates are independent of the adaption step, the arguments leading to (3.30), (3.31) have to be modified only slightly to deduce (3.32), (3.33).

5. Some auxiliary results. The present section contains some auxiliary results, which are needed in the calculations of section 4. Apart from Lemma 3 those results are both formulated and proved essentially independently of the dimension; i.e., they are discussed for arbitrary $\mathbb{R}^d$, $d \ge 1$.

First, we study the approximation of the gradient of some smooth function $f : \mathbb{R}^2 \to \mathbb{R}$ by expressions like those on the right side of (2.7). On the one hand, the quality of that approximation is estimated in terms of $\|\nabla^{\otimes 2} f\|_\infty$. On the other hand, if the components of the tensor $\nabla^{\otimes 2} f$ are replaced by discrete approximations like those introduced in step (E) in section 2, an estimate involving $\|\nabla^{\otimes 3} f\|_\infty$ and an expression like $\kappa_{N,\delta,h,\theta,\tau}(.)$, cf. (2.16), is obtained.

For the subsequent considerations we choose some region $Q \subseteq \mathbb{R}^2$ and a discretization parameter $\delta \in (0, \delta_0)$ according to steps (A) and (B) in section 2. We also select a rectangle $R = R^m$ in $\mathcal{L}_0$ or any of its refinements, which are constructed according to step (C) in section 2. As indicated in section 2, Figure 1, and discussed in steps (B) and (C) in section 2, some points $A = A^m$, $B = B^m$, $C = C^m$, $D = D^m$, $M = M^m$ and some triangles $\Delta_1 = \Delta_1^m$, $\Delta_2 = \Delta_2^m$, $\Delta_3 = \Delta_3^m$, $\Delta_4 = \Delta_4^m$ are associated with R. Moreover, the width of R in x_1- (x_2-) direction is denoted by $S_1 = S_1^m$ ($S_2 = S_2^m$). We note that S_1 and S_2 satisfy (2.5).

Let $f \in C_b^1(\mathbb{R}^2)$ and $x \in R$. Similarly as in (2.7) we introduce an *approximation* $\widehat{\nabla}_R f(x)$ to $\nabla f(x)$ in terms of $f(A)$, $f(B)$, $f(C)$, $f(D)$, and $f(M)$. More precisely, we define

$$\widehat{\nabla}_R f(x) = \begin{cases} \left(\dfrac{f(B)-f(A)}{S_1}, \dfrac{f(A)+f(B)-2f(M)}{S_2}\right)^T & \text{if } x \in \Delta_1, \\ \left(\dfrac{f(B)+f(C)-2f(M)}{S_1}, \dfrac{f(B)-f(C)}{S_2}\right)^T & \text{if } x \in \Delta_2, \\ \left(\dfrac{f(C)-f(D)}{S_1}, \dfrac{2f(M)-f(C)-f(D)}{S_2}\right)^T & \text{if } x \in \Delta_3, \\ \left(\dfrac{2f(M)-f(A)-f(D)}{S_1}, \dfrac{f(A)-f(D)}{S_2}\right)^T & \text{if } x \in \Delta_4. \end{cases} \tag{5.1}$$

If $f \in C_b^2(\mathbb{R}^2)$, we may also apply the considerations in step (E) in section 2 to determine approximations to the components of the tensor $\nabla^{\otimes 2} f(x)$, $x \in R$, which quite similarly as in (2.9)–(2.14) involve only $f(A)$, $f(B)$, $f(C)$, $f(D)$, $f(M)$, and $f(G)$, $G \in \mathcal{G}$, where $\mathcal{G}$ is a set of corners of some rectangles, which are adjacent to R. The absolute value of the product of such an approximation to some component of $\nabla^{\otimes 2} f(x)$, $x \in R$, with $\sqrt{S_1 S_2}$ is bounded above by a quantity $\kappa_R(f)$ defined by an immediate analogue of the right side of (2.16). Since a detailed description of $\kappa_R(f)$ would amount to a more or less literal repetition of the considerations between the paragraph before (2.15) and (2.16), it is omitted here.

Now, we may state our estimates on the error, which is associated with the approximation of ∇ by $\widehat{\nabla}_R$.

LEMMA 3. *Let R be some rectangle as discussed above. Moreover, for $f : \mathbb{R}^2 \to \mathbb{R}$ let $\widehat{\nabla}_R f$ and $\kappa_R(f)$ be determined by* (5.1) *and the considerations in the subsequent paragraph. Then,*

$$\sup_{x \in R} |\nabla f(x) - \widehat{\nabla}_R f(x)| \le \begin{cases} C\|\nabla^{\otimes 2} f\|_\infty \delta & \text{if } f \in C_b^2(\mathbb{R}^2), \\ C\kappa_R(f) + C'\|\nabla^{\otimes 3} f\|_\infty \delta^2 & \text{if } f \in C_b^3(\mathbb{R}^2), \end{cases} \tag{5.2}$$

where the constants C and C' are independent of R and f.

Proof. For $f \in C_b^2(\mathbb{R}^2)$ and $x \in R$ the values $f(A)$, $f(B)$, $f(C)$, $f(D)$, and $f(M)$ may be obtained by a Taylor expansion of f at x. For example,

$$\begin{aligned} f(A) = f(x) &+ (A-x)\cdot \nabla f(x) \\ &+ \frac{1}{2}\sum_{i,j=1}^{2} (A_i - x_i)(A_j - x_j)\nabla_i \nabla_j f\big(x + \vartheta(A,x,f)(A-x)\big), \end{aligned} \tag{5.3}$$

where $0 \le \vartheta(\ldots) \le 1$. Employing (5.3) and a corresponding expression for $f(B)$ in the case where $x \in \Delta_1$ we immediately get

$$\begin{aligned} \frac{f(B)-f(A)}{S_1} &= \frac{f(B)-f(A)}{B_1 - A_1} \\ &= \frac{1}{B_1 - A_1}(B-A)\cdot \nabla f(x) + g_1(A,B,x,f) = \nabla_1 f(x) + g_1(A,B,x,f) \end{aligned}$$

with

$$|g_1(A,B,x,f)| \le 2\frac{|A-x|^2 + |B-x|^2}{S_1}\|\nabla^{\otimes 2} f\|_\infty \le C\delta\|\nabla^{\otimes 2} f\|_\infty,$$

where (2.5) is also used. For $x \in \Delta_1$ we additionally observe

$$\frac{f(A)+f(B)-2f(M)}{S_2} = \nabla_2 f(x) + g_2(A,B,M,x,f),$$

where $|g_2(A,B,M,x,f)| \le C\delta \|\nabla^{\otimes 2} f\|_\infty$. These relations and analogues for the remaining cases $x \in \Delta_2$, $x \in \Delta_3$, $x \in \Delta_4$ on the right side of (5.1) prove the first part of (5.2).

If $f \in C_b^3(\mathbb{R}^2)$ and $x \in R$, as an extension of (5.3) the relation

$$\begin{aligned} f(A) = f(x) + (A-x)\cdot\nabla f(x) + \frac{1}{2}\sum_{i,j=1}^{2}(A_i-x_i)(A_j-x_j)\nabla_i\nabla_j f(x) \\ + \frac{1}{6}\sum_{i,j,k=1}^{2}(A_i-x_i)(A_j-x_j)(A_k-x_k)\nabla_i\nabla_j\nabla_k f\big(x+\vartheta'(A,x,f)(A-x)\big) \end{aligned} \tag{5.4}$$

with $0 \le \vartheta'(\ldots) \le 1$ holds. With a similar expansion for $f(B)$ and by taking into account (2.5) we now deduce

$$\begin{aligned} \frac{f(B)-f(A)}{S_1} &= \frac{f(B)-f(A)}{B_1-A_1} \\ &= \nabla_1 f(x) + \frac{1}{2(B_1-A_1)}\sum_{i,j=1}^{2}\big((B_i-x_i)(B_j-x_j)-(A_i-x_i)(A_j-x_j)\big)\nabla_i\nabla_j f(x) \\ &\quad + g_3(A,B,x,f), \quad x \in \Delta_1, \end{aligned} \tag{5.5}$$

where

$$|g_3(A,B,x,f)| \le C\delta^2 \|\nabla^{\otimes 3} f\|_\infty, \quad x \in \Delta_1. \tag{5.6}$$

The partial derivatives $\nabla_i\nabla_j f(.)$ in (5.5) may be approximated by expressions like those introduced on the right sides of (2.9)–(2.12). For example, suppose $i=j=1$. The considerations in the paragraph preceding (2.15) demonstrate that there exists at least one η_1 in the set $\mathcal{X} = \mathcal{X}^m$ associated with $R = R^m$ such that $P_{\eta_1}, M_{\eta_1}, E_{\eta_1} \in \Theta_{\eta_1}$ are contained in a line parallel to the x_1-axis; i.e., for the x_2-coordinates of these points the relations $P_{\eta_1,2} = M_{\eta_1,2} = E_{\eta_1,2}$ hold. Moreover, M_{η_1} is contained in the closure of R. For convenience we may also suppose $P_{\eta_1,1} < M_{\eta_1,1} < E_{\eta_1,1}$ for the x_1-coordinates. Then by choosing $x = M_{\eta_1}$ in (5.4) and furthermore $A = P_{\eta_1}$ and $A = E_{\eta_1}$, respectively, and by employing (2.5) we deduce

$$\begin{aligned} &\frac{1}{(M_{\eta_1,1}-P_{\eta_1,1})(E_{\eta_1,1}-M_{\eta_1,1})} \\ &\qquad \left(\frac{E_{\eta_1,1}-M_{\eta_1,1}}{E_{\eta_1,1}-P_{\eta_1,1}} f(P_{\eta_1}) + \frac{M_{\eta_1,1}-P_{\eta_1,1}}{E_{\eta_1,1}-P_{\eta_1,1}} f(E_{\eta_1}) - f(M_{\eta_1})\right) \\ &\quad = \frac{1}{2}\nabla_1\nabla_1 f(M_{\eta_1}) + g_{11}(P_{\eta_1}, M_{\eta_1}, E_{\eta_1}, f) \end{aligned} \tag{5.7}$$

with

$$|g_{11}(P_{\eta_1}, M_{\eta_1}, E_{\eta_1}, f)| \le C\delta \|\nabla^{\otimes 3} f\|_\infty. \tag{5.8}$$

Further relations corresponding to (5.7), (5.8) may be deduced for $(i,j) \neq (1,1)$, where again (2.9)–(2.12) are used as a guide. As indicated in the paragraph after (5.1) the respective products of the left sides of (5.7) and its analogues with $\sqrt{S_1 S_2}$ are bounded by a multiple of $\kappa_R(f)$. Hence, with (2.5), (5.1), (5.5), (5.6) and since $\sup_{x\in\Delta_1} |\nabla^{\otimes 2} f(x) - \nabla^{\otimes 2} f(M_{\eta_1})| \leq C\delta \|\nabla^{\otimes 3} f\|_\infty$ we obtain

$$\sup_{x\in\Delta_1} |\nabla_1 f(x) - \widehat{\nabla}_{R,1} f(x)| \leq C\kappa_R(f) + C' \|\nabla^{\otimes 3} f\|_\infty \delta^2,$$

where $\widehat{\nabla}_{R,1} f$ is the x_1-component of $\widehat{\nabla}_R f$.

By repeating now the considerations after (5.4) for $(f(A) + f(B) - 2f(M))/S_2$ instead of $(f(B) - f(A))/S_1$ and then for $x \in \Delta_2, \Delta_3, \Delta_4$ the verification of the second part of (5.2) may be finished. □

Next, we provide estimates for the distance between some smooth function and its convolutions with kernels like ϕ_θ or ϕ_θ^r.

LEMMA 4. *Let κ be some symmetric probability density on $\mathbb{R}^d$; i.e.,*

$$\kappa \geq 0, \quad \int_{\mathbb{R}^d} dx\ \kappa(x) = 1, \quad \kappa(-x) = \kappa(x),\ x \in \mathbb{R}^d, \tag{5.9}$$

which satisfies

$$\widetilde{\kappa} \in C_b^2(\mathbb{R}^d), \tag{5.10}$$

and let

$$\kappa_\theta(x) = \theta^d \kappa(\theta x), \quad x \in \mathbb{R}^d,\ \theta > 1, \tag{5.11}$$

be a family of rescaled versions of κ.

Then, we get

$$\|f - f * \kappa_\theta\|_2 \leq C\theta^{-2} \|\Delta f\|_2 \|\nabla^{\otimes 2} \widetilde{\kappa}\|_\infty, \quad f \in H_2^2(\mathbb{R}^d), \tag{5.12}$$

where the constant C is independent of f and κ.

Proof. Assumptions (5.9)–(5.11) yield

$$\begin{aligned}
\|f - f * \kappa_\theta\|_2^2 &= \int_{\mathbb{R}^d} d\lambda\ |\widetilde{f}(\lambda)|^2 |1 - (2\pi)^{d/2} \widetilde{\kappa_\theta}(\lambda)|^2 \\
&= \int_{\mathbb{R}^d} d\lambda\ |\widetilde{f}(\lambda)|^2 |1 - (2\pi)^{d/2} \widetilde{\kappa}(\lambda/\theta)|^2 \\
&= (2\pi)^d \int_{\mathbb{R}^d} d\lambda\ |\widetilde{f}(\lambda)|^2 \Big| \widetilde{\kappa}(0) - \Big(\widetilde{\kappa}(0) + \frac{1}{\theta} \lambda \cdot \nabla \widetilde{\kappa}(0) \\
&\qquad + \frac{1}{2\theta^2} \sum_{i,j=1}^d \lambda_i \lambda_j \nabla_i \nabla_j \widetilde{\kappa}(\eta(\lambda,\theta)\lambda/\theta) \Big) \Big|^2,
\end{aligned}$$

where $0 \leq \eta(\ldots) \leq 1$. As a consequence of the symmetry of κ, cf. (5.9), we get $\nabla \widetilde{\kappa}(0) = 0$, and therefore

$$\|f - f * \kappa_\theta\|_2^2 \leq C\theta^{-4} \|\nabla^{\otimes 2} \widetilde{\kappa}\|_\infty^2 \int_{\mathbb{R}^d} d\lambda\ |\widetilde{f}(\lambda)|^2 |\lambda|^4,$$

which proves (5.12). □

For the estimation of $A^5_{N,\delta,h,\theta}(s)$, cf. (4.4), we have to study an expression like

$$\langle \mu, (h * \phi^r_\theta) g \rangle, \quad \mu \in \mathcal{M}(\mathbb{R}^2),\ h \in L^2(\mathbb{R}^2),\ g \in C^\infty_b(\mathbb{R}^2),$$

where ϕ^r_θ, $\theta > 1$, is the family of convolution kernels introduced in (3.9) by using the function ϕ^r_1, which in particular satisfies (3.2) and (3.4)–(3.7). As demonstrated by Lemma 5 the absolute value of such expressions may be estimated from above in terms of $\|h\|_2$, L^∞-norms of partial derivatives of g, the L^2-norm of the regularization $\mu * \phi^r_\theta$ of μ, and the total variation $|\mu|(\mathbb{R}^2)$ of μ. We note that some slightly different variant of Lemma 5 can be found in [15, section 4 (ii)].

LEMMA 5. *Suppose that κ and κ_θ, $\theta > 1$, satisfy* (5.9) *and* (5.11) *and let $L \in \{\lceil d/2 \rceil, \lceil d/2 \rceil + 1, \ldots\}$ be fixed. Moreover, assume that the functions*

$$V^l_{k_1,\ldots,k_d}(x) = (-1)^{k_1+\cdots+k_d} \frac{x_1^{k_1} \ldots x_d^{k_d}}{k_1! \ldots k_d!} \kappa(x),$$
$$k_1, \ldots, k_d = 0, \ldots, L,\ 0 \le l = k_1 + \cdots + k_d \le L,\ x \in \mathbb{R}^d,$$

satisfy

$$\left| \widetilde{V^l_{k_1,\ldots,k_d}}(\lambda) \right| \le C |\widetilde{\kappa}(\lambda)|, \tag{5.13}$$
$$k_1, \ldots, k_d = 0, \ldots, L-1,\ 1 \le l = k_1 + \cdots + k_d \le L-1,\ \lambda \in \mathbb{R}^d,$$

and

$$|V^L_{k_1,\ldots,k_d}(x)| \le C \left(\frac{1}{1+|x|^{d+1}} \right)^{1/2}, \tag{5.14}$$
$$k_1, \ldots, k_d = 0, \ldots, L,\ k_1 + \cdots + k_d = L,\ x \in \mathbb{R}^d.$$

Then,

$$\langle \mu, (h * \kappa_\theta) g \rangle = \langle \mu * \kappa_\theta, hg \rangle + R_{\kappa,\theta,L}(\mu, h, g),$$
$$\mu \in \mathcal{M}(\mathbb{R}^d),\ h \in L^2(\mathbb{R}^d),\ g \in C^L_b(\mathbb{R}^d),\ \theta > 1,$$

where

$$|R_{\kappa,\theta,L}(\mu,h,g)| \le C \left(\|\mu * \kappa_\theta\|_2 \sum_{l=1}^{L-1} \theta^{-l} \|\nabla^{\otimes l} g\|_\infty + \theta^{(d/2)-L} |\mu|(\mathbb{R}^d) \|\nabla^{\otimes L} g\|_\infty \right) \|h\|_2,$$
$$\mu \in \mathcal{M}(\mathbb{R}^d),\ h \in L^2(\mathbb{R}^d),\ g \in C^L_b(\mathbb{R}^d),\ \theta > 1.$$

Proof. Similarly as in [15, section 4 (ii)], we first observe that

$$\langle \mu, (h * \kappa_\theta) g \rangle = \left\langle \mu, \int_{\mathbb{R}^d} dz\ \kappa_\theta(z) h(.-z) g(.) \right\rangle \tag{5.15}$$
$$= \left\langle \mu, \int_{\mathbb{R}^d} dz\ \kappa_\theta(z) \Big\{ g(.-z) + \sum_{l=1}^{L-1} \sum_{\substack{s_1,\ldots,s_d=0,\ldots,l \\ s_1+\cdots+s_d=l}} \frac{z_1^{s_1} \ldots z_d^{s_d}}{s_1! \ldots s_d!} \nabla_1^{s_1} \ldots \nabla_d^{s_d} g(.-z) \right.$$
$$\left. + \sum_{\substack{s_1,\ldots,s_d=0,\ldots,L \\ s_1+\cdots+s_d=L}} \frac{z_1^{s_1} \ldots z_d^{s_d}}{s_1! \ldots s_d!} \nabla_1^{s_1} \ldots \nabla_d^{s_d} g(.-z+\vartheta(.,z)z) \Big\} h(.-z) \right\rangle,$$

where $|\vartheta(\ldots)| \leq 1$. In (5.15) we may rearrange the brackets and the integral. Consequently, we get

$$\langle \mu, (h*\kappa_\theta)g\rangle = \sum_{l=0}^{L-1} \sum_{\substack{s_1,\ldots,s_d=0,\ldots,l \\ s_1+\cdots+s_d=l}} R^l_{\theta;s_1,\ldots,s_d}(\mu,h,g) + \sum_{\substack{s_1,\ldots,s_d=0,\ldots,L \\ s_1+\cdots+s_d=L}} R^{*,L}_{\theta;s_1,\ldots,s_d}(\mu,h,g), \tag{5.16}$$

where

$$R^l_{\theta;s_1,\ldots,s_d}(\mu,h,g) = \big\langle \mu * V^l_{\theta;s_1,\ldots,s_d}, h\nabla_1^{s_1}\ldots\nabla_d^{s_d} g\big\rangle, \tag{5.17}$$

$$R^{*,L}_{\theta;s_1,\ldots,s_d}(\mu,h,g) = \left\langle \mu, \int_{\mathbb{R}^d} dz\ \kappa_\theta(z)\frac{z_1^{s_1}\ldots z_d^{s_d}}{s_1!\ldots s_d!}\nabla_1^{s_1}\ldots\nabla_d^{s_d} g(.-z+\vartheta(.,z)z)h(.-z)\right\rangle, \tag{5.18}$$

with

$$V^l_{\theta;s_1,\ldots,s_d}(x) = \theta^{d-l}V^l_{s_1,\ldots,s_d}(\theta x), \quad x\in\mathbb{R}^d. \tag{5.19}$$

Of course, in (5.17)–(5.19) the respective range of l and $s_1,\ldots,s_d$ is determined by (5.15) and (5.16).

For $l=0$ we also have $s_1=\ldots=s_d=0$. Since $V^0_{\theta;0,\ldots,0}(x) = \theta^d V^0_{0,\ldots,0}(\theta x) = \kappa_\theta(x)$, we deduce

$$R^0_{\theta;0,\ldots,0}(\mu,h,g) = \langle \mu*\kappa_\theta, hg\rangle \tag{5.20}$$

in this case. Next, for $l=1,\ldots,L-1$ we get

$$|R^l_{\theta;s_1,\ldots,s_d}(\mu,h,g)| \leq \|\mu * V^l_{\theta;s_1,\ldots,s_d}\|_2 \|h\|_2 \|\nabla_1^{s_1}\ldots\nabla_d^{s_d} g\|_\infty.$$

Since (5.11), (5.13), and (5.19) yield

$$\begin{aligned}
\|\mu * V^l_{\theta;s_1,\ldots,s_d}\|_2^2 &= (2\pi)^d \int_{\mathbb{R}^d} d\lambda\ |\widetilde{\mu}(\lambda)|^2 \left|\widetilde{V^l_{\theta;s_1,\ldots,s_d}}(\lambda)\right|^2 \\
&= \frac{(2\pi)^d}{\theta^{2l}} \int_{\mathbb{R}^d} d\lambda\ |\widetilde{\mu}(\lambda)|^2 \left|\widetilde{V^l_{s_1,\ldots,s_d}}(\lambda/\theta)\right|^2 \\
&\leq \frac{C}{\theta^{2l}} \int_{\mathbb{R}^d} d\lambda\ |\widetilde{\mu}(\lambda)|^2 |\widetilde{\kappa}(\lambda/\theta)|^2 \leq \frac{C}{\theta^{2l}} \|\mu*\kappa_\theta\|_2^2,
\end{aligned}$$

we obtain

$$|R^l_{\theta;s_1,\ldots,s_d}(\mu,h,g)| \leq \frac{C}{\theta^l}\|\mu*\kappa_\theta\|_2\|h\|_2\|\nabla^{\otimes l} g\|_\infty, \quad l=1,\ldots,L-1. \tag{5.21}$$

From (5.9), (5.11), (5.14), (5.18), and (5.19) we finally derive

$$\begin{aligned}
&|R^{*,L}_{\theta;s_1,\ldots,s_d}(\mu,h,g)| \\
&\quad\leq \left\langle |\mu|, \int_{\mathbb{R}^d} dz\ \kappa_\theta(z)\frac{|z_1|^{s_1}\ldots|z_d|^{s_d}}{s_1!\ldots s_d!}|h(.-z)|\right\rangle \|\nabla_1^{s_1}\ldots\nabla_d^{s_d} g\|_\infty \\
&\quad\leq \big\langle |\mu| * |V^L_{\theta;s_1,\ldots,s_d}|, |h|\big\rangle \|\nabla^{\otimes L} g\|_\infty
\end{aligned} \tag{5.22}$$

$$\le \|\nabla^{\otimes L} g\|_\infty \int_{\mathbb{R}^d} |\mu|(dx) \int_{\mathbb{R}^d} dy\ |V^L_{\theta;s_1,\dots,s_d}(y-x)||h(y)|$$
$$\le C\theta^{d-L}\|\nabla^{\otimes L} g\|_\infty \int_{\mathbb{R}^d} |\mu|(dx) \int_{\mathbb{R}^d} dy\ \left(\frac{1}{1+(\theta|y-x|)^{d+1}}\right)^{1/2} |h(y)|$$
$$\le C\theta^{d-L}\|\nabla^{\otimes L} g\|_\infty \int_{\mathbb{R}^d} |\mu|(dx) \left\| \left(\frac{1}{1+(\theta|.-x|)^{d+1}}\right)^{1/2} \right\|_2 \|h\|_2$$
$$\le C\theta^{(d/2)-L}\|\nabla^{\otimes L} g\|_\infty |\mu|(\mathbb{R}^d)\|h\|_2.$$

Equations (5.16) and (5.20)–(5.22) suffice to complete the proof of Lemma 5. □

In our estimate of $A^2_{N,\delta,h,\theta}(s)$; cf. (4.4), the term $\|\gamma_{N,\delta,h,\theta}(.,s)\|_2$ appears; cf. (4.9). As indicated in (4.10), to determine an upper bound of that quantity an estimate for some particular functional of three independent diffusion processes with constant drift and diffusion matrix is needed. This estimate is provided now.

LEMMA 6. *Let*

$$X^j(t) = x^j + D^j t + B^j(t), \quad t \ge 0,\ j = 1,2,3, \tag{5.23}$$

where $x^1, x^2, x^3, D^1, D^2, D^3 \in \mathbb{R}^d$, and B^1, B^2, B^3 are independent, standard Brownian motions in $\mathbb{R}^d$. Then, for any positive $\psi \in L^1(\mathbb{R}^d)$ we get

$$\mathbb{E}\big[\psi(X^1(t)-X^2(t))(X^n(t)-x^n)^2\big]$$
$$\le C(d)t\exp\Big(\frac{t}{2}\big(|D^1|^2+|D^2|^2+|D^3|^2\big)\Big)(\psi * \sigma_{d;8t})(x^1-x^2), \quad t\ge 0,\ n=1,2,3.$$

Proof. The family of transition densities for the standard Brownian motion in $\mathbb{R}^d$ is given by $\{\sigma_{d;s}(.) : s > 0\}$. These kernels can also be used to describe the time evolution of the processes X^j, $j = 1,2,3$, which obviously are obtained from B^1, B^2 and B^3, respectively, by a simple transformation. In particular, for any $n = 1,2,3$ we get

$$\mathbb{E}\big[\psi(X^1(t)-X^2(t))(X^n(t)-x^n)^2\big]$$
$$= \frac{1}{(2\pi t)^{3d/2}} \int_{\mathbb{R}^d} dz^1 \int_{\mathbb{R}^d} dz^2 \int_{\mathbb{R}^d} dz^3\ \psi(z^1-z^2)(z^n-x^n)^2$$
$$\exp\left(-\frac{(x^1-z^1+D^1t)^2}{2t}\right)\exp\left(-\frac{(x^2-z^2+D^2t)^2}{2t}\right)\exp\left(-\frac{(x^3-z^3+D^3t)^2}{2t}\right)$$
$$\le \frac{1}{(2\pi t)^{3d/2}}\exp\left(\frac{t}{2}\big(|D^1|^2+|D^2|^2+|D^3|^2\big)\right)$$
$$\int_{\mathbb{R}^d} dz^1 \int_{\mathbb{R}^d} dz^2 \int_{\mathbb{R}^d} dz^3\ \psi(z^1-z^2)(z^n-x^n)^2$$
$$\exp\left(-\frac{(x^1-z^1)^2}{4t}\right)\exp\left(-\frac{(x^2-z^2)^2}{4t}\right)\exp\left(-\frac{(x^3-z^3)^2}{4t}\right)$$
$$\le \exp\left(\frac{t}{2}\big(|D^1|^2+|D^2|^2+|D^3|^2\big)\right)\frac{t}{(2\pi t)^{3d/2}}\sup_{y\in\mathbb{R}^d}\left(y^2\exp\left(-\frac{y^2}{8}\right)\right)$$
$$\int_{\mathbb{R}^d} dz^1 \int_{\mathbb{R}^d} dz^2 \int_{\mathbb{R}^d} dz^3\ \psi(z^1-z^2)$$
$$\exp\left(-\frac{(x^1-z^1)^2}{8t}\right)\exp\left(-\frac{(x^2-z^2)^2}{8t}\right)\exp\left(-\frac{(x^3-z^3)^2}{8t}\right),$$

which completes the proof of Lemma 6, since $\sigma_{d;\alpha} * \sigma_{d;\beta} = \sigma_{d;\alpha+\beta}$, $\alpha,\beta > 0$. □

6. Computer simulations. To illustrate our results summarized in section 3 we now discuss computer simulations of the many-particle system described in section 2. First, we present some details of our algorithms in order to explain how the general description in section 2 is transfered to our computer program. In the second subsection the results of our simulations are exhibited, in particular in terms of three-dimensional visualizations of *particle densities* and associated approximation errors. We also add some notes on computational expenses. The final part of this section contains some remarks on the hardware and software employed to perform the simulations.

6.1. Algorithmical details. In general, our purpose is to simulate for fixed parameters $N \in \mathbb{N}$, $\delta \in (0, \delta_0)$, $h > 0$, $\theta > \theta_0$ (and $\tau > 0$) the system (2.17) or (2.18) describing the individual motion of the particles, to visualize regularizations of the empirical processes $\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(.)$, cf. (2.1), and also to compare our results with the solution ρ of the limit dynamics (1.1). Within our computer program needed for the simulation we tried to follow the description in section 2 and the assumptions in section 3.1 as close as possible. A few departures, which are subsequently described, should have no essential influence on the relevance of our simulation results.

As far as the negative drift $D_{N,\delta,h,\theta(,\tau)}$ acting on the individual particles is concerned, we omit the cut-off step in (2.8). The kernels ϕ_θ, $\theta > \theta_0$, which characterize the interaction between the particles, cf. (2.6), (2.7), are obtained in our simulations by the scaling (1.8) from a Gaussian density ϕ_1 with mean 0 and variance 0.5. Obviously, the natural choice for ϕ_1^r, cf. (3.1), would then be a centered Gaussian density with variance 0.25. With these functions ϕ_1 and ϕ_1^r the conditions (3.1)–(3.3), (3.7), and (3.8) are satisfied, whereas (3.6) does not hold. We note that this condition appears only as a technical ingredient needed for the application of Lemma 5 to prove (4.14).

Of course, to implement the Gaussian random variables $Z^{k,p}$, $k = 1, \dots, N$, $p = 0, 1, 2, \dots$, in (2.17), which correspond to the increments of the Brownian motions $W^1, W^2, \dots$ in (2.18), *pseudorandom numbers* produced by some random number generator, which is provided by the computer system, are employed.

This random number generator is also applied to determine the initial positions $Y^k_{N,\delta,h,\theta(,\tau)}(0)$, $k = 1, \dots, N$, of the particles according to the method proposed in Remark 3(iii) in section 3, i.e., as i.i.d. random variables with density ρ_0. Since in all our simulations the initial state ρ_0 of (1.1) is chosen as a weighted superposition of Gaussian densities that method is particularly convenient.

Furthermore, the choice of the rectangle Q, cf. step (A) and Remark 1(ii) in section 2, is facilitated for these special initial conditions. In particular, the method described in Remark 3(vi) in section 3 is utilized here.

For the calculation of $\kappa^m_{N,\delta,h,\theta,\tau}(ph)$, which is employed to decide whether a given rectangle R^m in some mesh $\mathcal{L}'(ph)$, cf. step (E) in section 2, should be divided, we do not follow the line leading to (2.16) quite exactly. More precisely, after interchanging S^m_1 and S^m_2 in several places and taking into account (2.5) we finally arrive at a quantity $\widehat{\kappa}^m_{N,\delta,h,\theta,\tau}(ph)$ satisfying $C_3\kappa^m_{N,\delta,h,\theta,\tau}(ph) \le \widehat{\kappa}^m_{N,\delta,h,\theta,\tau}(ph) \le C_4\kappa^m_{N,\delta,h,\theta,\tau}(ph)$, $m = 1, \dots, |\mathcal{L}'(ph)|$, $p = 0, 1, 2, \dots$, where $C_3 = C_3(Q)$ and $C_4 = C_4(Q)$ depend only on the region Q. Obviously, by (2.3) the use of $\widehat{\kappa}_{N,\delta,h,\theta,\tau}$ instead of $\kappa_{N,\delta,h,\theta,\tau}$ is essentially equivalent to the replacement of the adaption parameter τ by some $\widehat{\tau} = C\tau$.

To assess the approximation error of our simulation results we also solve the limit equation (1.1) numerically. Since this way to obtain the solution ρ of (1.1) is not of particular interest in the present paper, we use an extremely simple algorithm, namely

an *Euler scheme* with a *finite-difference method* to approximate the spatial partial derivatives. This approach is justified, since the viscous porous medium equation is a well-behaved partial differential equation for which we also consider only smooth initial conditions. To obtain a reliable representation of ρ we choose a very fine discretization both in space and time. In particular, this discretization is much finer than the discretizations employed for the corresponding particle simulations. In our numerical solution of (1.1) we completely neglect the mass of ρ outside the rectangle Q, which anyhow is chosen in such a way that the mass of ρ is concentrated essentially within Q; cf. Remark 3(vi) in section 3. In particular, we employ *Dirichlet conditions* at the boundary ∂Q of Q.

The discretization used to solve (1.1) is also applied to compute the L^2-norms in the second row of Table 1 in section 6.2.

6.2. Results of the simulations. In order that the quantitative behavior of the particle simulations is clearly displayed, we first consider a special case of (1.1), where the graph of the solution ρ has a particularly simple shape. More precisely, we suppose that ρ_0 is a Gaussian density with mean 0 and variance 1. Later on, we also consider a more complicated initial condition, namely a superposition of four Gaussian densities.

Our estimates in Theorem 2 or (3.26) and (3.27) and also in the formal results (3.30) and (3.32) provide upper bounds for the expectation of a *squared approximation error* $\mathbb{E}[\sup_{t\le T} \|Y_{N,\delta,h,\theta(,\tau)}(t) * \phi_\theta^r - \rho(.,t)\|_2^2]$. On the respective right sides of those estimates only the dependence on the simulation parameters N, δ, h, θ (and τ) is displayed explicitly. In particular, any dependence on the solution ρ of the limit dynamics (1.1) or on the region Q is subsumed in the constants $C(T)$. Hence, our results do not yield precise upper bounds for the approximation error. However, they indicate how a given set $\mathcal{M}_A = \{N_A, \delta_A, h_A, \theta_A(, \tau_A)\}$ of parameters may be modified, in order that the expected squared approximation error, whose exact value is unknown, is reduced by a certain factor $\alpha \in (0,1)$. For this reason we provide several sequences of simulations, where the squared error obtained with some parameter set $\mathcal{M}_A$ should be reduced successively by factors $\alpha_1, \alpha_2, \ldots$.

We first present the results of our simulations by three-dimensional visualizations of the *densities* $y \to d_{N,\delta,h,\theta(,\tau)}(y,t_i) = (\mathbb{Y}_{N,\delta,h,\theta(,\tau)}(t_i) * \phi_\theta)(y)$ for several time points $t_i \in [0,1]$, $i = 1,2,\ldots$. They may be compared to corresponding representations of the numerical solution ρ of (1.1). We note that in all pictures pertaining to the same initial condition ρ_0 any function depicted there is restricted to the respective domain Q. Moreover, the values of these functions are scaled in the same way, and the lighting conditions and the viewpoints do not vary. Since the factor for the scaling in z-direction is considerably larger than 1, the differences between $d_{N,\delta,h,\theta(,\tau)}$ and ρ are overaccentuated. To assess the different simulations quantitatively the respective L^2-norms of the difference $d_{N,\delta,h,\theta(,\tau)}(.,1) - \rho(.,1)$ are also given.

6.2.1. A simply structured initial condition ρ_0. In this first example we suppose that ρ_0 is a Gaussian density with mean 0 and variance 1. The simulation results refering to that situation are presented in Figures 3–5, where in particular the region Q is chosen as $[-5.38, 5.38] \times [-5.38, 5.38]$.

First, in Figure 3 we provide for $t = 0$, $t = 0.5$, and $t = 1$ in the right row three-dimensional visualizations of a numerical solution of (1.1). Additionally, the results of three simulations of the many-particle system (2.17) or (2.18) without the adaption

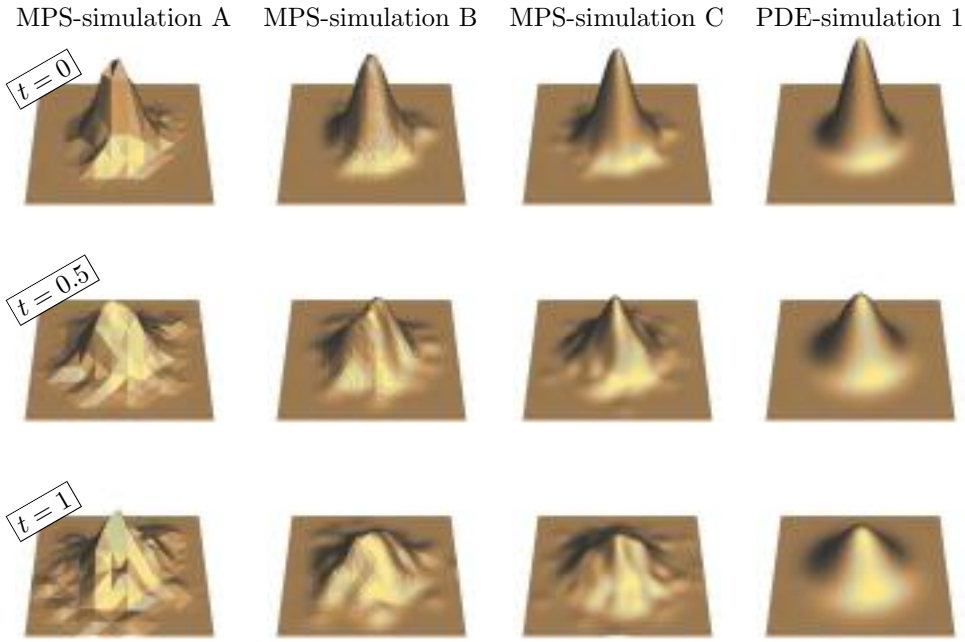


FIG. 3. *Simulations of* (1.1). *Reduction of the approximation error according to* (3.26) *and* (6.2).

step are shown. In MPS-simulation A the parameters

$$N = N_A = 100, \quad \delta = \delta_A = 1, \quad h = h_A = 0.1, \quad \theta = \theta_A = 2 \tag{6.1}$$

are used. In the remaining simulations these parameters are chosen in such a way that the expected squared approximation error estimated by the right side of (3.26) is reduced by the factors 0.75 and 0.5, respectively. In general, a reduction by $\alpha \in (0, 1)$ is achieved if the parameters $N = N_\alpha$, $\delta = \delta_\alpha$, $h = h_\alpha$, and $\theta = \theta_\alpha$ satisfy

$$\frac{1}{\theta_\alpha^2} = \frac{\alpha}{\theta_A^2}, \quad \frac{\theta_\alpha^4}{N_\alpha} = \alpha \frac{\theta_A^4}{N_A}, \quad \delta_\alpha^2 \theta_\alpha^8 = \alpha \delta_A^2 \theta_A^8, \quad h_\alpha \theta_\alpha^8 = \alpha h_A \theta_A^8,$$

i.e., if

$$N_\alpha = \frac{N_A}{\alpha^3}, \quad \delta_\alpha = \delta_A \sqrt{\alpha^5}, \quad h_\alpha = h_A \alpha^5, \quad \theta_\alpha = \frac{\theta_A}{\sqrt{\alpha}}. \tag{6.2}$$

Consequently, after some rounding we get

$$N_B = 300, \quad \delta_B = 0.487, \quad h_B = 0.0237, \quad \theta_B = 2.31$$

for MPS-simulation B with $\alpha = 0.75$, whereas

$$N_C = 800, \quad \delta_C = 0.177, \quad h_C = 0.0031, \quad \theta_C = 2.83$$

are employed for MPS-simulation C with $\alpha = 0.5$.

By (6.2) the discretization parameters δ_α and h_α with decreasing α get very small. Hence, the CPU-time for the particle method grows enormously if (3.26) and the

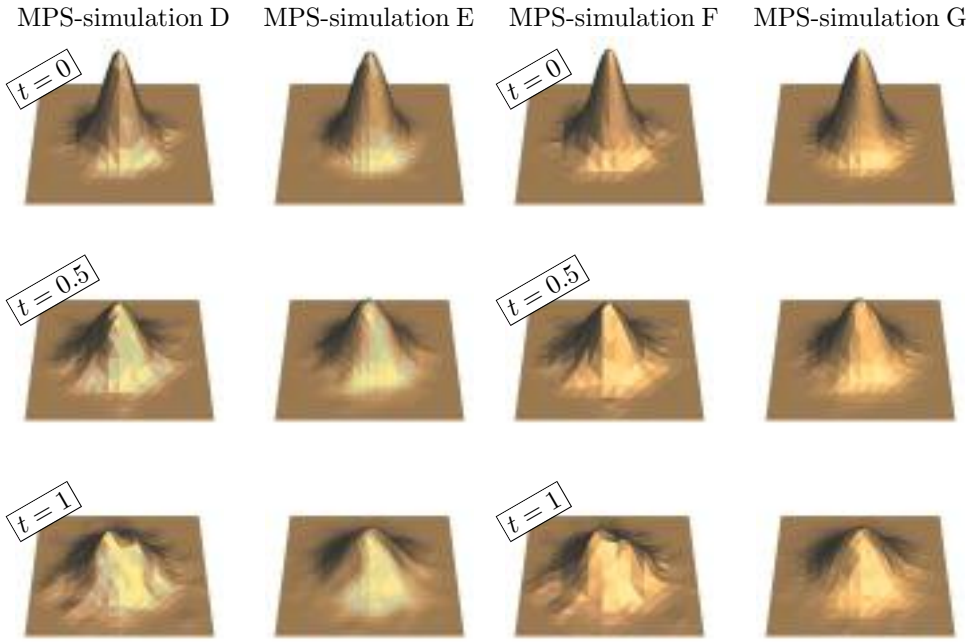


FIG. 4. *Simulations of* (1.1). *Reduction of the approximation error according to* (3.25), (3.30), *and* (6.3).

FIG. 5. *Simulations of* (1.1). *Reduction of the approximation error according to* (3.25), (3.32), *and* (6.5).

associated relations (6.2) are employed as the basis for the choice of the parameters in order to reduce the approximation error. In this respect a considerable improvement may be expected if (3.26) is replaced by the formal result (3.30) supplemented with (3.25). Then, instead of (6.2) the relations

$$N_\alpha = \frac{N_A}{\alpha^3}, \quad \delta_\alpha = \delta_A\sqrt{\alpha}, \quad h_\alpha = h_A\alpha, \quad \theta_\alpha = \frac{\theta_A}{\sqrt{\alpha}} \tag{6.3}$$

are obtained if a reduction of the expected squared approximation error by the factor α has to be attained.

Simulations of our many-particle system (2.17) or (2.18), where (6.3) is applied to improve the accuracy of MPS-simulation A, are presented in Figure 4. In particular, MPS-simulation D and MPS-simulation E refer to $\alpha = 0.5$ and $\alpha = 0.25$, respectively, where (6.3) suggests to use

$$N_D = 800, \quad \delta_D = 0.707, \quad h_D = 0.05, \quad \theta_D = 2.83$$

and

$$N_E = 6400, \quad \delta_E = 0.5, \quad h_E = 0.025, \quad \theta_E = 4. \tag{6.4}$$

The simulations related to Figure 4 are still unnecessarily expensive. In regions where the graph of the density is flat a larger spatial discretization parameter δ should suffice for our calculations. On the other hand, in regions with a large density curvature a finer mesh seems to be desirable, i.e., the use of a globally larger δ,

TABLE 1
Approximation errors and reduction factors for different variants of the particle method.

	$\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2$	α^R_{theor}	α^R_{obs}
MPS-simulation A	0.0440	1.000	1.000
MPS-simulation B	0.0331	0.866	0.752
MPS-simulation C	0.0266	0.707	0.605
MPS-simulation D	0.0217	0.707	0.493
MPS-simulation E	0.0083	0.500	0.189
MPS-simulation F	0.0216	0.707	0.491
MPS-simulation G	0.0086	0.500	0.195

which is reduced only locally in a few critical regions with high curvature, should be advantageous. Accordingly, we next present in Figure 5 some simulations including the adaption step (C) in section 2 in order to provide a local refinement of the spatial mesh.

Just as their counterparts in Figure 4 MPS-simulation F and MPS-simulation G are obtained by reducing the expected squared approximation error of MPS-simulation A by 0.5 and 0.25, respectively. Now, the determination of the simulation parameters is based on (3.25) and the formal estimate (3.32), which yield

$$N_\alpha = \frac{N_A}{\alpha^3}, \quad \delta_\alpha = \delta_A \alpha^{1/4}, \quad \tau_\alpha = \tau_A \sqrt{\alpha}, \quad h_\alpha = h_A \alpha, \quad \theta_\alpha = \frac{\theta_A}{\sqrt{\alpha}} \tag{6.5}$$

if a reduction by the factor α is desired. Therefore, with (6.1), which has been supplemented arbitrarily with $\tau = \tau_A = 0.1$, and (6.5) we get

$$N_F = 800, \quad \delta_F = 0.841, \quad \tau_F = 0.071, \quad h_F = 0.05, \quad \theta_F = 2.83$$

for MPS-simulation F and

$$N_G = 6400, \quad \delta_G = 0.707, \quad \tau_G = 0.05, \quad h_G = 0.025, \quad \theta_G = 4 \tag{6.6}$$

for MPS-simulation G.

So far in this section we have considered the application of several variants of our particle method to the same initial condition. The variants differ in the rule to modify the parameters N, δ, h, θ (and τ) in order to reduce the approximation error. In addition to their visual comparison given above we next in Tables 1 and 2 discuss them quantitatively.

First, in Table 1, in particular, the approximation errors $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2$ are collected. We also contrast the respective *theoretical reduction* α^R_{theor} of that error with its *observed reduction* α^R_{obs}. Here, for MPS-simulation X we use $\alpha^R_{theor} = \sqrt{\alpha}$, where α is the corresponding parameter for the *reduction of the expected squared approximation error* appearing in the preceding paragraphs in this section. Moreover, we define

$$\alpha^R_{obs} = \frac{\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2 \text{ for MPS-simulation X}}{\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2 \text{ for MPS-simulation A}}.$$

In any situation we get

TABLE 2
Comparison of the computational expenses for different variants of the particle method.

	CPU-time (Sec)	$SF(0)$ (Byte)	$SF(1)$ (Byte)
MPS-simulation C	21692.10	522445	518175
MPS-simulation D	113.70	32053	31693
MPS-simulation F	82.64	31851	23252

$$\frac{\alpha^R_{theor}}{\alpha^R_{obs}} > 1; \tag{6.7}$$

i.e., the observed reduction α^R_{obs} is always better than the predicted reduction α^R_{theor}. We additionally notice that the fraction on the left side of (6.7) seems to grow with decreasing $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2$. Perhaps this observation may be attributed to the fact that according to (3.26), (3.27), (3.30), or (3.32) also $\int_0^1 ds\ \|\nabla(d_{N,\delta,h,\theta(,\tau)}(.,s)-\rho(.,s))\|_2^2$ contributes to α and hence to α^R_{theor}. If $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2 \to 0$, this part of α^R_{theor} possibly becomes substantially larger than $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2^2$. To supplement Table 1 we also note that $\|\rho(.,1)\|_2 = 0.195$. In particular, we now may use the *absolute approximation error* $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2$ to determine the *relative approximation error* $\|d_{N,\delta,h,\theta(,\tau)}(.,1)-\rho(.,1)\|_2/\|\rho(.,1)\|_2$.

Table 2 may provide some insight into the computational expenses, which are associated with the different variants of our particle method. In particular, for MPS-simulations C, D, and F, which all serve to reduce a fixed expected squared approximation error by the factor $\alpha = 0.5$, we present the CPU-time for the simulation and additionally the size $SF(t)$ of the files used to store the particle density $d_{N,\delta,h,\theta(,\tau)}(.,t) = (1/N)\sum_{m=1}^N \phi_\theta(.-Y^m_{N,\delta,h,\theta(,\tau)}(t))$ for $t=0$ and $t=1$. As discussed in steps (B) and (C) in section 2 the domain of $d_{N,\delta,h,\theta(,\tau)}(.,t)$ consists of the set of corners and centers of the rectangles in the lattice $\mathcal{L}(t)$. Hence, for any $t = ph$, $p = 0,1,2,\ldots$, the quantity $SF(t)$ corresponds to the number of mesh points in $\mathcal{L}(t)$ and consequently reflects the amount of storage needed to perform the simulation.

Obviously, as a consequence of the small values of δ_C and h_C the expenses for MPS-simulation C are fairly large. On the other hand, Figures 3–5 and Table 1 indicate that apart from the higher resolution the quality of the simulation result associated with MPS-simulation C is not substantially better than the quality obtained with MPS-simulation D or MPS-simulation F, which are performed with much fewer computational efforts. In particular, the formal considerations from section 4.3, which present the background of MPS-simulation D and MPS-simulation F, seem to be quite reliable and useful when a higher computational efficiency is desired. We finally remark that the adaption step in MPS-simulation F in comparison to MPS-simulation D leads to reduced expenses, whereas the quality of the simulation result remains unchanged.

6.2.2. An initial condition with a nontrivial shape. To illustrate the performance of our particle method in a more complicated situation we now suppose that ρ_0 is a superposition of four Gaussian densities. These Gaussians have equal weight 0.25 and variance 1, and their centers span a "T"on the boundary of a rectangle Q' of size 2.5×5. As the domain of the densities obtained by our simulations we choose the rectangle Q of size 10.5×13, whose border lines have distance 4 to the corresponding borders of Q'.

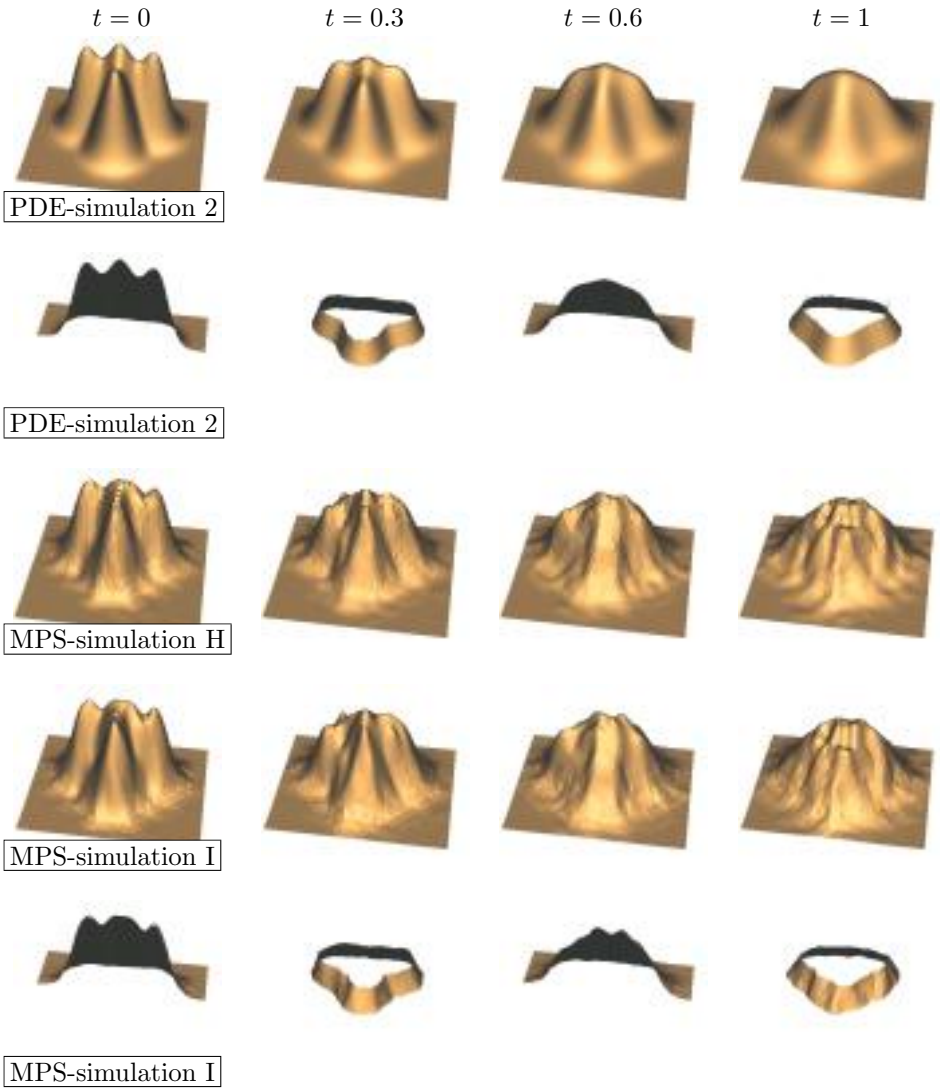


FIG. 6. *Simulations of* (1.1) *with "*T*"-initial condition.*

Our simulation results for $t = 0$, $t = 0.3$, $t = 0.6$, and $t = 1$ are presented in Figure 6. The first line contains three-dimensional visualizations of a numerical solution of (1.1). Modifications of these pictures can be found in the second line. More precisely, in the first and third row the graphs are cut along the line connecting those centers of the Gaussians, which characterize the top of the "T"-initial condition. In the second and fourth row cuts along two different level lines are performed. Visualizations of simulations of the many-particle system (2.17) or (2.18) are collected in the remaining lines of Figure 6. For the simulation parameters N, δ, h, θ (and τ) we essentially apply combinations also employed in the previous simulations described in section 6.2.1. In particular, for MPS-simulation H, where the adaption step is not included, the same parameters N_E, δ_E, h_E, and θ_E as in MPS-simulation E are taken;

cf. (6.4). Finally, the parameters N_G, δ_G, h_G, and θ_G from MPS-simulation G, cf. (6.6), are utilized again now in MPS-simulation I also featuring the adaption step. Since the adaption of the mesh did not begin to work, i.e., the resulting pictures did not differ from those in the third row, when $\tau_G = 0.05$ was chosen again, a smaller adaption parameter $\tau_I = 0.01$ was selected for this simulation. In the last row of Figure 6 modifications of the snapshots in the fourth row of MPS-simulation I are collected, where cuts are performed in exactly the same way as in the pictures in the second row.

6.3. Remarks on the hardware and software. Our simulations have been performed on a PC with an Intel Pentium processor working with 200 MHz, where first Slackware Linux by Patrick Volkerding and later on Redhat Linux was used as the operating system. The simulation program was written in C_{++}. In particular, we used the GNU C_{++}-library to obtain a random number generator and also in order to work with the vector classes and the set- and bag-class prototypes implemented there. To generate visualizations of our results we employed the Mesa 3D graphics library by Brian Paul. Finally, to control the various parts of the simulation program a graphical user interface generated with the XForms library by T. C. Zhao and Mark Overmars was utilized. Presently, efforts are made to replace the GNU C_{++}-library by the C_{++} standard library.

Concluding remarks. In this paper we employ a particular partial differential equation, namely the *viscous porous medium equation* (1.1), to discuss a special particle method based on the concept of *moderately interacting many-particle systems* to obtain an approximation to its solution. We propose an algorithm, which may be classified as a combination of a *smoothed particle hydrodynamics* and a *particle-mesh method*. This algorithm depends on several discretization parameters, namely on a particle number N, a spatial mesh size δ, a time step h, an inverse interaction range θ and, for some modified version, on an adaption parameter τ.

To assess this particle method we give upper bounds for the expected squared approximation error. In particular, our result provides some hints on how to modify the discretization parameters to obtain a desired reduction of the approximation error. Furthermore, the proof of our result suggests employing some regularity hypothesis (HS), cf. section 4.3, about the empirical processes associated with the particle method to deduce improved upper bounds for the approximation error. As a consequence of these formal considerations another approach to modify the discretization parameters is obtained.

To examine our mathematical considerations we perform some computer simulations, where in addition to tests of our particle method we also determine a numerical solution of (1.1) by a simple finite-difference method. We check the dependence of the performance of the particle method, i.e., its precision and its computational expenses, on the strategy to modify the discretization parameters. We observe that the strategy based on (HS) has a considerable advantage.

As mentioned in section 1 particle methods are established in several branches of computational physics. Studies like the present one may be helpful for their improvement. For example, our considerations demonstrate that a careful choice of discretization parameters can have a considerable influence on the performance of particle methods. In this context a mathematical justification of (HS) or an investigation of the dependence of the right side of estimates like (3.23) on the region Q, cf. step (A) in section 2, might be useful.

In our computer experiments the solution of (1.1) by finite differences turned out to be faster. More elaborate methods like finite elements should perform even better. Although in this paper we have not considered them in detail, we presently believe that for the solution of some more general reaction-diffusion equations in low-dimensional spaces those methods should also be preferred. An additional plus for finite elements is the availability of software packages like FEMLAB, cf. http://www.femlab.com/, or FREEFEM, cf. http://www-rocq.inria.fr/Frederic.Hecht/freefem++.htm or [7], for the solution of general partial differential equations. For reaction-diffusion equations in high-dimensional spaces further investigations seem to be needed. In such cases for finite differences or finite elements the number of nodes needed in the respective meshes grows enormously with increasing precision, which may lead to serious drawbacks for these methods.

Moreover, in many situations partial differential equations provide less useful models. For example, for the modelling of biological populations of moderate size, which typically exhibit stochastic fluctuations, many-particle models are more realistic than PDE models, which are conceived to describe the limit of infinite populations. Hence, in this field studies of the simulation of many-particle systems are important. In order to develop effective simulation methods, e.g., the "distance" between "true" many-particle systems like (1.2) and its modifications optimized for computer simulations as that considered in this paper should be investigated. As a first step in this direction the ideas exposed in section 2 have already been used to improve the computer program employed for obtaining the results described in section 6 such that many-particle systems as those described in [14] can be simulated.

REFERENCES

[1] G. V. Bicknell, *The equations of motion of particles in smoothed particle hydrodynamics*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1198–1206.

[2] M. Bossy and D. Talay, *A stochastic particle method for the McKean-Vlasov and the Burgers equation*, Math. Comp., 66 (1997), pp. 157–192.

[3] B. Chauvin and A. Rouault, *A stochastic simulation for solving scalar reaction-diffusion equations*, Adv. in Appl. Probab., 22 (1990), pp. 88–100.

[4] J. M. Dawson, *Particle simulation of plasmas*, Rev. Modern Phys., 55 (1983), pp. 403–447.

[5] R. A. Gingold and J. J. Monaghan, *Kernel estimates as a basis for general particle methods in hydrodynamics*, J. Comput. Phys., 46 (1982), pp. 429–453.

[6] K. J. Havlak and H. D. Victory, Jr., *The numerical analysis of random particle methods applied to Vlasov-Poisson–Fokker-Planck kinetic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 291–317.

[7] F. Hecht and O. Pironneau, *Multiple Unstructured Meshes and the Design of Freefem+*, preprint, 1998.

[8] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, McGraw-Hill, New York, 1981.

[9] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.

[10] D.-G. Long, *Convergence of the random vortex method in two dimensions*, J. Amer. Math. Soc., 1 (1988), pp. 779–804.

[11] H. Neunzert and J. Struckmeier, *Particle Methods for the Boltzmann Equation*, Acta Numer., Cambridge University Press, Cambridge, UK, 1995, pp. 417–457.

[12] K. Oelschläger, *A law of large numbers for moderately interacting diffusion processes*, Z. Wahrsch. Verw. Gebiete, 69 (1985), pp. 279–322.

[13] K. Oelschläger, *A fluctuation theorem for moderately interacting diffusion processes*, Probab. Theory Related Fields, 74 (1987), pp. 591–616.

[14] K. Oelschläger, *On the derivation of reaction-diffusion equations as limit dynamics of systems of moderately interacting stochastic processes*, Probab. Theory Related Fields, 82 (1989), pp. 565–586.

[15] K. Oelschläger, *On the connection between Hamiltonian many-particle systems and the hydrodynamical equations*, Arch. Rational Mech. Anal., 115 (1991), pp. 297–310.
[16] K. Oelschläger, *The description of many-particle systems by the equations for a viscous, compressible, barotropic fluid*, Math. Models Methods Appl. Sci., 5 (1995), pp. 887–922.
[17] K. Oelschläger, *A sequence of integro-differential equations approximating a viscous porous medium equation*, Z. Anal. Anwendungen, 20 (2001), pp. 55–91.
[18] E. G. Puckett, *Convergence of a random particle method to solutions of the Kolmogorov equation* $u_t = \nu u_{xx} + u(1-u)$, Math. Comp., 52 (1989), pp. 615–645.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1763–1774

DESCENT DIRECTIONS OF QUASI-NEWTON METHODS FOR SYMMETRIC NONLINEAR EQUATIONS*

GUANG-ZE GU†, DONG-HUI LI†‡, LIQUN QI‡, AND SHU-ZI ZHOU†

Abstract. In general, when a quasi-Newton method is applied to solve a system of nonlinear equations, the quasi-Newton direction is not necessarily a descent direction for the norm function. In this paper, we show that when applied to solve symmetric nonlinear equations, a quasi-Newton method with positive definite iterative matrices may generate descent directions for the norm function. On the basis of a Gauss–Newton based BFGS method [D. H. Li and M. Fukushima, *SIAM J. Numer. Anal.*, 37 (1999), pp. 152–172], we develop a norm descent BFGS method for solving symmetric nonlinear equations. Under mild conditions, we establish the global and superlinear convergence of the method. The proposed method shares some favorable properties of the BFGS method for solving unconstrained optimization problems: (a) the generated sequence of the quasi-Newton matrices is positive definite; (b) the generated sequence of iterates is norm descent; (c) a global convergence theorem is established without nonsingularity assumption on the Jacobian. Preliminary numerical results are reported, which positively support the method.

Key words. BFGS method, norm descent direction, global convergence, superlinear convergence

AMS subject classifications. 65H10, 90C53

PII. S0036142901397423

1. Introduction.

Let $F : R^n \to R^n$ be continuously differentiable. A general quasi-Newton method for solving the system of nonlinear equations

$$F(x) = 0 \tag{1.1}$$

generates a sequence of iterates $\{x_k\}$ by letting $x_{k+1} = x_k + d_k$, where d_k is a solution of the following system of linear equations:

$$B_k d + F(x_k) = 0. \tag{1.2}$$

If in (1.2), matrix B_k is replaced by $F'(x_k)$, the Jacobian of the function F at x_k, the method reduces to the well-known Newton method. An attractive feature of a quasi-Newton method is its local superlinear convergence property without computation of Jacobians. To enlarge the convergence domain of a quasi-Newton method, line search technique or trust region strategy can be exploited. In this paper, we use a backtracking line search technique to globalize a quasi-Newton method.

A line search step at iteration k of an iterative method determines a scalar $\lambda_k > 0$ which satisfies

$$\|F(x_k + \lambda_k d_k)\| < \|F(x_k)\|. \tag{1.3}$$

The next iterate is then determined by letting $x_{k+1} = x_k + \lambda_k d_k$. The scalar λ_k is called the steplength. Let θ be the norm function defined by

$$\theta(x) = \frac{1}{2}\|F(x)\|^2. \tag{1.4}$$

*Received by the editors November 5, 2001; accepted for publication (in revised form) May 28, 2002; published electronically November 14, 2002. This work was partially supported by the NSF (10171030) of China and the RGC of Hong Kong.
http://www.siam.org/journals/sinum/40-5/39742.html

†Institute of Applied Mathematics, Hunan University, Changsha 410082, China (szzhou@hunu.edu.cn).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (madhli@polyu.edu.hk, maqilq@polyu.edu.hk).

Then the nonlinear equation problem (1.1) is equivalent to the following global optimization problem:

$$\min \theta(x), \qquad x \in R^n, \tag{1.5}$$

and condition (1.3) is equivalent to

$$\theta(x_k + \lambda_k d_k) < \theta(x_k). \tag{1.6}$$

An iterative method that generates a sequence $\{x_k\}$ satisfying (1.3) or (1.6) is called a norm descent method. If d_k is a descent direction of θ at x_k, then inequality (1.6) holds for all $\lambda_k > 0$ sufficiently small. Accordingly, the related iterative method is a norm descent method. In particular, Newton's method with line search is norm descent. For a quasi-Newton method, however, d_k may not be a descent direction of θ at x_k even if B_k is symmetric and positive definite. To globalize a quasi-Newton method, Li and Fukushima [6] proposed an approximately norm descent line search technique and established global and superlinear convergence of a Gauss–Newton based BFGS method for solving symmetric nonlinear equations. The method in [6] is not norm descent. In addition, the global convergence theorem is established under the assumption that $F'(x)$ is uniformly nonsingular.

The purpose of this paper is to develop a norm descent Gauss–Newton based BFGS method. We adjust the steplength and the search direction simultaneously so that the generated iterate sequence satisfies (1.6). We update B_k by combining a modified BFGS formula [7] or the cautious BFGS update rule with the Gauss–Newton based BFGS method [6] such that B_{k+1} inherits positive definiteness of B_k no matter whatever line search is used. Under mild conditions, we establish a global convergence theorem which shows that there exists an accumulation point that is a stationary point of problem (1.5) even if $F'(x)$ is singular everywhere. We also get the superlinear convergence of the proposed method.

In the next section, we describe how to generate a quasi-Newton direction that is descent for θ. We also state the steps of the proposed method. In section 3, we establish the global and superlinear convergence of the proposed method. In section 4, we present some numerical results.

2. Descent direction in a quasi-Newton method. In this section, we describe a way to generate a descent quasi-Newton direction for θ and then propose a norm descent BFGS method for solving (1.1). We assume that the function F is continuously differentiable, and its Jacobian $F'(x)$ is symmetric for every $x \in R^n$.

Recall that in Newton's method, the Newton direction is a solution of the Newton equation

$$F'(x_k)d + F(x_k) = 0. \tag{2.1}$$

Equation (2.1) may have no solution if $F'(x_k)$ is singular. In the case where the solution set of (2.1) is empty, instead of solving (2.1), we may solve the least squares problem

$$\min \frac{1}{2}\|F'(x_k)d + F(x_k)\|^2$$

to get a direction d_k, which results in the so-called Gauss–Newton equation

$$F'(x_k)^2 d + F'(x_k)F(x_k) = 0. \tag{2.2}$$

Here we have used the symmetry of $F'(x_k)$. On the other hand, if $F'(x_k)$ is nonsingular, (2.2) is equivalent to (2.1). In [6], a Gauss–Newton based quasi-Newton method was proposed in which the quasi-Newton direction is the solution of the following system of linear equations:

$$B_k d + \bar{q}_k = 0, \tag{2.3}$$

where B_k is an approximation of matrix $F'(x_k)^2$, and $\bar{q}_k$ is an approximation of vector $F'(x_k)F(x_k)$. Specifically, let λ_{k-1} be the steplength used at the previous iteration. Then, vector $\bar{q}_k$ is defined by

$$\bar{q}_k = (F(x_k + \lambda_{k-1}F(x_k)) - F(x_k))/\lambda_{k-1} \approx F'(x_k)F(x_k),$$

and matrix B_k is updated by the BFGS formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{2.4}$$

where $s_k = x_{k+1} - x_k$, $y_k = F(x_k + \delta_k) - F(x_k)$, and $\delta_k = F(x_{k+1}) - F(x_k)$. It is clear that if $\|s_k\|$ is small, then $B_{k+1}s_k = y_k \approx F'(x_{k+1})^2 s_k$. Since the solution d_k of (2.3) may not be a descent direction of θ at x_k when x_k is far away from a solution of (1.1), it is generally not possible to get a steplength $\lambda_k > 0$ satisfying (1.6). Taking this into account, Li and Fukushima [6] proposed a nondescent line search in which the steplength $\lambda_k > 0$ satisfies the following inequality:

$$\theta(x_k + \lambda_k d_k) - \theta(x_k) \le -\sigma_1 \|\lambda_k d_k\|^2 - \sigma_2 \|\lambda_k F(x_k)\|^2 + \epsilon_k \|F(x_k)\|^2, \tag{2.5}$$

where σ_1 and σ_2 are positive constants, and $\epsilon_k > 0$ satisfies

$$\sum_{k=0}^{\infty} \epsilon_k < \infty.$$

Since ϵ_k is small, $\{x_k\}$ is approximately norm descent.

The purpose of this paper is to develop a norm descent BFGS method. In other words, we want to construct a system of linear equations like (2.3) such that its solution provides a descent direction of θ at x_k.

Observe that

$$\lim_{\lambda_{k-1} \to 0^+} \bar{q}_k = F'(x_k)F(x_k) \triangleq \tilde{q}_k.$$

Accordingly, the solution of (2.3) with $\tilde{q}_k$ instead of $\bar{q}_k$ is $\tilde{d}_k = -B_k^{-1}F'(x_k)F(x_k)$. If B_k is positive definite and $F'(x_k)$ is symmetric, then $\tilde{d}_k$ is a descent direction of θ at x_k. This observation prompts us to regard λ_{k-1} as a parameter. When this parameter is adjusted to be small enough, the solution of (2.3) is a descent direction of θ at x_k. The following process gives details to realize it.

Let

$$q_k(\lambda) = (F(x_k + \lambda F(x_k)) - F(x_k))/\lambda. \tag{2.6}$$

Consider the system of linear equations with parameter λ:

$$B_k d + q_k(\lambda) = 0. \tag{2.7}$$

Let $d(\lambda)$ be the solution of (2.7). The following lemma shows that when $\lambda > 0$ is sufficiently small, every solution of (2.7) is a descent direction of θ at x_k.

LEMMA 2.1. *Let σ_1 and σ_2 be positive constants and B_k be a symmetric and positive definite matrix. If x_k is not a stationary point of* (1.5)*, then there exists a constant $\bar{\lambda} > 0$ depending on k such that when $\lambda \in (0, \bar{\lambda})$, the unique solution $d(\lambda)$ of* (2.7) *satisfies*

$$\nabla\theta(x_k)^T d(\lambda) < 0. \tag{2.8}$$

Moreover, inequality

$$\theta(x_k + \lambda d(\lambda)) - \theta(x_k) \le -\sigma_1 \|\lambda d(\lambda)\|^2 - \sigma_2 \|\lambda F(x_k)\|^2 \tag{2.9}$$

holds for all $\lambda > 0$ sufficiently small.

Proof. It is clear that

$$\lim_{\lambda \to 0} q_k(\lambda) = F'(x_k)F(x_k).$$

Therefore, we get from (2.7) that

$$\begin{aligned}\lim_{\lambda \to 0^+} \nabla\theta(x_k)^T d(\lambda) &= -\lim_{\lambda \to 0^+} F(x_k)^T F'(x_k) B_k^{-1} q_k(\lambda) \\ &= -F(x_k)^T F'(x_k) B_k^{-1} F'(x_k) F(x_k).\end{aligned}$$

Since $F'(x_k)$ is symmetric and $F'(x_k)F(x_k) \neq 0$ as x_k is not a stationary point of (1.5), the last equality and the positive definiteness of B_k imply (2.8). We turn to verifying (2.9).

Notice that

$$\begin{aligned}\lim_{\lambda \to 0^+} (\theta(x_k + \lambda d(\lambda)) - \theta(x_k))/\lambda &= \lim_{\lambda \to 0^+} \nabla\theta(x_k)^T d(\lambda) \\ &= -F(x_k)^T F'(x_k) B_k^{-1} F'(x_k) F(x_k) < 0.\end{aligned}$$

However, the right-hand side of (2.9) is $o(\lambda)$. Therefore, inequality (2.9) holds for all $\lambda > 0$ sufficiently small. ∎

Lemma 2.1 motivates us to find a descent quasi-Newton direction by adjusting parameter λ.

Procedure 1. Let constant $\rho \in (0, 1)$ be given. Let i_k be the smallest nonnegative integer such that inequality (2.9) holds with $\lambda = \rho^i$, $i = 0, 1, \ldots$. Let $d_k = d(\rho^{i_k})$, and $q_k = q_k(\rho^{i_k})$.

Procedure 1 ensures that the value of θ at $x_k + \rho^{i_k} d_k$ is less than that of θ at x_k, though d_k may not necessarily be a descent direction of θ at x_k. It is reasonable to let the scalar ρ^{i_k} be the steplength. However, this steplength may be very small if i_k is large. To enlarge steplength, we exploit the following forward procedure.

Procedure 2. Let i_k and d_k be determined by Procedure 1. If $i_k = 0$, let $\lambda_k = 1$. Otherwise, let j_k be the largest positive integer $j \in \{0, 1, 2, \ldots, i_k - 1\}$ satisfying

$$\theta(x_k + \rho^{i_k - j} d_k) - \theta(x_k) \le -\sigma_1 \|\rho^{i_k - j} d_k\|^2 - \sigma_2 \|\rho^{i_k - j} F(x_k)\|^2. \tag{2.10}$$

Let $\lambda_k = \rho^{i_k - j_k}$.

Note that (2.10) is satisfied with $j = 0$. Therefore, Procedure 2 is well defined.

Procedures 1 and 2 describe a way to generate d_k and λ_k. It is easy to see from Procedures 1 and 2 that

$$\theta(x_k + \lambda_k d_k) - \theta(x_k) \le -\sigma_1 \|\lambda_k d_k\|^2 - \sigma_2 \|\lambda_k F(x_k)\|^2, \tag{2.11}$$

which corresponds to (2.5) with $\epsilon_k = 0$. It is also easy to see that if $\lambda_k \neq 1$, then $\lambda'_k = \lambda_k/\rho$ satisfies

$$\theta(x_k + \lambda'_k d_k) - \theta(x_k) > -\sigma_1 \|\lambda'_k d_k\|^2 - \sigma_2 \|\lambda'_k F(x_k)\|^2. \tag{2.12}$$

Notice that Procedure 1 generates a direction d_k which satisfies

$$B_k d_k + q_k = 0, \tag{2.13}$$

where $q_k = q_k(\rho^{i_k})$. Vector q_k differs from $q_k(\lambda_k)$ if $j_k \neq 0$.

Based on the above process, we propose a norm descent Gauss–Newton based BFGS method as follows.

Algorithm 1 (a descent BFGS method).

Initial Let $B_0 \in R^{n\times n}$ be symmetric and positive definite. Let $x_0 \in R^n$. Set $k = 0$.

Step 1 Determine d_k and λ_k by Procedures 1 and 2. Let $x_{k+1} = x_k + \lambda_k d_k$.

Step 2 Update B_k to get B_{k+1} by the modified BFGS formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{2.14}$$

where $s_k = x_{k+1} - x_k$,

$$y_k = \gamma_k + \Big(\max\Big\{0, -\frac{\gamma_k^T s_k}{\|s_k\|^2}\Big\} + \phi(\|F(x_k)\|)\Big)s_k,$$

$\gamma_k = F(x_k + \delta_k) - F(x_k)$, $\delta_k = F(x_{k+1}) - F(x_k)$, and function $\phi : R \to R$ satisfies (i) $\phi(t) > 0$ for all $t > 0$, (ii) $\phi(t) = 0$ if and only if $t = 0$, (iii) $\phi(t)$ is bounded if t is in a bounded set.

Step 3 Let $k := k + 1$ and go to Step 1.

In Step 2 of Algorithm 1, we use a modified BFGS update formula instead of the ordinary BFGS formula. The modified BFGS update formula was proposed by Li and Fukushima [7], where $\phi(t) = \mu t$ with some constant $\mu > 0$. A favorable property for this modification is that B_{k+1} inherits positive definiteness of B_k whatever line search is used [7]. Indeed, it is not difficult to get that

$$y_k^T s_k \ge \max\Big\{\gamma_k^T s_k,\ \phi(\|F(x_k)\|)\|s_k\|^2\Big\} > 0, \tag{2.15}$$

which is sufficient to guarantee positive definiteness of B_{k+1} as long as B_k is positive definite. Suppose that $\{x_k\}$ is contained in a bounded set at which F is continuously differentiable. It is not difficult to deduce that

$$\|y_k\| \le 2\|\gamma_k\| + \phi(\|F(x_k)\|)\ \|s_k\| \le 2L\|\delta_k\| + M\|s_k\| \le (2L^2 + M)\|s_k\|, \tag{2.16}$$

where $M > 0$ is an upper bound of $\phi(\|F(x)\|)$ and $L > 0$ is a Lipschitz constant of F. Inequalities (2.15) and (2.16) imply that

$$\max\Big\{\gamma_k^T s_k,\ \phi(\|F(x_k)\|)\|s_k\|^2\Big\} \le y_k^T s_k \le (2L^2 + M)\|s_k\|^2. \tag{2.17}$$

Another way to develop quasi-Newton methods is to adopt the so-called cautious update rule proposed by Li and Fukushima [8]. The steps of the related BFGS algorithm is stated as follows.

ALGORITHM 2 (a descent cautious BFGS method).

Initial Let $B_0 \in R^{n\times n}$ be symmetric and positive definite. Let $x_0 \in R^n$. Set $k = 0$.

Step 1 Determine d_k and λ_k by Procedures 1 and 2. Let $x_{k+1} = x_k + \lambda_k d_k$.

Step 2 Update B_k to get B_{k+1} by the cautious BFGS formula

$$B_{k+1} = \begin{cases} B_k - \dfrac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \dfrac{\gamma_k \gamma_k^T}{\gamma_k^T s_k} & \text{if } \dfrac{\gamma_k^T s_k}{\|s_k\|^2} \geq \phi(\|F(x_k)\|), \\ B_k & \text{otherwise,} \end{cases} \tag{2.18}$$

where γ_k and ϕ are the same as those in Algorithm 1.

Step 3 Let $k := k + 1$ and go to Step 1.

The only difference between Algorithms 1 and 2 is the update formula. The cautious BFGS method possesses similar properties of the modified BFGS method. For details, we refer to [8].

3. Global and superlinear convergence. In this section, we prove the global and superlinear convergence of Algorithm 1. The global convergence of Algorithm 2 can be obtained in a similar way. Without specification, we let $\{x_k\}$ and $\{B_k\}$ stand for the sequences of iterates and matrices generated by Algorithm 1, respectively. The following lemma is straightforward from Algorithm 1.

LEMMA 3.1. *The sequence $\{\theta(x_k)\}$ is strictly decreasing. In addition, the following inequalities hold:*

$$\sum_{k=0}^{\infty} \|s_k\|^2 < \infty, \qquad \sum_{k=0}^{\infty} \|\lambda_k F(x_k)\|^2 < \infty. \tag{3.1}$$

We summarize the condition needed for the global convergence of Algorithm 1 as follows.

Assumption A.

(i) The level set

$$\Omega = \{x \in R^n \mid \theta(x) \leq \theta(x_0)\}$$

is bounded.

(ii) Function F is continuously differentiable on Ω, and $F'(x)$ is symmetric for every $x \in \Omega$.

It is clear that under condition (i) in Assumption A, sequence $\{x_k\} \subset \Omega$ is bounded.

We are going to establish a global convergence theorem of Algorithm 1 to show that under Assumption A, there exists an accumulation point of $\{x_k\}$ which is a stationary point of (1.5), namely,

$$\liminf_{k\to\infty} \|\nabla\theta(x_k)\| = 0. \tag{3.2}$$

It is easy to see from Lemma 3.1 that if $\limsup_{k\to\infty} \lambda_k > 0$, then $\liminf_{k\to\infty} \|F(x_k)\| = 0$ and, hence, (3.2) holds. So, we need only to show (3.2) for the case $\lim_{k\to\infty} \lambda_k = 0$. We do it by assuming

$$\liminf_{k\to\infty} \|\nabla\theta(x_k)\| > 0 \tag{3.3}$$

to deduce a contradiction.

Notice that (3.3) particularly implies that there is a constant $\eta > 0$ such that $\|F(x_k)\| \geq \eta$ for all k. It follows from (2.17) and the properties of ϕ that if (3.3) holds, then there are positive constants $c \leq C$ such that

$$c\|s_k\|^2 \leq y_k^T s_k \leq C\|s_k\|^2. \tag{3.4}$$

Therefore, we get the following lemma from (2.16), (3.4), and Theorem 2.1 of [1].

LEMMA 3.2. *If* (3.3) *holds, then there are positive constants* β_i, $i = 1, 2, 3$, *such that for any positive integer* k, *inequalities*

$$\|B_i s_i\| \leq \beta_1 \|s_i\|, \qquad \beta_2 \|s_i\|^2 \leq s_i^T B_i s_i \leq \beta_3 \|s_i\|^2 \tag{3.5}$$

hold for at least $\lceil k/2 \rceil$ *many* $i \leq k$.

Inequalities (3.5) together with (2.13) imply that there are at least $\lceil k/2 \rceil$ many $i \leq k$ satisfying

$$\|q_i\| = \|B_i d_i\| \leq \beta_1 \|d_i\|, \qquad \|d_i\| \leq \beta_2^{-1} \|q_i\|. \tag{3.6}$$

We now prove the global convergence of Algorithm 1.

THEOREM 3.3. *Let Assumption* A *hold and* $\{x_k\}$ *be generated by Algorithm* 1. *Then* (3.2) *holds.*

Proof. We need only to show (3.2) for the case $\lim_{k\to\infty} \lambda_k = 0$. In this case, inequality (2.12) holds for all k sufficiently large. Suppose contrarily that (3.2) does not hold or, equivalently, (3.3) holds. Denote by K the set of indices i such that (3.5) holds. Then K is infinite. Since $\{x_k\} \subset \Omega$ is bounded, it is clear that sequences $\{q_k\}_{k\in K}$ and $\{d_k\}_{k\in K}$ are bounded. Let $K_1 \subset K$ and subsequences $\{x_k\}_{k\in K_1}$ and $\{d_k\}_{k\in K_1}$ converge to x^* and d^*, respectively. Then we have

$$\lim_{k\in K_1} q_k = \nabla\theta(x^*). \tag{3.7}$$

Dividing both sides of (2.12) by λ_k' and then taking limits as $k \to \infty$ with $k \in K_1$, we get

$$\nabla\theta(x^*)^T d^* \geq 0. \tag{3.8}$$

On the other hand, taking the inner product with d_k in (2.13), we get

$$0 = d_k^T B_k d_k + q_k^T d_k \geq \beta_2 \|d_k\|^2 + q_k^T d_k.$$

Taking limits in both sides as $k \to \infty$ with $k \in K_1$ yields

$$\nabla\theta(x^*)^T d^* \leq -\beta_2 \|d^*\|^2.$$

This together with (3.8) implies that $d^* = 0$. It then follows from (3.6) that $\lim_{k\in K_1} q_k = 0$, which together with (3.7) yields a contradiction with (3.3). The contradiction proves (3.2). □

Remark. In [2] the global convergence of Broyden's class of variable metric methods except for DFP was proved. The proof there depends on the convexity of the objective function. A similar result was obtained by Powell [10] when the BFGS method is applied to convex minimization problems. For nonconvex minimization problems, no theory exists to support the global convergence of the BFGS method.

On the contrary, an example has been constructed [3] recently, which shows that the ordinary BFGS method with the Wolfe line search may fail to converge to a stationary point of a nonconvex unconstrained minimization.

On the other hand, a modified BFGS method was proposed by Li and Fukushima [7]. In the modified BFGS method, the iterative matrix B_k is always positive definite whatever line search is used as long as B_0 is positive definite. Moreover, a liminf result was obtained for nonconvex unconstrained minimization. Besides, another modified BFGS method called the cautious BFGS method was proposed by Li and Fukushima [8]. The cautious BFGS method also possesses global convergence in the sense $\liminf_{k\to\infty} \nabla f(x_k) = 0$ when it is applied to $\min f(x)$. In both papers, the results were obtained without the requirement of nonsingular Hessian. These two papers show the possibility to improve the unconstrained minimization result by Byrd, Nocedal, and Yuan [2] and Powell [10].

This paper adopts a similar updating technique as used in [4] and [5]. Consequently, we established Theorem 3.3, which shows that the iterative sequence has an accumulation point which is a stationary point of problem $\min \theta(x) = \frac{1}{2}\|F(x)\|^2$. It may not be a solution of the nonlinear equation (1.1) if the Jacobian is singular at that point.

The next theorem shows a strong convergence property of Algorithm 1.

THEOREM 3.4. *Let Assumption* A *hold. Suppose that the sequence* $\{x_k\}$ *generated by Algorithm* 1 *has a subsequence converging to a stationary* x^* *at which* $F'(x^*)$ *is nonsingular. Then* x^* *is a solution of* (1.1)*. Moreover, the whole sequence* $\{x_k\}$ *converges to* x^**.*

Proof. Since x^* satisfies $\nabla\theta(x^*) = F'(x^*)F(x^*) = 0$, we obviously have $F(x^*) = 0$ if $F'(x^*)$ is nonsingular. Since $\{\theta(x_k)\}$ converges, every accumulation point of $\{x_k\}$ is a solution of (1.1). By the nonsingularity of $F'(x^*)$ again, x^* is an isolated limit point of $\{x_k\}$. However, we have from (3.1) that $x_{k+1} - x_k \to 0$ as $k \to \infty$. Therefore, the whole sequence $\{x_k\}$ converges to x^*. □

In a way similar to the proof of Theorem 3.8 in [7], it is not difficult to prove the superlinear convergence of Algorithm 1. We state the theorem as follows but omit the proof.

THEOREM 3.5. *Let the conditions of Theorem* 3.4 *hold. Suppose further that* F' *is Lipschitz continuous. Then* $\{x_k\}$ *is superlinearly convergent.*

Similar to the above argument, we can establish the global and superlinear convergence of Algorithm 2. We state the results as follows but omit the proof.

THEOREM 3.6. *Let Assumption* A *hold and* $\{x_k\}$ *be generated by Algorithm* 2*. Then* (3.2) *holds. If the sequence* $\{x_k\}$ *has a subsequence converging to a stationary* x^* *at which* $F'(x^*)$ *is nonsingular, then* x^* *is a solution of* (1.1)*. Moreover, the whole sequence* $\{x_k\}$ *converges to* x^**. If we further suppose that* F' *is Lipschitz continuous, then* $\{x_k\}$ *is superlinearly convergent.*

4. Numerical results. In this section, we test the proposed descent BFGS methods on nonlinear equation problems obtained from [6, 9] and the unconstrained optimization problems obtained from the website ftp://ftp.mathworks.com/pub/contrib/v4/optim/uncprobs/. We call Algorithms 1 and 2 the DBFGS (descent BFGS) method and the CBFGS (cautious BFGS) method, respectively, and call the BFGS method based on the Gauss–Newton approach and the nondescent line search [6] the NBFGS (nondescent BFGS) method. Then we compare their performance.

The parameters are specified as follows. We take $\rho = 0.1$ and $\sigma_1 = \sigma_2 = 10^{-5}$

in (2.9). The initial quasi-Newton matrices are set to be $B_0 = A$ [6] for nonlinear equation problems and $B_0 = I$ for unconstrained optimization problems. The function ϕ is determined by

$$\phi(t) = \begin{cases} Ct^2 & \text{if } t \le 1, \\ Ct^{0.1} & \text{otherwise,} \end{cases}$$

where $C = 10^{-5}$. For the NBFGS method, we update B_k by the BFGS formula [6] if $y_k^T s_k \ge 10^{-5}$. Otherwise, we let $B_{k+1} = B_k$. We stop the iteration process if $\|F(x_k)\| \le 10^{-4}$.

The tested results are listed in Tables 1 and 2. Table 3 gives the average performance of the three methods for solving nonlinear equation problems. The columns of the tables have the following meaning:

Dim: the dimension of the problem.
Method: the name of the algorithm.
Init: the initial point, namely, integer l in Table 1 meaning $x_0 = (l, l, \ldots, l)^T$.
Iter: the total number of iterations.
Inner: for the NBFGS method, the number of iterations at which $y_k^T s_k \ge 10^{-5}$ is satisfied; for the DBFGS method and the CBFGS method, the maximum number of inner iterations to generate the descent direction d_k.
Numf: the number of the function evaluations.
Fnorm: the final value of $\|F(x_k)\|$.

All the three methods terminate at solutions of nonlinear equation problems for all tested starting points. However, for the 33 unconstrained optimization problems, all the three methods fail to converge to a solution for at least 10 problems. The numbers of problems for which the NBFGS method, the DBFGS method, and the CBFGS method fail to converge are 16, 19, and 12, respectively.

The numerical results show that for low dimensional problems, the performance of these three methods is not different very much. For most of the test problems, the DBFGS method and the CBFGS method perform better than the NBFGS method in the iteration number, but worse in the number of the function evaluation. However, for high dimensional problems ($n = 200$ in Tables 1 and 3), both the DBFGS and the CBFGS methods perform much better than the NBFGS method in the iteration number as well as the number of the function evaluation. The maximum numbers of the inner iteration to generate a descent direction of a DBFGS method are generally very small. We also note that the performance of the DBFGS and CBFGS methods is almost the same if the both methods terminate regularly. For unconstrained optimization problems, the DBFGS method fails more frequently than the CBFGS method does.

In summary, the presented numerical results reveal that the DBFGS and CBFGS methods, compared with the NBFGS method, have potential advantages when applied to solve symmetric nonlinear equation whose function is not difficult to compute.

In Tables 1–3, we simply denote the NBFGS method as the BFGS method.

TABLE 1
Test results for nonlinear equation problems $B_0 = A$.

Dim	Method	Init	Iter	Inner	Numf	Fnorm	Dim	Method	Init	Iter	Inner	Numf	Fnorm
10	BFGS	0	6	0	19	2.9e-06	50	BFGS	0	25	0	76	3.2e-05
	DBFGS		6	1	19	2.9e-06		DBFGS		25	1	76	3.9e-05
	CBFGS		6	1	19	2.9e-06		CBFGS		25	1	76	3.2e-05
	BFGS	1	11	0	35	9.1e-05		BFGS	1	37	0	114	3.3e-05
	DBFGS		10	2	42	3.5e-05		DBFGS		36	2	120	1.6e-05
	CBFGS		10	2	42	3.5e-05		CBFGS		36	2	120	1.6e-05
	BFGS	-1	12	0	38	2.0e-05		BFGS	-1	37	0	114	1.8e-05
	DBFGS		10	2	36	1.1e-05		DBFGS		36	2	120	1.0e-05
	CBFGS		10	2	36	1.1e-05		CBFGS		36	2	120	1.0e-05
	BFGS	10	13	0	41	4.1e-05		BFGS	10	38	0	117	6.0e-05
	DBFGS		13	2	52	2.6e-05		DBFGS		38	2	133	4.9e-05
	CBFGS		13	2	52	2.6e-05		CBFGS		38	2	133	4.8e-05
	BFGS	-10	13	0	41	3.9e-04		BFGS	-10	38	0	117	6.1e-05
	DBFGS		13	2	52	2.5e-05		DBFGS		38	2	133	5.3e-05
	CBFGS		13	2	52	2.5e-05		CBFGS		38	2	133	4.9e-05
	BFGS	10^2	14	0	44	1.4e-05		BFGS	10^2	41	0	127	4.5e-05
	DBFGS		12	2	48	1.3e-05		BFGS		40	2	132	1.5e-05
	CBFGS		12	2	48	1.3e-05		CBFGS		40	2	132	1.5e-05
	BFGS	-10^2	14	0	44	2.0e-05		BFGS	-10^2	41	0	127	5.0e-05
	DBFGS		12	2	48	1.1e-05		DBFGS		38	2	126	8.6e-05
	CBFGS		12	2	48	1.1e-05		CBFGS		38	2	126	8.6e-05
	BFGS	10^3	16	0	50	1.8e-06		BFGS	10^3	44	0	136	4.1e-05
	DBFGS		13	2	51	3.9e-06		DBFGS		40	2	137	9.9e-05
	CBFGS		13	2	51	3.9e-06		CBFGS		40	2	137	9.9e-05
	BFGS	-10^3	16	0	50	7.5e-07		BFGS	-10^3	44	0	136	3.4e-05
	DBFGS		13	2	51	5.0e-06		DBFGS		40	2	137	7.9e-05
	CBFGS		13	2	51	5.0e-06		CBFGS		40	2	137	7.9e-05
	BFGS	10^4	16	0	50	1.1e-05		BFGS	10^4	49	0	152	5.4e-05
	DBFGS		14	2	54	5.7e-06		DBFGS		44	2	160	4.9e-05
	CBFGS		14	2	54	5.7e-06		CBFGS		44	2	160	4.9e-05
	BFGS	-10^4	16	0	50	8.8e-06		BFGS	-10^4	49	0	144	9.6e-05
	DBFGS		14	2	54	5.1e-06		DBFGS		44	2	152	6.9e-05
	CBFGS		14	2	54	5.1e-06		CBFGS		44	2	152	6.9e-05
100	BFGS	0	50	0	151	5.7e-06	200	BFGS	0	501	0	2181	8.9e-05
	DBFGS		50	1	151	5.7e-06		DBFGS		155	3	1142	9.5e-05
	CBFGS		50	1	151	5.7e-06		CBFGS		151	3	1104	9.5e-05
	BFGS	1	63	0	192	1.0e-04		BFGS	1	2191	1	8661	1.0e-04
	DBFGS		62	2	198	6.8e-05		DBFGS		116	3	383	2.2e-05
	CBFGS		62	2	198	6.8e-05		CBFGS		116	3	383	2.2e-05
	BFGS	-1	63	0	192	5.8e-05		BFGS	-1	1971	1	7835	1.0e-04
	DBFGS		62	2	198	5.4e-05		DBFGS		116	3	382	1.8e-05
	CBFGS		62	2	198	5.4e-05		CBFGS		116	3	382	1.8e-05
	BFGS	10	66	0	203	2.0e-05		BFGS	10	4547	1	18085	1.0e-04
	DBFGS		66	2	220	3.7e-05		DBFGS		119	3	405	4.1e-05
	CBFGS		66	2	217	9.0e-05		CBFGS		119	3	405	4.0e-05
	BFGS	-10	66	0	203	1.2e-05		BFGS	-10	4070	1	16177	1.0e-04
	DBFGS		66	2	221	1.8e-05		DBFGS		119	3	406	2.3e-06
	CBFGS		76	2	254	6.9e-05		CBFGS		119	3	407	2.3e-06
	BFGS	10^2	69	0	213	6.7e-05		BFGS	10^2	6095	1	24277	1.0e-04
	DBFGS		67	2	220	5.7e-05		DBFGS		125	4	447	6.9e-05
	CBFGS		66	2	217	9.0e-05		CBFGS		120	4	424	9.7e-05
	BFGS	-10^2	69	0	213	6.2e-05		BFGS	-10^2	6049	1	24093	1.0e-04
	DBFGS		65	2	211	9.8e-05		DBFGS		125	3	433	6.9e-05
	CBFGS		65	2	211	8.5e-05		CBFGS		124	3	439	5.6e-05
	BFGS	10^3	72	0	221	9.2e-05		BFGS	10^3	8330	1	33217	1.0e-04
	DBFGS		70	2	236	5.8e-05		DBFGS		150	3	543	7.3e-05
	CBFGS		67	2	222	2.3e-05		CBFGS		540	3	3704	8.5e-05
	BFGS	-10^3	71	0	220	4.3e-05		BFGS	-10^3	8296	1	33081	1.0e-04
	DBFGS		69	2	232	8.5e-05		DBFGS		145	3	511	9.8e-05
	CBFGS		67	2	222	2.7e-05		CBFGS		533	3	3662	8.3e-05
	BFGS	10^4	73	0	227	6.7e-05		BFGS	10^4	9962	1	39745	1.0e-04
	DBFGS		72	2	246	4.9e-05		DBFGS		185	3	704	8.9e-05
	CBFGS		69	2	233	8.3e-05		CBFGS		1299	2	10043	5.0e-05
	BFGS	-10^4	73	0	227	6.3e-05		BFGS	-10^4	9915	1	39557	1.0e-04
	DBFGS		72	2	242	4.3e-05		DBFGS		590	3	3123	8.5e-05
	CBFGS		69	2	232	5.6e-05		CBFGS		1268	2	9837	4.8e-05

TABLE 2
Test results for unconstrained optimization problems $B_0 = I$.

Method	Prob	Dim	Iter	Inner	Numf	Fnorm	Method	Prob	Dim	Iter	Inner	Numf	Fnorm
BFGS	rose	2	103	0	415	6.3e-005	BFGS	froth	2	-	-	-	-
DBFGS			-	-	-	-	DBFGS			-	-	-	-
CBFGS			668	7	6301	9.1e-05	CBFGS			282	7	3155	9.1e-06
BFGS	beale	2	347	0	1331	9.4e-05	BFGS	jensam	-	-	-	-	-
DBFGS			-	-	-	-	DBFGS			-	-	-	-
CBFGS			155	4	1330	2.6e-05	CBFGS			12	5	65	8.3e-05
BFGS	helix	3	279	0	1205	8.9e-05	BFGS	gulf	3	1	1	4	5.6e-086
DBFGS			-	-	-	-	DBFGS			1	1	4	1.9e-10
CBFGS			156	6	1413	3.1e-05	CBFGS			1	1	4	1.0e-10
BFGS	gauss	3	2	0	8	5.9e-006	BFGS	meyer	3	-	-	-	-
DBFGS			2	2	10	6.0e-06	DBFGS			1	4	14	4.2e-07
CBFGS			2	2	10	6.0e-06	CBFGS			1	4	14	4.2e-07
BFGS	sing	4	218	1	875	8.6e-05	BFGS	wood	4	-	-	-	-
DBFGS			214	9	1847	9.9e-05	DBFGS			-	-	-	-
CBFGS			97	6	650	9.7e-05	CBFGS			617	8	8971	6.9e-05
BFGS	kowosb	5	-	-	-	-	BFGS	biggs	6	59	0	211	4.4e-05
DBFGS			661	4	7031	1.0e-04	DBFGS			101	5	589	6.9e-05
CBFGS			661	4	7028	1.0e-04	CBFGS			101	5	589	6.9e-05
BFGS	osb2	11	225	1	775	4.0e-05	BFGS	watson	2	24	0	90	2.8e-06
DBFGS			-	-	-	-	DBFGS			18	5	124	1.1e-05
CBFGS			-	-	-	-	CBFGS			18	5	124	1.1e-05
BFGS	trid	10	152	0	609	1.8e-05	BFGS	singx	40	-	-	-	-
DBFGS			115	5	682	6.2e-05	DBFGS			-	-	-	-
CBFGS			115	5	682	6.2e-05	CBFGS			741	6	6372	1.0e-04
BFGS	pen1	10	248	0	1048	2.9e-05	BFGS	pen2	10	320	0	1499	5.0e-05
DBFGS			148	7	1235	4.5e-05	DBFGS			-	-	-	-
CBFGS			148	7	1235	4.5e-05	CBFGS			-	-	-	-
BFGS	bv	10	30	0	104	1.0e-05	BFGS	ie	10	5	0	17	2.6e-05
DBFGS			31	3	135	1.8e-05	DBFGS			4	2	18	2.6e-05
CBFGS			31	3	135	1.8e-05	CBFGS			4	2	18	2.6e-05
BFGS	lin	10	1	0	4	1.0e-13	BFGS	lin1	10	2	0	17	7.7e-06
DBFGS			1	1	4	8.9e-16	DBFGS			2	11	28	1.1e-10
CBFGS			1	1	4	8.9e-16	CBFGS			2	11	28	1.1e-10
BFGS	lin0	10	2	0	17	7.7e-07							
DBFGS			2	11	30	1.3e-11							
CBFGS			2	11	30	1.3e-11							

TABLE 3
Average performance for nonlinear equation problems.

Dim	Method	Iter	Inner	Numf	Dim	Method	Iter	Inner	Numf
10	BFGS	13.4	0	42	50	BFGS	46	0	123.6
	DBFGS	11.8	1.9	46.1		DBFGS	38	1.9	129.6
	CBFGS	11.8	1.9	46.1		CBFGS	38	1.9	129.6
100	BFGS	66.8	0	205.6	200	BFGS	5629.7	0.9	22446
	DBFGS	65.5	1.9	215.9		DBFGS	176.8	3.1	770.8
	CBFGS	65.2	1.9	214.1		CBFGS	409.5	2.6	2799.1

REFERENCES

[1] R. H. BYRD AND J. NOCEDAL, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.

[2] R. H. BYRD, J. NOCEDAL, AND Y.-X. YUAN, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171–1190.

[3] Y. DAI, *Convergence Properties of the BFGS Algorithm*, Technical report, The State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences, Beijing, China, 2001.

[4] J. E. DENNIS, JR., AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 1171–1190.

[5] J. E. DENNIS, JR., AND J. J. MORÉ, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[6] D. H. LI AND M. FUKUSHIMA, *A globally and superlinearly convergent Gauss–Newton-based BFGS method for symmetric nonlinear equations*, SIAM J. Numer. Anal., 37 (1999), pp. 152–172.

[7] D. H. Li and M. Fukushima, *A modified BFGS method and its global convergence in nonconvex minimization*, J. Comput. Appl. Math., 129 (2001), pp. 15–35.
[8] D. H. Li and M. Fukushima, *On the global convergence of the BFGS method for nonconvex unconstrained optimization problems*, SIAM J. Optim., 11 (2001), pp. 1054–1064.
[9] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[10] M. J. D. Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, SIAM-AMS Proceedings, Vol. IX, R. W. Cottle and C. E. Lemke, eds., AMS, Providence, RI, 1976, pp. 55–92.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1775–1797

NUMERICAL APPROXIMATION OF AN SQP-TYPE METHOD FOR PARAMETER IDENTIFICATION*

MARTIN BURGER† AND WOLFRAM MÜHLHUBER†

Abstract. This paper deals with the numerical approximation of the Levenberg–Marquardt SQP (LMSQP) method for parameter identification problems, which has been presented and analyzed in [M. Burger and W. Mühlhuber, *Inverse Problems*, 18 (2002), pp. 943–969]. It is shown that a Galerkin-type discretization leads to a convergent approximation and that the indefinite system arising from the Karush–Kuhn–Tucker (KKT) system is well-posed.

In addition, we present a multilevel version of the Levenberg–Marquardt method and discuss the simultaneous solution of the discretized KKT system by preconditioned iteration methods for indefinite problems. From a discussion of the numerical effort we conclude that these approaches may lead to a considerable speed-up with respect to standard iterative regularization methods that eliminate the underlying state equation. The numerical efficiency of the LMSQP method is confirmed by numerical examples.

Key words. parameter identification, sequential quadratic programming, iterative regularization, Galerkin methods, indefinite systems

AMS subject classifications. 65N21, 65N22, 65N30, 90C55

PII. S0036142901389980

1. Introduction.

Parameter identification denotes the procedure of determining unknown parameters appearing in an underlying state equation (usually a partial differential equation) from indirect measurements related to the solution of this equation. Such problems appear in many applications, where mathematical models of physical, chemical, biological, or economical processes are used (cf., e.g., [2, 13, 17] and the references therein).

Since such problems are *ill-posed* in general, i.e., the parameter to be reconstructed does not depend on the observation in a stable way, regularization methods have to be used in order to compute a stable approximation of the parameter in the presence of data noise. Due to the ill-posedness of the identification problem, the numerical approximation of such problems is not a simple task. The standard approach that can be found in literature is based on a priori elimination of the state equation, and an application of a discretized regularization method to the resulting operator equation involving the *parameter-to-output map*, which is the operator mapping the parameter to the corresponding observation. The main part in the evaluation of this map is the solution of the underlying state equation for a given parameter, which is numerically realized by standard discretizations such as finite elements.

The approach based on the parameter-to-output map, in particular combined with *iterative regularization methods* (cf. [18] for an overview), has been applied with success even to rather complicated parameter identification problems (cf., e.g., [10, 25, 26]). However, since these methods need a high number of direct solves (i.e., solutions of the state equation), fine discretizations of the parameter lead to a considerable compu-

*Received by the editors May 30, 2001; accepted for publication (in revised form) May 8, 2002; published electronically November 22, 2002. This work was supported by the Austrian National Science Foundation FWF under project grants F 13/08 and F 13/09.

http://www.siam.org/journals/sinum/40-5/38998.html

†SFB F 013 Numerical and Symbolic Scientific Computing, Johannes Kepler Universität, Freistädterstr. 313, A-4040 Linz, Austria (burger@indmath.uni-linz.ac.at, wmuehlhu@sfb013.uni-linz.ac.at).

tational effort, which results in high CPU times or even in the impossibility of using fine discretizations. Another drawback of this approach is that the discretizations of state and parameter are rather independent, which makes the numerical analysis extremely difficult. Therefore, fundamentally different methods for the solution of parameter identification problems have been investigated recently, whose common idea is to avoid a priori elimination of the state equation (cf. [11, 21, 27]). The aim of this paper is to discuss the numerical approximation of an iterative regularization method based on the idea of SQP (cf. [11]). We investigate Galerkin-type discretizations in the product space for parameter, state variable, and a corresponding Lagrangian variable, which leads to a sequence of well-posed indefinite systems. With this approach we are able to show convergence of the numerical approximation both for the quadratic programming problem arising in each iteration step and for the overall minimization procedure.

The general setup in this paper is as follows: we assume that we are given a noisy measurement z^δ satisfying

$$\|\hat{z} - z^\delta\|_Z \leq \delta, \tag{1.1}$$

where the exact data satisfy

$$\hat{z} := E\hat{u}, \tag{1.2}$$

with $E \in \mathcal{L}(X, Z)$ and $\hat{u} \in X$ solving $e(\hat{u}, \hat{q}) = f$ for some $\hat{q} \in Q_{ad} \subset Q$ (where Q_{ad} is a closed subset of Q with a nonempty interior). Our aim is to identify the parameter $q \in Q_{ad}$ in the underlying equation

$$e(u, q) = f, \tag{1.3}$$

where $e : X \times Q \to X^*$ is a continuous nonlinear operator with

$$e(0, 0) = 0. \tag{1.4}$$

In this setup X, X^*, Q, and Z are Hilbert spaces, and X^* can be identified with the dual of X. Finally, we assume that e is continuously Fréchet-differentiable on $X \times Q$ and that the partial derivative $e_u \in \mathcal{L}(X, X^*)$ is self-adjoint and satisfies the coercivity condition

$$\langle e_u(u, q)v, v\rangle \geq \alpha_e \|v\|_X^2 \qquad \forall (u, q, v) \in X \times Q_{ad} \times X \tag{1.5}$$

for some $\alpha_e \in \mathbb{R}^+$.

The above setup is typical for partial differential equations of elliptic type, which is also the main type of application we have in mind. We want to mention that the infinite-dimensional analysis carried out in the preceding paper [11] was not restricted to elliptic problems but only assumed the well-posedness of the state equation for a given parameter. However, since the numerical approximation techniques for elliptic problems differ from the ones for parabolic or hyperbolic problems (cf., e.g., [33] for an overview), one cannot expect a successful unified approach to corresponding parameter identification problems. For this reason we start with an investigation of the elliptic case in this paper, but we want to mention that the numerical identification of parameters in transient equations or even mixed systems of equations is an important and challenging problem for future research.

In [11], it has been mentioned that the parameter identification problem in the above setup is an *ill-posed inverse problem,* and we have proposed the following iterative regularization method based on the idea of SQP.

METHOD 1 (Levenberg–Marquardt SQP method). *Let*

$$(u_0, q_0) \in X \times Q$$

be a given initial value, and let $(\beta_k)_{k\in\mathbb{N}}$ *be a bounded sequence of positive real numbers. The* Levenberg–Marquardt SQP *(LMSQP) method consists of the iteration procedure*

$$(u_{k+1}, q_{k+1}) = (\overline{u}_k, \overline{q}_k), \tag{1.6}$$

where $(\overline{u}_k, \overline{q}_k)$ *is the minimizer of the quadratic programming problem*

$$\frac{1}{2}\|Eu - z^\delta\|_Z^2 + \frac{\beta_k}{2}\|q - q_k\|_Q^2 \to \min_{(u,q)\in X\times Q} \tag{1.7}$$

subject to the linear constraint

$$e(u_k, q_k) + e'(u_k, q_k)(u - u_k, q - q_k) = f. \tag{1.8}$$

The iteration procedure is stopped as soon as $k = k_*$*, where*

$$\|Eu_{k_*} - z^\delta\|_Z \le \tau\delta < \|Eu_k - z^\delta\| \qquad \forall\ k < k_*, \tag{1.9}$$

with appropriately chosen $\tau > 1$.

The motivation for the LMSQP method comes from two sources: a first guide for the design of this method are classical SQP methods, where one would minimize in each step a functional of the form

$$\frac{1}{2}\|Eu - z^\delta\|_Z^2 + \langle e''(u_k, q_k)(u - u_k, q - q_k)^2, \lambda_k\rangle \to \min_{(u,q)\in X\times Q}, \tag{1.10}$$

subject to the constraint (1.8), in order to obtain the new values u_{k+1}, q_{k+1} with corresponding Lagrangian variable λ_{k+1} (cf. [1, 14] and the references therein for further details on SQP methods). Since this approach results in a sequence of ill-posed quadratic minimization problems for typical cases in parameter identification (cf. [11]), it seems natural to add a regularizing term. This leads us to the second guide for the construction of the LMSQP method, namely the classical Levenberg–Marquardt method for nonlinear least-squares problems. For the solution of an unconstrained nonlinear least-squares problem of the form $\|F(q) - \hat{z}\| \to \min$ (which would arise, e.g., by a priori elimination of the state equation), the Levenberg–Marquardt method can be characterized via the sequence of minimization problems

$$\frac{1}{2}\|F(q_k) + F'(q_k)(q - q_k) - \hat{z}\|_Z^2 + \frac{\beta_k}{2}\|q - q_k\|_Q^2 \to \min_{q\in Q}. \tag{1.11}$$

The synthesis of Levenberg–Marquardt and SQP methods for ill-posed parameter identification problems is not unique; one possibility would be to minimize

$$\frac{1}{2}\|Eu - z^\delta\|_Z^2 + \langle e''(u_k, q_k)(u - u_k, q - q_k)^2, \lambda_k\rangle + \frac{\beta_k}{2}\|q - q_k\|_Q^2 \to \min_{(u,q)\in X\times Q}, \tag{1.12}$$

subject to (1.8), and another one would be the method given by (1.7), (1.8). Both of these methods have been introduced in [11], but since for least-squares systems the Lagrangian variable must be small close to a solution (note that we are interested in the case of attainable data here) we restrict our attention to the LMSQP method (1.7), (1.8).

Due to the results of [11], the LMSQP method is a convergent regularization method if the condition (1.9) is used as a stopping rule. We note that (1.9) is a standard stopping criterion for ill-posed inverse problems but a nonstandard one for SQP-type methods. However, the analysis in [11] and in the remainder of this paper shows that in this context it does not cause termination of the method at (largely) infeasible points. As a particular consequence of the results in [11], the quadratic programming problems of the form (1.7), (1.8), which have to be solved in each iteration step, are well-posed. Our aim in this paper is to investigate the numerical approximation of the LMSQP method by a Galerkin-type approach. We shall show below that this leads to an indefinite system in each iteration step, whose solution is an approximation of optimal order to the solution of (1.7), (1.8). Moreover, we show that the reconstructions obtained with the discretized LMSQP method converge to a solution of the parameter identification problem as the noise level and the discretization size tend to zero, if an appropriate stopping rule is used, which relates the residual to the noise level and some measures for the discretization.

Moreover, we shall discuss the solution of the discretized Karush–Kuhn–Tucker (KKT) system, which is an indefinite linear system to be solved for the discretized equivalents of state, parameter, and Lagrangian variable. The standard approaches to the solution of such discretized problems arising from partial differential equations are *reduced SQP methods*, where state and Lagrangian variable are eliminated a priori. We recall the basic properties of the reduced SQP approach, but we mainly focus on the iterative solution of the whole system with appropriate preconditioning. This promising approach has been employed recently for parameter identification (cf. [21, 27]) and optimal control problems (cf. [3, 4, 5, 6]) with good numerical results, in particular with respect to efficiency.

The paper is organized as follows: in section 2 we investigate the numerical approximation of the LMSQP method by a Galerkin-type approach and discuss the well-posedness, stability, and approximation properties of the discretized KKT system; the convergence of the discretized solutions is shown in section 3. Some further numerical methods and the implementation of the SQP iteration are examined in section 4. We briefly discuss the correct scaling of variables, the solution of the KKT system, and globalization strategies. Moreover, we present and analyze a multilevel approach, which leads to a further speed-up of the method. As a first application we investigate the identification of a potential in an elliptic boundary value problem, where we can give quantitative error estimates in terms of the discretization sizes. Some numerical experiments related to this identification problem are presented in section 6, before we conclude and give an outlook for further interesting problems related to this topic in section 7.

2. Discretization techniques. In the following we investigate the discretization of the LMSQP method by a Galerkin approach. First of all, we assume that we have discretized data $z^{\delta,\eta} \in Z_\eta \subset Z$ of the form

$$z^{\delta,\eta} = R_\eta z^\delta, \tag{2.1}$$

where $R_\eta : Z \to Z_\eta$ is the orthogonal projector onto the finite-dimensional subspace Z_η. Note that we can give an error estimate for $z^{\delta,\eta}$ using (1.1) and $\|R_\eta\| = 1$, which yields

$$\delta_\eta := \|R_\eta z^\delta - \hat{z}\|_Z \le \|R_\eta(z^\delta - \hat{z})\|_Z + \|R_\eta \hat{z} - \hat{z}\|_Z \le \delta + \inf_{y \in Z_\eta} \|y - \hat{z}\|_Z. \tag{2.2}$$

Now let $X_h \subset X$, $Q_h \subset Q$ be finite-dimensional subspaces of X and Q, with the corresponding orthogonal projectors $P_h : X \to X_h$ and $\tilde{P}_h : Q \to Q_h$. Then we can discretize the LMSQP method as follows.

METHOD 2 (Galerkin LMSQP method). *Let X_h, Q_h, and Z_η be as above, and let*

$$(u_0, q_0) \in X_h \times Q_h$$

be a given initial value. Moreover, let $(\beta_k)_{k\in\mathbb{N}}$ be a bounded sequence of positive real numbers. The Galerkin LMSQP *(GLMSQP) method consists of the iteration procedure*

$$(u_{k+1}, q_{k+1}) = (\overline{u}_k, \overline{q}_k), \tag{2.3}$$

where $(\overline{u}_k, \overline{q}_k) \in X_h \times Q_h$ is the minimizer of the quadratic programming problem

$$\frac{1}{2}\|R_\eta(Eu - z^\delta)\|_Z^2 + \frac{\beta_k}{2}\|q - q_k\|_Q^2 \to \min_{(u,q)\in X_h\times Q_h} \tag{2.4}$$

subject to the linear constraint

$$\langle e(u_k, q_k) + e'(u_k, q_k)(u - u_k, q - q_k), \varphi\rangle = \langle f, \varphi\rangle \qquad \forall\ \varphi \in X_h. \tag{2.5}$$

Note that the constraint (2.5) can be rewritten in operator form as

$$P_h^* K_k P_h(u - u_k) + P_h^* L_k \tilde{P}_h(q - q_k) = P_h^*(f - e(u_k, q_k)), \tag{2.6}$$

to be solved for $(u, q) \in X_h \times Q_h$, with the notation

$$K_k : X \to X^*, \qquad K_k u = e_u(u_k, q_k)u \quad \forall\ u \in X, \tag{2.7}$$

$$L_k : Q \to X^*, \qquad L_k q = e_q(u_k, q_k)q \quad \forall\ q \in Q, \tag{2.8}$$

and $P_h^* : X_h^* \to X^*$ is the adjoint of P_h. Under the assumption (1.5), we obtain that

$$\langle P_h^* K_k P_h v, v\rangle = \langle K_k P_h v, P_h v\rangle = \langle K_k v, v\rangle \ge \alpha_e \|v\|_X^2 \tag{2.9}$$

for all $v \in X_h$; i.e., the discrete bilinear form associated with the operator $P_h^* K_k P_h$ is coercive on X_h. This implies by the Lax–Milgram theorem that (2.6) is uniquely solvable with respect to u for given $q \in Q_h$. Consequently, in an analogous way to the proof of Proposition 2.1 in [11] we may show the following result on the well-posedness of the quadratic programming problem that has to be solved in each step of Method 2.

PROPOSITION 2.1. *Let e be continuously Fréchet-differentiable, let* (1.5) *hold, and let $\beta_k > 0$. Then the quadratic programming problem* (2.4), (2.5) *has a unique solution $(\overline{u}_k, \overline{q}_k) \in X_h \times Q_h$, which is also the only local minimum.*

2.1. The discretized KKT system. In [11], the KKT system for the infinite-dimensional version of the LMSQP method has been derived and analyzed in the framework of linear saddle-point problems. Now we will discuss the discretized analogue of this system, namely the first-order optimality conditions for the quadratic programming problem (2.4), (2.5).

The Lagrangian of (2.4), (2.5) is given by

$$\begin{aligned}\mathcal{L}_k(u,q;\lambda) &= \frac{1}{2}\|R_\eta(Eu - z^\delta)\|_Z^2 + \frac{\beta_k}{2}\|q - q_k\|_Q^2 \\ &\quad + \langle \lambda, e'(u_k,q_k)(u-u_k, q-q_k) + e(u_k,q_k) - f\rangle \end{aligned} \tag{2.10}$$

for $(u,q,\lambda) \in X_h \times Q_h \times X_h$. Since P_h and $\tilde{P}_h$ are equal to the identity on X_h and Q_h, respectively, we can rewrite the Lagrangian as

$$\begin{aligned}\mathcal{L}_k(u,q;\lambda) &= \frac{1}{2}\|R_\eta(EP_hu - z^\delta)\|_Z^2 + \frac{\beta_k}{2}\|\tilde{P}_h(q - q_k)\|_Q^2 \\ &\quad + \langle P_h\lambda, K_kP_h(u-u_k) + L_k\tilde{P}_h(q-q_k) + e(u_k,q_k) - f\rangle, \end{aligned} \tag{2.11}$$

with the operators K_k and L_k defined by (2.7), (2.8). The KKT system can now be deduced by computing the partial derivatives of the Lagrangian with respect to u, q, and λ; i.e., $(u_{k+1} - u_k, q_{k+1} - q_k, \lambda_{k+1})$ solves the linear saddle-point problem

$$\begin{pmatrix} P_h^*E^*R_\eta^*R_\eta EP_h & 0 & P_h^*K_k^*P_h \\ 0 & \beta_k\tilde{P}_h^*\tilde{P}_h & \tilde{P}_h^*L_k^*P_h \\ P_h^*K_kP_h & P_h^*L_k\tilde{P}_h & 0 \end{pmatrix} \begin{pmatrix} u \\ q \\ \lambda \end{pmatrix} = \begin{pmatrix} P_h^*E^*R_\eta^*R_\eta(z^\delta - Eu_k) \\ 0 \\ P_h^*(f - e(u_k,q_k)) \end{pmatrix}. \tag{2.12}$$

As in [11], we define the symmetric bilinear form $a_k : (X \times Q)^2 \to \mathbb{R}$ by

$$a_k^\eta(u,q;\varphi,\sigma) := \langle R_\eta Eu, R_\eta E\varphi\rangle_Z + \beta_k\langle q,\sigma\rangle_Q \tag{2.13}$$

and the bilinear form $b_k : (X \times Q) \times X \to \mathbb{R}$ by

$$b_k(u,q;\lambda) := \langle K_ku,\lambda\rangle + \langle L_ku,\lambda\rangle. \tag{2.14}$$

Moreover, we use the right-hand sides

$$f_k := f - e(u_k,q_k) \in X^*, \tag{2.15}$$

$$g_k^\eta := (E^*R_\eta^*R_\eta(z^\delta - Eu_k), 0) \in X^* \times Q. \tag{2.16}$$

Then the KKT system (2.12) can be interpreted as the Galerkin approximation of an indefinite variational problem; i.e., $(u,q,\lambda) \in X_h \times Q_h \times X_h$ is the solution of

$$a_k^\eta(u,q;\varphi,\sigma) + b_k(\varphi,\sigma;\lambda) = \langle g_k^\eta, (\varphi,\sigma)\rangle \qquad \forall\ (\varphi,\sigma) \in X_h \times Q_h, \tag{2.17}$$

$$b_k(u,q;\mu) = \langle f_k,\mu\rangle \qquad \forall\ \mu \in X_h. \tag{2.18}$$

In an analogous way to the proof of Theorem 2.3 in [11] we can show that the bilinear form a satisfies the kernel-ellipticity condition on $X_h \times Q_h$; i.e., there exists a constant $\alpha_a > 0$ such that

$$a_k^\eta(u,q;u,q) \geq \alpha_a\|(u,q)\|^2$$
$$\forall\ (u,q) \in \mathcal{K}_b^h := \{\ (v,s) \in X_h \times Q_h \mid b(v,s;\lambda) = 0,\ \forall\ \lambda \in X_h\ \}$$

and that b satisfies the LBB condition

$$\inf_{\lambda\in X_h}\sup_{(u,q)\in X_h\times Q_h}\frac{b_k(u,q;\lambda)}{\|(u,q)\|\,\|\lambda\|}\geq\alpha_b$$

for some $\alpha_b>0$. This implies the following well-posedness result (cf. [8, 9]) for the discretized problem (2.17), (2.18).

THEOREM 2.2. *Let e be continuously Fréchet-differentiable, let* (1.5) *hold, and let $\beta_k>0$. Then the indefinite system* (2.17), (2.18) *has a unique solution $(u,q,\lambda)\in X_h\times Q_h\times X_h$, which depends continuously on the right-hand sides f_k and g_k^η.*

Since the constants α_a and α_b are the same as in the corresponding infinite-dimensional conditions in $X\times Q$, they are, in particular, independent of the discrete subspaces X_h and Q_h. This allows us to deduce an approximation result for the solutions of (2.17), (2.18) to the solution $(u,q,\lambda)\in X\times Q\times X$ of the infinite-dimensional KKT system, given in variational form as

$$a_k(u,q;\varphi,\sigma)+b_k(\varphi,\sigma;\lambda)=\langle g_k,(\varphi,\sigma)\rangle\qquad\forall\,(\varphi,\sigma)\in X\times Q, \tag{2.19}$$

$$b_k(u,q;\mu)=\langle f_k,\mu\rangle\qquad\forall\,\mu\in X, \tag{2.20}$$

with a_k given by

$$a_k(u,q;\varphi,\sigma):=\langle Eu,E\varphi\rangle_Z+\beta_k\langle q,\sigma\rangle_Q, \tag{2.21}$$

b_k, f_k as above, and g_k defined by

$$g_k:=(E^*(z^\delta-Eu_k),0)\in X^*\times Q. \tag{2.22}$$

THEOREM 2.3. *Suppose that the assumptions of Theorem* 2.2 *are satisfied, and let*

$$(u_h,q_h,\lambda_h)\in X_h\times Q_h\times X_h$$

denote the unique solution of (2.17), (2.18). *Then there exists a constant $c>0$ independent of X_h and Q_h such that*

$$\|(u-u_h,q-q_h,\lambda-\lambda_h)\|\leq c\left(r_{\eta,h}^\delta+\inf_{(v,s,\mu)\in X_h\times Q_h\times X_h}\|(u-v,q-s,\lambda-\mu)\|\right), \tag{2.23}$$

where (u,q,λ) denotes the unique solution of (2.19), (2.20) *and*

$$r_{\eta,h}^\delta:=\|(R_\eta-I)z^\delta\|_Z+\sup_{v\in X_h,\|v\|=1}\|(R_\eta-I)Ev\|_Z. \tag{2.24}$$

Proof. First, let $(\tilde u_h,\tilde q_h,\tilde\lambda_h)$ denote the solution of (2.17), (2.18) with a_k^η, g_k^η replaced by a_k, g_k. Then Theorem 2.1 in [9] implies the existence of a constant $c_1>0$ (independent of X_h and Q_h) such that

$$\|(u-\tilde u_h,q-\tilde q_h,\lambda-\tilde\lambda_h)\|\leq c_1\inf_{(v,s,\mu)\in X_h\times Q_h\times X_h}\|(u-v,q-s,\lambda-\mu)\|.$$

Moreover, the stable dependence of the solutions of (2.17), (2.18) on the right-hand side implies the existence of $c_2 > 0$, $c_3 > 0$ with

$$\begin{aligned}
&\|(u_h - \tilde{u}_h, q_h - \tilde{q}_h, \lambda_h - \tilde{\lambda}_h)\| \\
&\quad \le c_2 \left(\sup_{v \in X_h, \|v\|=1} \langle g_k^\eta - g_k, (v,0) \rangle + \sup_{\varphi \in X_h, \|\varphi\|=1} |a_k^\eta(\tilde{u}_h, \tilde{q}_h, \varphi) - a_k(\tilde{u}_h, \tilde{q}_h, \varphi)| \right) \\
&\quad \le c_2 \left(\sup_{v \in X_h, \|v\|=1} \langle Ev, (R_\eta^* R_\eta - I)(z^\delta - Eu_k) \rangle + \sup_{\varphi \in X_h, \|\varphi\|=1} \langle E\varphi, (R_\eta^* R_\eta - I) E\tilde{u}_h \rangle \right) \\
&\quad \le c_3 \left(\|E\| \, \|(R_\eta - I) z^\delta\|_Z + \sup_{v \in X_h, \|v\|=1} \|(R_\eta - I) Ev\|_Z \right),
\end{aligned}$$

and with the triangle inequality we may conclude (2.23). ∎

Theorem 2.3 provides an error estimate for the solutions of the discretized saddle-point problem (2.17), (2.18), consisting of two parts corresponding to the numerical approximation in the image space Z and in the preimage spaces X and Q. An obvious estimate for the first term is

$$r_{\eta,h}^\delta \le \inf_{y \in Z_\eta} \|y - z^\delta\|_Z + \sup_{v \in X_h, \|v\|_X = 1} \inf_{\tilde{y} \in Z_\eta} \|\tilde{y} - Ev\|_Z,$$

which possibly does not lead to a quantitative estimate, since there is no additional information on the smoothness of the noisy data. An alternative estimate is

$$r_{\eta,h}^\delta \le \delta + \inf_{y \in Z_\eta} \|y - \hat{z}\|_Z + \sup_{v \in X_h, \|v\|_X = 1} \inf_{\tilde{y} \in Z_\eta} \|\tilde{y} - Ev\|_Z.$$

The infimum of $\|y - \hat{z}\|_Z$ can usually be estimated more easily, since the exact data $\hat{z}$ are smoother due to the fact that $\hat{u}$ is the solution of the state equation for some parameter $\hat{q}$. For example, if the state equation is of elliptic type with solution $\hat{u} \in H^1(\Omega)$, $E : H^1(\Omega) \to L^2(\Omega)$ is the embedding operator, and R_η results from a standard finite element discretization on a grid with fineness η, then we have at least

$$\inf_{y \in Z_\eta} \|y - \hat{z}\| = \mathcal{O}(\eta).$$

Another important observation is that the last term vanishes if the discrete spaces Z_η and X_h are equal, which can be achieved in some applications.

The second term in (2.23) shows that the Galerkin approximation of the KKT system is of optimal order in $X_h \times Q_h \times X_h$; it can be estimated by standard methods for finite element discretizations; quantitative estimates can be obtained using the regularity of the iterates. This part depends, of course, strongly on the specific application.

3. Convergence analysis. In this section we will analyze the GLMSQP method with respect to convergence, i.e., the convergence of the reconstruction obtained with an appropriate stopping rule as the noise level and the measure for the discretization fineness tend to zero. With $\eta = 0$, $h = 0$ we will identify the infinite-dimensional case, i.e., $X_0 = X$, $Q_0 = Q$, and $Z_0 = Z$. We assume that the discrete subspaces satisfy

$$\overline{\bigcup_{h>0} X_h} = X, \qquad \overline{\bigcup_{\eta>0} Z_\eta} = Z, \qquad \overline{\bigcup_{h>0} Q_h} = Q.$$

If we denote by d_k and f_k the error terms (note that $e(\hat{u},\hat{q}) = f$)

$$(d_k, f_k) := (P_h(u_k - \hat{u}), \tilde{P}_h(q_k - \hat{q})), \tag{3.1}$$

we can rewrite the KKT system (2.12) as

$$\begin{pmatrix} P_h^* E^* R_\eta^* R_\eta E P_h & 0 & P_h^* K_k^* P_h \\ 0 & \beta_k \tilde{P}_h^* \tilde{P}_h & \tilde{P}_h^* L_k^* P_h \\ P_h^* K_k P_h & P_h^* L_k \tilde{P}_h & 0 \end{pmatrix} \begin{pmatrix} d_{k+1} \\ f_{k+1} \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} P_h^* E^* R_\eta^* R_\eta (z^\delta - E P_h \hat{u}) \\ \beta_k \tilde{P}_h^* \tilde{P}_h (q_k - \hat{q}) \\ r_k \end{pmatrix}, \tag{3.2}$$

where the r_k denotes the remainder

$$r_k := P_h^* \left(e(\hat{u},\hat{q}) - e(u_k, q_k) + e'(u_k, q_k)(P_h d_k, \tilde{P}_h f_k) \right). \tag{3.3}$$

As in [11], we require a condition on the nonlinearity, which is summarized in the following.

ASSUMPTION 1. *Let* (1.5) *be satisfied and define the remainder* $r(u,q)$ *by*

$$r(u,q) := e(\hat{u},\hat{q}) - e(u,q) - e'(u,q)(\hat{u} - u, \hat{q} - q). \tag{3.4}$$

Then we assume that there exists a constant $\gamma_1 < 1$ *such that*

$$\|E e_u(u,q)^{-1} r(u,q)\|_Z \le \gamma_1 \|Eu - \hat{z}\|_Z \qquad \forall\ (u,q) \in B_{2\zeta}(u_0) \times B_{2\rho}(q_0) \tag{3.5}$$

and that there exists a solution $(\hat{u},\hat{q}) \in B_\zeta(u_0) \times B_\rho(q_0)$ *of the parameter identification problem.*

For a discussion of the nonlinearity condition (3.5) and a comparison with standard conditions used in the convergence analysis of nonlinear ill-posed problems we refer the reader to [11].

If we define the discretization measures ϵ_h, κ_h by

$$\epsilon_h = \|E(I - P_h)\hat{u}\|_Z, \qquad \kappa_h = c_{\zeta,\rho} \|(I - \tilde{P}_h)\hat{q}\|_Q, \tag{3.6}$$

where

$$c_{\zeta,\rho} = \sup_{(u,q) \in B_{2\zeta}(u_0) \times B_{2\rho}(q_0)} \|E e_u(u,q)^{-1} e_q(u,q)\|_Z, \tag{3.7}$$

and ϵ_η by

$$\epsilon_\eta = \gamma_1 \zeta^{-1} \sup_{\|v\|_X = 1} \|(R_\eta - I)Ev\|_Z, \tag{3.8}$$

then for all $(u,q) \in B_{2\zeta}(u_0) \times B_{2\rho}(q_0)$ the estimate

$$\|R_\eta E e_u(u,q)^{-1} r_h(u,q)\|_Z \le \gamma_1 \|R_\eta(Eu - \hat{z})\|_Z + \epsilon_\eta + \epsilon_h + \kappa_h \tag{3.9}$$

holds, where

$$r_h(u,q) := e(\hat{u},\hat{q}) - e(u,q) - e'(u,q)(P_h\hat{u} - u, \tilde{P}_h\hat{q} - q). \tag{3.10}$$

Remark 1. If X_h, Z_η, and Q_h are standard finite element spaces on some triangulations, then ϵ_h, ϵ_η, and κ_h can be estimated by the approximation error of these elements. In particular, if the discretization parameter (i.e., the maximal size of a triangle) tends to zero and if the triangulation is regular, one can guarantee that ϵ_h, ϵ_η, and κ_h tend to zero (cf. [33] for further details).

For the choice of the stopping index we use a numerical version of (1.9), which involves the discretization measures defined above:

$$\|R_\eta(Eu_{k_*} - z^\delta)\|_Z \le \tau(\delta_\eta + 2\epsilon_h + \kappa_h) < \|R_\eta(Eu_k - z^\delta)\|_Z \qquad \forall\ k < k_*. \tag{3.11}$$

For an appropriate choice of τ, this allows us to prove the following monotonicity property of the iterates.

LEMMA 3.1. *Let Assumption* 1 *be fulfilled, let the noise be bounded by* (1.1)*, and assume that*

$$\beta_0^{-1}(\|Ed_0\|_Z - \delta - \epsilon_h)^2 + \|f_0\|_Q^2 \le \rho^2. \tag{3.12}$$

In addition, β_k is chosen such that $\beta_k \le \beta_{k-1}$ for all $k \in \mathbb{N}$ and that

$$\overline{\gamma}_1 := \gamma_1 \sup_{k\in\mathbb{N}} \sqrt{\frac{\beta_{k-1}}{\beta_k}} < 1, \tag{3.13}$$

and the stopping index k_* is chosen according to the generalized discrepancy principle (3.11) *with*

$$\tau > 1 + \frac{\gamma_1 + \overline{\gamma}_1}{\gamma_1(1 - \overline{\gamma}_1)}; \tag{3.14}$$

then $q_k \in B_{2\rho}(q_0)$ and the estimates

$$\begin{aligned}(\|R_\eta Ed_{k+1}\|_Z - \delta - \epsilon_h)^2 + \beta_k\|f_{k+1}\|_Q^2 + \beta_k\|q_{k+1} - q_k\|_Q^2 \\ \le (\gamma_1\|R_\eta Ed_k\|_Z + \delta + \epsilon_\eta + 2\epsilon_h + \kappa_h)^2 + \beta_k\|f_k\|_Q^2\end{aligned} \tag{3.15}$$

and

$$\beta_k^{-1}(\|R_\eta Ed_{k+1}\|_Z - \delta - \epsilon_h)^2 + \|f_{k+1}\|_Q^2 \le \beta_{k-1}^{-1}(\|R_\eta Ed_k\|_Z - \delta - \epsilon_h)^2 + \|f_k\|_Q^2 \tag{3.16}$$

hold for all $k < k_*$.

Proof. Assume that $q_k \in B_{2\rho}(q_0)$. Then, with (3.2) and

$$\lambda_{k+1} = -(P_h^* K_k^* P_h)^{-1} P_h^* E^* R_\eta^* R_\eta(Eu_{k+1} - z^\delta)$$

we deduce the identity

$$\begin{aligned}2\|R_\eta Ed_{k+1}\|_Z^2 &+ \beta_k\|f_{k+1}\|_Q^2 + \beta_k\|q_{k+1} - q_k\|_Q^2 \\ &= 2\|R_\eta Ed_{k+1}\|_Z^2 + \beta_k\|f_k\|_Q^2 + 2\beta_k\langle f_{k+1}, q_{k+1} - q_k\rangle \\ &= 2\langle R_\eta(z^\delta - EP_h\hat{u}), R_\eta Ed_{k+1}\rangle_Z + \beta_k\|f_k\|_Q^2 \\ &\quad + 2\langle R_\eta(Eu_{k+1} - z^\delta), R_\eta EP_h(P_h^* K_k P_h)^{-1} P_h^* r_h(u_k, q_k)\rangle_Z.\end{aligned}$$

The noise bound (1.1) implies that

$$\|R_\eta(z^\delta - P_h\hat{u})\|_Z \le \|R_\eta(z^\delta - \hat{z})\|_Z + \|E(I - P_h)\hat{u}\|_Z \le \delta + \epsilon_h,$$

and using the Cauchy–Schwarz inequality together with (3.9) we obtain the estimate

$$(\|R_\eta E d_{k+1}\|_Z - \delta - \epsilon_h)^2 + \beta_k \|f_{k+1}\|_Q^2 + \beta_k \|q_{k+1} - q_k\|_Q^2 \\ \leq (\gamma_1 \|R_\eta E d_k\|_Z + \delta + \epsilon_\eta + 2\epsilon_h + \kappa_h)^2 + \beta_k \|f_k\|_Q^2.$$

Equation (3.16) follows from dividing (3.15) by β_k and the fact that

$$\sqrt{\frac{\beta_{k-1}}{\beta_k}}(\gamma_1 \|R_\eta E d_k\|_Z + \delta + \epsilon_\eta + 2\epsilon_h + \kappa_\eta) \leq \|R_\eta E d_k\|_Z - \delta - \epsilon_h.$$

By induction we can now show that $q_k \in B_\rho(q_0)$ for $k < k_*$ and τ satisfying (3.14). ∎

In an analogous way to the proof of Lemma 3.2 in [11] we can prove the following statement on the finiteness of the stopping index k_* if $\delta > 0$.

LEMMA 3.2. *Under the assumptions of Lemma* 3.1, *the discrepancy principle* (3.11) *yields a finite stopping index* k_* *if*

$$\delta_{\eta,h} := \delta + \epsilon_\eta + 2\epsilon_h + \kappa_h > 0, \tag{3.17}$$

and τ *is chosen according to* (3.14).

One observes that in the above estimates the term $\delta_{\eta,h}$ now plays the same role as the noise level δ in the infinite-dimensional setup. Therefore it is also possible to prove convergence as $\delta_{\eta,h} \to 0$ in the same way as convergence in the infinite-dimensional case for $\delta \to 0$ (cf. [11, Theorem 3.5]). Consequently, we do not give the detailed convergence proof but refer to [11] for further details on the technique of the proof. We recall only the basic assumptions on e and give the final convergence result, where we use the notation $(u_k^{\delta,\eta,h}, q_k^{\delta,\eta,h})$ for the iteration according to (2.12) with initial value $(P_h u_0, \tilde{P}_h q_0)$, noise level δ, and discretization parameters h and η.

ASSUMPTION 2. *In addition to Assumption* 1, *assume that* e *is of the form*

$$e(u,q) = A(u) + N(u,q) \qquad \forall\ (u,q) \in X \times Q, \tag{3.18}$$

with continuously Fréchet-differentiable (nonlinear) operators $A : X \to X^*$ *and* $N : X \times Q \to X^*$ *such that*

$$N(u,.) \in \mathcal{L}(Q, X^*) \qquad \forall\ u \in X. \tag{3.19}$$

Moreover, we assume that A *and* N *satisfy the nonlinearity conditions*

$$\|E e_u(u,q)^{-1} A'(v) w\|_{X^*} \leq \gamma_2 \|Ew\|_{X^*} \quad \forall\ (u,v,w,q) \in B_{2\zeta}(u_0)^2 \times X \times B_{2\rho}(q_0) \tag{3.20}$$

and

$$\|E e_u(u,q)^{-1} N_u(v,s) w\|_Y \leq \gamma_3 \|Ew\|_Y \quad \forall\ (u,v,w,q,s) \in B_{2\zeta}(u_0)^2 \times X \times B_{2\rho}(q_0)^2 \tag{3.21}$$

for some positive constants γ_2 *and* γ_3.

THEOREM 3.3 (convergence). *Let Assumption* 2 *and* (3.12) *be fulfilled with* ζ, ρ *sufficiently small, and let the noise be bounded by* (1.1). *Moreover, let* β_k *be chosen such that* $\beta_k \leq \beta_0$ *for all* $k \in \mathbb{N}$ *and that* (3.13) *is satisfied. If the perturbed iteration*

is stopped with $k_* = k_*(\delta, R_\eta z^\delta, h)$ *according to the generalized discrepancy principle* (3.11) *with* $\tau = \tau(h,\eta)$ *(uniformly bounded in* h *and* η*) satisfying* (3.14)*, then*

$$(q^{\delta,\eta,h}_{k_*(\delta,R_\eta z^\delta,h)}, u^{\delta,\eta,h}_{k_*(\delta,R_\eta z^\delta,h)}) \to (\overline{q},\overline{u}) \qquad \text{in } X\times Q \quad \text{as } \max\{\delta_\eta,\epsilon_h,\kappa_h\}\to 0, \tag{3.22}$$

where $(\overline{u},\overline{q})$ *is a solution of* (1.3) *with* $E\overline{u} = \hat{z}$*.*

Proof. The proof is analogous to the proof of Theorem 3.5 in [11]. ∎

4. Numerical realization. In the following we want to discuss some numerical methods and variants for the GLMSQP algorithm. We split this discussion into two parts, the first related to the *inner iteration*, i.e., the numerical solution of the discretized KKT system (2.12) for fixed iteration number k, and the second related to the *outer iteration*, i.e., the iteration in k defined by the LMSQP method.

4.1. The inner iteration. Using appropriate bases for all finite-dimensional subspaces, the discretized KKT system (with penalty parameter $\beta = \beta_k$) can be written as

$$\begin{pmatrix} G & 0 & K^T \\ 0 & \beta H & L^T \\ K & L & 0 \end{pmatrix} \begin{pmatrix} V \\ S \\ \Lambda \end{pmatrix} = \begin{pmatrix} f_1 \\ 0 \\ f_3 \end{pmatrix}, \tag{4.1}$$

to be solved for $V \in \mathbb{R}^m$, $S \in \mathbb{R}^n$, and $\Lambda \in \mathbb{R}^m$. One notices that due to the properties of the corresponding infinite-dimensional operators, the matrices $G \in \mathbb{R}^{m\times m}$ and $H \in \mathbb{R}^{n\times n}$ are symmetric and positive definite, $K \in \mathbb{R}^{m\times m}$ is regular, and $L \in \mathbb{R}^{m\times n}$ is of rather general form. An alternative formulation of (4.1) is

$$MX = F, \tag{4.2}$$

with

$$M = \begin{pmatrix} G & 0 & K^T \\ 0 & \beta H & L^T \\ K & L & 0 \end{pmatrix}, \qquad X = \begin{pmatrix} V \\ S \\ \Lambda \end{pmatrix}, \qquad F = \begin{pmatrix} f_1 \\ 0 \\ f_3 \end{pmatrix}.$$

From the analysis in section 2 we may conclude that M is a regular, indefinite matrix, whose further properties (such as distribution of eigenvalues) rely on the specific form of the state equation and the objective. For a first application, we will investigate these properties in section 5.

For the solution of the discretized KKT system (4.2), there are two basic possibilities. The first one is the so-called *reduced SQP* approach, which consists of eliminating the state and Lagrangian variable (using the regularity of K) and then solving the arising lower-dimensional system for the parameter, which has a symmetric positive definite system matrix $M_r \in \mathbb{R}^{n\times n}$. This approach is frequently used in optimal control and parameter identification (cf., e.g., [36, 37, 38]) and seems attractive, since the problem dimension is reduced significantly. However, the reduced matrix is of the same structure as the matrix arising from a discretization of a Newton method following the feasible path (in particular, it involves the inverses of K and K^T), and therefore one has to expect that the numerical effort is of the same order as for such well-known methods. Recently, the simultaneous solution of KKT systems by iterative methods has been investigated, in particular in connection with optimal control problems (cf. [3, 5, 6, 21]). Compared to the reduced SQP approach, a simultaneous solution strategy has the obvious advantage that the allocation and evaluation of the

system matrix M is much cheaper than of M_r. Also matrix-vector products with M_r are by far more effort than matrix-vector products with M.

At first glance, it seems rather straightforward to solve (4.2) by a standard iterative method for indefinite systems such as inexact Uzawa methods (cf. [7, 16, 40]) or Krylov-subspace correction methods such as GMRES (cf. [35]), MINRES (cf. [32]), and QMR (cf. [19]). However, in the case of large-scale problems, we have to expect a large condition number (note that β is usually small and that M is singular for $\beta = 0$) and a complicated eigenvalue pattern of the matrix M, which might cause iterative methods to diverge or to need a high number of iterations. Therefore, an appropriate preconditioning technique seems necessary for any of the methods. We do not go into detail here but refer to the forthcoming paper [12] for a discussion and comparison of different preconditioners.

4.2. The outer iteration. As usual for nonlinear optimization we have to take care of the following two aspects for the discretized LMSQP method:

- *Scaling* of the state variable, parameter, and Lagrangian variable is needed in order to ensure that all variables and all sensitivities are of the same magnitude. In addition, appropriate scaling is needed for balancing the set of constraints and the objective. Since this topic is of high practical importance for any optimization problem and well investigated, we refer to monographs on nonlinear optimization for a detailed discussion (cf., e.g., [20, 31]).
- *Globalization strategies* are important for any locally convergent optimization method such as Newton-type or SQP-type methods. The two most popular classes of globalization techniques in optimization are *trust region methods* and *line search strategies*, which can both be applied for a globalization of the LMSQP algorithm. For a comprehensive overview of trust region methods we refer to Conn, Gould, and Toint [14], and for details on line search strategies we refer the reader to Nocedal and Wright [31].

Important tools for the efficient numerical approximation of infinite-dimensional optimization problems are *multilevel optimization methods*. In the nested multilevel setup, one starts the optimization procedure at a coarse level $X_{h_1} \times Q_{h_1}$, where the iteration procedure can be carried out efficiently. If an appropriate stopping rule is satisfied, one interpolates the state and parameter obtained in this way to a finer level $X_{h_2} \times Q_{h_2}$ (for $h_2 < h_1$), serving now as a starting value on this level. This procedure is repeated until the finest level is reached. Usually, nested spaces are used in this approach, i.e., $X_{h_1} \subset X_{h_2}$, $Q_{h_1} \subset Q_{h_2}$ (for $h_2 < h_1$), which leads to simple interpolation operators. Since one cannot choose the discretization of the data arbitrarily, in general, we consider only the case of fixed η here, but a multilevel approach in η can be realized in an analogous way, if necessary.

Nested multilevel methods outperform standard discretization techniques in many cases (cf., e.g., [22, 23, 30]); usually a considerable number of iterations is needed on the coarse level only, where the numerical effort per iteration is very low. On the finest levels, the stopping rule is often satisfied already after one iteration step and so the overall effort is less than for a direct discretization on the finest level. For the GLMSQP method, we can formulate a multilevel algorithm as follows.

ALGORITHM 4.1 (nested multilevel GLMSQP).

Given a decreasing sequence $\{h_\ell\}_{\ell=1,\dots,L}$ with nested spaces $X_{h_\ell} \subset X_{h_{\ell+1}}$, $Q_{h_\ell} \subset Q_{h_{\ell+1}}$ (e.g., $h_\ell = 2^{-\ell} h_0$), and a nonincreasing sequence τ_ℓ satisfying (3.14)*, the nested multilevel GLMSQP method consists of the following iterative procedure:*

1. *Set $\ell = 1$, $h = h_1$ and start with $(u_0^1, q_0^1) \in X_{h_1} \times Q_{h_1}$.*

2. *Perform the GLMSQP method until the stopping criterion* (3.11) *is satisfied with stopping index* $k_*(\ell)$.
3. *If* $\ell = L$ *stop the iteration, else prolongate the iterate* $(u^\ell_{k_*}, q^\ell_{k_*})$ *to the finer level* $X_{h_{\ell+1}} \times Q_{h_{\ell+1}}$, *which results in a new starting value* $(u_0^{\ell+1}, q_0^{\ell+1})$. *Set* $h = h_{\ell+1}$, $\ell = \ell + 1$, *and go to step* 2.

The analysis in section 3 shows that for $\beta_0^\ell \geq \beta^{\ell-1}_{k_*(\ell-1)}$ the estimate

$$\begin{aligned}(\beta^\ell_{k_*(\ell)})^{-1}\|R_\eta E d^\ell_{k_*(\ell)}\|_Z^2 + \|f^\ell_{k_*(\ell)}\|_Q^2 + \sum_{j=0}^{k_*(\ell)-1} \|q^\ell_{j+1} - q^\ell_j\|_Q^2 \\ \leq (\beta_0^\ell)^{-1}\|R_\eta E d_0^\ell\|_Z^2 + \|f_0^\ell\|_Q^2 \\ \leq (\beta^{\ell-1}_{k_*(\ell-1)})^{-1}\|R_\eta E d^{\ell-1}_{k_*(\ell-1)}\|_Z^2 + \|f^{\ell-1}_{k_*(\ell-1)}\|_Q^2 + \theta_\ell\end{aligned}$$

holds, where θ_ℓ is the error corresponding to the interpolation of the iterates from level $\ell - 1$ to level ℓ, i.e.,

$$\theta_\ell = (\beta_0^\ell)^{-1}\left(\|R_\eta E d_0^\ell\|_Z^2 - \|R_\eta E d^{\ell-1}_{k_*(\ell-1)}\|_Z^2\right) + \left(\|f_0^\ell\|_Q^2 - \|f^{\ell-1}_{k_*(\ell-1)}\|_Q^2\right). \tag{4.3}$$

This monotonicity estimate corresponds very well to the intuition that only few iterations are needed on the fine levels, in particular if β_k^ℓ is decreasing, which leads to

$$\|R_\eta E d^\ell_{k_*(\ell)}\|_Z^2 \leq \beta^\ell_{k_*(\ell)}\left((\beta_0^\ell)^{-1}\tau_{\ell-1}\delta_{\eta,h_{\ell-1}} + \|f^{\ell-1}_{k_*(\ell-1)}\|_Q^2 + \theta_\ell\right).$$

For a fine level with small β, we can expect that

$$\beta^\ell_{k_*(\ell)}(\beta_0^\ell)^{-1}\tau_{\ell-1}\delta_{\eta,h_{\ell-1}} \approx \tau_\ell \delta_{\eta,h_\ell},$$

and the second term $\beta^\ell_{k_*(\ell)}(\|f^{\ell-1}_{k_*(\ell-1)}\|_Q^2 + \theta_\ell)$ can be expected to be negligible. In other words, the stopping rule at level ℓ is probably satisfied with $k_*(\ell) = 1$.

Under typical conditions, where X_{h_ℓ} and Q_{h_ℓ} correspond to standard finite element spaces on different refinement levels of an initial triangulation of a domain Ω, one can show that at least $\theta_\ell = \mathcal{O}(h_{\ell-1})$, and consequently

$$\sum_{\ell=2}^{L} \theta_\ell \leq c h_1 \left(1 + \sum_{j=0}^{L-2} r^j\right) \leq c h_1 \frac{2}{1-r}$$

for some constant $c \in \mathbb{R}_+$, where

$$r = \max_{1\leq \ell \leq L-1} \frac{h_{\ell+1}}{h_\ell} < 1.$$

Together with the above estimate one can show with a standard proof technique that the pair $(u^L_{k_*(L)}, q^L_{k_*(L)})$ converges to a solution $(\overline{u}, \overline{q})$ of the parameter identification problem for $\delta_{\eta,h_L} \to 0$.

5. Application to potential reconstruction. As a first application we investigate the identification of the potential q in the elliptic boundary value problem

$$-\Delta u + qu = f \qquad \text{in } \Omega, \tag{5.1}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.2}$$

from a state observation in $L^2(\Omega)$, which is a well-studied problem in literature (cf., e.g., [34]). In [11], it has been shown that in the setup (d denotes the space dimension)

$$X = H_0^1(\Omega), \qquad X^* = H^{-1}(\Omega), \qquad Q = H^d(\Omega), \qquad Z = L^2(\Omega), \tag{5.3}$$

the operators

$$e : X \times Q \to X^*, (u, q) \mapsto (-\Delta u + qu) \tag{5.4}$$

$$E : X \to Z, u \mapsto u \tag{5.5}$$

satisfy all assumptions needed for the convergence analysis of the LMSQP method. Now we shall study a concrete finite element discretization of the KKT system and the derivation of estimates for the numerical errors ϵ_η, ϵ_h, and κ_h.

5.1. Error estimates for the discretized KKT system. If we denote the iterates for state and Lagrangian variable at step $k+1$ by u and λ, respectively, then we can write the whole KKT system in classical form as

$$-\Delta u + q_k u + q u_k = f + q_k u_k \qquad \text{in } \Omega, \tag{5.6}$$

$$-\Delta \lambda + q_k \lambda + u - z^\delta = 0 \qquad \text{in } \Omega, \tag{5.7}$$

$$\beta L_d(q - q_k) + u_k \lambda = 0 \qquad \text{in } \Omega, \tag{5.8}$$

again with homogenous Dirichlet boundary conditions upon u and λ on $\partial\Omega$, where L_d is a dimension-dependent differential operator of order $2d$ corresponding to the norm in $H^d(\Omega)$; e.g., we have

$$L_1 q = -q_{xx} + q, \tag{5.9}$$

$$L_2 q = \Delta(\Delta q + q) + q, \tag{5.10}$$

supplemented by homogenous boundary conditions up to order $d-1$. If $f \in L^2(\Omega)$ and $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, a standard elliptic regularity argument shows that $\hat{u}, u_k \in H^2(\Omega) \cap H_0^1(\Omega)$ (where $\hat{u}$ is the exact solid state) for all $k \in \mathbb{N}$. In the same way we can show that $\lambda_k \in H^2(\Omega) \cap H_0^1(\Omega)$ and $s_{k+1} - s_k \in H^{2d}(\Omega)$. This additional regularity can be employed to derive standard error estimates for finite element discretizations of the KKT system (2.12).

If we use piecewise linear finite elements on regular triangulations $\mathcal{T}_\eta$ and $\mathcal{T}_h$ for the discretization spaces Z_η and X_h, where η and h represent the fineness of the grids, then a classical approximation result for finite elements (cf. [33, p. 96]) implies that

$$\epsilon_\eta = \mathcal{O}(\eta^2) \qquad \text{and} \qquad \epsilon_h = \mathcal{O}(h). \tag{5.11}$$

Of course, one could also use piecewise constant elements on $\mathcal{T}_\eta$, which would yield $\epsilon_\eta = \mathcal{O}(\eta)$. However, in practical applications a higher-order approximation in η is often desirable, since η can be significantly larger than a reasonable choice of h. A canonical approximation of the parameter q is a finite element space of order greater than or equal to d on a regular triangulation $\mathcal{T}_{\tilde{h}} \subset \mathcal{T}_h$; under a priori assumptions on the exact solution $\hat{q}$ one can obtain quantitative estimates for κ_h in terms $\tilde{h}$. At first glance it seems surprising that one needs a priori assumptions on the parameter but not on the state in order to derive error estimates. However, due to the ill-posedness of the identification problem with respect to the parameter q, such a priori knowledge seems to be necessary. The approximation of the state corresponds rather

to the approximation of the underlying elliptic state equation, which is well-posed with respect to u and yields further regularity. We finally want to mention that according to the theory developed above, one could choose $\mathcal{T}_{\tilde{h}}$ independent of $\mathcal{T}_h$, but this would cause unnecessary complications in the implementation of the method.

We note that alternatively one can use the space $Q = L^2(\Omega)$ for $d \leq 3$, which yields $L_d = I$; i.e., (5.8) becomes

$$\beta(q - q_k) + u_k\lambda = 0. \tag{5.12}$$

An appropriate discretization strategy is, e.g., to choose Q_h as the space of piecewise constant elements on an underlying grid $\mathcal{T}_{\tilde{h}}$. The advantage of this approach is that elements of order greater than one, which are necessary for $Q = H^d(\Omega)$ $(d \geq 2)$, can be avoided.

5.2. Structure of the system matrix. For the potential identification problem, some parts of the system matrix M in (4.1) are well-understood. First of all, G is an L^2-mass matrix and it is positive definite if the triangulations $\mathcal{T}_\eta$ and $\mathcal{T}_h$ coincide, which we assume in the following. The eigenvalues of G are then all of order h^d. The matrix H is the stiffness matrix for the differential operator L_d, with minimal eigenvalue of order h^d and maximal eigenvalue of order h^{-d}.

The matrix K is the sum of a stiffness matrix for the Laplacian and a weighted mass matrix (with weight q_k in the L^2-scalar product), where one can expect the first part in this sum to be dominating. Thus, the stiffness matrix $\hat{K}$ for the Laplacian will be a good preconditioner for K. The maximal and minimal eigenvalues of K and $\hat{K}$ are of order h^{d-2} and h^d, respectively. The remaining part in the system matrix, namely the matrix L, is difficult to understand, since its elements are weighted L^2-scalar products of basis functions of different finite element spaces. However, the spectral norm of L can be estimated; it is of order $\tilde{h}^d$.

The construction of preconditioners for G and H is well investigated; even exact preconditioning seems to be applicable. For K it seems reasonable to use a preconditioner $\hat{K}$ for the Laplacian, e.g., a multigrid preconditioner. With preconditioning for K, the system matrix can be transformed to

$$\tilde{M} = \begin{pmatrix} G & 0 & K\hat{K}^{-1} \\ 0 & \beta H & L^T\hat{K}^{-1} \\ \hat{K}^{-1}K & \hat{K}^{-1}L & 0 \end{pmatrix}, \tag{5.13}$$

with the corresponding Schur complement

$$\tilde{C} = \hat{K}^{-1}KG^{-1}K\hat{K}^{-1} + \beta^{-1}\hat{K}^{-1}LH^{-1}L^T\hat{K}^{-1}. \tag{5.14}$$

If $\hat{K}$ is an appropriate preconditioner for K, then we can estimate the minimal eigenvalue by

$$\lambda_{\min}(\tilde{C}) \geq \lambda_{\min}(\hat{K}^{-1}KG^{-1}K\hat{K}^{-1}) = \mathcal{O}(h^{-d}) \tag{5.15}$$

and the maximal eigenvalue by

$$\lambda_{\max}(\tilde{C}) \leq \|\hat{K}^{-1}KG^{-1}K\hat{K}^{-1}\|_2 + \beta^{-1}\|\hat{K}^{-1}LH^{-1}L^T\hat{K}^{-1}\|_2 \tag{5.16}$$

$$= \mathcal{O}\left(h^{-d}(1 + \beta^{-1}h^{-2d}\tilde{h}^{2d})\right). \tag{5.17}$$

Hence, the condition number of $\tilde{C}$ is independent of h but depends only on β and $\frac{\tilde{h}}{h}$. One observes that the condition number is decreasing as $\tilde{h}$ tends to h from above. (Note that usually $\tilde{h} \geq h$.) For the Uzawa iteration, one can choose the preconditioner $\hat{C}$ in this case as a multiple of $\hat{K}^{-1}KG^{-1}K\hat{K}^{-1}$ or even of G^{-1}. If $\tilde{h} \gg h$, the Uzawa iteration seems not to be optimal; in this case one can apply either a reduced SQP approach or use Krylov-subspace methods with different preconditioning strategies. For the details on the latter we refer the reader to [12].

6. Numerical experiments. In order to test our theoretical results, we numerically solve some model problems, which have already been investigated with respect to the convergence behavior of the LMSQP method in [11].

Example 6.1. Our first example is the identification of the potential q in (5.1), (5.2) from a state observation $u \in L^2(\Omega)$, with $\Omega = (0,1)$, homogeneous Dirichlet values for u, and

$$f(x) = \frac{1}{2} + \sin x, \qquad x \in \Omega.$$

The exact potential is given by

$$q(x) = x(1-x),$$

which is an element of $Q = H^1(\Omega)$.

This problem was implemented in the software system MATLAB as follows. The data are generated by solving the state equation on a fine grid and subsequent interpolation to a coarser grid; the noise is an additive high-frequency perturbation. We used uniform grids with m nodes for the discretization of the state u and the Lagrange parameter λ and n nodes for the parameter q, i.e., $h = (m-1)^{-1}$ and $\tilde{h} = (n-1)^{-1}$. This implies a rather simple structure of the KKT submatrices, in particular $G = g_0 I$ with $g_0 \in \mathbb{R}^+$ and I the identity matrix, and H is the H^1-stiffness matrix. The parameters β_k are chosen according to $\beta_{k+1} = 0.9\beta_k$, with $\beta_0 = 10^{-6}$, which led to convergence of the method even for starting value $q \equiv 0$. Roughly speaking, this choice of the regularization parameters corresponds to a globalization strategy of trust-region type, since it is well known that the diameter of the trust region is negative proportional to the parameter β_k for Levenberg–Marquardt methods. The KKT system (4.1) is solved using a direct solver in this case, which is probably not the best choice with respect to the numerical effort for fine discretizations, but it still leads to reasonable results in our case.

The convergence results for the overall LMSQP method have been shown in [11] and compared to a Levenberg–Marquardt method following the feasible path. It turned out that both methods lead to almost the same iteration sequence q_k. In particular, the number of iterations needed until the stopping rule is satisfied is the same for both methods. Now we compare the numerical efficiency of the LMSQP method with feasible path approaches, namely the Levenberg–Marquardt (LM) method on the feasible path (with the same Galerkin discretization as for LMSQP and solution of the Gauss–Newton system by a CG method) and a Broyden-type variant of the LM method (cf. [24] for further details).

For this sake we choose different discretization levels (fixed during the iteration) and measure the CPU time needed for the LMSQP method until the stopping rule is satisfied (for fixed noise level δ). From the results shown in Table 1 one observes that the LMSQP method with simultaneous solution of the KKT system outperforms

TABLE 1
CPU time (in seconds) needed for the LMSQP method, the LM method, and a Broyden-type variant of the LM method.

m	n	LMSQP	LM	Broyden
201	41	0.07	1.37	0.51
201	101	0.18	3.44	1.34
201	201	0.36	6.94	2.88
401	201	0.51	24.83	9.09
401	401	1.39	50.39	20.48
801	401	2.61	193.21	70.69
801	801	5.66	392.54	158.69
1601	801	7.91	1564.50	600.66
1601	1601	22.86	3144.40	1356.60

the feasible-path approaches for all different discretizations. Since the LMSQP and the LM method need the same number of outer iterations, the difference in the numerical effort is caused by the fact that the effort for matrix-vector products with the system matrix in the LM method is significantly higher than the preconditioning and performing of matrix-vector products with the system matrix in the simultaneous LMSQP method. Obviously, the gain in the numerical effort for the evaluation of the system matrix increases with the number of discretization points, which explains the extremely large CPU time for the LM method at the finest discretization level ($m = 1601$). For small m and n, the Broyden variant is much faster than the LM method, which is again caused by the fact that the evaluation of the system matrix can be carried out efficiently. However, the number of iterations needed for the Broyden-type variant is much larger than for the other two methods, which use the full information about the derivatives.

Finally, we investigate the spectral condition of the system matrix M as well as of the matrix $\tilde{M}$ defined by (5.13), where we use a Jacobi preconditioner for the Laplacian as $\hat{K}$. From the left picture in Figure 1, which shows the condition number as a function of the discretization size h (in logarithmic scale) for fixed $\beta = 10^{-5}$, one observes that the condition number of M grows quadratically with h^{-1}, while the condition number of $\tilde{M}$ is much smaller and almost independent of h. The second part of Figure 1 shows a plot of the condition numbers vs. the parameter β in doubly logarithmic scale, from which it seems that the growth of the condition number as $\beta \to 0$ is slower for $\tilde{M}$ than for the original matrix M. In both cases, the condition number seems to be a convex function of β, which has a unique minimum at some $\overline{\beta}$. However, this value $\overline{\beta}$ is rather large, and values of β that are significantly larger than $\overline{\beta}$ are not of interest for our purpose, since they would cause a tremendous slowdown of the outer iteration. Therefore we can focus our attention on the case $\beta < \overline{\beta}$, where the condition number increases monotonically with β^{-1}.

Example 6.2. Our second numerical example is the identification of the conductivity $q \in L^\infty(\Omega)$ in

$$-\operatorname{div}(q\nabla u) = f \qquad \text{in } \Omega, \tag{6.1}$$

$$u = g \qquad \text{on } \partial\Omega \tag{6.2}$$

from a state observation $u \in L^2(\Omega)$. The domain Ω is a ball in $\mathbb{R}^2$ with a missing first quadrant; i.e., in radial coordinates

$$\Omega = \{\ (r\cos\theta, r\sin\theta) \mid r \in [0,1), \theta \in (\pi/2, 2\pi)\ \}. \tag{6.3}$$

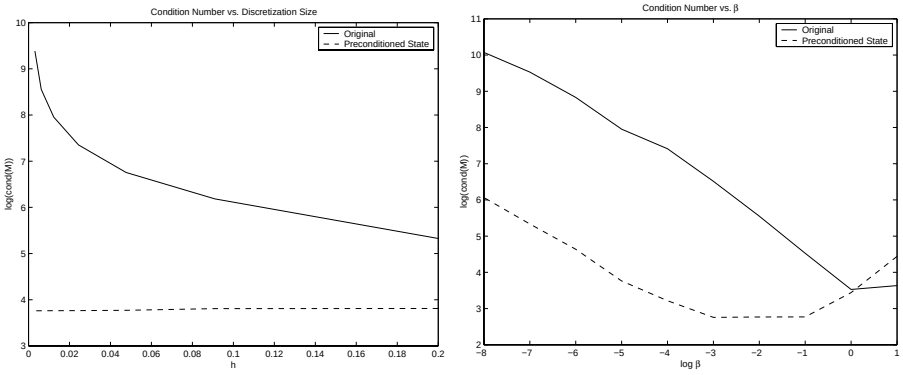


FIG. 1. *Plot of the spectral condition of the matrix M vs. the discretization size h (in logarithmic scale, left) and vs. the parameter β (in doubly logarithmic scale, right). The solid line shows the condition number of the original matrix M, the dashed line of the matrix $\tilde{M}$ with the preconditioned state equation.*

The exact parameter to be reconstructed is $\hat{q} \equiv 1$, and the right-hand side in (6.1) is given by

$$f = \frac{3\pi}{4}\left(3\pi\cos\left(\frac{3\pi}{2}r\right) + \frac{2}{r}\sin\left(\frac{3\pi}{2}r\right)\right) \qquad \text{with } r = \sqrt{x^2+y^2}.$$

The corresponding solution of the state equation is $\hat{u} = \cos(\frac{3\pi}{2}r)$. The data are generated using the exact solution $\hat{u}$ perturbed by uniformly distributed random noise. For the discretization we used triangular finite elements with piecewise quadratic shape functions for the state u and the Lagrange parameter λ and piecewise constant shape functions for the parameter q. The results were calculated using the finite element code FEPP [28], developed at the Institute of Computational Mathematics of the University of Linz.

We want to mention that this identification problem is quite challenging not only due to the complicated geometry, but also due to the fact that q is not identifiable along a level line in the interior, where u attains an extremum. This does not destroy the theoretical identifiability results, because it is a set of Lebesgue-measure zero, but it can be expected to create numerical difficulties.

The KKT system (4.1) was solved using a preconditioned QMR method with a block-factorization type preconditioner of the form

$$\hat{M} = \begin{pmatrix} G\hat{K}^{-1} & 0 & I \\ 0 & I & L^T\hat{K}^{-T} \\ I & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{K} & L & 0 \\ 0 & S_c & 0 \\ 0 & -G\hat{K}^{-1}L & \hat{K}^T \end{pmatrix} \tag{6.4}$$

(cf. [5, 6, 21]) with a multigrid preconditioner $\hat{K}$ and no preconditioning of the Schur complement S_c. Results for exact data can be found in Table 2. The good performance of the method with respect to both CPU time and the number of outer iterations can be observed clearly. Especially for problems with fine discretizations of the parameter q, this method can still be realized efficiently, while classical approaches do not yield results in reasonable time. A plot of the parameter q can be found in Figure 2, from which one observes that the parameter is reconstructed very well except in a neighborhood of the level curve $\{\nabla u = 0\}$.

TABLE 2
CPU time and number of inner (QMR) and outer (SQP) iterations for exact data

Level	dim q	dim u	Avg QMR it	SQP it	Time
2	92	215	200	9	8 sec
3	368	797	200	4	15 sec
4	1472	3065	180	5	77 sec
5	5888	12017	142	6	450 sec

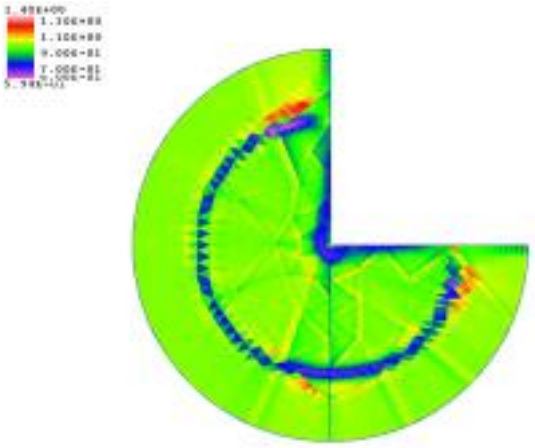


FIG. 2. *Parameter distribution for exact data at level* 4, $q_{min} = 0.59$, $q_{max} = 1.4$.

Additional speed-up can be gained using a multilevel approach as described in subsection 4.2. We used nested spaces for q and u by subdividing each triangular element into four smaller elements, when refining the mesh. Table 3 presents results for this approach. It can be seen that on fine discretization levels one SQP step is sufficient for fulfilling the stopping criterion, which corresponds very well to the theoretical predictions made in section 4.2. A comparison of the results to the ones in Table 2 shows that for fixed discretization level the solution of the identification problem on level 5 is only slightly faster than the identification of q on level 6 (with about the fourfold number of parameters) using a multilevel approach. A plot of the parameter can be found in Figure 3. Here the approximation of the parameter in the area where it cannot be identified is by far better than in the classical approach using only one discretization level (compare Figure 2). A possible explanation for this effect is the following: the influence of the level line $\{\nabla u = 0\}$ where q cannot be identified on the solution is smaller the coarser the discretization. The prolongation from coarse levels to finer ones adds information to the region where the parameter is not identifiable from its surrounding region. As long as the parameter is smooth this helps to improve the quality of the numerical results where the parameter cannot be identified.

7. Conclusions and outlook. We have developed a framework for Galerkin-type approximations of the LMSQP method for parameter identification problems in elliptic partial differential equations, and we have discussed the implementation of the GLMSQP method with iterative solution of the KKT system. The numerical results show that the resulting iteration method clearly outperforms state-of-the-art methods for iterative regularization and provides a tool for the efficient solution of identification

TABLE 3
CPU time per level, accumulated time, and number of inner (QMR) and outer (SQP) iterations for exact data using a nested multilevel approach.

Level	dim q	dim u	Avg QMR it	SQP it	Time	Acc. time
2	92	215	200	9	8 sec	8 sec
3	368	797	200	4	15 sec	23 sec
4	1472	3065	175	2	24 sec	47 sec
5	5888	12017	80	1	47 sec	94 sec
6	23552	47585	121	1	425 sec	520 sec

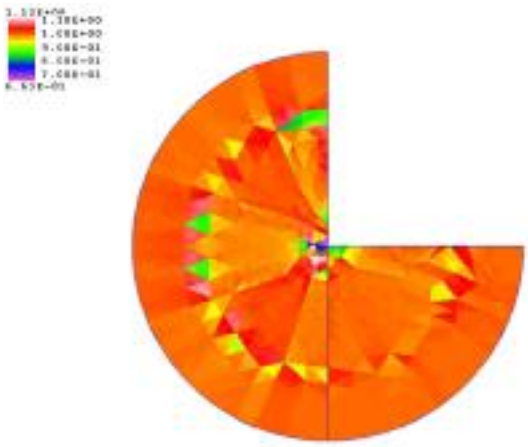


FIG. 3. *Parameter distribution for exact data at level* 4 *using a nested multilevel approach,* $q_{min} = 0.66$, $q_{max} = 1.13$.

problems with fine discretizations. Moreover, we have developed a multilevel version of the GLMSQP method, which yields a further speed-up.

The crucial point for the possibility to obtain an efficient implementation of the LMSQP method is the preconditioning of the KKT system, which is then solved iteratively as an indefinite problem in the product space for state, parameter, and Lagrangian variable. The construction of such preconditioners is not a simple task and has not been discussed in detail in the present paper, but will be investigated in [12], where different preconditioning techniques will be compared.

Other numerical aspects to be investigated in future research are adaptive discretization strategies and fast parallel solvers based on domain decomposition techniques. The adaptive discretization of optimal control problems, which is a closely related subject, has been discussed by Becker, Kapp, and Rannacher [4]; possibly the ideas of this work can be carried over to identification problems, too. The parallel solution of optimal control problems has been investigated by Lions and Pironneau [29] in the case of quadratic problems; recently Biros and Ghattas [5, 6] performed a numerical study of a parallel solver with an SQP method for the outer and preconditioned Krylov-subspace methods for the inner iteration. Many of their ideas seem to be applicable also for parameter identification problems that are solved with the LMSQP method, which raises the hope that efficient parallel versions of the LMSQP method can be designed also for large-scale identification problems such as impedance tomography.

Finally, we want to recall that the framework of this problem does not apply

to transient problems of parabolic or hyperbolic type. Since numerical methods for different types of partial differential equations have many type-specific features, in general, it is not surprising that also the numerical treatment of parameter identification problems should depend on the type of the underlying state equation. However, it seems possible to construct efficient and convergent discretized methods at least in the case of parabolic equations, which is an important task for future research.

Acknowledgments. The authors thank Dr. Walter Zulehner (University of Linz) and Dr. Joachim Schöberl (currently at Texas A & M University) for useful and stimulating discussions on the preconditioning of the indefinite system (4.1).

REFERENCES

[1] W. Alt, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Basel, Boston, Berlin, 1989.

[3] A. Battermann and M. Heinkenschloss, *Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems*, in Optimal Control of Partial Differential Equations, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser, Basel, Boston, Berlin, 1998, pp. 15–32.

[4] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[5] G. Biros and O. Ghattas, *Parallel Lagrange-Newton-Krylov-Schur methods for PDE-Constrained Optimization Problems. Part* 1*: The Krylov–Schur Solver*, preprint, Carnegie Mellon University, Pittsburgh, PA, 2000.

[6] G. Biros and O. Ghattas, *Parallel Lagrange-Newton-Krylov-Schur Methods for PDE-Constrained Optimization Problems. Part* 2*: The Lagrange-Newton Solver and Its Application to Optimal Control of Steady Viscous Flow*, preprint, Carnegie Mellon University, Pittsburgh, PA, 2001.

[7] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev, *Analysis of the inexact Uzawa algorithm for saddle point problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1072–1092.

[8] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 8 (1974), pp. 129–151.

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer, New York, 1991.

[10] M. Burger, *Iterative regularization of a parameter identification problem occurring in polymer crystallization*, SIAM J. Numer. Anal., 39 (2001), pp. 1029–1055.

[11] M. Burger and W. Mühlhuber, *Iterative regularization of parameter identification problems by SQP-methods*, Inverse Problems, 18 (2002), pp. 943–969.

[12] M. Burger and W. Mühlhuber, *Iterative solution of saddle-point problems arising in the identification and control of distributed parameters*, in preparation.

[13] D. Colton, R. Ewing, and W. Rundell, eds., *Inverse Problems in Partial Differential Equations*, SIAM, Philadelphia, 1990.

[14] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust Region Methods*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, 2000.

[15] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point SQP algorithms for a class of nonlinear programming problems*, SIAM J. Control. Optim., 36 (1998), pp. 1750–1794.

[16] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[17] H. W. Engl and W. Rundell, eds., *Inverse Problems in Diffusion Processes. Proceedings of the GAMM-SIAM Symposium*, SIAM, Philadelphia, 1995.

[18] H. W. Engl and O. Scherzer, *Convergence rate results for iterative methods for solving nonlinear ill-posed problems*, in Solution Methods for Inverse Problems D. Colton, H. W. Engl, J. McLaughlin, A. Louis, and W. Rundell, eds., Springer, Vienna, New York, 2000, pp. 7–34.

[19] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[20] P. E. Gill, W. Murray, and M. H.Wright, *Practical Optimization*, Academic Press, London, San Diego, New York, 1981.
[21] E. Haber and U. Ascher, *Preconditioned all-at-once methods for large, sparse parameter estimation problems*, Inverse Problems, 17 (2001), pp. 1847–1864.
[22] M. Heinkenschloss, *The numerical solution of a control problem governed by a phase field model*, Optim. Methods Softw., 7 (1997), pp. 211–263.
[23] B. Heise, *Nonlinear field calculations with multigrid Newton methods*, Impact Comput. Sci. Engrg., 5 (1993), pp. 75–110.
[24] B. Kaltenbacher, *On Broyden's method for the regularization of nonlinear ill-posed problems*, Numer. Funct. Anal. Optim., 19 (1998), pp. 807–833.
[25] B. Kaltenbacher, *A projection-regularized Newton method for nonlinear ill-posed problems and its application to parameter identification problems with finite element discretization*, SIAM J. Numer. Anal., 37 (2000), pp. 1885–1908.
[26] B. Kaltenbacher, M. Kaltenbacher, and S. Reitzinger, *Identification of nonlinear B-H curves based on magnetic field computations and multigrid methods for ill-posed problems*, European J. Appl. Math., to appear.
[27] B. Kaltenbacher and J. Schöberl, *A saddle point variational formulation for projection-regularized parameter identification*, Numer. Math., 91 (2002), pp. 675–697.
[28] M. Kuhn, U. Langer, and J. Schöberl, *Scientific computing tools for* 3*D magnetic field problems*, in The Mathematics of Finite Elements and Applications X, J. R. Whiteman, ed., Elsevier, Amsterdam, 2000, pp. 239–258.
[29] J. L. Lions and O. Pironneau, *Sur le controle parallele des systemes distribues*, C. R. Acad. Sci. Paris Ser. I Math., 327 (1998), pp. 993–998.
[30] D. Lukáš, *Shape Optimization of Homogeneous Electromagnets*, SFB-Report 00-30, University of Linz, Linz, Austria, 2000.
[31] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer, New York, Berlin, Heidelberg, 1999.
[32] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite linear systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[33] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Series in Comput. Math. 23, Springer, Berlin, Heidelberg, New York, 1994.
[34] D. E. Reeve and M. Spivack, *Determination of a source term in the linear diffusion equation*, Inverse Problems, 10 (1994), pp. 1335–1344.
[35] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
[36] E. W. Sachs, *Control applications of reduced SQP methods*, in Computational Optimal Control, R. Bulirsch and D. Kraft, eds., Birkhäuser, Basel, 1994, pp. 89–104.
[37] V. Schulz and H. G. Bock, *Partially reduced SQP methods for large-scale nonlinear optimization problems*, Nonlinear Anal., 30 (1997), pp. 4723–4734.
[38] V. Schulz, *Solving discretized optimization problems by partially reduced SQP methods*, Comp. and Vis. in Science, 1 (1998), pp. 83–96.
[39] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim., 37 (1999), pp. 731–764.
[40] W. Zulehner, *Analysis of iterative methods for saddle point problems: a unified approach*, Math. Comput., 71 (2002), pp. 479–505.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1798–1822

THE SPECTRUM OF CIRCULANT-LIKE PRECONDITIONERS FOR SOME GENERAL LINEAR MULTISTEP FORMULAS FOR LINEAR BOUNDARY VALUE PROBLEMS*

DANIELE BERTACCINI†

Abstract. The spectrum of the eigenvalues, the conditioning, and other related properties of circulant-like matrices used to build up block preconditioners for the nonsymmetric algebraic linear equations of time-step integrators for linear boundary value problems are analyzed. Moreover, results concerning the entries of a class of Toeplitz matrices related to the latter are proposed. Generalizations of implicit linear multistep formulas in boundary value form are considered in more detail.

It is proven that there exists a new class of approximations which are well conditioned and whose eigenvalues have positive and bounded real and bounded imaginary part. Moreover, it is observed that preconditioners based on other circulant-like approximations, which are well suited for Hermitian linear systems, can be severely ill conditioned even if the matrices of the nonpreconditioned system are well conditioned.

Key words. trigonometric preconditioners, nonsymmetric Toeplitz matrices, eigenvalues, linear systems of time-step integrators, general linear multistep formulas in boundary value form, boundary value problems

AMS subject classifications. 65F10, 65F15, 65N22, 15A18, 15A48

PII. S0036142901397447

1. Introduction.

In this paper we investigate the properties of some classes of circulant approximations and some generalizations used in the preconditioners for (small rank perturbations of) block nonsymmetric Toeplitz matrices introduced in [4]. These matrices arise in the numerical approximation of time-dependent partial differential equations by means of generalizations of implicit linear multistep formulas.

An $n \times n$ matrix $A_n = (a_{j,k})$ is said to be Toeplitz if $a_{j,k} = a_{j-k}$, $j, k = 1, \dots, n$, i.e., A_n is constant along its diagonals, quasi Toeplitz if it is a small rank perturbation of a Toeplitz matrix. An $n \times n$ matrix $\breve{A}_n$ is said to be circulant if it is Toeplitz and its diagonals satisfy $\breve{a}_{n-j} = \breve{a}_{-j}$, $j = 1, \dots, n-1$. The circulant matrices $\breve{A}_n$ are diagonalized by the Fourier matrix $F = (F_{j,k})$, $F_{j,k} = e^{2\pi \mathbf{i} jk/n}/\sqrt{n}$, $j, k = 0, \dots, n-1$, $\mathbf{i}$ is the imaginary unit; see [13]. From the previous arguments, it follows that such matrices are easily and efficiently invertible using the fast Fourier transform (FFT); see, e.g., [11]. Other circulant-like matrices will be mentioned in section 4.

The matrices of the underlying linear systems can be written as follows:

$$M = A \otimes I - h\, B \otimes J, \tag{1.1}$$

where A and B are $n \times n$ (small rank perturbations of) band Toeplitz matrices whose entries are given by the coefficients of the scheme involved, I is the identity, and J is an $m \times m$ matrix which can be large and sparse. More precisely, J is the Jacobian matrix of a system of ordinary or partial differential equations discretized in space by

*Received by the editors November 5, 2001; accepted for publication (in revised form) June 9, 2002; published electronically November 22, 2002. This work was partially supported by an INdAM-GNCS project and by a grant from the MURST project "Progetto giovani ricercatori anno 2000."
http://www.siam.org/journals/sinum/40-5/39744.html
†Università di Roma "La Sapienza," Dipartimento di Matematica, P.le A. Moro 2, 00185 Roma, Italy (bertaccini@mat.uniroma1.it).

finite differences; see [4] for details. It is worth noting that J can have a (multilevel) structure as well. For example, J can be block-banded, block-Toeplitz, etc.

Unfortunately, when m and/or n are (even moderately) large, iterative solvers for (1.1), used without preconditioners or with general purpose preconditioners such as those based on incomplete factorizations, often converge very slowly or do not converge at all; see [4, section 5]. Moreover, direct methods are not appropriate because they cannot exploit the block structure of (1.1). On the other hand, the preconditioners we consider here take into account the block structure in (1.1). More precisely, they are block-circulant and, in matrix form, can be written as

$$P = \breve{A} \otimes I - h\,\breve{B} \otimes \tilde{J}, \tag{1.2}$$

where $\breve{A}$ and $\breve{B}$ are circulant-like approximations for A, B, respectively, and $\tilde{J}$ is a suitable approximation for J.

In [4] we have observed that $P^{-1}M$, the preconditioned matrix, can be written as a perturbation of the identity matrix (see also section 5.3), which can result in fast convergence of Krylov subspace methods for nonsymmetric linear systems such as GMRES and BiCG-like methods such as BiCGstab. The computational cost for a possible implementation has been considered in detail in [4, section 4.1], showing that the cost per iteration is of the order of $O(mn \log n)$ if J is banded, say.

Here we will prove that there exists a class of circulant approximations, introduced in [4], which have a moderate 2-norm condition number increasing at most linearly with their size n. Moreover, the spectrum of the eigenvalues of several of the possible approximations for the nonsymmetric matrices A, B in (1.1) will be investigated as well, showing that it lies in the right half plane. It is worth noting that this holds true for the original matrices A, B in (1.1) considered here; see [6].

We stress that the condition number and the spectrum of the component matrices $\breve{A}$ and $\breve{B}$ of the preconditioner (1.2) are very important to have fast convergence. Indeed, as observed in [4, 5, 6], the matrix J in (1.1) can have very small (and/or very large) singular values in different subintervals of integration (see [5]), and this is difficult (if not impossible) to know in advance. Recall that J is the Jacobian matrix of the given continuous time-dependent problem; see [15, 19]. Thus, if $\breve{A}$ or $\breve{B}$ are ill conditioned, we can have an ill conditioned preconditioner even if the original matrix M is well conditioned; see, e.g., the end of section 5.2 and Figure 5.3. As observed in [4, 5], this can slow down the convergence process (see [14]), giving unacceptably slowly convergent (or even divergent!) preconditioned iterations.

We observe that, in the case of nonsymmetric linear systems, solved by Krylov accelerators, the condition number of the *preconditioned* matrix $P^{-1}M$ (say), assumed to be not too large, is much less important for the convergence than the clustering of the spectrum of its eigenvalues; see, e.g., [18, 14]. On the other hand, the condition number of the matrix M in (1.1) is crucial for the rate of convergence of conjugate gradients preconditioned iterations for the normal linear system; see [6].

Here we will consider certain general linear multistep methods (or GLMs, see [19]) used in boundary value form and called boundary value methods. These methods are used to solve continuous boundary value problems for differential equations (see [2, 9] and references therein). However, the asymptotic techniques considered here could be adapted, at least in principle, to other discretization schemes.

Notice that, in this paper, we will consider multistep formulas of arbitrarily high order merely to state bounds and the asymptotic behavior of the spectrum of the underlying circulant(-like) approximations involved in (1.2). In practice, the best

performance of the underlying preconditioners seems to be achieved for formulas (2.3) whose number of steps k is not too large (typically 3 to 9, say). On the other hand, we have observed in [4, 5] that the preconditioners (1.2) can be effective for any order of magnitude of n, either if it is small (4 to 8, say) or (very) large ($n > 1024$, say) as well.

In section 2 we summarize some information on numerical integrators based on linear multistep formulas in boundary value form. Section 3 contains some introductory lemmas. In section 4 we recall some circulant approximations. Section 5 is devoted to the investigation on the spectrum and the conditioning of the underlying matrices. Finally, some remarks on the convergence of preconditioned iterations and the use of different approximations in (1.2) are given in section 5.3.

2. Families of numerical integrators.

2.1. Formulas in boundary value form. The boundary value methods for differential equations are a generalization of implicit linear multistep formulas; see [2, 9] and references therein. They approximate the solution of a continuous differential boundary value problem by means of a discrete boundary value problem. For simplicity, let us consider the linear boundary value problem

$$\begin{cases} y'(t) = f(t, y(t)) := J\, y(t) + g(t), \quad t \in (t_0, T], \\ y(t_0) = \eta_1, \quad y(T) = \eta_2, \end{cases} \tag{2.1}$$

where $y(t), g(t) : \mathbb{R} \to \mathbb{R}^m$, $J \in \mathbb{R}^{m\times m}$, $\eta_j \in \mathbb{R}^m$, $j = 1, 2$. The continuous problem (2.1) can be reduced to a discrete boundary value problem by the following k-step linear multistep formula of order p used with $\nu > 0$ initial and $k-\nu > 0$ final conditions over a uniform mesh $t_j = t_0 + j\, h,\ j = 0, \ldots, s$:

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f_{n+i}, \qquad n = 0, \ldots, s-k, \tag{2.2}$$

where y_n is the discrete approximation to $y(t_n)$, $f_n = f(t_n, y_n) \equiv J\, y_n + g_n$, $g_n = g(t_n)$, while the values $y_0, \ldots, y_{\nu-1}, y_{s-k+\nu+1}, \ldots, y_s$ of the approximation computed in the mesh points $t_0, \ldots, t_{\nu-1}, t_{s-k+\nu+1}, \ldots, t_s$, respectively, are assumed to be given. We observe that the boundary value problem (2.1) provides only the initial and final values y_0 and y_s, respectively. The missing values are supplied by coupling the method (2.2) with other difference schemes of order p, sometimes called additional methods, which provide an additional set of equations, independent of those in (2.2). For simplicity, we can assume that these formulas have the same number of steps as (2.2) but different coefficients $\alpha_j^{(r)}$, $\beta_j^{(r)}$, $r = 1, \ldots, \nu-1, s-k+\nu+1, \ldots, s-1$, $j = 0, \ldots, k > \nu$; see [4] for details.

In order to stress the dependence of the formula on the ν initial and $k-\nu$ final values, it is useful to rewrite (2.2) in the following shifted form:

$$\sum_{i=-\nu}^{k-\nu} \alpha_{i+\nu} y_{n+i} = h \sum_{i=-\nu}^{k-\nu} \beta_{i+\nu} f_{n+i}, \qquad n = \nu, \ldots, s-k+\nu. \tag{2.3}$$

To have order $p \geq 1$, the coefficients α_j, β_j in (2.3) should satisfy the order conditions (see, e.g., [19])

$$\sum_{j=0}^{k} \left(j^i \alpha_j - i j^{i-1} \beta_j \right) = 0, \qquad i = 0, \ldots, p, \tag{2.4}$$

where the first two equations of (2.4) ($i = 0, 1$) are usually called consistency conditions. By rewriting (2.4) in shifted form we have

$$\sum_{j=-\nu}^{k-\nu} \left(j^r \alpha_{j+\nu} - r j^{r-1} \beta_{j+\nu} \right) = 0, \qquad r = 0, \dots, p, \tag{2.5}$$

where we assume, as is customary, that $0 j^{-1} = 0$, $j^0 = 1$ for all j.

For practical implementation, we cast the above discrete problem in matrix form:

$$\begin{aligned} &MY = b,\ Y = \left(y_0^T, y_1^T, \dots, y_s^T \right)^T, \qquad M = A \otimes I_m - hB \otimes J, \\ &b = e_1 \otimes \eta_1 + e_{s+1} \otimes \eta_2 + h(B \otimes I_m) g, \qquad g = (g(t_0) \cdots g(t_s))^T, \end{aligned} \tag{2.6}$$

where $e_i \in \mathbb{R}^{s+1}$, $i = 1, \dots, s+1$, is the ith column of the identity matrix and A, $B \in \mathbb{R}^{(s+1)\times(s+1)}$ are quasi-Toeplitz matrices whose pattern is

$$A = \begin{pmatrix} 1 & \cdots & 0 & & & \\ \alpha_0^{(1)} & \cdots & \alpha_k^{(1)} & & & \\ \vdots & \vdots & \vdots & & & \\ \alpha_0^{(\nu-1)} & \cdots & \alpha_k^{(\nu-1)} & & & \\ \alpha_0 & \cdots & \alpha_k & & & \\ & \alpha_0 & \cdots & \alpha_k & & \\ & & \ddots & \ddots & \ddots & \\ & & & \alpha_0 & \cdots & \alpha_k \\ & & & \alpha_0^{(s-k+\nu+1)} & \cdots & \alpha_k^{(s-k+\nu+1)} \\ & & & \vdots & \vdots & \vdots \\ & & & \alpha_0^{(s-1)} & \cdots & \alpha_k^{(s-1)} \\ & & & 0 & \cdots & 1 \end{pmatrix}, \tag{2.7}$$

where $\alpha_j^{(r)}$, $j = 0, \dots, k$, are the coefficients of the additional formulas. The matrix B is similarly defined, but with β_j ($\beta_j^{(r)}$) instead of α_j ($\alpha_j^{(r)}$) and with the entries of the first and last rows set to zero; see [4] for details. We stress that the matrix M in (2.6) is usually nonsymmetric, nondiagonally dominant, and large and sparse if, e.g., n or m are large and J is sparse.

2.2. Some linear multistep schemes. Let us recall some families of formulas (2.3) that will be considered in subsequent sections.

A minimal requirement for a multistep formula (2.3) is consistency, see, e.g., [19], i.e., they must satisfy the conditions $\rho(1) = 0$, $\rho'(1) = \sigma(1)$, where $\rho(z)$ and $\sigma(z)$ denote the two characteristic polynomials associated with the given method, i.e.,

$$\rho(z) = \sum_{j=0}^{k} \alpha_j \ z^j, \quad \sigma(z) = \sum_{j=0}^{k} \beta_j \ z^j$$

or, in shifted form,

$$\rho(z) = z^\nu \sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} \ z^j, \quad \sigma(z) = z^\nu \sum_{j=-\nu}^{k-\nu} \beta_{j+\nu} \ z^j. \tag{2.8}$$

The backward differentiation formulas are a class of well-known initial value methods for the numerical integration of stiff problems (see, e.g., [15, 19]). A generalization of these as a boundary value scheme, called generalized backward differentiation formulas, has been proposed in [9] and can be written in the form

$$\sum_{i=-\nu}^{k-\nu} \alpha_{i+\nu}\, y_{n+i} = h\, f_n, \qquad n = \nu, \dots, s-k+\nu, \tag{2.9}$$

where $\nu = (k+2)/2$ if k is even, and $\nu = (k+1)/2$ if k is odd; see [9]. Notice that backward differentiation formulas have $\nu = k$ in (2.9). The coefficients $\{\alpha_i\}$ are determined by imposing maximum order for (2.9), i.e., order k, $k \geq 1$.

Another popular class of initial value methods is the Adams–Moulton formulas; see, e.g., [15, 19]. Let us consider their generalization in the boundary value form, proposed in [9], called generalized Adams–Moulton methods, that can be written in the following form:

$$y_n - y_{n-1} = h \sum_{i=-\nu}^{k-\nu} \beta_{i+\nu}\, f_{n+i}, \qquad n = \nu, \dots, s-k+\nu, \tag{2.10}$$

i.e., the only nonzero coefficients in the first characteristic polynomial are $\alpha_\nu = 1$ and $\alpha_{\nu-1} = -1$, $\nu = k/2$ if k is even, and $\nu = (k+1)/2$ if k is odd. The coefficients $\{\alpha_i\}$ are determined by imposing that the method has maximum order, i.e., $k+1$. Notice that the classical Adams–Moulton methods have $\nu = k$; see, e.g., [19]. When k is odd, the scheme shares the same stability properties of the trapezoidal rule. Such methods can be suitable for approximating Hamiltonian problems and continuous boundary value problems.

Another generalization of the trapezoidal rule proposed in [9] is given by the following formula:

$$\sum_{i=-\nu}^{\nu-1} \alpha_{i+\nu}\, y_{n+i} = \frac{h}{2}(f_n + f_{n-1}), \qquad n = \nu, \dots, s-k+\nu, \tag{2.11}$$

where $\nu = (k+1)/2$ if k is odd and $\nu = k/2$ if k is even. The coefficients $\{\alpha_i\}$ are determined by imposing that the above formula has maximum order, i.e., $k+1$. Such methods can be suitable for approximating Hamiltonian problems and continuous boundary value problems.

It will be useful in the following sections to have some of the order conditions (2.4) for the above mentioned schemes written in a different form. For the formulas (2.11), we consider (2.4) with $\beta_\nu = \beta_{\nu-1} = 1/2$. Therefore, we have

$$\sum_{j=-\nu}^{\nu-1} j^r \alpha_{j+\nu} = (-1)^{r+1}\frac{r}{2}, \quad r = 0, 2, \dots, k+1, \qquad \sum_{j=-\nu}^{\nu-1} j\alpha_{j+\nu} = 1. \tag{2.12}$$

Similarly, for (2.9), $\beta_\nu = 1$, $\beta_j = 0$ for $j \neq \nu$. Thus, the α_j, $j = 0 \dots, k$, satisfy consistency conditions and

$$\sum_{j=-\nu}^{k-\nu} j^r \alpha_{j+\nu} = 0, \qquad r = 0, 2, \dots, k. \tag{2.13}$$

Finally, for (2.10), $\alpha_\nu = -\alpha_{\nu-1} = 1$ while the coefficients β_j, $j = 0 \dots, k$, satisfy

$$\sum_{j=-\nu}^{k-\nu} j^r \beta_{j+\nu} = (-1)^r \frac{1}{r+1}, \qquad r = 0, 1, \dots, k. \tag{2.14}$$

3. The entries of a class of Toeplitz matrices. Let us consider the $n \times n$ band Toeplitz matrix $\hat{A}_n = (\alpha_j)$, $n > k$,

$$\hat{A}_n = \begin{pmatrix} \alpha_\nu & \dots & \alpha_{k-1} & \alpha_k & 0 & \dots & 0 \\ \vdots & \alpha_\nu & \ddots & \alpha_{k-1} & \alpha_k & \ddots & \vdots \\ \alpha_0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \alpha_{k-1} & \alpha_k \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \alpha_{k-1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \alpha_0 & \dots & \alpha_\nu \end{pmatrix}, \tag{3.1}$$

and $\hat{B}_n = (\beta_j)$ having similar pattern, but with β_j instead of α_j, $j = 0, \dots, k$. If α_j, β_j, $j = 0, \dots, k$, are the coefficients of (2.3), $E_n^{(A)} = A_n - \hat{A}_n$, $E_n^{(B)} = B_n - \hat{B}_n$ are small rank matrices if $n \gg k$, where $A \equiv A_n$ is defined in (2.7) and similarly for $B \equiv B_n$.

It can be checked that all such matrices are, in general, nonsymmetric, nondiagonally dominant, with real entries of nonconstant sign. Moreover, let us associate to the matrices $\hat{A}_n$, $\hat{B}_n$ (A_n and B_n) as above the functions $g_A(z)$, $g_B(z)$, respectively. It is customary to call $g_A(z)$ the symbol of the matrix A_n, see, e.g., [8], where

$$g_A(z) = z^{-\nu}\rho(z) = \sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} z^j, \qquad z \in \mathbb{C}, \tag{3.2}$$

and $\rho(z)$ is the characteristic polynomial of $\hat{A}_n$ while $g_B(z)$ is defined similarly for $\hat{B}_n$ from $\sigma(z)$ in (2.8). The set $\{q \in \mathbb{C} : q = g_A(e^{\mathbf{i}\theta}), 0 \le \theta < 2\pi\}$ is called the boundary locus of the Toeplitz matrix $\hat{A}_n$. It is worth noting that $g_A(e^{\mathbf{i}j\theta})$, $g_B(e^{\mathbf{i}j\theta})$ are the generating functions of the band Toeplitz matrices $\hat{A}_n$, $\hat{B}_n$, respectively; see, e.g., [8, 11].

The Toeplitz matrices related to the linear multistep formulas considered in section 2.2 have the boundary locus and their spectrum of eigenvalues in the right half plane; see [9]. We recall that the families of matrices $\{\hat{A}_n\}$, $\{\hat{B}_n\}$ are such that their entries α_j, β_j, $j = 0, \dots, k$, satisfy the system of linear equations (2.4), where p in (2.4) is the largest integer such that those equations are independent. Notice that the choice of ν is strictly related to the condition number of the underlying Toeplitz matrices; see [6, 8, 9]. For example, with the choice suggested in section 2.2, the matrices $\{\hat{A}_n\}$, $\{\hat{B}_n\}$ related to the formulas (2.9), (2.10), (2.11) have a condition number which increases at most linearly with their size; see [6]. On the other hand, the boundary locus and the spectrum of the eigenvalues of the matrices related to linear multistep formulas are not necessarily contained in one half plane for all k. For example, one can consider the matrices associated with well-known families of formulas used as initial value methods for a sufficiently large value of k in (2.3). This is the

case of the backward differentiation formulas for $k > 2$ and of the Adams–Moulton methods for $k > 1$. However, if ν is chosen differently from the choice suggested in section 2.2, the eigenvalues of $\{\hat{A}_n\}$, $\{\hat{B}_n\}$ can have both positive and negative (or zero!) real part. Indeed, it is easy to check that this is the case of formulas (2.10) used with $k = 5$ but with $\nu = 4$ instead of $\nu = (k+1)/2 = 3$.

We will assume, as is the case in practice for the methods described in section 2.2, that the influence of the small rank perturbations $E_n^{(A)} = A_n - \hat{A}_n$, $E_n^{(B)} = B_n - \hat{B}_n$ on the spectral properties of $\hat{A}_n$, $\hat{B}_n$ is moderate. More precisely, here we refer to suitably chosen additional schemes such that their related matrices A_n, B_n have the spectrum of eigenvalues in the right half plane, and the condition number of these is still of the order of $O(n)$, where n is their size. These hypotheses are usually reasonable; see [6]. On the other hand, notice that, in general, the influence of low rank modifications in the non-Hermitian case can very much change the spectral properties of a given matrix; see [22]. However, in this paper we will focus mainly on the preconditioner and on the spectrum of the component matrices of the preconditioner (1.2), which are normal and defined by using the coefficients of (2.3), i.e., by the entries of $\hat{A}_n$, $\hat{B}_n$ only.

We will need an explicit expression of the coefficients of the formulas (2.9) and (2.10). To this end, there are at least two (equivalent) strategies. In the first one, the coefficients can be computed by writing the formula of the GLM in backward difference form; see, e.g., [19, chapter 3]. Thus, expanding the backward differences of y_{n+j} for the formulas (2.9) and of f_{n+j} for the formulas (2.10), equating the coefficients of y_{n+j} and of f_{n+j}, $j = 0, \dots, k$, to the corresponding expressions, and using an induction argument gives the coefficients α_j, β_j, $j = 0, \dots, k$.

PROPOSITION 3.1. *The coefficients of the formulas* (2.9) *are given by*

$$\alpha_i = (-1)^{k-i} \sum_{j=k-i}^{k} \binom{j}{k-i} \delta_j, \qquad i = 0, \dots, k, \tag{3.3}$$

where

$$\delta_i = \begin{cases} 0, & i = 0; \\ 1, & i = 1; \\ \dfrac{1}{i!} \displaystyle\sum_{s=0}^{i-1} \prod_{\substack{j=0, \\ j \neq s}}^{i-1} (-(k-\nu)+j), & i \leq k-\nu,\ i \geq 1; \\ \dfrac{1}{i!} \displaystyle\prod_{\substack{j=0, \\ j \neq k-\nu}}^{i-1} (-(k-\nu)+j), & i > (k-\nu) \geq 1. \end{cases}$$

The coefficients of the formulas (2.10) *are given by*

$$\beta_i = \frac{(-1)^{k-i}}{(k-i)!} \sum_{j=k-i}^{k} \frac{1}{(j-(k-i))!} \int_{-(k-\nu)-1}^{-(k-\nu)} \prod_{m=0}^{j-1} (r+m)\, dr, \quad i = 0, \dots, k. \tag{3.4}$$

Proof. The expression (3.3) is derived by writing (2.9) in backward difference form (see [19, chapter 3]), i.e.,

$$\sum_{j=0}^{k} \delta_i \nabla^i y_{n+k-\nu} = h\, f_n,$$

where

$$\delta_i = (-1)^i \frac{d}{dr}\binom{-r}{i},$$

and the above derivative is computed at $r = k - \nu$. Thus, by observing that

$$\binom{-r}{j} = \frac{(-r-j+1)\cdots(-r)}{j!} = \frac{(-1)^j}{j!}\prod_{m=0}^{j-1}(r+m), \tag{3.5}$$

we have (3.3). The other expression, i.e., (3.4), is derived by observing that (2.10) can be written as (see [19, chapter 3])

$$y_{n+1} - y_n = h\sum_{j=0}^{k}\gamma_i \nabla^i f_{n+k-\nu+1}, \tag{3.6}$$

where

$$\gamma_j = (-1)^j \int_{-(k-\nu)-1}^{-(k-\nu)} \binom{-r}{j}\, dr.$$

Thus, from (3.5), we have (3.4). ∎

The other strategy is based on the explicit solution of linear equations in the unknowns α_j, β_j, $j = 0, \ldots, k$, by writing (2.4) in matrix form. Thus, we have to solve a linear system whose matrix is a Vandermonde-like one, and several combinatorial identities can be used. The coefficients of (2.9) and of (2.10) were computed following this strategy in [3]. We stress that the derivation of a useful expression can be rather lengthy. Full details can be found in [3, pp. 46–50, 66–69].

PROPOSITION 3.2. *The coefficients of the formula* (2.9) *are given by*

$$\begin{aligned}
\alpha_i &= \frac{(-1)^{\nu-i}}{\nu-i}\frac{\nu!(k-\nu)!}{i!(k-i)!}, \qquad i \neq \nu,\ i = 0, \ldots, k,\\
&= \frac{1}{\nu} = \frac{2}{k+1}, \qquad i \equiv \nu,\ k\ \textit{odd},\\
&= \frac{2\nu-1}{\nu(\nu-1)} = \frac{4(k+1)}{k(k+2)}, \qquad i \equiv \nu,\ k\ \textit{even},\ k \geq 1.
\end{aligned} \tag{3.7}$$

The coefficients of the formula (2.10) *are given by*

$$\beta_i = \frac{(-1)^{k-i}}{i!\,(k-i)!}\int_{\nu}^{\nu+1}\prod_{\substack{m=0,\\ m\neq i}}^{k}(t-m)\, dt, \qquad i = 0, \ldots, k. \tag{3.8}$$

Proof. The proof follows after some manipulations of the results in Theorem 4.1.1 for (3.7) and in Remark 4.2.2 for (3.8) in [3] by recalling that $\nu = (k+1)/2$ if k is odd, while $\nu = (k+2)/2$ for (2.9), $\nu = k/2$ for (2.10) if k is even. ∎

We remark that there is a third approach to derive the coefficients of the formulas (2.9) and (2.10) that is simpler than the other two. It is based on generating functions and symbolic operators; see, e.g., [19, sections 3.9–3.12]. Using that approach, we have that the generating function for δ_i in (3.3) is given by

$$G_1(z) = -(1-z)^{k-\nu}\log(1-z) = \sum_{i=0}^{\infty}\delta_i z^i.$$

Therefore,

$$\delta_i = \sum_{s=0}^{i-1} (-1)^s \binom{k-\nu}{s} \frac{1}{i-s}.$$

Similarly, the generating function for γ_i in (3.6) is given by

$$G_2(z) = \frac{-z(1-z)^{k-\nu}}{\log(1-z)} = \sum_{i=0}^{\infty} \gamma_i z^i,$$

and an explicit expression for γ_i can be derived accordingly.

Obviously, suitably manipulating the expressions derived by one strategy (e.g., (3.3), (3.4)) gives the expressions derived by the others (see, e.g., (3.7), (3.8), respectively).

COROLLARY 3.3. *The coefficients of the formulas* (2.9) *are uniformly bounded by* 2 *for all* $k \geq 1$. *Moreover,* $|\alpha_{i+1}| < |\alpha_i|$ *for* $i = \nu+1, \ldots, k-1$; $|\alpha_{i+1}| > |\alpha_i|$ *for* $i = 0, \ldots, \nu-2$; $|\alpha_\nu| < |\alpha_{\nu+1}|$; $|\alpha_{\nu-1}| > |\alpha_\nu|$; *and* $\lim_{k\to\infty} \alpha_j = 0$, $j = 0, k, \nu$.

Proof. By considering the expression (3.7) we have that $|\alpha_{i+1}| < |\alpha_i|$ for $i = \nu+1, \ldots, k-1$ and $|\alpha_{i+1}| > |\alpha_i|$ for $i = 0, \ldots, \nu-2$. For k odd, $\nu = (k+1)/2$, and, by (3.7), we have $\alpha_{\nu-1} = -\frac{(k+1)/2}{k-(k+1)/2+1} = -1$ while, for k even, $\nu = (k+2)/2$ and $\alpha_{\nu-1} = \frac{k+2}{k} \leq 2$ for $k \geq 2$. Similarly, for k odd, we have $\alpha_{\nu+1} = (k-1)/(k+4) < 1$, otherwise $\alpha_{\nu+1} = (k-2)/(k+4) < 1$, $k \geq 2$, and the proof is complete by recalling (3.7) again for $i = 0, \nu, k$. ∎

On the other hand, we can observe that for many families of linear multistep formulas the above results do not hold. This is the case of popular initial value methods such as the backward differentiation formula and the Adams–Moulton methods or of the schemes in section 2.2 with some choices of ν different from those suggested there. More precisely, some of the coefficients α_j and β_j, $j = 0, \ldots, k$, for the methods above, can grow boundlessly very fast for $k \to \infty$.

4. Circulant approximations for general linear multistep formulas. Let us consider the block preconditioners in (1.2) for the linear systems in (2.6) based on circulant-like matrices introduced in [4, 5, 7]. The approximating operators $\breve{A}$, $\breve{B}$ in (1.2) are computed by taking into account the coefficients of the formula (2.3), i.e., they are defined for the Toeplitz matrices $\hat{A}_n$, $\hat{B}_n$.

In what follows, we will recall in brief the main trigonometric approximations for the nonsymmetric matrices $\hat{A}_n$, $\hat{B}_n$ (and for A, B in (1.1)) we have found effective for the preconditioner (1.2); see also section 5.3. To this end, let $T_n = (t_j)$ be an $n \times n$ Toeplitz matrix whose diagonal entries are t_j, $j = -(n-1), \ldots, n-1$.

Strang's $s(T_n)$ (see [21]), sometimes called simple circulant approximation, is such that if $s_0, \ldots, s_{n-1}$ are the entries of the first row of the corresponding $n \times n$ preconditioner for T_n, we have

$$s_j = \begin{cases} t_j, & 0 < j \leq \left\lfloor \frac{n}{2} \right\rfloor, \\ \\ t_{j-n}, & \left\lfloor \frac{n}{2} \right\rfloor < j < n, \end{cases} \qquad j = 0, \ldots, n-1. \tag{4.1}$$

The spectrum of the Hermitian Toeplitz matrices preconditioned using Strang's preconditioner was analyzed in [10]. Notice that $s(T_n)$ is singular for the Toeplitz matrices

T_n whose generating function $f(\theta)$ is zero in $\theta = 0$, as observed, e.g., in [5, 24]. Unfortunately, the generating function of the matrix $\hat{A}_n$ always has a zero of multiplicity one in $\theta = 0$ because of the consistency condition $0 = \rho(1) = \sum_{j=0}^{k} \alpha_j$. Thus, as observed in [5], the approximation (4.1) cannot be safely used in the preconditioner (1.2), e.g., when the Jacobian matrix J in (2.6) has some very small or zero eigenvalues; see [5] for more details.

T. Chan's circulant preconditioner for the Toeplitz matrix T_n, denoted by $c(T_n)$, is defined such that $\|c(T_n) - T_n\|_F$ is minimum, where $c(T_n)$ is chosen in the set of $n \times n$ circulant matrices and $\|\cdot\|_F$ is the Frobenius norm. If $c_0, \ldots, c_{n-1}$ are the entries of the first row of $c(T_n)$ and t_j, $j = -(n-1), \ldots, n-1$, are the elements on the diagonals of the Toeplitz matrix T_n, we have (see [12])

$$c_j = \frac{(n-j)t_j + jt_{j-n}}{n}, \qquad j = 0, \ldots, n-1. \tag{4.2}$$

If the Toeplitz matrix T_n is Hermitian and positive definite, then these properties hold true for $c(T_n)$ as well; see [23]. Unfortunately, if T_n is nonsymmetric, $s(T_n)$ and $c(T_n)$ can have eigenvalues in the right and left half plane or zero as well, even for those matrices T_n whose eigenvalues have strictly positive real part. For example, this holds true for the underlying linear systems based on the formulas in section 2. Moreover, there are families of formulas (2.3) such that the circulant approximation (4.2) can be ill conditioned or even singular (e.g., those based on the midpoint method in boundary value form; see at the end of section 5.2).

Let us consider the P-circulant approximation introduced in [4]. Again, if T_n is a Toeplitz matrix whose entries of the diagonals are $t_{-(n-1)}, \ldots, t_{n-1}$, we have that the entries $p_0, \ldots, p_{n-1}$ of the first row of the P-circulant preconditioner $p(T_n)$ for T_n are given by

$$p_j = \frac{(n+j)t_j + jt_{j-n}}{n}, \qquad j = 0, \ldots, n-1. \tag{4.3}$$

Notice that the P-circulant and simple and T. Chan's circulants are equivalent in the sense of the linear approximation processes; see [20]. In practice, P-circulant matrices come from using the Frobenius norm weight $(n-j)/n$ for the lower and the weight $(n+j)/n$ for the upper diagonals, respectively. The circulant matrices, whose entries are defined in (4.3), have been called P-circulant in [4] because, for some classes of Toeplitz matrices (and thus for formulas (2.3)), their eigenvalues have positive real part; see section 5.2. This property can speed up the convergence process with respect to the other basic approximations described here; see section 5.3. We stress that P-circulants neither preserve symmetry (but for our purpose this is not essential) nor minimize the "distance" with the original Toeplitz matrix. More precisely, $||p(T_n) - T_n||$ is not minimized with respect to the p-norms (e.g., $p = 1, 2, \infty$) nor the Frobenius norm.

The MS-circulant approximation for T_n is given by a rank-one perturbation of the simple circulant preconditioner whose zero eigenvalue in (5.1) is set to a suitable nonzero value c for those Toeplitz matrices T_n whose generating functions have a zero. We achieved interesting results in [5] by setting $c = 1/n$ and $c = \min_r\{\mathrm{Re}(\phi_r)\} > 0$, where ϕ_r, $r = 1, \ldots, n$, are the eigenvalues of $s(T_n)$.

Finally, the $\{\omega\}$-circulant approximation can be considered as another extension of the simple circulant approximation. Let T_n be a n_1-band Toeplitz matrix, $n_1 < \lceil n/2 \rceil$. The $\{\omega\}$-circulant matrix $\tilde{s}(t_n)$ differs from the simple circulant $s(T_n)$ because

the entries outside the diagonals $-n_1, \dots, n_1$ of $\tilde{s}(t_n)$ are given by those of $s(T_n)$ multiplied by $\omega = \exp(\mathbf{i}\theta)$, $0 < \theta \leq \pi$, and $\tilde{s}(T_n)$ is nonsingular even if the generating function of T_n has a zero for $\theta = 0$; see [7].

We observe that some combinations of the above approximations can give further useful preconditioners as well. For example, it is straightforward to define $\{\omega\}$–P-circulant preconditioners by using (4.3) and $\{\omega\}$-circulant matrices instead of circulant matrices. The arguments used in the following sections can be adapted for these preconditioners as well, in general, and we will focus only on the "basic" approximations above.

5. The spectrum of the circulant approximations. The Toeplitz matrices $\hat{A}_n$, $\hat{B}_n$ in (3.1) are positive stable for the linear multistep formulas (2.9), (2.10), (2.11); see [9, chapter 11]. We recall that a square matrix is said to be (semi)positive stable if its eigenvalues have positive (nonnegative) real part; see, e.g., [16]. It is straightforward to note that positive stable matrices are nonsingular.

Let us consider $n = s + 1$ and the $(s + 1) \times (s + 1)$ P-circulant matrix $p(A)$ defined in (4.3) for the Toeplitz matrix A in (2.7) (and then for $\hat{A}_{s+1}$ in (3.1)). The eigenvalues ϕ_j, $j = 0, \dots, s$, of $p(A)$ can be computed by a linear combination of the entries of the first row (see Davis [13]):

$$\phi_l = \sum_{j=0}^{s} p_j \epsilon^{jl}, \qquad l = 0, \dots, s, \quad \epsilon = e^{2\pi\,\mathbf{i}/(s+1)}. \tag{5.1}$$

From (4.3) we have

$$\phi_l = \sum_{j=0}^{s} \alpha_{j+\nu} \left(1 + \frac{j}{s+1}\right) \epsilon^{jl} + \sum_{j=0}^{s} \left(\frac{j}{s+1} \alpha_{j+\nu-(s+1)}\right) \epsilon^{jl}, \qquad l = 0, \dots, s.$$

Therefore,

$$\phi_l = \sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} \left(1 + \frac{j}{s+1}\right) \epsilon^{jl}, \qquad l = 0, \dots, s. \tag{5.2}$$

A similar expression holds for the eigenvalues $\psi_0, \dots, \psi_s$ of $p(B)$:

$$\psi_l = \sum_{j=-\nu}^{k-\nu} \beta_{j+\nu} \left(1 + \frac{j}{s+1}\right) \epsilon^{jl}, \qquad l = 0, \dots, s. \tag{5.3}$$

Notice that (5.2) and (5.3) are trigonometric sums. Let us define

$$\hat{\Phi}_k(x) = \sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} \left(1 + \frac{j}{s+1}\right) \cos(j\,x), \qquad x \in \mathbb{R}, \tag{5.4}$$

$$\hat{\Psi}_k(x) = \sum_{j=-\nu}^{k-\nu} \beta_{j+\nu} \left(1 + \frac{j}{s+1}\right) \cos(j\,x), \qquad x \in \mathbb{R}. \tag{5.5}$$

We observe that (5.4) and (5.5) are analytic functions (for $k < \infty$). From (5.2), we have that

$$\hat{\Phi}_k\left(\frac{2\pi l}{s+1}\right) = \mathrm{Re}(\phi_l), \quad \hat{\Psi}_k\left(\frac{2\pi l}{s+1}\right) = \mathrm{Re}(\psi_l), \qquad l = 0, \dots, s.$$

Thus, it is straightforward to see that if $\hat{\Phi}_k(x)$, $\hat{\Psi}_k(x)$ are positive for real values of x, then $p(A)$ and $p(B)$ are positive stable.

By using similar arguments, we can derive the expression of the eigenvalues of $s(A_{s+1})$, $s(B_{s+1})$ and $c(A_{s+1})$, $c(B_{s+1})$, respectively:

$$\gamma_l = \sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu}\epsilon^{jl}, \qquad \delta_l = \sum_{j=-\nu}^{k-\nu} \beta_{j+\nu}\epsilon^{jl}, \qquad l = 0, \ldots, s, \tag{5.6}$$

and

$$\sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu}\left(1 - \frac{|j|}{s+1}\right)\epsilon^{jl}, \qquad \sum_{j=-\nu}^{k-\nu} \beta_{j+\nu}\left(1 - \frac{|j|}{s+1}\right)\epsilon^{jl}, \qquad l = 0, \ldots, s. \tag{5.7}$$

Notice that the eigenvalues of $s(A_{s+1})$, $s(B_{s+1})$ lie on the boundary locus of A_{s+1}, B_{s+1}, respectively.

5.1. Preliminary results. First, let us give some properties of the trigonometric sums (5.4), (5.5).

We recall that a sequence $\{c_j\}$ is of bounded variation (see [25]) if the series $\sum_{j=0}^{\infty} |c_{j+1} - c_j|$ converges. If $\{c_j\}$ tends monotonically to zero, then $\{c_j\}$ is of bounded variation. It is useful to simplify the expressions (5.4) and (5.5) by observing that $\cos(x)$ is an even function.

LEMMA 5.1. *The function* $\hat{\Phi}_k(x)$ *in* (5.4) *can be expressed for* (2.9) *as*

$$\hat{\Phi}_k(x) = \frac{a_0}{2} + \sum_{n=1}^{k-\nu} (-1)^n a_n \cos(n\,x), \tag{5.8}$$

where $a_0 = 2\alpha_\nu$, $\nu = (k+1)/2$ *if* k *is odd,* $\nu = (k+2)/2$ *if* k *is even, and* $a_n = (-1)^n \tilde{a}_n$,

$$\tilde{a}_n = \alpha_{n+\nu}\left(1 + \frac{n}{s+1}\right) + \alpha_{-n+\nu}\left(1 - \frac{n}{s+1}\right), \qquad n = 1, \ldots, k-\nu. \tag{5.9}$$

It is intended that α_j *is zero if* $j < 0$ *or* $j > k$. *The sequence* $\{a_n\}$ *has the following properties:*

(1) $a_n \geq 0$, $n \geq 0$;
(2) a_n *tends to zero if* $n \to \infty$;
(3) a_n *is uniformly bounded (i.e.,* $0 \leq a_n < 2$, $n \geq 0$*);*
(4) $\{a_n\}$ *is monotonic decreasing;*
(5) $\{a_n\}$ *is of bounded variation.*

Proof. The expression (5.8) follows by observing that from (3.7), (5.4), and (5.9) we have

$$\tilde{a}_n = \frac{(-1)^n}{n}\left\{ -\frac{\binom{k}{\nu+n}}{\binom{k}{\nu}}\left(1 + \frac{n}{s+1}\right) + \frac{\binom{k}{\nu-n}}{\binom{k}{\nu}}\left(1 - \frac{n}{s+1}\right)\right\} = (-1)^n \cdot a_n.$$

(1) Let us check for first that $a_n > 0$ for $n \geq 1$, $n \leq k - \nu$ (recall that $a_0 = 2\alpha_\nu$ is positive; see (3.7)). From here on, it is intended that $a_n = 0$ if $n > k - \nu$. Again,

from the expression (3.7), we have

$$
\begin{aligned}
n\cdot a_n &= \frac{\binom{k}{\nu-n}}{\binom{k}{\nu}}\left(1-\frac{n}{s+1}\right)-\frac{\binom{k}{\nu+n}}{\binom{k}{\nu}}\left(1+\frac{n}{s+1}\right)\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\cdot\left(1-\frac{n}{s+1}\right)\right]-\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-n}{\nu+n}\cdot\left(1+\frac{n}{s+1}\right)\right]\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\left[\left(1-\frac{n}{s+1}\right)-\frac{\nu-n}{\nu+n}\left(1+\frac{n}{s+1}\right)\right]\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\frac{2n}{(\nu+n)(s+1)}\cdot(s+1-\nu)>0.
\end{aligned}
$$

Indeed, notice that the term in square brackets above can assume values in $(0,1)$, and $(s+1-\nu)$ is greater than zero because $s\geq k\geq\nu\geq 1$ by hypothesis; see section 2.2.

(2) Now, let us check that a_n converges to zero for $n\to\infty$. From the last expression, we have

$$
\begin{aligned}
a_n &= \left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\frac{2(s+1-\nu)}{(\nu+n)(s+1)}\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\frac{2}{\nu+n}\cdot\left(1-\frac{\nu}{s+1}\right)\\
&\leq\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\frac{2}{n}\leq\frac{2}{n} \qquad (5.10)
\end{aligned}
$$

because the term in square brackets above assumes values in $(0,1)$, $n\leq\nu$, and $0<1-\nu/(s+1)<1$ because $s\geq k\geq\nu\geq 1$.

(3) It is an immediate consequence of the bound in (5.10).

(4) $\{a_n\}$ is monotonic (decreasing). Indeed,

$$
\begin{aligned}
&a_{n+1}-a_n\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot\left[\frac{2\nu(\nu-n)}{(\nu+n+1)(s+1)(\nu+n)}-\frac{2\nu}{(\nu+n)(s+1)}\right]\\
&=\left[\frac{\nu-1}{\nu+1}\cdot\ldots\cdot\frac{\nu-(n-1)}{\nu+(n-1)}\right]\cdot-\frac{2\nu}{s+1}\cdot\frac{2n+1}{(\nu+n+1)(\nu+n)}<0
\end{aligned}
$$

by using similar arguments as in (1) and (2).

(5) Finally, for (1)–(4), $\{a_n\}$ is of bounded variation. □

We recall that a sequence of functions is said to converge locally uniformly on a set $\mathcal{S}$ if it converges uniformly on every compact subset of $\mathcal{S}$; see, e.g., [17, p. 160].

PROPOSITION 5.2. *The sequence of functions $\{\hat{\Phi}_k(x)\}$ for (2.9) converges locally uniformly with respect to k (and then with respect to $\nu=O(k)$, see section 2) in $(-\pi,\pi)$.*

Proof. It is a consequence of Lemma 5.1 and of [25, Theorem 2.7, p. 4]. □

COROLLARY 5.3. *Under the hypotheses of Proposition 5.2, the function*

$$\hat{\Phi}(x)=\lim_{k\to\infty}\hat{\Phi}_k(x) \qquad (5.11)$$

is an analytic function for $x\in(-\pi,\pi)$ and continuous for $x\in\mathbb{R}$.

Proof. It is a consequence of Proposition 5.2 and of [17, Corollary 3.4c, p. 161]. The continuity over the whole real axis derives from Abel's limit theorem applied in $x = \pm\pi$ and considering that, for n integer, we have

$$\hat{\Phi}_k(\pm x + 2n\pi) = \hat{\Phi}_k(x), \qquad \hat{\Phi}_k(2\pi - x)' = -\hat{\Phi}_k(x)', \tag{5.12}$$

$$\hat{\Psi}_k(\pm x + 2n\pi) = \hat{\Psi}_k(x), \qquad \hat{\Psi}_k(2\pi - x)' = -\hat{\Psi}_k(x)', \qquad x \in [0, 2\pi],\ k \geq 1. \tag{5.13}$$

Similar expressions hold true for $\hat{\Phi}(x)$ and $\hat{\Psi}(x)$. □

LEMMA 5.4. *Let k be an integer and $\nu = \lceil (k+1)/2 \rceil$ (k even or odd) or $\nu = (k+2)/2$ (k even). For $-(k-\nu) \leq n \leq (k-\nu)$, we have*

$$(\nu + n)!\,(k - \nu - n)! \geq (\nu - n)!\,(k - \nu + n)!, \tag{5.14}$$

$$\nu!\,(k-\nu)! \leq (\nu + n)!\,(k - \nu - n)!. \tag{5.15}$$

Proof. Let us consider (5.14). We have

$$\frac{(\nu-n)!(k-\nu+n)!}{(\nu+n)!(k-\nu-n)!} = \frac{(\nu-n)!(k-(\nu-n))!}{k!} \cdot \frac{k!}{(\nu+n)!(k-(\nu+n))!} = \frac{\binom{k}{\nu+n}}{\binom{k}{\nu-n}},$$

where the ratio above is equal to 1 if k is even and it is less than 1 otherwise. Indeed,

$$\binom{k}{\nu-n} = \binom{k}{k-(\nu-n)} = \begin{cases} \binom{k}{\nu+n} & \text{if } k-\nu = \nu \text{ (and } k \text{ is even)}, \\ \binom{k}{\nu+n-1} & \text{if } k-\nu = \nu-1 \text{ (and } k \text{ is odd)}\ . \end{cases}$$

Thus, for k odd, we have

$$\binom{k}{\nu+n-1} = \binom{k+1}{\nu+n} - \binom{k}{\nu+n} \Rightarrow \frac{\binom{k}{\nu+n}}{\binom{k}{\nu-n}} = \frac{1}{\frac{\nu+n-1}{\nu-n}} < 1, \quad 1 \leq n \leq (k-\nu).$$

Using similar arguments, we can see that (5.14) holds true for $\nu = (k+2)/2$ and k even as well. Now, let us consider (5.15) for $n \geq 1$ (for negative values of n a similar argument can be used). We have

$$\frac{\nu!\,(k-\nu)!}{(\nu+n)!\,(k-\nu-n)!} = \frac{k-\nu}{\nu+1} \cdot \frac{k-\nu-1}{\nu+2} \cdots \frac{k-\nu-(n-1)}{\nu+n},$$

where we have that the above expression is equal to

$$\begin{cases} \dfrac{\nu-1}{\nu+1} \cdots \dfrac{\nu-n}{\nu+n} < 1 & \text{if } k-\nu = \nu \text{ and } k \text{ is odd}, \\ \dfrac{\nu}{\nu+1} \cdots \dfrac{\nu-(n-1)}{\nu+n} < 1 & \text{if } k-\nu = \nu-1 \text{ and } k \text{ is even}\ . \end{cases}$$

Using similar arguments, we can see that (5.15) holds true for $\nu = (k+2)/2$ and k even as well. □

LEMMA 5.5. *The function* $\hat{\Psi}_k(x)$ *can be expressed for* (2.10) *as*

$$\hat{\Psi}_k(x) = \frac{b_0}{2} + \sum_{n=1}^{k-\nu} (-1)^{n+1} b_n \cos(n\,x), \tag{5.16}$$

where $b_0 = 2\beta_\nu$, $\nu = \lceil (k+1)/2 \rceil$, *and* $b_n = (-1)^{n+1}\tilde{b}_n$, $n \geq 1$,

$$\tilde{b}_n = \beta_{n+\nu}\left(1 + \frac{n}{s+1}\right) + \beta_{-n+\nu}\left(1 - \frac{n}{s+1}\right), \qquad n = 1, \ldots, k-\nu. \tag{5.17}$$

It is intended that β_j *is zero if* $j < 0$ *or* $j > k$. *The sequence* $\{b_n\}$ *has the following properties:*

(1) $b_n \geq 0$, $n \geq 0$;
(2) b_n *tends to zero if* $n \to \infty$;
(3) b_n *is uniformly bounded (i.e.,* $0 \leq b_n < 2$, $n \geq 0$*);*
(4) $\{b_n\}$ *is monotonic decreasing;*
(5) $\{b_n\}$ *is of bounded variation.*

Proof. (1) By expanding (5.17), we have

$$\begin{aligned}\tilde{b}_n = &\left[\frac{(-1)^{k-\nu-n}}{(\nu+n)!(k-\nu-n)!}\int_\nu^{\nu+1} \prod_{\substack{m=0,\\ m\neq\nu+n}}^{k} (t-m)\,dt\right]\left(1 + \frac{n}{s+1}\right)\\ &+ \left[\frac{(-1)^{k-\nu+n}}{(\nu-n)!(k-\nu+n)!}\int_\nu^{\nu+1} \prod_{\substack{m=0,\\ m\neq\nu-n}}^{k} (t-m)\,dt\right]\left(1 - \frac{n}{s+1}\right).\end{aligned} \tag{5.18}$$

Thus, for $n \geq 1$, we have

$$\begin{aligned}\tilde{b}_n = (-1)^{n+1}\int_\nu^{\nu+1} &\left[\frac{1+n/(s+1)}{(\nu+n)!(k-\nu-n)!}\frac{1}{|t-\nu-n|}\right.\\ &\left. - \frac{1-n/(s+1)}{(\nu-n)!(k-\nu+n)!}\frac{1}{(t-\nu+n)}\right]\prod_{m=0}^{k}|t-m|\,dt = (-1)^{n+1}\cdot b_n,\end{aligned} \tag{5.19}$$

while, by observing that $t-m$, $m = 0, \ldots, k$, do not change sign for $t \in (\nu, \nu+1)$,

$$\tilde{b}_0 \equiv b_0 = 2\beta_\nu = \frac{2(-1)^{k-\nu}}{\nu!(k-\nu)!}\int_\nu^{\nu+1}\prod_{\substack{m=0,\\ m\neq\nu}}^{k}(t-m)\,dt = \frac{2}{\nu!(k-\nu)!}\int_\nu^{\nu+1}\prod_{\substack{m=0,\\ m\neq\nu}}^{k}|t-m|\,dt, \tag{5.20}$$

therefore (5.20) is positive. To check that $b_n > 0$, $n \geq 1$, and $n \leq k-\nu$ (it is intended that $b_n = 0$ for $n > k-\nu$), it is enough to see that the part in square bracket in (5.19) is positive or zero. For brevity, let us consider k even. By Lemma 5.4, the part in square brackets in (5.19) can be rewritten as

$$\frac{1}{(\nu+n)!(k-\nu-n)!}\left(\frac{1+n/(s+1)}{n-(t-\nu)} - \frac{1-n/(s+1)}{n+(t-\nu)}\right)$$

$$= \frac{(s+1+n)\,(n+(t-\nu)) - (s+1-n)\,(n-(t-\nu))}{(\nu+n)!\,(k-\nu-n)!\,(n-(t-\nu))\,(n+(t-\nu))\,(s+1)}.$$

Then, we can observe that the ratio above is positive because the denominator of the related expression is positive, $s+1+n>0$, $n+(t-\nu)>0$, and

$$\frac{(s+1-n)(n-(t-\nu))}{(s+1+n)(n+(t-\nu))} < 1, \quad n \geq 1, \quad 0 \leq t \leq \nu+1.$$

For k odd a similar argument can be used, and (1) and the expression (5.16) are verified.

(2) Let us check first that b_0 is bounded. We observe that, for $t = t^* = \nu + \epsilon(\nu)$, $0 < \epsilon(\nu) \to 0$ for $k, \nu \to \infty$ (recall that $\nu = O(k)$), the following function

$$f(t) = \prod_{\substack{m=0,\\ m\neq\nu}}^{k} |t-m|, \qquad \nu \leq t \leq (\nu+1), \tag{5.21}$$

reaches its (unique) maximum in the segment $\nu \leq t \leq (\nu+1)$. This can be checked by considering the derivative df/dt of (5.21) in $(\nu, \nu+1)$ and applying an induction argument on k. Thus, $f(t^*) = c\cdot\nu!\,(k-\nu)!$, where $c = c(\nu)$ is a parameter of the order of 1 that converges fast to 1 as $\nu \to \infty$ and, by (5.20), b_0 is uniformly bounded above by 2. As a corollary of the above result, by (3.8), β_ν is uniformly bounded above by 1. Now, to check that b_n is bounded for $n \geq 1$, it is enough to observe that both the following factors in (5.19)

$$\frac{1}{(\nu+n)!(k-\nu-n)!}\int_{\nu}^{\nu+1}\prod_{m=0}^{k}|t-m|\,dt, \quad \frac{1}{(\nu-n)!(k-\nu+n)!}\int_{\nu}^{\nu+1}\prod_{m=0}^{k}|t-m|\,dt$$

are positive and bounded by a constant of the order of unity. To this end, notice that

$$G(t) = \prod_{m=0}^{k}(t-m) = \begin{cases} (-1)^{k+1}\cdot\dfrac{\Gamma(k+1-t)}{\Gamma(-t)}, & \nu < t < \nu+1, \\ 0, & t = \nu \text{ or } t = \nu+1, \end{cases} \tag{5.22}$$

where $\Gamma(z)$ is the Gamma function (see [1] for definitions and some properties). The equality (5.22) can be derived by using arguments in [1, pp. 12–13]. It is straightforward to observe that

$$\int_{\nu}^{\nu+1}\prod_{m=0}^{k}|t-m| \leq \sup_{\nu<t<\nu+1}\prod_{m=0}^{k}|t-m|.$$

Let us denote by t^* the maximum of the function $|G(t)|$ in $\nu \leq t \leq \nu+1$; $G(t)$ is defined in (5.22). By considering dG/dt in $(\nu, \nu+1)$ and using an induction argument on k, we have that $t^* = \nu + 1/2 + \epsilon(\nu)$, where $\epsilon(\nu) \to 0$ as $\nu \to \infty$ ($k \to \infty$). By using the definition of $\Gamma(z)$, we have $\Gamma(x+1) = x\,\Gamma(x) \Rightarrow \Gamma(x) = \Gamma(x+1)/x$, where x cannot be a negative integer or zero. Applying repeatedly $\Gamma(x) = \Gamma(x+1)/x$, we have

$$|\Gamma(-t)|^{-1} \leq \nu!\,(t^*-\nu)\,\frac{1}{|\Gamma(-t^*+\nu+1)|}. \tag{5.23}$$

By observing that

$$(k-\nu-1)! < \Gamma(k+1-t^*) < (k-\nu)!,\ 0 < c \equiv \Gamma(-t^*+\nu+1) < 1,\ 1 < 1/(t^*-\nu) < 2$$

and by recalling Lemma 5.4, we can write

$$|G(t)| \le c \cdot (t^*-\nu)^{-1} \cdot \nu! \cdot (k-\nu)!;$$

thus, b_n is bounded above by 2.

(3) To check that $\{b_n\}$ converges to zero as $n \to \infty$, arguments similar to those used to prove (1), (2) give that b_n in (5.19) can be written as $b_n = \frac{1}{n} \cdot h_n$, where $\{h_n\}$ is a uniformly bounded sequence.

(4) To check that $\{b_n\}$ is a monotonic nonnegative decreasing sequence, we observe that the expression (5.19) for $n > 1$ and for k even gives

$$\begin{aligned} b_n - b_{n+1} = \frac{1}{(\nu+n)!(k-\nu-n)!} \int_\nu^{\nu+1} \left\{ \left[\frac{1+n/(s+1)}{n-(t-\nu)} - \frac{1-n/(s+1)}{n+(t-\nu)} \right] \right. \\ \left. - \left[\frac{1+(n+1)/(s+1)}{(n+1)-(t-\nu)} - \frac{1-(n+1)/(s+1)}{(n+1)+(t-\nu)} \right] \right\} \prod_{m=0}^{k} |t-m|\, dt. \end{aligned} \tag{5.24}$$

Let us consider the expression in curly brackets in (5.24). We have the following lower bound:

$$\begin{aligned} \{\cdot\} &> \frac{1+n/(s+1)}{n-1} - \frac{1+(n+1)/(s+1)}{n+1} - \frac{1-n/(s+1)}{n-1} + \frac{1-(n+1)/(s+1)}{n+1} \\ &= \frac{2n}{(s+1)(n-1)} - \frac{2(n+1)}{(s+1)(n+1)} = \frac{2}{s+1}\left(\frac{n}{n-1} - 1\right) > 0, \end{aligned} \tag{5.25}$$

and thus, by (5.24), $b_n - b_{n+1} > 0$ and the sequence $\{b_n\}$ is monotonic decreasing.

(5) For (1)–(4), $\{b_n\}$ is of bounded variation. ∎

Finally, by using similar arguments such as in the Proposition 5.2 and Corollary 5.3, we have the following results.

PROPOSITION 5.6. *The sequence of functions* $\{\hat{\Psi}_k(x)\}$ *for* (2.10) *converges locally uniformly with respect to* k *(and thus with respect to* $\nu = O(k)$*) in* $(-\pi, \pi)$.

Proof. It is a consequence of Lemma 5.5 and of [25, Theorem 2.7, p. 4]. ∎

COROLLARY 5.7. *Under the hypotheses of Proposition* 5.6, *the function*

$$\hat{\Psi}(x) = \lim_{k\to\infty} \hat{\Psi}_k(x) \tag{5.26}$$

is analytic for $x \in (-\pi, \pi)$ *and continuous for* $x \in \mathbb{R}$. ∎

5.2. Main results. As a consequence of the results in the previous section, we can give bounds for the eigenvalues for some of the underlying approximations.

THEOREM 5.8. *The P-circulant matrices* $p(A_{s+1})$, $p(B_{s+1})$ *related to the formulas* (2.9), (2.10) *are positive stable and, if* ϕ_j, ψ_j, $j = 0, \dots, s$, *are the eigenvalues of* $p(A_{s+1})$, $p(B_{s+1})$, *respectively, we have*

$$\frac{1}{s+1} \le \mathrm{Re}(\phi_j) < 2, \qquad \mathrm{Re}(\psi_j) = 1 \text{ for } (2.9), \tag{5.27}$$

$$\frac{1}{s+1} \le \mathrm{Re}(\phi_j) < \frac{2s+1}{s+1} < 2, \qquad \frac{2}{\pi^2(s+1)} < \mathrm{Re}(\psi_j) < 1 \ \text{ for } (2.10). \tag{5.28}$$

Proof. Let us observe that, from (4.3) and (5.2), considering the scaling and consistency conditions $\rho(1) = 0$, $\sigma(1) = 1$, we have

$$\sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} = 0, \quad \sum_{j=-\nu}^{k-\nu} j\alpha_{j+\nu} = 1. \tag{5.29}$$

Thus, the expression (5.2) gives $\phi_0 = \hat{\Phi}(0) = 1/(s+1)$ for the formulas (2.3).

Let us check (5.27). To this end, we consider formulas (2.9) and expand $\cos(jx)$ in the right-hand side of (5.4) using power series in a neighborhood of the origin. If $\mathcal{P}(f)$ is the formal power series expansion of a function f, we can write

$$\mathcal{P}\left(\hat{\Phi}_k(x)\right) = \frac{1}{s+1} + \sum_{n=1}^{\infty} \left\{ (-1)^n \frac{x^{2n}}{(2n)!} \sum_{j=-\nu}^{k-\nu} j^{2n} \alpha_{j+\nu} \left(1 + \frac{j}{s+1}\right) \right\}. \tag{5.30}$$

For brevity, we consider k odd ($\Rightarrow \nu = (k+1)/2$). From (2.13), it is worth noting that (5.30) is equivalent to the following expression:

$$\mathcal{P}\left(\hat{\Phi}_k(x)\right) = \frac{1}{s+1} + \sum_{n=\nu}^{\infty} \left\{ (-1)^n \frac{x^{2n}}{(2n)!} \sum_{j=-\nu}^{k-\nu} j^{2n} \alpha_{j+\nu} \left(1 + \frac{j}{s+1}\right) \right\}. \tag{5.31}$$

However, in Proposition 5.2, we observed that $\{\hat{\Phi}_k(x)\}$ converges locally uniformly in $\mathcal{S} = (-\pi, \pi)$ with respect to k (i.e., to ν because $k = 2\nu - 1$) for (2.9). Moreover, the functions $f_n(x) = (-1)^n a_n \cos(n\,x)$ in (5.8) are analytic in $\mathcal{S}$. Then, by [17, Corollary 3.4c, p. 161], we have that $\sum f_n(x)$, i.e., $\hat{\Phi}(x)$ in (5.11), is analytic in $\mathcal{S}$ and that the sequence $\{\hat{\Phi}_k(x)\}$ converges in $\mathcal{S}$ and the series related to (5.30) (and (5.31)) is the Taylor series of $\hat{\Phi}(x)$. However, by Abel's limit theorem for the power expansions, we have that the Taylor expansion in (5.30) (and (5.31)) converges for $x = \pm\pi$ as well,

$$\hat{\Phi}(x) := \lim_{k\to+\infty} \hat{\Phi}_k(x) = \lim_{k\to+\infty} \mathcal{P}(\hat{\Phi}_k(x)), \qquad x \in [-\pi, \pi], \tag{5.32}$$

and thus, by (5.12), for $x \in \mathbb{R}$. To conclude the first part of the proof, we observe that the quantity in the curly brackets in (5.31) is positive for $n = \nu$ (i.e., the first term of the sum), vanishes fast for $k \to \infty$ (recall (2.4) and (2.13)), and we can see that (see Figure 5.1, right)

$$\frac{1}{s+1} \le \Phi_k(x) \le \Phi_1(x) \le 2, \qquad k \ge 1, \qquad -\pi \le x \le \pi.$$

Let us now check (5.28). We can expand $\cos(jx)$ in the expression (5.5) in Taylor series in a neighborhood of the origin, and, for $k < \infty$, we have

$$\hat{\Psi}_k(x) = \sum_{n=0}^{\infty} \left\{ (-1)^n \frac{x^{2n}}{(2n)!} \sum_{j=-\nu}^{k-\nu} j^{2n} \beta_{j+\nu} \left(1 + \frac{j}{s+1}\right) \right\}. \tag{5.33}$$

By considering the order conditions (2.14), recalling the power series expansion of $\sin(x)$ and of $\cos(x)$ in a neighborhood of the origin and arguments similar to those

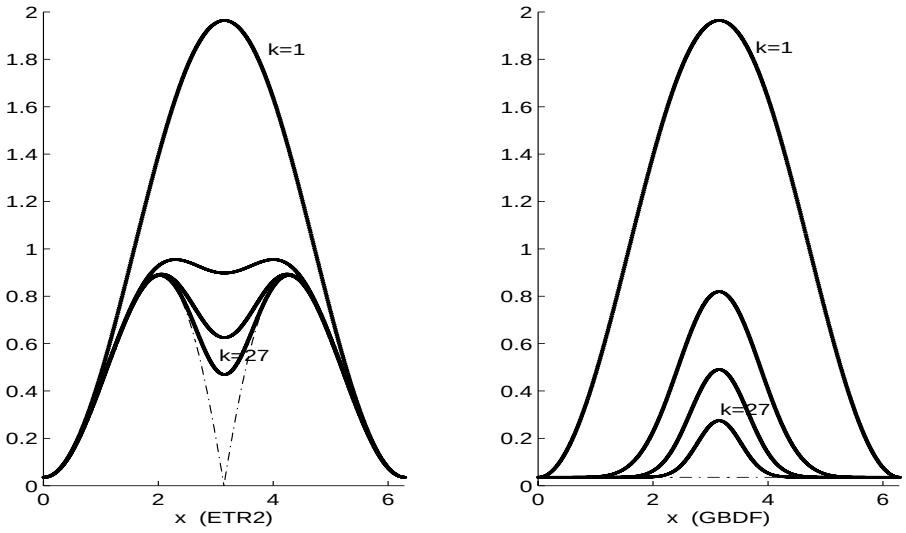


FIG. 5.1. $\hat{\Phi}_k(x)$, $k = 1, 7, 15, 27$, $s = 30$ *for formula* (2.11) *(left) and for formula* (2.9) *(right). The dashed curves give* $\hat{\Phi}(x)$.

used to prove (5.27), we have that, for $k \to \infty$ (i.e., $\nu \to \infty$ because $\nu = O(k)$),

$$\lim_{k\to\infty} \hat{\Psi}_k(x) = \hat{\Psi}(x) = \begin{cases} \dfrac{\sin(x)}{x} - \dfrac{1}{s+1}\left(\dfrac{\sin(x)}{x} - \dfrac{1}{2}\dfrac{\sin^2(x/2)}{(x/2)^2}\right), \\ \quad x \in (-\pi, 0) \cup (0, \pi), \\ \\ 1 - \dfrac{1}{2(s+1)}, \qquad x = 0. \end{cases}$$

Thus, by using similar arguments as before, the expressions (5.12), (5.13), and Abel's limit theorem, we see that the following inequalities hold true for (2.10):

$$\hat{\Psi}_k(x) \geq \hat{\Psi}(\pi) = \frac{2}{\pi^2(s+1)} > 0, \qquad x \in \mathbb{R},$$

$$\frac{1}{s+1} \leq \hat{\Phi}_k(x) \equiv \hat{\Phi}_1(x) \leq \frac{2s+1}{s+1} < 2, \qquad x \in \mathbb{R},\ k \geq 1.$$

The behavior of $\hat{\Psi}_k(x)$ for some values of k is displayed in Figure 5.2. □

We observe that the imaginary parts of the eigenvalues of the circulant approximations (4.1) and (4.3) for the matrices A and B in (2.6) for the formulas (2.9) and (2.10) are uniformly bounded by constants of the order of unity.

THEOREM 5.9. *If* ϕ_j, ψ_j, $j = 0, \ldots, s$, *are the eigenvalues of* $p(A_{s+1})$, $p(B_{s+1})$, *respectively, we have*

$$-\pi < \mathrm{Im}(\phi_j) < \pi, \qquad \mathrm{Im}(\psi_j) = 0 \qquad \textit{for } (2.9), \tag{5.34}$$

$$-\frac{s}{s+1} \leq \mathrm{Im}(\phi_j) < \frac{s}{s+1}, \qquad -c < \mathrm{Im}(\psi_j) < c \qquad \textit{for } (2.10),\ j = 0, \ldots, s, \tag{5.35}$$

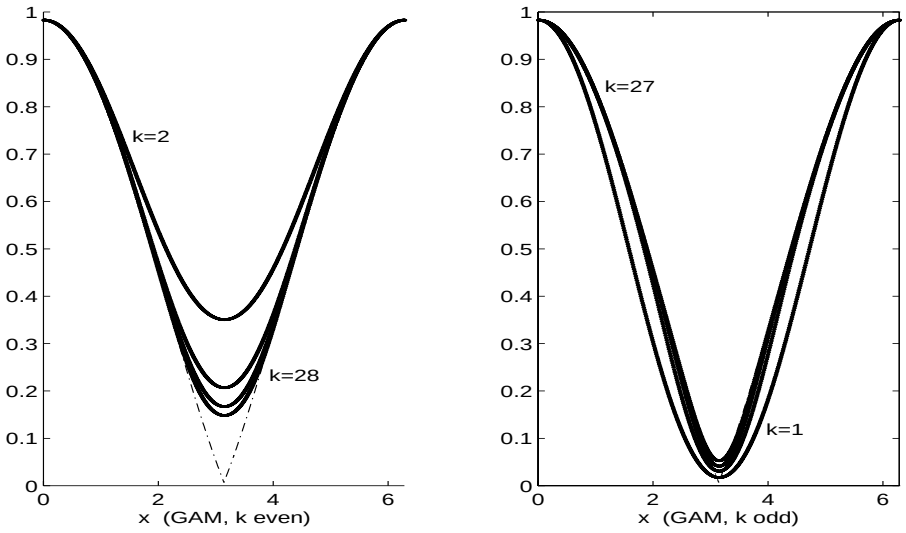


FIG. 5.2. *Left:* $\hat{\Psi}_k(x)$, $k = 2, 6, 14, 28$ *(k even); right:* $\hat{\Psi}_k(x)$, $k = 1, 7, 15, 27$ *(k odd) for formula* (2.10), $s = 28$. *The dashed curve gives* $\hat{\Psi}(x)$.

where

$$c = \max_{0<x<\pi} \left| \frac{\cos(x) - 1}{x} \right| \quad (\Rightarrow 0.7246 < c < 0.7247).$$

Proof. The proof uses arguments similar to those in the proof of Theorem 5.8. □

Again, as a corollary of Theorem 5.8, we have the following results.

THEOREM 5.10. *The preconditioners* $s(A_{s+1})$, $s(B_{s+1})$ *defined in* (4.1) *and related to the formulas* (2.9), (2.10) *are semipositive stable. If* γ_j, δ_j, $j = 0, \ldots, s$, *are the eigenvalues of* $s(A_{s+1})$, $s(B_{s+1})$, *respectively, we have*

$$0 \le \mathrm{Re}(\gamma_j) \le 2, \qquad \mathrm{Re}(\delta_j) = 1 \qquad \textit{for (2.9)}, \tag{5.36}$$

$$0 \le \mathrm{Re}(\gamma_j) \le 2, \qquad 0 \le \mathrm{Re}(\delta_j) \le 1 \qquad \textit{for (2.10)}, \ j = 0, \ldots, s. \tag{5.37}$$

THEOREM 5.11. *The* $\{\omega\}$*-circulant preconditioners* $\tilde{s}(A_{s+1})$, $\tilde{s}(B_{s+1})$ *for* $\omega = \exp(\mathbf{i}\theta)$, $0 < \theta \le \pi$, *and the MS-circulant preconditioners defined in section* 4, *related to the formulas* (2.9), (2.10), *are positive stable. If* $\tilde{\gamma}_j$, $\tilde{\delta}_j$, $j = 0, \ldots, s$, *are the eigenvalues of* $\tilde{s}(A_{s+1})$, $\tilde{s}(B_{s+1})$, *respectively, we have*

$$0 < \mathrm{Re}(\tilde{\gamma}_j) \le 2, \qquad \mathrm{Re}(\tilde{\delta}_j) = 1 \qquad \textit{for (2.9)}, \tag{5.38}$$

$$0 < \mathrm{Re}(\tilde{\gamma}_j) \le 2, \qquad 0 \le \mathrm{Re}(\tilde{\delta}_j) \le 1 \qquad \textit{for (2.10)}, \ j = 0, \ldots, s. \tag{5.39}$$

The bounds (5.38), (5.39) *hold true for the eigenvalues of the MS-circulant approximations as well.*

Similarly, it is straightforward to derive a result analogous to Theorem 5.9 for simple $\{\omega\}$-circulant and MS-circulant preconditioners by using the results in [7, 5].

It is worth noting that Theorems 5.8 and 5.10 can give results beyond linear algebra. The following corollary suggests a proof for the A-stability of formulas (2.9) using different tools, shorter than in [3, pp. 50–65].

COROLLARY 5.12. *The formulas* (2.9)*, used in boundary value form with* ν *initial and* $k-\nu$ *final conditions, are A-stable.*

Proof. As observed in [9], a linear multistep formula used in boundary value form is A-stable if its boundary locus is in the right half plane. In fact, the expression of the real part of the boundary locus of the formulas (2.9) is given by

$$\sum_{j=-\nu}^{k-\nu} \alpha_{j+\nu} \cos(j\,x), \qquad x \in \mathbb{R},\ k \geq 1;$$

see (5.6). Thus, by using the bound (5.36), we have that the boundary locus of formulas (2.9) is in the right half plane. ∎

Notice that the condition number of the underlying P-circulant approximations has a favorable behavior, e.g., for the methods based on formulas (2.9) and (2.10).

COROLLARY 5.13. *Consider the sequences* $\{K_2(p(A_{s+1}))\}$, $\{K_2(p(B_{s+1}))\}$. *We have that*

$$K_2(p(A_{s+1})) < (s+1)\sqrt{\pi^2+1}, \qquad K_2(p(B_{s+1})) = 1 \qquad \textit{for } (2.9),$$

and

$$K_2(p(A_{s+1})) < 2(s+1), \qquad K_2(p(B_{s+1})) < (s+1)\frac{\pi^2}{2} \qquad \textit{for } (2.10),$$

where $K_2(\cdot)$ *is the* 2*-norm condition number.*

Proof. The proof follows from Theorems 5.8 and 5.9 by considering that circulant matrices are normal (see [13]), and thus the singular values are given by the modulus of the eigenvalues. ∎

We observe that the bounds in the Corollary 5.13 could be not very tight for all values of k and s. However, for our purposes, it is enough to stress the linear dependence of the condition number from the size of the underlying matrices. Again, recall that $K_2(A_{s+1}) = O(s)$ and $K_2(B_{s+1}) = O(s)$ as well; see [6].

On the other hand, we cannot give an upper bound for $K_2(p(A_{s+1}))$ for the matrices related to the formulas in (2.11). Indeed, applying arguments similar to those used in the proofs of Theorem 5.8 and of Corollary 5.13, we would have $\hat{\Phi}(n\,\pi) = 0$, $n \neq 0$ integer (see Figure 5.1, left). Therefore, $K_2(p(A_{s+1}))$ cannot be bounded, and we have not considered in detail formulas (2.11). Moreover, we experienced that the methods based on formulas (2.9) and (2.10) can perform better than those based on (2.11) with the underlying preconditioners. For example, less preconditioned iterations are often required to solve the linear systems (2.6) for (2.9) and (2.10).

Notice that, for several families of non-Hermitian Toeplitz matrices, the real parts of the eigenvalues of their circulant approximations can be positive, negative, or zero even when the nonpreconditioned matrix is positive stable. For example, this is the case of backward differentiation formulas, Adams–Moulton methods, formulas in section 2.2 for choices of ν different from those suggested there. Moreover, for non-Hermitian matrices, the circulant approximation in (4.2) may give ill conditioned preconditioners as well. For example, the condition number of $c(A_{s+1})$ can grow fast with k, e.g., for the families of k-step formulas in section 2.2; see Figure 5.3. Moreover, the block preconditioners using the approximations (4.2) can be singular

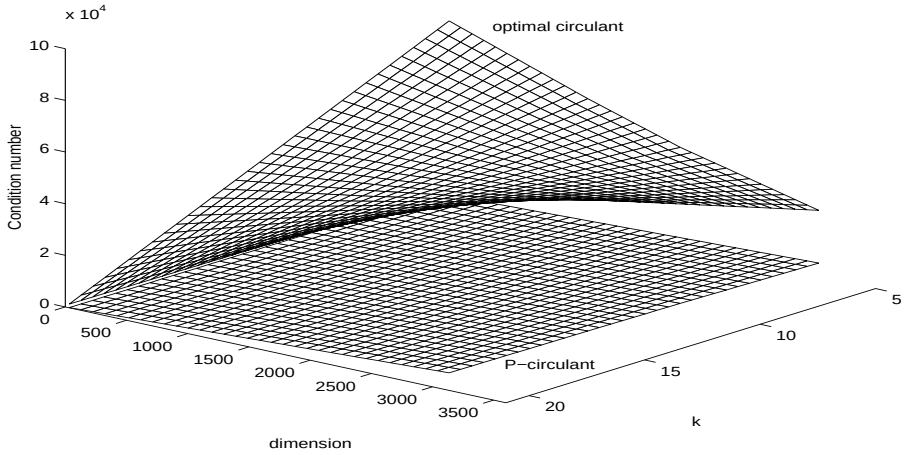


FIG. 5.3. *Condition number of the P-circulant and of the circulant approximation based on* (4.2) *for the matrices A as in* (2.7) *related to k-step formulas* (2.9).

for stable multistep formulas in boundary value form whose component matrices are nonsingular. This is the case of the midpoint method using with one initial and one final condition ($k = 2$, $\nu = 1$ in (2.3)) introduced in [2]. Indeed, by using (5.7), and by recalling that $\alpha_2 = 1 = -\alpha_0$ in (2.3), we have the expression of the eigenvalues of $c(A_{s+1})$ for the above-mentioned method:

$$\left(1 - \frac{1}{s+1}\right)(\epsilon^l - \epsilon^{-l}) = 2\mathbf{i}\left(1 - \frac{1}{s+1}\right)\sin\left(\frac{2\pi\mathbf{i}l}{s+1}\right), \qquad l = 0, \ldots, s,$$

which is zero for $l = 0$ for any $s \geq 2$ and, if s is odd, for $l = (s+1)/2$ as well. On the other hand, by (5.2), the eigenvalues of the corresponding P-circulant matrix $p(A_{s+1})$ are given by

$$\frac{1}{s+1}\cos\left(\frac{2\pi l}{s+1}\right) + \mathbf{i}\sin\left(\frac{2\pi l}{s+1}\right), \qquad l = 0, \ldots, s,$$

which cannot be zero. However, we have experienced that, for low order formulas in section 2.2, both the preconditioners based on P-circulants and on T. Chan's circulants (4.2) can be effective to solve (2.6) with Krylov subspace accelerators; see [4, 5].

5.3. How to choose the approximations and convergence of preconditioned iterations. In the previous sections, we have considered the spectrum of the component matrices of the block preconditioners in (1.2). Further information on the spectrum of the matrix M in (2.6) and its component matrices can be found in [2, 6, 9]. On the other hand, the convergence of preconditioned iterations using, e.g., GMRES, BiCGstab, and some other BiCG-like methods, is essentially decided by the distribution of the spectrum of the eigenvalues and by the eigenvectors of the preconditioned matrix; see, e.g., [18]. Notice that the analysis of the preconditioned linear system, in the nonsymmetric case, cannot be performed by using the arguments in the previous sections. In fact, one should explicitly manipulate the related characteristic polynomial. For example, if we consider the preconditioner (1.2) and the linear

systems (2.6), we need to derive an analytic expression for λ from

$$\det((1-\lambda)I_{m(s+1)} + P^{-1}E) = 0, \qquad E = (A - \breve{A}) \otimes I_m - h(B - \breve{B}) \otimes J, \tag{5.40}$$

where we recall that $I_{m(s+1)} + P^{-1}E = P^{-1}M$; see [4, Theorem 4.1] for details. Unfortunately, the above approach can fail to give complete and useful information on the convergence process. Indeed, the above analysis must consider a specific Jacobian matrix J and a formula (2.3) with k fixed. Moreover, the derivation of λ from (5.40) is usually rather lengthy even for low order schemes and sometimes it is difficult to handle in view of the behavior of the eigenvalues.

Therefore, in what follows, we will give some general suggestions in order to decide whether approximation could be more suitable to precondition the underlying linear system. These hints could be adapted for the solution of other problems based on nonsymmetric (block-)Toeplitz-like matrices.

Recall that, for Krylov subspace methods, we expect fast convergence of preconditioned iterations if the spectrum of the eigenvalues of the block preconditioned matrix is clustered around $(1, 0) \in \mathbb{C}$. Let T_n be a nonsymmetric band Toeplitz matrix, $p(T_n)$ the P-circulant approximation for T_n, and $l(T_n)$ a trigonometric approximation, e.g., one of those described in section 4. Defining $E_p = T_n - p(T_n)$, $E_l = T_n - l(T_n)$, and using similar arguments as in [4], we can write

$$p(T_n)^{-1}T_n = I + p(T_n)^{-1}E_p = I + p(T_n)^{-1}(E_p^{(1)} + E_p^{(2)}), \tag{5.41}$$

$$l(T_n)^{-1}T_n = I + l(T_n)^{-1}E_l = I + l(T_n)^{-1}(E_l^{(1)} + E_l^{(2)}), \tag{5.42}$$

where $E_p^{(2)}$, $E_l^{(2)}$ have small rank with respect to n and $E_p^{(1)}$, $E_l^{(1)}$ have small norm (with respect to T_n, say). From (5.41), (5.42), we expect that the P-circulant-based approximation will perform better than the other if, e.g.,

(C1) $||p(T_n)^{-1}||_2 < ||l(T_n)^{-1}||_2$;

(C2) $||p(T_n)^{-1}E_p^{(1)}||_2 < ||l(T_n)^{-1}E_l^{(1)}||_2$ (if the underlying approximation $l(T_n)$ is such that $||E_l^{(1)}|| \neq 0$; otherwise we require that $||p(T_n)^{-1}E_p^{(1)}||_2$ is moderate);

(C3) the outlying eigenvalues of $p(T_n)^{-1}T_n$ (i.e., the eigenvalues outside the cluster in $(1, 0) \in \mathbb{C}$) have positive real part whereas some of $l(T_n)^{-1}T_n$ have negative real part.

Notice that condition (C1) is equivalent to, say, that of $K_2(p(T_n)) < K_2(l(T_n))$ because $||p(T_n)||_2$, $||l(T_n)||_2$ are uniformly bounded with n. (We assume, as is customary, that the entries of T_n are uniformly bounded with respect to n.) By condition (C2) alone and (5.41), (5.42), it would appear that preconditioners based on simple circulant-like approximations such as Strang's, MS-circulant, and $\{\omega\}$-circulant will perform definitively better than a P-circulant based one (or, e.g., better than (4.2)) because they have $E_l^{(1)} = 0$ in (5.42), i.e., no small norm perturbation. Unfortunately, this is false in general. Finally, the third condition (C3) can be very important for the convergence of GMRES and BiCG-like Krylov methods. Indeed, as observed in [18], if the convex hull of the eigenvalues includes the origin of the complex plane, then the convergence can be slow.

Let us consider some examples in which P-circulant-like block preconditioners in (1.2) can outperform preconditioners based on other approximations for the linear systems (2.6). For simplicity, we assume $J = VDV^{-1}$ diagonalizable, $D =$

$diag(\mu_1, \dots, \mu_m)$, and $\mathrm{Re}(\mu_r) \leq 0$. By using the notation of the previous sections, we have the following decomposition for P as in (1.2):

$$P = (F^* \otimes V)\, \mathrm{diag}\,(\phi_0 - h\psi_0\mu_1, \dots, \phi_0 - h\psi_0\mu_m, \dots \\ \dots, \phi_s - h\psi_s\mu_1, \dots, \phi_s - h\psi_s\mu_m)\,(F \otimes V^{-1}). \tag{5.43}$$

Then, the eigenvalues of the block preconditioner are given by $\phi_j - h\psi_j\mu_r$, $j = 0, \dots, s$, $r = 1, \dots, m$, and

$$||P^{-1}||_2 \leq K_2(V) \min_{j,r}\{|\phi_j - h\psi_j\mu_r|\}^{-1},$$

where $K_2(V)$ does not depend on s. If we consider the matrices related to the schemes (2.9), we have $\psi_j \equiv 1$, $j = 0, \dots, s$, and using P-circulant approximations for $\breve{A}$, $\breve{B}$ in (1.2) gives

$$||P^{-1}||_2 \leq K_2(V) \frac{s+1}{1 + (T - t_0)\tilde{\mu}} = O(s), \qquad \tilde{\mu} = \min_r\{|\mu_r|\}.$$

On the other hand, similar bounds cannot be stated for non-P-circulant-like approximations because, in general, we have

$$||P^{-1}||_2 \leq K_2(V) \frac{s+1}{(T - t_0)\tilde{\mu}}, \qquad \tilde{\mu} = \min_r\{|\mu_r|\},$$

which can be unbounded if some eigenvalues of J are very small or zero in modulus. A similar effect can be observed for some classes of matrices J with purely imaginary eigenvalues and other matrices A, B in (2.6), T_n in (5.41), (5.42) as well.

Notice that, by using similar arguments as before, we can write $P^{-1}M = I + P^{-1}(E^{(1)} + E^{(2)})$; see, [4, Theorem 4.1]. Therefore, if we take the 2-norm of the perturbation of the identity in the right-hand side above, we get

$$||P^{-1}(E^{(1)} + E^{(2)})|| \leq ||P^{-1}|| \cdot (||E^{(1)}|| + ||E^{(2)}||).$$

By the above arguments, $||P^{-1}||$ can be larger for the preconditioner not based on P-circulant matrices. As a result, the amplification of the perturbations $E^{(1)} + E^{(2)}$ given by the multiplication by P^{-1} can give (C3); see (5.41), (5.42). Moreover, recall that the spectrum of the eigenvalues can be much more sensitive to perturbations with respect to the Hermitian case; see, e.g., [22].

On the other hand, if the eigenvalues μ_r, $r = 1, \dots, m$, are, e.g., negative and bounded from below by a constant $c < 0$, then preconditioners based on simple circulant-like approximations (i.e., based on Strang's, $\{\omega\}$-circulant, and MS-circulant matrices) may give better performances for large s as well. For numerical examples, see [4, 5, 7].

Acknowledgments. The author would like to thank three anonymous referees, Lionello Pasquini, and the editor for helpful comments and useful suggestions which have improved this presentation. This work is dedicated to my wife Vittoria.

REFERENCES

[1] E. Artin, *The Gamma Function*, Holt, Rinehart and Winston, New York, 1964.
[2] A. O. H. Axelsson and J. G. Verwer, *Boundary Value Techniques for Initial Value Problems in Ordinary Differential Equations*, Math. Comp., 45 (1985), pp. 153–171.
[3] L. Aceto, *On the Stability Problem Arising in Numerical Methods for ODEs*, Ph.D thesis, Università di Genova, Italy, 1999.
[4] D. Bertaccini, *A circulant preconditioner for the systems of LMF-based ODE codes*, SIAM J. Sci. Comput., 22 (2000), pp. 767–786.
[5] D. Bertaccini, *Reliable preconditioned iterative linear solvers for some numerical integrators*, Numer. Linear Algebra Appl., 8 (2001), pp. 111–125.
[6] D. Bertaccini and M. K. Ng, *The convergence rate of block preconditioned systems arising from LMF-based ODE codes*, BIT, 41 (2001), pp. 433–450.
[7] D. Bertaccini and M. K. Ng, *Skew-circulant preconditioners for systems of LMF-based ODE codes,* Lecture Notes in Comp. Sci. 1988, Springer-Verlag, Berlin, 2001, pp. 93–101.
[8] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1998.
[9] L. Brugnano and D. Trigiante, *Solving ODE by Linear Multistep Methods: Initial and Boundary Value Methods*, Gordon & Breach, Reading, UK, 1998.
[10] R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 104–119.
[11] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[12] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 766–771.
[13] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.
[14] T. A. Driscoll, K. C. Toh L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.
[15] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.
[16] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1994.
[17] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 1, Wiley, New York, 1974.
[18] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Front. Appl. Math. 17, SIAM, Philadelphia, 1997.
[19] J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*, John Wiley, New York, 1991.
[20] S. Serra Capizzano, *Toeplitz preconditioners constructed from linear approximation processes*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 446–465.
[21] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.
[22] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
[23] E. E. Tyrtyshnikov, *Optimal and superoptimal circulant preconditioners*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 459–473.
[24] E. E. Tyrtyshnikov, *Circulant preconditioners with unbounded inverses*, Linear Algebra Appl., 216 (1995), pp. 1–23.
[25] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1823–1849

A FINITE ELEMENT METHOD WITH LAGRANGE MULTIPLIERS FOR LOW-FREQUENCY HARMONIC MAXWELL EQUATIONS*

ALFREDO BERMÚDEZ†, RODOLFO RODRÍGUEZ‡, AND PILAR SALGADO†

Abstract. The aim of this paper is to analyze a finite element method to solve the low-frequency harmonic Maxwell equations in a bounded domain containing conductors and dielectrics. This system of partial differential equations is a model for the so-called eddy currents problem. After writing this problem in terms of the magnetic field, it is discretized by Nédélec edge finite elements on a tetrahedral mesh. Error estimates are easily obtained if the curl-free condition is imposed on the elements in the dielectric domain.

Then, the curl-free condition is imposed, at a discrete level, by introducing a piecewise linear multivalued potential. The resulting problem is shown to be a discrete version of other continuous formulation in which the magnetic field in the dielectric part of the domain has been replaced by a magnetic potential. Moreover, this approach leads to an important saving in computational effort. Problems related to the topology are also considered in that the possibility of having a nonsimply connected dielectric domain is taken into account.

Implementation issues are discussed, including an amenable procedure to impose the boundary conditions by means of a Lagrange multiplier. Finally, the method is applied to solve a three-dimensional model problem: a cylindrical electrode surrounded by dielectric.

Key words. low-frequency harmonic Maxwell equations, eddy currents problems, finite element computational electromagnetism

AMS subject classifications. 78M10, 65N30

PII. S0036142901390780

1. Introduction.

In this paper we analyze a finite element method with Lagrange multipliers to solve the eddy currents model in a bounded domain including conductors and dielectrics. This model can be obtained from Maxwell equations by assuming that all fields are harmonic and that the current frequency is low enough so that the term involving the displacement current in Ampère's law can be neglected. Such a situation happens, for instance, in problems related to machines working at power frequencies. In particular, this paper is motivated by the need for a three-dimensional numerical simulation of a metallurgical furnace. (See [6, 7] for related works concerning axisymmetric models.)

Because of many interesting applications in electrical engineering, numerical solution of eddy currents problems became an important research area, leading to a great number of publications in recent years. (See, for instance, [2, 3, 8, 9, 10, 12, 13, 16, 20, 29].) The books by Bossavit [11] and Silvester and Ferrari [28] also contain valuable material on this subject and include large reference lists.

While several papers deal with the mathematical and numerical analysis of the full harmonic Maxwell equations (see, for instance, the papers by Monk [23, 24, 25]

*Received by the editors June 13, 2001; accepted for publication (in revised form) May 31, 2002; published electronically November 22, 2002. This work was partially supported by Programa de Cooperación Científica con Iberoamérica, Ministerio de Educación y Ciencia, Spain.

http://www.siam.org/journals/sinum/40-5/39078.html

†Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, 15706, Santiago de Compostela, Spain (mabermud@usc.es, mpilar@usc.es). The research of these authors was partially supported by Xunta de Galicia research project PGIDT00PXI20701PR and grant FEDER-CICYT 1FD97.0280 (Spain).

‡GI²MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (rodolfo@ing-mat.udec.cl). The research of this author was partially supported by FONDAP in Applied Mathematics (Chile).

and Fernandes and Gilardi [17]), the number of papers concerning analysis of the eddy currents model is much smaller. Significant mathematical and numerical results have been obtained by MacCamy and coauthors [18, 21, 22] for a two-dimensional eddy currents problem. In the three-dimensional case, let us mention the article by Ammari, Buffa, and Nédélec [4], where a thorough justification of the eddy current model is given.

The above mentioned papers deal with the eddy currents problem in the whole space, the infinity being usually taken into account by means of integral equations. A useful alternative approach is considered by Alonso and Valli [2, 3], where the problem in a bounded domain is considered, including appropriate boundary conditions. In these papers, a formulation involving only the electric field is given and then numerically solved by using a domain decomposition technique and Nédélec edge finite elements.

In the present paper we also consider the eddy currents problem in a bounded domain which includes conductors and dielectrics. The conductors are not assumed to be totally included in the problem domain. We consider a formulation in terms of magnetic field with mixed Neumann and Dirichlet boundary conditions. The former are the natural conditions for the conducting part of the boundary. The latter are imposed on the dielectric part and allow taking into account all the electromagnetics effects outside the domain.

Then, following Bossavit and Vérité [13], we introduce a scalar magnetic potential in the domain occupied by the dielectric. This hybrid formulation is discretized by using Nédélec edge elements for the magnetic field and standard piecewise linear continuous elements for the magnetic potential.

The outline of the paper is as follows: In section 2, we recall the eddy currents model and obtain a weak formulation involving the magnetic field only. Section 3 concerns existence and uniqueness of a solution which are proved by using classical tools. Then, in section 4, we introduce a scalar magnetic potential in the dielectric domain and show that the resulting problem is completely equivalent to the previous one. The numerical discretization is introduced in section 5, where error estimates are obtained under mild regularity assumption on the solution.

In order to solve the discretized problem, a Lagrange multiplier is proposed in section 6 to impose the Dirichlet boundary conditions. The resulting mixed problem is shown to attain a unique solution and to be equivalent to the original discrete one. Finally, in section 7, we report numerical results for a test with known analytical solution; these results confirm the predicted order of convergence of the method.

2. The eddy currents problem. Eddy currents are usually modeled by the low-frequency harmonic Maxwell equations. First let us recall the governing equations of electromagnetism: Maxwell equations,

$$\frac{\partial \boldsymbol{\mathcal{D}}}{\partial t} - \mathbf{curl}\, \boldsymbol{\mathcal{H}} = -\boldsymbol{\mathcal{J}}, \tag{2.1}$$

$$\frac{\partial \boldsymbol{\mathcal{B}}}{\partial t} + \mathbf{curl}\, \boldsymbol{\mathcal{E}} = \mathbf{0}, \tag{2.2}$$

$$\operatorname{div} \boldsymbol{\mathcal{B}} = 0, \tag{2.3}$$

$$\operatorname{div} \boldsymbol{\mathcal{D}} = \rho; \tag{2.4}$$

constitutive laws,

$$\boldsymbol{\mathcal{B}} = \mu \boldsymbol{\mathcal{H}}, \tag{2.5}$$

$$\boldsymbol{\mathcal{D}} = \epsilon \boldsymbol{\mathcal{E}}; \tag{2.6}$$

and Ohm's law in conductors,

$$\boldsymbol{\mathcal{J}} = \sigma \boldsymbol{\mathcal{E}}. \tag{2.7}$$

We have used notations which are standard in electromagnetism:

- $\boldsymbol{\mathcal{D}}$ is the electric displacement,
- $\boldsymbol{\mathcal{E}}$ is the electric field,
- $\boldsymbol{\mathcal{B}}$ is the magnetic induction,
- $\boldsymbol{\mathcal{H}}$ is the magnetic field,
- $\boldsymbol{\mathcal{J}}$ is the current density,
- ρ is the electric charge density,
- μ is the magnetic permeability,
- ϵ is the electric permittivity,
- σ is the electric conductivity.

We use boldface letters to denote vector fields and variables, as well as vector-valued operators, throughout the paper.

When alternating currents are considered, all the fields have the following steady-state form:

$$\boldsymbol{\mathcal{F}}(\mathbf{x}, t) = \mathrm{Re}\left[\mathrm{e}^{\mathrm{i}\omega t}\mathbf{F}(\mathbf{x})\right],$$

where ω is the angular frequency. Moreover, in the low-frequency harmonic regime, the term in (2.1) including the electric displacement can be neglected. Under these assumptions, (2.1)–(2.7) reduce to the so-called eddy currents model:

$$\mathbf{curl}\,\mathbf{H} = \mathbf{J}, \tag{2.8}$$

$$\mathrm{i}\omega\mu\mathbf{H} + \mathbf{curl}\,\mathbf{E} = \mathbf{0}, \tag{2.9}$$

$$\mathrm{div}\,\mathbf{B} = 0, \tag{2.10}$$

$$\mathrm{div}\,\mathbf{D} = \rho, \tag{2.11}$$

with

$$\mathbf{B} = \mu\mathbf{H}, \tag{2.12}$$

$$\mathbf{D} = \epsilon\mathbf{E}, \tag{2.13}$$

$$\mathbf{J} = \sigma\mathbf{E}. \tag{2.14}$$

We are interested in solving these equations in a bounded domain Ω, which consists of two parts, Ω_{C} and Ω_{D}, occupied by conductors and dielectrics, respectively. The electric conductivity σ vanishes in the dielectric domain. The boundary of the domain Ω also splits into two parts: $\Gamma_{\mathrm{C}} := \partial\Omega_{\mathrm{C}} \cap \partial\Omega$ and $\Gamma_{\mathrm{D}} := \partial\Omega_{\mathrm{D}} \cap \partial\Omega$. Finally, we denote $\Gamma_{\mathrm{I}} := \partial\Omega_{\mathrm{C}} \cap \partial\Omega_{\mathrm{D}}$, the interface between dielectric and conductors.

Boundary conditions must be added to solve the eddy currents model in the bounded domain Ω. We consider

$$\mathbf{E} \times \mathbf{n} = \mathbf{0} \quad \text{on } \Gamma_{\mathrm{C}}, \tag{2.15}$$

$$\mathbf{H} \times \mathbf{n} = \mathbf{f} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{2.16}$$

with $\mathbf{f}$ being a given tangential vector field (i.e., satisfying $\mathbf{f} \cdot \mathbf{n} = 0$ on Γ_{D}).

In the equations above, $\mathbf{n}$ denotes the outer unit normal vector to $\partial\Omega$. Throughout the paper, $\mathbf{n}$ will denote a unit vector normal to a given surface, not necessarily the same at each occurrence. In general, it will not be explicitly mentioned which surface this is, provided this is sufficiently clear from the context.

To obtain a weak formulation of the boundary value problem (2.8)–(2.16), consider a test function $\mathbf{G}$ such that $\mathbf{G} \times \mathbf{n} = \mathbf{0}$ on Γ_{D}. From (2.9) we have

$$\mathrm{i}\omega \int_\Omega \mu \mathbf{H} \cdot \bar{\mathbf{G}} + \int_\Omega \mathbf{curl}\, \mathbf{E} \cdot \bar{\mathbf{G}} = 0. \tag{2.17}$$

Now, we can transform the second term above by using Green's formula:

$$\begin{aligned} \int_\Omega \mathbf{curl}\, \mathbf{E} \cdot \bar{\mathbf{G}} &= \int_\Omega \mathbf{E} \cdot\, \mathbf{curl}\, \bar{\mathbf{G}} - \int_{\Gamma_{\mathrm{C}}} \mathbf{E} \cdot \mathbf{n} \times \bar{\mathbf{G}}\, d\Gamma \\ &= \int_\Omega \mathbf{E} \cdot\, \mathbf{curl}\, \bar{\mathbf{G}} - \int_{\Gamma_{\mathrm{C}}} \mathbf{E} \times \mathbf{n} \cdot \bar{\mathbf{G}}\, d\Gamma \;=\; \int_\Omega \mathbf{E} \cdot\, \mathbf{curl}\, \bar{\mathbf{G}}, \end{aligned} \tag{2.18}$$

where we have used the boundary condition (2.15) to obtain the last equality. We observe that (2.8) and (2.14), and the fact that σ is null in the dielectric domain, lead to

$$\mathbf{curl}\, \mathbf{H} = \mathbf{0} \quad \text{in } \Omega_{\mathrm{D}}.$$

Because of this, we need only to take test functions $\mathbf{G}$ satisfying $\mathbf{curl}\, \mathbf{G} = \mathbf{0}$ in Ω_{D}. By doing so, (2.17) and (2.18) yield

$$\mathrm{i}\omega \int_\Omega \mu \mathbf{H} \cdot \bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \mathbf{E} \cdot\, \mathbf{curl}\, \bar{\mathbf{G}} = 0.$$

Instead, in the conductors, (2.8) and (2.14) lead to $\mathbf{E} = \frac{1}{\sigma}\, \mathbf{curl}\, \mathbf{H}$, which allows us to eliminate $\mathbf{E}$ in the equation above. Thus, finally we obtain

$$\mathrm{i}\omega \int_\Omega \mu \mathbf{H} \cdot \bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\, \mathbf{curl}\, \mathbf{H} \cdot\, \mathbf{curl}\, \bar{\mathbf{G}} = 0.$$

3. Analysis of the magnetic field formulation of the eddy currents problem. Let us assume that Ω is simply connected, with a Lipschitz-continuous connected boundary. The subdomains Ω_{C} and Ω_{D} are also assumed to have Lipschitz-continuous boundaries, although not necessarily connected. Finally, the boundaries of Γ_{C}, Γ_{D}, and Γ_{I} are assumed to be Lipschitz-continuous, too.

We use standard notation for Sobolev spaces and norms. Moreover, we recall the definition of some functional spaces. Let

$$\mathrm{H}(\mathbf{curl}, \Omega) := \left\{ \mathbf{G} \in \mathrm{L}^2(\Omega)^3 : \;\; \mathbf{curl}\, \mathbf{G} \in \mathrm{L}^2(\Omega)^3 \right\}$$

endowed with the norm

$$\|\mathbf{G}\|_{\mathrm{H}(\mathbf{curl},\Omega)} := \left[\|\mathbf{G}\|^2_{\mathrm{L}^2(\Omega)^3} + \|\, \mathbf{curl}\, \mathbf{G}\|^2_{\mathrm{L}^2(\Omega)^3} \right]^{1/2},$$

and, for each positive real number r, let

$$\mathrm{H}^r(\mathbf{curl}, \Omega) := \left\{ \mathbf{G} \in \mathrm{H}^r(\Omega)^3 : \;\; \mathbf{curl}\, \mathbf{G} \in \mathrm{H}^r(\Omega)^3 \right\}$$

endowed with the norm

$$\|\mathbf{G}\|_{\mathrm{H}^r(\mathbf{curl},\Omega)} := \left[\|\mathbf{G}\|^2_{\mathrm{H}^r(\Omega)^3} + \|\,\mathbf{curl}\,\mathbf{G}\|^2_{\mathrm{H}^r(\Omega)^3}\right]^{1/2}.$$

Consider the following closed subspaces of $\mathrm{H}(\mathbf{curl},\Omega)$:

$$\boldsymbol{\mathcal{V}} = \{\mathbf{G} \in \mathrm{H}(\mathbf{curl},\Omega) :\ \ \mathbf{curl}\,\mathbf{G} = \mathbf{0} \text{ in } \Omega_{\mathrm{D}}\},$$
$$\boldsymbol{\mathcal{V}}^0 = \left\{\mathbf{G} \in \boldsymbol{\mathcal{V}} :\ \mathbf{G}\times\mathbf{n} = \mathbf{0} \text{ in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3\right\},$$

where $\mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3$ denotes the dual space of $\mathrm{H}_{00}^{1/2}(\Gamma_{\mathrm{D}})^3$, which, in its turn, is the space of functions defined on Γ_{D} that, extended by $\mathbf{0}$ on $\partial\Omega\setminus\Gamma_{\mathrm{D}}$, belong to $\mathrm{H}^{1/2}(\partial\Omega)^3$.

We assume that $\mu,\epsilon,\sigma \in \mathrm{L}^\infty(\Omega)$ and that there exist constants $\underline{\mu}$, $\underline{\epsilon}$, and $\underline{\sigma}$ such that

$$\begin{aligned}
&\mu(\mathbf{x}) \geq \underline{\mu} > 0, \quad \text{a.e. in } \Omega,\\
&\epsilon(\mathbf{x}) \geq \underline{\epsilon} > 0, \quad \text{a.e. in } \Omega,\\
&\sigma(\mathbf{x}) \geq \underline{\sigma} > 0, \quad \text{a.e. in } \Omega_{\mathrm{C}}, \qquad \sigma(\mathbf{x}) = 0 \quad \text{in } \Omega_{\mathrm{D}}.
\end{aligned}$$

Concerning the boundary data $\mathbf{f}$, we suppose there exists a field $\mathbf{H_f} \in \boldsymbol{\mathcal{V}}$ such that

$$\mathbf{H_f}\times\mathbf{n} = \mathbf{f} \text{ in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3. \tag{3.1}$$

Remark 3.1. We refer to [1] for necessary and sufficient conditions on $\mathbf{f}$ to ensure that there exists $\mathbf{H_f} \in \mathrm{H}(\mathbf{curl},\Omega)$ such that $\mathbf{H_f}\times\mathbf{n} = \mathbf{f}$ on Γ_{D} in a weak sense, in the case $\Gamma_{\mathrm{C}} = \emptyset$, $\Gamma_{\mathrm{D}} = \partial\Omega$ (i.e., when the conductors $\overline{\Omega}_{\mathrm{C}}$ are fully contained in Ω). We also refer to [14, 15] for similar conditions in the case that Ω is a Lipschitz polyhedron, and Γ_{C} and Γ_{D} are polyhedral surfaces with piecewise smooth boundaries.

Equation (3.1) implies an additional constraint on the data $\mathbf{f}$, since $\mathbf{H_f}$ has to be curl-free in Ω_{D}. A necessary condition for the existence of such $\mathbf{H_f}$ is that $\mathrm{div}_\Gamma\,\mathbf{f} = 0$ on Γ_{D}, where div_Γ stands for the tangential divergence operator. (See [1] for the result and a precise definition of div_Γ.) In the case $\Gamma_{\mathrm{C}} = \emptyset$, $\Gamma_{\mathrm{D}} = \partial\Omega$; then $\mathrm{div}_\Gamma\,\mathbf{f} = 0$ on Γ_{D} is also a sufficient condition, when Ω has a smooth boundary (see Theorem 4.1 of [1]).

Now, we can state a variational formulation of our problem in terms of the magnetic field $\mathbf{H}$.

Problem **MP**. *Find* $\mathbf{H} \in \boldsymbol{\mathcal{V}}$ *such that*

$$\mathbf{H}\times\mathbf{n} = \mathbf{f} \text{ in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3, \tag{3.2}$$

$$\mathrm{i}\omega\int_\Omega \mu\mathbf{H}\cdot\bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\mathbf{H}\cdot\,\mathbf{curl}\,\bar{\mathbf{G}} = 0 \quad \forall\mathbf{G} \in \boldsymbol{\mathcal{V}}^0. \tag{3.3}$$

Let $a\colon \mathrm{H}(\mathbf{curl},\Omega)\times\mathrm{H}(\mathbf{curl},\Omega) \longrightarrow \mathbb{C}$ be the sesquilinear continuous form defined by

$$a(\mathbf{H},\mathbf{G}) := \mathrm{i}\omega\int_\Omega \mu\mathbf{H}\cdot\bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\mathbf{H}\cdot\,\mathbf{curl}\,\bar{\mathbf{G}}.$$

This form clearly satisfies

$$|a\,(\mathbf{G},\mathbf{G})| \geq \alpha\|\mathbf{G}\|^2_{\mathrm{H}(\mathbf{curl},\Omega)} \quad \forall\mathbf{G} \in \boldsymbol{\mathcal{V}}. \tag{3.4}$$

Hence, the following existence result is immediately derived.

THEOREM 3.1. *If there exists* $\mathbf{H_f} \in \boldsymbol{\mathcal{V}}$ *such that* $\mathbf{H_f} \times \mathbf{n} = \mathbf{f}$ *in* $\mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3$*, then Problem* **MP** *attains a unique solution.*

Proof. Consider the translation $\hat{\mathbf{H}} = \mathbf{H} - \mathbf{H_f}$. Then Problem **MP** is equivalent to finding $\hat{\mathbf{H}} \in \boldsymbol{\mathcal{V}}^0$ such that

$$a(\hat{\mathbf{H}}, \mathbf{G}) = -a(\mathbf{H_f}, \mathbf{G}) \quad \forall \mathbf{G} \in \boldsymbol{\mathcal{V}}^0,$$

and this problem has a unique solution because of inequality (3.4) and the Lax–Milgram lemma. ∎

Once the magnetic field $\mathbf{H}$ is known, the current density $\mathbf{J}$ and the electric field $\mathbf{E}$ can be readily computed in the conductors by means of (2.8) and (2.14), respectively. These are the magnitudes actually needed in most applications.

In the following theorem we show that the solution of Problem **MP** satisfies some of the Maxwell equations (2.8)–(2.11) and the boundary conditions (2.15)–(2.16) in a weak sense.

THEOREM 3.2. *Let* $\mathbf{H} \in \boldsymbol{\mathcal{V}}$ *be the solution of Problem* **MP**. *Let* $\mathbf{B} = \mu\mathbf{H} \in \mathrm{L}^2(\Omega)$, $\mathbf{J} = \mathbf{curl}\,\mathbf{H} \in \mathrm{L}^2(\Omega)$, *and* $\mathbf{E} = (\frac{1}{\sigma}\mathbf{J})|_{\Omega_{\mathrm{C}}} \in \mathrm{L}^2(\Omega_{\mathrm{C}})$. *Then the following properties hold true:*

$$\operatorname{div}\mathbf{B} = 0 \quad \text{in } \Omega, \tag{3.5}$$

$$\mathrm{i}\omega\mu\mathbf{H} + \mathbf{curl}\,\mathbf{E} = \mathbf{0} \quad \text{in } \Omega_{\mathrm{C}}, \tag{3.6}$$

$$\mathbf{E} \times \mathbf{n} = \mathbf{0} \quad \text{in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{C}})^3, \tag{3.7}$$

$$\mathbf{H} \times \mathbf{n} = \mathbf{f} \quad \text{in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3, \tag{3.8}$$

$$\mathbf{J} = \mathbf{0} \quad \text{in } \Omega_{\mathrm{D}}. \tag{3.9}$$

Proof. Given $\Psi \in \mathcal{D}(\Omega) := \{\Psi \in \mathcal{C}^\infty(\Omega) : \operatorname{supp}\Psi \subset \Omega\}$, let $\mathbf{G} = \mathbf{grad}\,\Psi \in \boldsymbol{\mathcal{V}}^0$. Then, (3.3) yields

$$\int_\Omega \mu\mathbf{H} \cdot \mathbf{grad}\,\bar{\Psi} = 0.$$

Consequently, $\mathbf{B} = \mu\mathbf{H} \in \mathrm{H}(\mathrm{div}, \Omega)$, and (3.5) holds true.

Now, let $\mathbf{G} \in \mathcal{D}(\Omega)^3$ be such that $\operatorname{supp}\mathbf{G} \subset \Omega_{\mathrm{C}}$. Then $\mathbf{G} \in \boldsymbol{\mathcal{V}}^0$, and (3.3) yields

$$\mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu\mathbf{H} \cdot \bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\mathbf{H} \cdot \mathbf{curl}\,\bar{\mathbf{G}} = 0.$$

Hence $\mathbf{E} = \frac{1}{\sigma}\mathbf{J} = \frac{1}{\sigma}\,\mathbf{curl}\,\mathbf{H} \in \mathrm{H}(\mathbf{curl}, \Omega_{\mathrm{C}})$, and (3.6) holds true.

To prove (3.7), given $\boldsymbol{\varphi} \in \mathrm{H}_{00}^{1/2}(\Gamma_{\mathrm{C}})^3$, we will show that $\langle \mathbf{E} \times \mathbf{n}, \widetilde{\boldsymbol{\varphi}} \rangle_{\partial\Omega_{\mathrm{C}}} = 0$, where $\langle \cdot, \cdot \rangle_{\partial\Omega_{\mathrm{C}}}$ denotes the duality pairing in $\mathrm{H}^{-1/2}(\partial\Omega_{\mathrm{C}})^3 \times \mathrm{H}^{1/2}(\partial\Omega_{\mathrm{C}})^3$, and $\widetilde{\boldsymbol{\varphi}} \in \mathrm{H}^{1/2}(\partial\Omega_{\mathrm{C}})^3$ is the natural extension of $\boldsymbol{\varphi}$ by $\mathbf{0}$ on $\partial\Omega_{\mathrm{C}} \setminus \Gamma_{\mathrm{C}}$. To this aim, let $\mathbf{G} \in \mathrm{H}^1(\Omega_{\mathrm{C}})^3$ be such that $\mathbf{G}|_{\partial\Omega_{\mathrm{C}}} = \widetilde{\boldsymbol{\varphi}}$, and $\widetilde{\mathbf{G}}$ is the extension by $\mathbf{0}$ of $\mathbf{G}$ to $\Omega \setminus \Omega_{\mathrm{C}}$. Then $\widetilde{\mathbf{G}} \in \boldsymbol{\mathcal{V}}^0$, and (3.3) yields

$$\begin{aligned} 0 &= \mathrm{i}\omega \int_\Omega \mu\mathbf{H} \cdot \bar{\widetilde{\mathbf{G}}} + \int_{\Omega_{\mathrm{C}}} \mathbf{E} \cdot \mathbf{curl}\,\bar{\widetilde{\mathbf{G}}} \\ &= \mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu\mathbf{H} \cdot \bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \mathbf{curl}\,\mathbf{E} \cdot \bar{\mathbf{G}} + \left\langle \mathbf{E} \times \mathbf{n}, \mathbf{G}|_{\partial\Omega_{\mathrm{C}}} \right\rangle_{\partial\Omega_{\mathrm{C}}} = \langle \mathbf{E} \times \mathbf{n}, \widetilde{\boldsymbol{\varphi}} \rangle_{\partial\Omega_{\mathrm{C}}}, \end{aligned}$$

where we have used that $\mathbf{E} = \frac{1}{\sigma}\,\mathbf{curl}\,\mathbf{H}$ in Ω_{C} and (3.6).

Finally, (3.8) and (3.9) arise explicitly in Problem **MP**. ∎

Remark 3.2. The theorem above shows that Problem **MP** allows us to determine uniquely the electric field $\mathbf{E}$ in the conductors. In its turn, $\mathbf{E}$ and Maxwell equation (2.11) determine the charge density ρ in Ω_{C}. In particular, in the interior of any homogeneous subdomain Ω' of Ω_{C} (i.e., $\Omega' \subset \Omega_{\mathrm{C}}$ such that $\epsilon|_{\Omega'}$ and $\sigma|_{\Omega'}$ are constant), $\rho|_{\Omega'} = \operatorname{div}\left(\frac{\epsilon}{\sigma}\,\mathbf{curl}\,\mathbf{H}\right)\big|_{\Omega'} = 0$.

Instead, the electric field $\mathbf{E}$ is not uniquely determined in the dielectric. Indeed, from the eddy currents model (2.8)–(2.14) we obtain the following equations for $\mathbf{E}|_{\Omega_{\mathrm{D}}}$:

$$\mathbf{curl}\,\mathbf{E} = -\mathrm{i}\omega\mu\mathbf{H} \quad \text{in } \Omega_{\mathrm{D}}, \tag{3.10}$$

$$\operatorname{div}(\epsilon\mathbf{E}) = \rho \quad \text{in } \Omega_{\mathrm{D}}, \tag{3.11}$$

$$\mathbf{E}\times\mathbf{n} = \mathbf{E}|_{\Omega_{\mathrm{C}}}\times\mathbf{n} \quad \text{on } \Gamma_{\mathrm{I}}. \tag{3.12}$$

The latter arises from the facts that $\mathbf{E}|_{\Omega_{\mathrm{C}}}$ is already known and that $\mathbf{E}$ is globally in $\mathrm{H}(\mathbf{curl},\Omega)$.

A boundary condition on Γ_{D} is needed to determine a unique solution, even in the simplest case of a topologically trivial Ω_{D} (i.e., when Ω_{D} is simply connected with a connected boundary). A natural condition would be to impose the normal component of the electric displacement $\mathbf{D}$ on Γ_{D}; namely,

$$\epsilon\mathbf{E}\cdot\mathbf{n} = \psi \quad \text{on } \Gamma_{\mathrm{D}}. \tag{3.13}$$

The data ψ amounts to eventual surface charges on the outer boundary of the dielectric domain.

Existence of the solution to (3.10)–(3.13) has been proved in Theorem 4.2 of [1] in the case that $\partial\Omega_{\mathrm{D}}$ is smooth and that $\Gamma_{\mathrm{I}}\cap\Gamma_{\mathrm{D}} = \emptyset$ (for instance, when $\overline{\Omega}_{\mathrm{C}} \subset \Omega$). Even in this simpler case, a number of additional constraints related to the topology of Ω_{D} must be added to have uniqueness, as can be seen in this reference.

To the best of the authors' knowledge, a similar result has not been proved for the general case of Ω_{D} being a Lipschitz polyhedron with $\Gamma_{\mathrm{I}}\cap\Gamma_{\mathrm{D}} \neq \emptyset$. Nevertheless, this is not a drawback for the application of this eddy currents model, since typically the goal of these problems is to compute the electric field only in the conductors, as mentioned above.

4. Introducing a magnetic potential. In this section we show how Problem **MP** can be transformed by replacing the magnetic field in the dielectric domain Ω_{D} by a (scalar) magnetic potential.

We recall that Ω is assumed to be simply connected with connected boundary $\partial\Omega$. Let $\Omega_{\mathrm{C}} = \bigcup_{j=0}^{j=J}\Omega_{\mathrm{C}}^j$, with Ω_{C}^0 being the union of all the connected components of Ω_{C} such that $\Omega\backslash\Omega_{\mathrm{C}}^0$ is simply connected, and Ω_{C}^j, $j = 1,\dots,J$, are the remaining connected components of Ω_{C} (see Figure 4.1).

We assume that for each Ω_{C}^j, $j = 1,\dots,J$, there exists an open "cut" surface $\Sigma_j \subset \Omega_{\mathrm{D}}$ such that $\partial\Sigma_j \subset \partial\Omega_{\mathrm{D}}$ and $\widetilde{\Omega}_{\mathrm{D}} := \Omega_{\mathrm{D}} \setminus \bigcup_{j=0}^{j=J}\overline{\Sigma}_j$ is pseudo-Lipschitz and simply connected (see Figure 4.1). We also assume that each one of these surfaces Σ_j is connected, and $\overline{\Sigma}_j \cap \overline{\Sigma}_k = \emptyset$ for $j \neq k$ (see, for instance, [5]).

Let us arrange the conductors Ω_{C}^j in such a way that the inner ones are numbered from $j = 1$ to K, and those going through $\partial\Omega$, from $j = K+1$ to J. In Figure 4.1, Ω_{C}^1 is an example of a conductor of the first kind, and Ω_{C}^2 of the second kind.

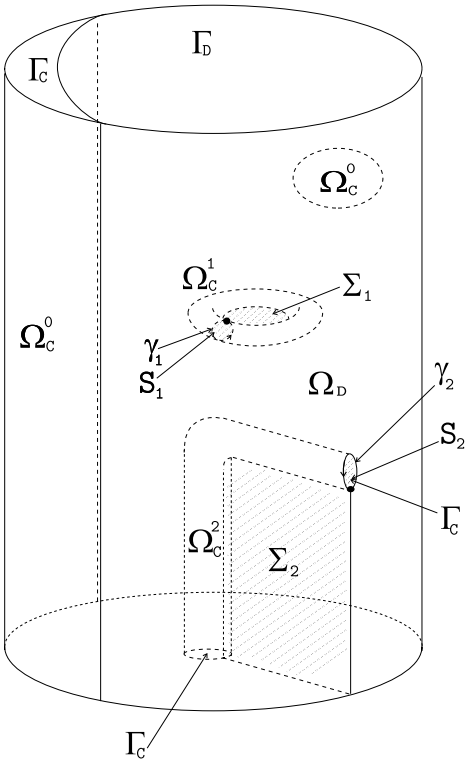


FIG. 4.1. *Sketch of the domain.*

We also assume that there exist cross sections of Ω_C^j, $j = 1 \dots J$; namely, open surfaces $S_j \subset \overline{\Omega}_C^j$, with respective boundaries $\partial S_j = \overline{S}_j \cap \Gamma_I$, which are assumed to be closed simple curves. We denote these curves γ_j. Moreover, for $j = K+1, \dots, J$, we take $S_j \subset \Gamma_C$ and $\gamma_j \subset \Gamma_C \cap \Gamma_D$ (see Figure 4.1 again).

Let $\widetilde{\Omega}_D^j := \Omega_D \setminus \overline{\Sigma}_j$, $j = 1 \dots J$. We fix a unit normal $\mathbf{n}_j$ on each Σ_j and denote its two faces by Σ_j^- and Σ_j^+, with $\mathbf{n}_j$ being the "outer" normal to $\widetilde{\Omega}_D^j$ along Σ_j^+. We choose an orientation for each γ_j by taking its initial and end points on $\overline{\Sigma}_j^-$ and $\overline{\Sigma}_j^+$, respectively. We denote by $\mathbf{t}_j$ the unit vector tangent to γ_j.

For any function $\widetilde{\Psi} \in \mathrm{H}^1(\widetilde{\Omega}_D)$, we denote by

$$[\![\widetilde{\Psi}]\!]_{\Sigma_j} := \widetilde{\Psi}|_{\Sigma_j^-} - \widetilde{\Psi}|_{\Sigma_j^+}$$

the jump of $\widetilde{\Psi}$ through Σ_j along $\mathbf{n}_j$. The gradient of $\widetilde{\Psi}$ in $\mathcal{D}'(\widetilde{\Omega}_D)$ can be extended to $\mathrm{L}^2(\Omega_D)^3$ and will be denoted by $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}$.

Let Θ be the linear space of $\mathrm{H}^1(\widetilde{\Omega}_D)$ defined by

$$\Theta = \left\{ \widetilde{\Psi} \in \mathrm{H}^1(\widetilde{\Omega}_D) : \ [\![\widetilde{\Psi}]\!]_{\Sigma_j} = \text{constant}, \ j = 1, \dots, J \right\}.$$

Then, for $\widetilde{\Psi} \in \mathrm{H}^1(\widetilde{\Omega}_D)$, we have that $\widetilde{\mathbf{grad}}\,\widetilde{\Psi} \in \mathrm{H}(\mathbf{curl}, \Omega_D)$ if and only if $\widetilde{\Psi} \in \Theta$, in which case $\mathbf{curl}\,(\widetilde{\mathbf{grad}}\,\widetilde{\Psi}) = \mathbf{0}$ (see Lemma 3.11 in [5]). Actually, the kernel of the operator $\mathbf{curl} : \ \mathrm{H}(\mathbf{curl}, \Omega_D) \longrightarrow \mathrm{L}^2(\Omega_D)^3$ is given by

$$\mathrm{Ker}\,(\mathbf{curl}) = \widetilde{\mathbf{grad}}\,\Theta = \mathbf{grad}\,\mathrm{H}^1(\Omega_D) \oplus \boldsymbol{\mathcal{C}}, \tag{4.1}$$

where $\boldsymbol{\mathcal{C}}$ is the space of the so-called *Neumann harmonic fields* in Ω_{D} defined by

$$\boldsymbol{\mathcal{C}} := \left\{ \mathbf{G} \in \mathrm{L}^2(\Omega_{\mathrm{D}})^3 : \ \mathbf{curl}\, \mathbf{G} = \mathbf{0}, \, \mathrm{div}\, (\mu \mathbf{G}) = 0 \text{ in } \Omega_{\mathrm{D}}, \text{ and } \mu \mathbf{G} \cdot \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega_{\mathrm{D}} \right\}.$$

A basis of the space $\boldsymbol{\mathcal{C}}$ is given by the set of functions $\{ \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_j, \ j = 1, \ldots, J\}$, where, for each j, $\widetilde{\Phi}_j \in \mathrm{H}^1(\widetilde{\Omega}^j_{\mathrm{D}})$ is the solution of

$$\int_{\widetilde{\Omega}^j_{\mathrm{D}}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_j \cdot \mathbf{grad}\, \bar{\Psi} = 0 \quad \forall \Psi \in \mathrm{H}^1(\Omega_{\mathrm{D}}), \tag{4.2}$$

$$[\![\widetilde{\Phi}_j]\!]_{\Sigma_j} = 1. \tag{4.3}$$

By using the Lax–Milgram lemma, it is straightforward to see that $\widetilde{\Phi}_j$ is uniquely defined in $\mathrm{H}^1(\widetilde{\Omega}^j_{\mathrm{D}})/\mathbb{C}$. (See, for instance, [5] again.)

Therefore, according to (4.1), for all $\mathbf{G} \in \boldsymbol{\mathcal{V}}$, there exist unique constants c_j, $j = 1, \ldots, J$, and a unique scalar field $\Psi \in \mathrm{H}^1(\Omega_{\mathrm{D}})/\mathbb{C}$ such that $\mathbf{G}|_{\Omega_{\mathrm{D}}} = \widetilde{\mathbf{grad}}\, \widetilde{\Psi}$, with $\widetilde{\Psi} \in \Theta$ given by $\widetilde{\Psi} = \Psi + \sum_{j=1}^{J} c_j \widetilde{\Phi}_j$. Furthermore, because of (4.3), the constants c_j are the jumps of $\widetilde{\Psi}$ across the respective cuts Σ_j. Consequently, given $\widetilde{\Psi} \in \Theta$, we have that $\widetilde{\Psi} \in \mathrm{H}^1(\Omega)$ if and only if $[\![\widetilde{\Psi}]\!]_{\Sigma_j} = 0$ for $j = 1, \ldots, J$.

Remark 4.1. These jumps have a precise physical meaning. For instance, for the solution $\mathbf{H}$ of Problem **MP**, let us write $\mathbf{H}|_{\Omega_{\mathrm{D}}} = \widetilde{\mathbf{grad}}\, \widetilde{\Phi}$ with $\widetilde{\Phi} \in \Theta$. If $\mathbf{H}$ is sufficiently smooth, by using the Stokes theorem and (2.8) we have

$$\begin{aligned} [\![\widetilde{\Phi}]\!]_{\Sigma_j} &= \int_{\gamma_j} \widetilde{\mathbf{grad}}\, \widetilde{\Phi} \cdot \mathbf{t}_j \, d\gamma = \int_{\gamma_j} \mathbf{H}|_{\Omega_{\mathrm{D}}} \cdot \mathbf{t}_j \, d\gamma = \int_{\gamma_j} \mathbf{H}|_{\Omega_{\mathrm{C}}} \cdot \mathbf{t}_j \, d\gamma \\ &= \int_{S_j} \mathbf{curl}\, \mathbf{H}|_{\Omega_{\mathrm{C}}} \cdot \mathbf{n} \, d\Gamma = \int_{S_j} \mathbf{J} \cdot \mathbf{n} \, d\Gamma =: I_j, \qquad j = 1, \ldots, J. \end{aligned}$$

Thus, the jump of the magnetic potential $\widetilde{\Psi}$ across each cut surface Σ_j is exactly the current intensity I_j through the cross section S_j of the conductor Ω^j_{C} (as defined above).

We introduce the following notation: for $\mathbf{G}_{\mathrm{C}} \in \mathrm{L}^2(\Omega_{\mathrm{C}})^3$ and $\mathbf{G}_{\mathrm{D}} \in \mathrm{L}^2(\Omega_{\mathrm{D}})^3$, we denote by $(\mathbf{G}_{\mathrm{C}} | \mathbf{G}_{\mathrm{D}})$ the field $\mathbf{G} \in \mathrm{L}^2(\Omega)^3$ defined a.e. by

$$\mathbf{G}(\mathbf{x}) := \begin{cases} \mathbf{G}_{\mathrm{C}}(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_{\mathrm{C}}, \\ \mathbf{G}_{\mathrm{D}}(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_{\mathrm{D}}. \end{cases}$$

Let us denote by $\boldsymbol{\mathcal{W}}$ the linear space given by

$$\boldsymbol{\mathcal{W}} := \left\{ (\mathbf{G}, \widetilde{\Psi}) \in \mathrm{H}(\mathbf{curl}, \Omega_{\mathrm{C}}) \times (\Theta/\mathbb{C}) : \ (\mathbf{G} | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}) \in \mathrm{H}(\mathbf{curl}, \Omega) \right\}.$$

Clearly, the following application is an isomorphism:

$$\begin{array}{ccc} \boldsymbol{\mathcal{W}} & \longrightarrow & \boldsymbol{\mathcal{V}}, \\ (\mathbf{G}, \widetilde{\Psi}) & \longmapsto & (\mathbf{G} | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}). \end{array}$$

Similarly, we define the closed subspace of $\boldsymbol{\mathcal{W}}$

$$\boldsymbol{\mathcal{W}}^0 := \left\{ (\mathbf{G}, \widetilde{\Psi}) \in \boldsymbol{\mathcal{W}} : \ \widetilde{\mathbf{grad}}\, \widetilde{\Psi} \times \mathbf{n} = \mathbf{0} \text{ in } \mathrm{H}^{-1/2}_{00}(\Gamma_{\mathrm{D}})^3 \right\},$$

which is isomorphically equivalent to $\boldsymbol{\mathcal{V}}^0$.

Thus, we are led to define the following problem.

PROBLEM **HP**. *Find* $(\mathbf{H}, \widetilde{\Phi}) \in \boldsymbol{\mathcal{W}}$ *such that*

$$\widetilde{\mathbf{grad}}\, \widetilde{\Phi} \times \mathbf{n} = \mathbf{f} \text{ in } \mathrm{H}_{00}^{-1/2}(\Gamma_{\mathrm{D}})^3,$$
$$\mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu \mathbf{H} \cdot \bar{\mathbf{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \mathbf{curl}\, \mathbf{H} \cdot \mathbf{curl}\, \bar{\mathbf{G}} + \mathrm{i}\omega \int_{\Omega_{\mathrm{D}}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi} \cdot \widetilde{\mathbf{grad}}\, \bar{\widetilde{\Psi}} = 0 \quad \forall (\mathbf{G}, \widetilde{\Psi}) \in \boldsymbol{\mathcal{W}}^0.$$

This is the well-known magnetic field/magnetic potential hybrid formulation of the eddy currents problem introduced by Bossavit and Vérité [13]. One main advantage with respect to formulation (3.2)–(3.3) lies in the fact that a vector field is replaced by a scalar one in the dielectric domain.

The following lemma is an immediate consequence of the isomorphisms between $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{V}}$, and between $\boldsymbol{\mathcal{W}}^0$ and $\boldsymbol{\mathcal{V}}^0$.

LEMMA 4.1. *The pair* $(\mathbf{H}, \widetilde{\Phi})$ *is solution of Problem* **HP** *if and only if* $(\mathbf{H}|\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi})$ *is solution of Problem* **MP**.

As a consequence of this lemma, Theorem 3.1 yields existence and uniqueness of solution for Problem **HP**.

COROLLARY 4.2. *Under the assumptions of Theorem* 3.1*, Problem* **HP** *has a unique solution* $(\mathbf{H}, \widetilde{\Phi})$*, with* $(\mathbf{H}|\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi})$ *being the unique solution of Problem* **MP**.

5. Numerical solution. In this section we first introduce a discretization of Problem **MP** and prove its convergence. Then we prove that the obtained discrete problem is completely equivalent to a convenient discrete version of Problem **HP**.

5.1. Discretizing the magnetic field. We employ "edge" finite elements to approximate the magnetic field, more precisely, the lowest-order finite element of the family introduced by Nédélec in [26]. This element belongs to the family of the so-called Whitney elements (see [9]).

We assume Ω, Ω_{C}, and Ω_{D} are Lipschitz polyhedra and consider a family of regular tetrahedral meshes $\{\mathcal{T}_h\}$ of Ω such that, for every mesh $\mathcal{T}_h$, each element $K \in \mathcal{T}_h$ is contained either in $\overline{\Omega}_{\mathrm{C}}$ or in $\overline{\Omega}_{\mathrm{D}}$. ($h$ stands as usual for the corresponding mesh-size.)

The magnetic field is approximated in each tetrahedron K by a polynomial vector field in the space

$$\boldsymbol{\mathcal{N}}(K) := \left\{ \mathbf{G}_h \in \mathcal{P}_1(K)^3 :\ \mathbf{G}_h(\mathbf{x}) = \mathbf{a} \times \mathbf{x} + \mathbf{b},\ \mathbf{a}, \mathbf{b} \in \mathbb{C}^3,\ \mathbf{x} \in K \right\}.$$

An explicit computation shows that vector fields of this type have constant tangential components along each straight line in the Euclidean space. Moreover, given six complex numbers β_n, $n = 1, \ldots, 6$, there exists a unique $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}(K)$ (i.e., unique $\mathbf{a}, \mathbf{b} \in \mathbb{C}^3$) such that its tangential component along the nth edge of K coincides with β_n for $n = 1 \ldots 6$, respectively. Thus, these tangential components along the edges of K can be taken as the degrees of freedom defining the elements in $\boldsymbol{\mathcal{N}}(K)$.

These elements are H(**curl**)-conforming in the sense that, for all $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}(K)$, their tangential traces on each triangular face T of K depend only on the degrees of freedom of $\mathbf{G}_h$ on the three edges of T. So, if we set

$$\boldsymbol{\mathcal{N}}_h(\Omega) := \{ \mathbf{G}_h \in \mathrm{H}(\mathbf{curl}, \Omega) :\ \mathbf{G}_h|_K \in \boldsymbol{\mathcal{N}}(K)\ \forall K \in \mathcal{T}_h \},$$

the elements in this space are piecewise linear vector fields with tangential traces that are continuous through the faces of the mesh. This is the lowest-order Nédélec

finite element space introduced in [26]. See [19] for a detailed mathematical analysis and [11] for useful implementation issues.

If $\mathbf{G}$ is smooth enough (e.g., $\mathbf{G} \in \mathrm{H}^2(\Omega)^3$), then its Nédélec interpolant $\mathbf{G}^{\mathrm{I}}$ is defined by

$$\mathbf{G}^{\mathrm{I}} \in \boldsymbol{\mathcal{N}}_h(\Omega): \quad \int_\ell \mathbf{G}^{\mathrm{I}} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{G} \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h, \tag{5.1}$$

where, from now on, $\mathbf{t}_\ell$ denotes a unit vector tangent to the edge ℓ. The Nédélec interpolation operator

$$\begin{array}{rcl} \mathrm{H}^2(\Omega)^3 & \longrightarrow & \boldsymbol{\mathcal{N}}_h(\Omega), \\ \mathbf{G} & \longmapsto & \mathbf{G}^{\mathrm{I}}, \end{array} \tag{5.2}$$

with $\mathbf{G}^{\mathrm{I}}$ defined by (5.1), extends uniquely to $\mathrm{H}^r(\mathbf{curl}, \Omega)$ with $r > 1/2$. Indeed, according to the Sobolev imbedding theorem and a trace theorem, for each $K \in \mathcal{T}_h$, $\mathbf{G}|_K \in \mathrm{L}^p(K)^3$, $\mathbf{curl}\, \mathbf{G}|_K \in \mathrm{L}^p(K)^3$, and $\mathbf{G} \times \mathbf{n}|_{\partial K} \in \mathrm{L}^p(\partial K)^3$, with $p = 4/(3-2r) > 2$. Then, the result follows by applying Lemma 4.7 of [5].

However, the solution $\mathbf{H}$ of Problem **MP** does not satisfy, in general, $\mathbf{curl}\, \mathbf{H} = \mathbf{J} \in \mathrm{H}^r(\Omega)^3$ with $r > 1/2$. In fact, $\mathbf{J}|_{\Omega_{\mathrm{D}}} = 0$, whereas $\mathbf{J}|_{\Omega_{\mathrm{C}}} \times \mathbf{n} = (\frac{1}{\sigma}\mathbf{E})|_{\Omega_{\mathrm{C}}} \times \mathbf{n}$ in general does not vanish on Γ_{I}; thus, $\mathbf{J} \times \mathbf{n}$ has a jump across Γ_{I}. (See, for instance, the problem in section 7.) Nevertheless, typically $\mathbf{H}|_{\Omega_{\mathrm{C}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})$ and $\mathbf{H}|_{\Omega_{\mathrm{D}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$ with $r > 1/2$. This is enough for $\mathbf{H}^{\mathrm{I}}$ to be well defined as shown in the following lemma, which also provides an error estimate for the Nédélec interpolant under these assumptions. (Here and thereafter, C denotes a generic constant, not necessarily the same at each occurrence, but always independent of the mesh-size h.)

LEMMA 5.1. *Let $r \in (\frac{1}{2}, 1]$. The operator defined by* (5.2)–(5.1) *extends uniquely to the space* $\{\mathbf{G} \in \mathrm{H}(\mathbf{curl}, \Omega) : \ \mathbf{G}|_{\Omega_{\mathrm{C}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})$ *and* $\mathbf{G}|_{\Omega_{\mathrm{D}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})\}$. *Furthermore, for all* $\mathbf{G}$ *in this space,*

$$\|\mathbf{G} - \mathbf{G}^{\mathrm{I}}\|_{\mathrm{H}(\mathbf{curl},\Omega)} \leq Ch^r \left[\|\mathbf{G}\|_{\mathrm{H}^r(\mathbf{curl},\Omega_{\mathrm{C}})} + \|\mathbf{G}\|_{\mathrm{H}^r(\mathbf{curl},\Omega_{\mathrm{D}})} \right].$$

Proof. According to the discussion above, since $\mathbf{G}|_{\Omega_{\mathrm{C}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})$ and $\mathbf{G}|_{\Omega_{\mathrm{D}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$, with $r > 1/2$, the Nédélec interpolants of $\mathbf{G}|_{\Omega_{\mathrm{C}}}$ and $\mathbf{G}|_{\Omega_{\mathrm{D}}}$ are well defined in $\boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{C}})$ and $\boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$, respectively. Moreover, since $\mathbf{G} \in \mathrm{H}(\mathbf{curl}, \Omega)$, a density argument shows that the degrees of freedom corresponding to the edges $\ell \subset \Gamma_{\mathrm{I}}$ coincide for both interpolants. Thus the global interpolant $\mathbf{G}^{\mathrm{I}} \in \boldsymbol{\mathcal{N}}_h(\Omega)$ is well defined also in this case.

On the other hand, the arguments in the proof of Theorem 5.4 in [19] can be extended to this case to prove the error estimate above. ∎

In order to use these elements to discretize Problem **MP**, we have to use an approximant $\mathbf{f}_{\mathrm{I}}$ of the boundary data $\mathbf{f}$ such that a discrete version of (3.2) can hold true, namely, such that there exists $\mathbf{H}_h \in \boldsymbol{\mathcal{N}}_h(\Omega)$ satisfying $\mathbf{H}_h \times \mathbf{n} = \mathbf{f}_{\mathrm{I}}$.

To attain this goal, we will use the two-dimensional Nédélec interpolant of $\mathbf{n} \times \mathbf{f}$ on the triangular mesh induced by $\mathcal{T}_h$ on the polyhedral surface Γ_{D}. To introduce this interpolant, let $\mathcal{T}_h^{\Gamma_{\mathrm{D}}} := \{T \subset \Gamma_{\mathrm{D}} : \ T \text{ face of } K \in \mathcal{T}_h\}$. For each triangle $T \in \mathcal{T}_h^{\Gamma_{\mathrm{D}}}$, consider local orthogonal coordinates (ξ, η, ζ) such that T is contained in the plane $\zeta = 0$. Let

$$\boldsymbol{\mathcal{N}}^2(T) := \left\{ \boldsymbol{\varphi}_h \in \mathcal{P}_1(T)^3 : \ \boldsymbol{\varphi}_h(\xi, \eta, 0) = (a - c\eta, b + c\xi, 0), \ a, b, c \in \mathbb{C}, \ (\xi, \eta, 0) \in T \right\}.$$

This is the lowest-order two-dimensional Nédélec finite element (see [26]) on the plane $\zeta = 0$. The tangential components of these vector fields along the three edges of the triangle T can also be taken as the degrees of freedom defining them. Therefore, we define

$$\boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}}) := \Big\{ \boldsymbol{\varphi}_h \in \mathrm{L}^2(\Gamma_{\mathrm{D}})^3 : \ \boldsymbol{\varphi}_h|_T \in \boldsymbol{\mathcal{N}}^2(T) \ \forall T \in \mathcal{T}_h^{\Gamma_{\mathrm{D}}} \\ \text{and } \boldsymbol{\varphi}_h \cdot \mathbf{t}_\ell \text{ is continuous on } \ell \ \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}} \Big\}.$$

Let $\boldsymbol{\varphi}$ be a tangential vector field on Γ_{D} (i.e., satisfying $\boldsymbol{\varphi} \cdot \mathbf{n} = 0$ on Γ_{D}). If $\boldsymbol{\varphi}$ is sufficiently smooth (e.g., $\boldsymbol{\varphi} \in \mathrm{H}^1(\Gamma_{\mathrm{D}})^3$), then its Nédélec interpolant on Γ_{D}, which we denote by $\boldsymbol{\varphi}^{\mathrm{I}_2}$, is defined by

$$\boldsymbol{\varphi}^{\mathrm{I}_2} \in \boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}}) : \quad \int_\ell \boldsymbol{\varphi}^{\mathrm{I}_2} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \boldsymbol{\varphi} \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}}. \tag{5.3}$$

If $\mathbf{G}$ is smooth enough in Ω_{D} (e.g., $\mathbf{G} \in \mathrm{H}^2(\Omega_{\mathrm{D}})^3$), then its tangential trace on Γ_{D}, $\mathbf{n} \times (\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n})$, is smooth too and satisfies

$$\Big[\mathbf{n} \times \Big(\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \Big]^{\mathrm{I}_2} = \mathbf{n} \times \Big(\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \quad \text{on } \Gamma_{\mathrm{D}}. \tag{5.4}$$

Indeed, a straightforward computation shows that the right-hand side above also belongs to $\boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}})$. On the other hand, (5.1) implies

$$\int_\ell \mathbf{n} \times \Big(\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{G}^{\mathrm{I}} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{G} \cdot \mathbf{t}_\ell \, d\gamma \\ = \int_\ell \mathbf{n} \times \Big(\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}}.$$

Thus, the degrees of freedom defining both sides of (5.4) coincide and, consequently, (5.4) holds true.

The following lemma shows that a similar result is valid for $\mathbf{G} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$.

LEMMA 5.2. *Let $r \in (\frac{1}{2}, 1]$. The linear operator*

$$\begin{array}{rcl} \mathrm{H}^2(\Omega_{\mathrm{D}})^3 & \longrightarrow & \boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}}), \\ \mathbf{G} & \longmapsto & \big[\mathbf{n} \times (\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}) \big]^{\mathrm{I}_2}, \end{array} \tag{5.5}$$

with $(\cdot)^{\mathrm{I}_2}$ defined by (5.3)*, extends uniquely to $\mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$. Furthermore,* (5.4) *holds true for all $\mathbf{G}$ in this space.*

Proof. As said above, if $\mathbf{G} \in \mathrm{H}^2(\Omega_{\mathrm{D}})^3$, then $\big[\mathbf{n} \times (\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}) \big]^{\mathrm{I}_2} \in \boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}})$ is defined by

$$\int_\ell \Big[\mathbf{n} \times \Big(\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \Big]^{\mathrm{I}_2} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{G} \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}}.$$

Then, by repeating the arguments in the proof of Lemma 5.1 (i.e., using the Sobolev imbedding theorem and Lemma 4.7 of [5]) we prove that the operator defined by (5.5) and (5.3) extends uniquely to $\mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$ for $r > 1/2$.

Furthermore, we have also shown above that, for $\mathbf{G} \in \mathrm{H}^2(\Omega_{\mathrm{D}})$,

$$\int_\ell \Big[\mathbf{n} \times \Big(\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \Big]^{\mathrm{I}_2} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{G} \cdot \mathbf{t}_\ell \, d\gamma \\ = \int_\ell \mathbf{n} \times \Big(\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n} \Big) \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}}.$$

Then, a density argument and the fact that $\mathbf{n} \times (\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}) \in \boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}})$ allow us to conclude that (5.4) holds true for all $\mathbf{G} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$. ∎

If the data $\mathbf{f}$ of Problem **MP** are sufficiently smooth, we define

$$\mathbf{f}_{\mathrm{I}} := (\mathbf{n} \times \mathbf{f})^{\mathrm{I}_2} \times \mathbf{n}; \tag{5.6}$$

that is, $\mathbf{f}_{\mathrm{I}}$ is such that $\mathbf{n} \times \mathbf{f}_{\mathrm{I}} = (\mathbf{n} \times \mathbf{f})^{\mathrm{I}_2}$, which means

$$\mathbf{n} \times \mathbf{f}_{\mathrm{I}} \in \boldsymbol{\mathcal{N}}_h^2(\Gamma_{\mathrm{D}}): \quad \int_\ell \mathbf{n} \times \mathbf{f}_{\mathrm{I}} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{n} \times \mathbf{f} \cdot \mathbf{t}_\ell \, d\gamma \quad \forall \ell \text{ edge of } \mathcal{T}_h^{\Gamma_{\mathrm{D}}}.$$

The following lemma shows that this definition also works under the same weak smoothness assumptions of the previous lemmas.

LEMMA 5.3. *Let* $\mathbf{G} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$*, with* $r > 1/2$*, and let* $\mathbf{g} = \mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}$*. Then* $\mathbf{g}_{\mathrm{I}} := (\mathbf{n} \times \mathbf{g})^{\mathrm{I}_2} \times \mathbf{n}$ *is well defined and satisfies*

$$\mathbf{n} \times \mathbf{g}_{\mathrm{I}} = \mathbf{n} \times (\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}) \quad \text{on } \Gamma_{\mathrm{D}}.$$

Proof. As a consequence of Lemma 5.2, $(\mathbf{n} \times \mathbf{g})^{\mathrm{I}_2} = \left[\mathbf{n} \times (\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n})\right]^{\mathrm{I}_2}$ is well defined. Hence $\mathbf{g}_{\mathrm{I}} := (\mathbf{n} \times \mathbf{g})^{\mathrm{I}_2} \times \mathbf{n}$ is well defined, too. Moreover, according to this lemma, (5.4) holds true for $\mathbf{G} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$; thus,

$$\mathbf{n} \times \mathbf{g}_{\mathrm{I}} = (\mathbf{n} \times \mathbf{g})^{\mathrm{I}_2} = \left[\mathbf{n} \times (\mathbf{G}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n})\right]^{\mathrm{I}_2} = \mathbf{n} \times (\mathbf{G}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}) \quad \text{on } \Gamma_{\mathrm{D}}. \qquad ∎$$

Now we are in a position to discretize Problem **MP**. We introduce the following finite-dimensional spaces:

$$\begin{aligned} \boldsymbol{\mathcal{V}}_h &:= \{\mathbf{G}_h \in \boldsymbol{\mathcal{N}}_h(\Omega): \ \mathbf{curl}\, \mathbf{G}_h = \mathbf{0} \ \text{ in } \Omega_{\mathrm{D}}\}, \\ \boldsymbol{\mathcal{V}}_h^0 &:= \{\mathbf{G}_h \in \boldsymbol{\mathcal{V}}_h: \ \mathbf{G}_h \times \mathbf{n} = \mathbf{0} \ \text{ on } \Gamma_{\mathrm{D}}\}. \end{aligned}$$

Finally, we define the discrete magnetic problem as follows.

PROBLEM **DMP**. *Find* $\mathbf{H}_h \in \boldsymbol{\mathcal{V}}_h$ *such that*

$$\begin{aligned} &\mathbf{H}_h \times \mathbf{n} = \mathbf{f}_{\mathrm{I}} \quad \text{on } \Gamma_{\mathrm{D}}, \\ &\mathrm{i}\omega \int_\Omega \mu \mathbf{H}_h \cdot \bar{\mathbf{G}}_h + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \mathbf{curl}\, \mathbf{H}_h \cdot \mathbf{curl}\, \bar{\mathbf{G}}_h = 0 \quad \forall \mathbf{G}_h \in \boldsymbol{\mathcal{V}}_h^0. \end{aligned}$$

It is straightforward to prove existence and uniqueness of solution for this problem under mild smoothness assumptions on the solution of Problem **MP**. Moreover, an error estimate can be deduced from the standard finite element approximation theory.

THEOREM 5.4. *Let us assume that the solution* $\mathbf{H}$ *of Problem* **MP** *satisfies* $\mathbf{H}|_{\Omega_{\mathrm{C}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})$ *and* $\mathbf{H}|_{\Omega_{\mathrm{D}}} \in \mathrm{H}^r(\Omega_{\mathrm{D}})^3$*, with* $r \in (\frac{1}{2}, 1]$*. Then,* $\mathbf{f}_{\mathrm{I}}$ *is well defined by* (5.6)*, Problem* **DMP** *attains a unique solution* $\mathbf{H}_h$*, and*

$$\|\mathbf{H} - \mathbf{H}_h\|_{\mathrm{H}(\mathbf{curl},\Omega)} \leq C h^r \left[\|\mathbf{H}\|_{\mathrm{H}^r(\mathbf{curl},\Omega_{\mathrm{C}})} + \|\mathbf{H}\|_{\mathrm{H}^r(\Omega_{\mathrm{D}})^3}\right].$$

Proof. Since $\mathbf{H} \in \boldsymbol{\mathcal{V}}$, $\mathbf{curl}\, \mathbf{H} = \mathbf{0}$ in Ω_{D}. Hence, $\mathbf{H}|_{\Omega_{\mathrm{D}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{D}})$. Therefore, according to Lemma 5.1, its Nédélec interpolant $\mathbf{H}^{\mathrm{I}} \in \boldsymbol{\mathcal{N}}_h(\Omega)$ is well defined and satisfies

$$\|\mathbf{H} - \mathbf{H}^{\mathrm{I}}\|_{\mathrm{H}(\mathbf{curl},\Omega)} \leq C h^r \left[\|\mathbf{H}\|_{\mathrm{H}^r(\mathbf{curl},\Omega_{\mathrm{C}})} + \|\mathbf{H}\|_{\mathrm{H}^r(\Omega_{\mathrm{D}})^3}\right]. \tag{5.7}$$

Moreover, the arguments of Remark 5.6 in [19] can be extended to this case to prove that $\mathbf{curl}\,\mathbf{H}|_{\Omega_{\mathrm{D}}} = \mathbf{0}$ implies that $\mathbf{curl}\,\mathbf{H}^{\mathrm{I}}|_{\Omega_{\mathrm{D}}} = \mathbf{0}$. Therefore, $\mathbf{H}^{\mathrm{I}} \in \boldsymbol{\mathcal{V}}_h$.

On the other hand, because of Lemma 5.3, $\mathbf{f}_{\mathrm{I}}$ is well defined by (5.6) and satisfies $\mathbf{f}_{\mathrm{I}} = \mathbf{H}^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}} \times \mathbf{n}$ on Γ_{D}. Thus, we have proved that there exists $\mathbf{H}^{\mathrm{I}} \in \boldsymbol{\mathcal{V}}_h$ such that $\mathbf{H}^{\mathrm{I}} \times \mathbf{n} = \mathbf{f}_{\mathrm{I}}$ on Γ_{D}. Hence, since $\boldsymbol{\mathcal{V}}_h^0 \subset \boldsymbol{\mathcal{V}}^0$, the arguments in the proof of Theorem 3.1 also apply to Problem **DMP**, allowing us to prove existence and uniqueness of a solution $\mathbf{H}_h$ of this problem.

Finally, to prove the error estimate, notice that since $\boldsymbol{\mathcal{V}}_h^0 \subset \boldsymbol{\mathcal{V}}^0$,

$$a(\mathbf{H} - \mathbf{H}_h, \mathbf{G}_h) = 0 \quad \forall \mathbf{G}_h \in \boldsymbol{\mathcal{V}}_h^0.$$

Hence, since $\mathbf{H}_h \times \mathbf{n} = \mathbf{f}_{\mathrm{I}} = \mathbf{H}^{\mathrm{I}} \times \mathbf{n}$ on Γ_{D}, $\mathbf{H}_h - \mathbf{H}^{\mathrm{I}} \in \boldsymbol{\mathcal{V}}_h^0$. Therefore, because of this and (3.4),

$$\begin{aligned}\alpha\|\mathbf{H} - \mathbf{H}_h\|^2_{\mathrm{H}(\mathbf{curl},\Omega)} &\le |a(\mathbf{H} - \mathbf{H}_h, \mathbf{H} - \mathbf{H}_h)| = \left|a(\mathbf{H} - \mathbf{H}_h, \mathbf{H} - \mathbf{H}^{\mathrm{I}})\right| \\ &\le C\|\mathbf{H} - \mathbf{H}_h\|_{\mathrm{H}(\mathbf{curl},\Omega)}\left\|\mathbf{H} - \mathbf{H}^{\mathrm{I}}\right\|_{\mathrm{H}(\mathbf{curl},\Omega)},\end{aligned}$$

which together with estimate (5.7) allow us to conclude the proof. ∎

5.2. Discretizing the magnetic potential. Problem **DMP** is actually just a "theoretical" method in that its solution requires to impose somehow the curl-free condition in the definition of $\boldsymbol{\mathcal{V}}_h$ to trial and test functions. In what follows we show how to deal efficiently with this curl-free condition by introducing a discrete multivalued magnetic potential in the dielectric domain.

We assume that the cut surfaces Σ_j are polyhedral and that the meshes are compatible with them, in the sense that each $\overline{\Sigma}_j$ is a union of faces of tetrahedra $K \in \mathcal{T}_h$ for each mesh $\mathcal{T}_h$. Therefore, $\mathcal{T}_h^{\Omega_{\mathrm{D}}} := \{K \in \mathcal{T}_h : \ K \subset \overline{\Omega}_{\mathrm{D}}\}$ can also be seen as a mesh of $\widetilde{\Omega}_{\mathrm{D}}$.

First, we introduce an approximation of the space Θ. Let us denote

$$\mathcal{L}_h(\widetilde{\Omega}_{\mathrm{D}}) := \left\{\widetilde{\Psi}_h \in \mathrm{H}^1(\widetilde{\Omega}_{\mathrm{D}}) : \ \widetilde{\Psi}_h|_K \in \mathcal{P}_1(K)\ \forall K \in \mathcal{T}_h^{\Omega_{\mathrm{D}}}\right\}.$$

Then, we consider the family of finite-dimensional subspaces of Θ given by

$$\Theta_h := \{\widetilde{\Psi}_h \in \mathcal{L}_h(\widetilde{\Omega}_{\mathrm{D}}) : \ [\![\widetilde{\Psi}_h]\!]_{\Sigma_j} = \text{constant},\ j = 1, \ldots, J\}.$$

The following lemma shows that the curl-free vector fields in $\boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$ admit a multivalued potential in Θ_h.

LEMMA 5.5. *Let $\mathbf{G}_h \in \mathrm{L}^2(\Omega_{\mathrm{D}})^3$. Then $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$ with $\mathbf{curl}\,\mathbf{G}_h = \mathbf{0}$ in Ω_{D} if and only if there exists $\widetilde{\Psi}_h \in \Theta_h$ such that $\mathbf{G}_h = \widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h$ in Ω_{D}. Such $\widetilde{\Psi}_h$ is unique up to an additive constant.*

Proof. According to (4.1), $\mathbf{curl}\,\mathbf{G}_h = \mathbf{0}$ in Ω_{D} if and only if there exists $\widetilde{\Psi}_h \in \Theta$ such that $\mathbf{G}_h = \widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h$ in $\widetilde{\Omega}_{\mathrm{D}}$. Moreover, since $\widetilde{\Omega}_{\mathrm{D}}$ is connected, $\widetilde{\Psi}_h$ is unique up to an additive constant. Now, let $K \in \mathcal{T}_h^{\Omega_{\mathrm{D}}}$ be a tetrahedron of the mesh. A direct calculation shows that $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}(K)$ with $\mathbf{curl}\,\mathbf{G}_h|_K = \mathbf{0}$ if and only if $\mathbf{G}_h|_K \in \mathcal{P}_0(K)^3$, or, equivalently, if and only if $\widetilde{\Psi}_h|_K \in \mathcal{P}_1(K)^3$. Thus the lemma follows from the definition of Θ_h. ∎

Let us introduce the following families of finite-dimensional approximations of $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{W}}^0$, respectively:

$$\boldsymbol{\mathcal{W}}_h := \left\{ (\mathbf{G}_h, \widetilde{\Psi}_h) \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{C}}) \times (\Theta_h/\mathbb{C}) : \ (\mathbf{G}_h | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h) \in \mathrm{H}(\mathbf{curl}, \Omega) \right\},$$
$$\boldsymbol{\mathcal{W}}_h^0 := \left\{ (\mathbf{G}_h, \widetilde{\Psi}_h) \in \boldsymbol{\mathcal{W}}_h : \ \widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h \times \mathbf{n} = \mathbf{0} \text{ on } \Gamma_{\mathrm{D}} \right\}.$$

By virtue of Lemma 5.5, $\boldsymbol{\mathcal{W}}_h$ and $\boldsymbol{\mathcal{W}}_h^0$ are isomorphically equivalent to $\boldsymbol{\mathcal{V}}_h$ and $\boldsymbol{\mathcal{V}}_h^0$, respectively. Thus, we define the following discrete problem which turns out to be equivalent to Problem **DMP**.

PROBLEM **DHP**. *Find* $(\mathbf{H}_h, \widetilde{\Phi}_h) \in \boldsymbol{\mathcal{W}}_h$ *such that*

$$\widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \times \mathbf{n} = \mathbf{f}_{\mathrm{I}} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{5.8}$$

$$\mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu \mathbf{H}_h \cdot \bar{\mathbf{G}}_h + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \mathbf{curl}\, \mathbf{H}_h \cdot \mathbf{curl}\, \bar{\mathbf{G}}_h + \mathrm{i}\omega \int_{\Omega_{\mathrm{D}}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \cdot \widetilde{\mathbf{grad}}\, \bar{\widetilde{\Psi}}_h = 0 \qquad \forall (\mathbf{G}_h, \widetilde{\Psi}_h) \in \boldsymbol{\mathcal{W}}_h^0. \tag{5.9}$$

Clearly, the following discrete analogue of Lemma 4.1 holds true.

LEMMA 5.6. *The pair* $(\mathbf{H}_h, \widetilde{\Phi}_h)$ *is a solution of Problem* **DHP** *if and only if* $(\mathbf{H}_h | \, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h)$ *is a solution of Problem* **DMP**.

As an immediate consequence of these two lemmas, Theorem 5.4 yields an error estimate for the approximation obtained from Problem **DHP**.

COROLLARY 5.7. *Let us assume that the solution* $(\mathbf{H}, \widetilde{\Phi})$ *of Problem* **HP** *satisfies* $\mathbf{H} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})$ *and* $\widetilde{\mathbf{grad}}\, \widetilde{\Phi} \in \mathrm{H}^r(\Omega_{\mathrm{D}})^3$, *with* $r \in (\frac{1}{2}, 1]$. *Then, Problem* **DHP** *is well posed, it attains a unique solution* $(\mathbf{H}_h, \widetilde{\Phi}_h)$, *and*

$$\|\mathbf{H} - \mathbf{H}_h\|_{\mathrm{H}(\mathbf{curl}, \Omega_{\mathrm{C}})} + \|\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi} - \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h\|_{\mathrm{L}^2(\Omega_{\mathrm{D}})^3} \leq Ch^r \left[\|\mathbf{H}\|_{\mathrm{H}^r(\mathbf{curl}, \Omega_{\mathrm{C}})} + \|\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}\|_{\mathrm{H}^r(\Omega_{\mathrm{D}})^3} \right].$$

6. Computer implementation. For Problem **DHP** to be useful for computational purposes, we have to introduce effective procedures to impose the following constraints:

1. $(\mathbf{G}_h | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h) \in \mathrm{H}(\mathbf{curl}, \Omega)$, which arises in the definition of $\boldsymbol{\mathcal{W}}_h$;
2. $[\![\widetilde{\Psi}_h]\!]_{\Sigma_j} =$ constant, which arises in the definition of Θ_h;
3. the boundary condition $\widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \times \mathbf{n} = \mathbf{f}_{\mathrm{I}}$ on Γ_{D}.

We fix some notation to deal with these constraints. We choose an orientation for each edge ℓ of the mesh $\mathcal{T}_h$ and denote P_ℓ^- and P_ℓ^+ its initial and end points, respectively, and $\mathbf{t}_\ell$ its unit tangent vector pointing from P_ℓ^- to P_ℓ^+.

Regarding the first constraint we have the following result.

LEMMA 6.1. *Let* $(\mathbf{G}_h, \widetilde{\Psi}_h) \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{C}}) \times (\Theta_h/\mathbb{C})$. *Then,* $(\mathbf{G}_h, | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h) \in \mathrm{H}(\mathbf{curl}, \Omega)$ *if and only if*

$$\int_\ell \mathbf{G}_h \cdot \mathbf{t}_\ell \, d\gamma = \widetilde{\Psi}_h(P_\ell^+) - \widetilde{\Psi}_h(P_\ell^-) \quad \forall \ell \text{ edge of } \mathcal{T}_h : \ \ell \subset \Gamma_{\mathrm{I}}.$$

Proof. Since $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{C}})$ and $\widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$, $(\mathbf{G}_h | \, \widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h) \in \mathrm{H}(\mathbf{curl}, \Omega)$ if and only if the tangential traces on Γ_{I} of $\mathbf{G}_h$ and $\widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h$ coincide; that is, if and

only if

$$\mathbf{n}\times(\mathbf{G}_h\times\mathbf{n}) = \mathbf{n}\times\Big(\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h\times\mathbf{n}\Big) \quad \text{on } \Gamma_{\rm I}.$$

Now, the equation above holds true if and only if the degrees of freedom of $\mathbf{G}_h$ and $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h$ coincide on all the edges $\ell\subset\Gamma_{\rm I}$, and this reads as

$$\int_\ell \mathbf{G}_h\cdot\mathbf{t}_\ell\,d\gamma = \int_\ell \widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h\cdot\mathbf{t}_\ell\,d\gamma = \widetilde{\Psi}_h(P_\ell^+)-\widetilde{\Psi}_h(P_\ell^-). \qquad \square$$

This lemma shows that the constraint $(\mathbf{G}_h|\,\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h)\in \mathrm{H}(\mathbf{curl},\Omega)$ can be readily imposed by eliminating the degrees of freedom of $\mathbf{G}_h$ associated with the edges $\ell\subset\Gamma_{\rm I}$, in terms of those of $\widetilde{\Phi}_h$ corresponding to the vertices of the mesh on this interface.

Regarding the second constraint, for $\widetilde{\Psi}_h\in\Theta_h$, let us denote

$$c_{hj} := [\![\widetilde{\Psi}_h]\!]_{\Sigma_j}, \qquad j=1,\dots,J.$$

In order to handle the multivalued character of the functions $\widetilde{\Psi}_h\in\Theta_h$, for each cut surface Σ_j, we in principle distinguish the degrees of freedom of $\widetilde{\Psi}_h$ on $\overline{\Sigma}_j^+$ from those on $\overline{\Sigma}_j^-$. Then, the latter can be eliminated by using

$$\widetilde{\Psi}_h|_{\overline{\Sigma}_j^-} = \widetilde{\Psi}_h|_{\overline{\Sigma}_j^+} + [\![\widetilde{\Psi}_h]\!]_{\overline{\Sigma}_j} = \widetilde{\Psi}_h|_{\overline{\Sigma}_j^+} + c_{hj}, \qquad j=1,\dots,J.$$

This elimination must be carried out for the solution $(\mathbf{H}_h,\widetilde{\Phi}_h)\in\boldsymbol{\mathcal{W}}_h$ of Problem **DHP** as well as for the test functions $(\mathbf{G}_h,\widetilde{\Psi}_h)\in\boldsymbol{\mathcal{W}}_h^0$. For the former, the arguments in Remark 4.1 can be repeated at discrete level to show that each jump $[\![\widetilde{\Phi}_h]\!]_{\Sigma_j}$ represents the current intensity through the conductor $\Omega_{\rm C}^j$ corresponding to the discrete solution $(\mathbf{H}_h,\widetilde{\Phi}_h)$. Because of this, we denote these jumps

$$I_{hj} := [\![\widetilde{\Phi}_h]\!]_{\Sigma_j}, \qquad j=1,\dots,J.$$

For $j=1,\dots,K$ (i.e., for inner conductors $\Omega_{\rm C}^j$), I_{hj} are additional unknowns of the discrete problem. Instead, for $j=K+1,\dots,J$ (i.e., for conductors $\Omega_{\rm C}^j$ going through $\partial\Omega$), I_{hj} can be computed in advance from the data of the discrete problem. Indeed, since $\gamma_j\subset\Gamma_{\rm D}$ and $\widetilde{\mathbf{grad}}\,\widetilde{\Phi}_h\times\mathbf{n}=\mathbf{f}_{\rm I}$ on $\Gamma_{\rm D}$,

$$I_{hj} = \int_{\gamma_j}\widetilde{\mathbf{grad}}\,\widetilde{\Phi}_h\cdot\mathbf{t}_j\,d\gamma = \int_{\gamma_j}\mathbf{n}\times\Big(\widetilde{\mathbf{grad}}\,\widetilde{\Phi}_h\times\mathbf{n}\Big)\cdot\mathbf{t}_j\,d\gamma = \int_{\gamma_j}\mathbf{n}\times\mathbf{f}_{\rm I}\cdot\mathbf{t}_j\,d\gamma\,. \tag{6.1}$$

For the test functions $(\mathbf{G}_h,\widetilde{\Psi}_h)\in\boldsymbol{\mathcal{W}}_h^0$, by repeating these arguments and using that $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h\times\mathbf{n}=\mathbf{0}$ on $\Gamma_{\rm D}$, we have

$$c_{hj} = \int_{\gamma_j}\mathbf{n}\times\Big(\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h\times\mathbf{n}\Big)\cdot\mathbf{t}_j\,d\gamma = 0, \qquad j=K+1,\dots,J. \tag{6.2}$$

Hence, only the constants c_{hj}, for $j=1,\dots,K$, must be taken into account as genuine degrees of freedom in the definition of $\boldsymbol{\mathcal{W}}_h^0$.

Remark 6.1. The computed and exact intensities through the conductors $\Omega_{\rm C}^j$, $j = K+1, \ldots, J$, coincide. Indeed, because of (6.1), Lemma 5.3, (5.1), the fact that each γ_j is union of edges ℓ in $\mathcal{T}_h$, and Remark 4.1, we have

$$\begin{aligned} I_{hj} &= \int_{\gamma_j} \mathbf{n} \times \mathbf{f}_{\rm I} \cdot \mathbf{t}_j \, d\gamma = \int_{\gamma_j} \mathbf{n} \times \left[(\widetilde{\mathbf{grad}}\, \widetilde{\Phi})^{\rm I}|_{\Gamma_{\rm D}} \times \mathbf{n} \right] \cdot \mathbf{t}_j \, d\gamma \\ &= \int_{\gamma_j} (\widetilde{\mathbf{grad}}\, \widetilde{\Phi})^{\rm I} \cdot \mathbf{t}_j \, d\gamma = \int_{\gamma_j} \widetilde{\mathbf{grad}}\, \widetilde{\Phi} \cdot \mathbf{t}_j \, d\gamma = [\![\widetilde{\Phi}]\!]_{\Sigma_j} = I_j. \end{aligned}$$

Regarding the third constraint, we impose the boundary condition by means of a Lagrange multiplier. Let $\widetilde{\Gamma}_{\rm D}$ be the pseudo-Lipschitz connected polyhedral surface defined by

$$\widetilde{\Gamma}_{\rm D} := \Gamma_{\rm D} \setminus \bigcup_{j=K+1}^{J} \left(\overline{\Sigma}_j \cap \Gamma_{\rm D} \right).$$

Let

$$\begin{aligned} \mathcal{L}_h(\widetilde{\Gamma}_{\rm D}) &:= \left\{ \nu_h \in {\rm H}^1(\widetilde{\Gamma}_{\rm D}) : \ \nu_h|_T \in \mathcal{P}_1(T) \ \forall T \in \mathcal{T}_h^{\Gamma_{\rm D}} \right\}, \\ \mathcal{L}_h(\Gamma_{\rm D}) &:= \left\{ \nu_h \in {\rm H}^1(\Gamma_{\rm D}) : \ \nu_h|_T \in \mathcal{P}_1(T) \ \forall T \in \mathcal{T}_h^{\Gamma_{\rm D}} \right\}. \end{aligned}$$

Hence, given $\nu_h \in \mathcal{L}_h(\widetilde{\Gamma}_{\rm D})$, we have that $\nu_h \in \mathcal{L}_h(\Gamma_{\rm D})$ if and only if $[\![\nu_h]\!]_{\overline{\Sigma}_j \cap \Gamma_{\rm D}} = 0$ for $j = K+1, \ldots, J$.

Let $\widetilde{\mathbf{grad}}_{\,\Gamma}$ denote the surface gradient operator. Since we will use this operator acting only on piecewise linear functions, we give a definition valid in this case (for its general definition on polyhedral surfaces, see [14, 15]):

$$\widetilde{\mathbf{grad}}_{\,\Gamma} \ : \ \mathcal{L}_h(\widetilde{\Gamma}_{\rm D}) \longrightarrow {\rm L}^2_{\mathbf{t}}(\Gamma_{\rm D})^3 := \left\{ \boldsymbol{\varphi} \in {\rm L}^2(\Gamma_{\rm D})^3 : \ \boldsymbol{\varphi} \cdot \mathbf{n} = 0 \text{ on } \Gamma_{\rm D} \right\}$$

is defined, on each element $T \in \mathcal{T}_h^{\Gamma_{\rm D}}$, by $(\widetilde{\mathbf{grad}}_{\,\Gamma}\, \nu_h)|_T = \nabla_2(\nu_h|_T)$, where ∇_2 is the usual gradient of a function of two variables; i.e., using local coordinates (ξ, η, ζ) such that T is in the plane $\zeta = 0$,

$$(\widetilde{\mathbf{grad}}_{\,\Gamma}\, \nu_h)|_T := \left(\frac{\partial(\nu_h|_T)}{\partial \xi}, \frac{\partial(\nu_h|_T)}{\partial \eta}, 0 \right).$$

For all $\widetilde{\Psi}_h \in \mathcal{L}_h(\widetilde{\Omega}_{\rm D})$, we have $\widetilde{\Psi}_h|_{\Gamma_{\rm D}} \in \mathcal{L}_h(\widetilde{\Gamma}_{\rm D})$, and it is straightforward to show that

$$\widetilde{\mathbf{grad}}_{\,\Gamma}\, (\widetilde{\Psi}_h|_{\widetilde{\Gamma}_{\rm D}}) = \mathbf{n} \times \left[(\widetilde{\mathbf{grad}}\, \widetilde{\Psi}_h)|_{\Gamma_{\rm D}} \times \mathbf{n} \right] \quad \text{a.e. in } \Gamma_{\rm D}. \tag{6.3}$$

The following lemma provides a weak formulation of the boundary condition (5.8) in Problem **DHP**.

LEMMA 6.2. *Let* $\widetilde{\Psi} \in \Theta$ *be such that* $\widetilde{\mathbf{grad}}\, \widetilde{\Psi} \in {\rm H}^r(\Omega_{\rm D})$ *with* $r > 1/2$. *Let* $\mathbf{g} = \widetilde{\mathbf{grad}}\, \widetilde{\Psi}|_{\Gamma_{\rm D}} \times \mathbf{n}$ *and* $\mathbf{g}_{\rm I} = (\mathbf{n} \times \mathbf{g})^{{\rm I}_2} \times \mathbf{n}$ *(well defined because of Lemma* 5.3*). Let* $\widetilde{\Psi}_h \in \Theta_h$ *be such that*

$$[\![\widetilde{\Psi}_h]\!]_{\Sigma_j} = \int_{\gamma_j} \mathbf{n} \times \mathbf{g}_{\rm I} \cdot \mathbf{t}_j \, d\gamma, \qquad j = K+1, \ldots, J. \tag{6.4}$$

Then, $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h \times \mathbf{n} = \mathbf{g}_{\mathrm{I}}$ *on* Γ_{D} *if and only if*

$$\int_{\Gamma_{\mathrm{D}}} \widetilde{\mathbf{grad}}_{\,\Gamma}\,\widetilde{\Psi}_h \cdot \mathbf{grad}_{\Gamma}\,\bar{\nu}_h\,d\Gamma = \int_{\Gamma_{\mathrm{D}}} \mathbf{n}\times\mathbf{g}_{\mathrm{I}}\cdot\mathbf{grad}_{\Gamma}\,\bar{\nu}_h\,d\Gamma \quad \forall \nu_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}. \tag{6.5}$$

Proof. If $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}_h \times \mathbf{n} = \mathbf{g}_{\mathrm{I}}$ on Γ_{D}, then, because of (6.3), we have (6.5).

Conversely, let us assume that (6.5) holds true. Since $\widetilde{\Psi} \in \mathrm{H}^{1+r}(\widetilde{\Omega}_{\mathrm{D}})/\mathbb{C}$ with $r > 1/2$, its Lagrange interpolant $\widetilde{\Psi}^{\mathrm{L}} \in \mathcal{L}_h(\widetilde{\Omega}_{\mathrm{D}})/\mathbb{C}$ is well defined. Given a cut surface Σ_j, $j = 1,\dots,J$, for all the vertices P of $\mathcal{T}_h$ such that $P \in \Sigma_j$, we have $[\![\widetilde{\Psi}^{\mathrm{L}}(P)]\!]_{\Sigma_j} = [\![\widetilde{\Psi}(P)]\!]_{\Sigma_j} =$ constant (the same for all such P). Then $[\![\widetilde{\Psi}^{\mathrm{L}}]\!]_{\Sigma_j} =$ constant, and hence $\widetilde{\Psi}^{\mathrm{L}} \in \Theta_h/\mathbb{C}$. Thus, because of Lemma 5.5, $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}} \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$.

On the other hand, let $(\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}} \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$ be the Nédélec interpolant of $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}$. We have $\int_\ell (\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}}\cdot\mathbf{t}_\ell\,d\gamma = \int_\ell \widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}}\cdot\mathbf{t}_\ell\,d\gamma$ for all edges ℓ of $\mathcal{T}_h^{\Omega_{\mathrm{D}}}$. Indeed, if $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}$ were smooth (e.g., $\widetilde{\mathbf{grad}}\,\widetilde{\Psi} \in \mathrm{H}^2(\Omega_{\mathrm{D}})^3$), then

$$\begin{aligned}\int_\ell (\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}}\cdot\mathbf{t}_\ell\,d\gamma &= \int_\ell \widetilde{\mathbf{grad}}\,\widetilde{\Psi}\cdot\mathbf{t}_\ell\,d\gamma = \widetilde{\Psi}(P_\ell^+) - \widetilde{\Psi}(P_\ell^-)\\ &= \widetilde{\Psi}^{\mathrm{L}}(P_\ell^+) - \widetilde{\Psi}^{\mathrm{L}}(P_\ell^-) = \int_\ell \widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}}\cdot\mathbf{t}_\ell\,d\gamma\,.\end{aligned}$$

Hence, because of a density argument, this is also true for $\widetilde{\mathbf{grad}}\,\widetilde{\Psi} \in \mathrm{H}^r(\mathbf{curl},\Omega_{\mathrm{D}})^3$.

Therefore, since we have shown that $\widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}} \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{D}})$ and that its degrees of freedom coincide with those of $(\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}}$ for all edges ℓ of $\mathcal{T}_h^{\Omega_{\mathrm{D}}}$, we have

$$(\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}} = \widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}} \quad \text{in } \overline{\Omega}_{\mathrm{D}}. \tag{6.6}$$

Consequently,

$$\begin{aligned}[\![\widetilde{\Psi}^{\mathrm{L}}]\!]_{\Sigma_j} &= \int_{\gamma_j} \widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}}\cdot\mathbf{t}_j\,d\gamma = \int_{\gamma_j} (\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}}\cdot\mathbf{t}_j\,d\gamma\\ &= \int_{\gamma_j} \mathbf{n}\times\Big[(\widetilde{\mathbf{grad}}\,\widetilde{\Psi})^{\mathrm{I}}|_{\Gamma_{\mathrm{D}}}\times\mathbf{n}\Big]\cdot\mathbf{t}_j\,d\gamma = \int_{\gamma_j}\mathbf{n}\times\mathbf{g}_{\mathrm{I}}\cdot\mathbf{t}_j\,d\gamma\,,\quad j = K+1,\dots,J,\end{aligned}$$

the last equality because of Lemma 5.3.

Let $\nu_h := (\widetilde{\Psi}_h - \widetilde{\Psi}^{\mathrm{L}})|_{\widetilde{\Gamma}_{\mathrm{D}}} \in \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{D}})/\mathbb{C}$. Because of the equation above and (6.4), we have

$$[\![\nu_h]\!]_{\overline{\Sigma}_j\cap\Gamma_{\mathrm{D}}} = [\![\widetilde{\Psi}_h]\!]_{\Sigma_j} - [\![\widetilde{\Psi}^{\mathrm{L}}]\!]_{\Sigma_j} = 0, \qquad j = K+1,\dots,J.$$

Then, $\nu_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}$, and, because of (6.3), (6.6), and Lemma 5.3, we have

$$\mathbf{grad}_{\Gamma}\,\nu_h = \widetilde{\mathbf{grad}}_{\,\Gamma}\,(\widetilde{\Psi}_h|_{\widetilde{\Gamma}_{\mathrm{D}}}) - \mathbf{n}\times\Big[(\widetilde{\mathbf{grad}}\,\widetilde{\Psi}^{\mathrm{L}})|_{\Gamma_{\mathrm{D}}}\times\mathbf{n}\Big] = \widetilde{\mathbf{grad}}_{\,\Gamma}\,(\widetilde{\Psi}_h|_{\widetilde{\Gamma}_{\mathrm{D}}}) - \mathbf{n}\times\mathbf{g}_{\mathrm{I}}.$$

Thus, using this ν_h in (6.5), we obtain

$$\widetilde{\mathbf{grad}}_{\,\Gamma}\,(\widetilde{\Psi}_h|_{\widetilde{\Gamma}_{\mathrm{D}}}) - \mathbf{n}\times\mathbf{g}_{\mathrm{I}} = \mathbf{0}.$$

Hence, by using again (6.3), we conclude the proof. ∎

Now we are in a position to set a new discrete problem including the three constraints as we have just described. To this end, we introduce the following discrete spaces:

$$\boldsymbol{\mathcal{Z}}_h := \Big\{ (\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{N}}_h(\Omega_{\mathrm{C}}) \times (\Theta_h/\mathbb{C}) \times \mathbb{C}^J : \ [\![\widetilde{\Psi}_h]\!]_{\Sigma_j} = c_{hj},\ j = 1, \dots, J, \\ \text{and} \ \int_\ell \mathbf{G}_h \cdot \mathbf{t}_\ell \, d\gamma = \widetilde{\Psi}_h(P_\ell^+) - \widetilde{\Psi}_h(P_\ell^-) \ \ \forall \ell \text{ edge } \mathcal{T}_h : \ \ell \subset \Gamma_{\mathrm{I}} \Big\},$$

$$\boldsymbol{\mathcal{Z}}_h^0 := \Big\{ (\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{Z}}_h : \ c_{hj} = 0,\ j = K+1, \dots, J \Big\}.$$

The new discrete problem, which will be shown to be equivalent to Problem **DHP** in the next theorem, is the following one.

PROBLEM **DLP**. *Find* $(\mathbf{H}_h, \widetilde{\Phi}_h, I_h) \in \boldsymbol{\mathcal{Z}}_h$ *and* $\lambda_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}$ *such that*

$$I_{hj} = \int_{\gamma_j} (\mathbf{n} \times \mathbf{f}_{\mathrm{I}}) \cdot \mathbf{t}_j \, d\gamma, \qquad j = K+1, \dots, J, \tag{6.7}$$

$$\mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu \mathbf{H}_h \cdot \bar{\mathbf{G}}_h + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \, \mathbf{curl}\, \mathbf{H}_h \cdot \, \mathbf{curl}\, \bar{\mathbf{G}}_h + \mathrm{i}\omega \int_{\Omega_{\mathrm{D}}} \mu \, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \cdot \, \widetilde{\mathbf{grad}}\, \bar{\widetilde{\Psi}}_h \\ + \int_{\Gamma_{\mathrm{D}}} \mathbf{grad}_\Gamma \, \lambda_h \cdot \, \widetilde{\mathbf{grad}}_\Gamma \, \bar{\widetilde{\Psi}}_h \, d\Gamma = 0 \quad \forall (\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{Z}}_h^0, \tag{6.8}$$

$$\int_{\Gamma_{\mathrm{D}}} \widetilde{\mathbf{grad}}_\Gamma \, \widetilde{\Phi}_h \cdot \, \mathbf{grad}_\Gamma \, \bar{\nu}_h \, d\Gamma = \int_{\Gamma_{\mathrm{D}}} (\mathbf{n} \times \mathbf{f}_{\mathrm{I}}) \cdot \, \mathbf{grad}_\Gamma \, \bar{\nu}_h \, d\Gamma \quad \forall \nu_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}. \tag{6.9}$$

First we prove that Problem **DLP** is well posed.

THEOREM 6.3. *Let* $\mathbf{f}_{\mathrm{I}}$ *be any vector field defined on* Γ_{D}*, such that the integrals in the right-hand sides of* (6.7) *and* (6.9) *are well defined. Then, Problem* **DLP** *attains a unique solution.*

Proof. Problem **DLP** reduces to a linear system with the same number of equations and unknowns. Then, it is enough to prove that, for $\mathbf{f}_{\mathrm{I}} = \mathbf{0}$, this problem attains only the null solution. So let $(\mathbf{H}_h, \widetilde{\Phi}_h, I_h) \in \boldsymbol{\mathcal{Z}}_h$, and $\lambda_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}$ satisfying (6.7)–(6.9) with $\mathbf{f}_{\mathrm{I}} = \mathbf{0}$.

Equation (6.7) implies that $I_{hj} = 0$ for $j = K+1, \dots, J$; hence, $(\mathbf{H}_h, \widetilde{\Phi}_h, I_h) \in \boldsymbol{\mathcal{Z}}_h^0$, and $[\![\widetilde{\Phi}_h]\!]_{\Sigma_j} = I_{hj} = 0$, too. Thus, if we define $\nu_h := \widetilde{\Phi}_h|_{\widetilde{\Gamma}_{\mathrm{D}}} \in \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{D}})/\mathbb{C}$, we have $[\![\nu_h]\!]_{\overline{\Sigma}_j \cap \Gamma_{\mathrm{D}}} = 0$ for $j = K+1, \dots, J$, and, then, $\nu_h \in \mathcal{L}_h(\Gamma_{\mathrm{D}})/\mathbb{C}$.

Now, by testing (6.9) with this ν_h, we obtain $\widetilde{\mathbf{grad}}_\Gamma \, (\widetilde{\Phi}_h|_{\Gamma_{\mathrm{D}}}) = \mathbf{0}$. Hence, by testing (6.8) with $(\mathbf{H}_h, \widetilde{\Phi}_h, I_h)$ (which was already shown to belong to $\boldsymbol{\mathcal{Z}}_h^0$), we obtain

$$\mathrm{i}\omega \int_{\Omega_{\mathrm{C}}} \mu |\mathbf{H}_h|^2 + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} |\, \mathbf{curl}\, \mathbf{H}_h|^2 + \mathrm{i}\omega \int_{\Omega_{\mathrm{D}}} \mu |\, \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h|^2 = 0.$$

Hence, $\mathbf{H}_h = \mathbf{0}$ in Ω_{C}, and $\widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h = \mathbf{0}$ in Ω_{D}. Consequently, $\widetilde{\Phi}_h = 0$ in $\mathcal{L}_h(\widetilde{\Omega}_{\mathrm{D}})/\mathbb{C}$, and $I_{hj} = [\![\widetilde{\Phi}_h]\!]_{\Sigma_j} = 0$ for $j = 1, \dots, K$, too.

Thus, it remains only to prove that $\lambda_h = 0$. To do this, first let us show that there exists $(\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{Z}}_h^0$ satisfying $\widetilde{\Psi}_h|_{\Gamma_{\mathrm{D}}} = \lambda_h$. Indeed, let $\widetilde{\Psi}_h \in \mathcal{L}_h(\widetilde{\Omega}_{\mathrm{D}})/\mathbb{C}$ be the unique function in this space satisfying for each vertex P of $\mathcal{T}_h^{\Omega_{\mathrm{D}}}$,

$$\begin{aligned} &\widetilde{\Psi}_h(P) = \lambda_h(P) && \text{if } P \in \Gamma_{\mathrm{D}}, \\ &\widetilde{\Psi}_h(P) = 0 && \text{if } P \notin \Gamma_{\mathrm{D}}. \end{aligned}$$

Let $c_{hj} = [\![\widetilde{\Psi}_h]\!]_{\Sigma_j}$, $j = 1, \ldots, J$. Because of the definition of $\widetilde{\Psi}_h$, clearly $c_h = 0$. Then $\widetilde{\Psi}_h \in \Theta_h$. Finally, let $\mathbf{G}_h \in \boldsymbol{\mathcal{N}}_h(\Omega_{\rm C})$ be the unique function in this space satisfying for each edge ℓ of $\mathcal{T}_h$ such that $\ell \subset \overline{\Omega}_{\rm C}$,

$$\int_\ell \mathbf{G}_h \cdot \mathbf{t}_\ell \, d\gamma = \widetilde{\Psi}_h(P_\ell^+) - \widetilde{\Psi}_h(P_\ell^-) \quad \text{if } \ell \subset \Gamma_{\rm I},$$
$$\int_\ell \mathbf{G}_h \cdot \mathbf{t}_\ell \, d\gamma = 0 \quad \text{if } \ell \not\subset \Gamma_{\rm I}.$$

Therefore, $(\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{Z}}_h^0$, and $\widetilde{\Psi}_h|_{\Gamma_{\rm D}} = \lambda_h$. Now, by testing (6.8) with this $(\mathbf{G}_h, \widetilde{\Psi}_h, c_h)$, since we already know that $\mathbf{H}_h = \mathbf{0}$ in $\Omega_{\rm C}$ and that $\mathbf{grad}\, \widetilde{\Phi}_h = 0$ in $\Omega_{\rm D}$, we obtain

$$\int_{\Gamma_{\rm D}} |\, \mathbf{grad}_\Gamma \, \lambda_h|^2 \, d\Gamma = 0.$$

Then, $\lambda_h = 0$ in $\mathcal{L}_h(\Gamma_{\rm D})/\mathbb{C}$, and we conclude the proof. ∎

Now it is very simple to show that Problems **DHP** and **DLP** are equivalent.

THEOREM 6.4. *Let us assume that the solution* $(\mathbf{H}, \widetilde{\Phi})$ *of Problem* **HP** *satisfies* $\mathbf{H} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\rm C})$ *and* $\mathbf{grad}\, \widetilde{\Phi} \in \mathrm{H}^r(\Omega_{\rm D})^3$, *with* $r \in (\frac{1}{2}, 1]$.

If $((\mathbf{H}_h, \widetilde{\Phi}_h, I_h), \lambda_h)$ *is a solution of Problem* **DLP**, *then* $(\mathbf{H}_h, \widetilde{\Phi}_h)$ *is a solution of Problem* **DHP**.

Conversely, if $(\mathbf{H}_h, \widetilde{\Phi}_h)$ *is a solution of Problem* **DHP**, *and* $I_{hj} = [\![\widetilde{\Phi}_h]\!]_{\Sigma_j}$, $j = 1, \ldots, J$, *then there exists* $\lambda_h \in \mathcal{L}_h(\Gamma_{\rm D})/\mathbb{C}$ *such that* $((\mathbf{H}_h, \widetilde{\Phi}_h, I_h), \lambda_h)$ *is a solution of Problem* **DLP**.

Proof. Let $((\mathbf{H}_h, \widetilde{\Phi}_h, I_h), \lambda_h)$ be a solution of Problem **DLP**. Since $(\mathbf{H}_h, \widetilde{\Phi}_h, I_h) \in \boldsymbol{\mathcal{Z}}_h$, $(\mathbf{H}_h, \widetilde{\Phi}_h) \in \boldsymbol{\mathcal{W}}_h$ (because of Lemma 6.1), and $I_{hj} = [\![\widetilde{\Phi}_h]\!]_{\Sigma_j}$, $j = 1, \ldots, J$. Therefore, because of (6.7), $\widetilde{\Phi}_h$ satisfies assumption (6.4) in Lemma 6.2. Then, (6.9) implies (5.8).

On the other hand, let $(\mathbf{G}_h, \widetilde{\Psi}_h) \in \boldsymbol{\mathcal{W}}_h^0$ and $c_{hj} = [\![\widetilde{\Psi}_h]\!]_{\Sigma_j}$, $j = 1, \ldots, J$. Because of Lemma 6.1 and (6.2), $(\mathbf{G}_h, \widetilde{\Psi}_h, c_h) \in \boldsymbol{\mathcal{Z}}_h^0$. Since $\mathbf{grad}\, \widetilde{\Psi}_h \times \mathbf{n} = \mathbf{0}$ on $\Gamma_{\rm D}$, because of (6.3) we have $\mathbf{grad}_\Gamma\, (\widetilde{\Psi}_h|_{\Gamma_{\rm D}}) = \mathbf{0}$. Therefore, by testing (6.8) with such $(\mathbf{G}_h, \widetilde{\Psi}_h, c_h)$, we obtain (5.9). Thus, $(\mathbf{H}_h, \widetilde{\Phi}_h)$ is a solution of Problem **DHP**.

Conversely, let $(\mathbf{H}_h, \widetilde{\Phi}_h)$ be a solution of Problem **DHP**. Since the assumptions of Corollary 5.7 are fulfilled, this solution is unique. On the other hand, Theorem 6.3 shows that there also exists a unique solution $((\mathbf{H}'_h, \widetilde{\Phi}'_h, I_h), \lambda_h)$ of Problem **DLP**. But then we have already proved that $(\mathbf{H}'_h, \widetilde{\Phi}'_h)$ is a solution of Problem **DHP**. Hence, $(\mathbf{H}'_h, \widetilde{\Phi}'_h) = (\mathbf{H}_h, \widetilde{\Phi}_h)$, and we conclude the proof. ∎

Problem **DLP** is the one that we have actually implemented. The degrees of freedom for this problem are the following ones:

- for $\mathbf{H}_h \in \boldsymbol{\mathcal{N}}_h(\Omega_{\rm C})$: $\int_\ell \mathbf{H}_h \cdot \mathbf{t}_\ell \, d\gamma$, for all edge $\ell \subset \overline{\Omega}_{\rm C} \setminus \Gamma_{\rm I}$;
- for $\widetilde{\Phi}_h \in \mathcal{L}_h(\widetilde{\Omega}_{\rm D})/\mathbb{C}$: $\widetilde{\Phi}_h(P)$, for all vertices $P \in \overline{\Omega}_{\rm D}$ (one of them set to zero);
- for $I_h \in \mathbb{C}^J$: I_{hj}, $j = 1, \ldots, J$ ($I_{hK+1}, \ldots, I_{hJ}$ are directly computed from $\mathbf{f}$);
- for $\lambda_h \in \mathcal{L}_h(\Gamma_{\rm D})/\mathbb{C}$: $\lambda_h(P)$, for all vertices $P \in \Gamma_{\rm D}$ (one of them set to zero).

Remark 6.2. We have imposed the boundary condition of Problem **DHP** by means of a Lagrange multiplier. However, this is not the only way of doing it. An

alternative procedure consists of using the fact that, for each edge $\ell \subset \Gamma_{\mathrm{D}}$,

$$\widetilde{\Phi}_h(P_\ell^+) - \widetilde{\Phi}_h(P_\ell^-) = \int_\ell \widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{n} \times \left(\widetilde{\mathbf{grad}}\, \widetilde{\Phi}_h \times \mathbf{n} \right) \cdot \mathbf{t}_\ell \, d\gamma$$
$$= \int_\ell \mathbf{n} \times \mathbf{f}_{\mathrm{I}} \cdot \mathbf{t}_\ell \, d\gamma = \int_\ell \mathbf{n} \times \mathbf{f} \cdot \mathbf{t}_\ell \, d\gamma \,.$$

Therefore, the values of $\widetilde{\Phi}_h(P)$ can be obtained for each vertex $P \in \Gamma_{\mathrm{D}}$ by the following procedure:

1. Fix arbitrarily the value of $\widetilde{\Phi}_h$ at a given vertex $P_0 \in \Gamma_{\mathrm{D}}$:

$$\widetilde{\Phi}_h(P_0) = 0$$

   (this can be done because $\widetilde{\Phi}_h \in \Theta_h/\mathbb{C}$).
2. For each other vertex $P \in \Gamma_{\mathrm{D}}$ (those on $\Gamma_{\mathrm{D}} \cap \Sigma_j$, $j = K+1, \ldots, J$, must be counted twice),
   (a) find a path Γ_P joining P_0 with P, which does not cross any $\Sigma_j \cap \Gamma_{\mathrm{D}}$, $j = K+1, \ldots, J$, and which consists of adequately oriented edges $\ell \subset \Gamma_{\mathrm{D}}$:

$$\Gamma_P := \pm\ell_1 \cup \cdots \cup \pm\ell_{N_P};$$

   (b) evaluate

$$\widetilde{\Phi}_h(P) = \pm \int_{\ell_1} \mathbf{n} \times \mathbf{f} \cdot \mathbf{t}_{\ell_1} \, d\gamma \pm \ldots \pm \int_{\ell_{N_P}} \mathbf{n} \times \mathbf{f} \cdot \mathbf{t}_{\ell_{N_P}}.$$

The main drawback of this procedure is that step 2(a) is rather complicated to implement (see [11]). The strategy we have proposed is more expensive than this one in terms of degrees of freedom (one unknown per vertex on Γ_{D} is added, instead of being eliminated). Nevertheless, one neat advantage is that its implementation is quite straightforward.

7. Numerical experiments. In this section we present some numerical results obtained with a code developed by us, which implements in MATLAB the method described above.

We have solved a particular problem with a known analytical solution to validate the computer code and to test the performance and convergence properties of the method. The geometry of the domain is similar to that of an electric furnace with only one electrode.

More precisely, we consider a domain Ω containing a conductor Ω_{C} and dielectric Ω_{D}, as shown in Figure 7.1.

We assume that $\overline{\Omega}_{\mathrm{C}}$ and $\overline{\Omega} = \overline{\Omega}_{\mathrm{C}} \cup \overline{\Omega}_{\mathrm{D}}$ are coaxial cylinders of radius R_{C} and R_{D}, respectively, with height L. To obtain the data for a test problem in this domain with a known analytical solution, we consider that Ω_{C} and Ω are bounded sections of respective infinite cylinders. The electric conductivity σ is taken as a constant in Ω_{C}, and the magnetic permeability μ is constant in the whole Ω. We consider that an alternating current $\mathbf{J}$ goes through the conductor Ω_{C} in the direction of its axis; this current is assumed to be axially symmetric with an intensity $I(t) = I_0 \cos(\omega t)$.

We analyze this problem using a cylindrical coordinate system (r, θ, z) with the z-axis coinciding with the common axis of both cylinders (see Figure 7.1). We denote $\mathbf{e}_r$, $\mathbf{e}_\theta$, and $\mathbf{e}_z$ the unit vectors in the corresponding coordinate directions.

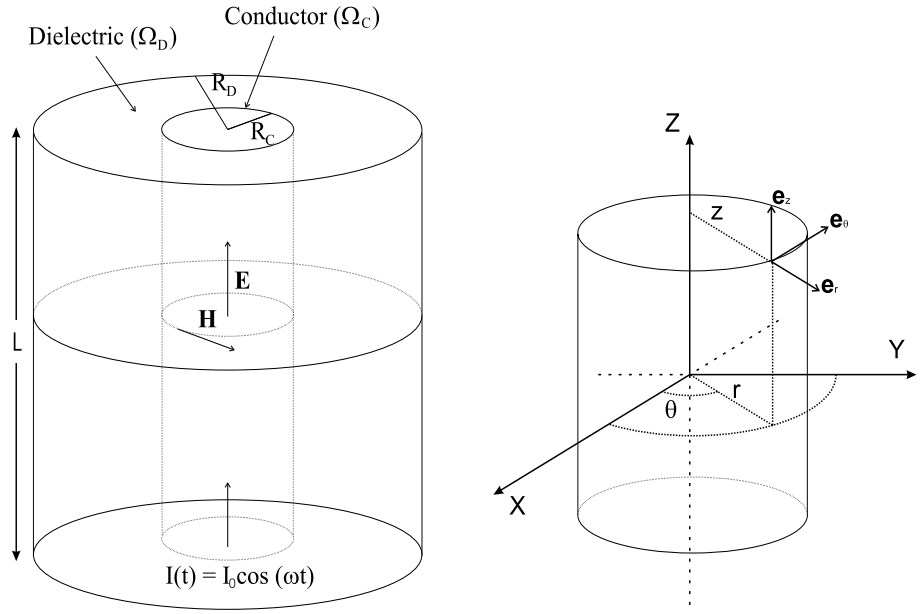


FIG. 7.1. *Sketch of the domain. Coordinate system.*

Because of the assumed conditions on $\mathbf{J}$, only the z-component of the electric field $\mathbf{E} = \frac{1}{\sigma}\mathbf{J}$ does not vanish in the conductor. Moreover, it depends on the radial coordinate r, but it is independent of the other two coordinates z and θ. Consequently, only the θ-component of the magnetic field $\mathbf{H} = \frac{\mathrm{i}}{\omega\mu}\,\mathbf{curl}\,\mathbf{E}$ does not vanish, and it also depends only on the coordinate r. Then, taking into account the expression of the **curl** operator in cylindrical coordinates, we have $\mathbf{H}(r,\theta,z) = H_\theta(r)\mathbf{e}_\theta$, with H_θ satisfying the equation

$$\mathrm{i}\omega\mu H_\theta(r) - \frac{d}{dr}\left\{\frac{1}{\sigma r}\frac{d}{dr}[rH_\theta(r)]\right\} = 0, \qquad r \in (0, R_C),$$

and the boundary conditions

$$|H_\theta(0)| < \infty, \qquad H_\theta(R_C) = \frac{I_0}{2\pi R_C}.$$

To solve this problem, we perform the change of variable $x = \gamma r$, where $\gamma = \sqrt{i\omega\mu\sigma} \in \mathbb{C}$. Then, we obtain the equation

$$x^2\frac{d^2}{dx^2}\widetilde{H}_\theta(x) + x\frac{d}{dx}\widetilde{H}_\theta(x) - (x^2+1)\widetilde{H}_\theta(x) = 0, \qquad x \in (0, \gamma R_C),$$

where $\widetilde{H}_\theta(x) = H_\theta(x/\gamma)$.

This is a Bessel equation, the solution of which is given by $\widetilde{H}_\theta(x) = \alpha \mathrm{I}_1(x)$, with I_1 being the modified Bessel function of the first kind, and α is a constant to be obtained from the boundary condition at $x = \gamma R_C$. Thus, the magnetic field in the conductor is given by

$$\mathbf{H}(r,\theta,z) = \frac{I_0}{2\pi R_C}\frac{\mathrm{I}_1(\gamma r)}{\mathrm{I}_1(\gamma R_C)}\mathbf{e}_\theta, \qquad r \in (0, R_C), \quad \theta \in [0, 2\pi], \quad z \in \mathbb{R}.$$

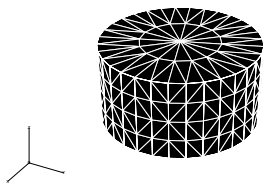


FIG. 7.2. *Coarsest mesh on the conductor domain.*

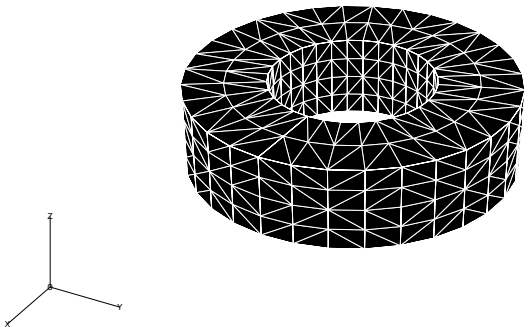


FIG. 7.3. *Coarsest mesh on the dielectric domain.*

On the other hand, the magnetic field created by an infinite circular cylindrical conductor of radius R_{C} carrying an axially aligned and symmetric current of intensity I_0 is computed using Ampère's circuital law (see, for instance, [27]). In cylindrical coordinates it is given by $\mathbf{H}(r,\theta,z) = H_\theta(r)\mathbf{e}_\theta$, with

$$H_\theta(r) = \frac{I_0}{2\pi r}, \qquad r \geq R_{\mathrm{C}}.$$

Once more, the magnitude of H_θ depends only on the radial coordinate r.

Moreover, from this expression, it is also possible to know the multivalued magnetic potential $\widetilde{\Phi}$, which corresponds to the magnetic field in the dielectric domain. Indeed, taking into account the expression of the gradient operator in cylindrical coordinates, we obtain

$$\widetilde{\Phi}(r,\theta,z) = \frac{I_0}{2\pi}\theta, \qquad r > R_{\mathrm{C}}, \quad \theta \in [0,2\pi], \quad z \in \mathbb{R}.$$

Notice that the scalar potential depends only on the variable θ and experiences a jump of magnitude I_0 across the cut surface Σ placed at $\theta = 0$.

Now, we again consider the bounded cylinder of Figure 7.1. The boundary conditions added to define properly this problem are the following:

- On the exterior boundary of the dielectric domain (i.e., the lateral surface of the cylinder Ω and the outer part of its top and bottom surfaces), we consider

TABLE 7.1
$\mathrm{H}(\mathbf{curl},\Omega)$-*norm of errors and exact solution.*

Mesh-size	Number d.o.f.	Computed solution	Error
h	1699	62257.66	36504.31
$h/2$	11285	64063.20	23591.63
$h/3$	35671	64296.90	16646.00
$h/4$	81769	64363.90	12709.90
$h/5$	156491	64391.83	10228.25

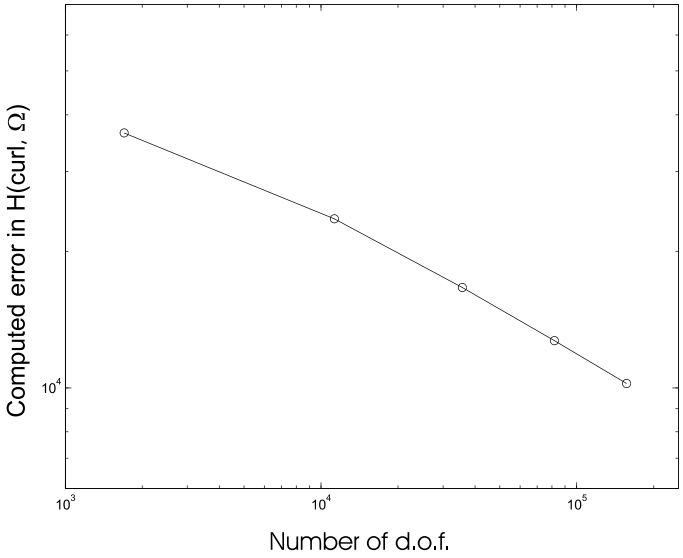


FIG. 7.4. *Error versus number of d.o.f. (log-log scale).*

the condition $\mathbf{H}\times\mathbf{n}=\mathbf{f}$, with $\mathbf{f}$ being obtained from the analytical solution:

$$\mathbf{f}(R_{\mathrm{D}},\theta,z)=-\frac{I_0}{2\pi R_{\mathrm{D}}}\mathbf{e}_z,\qquad \theta\in[0,2\pi],\quad z\in\left[-\tfrac{L}{2},\tfrac{L}{2}\right],$$

$$\mathbf{f}\left(r,\theta,\pm\tfrac{L}{2}\right)=\pm\frac{I_0}{2\pi R_{\mathrm{C}}}\frac{\mathrm{I}_1(\gamma r)}{\mathrm{I}_1(\gamma R_{\mathrm{C}})}\mathbf{e}_r,\qquad r\in[R_{\mathrm{C}},R_{\mathrm{D}}],\quad \theta\in[0,2\pi].$$

- On the top and bottom surfaces of the conducting cylinder Ω_{C}, we impose

$$\mathbf{E}\left(r,\theta,\pm\tfrac{L}{2}\right)\times\mathbf{n}=\mathbf{0},\qquad r\in(0,R_{\mathrm{D}}),\quad \theta\in[0,2\pi],$$

  which is true in this case, because the electric field has vanishing r- and θ-components, and, thus, it aligns with the normal vector $\mathbf{n}$ on these surfaces.

Finally, we have used the following geometrical and physical data:

- $R_{\mathrm{C}}=1\,\mathrm{m}$;
- $R_{\mathrm{D}}=2\,\mathrm{m}$;
- $L=1\,\mathrm{m}$;
- $\sigma=151565.8\,(\Omega\mathrm{m})^{-1}$;
- $\mu=\mu_0=4\pi\,10^{-7}\,\mathrm{Hm}^{-1}$ (magnetic permeability of free space);
- $I_0=62000\,\mathrm{A}$;
- $\omega=50\,\mathrm{H}z$.

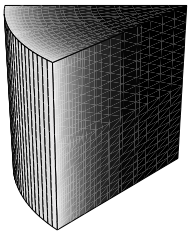


FIG. 7.5. *Intensity of the magnetic field* $|\mathbf{H}_h|$ *in the conductor.*

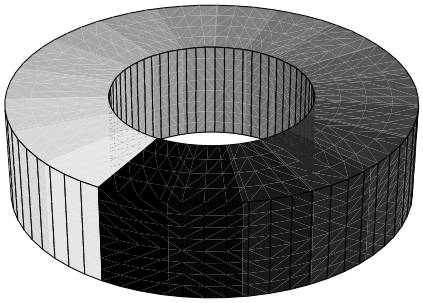


FIG. 7.6. *Magnetic potential* $\widetilde{\Phi}_h$ *in the dielectric.*

To determine the order of convergence, the numerical method has been used on several successively refined meshes, and we have compared the obtained numerical solutions with the analytical one. Figures 7.2 and 7.3 show the coarsest meshes used for conductor and dielectric domains, respectively.

Table 7.1 shows the $\mathrm{H}(\mathbf{curl}, \Omega)$ norms of the approximate solutions $\mathbf{H}_h$ computed on several meshes and their corresponding errors. The total number of degrees of freedom for each mesh are also included.

Figure 7.4 shows a log-log plot of the errors measured in $\mathrm{H}(\mathbf{curl}, \Omega)$-norm versus the number of degrees of freedom for the same meshes. A linear dependence on the mesh-size is obtained by calculating the slope of the line. These $\mathcal{O}(h)$ errors agree with the theoretical results, since the solution is smooth, and, hence, the hypotheses of Theorem 5.4 are fulfilled for $r = 1$.

Finally, Figures 7.5 and 7.6 show the intensity of the computed magnetic field $|\mathbf{H}_h|$ in the conductor domain Ω_{C} and the computed magnetic potential $\widetilde{\Phi}_h$ in the dielectric domain Ω_{D}. The former is presented in a section of Ω_{C} to show its behavior in the interior of this domain.

Acknowledgment. The authors thank Professor Alain Bossavit for many valuable discussions.

REFERENCES

[1] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of* $\mathrm{H}(\mathbf{curl},\Omega)$ *and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.

[2] A. Alonso and A. Valli, *A domain decomposition approach for heterogeneous time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 97–112.

[3] A. Alonso and A. Valli, *An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations*, Math. Comp., 68 (1999), pp. 607–631.

[4] H. Ammari, A. Buffa, and J.-C. Nédélec, *A justification of eddy currents model for the Maxwell equations*, SIAM J. Appl. Math., 60 (2000), pp. 1805–1823.

[5] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[6] A. Bermúdez, J. Bullón, and F. Pena, *A finite element method for the thermoelectrical modelling of electrodes*, Comm. Numer. Methods Engrg., 14 (1998), pp. 581–593.

[7] A. Bermúdez, M. C. Muñiz, F. Pena, and J. Bullón, *Numerical computation of the electromagnetic field in the electrodes of a three-phase arc furnace*, Internat. J. Numer. Methods Engrg., 46 (1999), pp. 649–658.

[8] A. Bossavit, *Magnetostatic problems in multiply connected regions: Some properties of the curl operator*, Proc. IEE-A, 135 (1988), pp. 179–187.

[9] A. Bossavit, *Whitney forms: A class of finite elements for three-dimensional computations in electromagnetism*, Proc. IEE-A, 135 (1988), pp. 493–500.

[10] A. Bossavit, *The computation of eddy-currents in dimension* 3 *by using mixed finite elements and boundary elements in association*, Math. Comput. Modelling, 15 (1991), pp. 33-42.

[11] A. Bossavit, *Computational Electromagnetism. Variational Formulations, Complementarity, Edge Elements*, Academic Press, San Diego, CA, 1998.

[12] A. Bossavit, *"Hybrid" electric-magnetic methods in eddy-current problems*, Comput. Methods Appl. Mech. Engrg., 178 (1999), pp. 383–391.

[13] A. Bossavit and J. C. Vérité, *A mixed FEM-BIEM method to solve* 3*-D eddy current problems*, IEEE Trans. Mag., 18 (1982), pp. 431–435.

[14] A. Buffa and P. Ciarlet, Jr., *On traces for functional spaces related to Maxwell's equations.* I. *An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[15] A. Buffa and P. Ciarlet, Jr., *On traces for functional spaces related to Maxwell equations.* II. *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.

[16] M. Costabel, V. J. Ervin, and E. P. Stephan, *Symmetric coupling of finite elements and boundary elements for a parabolic-elliptic interface problem*, Quart. Appl. Math., 48 (1990), pp. 265–279.

[17] P. Fernandes and G. Gilardi, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.

[18] S. I. Hariharan and R. C. MacCamy, *An integral equation procedure for eddy current problems*, J. Comput. Phys., 45 (1982), pp. 80–99.

[19] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer–Verlag, Berlin, 1986.

[20] L. Kettunen, K. Forsman, and A. Bossavit, *Formulation of the eddy current problem in multiple connected regions in terms of* h, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 935–954.

[21] R. C. MacCamy and E. Stephan, *A skin effect approximation for eddy current problems*, Arch. Ration. Mech. Anal., 90 (1985), pp. 87–98.

[22] R. C. MacCamy and M. Suri, *A time-dependent interface problem for two-dimensional eddy currents*, Quart. Appl. Math., 44 (1987), pp. 675–690.

[23] P. B. Monk, *A mixed method for approximating Maxwell's equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1610–1634.

[24] P. Monk, *A finite element method for approximating the time-harmonic Maxwell equations*, Numer. Math., 63 (1992), pp. 243–261.

[25] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal.,

29 (1992), pp. 714–729.
[26] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[27] B. D. Popović, *Introductory Engineering Electromagnetics*, Addison–Wesley, Reading, MA, 1971.
[28] P. P. Silvester and R. L. Ferrari, *Finite Elements for Electrical Engineers*, Cambridge University Press, Cambridge, UK, 1996.
[29] I. A. Tsukerman, *Error estimation for finite element solutions of the eddy currents problem*, COMPEL, 9 (1990), pp. 83–98.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1850–1869

A POSTERIORI ERROR ESTIMATES FOR CONTROL PROBLEMS GOVERNED BY STOKES EQUATIONS*

WENBIN LIU† AND NINGNING YAN‡

Abstract. In this paper, we derive a posteriori error estimates for the finite element approximation of distributed optimal control problems governed by the Stokes equations. We obtain a posteriori error estimators for both the state and the control approximation in the L^2 norm and the H^1 norm. These estimates can be used to construct reliable adaptive finite element approximation for the control problems.

Key words. distributed optimal control, finite element approximation, adaptive finite element methods, a posteriori error estimates, Stokes equations

AMS subject classifications. 49J20, 65N30

PII. S0036142901384009

1. Introduction.

Flow control problems are crucial to many engineering applications. For the last decade, it has become a very active and successful research area. Extensive research has been carried out on various theoretical aspects of flow control problems; see, for example, [1], [14], [23], [20], [31], [32], [42], [45], and the references cited therein, for existence of optimal control, optimality conditions, regularity of the optimal solutions, and existence of Lagrange multipliers. The literature for engineering applications is too huge to give even a very brief review here. One can find some useful model optimal control problems of flow motion with the purpose of achieving some desired objective in real-life applications in, for example, [19], [20], [25], [42], [45], and some flavors of aeronautical and chemical engineering problems can be found in, e.g., [26], [27], and [28]. In flow control problems, boundary and shape control are widely used, though body (i.e., distributed) control is also available through a magnetic field, a heat source using radiation, or laser technology; see [20], [25], and [40].

It is obvious that efficient numerical methods are essential to successful applications of flow control (indeed of any other control). Nowadays, the finite element method is undoubtedly the most widely used numerical method in computing optimal control problems, including flow control problems. The literature on finite element approximation of optimal control is huge. Systematic introduction of the finite element method for PDEs and optimal control problems can be found in, for example, [11], [24], [43], and [45]. There have been extensive theoretical studies for finite element approximation of various optimal control problems. For instance, a priori error estimates of finite element approximation were established long ago for the optimal control problems governed by linear elliptic or parabolic state equations; see, for example, [13], [30], and [38]. Furthermore, finite element approximation of some

*Received by the editors January 23, 2001; accepted for publication (in revised form) June 3, 2002; published electronically November 22, 2002. This research was supported by EPSRC research grant GR/R31980, the Special Funds for Major State Basic Research Projects (G2000067102), National Natural Science Foundation of China (19931030), and Innovation Funds of Academy of Mathematics and System Sciences, CAS.

http://www.siam.org/journals/sinum/40-5/38400.html

†CBS and Institute of Mathematics and Statistics, University of Kent, Canterbury, CT2 7NF, United Kingdom (w.b.liu@ukc.ac.uk).

‡Institute of System Sciences, Chinese Academy of Sciences, Beijing, China (yan@staff.iss.ac.cn).

flow control has been studied, and a priori error estimates have been established; see [22], [23], [20], and [25]. A priori error estimates have also been obtained for a class of state constrained control problems in [48], though the state equation is assumed to be linear. In [33] this assumption has been removed by reformulating the control problem as an abstract optimization problem in some Banach spaces and then applying nonsmooth analysis. In fact, the state equation there can be a variational inequality.

In recent years, the adaptive finite element method has been extensively investigated, beginning with the pioneering work in [7]. Adaptive finite element approximation is among the most important means to boost the accuracy and efficiency of finite element discretizations. It ensures a higher density of nodes in a certain area of the given domain, where the solution is more difficult to approximate, using an a posteriori error indicator. The decision of whether further refinement of the meshes is necessary is based on the estimate of the discretization error. If further refinement is to be performed, then the error indicator is used as a guide to show how the refinement might be accomplished most efficiently. The literature in this area is huge. Some of techniques directly relevant to our work can be found in [6], [37], [46], [49], [50], and [51]. It is our belief that adaptive finite element enhancement is one of the future directions to pursue in developing sophisticated numerical methods for optimal design problems.

Although adaptive finite element approximation is widely used in numerical simulations, it has not yet been *fully* utilized in optimal design. Initial attempts in this aspect have only been reported recently for some design problems; see, e.g., [2], [4], [39], and [47]. However, a posteriori error indicators of a heuristic nature are widely used in most applications. For instance, in some existing work on adaptive finite element approximation of optimal design, the mesh refinement is guided by a posteriori error estimators based on a posteriori error estimates *solely* from the state equation for a fixed control. Thus error information from approximation of the control (design) is not utilized. This strategy was found to be inefficient in recent numerical experiments (see [8]). In other work (see [12]), a preassigned mesh refinement scheme is applied around the possible singularity points of the state equation. Although these methods may work well in some particular applications, they cannot be applied confidently in general. It is unlikely that the potential power of adaptive finite element approximation has been fully utilized due to the lack of more sophisticated a posteriori error indicators.

It is not straightforward to rigorously derive suitable a posteriori error estimators or monitors for general optimal control problems. For instance, it seems difficult to apply the gradient recovery techniques as the control is normally not differentiable. Recovering approximation in function value is in general difficult. For a similar reason, it also seems difficult to apply the local solution strategy.

Very recently, some error indicators of residual type were developed in [8], [9], [34], and [36]. These error estimators are based on a posteriori estimation of the discretization error for the state *and* the control (design). When there exists no inequality control constraint in a control problem, normally the optimality conditions consist of coupled partial differential equations only. Consequently, one may be able to write down the dual system of the *whole* optimality conditions and then apply the adjoint approach (see [16], [17], [44]) to obtain some error estimators. For example, one can then apply the weighted a posteriori error estimation technique to obtain a posteriori estimators for *objective functional* approximation error; see [8] and [9]. Such estimators have indeed been derived for some unconstrained elliptic control problems, and

have proved quite efficient in the numerical tests carried out in [8].

However, there frequently exist some inequality constraints for the control in applications. In such cases, the optimality conditions often contain a variational inequality and then have some very different properties. For example, the optimal control and the state may have very different regularity. Thus, it is not clear how to apply the adjoint techniques used in [8] and [9] to general constrained control problems. In our work, constrained cases are studied via residual estimation using the norms of energy type. A posteriori error estimators are derived for some constrained control problems governed by elliptic equations; see [34] and [36]. However, to our best knowledge, there has been a lack of a posteriori error indicators for finite element approximation of any constrained flow control problem, which is immensely important and yet far more complicated to analyze than an elliptic control problem.

In this work we investigate optimal control governed by stationary Stokes flows with general control constraints. The purpose of studying control for Stokes flows is twofold: On the one hand, the Stokes equations model the flows with low velocity, or very viscous fluids, e.g., many biological flows and non-Newtonian flows. Thus, such a control model can be used as the first approximation of more complex control problems; see [41], [42], and [45] for some examples. On the other hand, some of the problems encountered when studying control for the full Navier–Stokes equations are already present in this simpler model. Thus this investigation will pave the way for further research on adaptive finite element approximation of optimal control for the Navier–Stokes equations; see Remark 3.2 for some details.

In this paper we derive a posteriori error estimates for the conforming finite element approximation of distributed optimal control governed by the two-dimensional Stokes equations. The case of boundary control can be dealt with similarly by combining the ideas and techniques used in [36]. The obtained error estimates can then be used as a posteriori error indicators to construct reliable adaptive finite element methods.

The plan of the paper is as follows: In section 2 we shall give a weak formula for the control problem and then discuss the finite element approximation of the control problem. In section 3, a posteriori error bounds are derived for the control problem in L^2 and H^1 norms. Some applications are presented in section 4.

Let Ω and $\Omega_{\mathbf{U}}$ be two bounded open sets in R^2 with Lipschitz boundaries $\partial\Omega$ and $\partial\Omega_U$, respectively. In this paper we adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with norm $\|\cdot\|_{W^{m,q}(\Omega)}$ and seminorm $|\cdot|_{W^{m,q}(\Omega)}$ (or $\|\cdot\|_{m,q,\Omega}$, $|\cdot|_{m,q,\Omega}$ for simplification). We shall extend these (semi)norms to vector functions whose components belong to $W^{m,p}(\Omega)$. We set $W_0^{m,q}(\Omega) \equiv \{w \in W^{m,q}(\Omega) : w|_{\partial\Omega} = 0\}$. We denote $W^{m,2}(\Omega)$ ($W_0^{m,2}(\Omega)$) by $H^m(\Omega)$ ($H_0^m(\Omega)$). In addition, c or C denotes a general positive constant independent of h.

2. Finite element approximation of optimal control problems. In this section, we discuss the finite element approximation of distributed convex optimal control problems governed by the Stokes equations. Let $\mathbf{Y} = (H_0^1(\Omega))^2$, $\mathbf{U} = (L^2(\Omega_U))^2$, $\mathbf{H} = (L^2(\Omega))^2$, and $Q = L_0^2(\Omega) = \{q \in L^2(\Omega), \int_\Omega q = 0\}$. In this paper, the state space and the control space will be $\mathbf{Y} \times Q$ and $\mathbf{U}$, respectively. Let B be a linear continuous operator from $\mathbf{U}$ to $\mathbf{H}$, and let $\mathbf{K}$ be a closed convex subset of $\mathbf{U}$. Assume that g and h are strictly convex functionals which are differentiable on $\mathbf{H}$ and that $h(\mathbf{u}) \to +\infty$ as $\|\mathbf{u}\|_{\mathbf{U}} \to \infty$.

We are interested in the following optimal control problem: find $(\mathbf{y}, r, \mathbf{u}) \in$

$\mathbf{Y} \times Q \times \mathbf{U}$ such that

$$\begin{aligned} &\min_{\mathbf{u}\in\mathbf{K}\subset\mathbf{U}}\{g(\mathbf{y})+h(\mathbf{u})\},\\ -\Delta\mathbf{y}+\nabla r &= \mathbf{f}+B\mathbf{u} \text{ in } \Omega,\\ \operatorname{div}\mathbf{y} &= 0 \text{ in } \Omega,\\ \mathbf{y} &= 0 \text{ on } \partial\Omega, \end{aligned}$$

where $\mathbf{f} \in \mathbf{L} = (L^2(\Omega))^2$, $\mathbf{K}$ is a closed convex set in $\mathbf{U}$, and B is a continuous linear operator from $\mathbf{U}$ to $\mathbf{H}$. To consider the finite element approximation of the above optimal control problem, we have to give a weak formula for the state equations. Let

$$a(\mathbf{y},\mathbf{w}) = \int_\Omega \nabla\mathbf{y}\cdot\nabla\mathbf{w} \qquad \forall \mathbf{y},\mathbf{w}\in\mathbf{Y},$$

$$b(\mathbf{v},r) = \int_\Omega r\,\operatorname{div}\mathbf{v} \qquad \forall(\mathbf{v},r)\in\mathbf{Y}\times Q,$$

$$(\mathbf{f}+B\mathbf{u},\mathbf{w}) = \int_\Omega (\mathbf{f}+B\mathbf{u})\cdot\mathbf{w} \qquad \forall \mathbf{u},\mathbf{w}\in\mathbf{H}\times\mathbf{Y}.$$

Then the standard weak formula for the state equations reads as follows: Given $\mathbf{f}\in\mathbf{L}$, find $(\mathbf{y}(\mathbf{u}), r(\mathbf{u}))\in\mathbf{Y}\times Q$ such that

$$a(\mathbf{y}(\mathbf{u}),\mathbf{w}) - b(\mathbf{w},r(\mathbf{u})) = (\mathbf{f}+B\mathbf{u},\mathbf{w}) \qquad \forall\mathbf{w}\in\mathbf{Y}, \tag{2.1}$$

$$b(\mathbf{y}(\mathbf{u}),\phi) = 0 \qquad \forall\phi\in Q. \tag{2.2}$$

For the above problem, it is well known that the following Babuska–Brezzi condition holds (see [18], for example).

LEMMA 2.1. *Let* $ß = (H_0^1(\Omega))^2 \times L_0^2(\Omega)$, *and define a bilinear form* $Ł$ *on* $ß\times ß$ *by* $Ł([\mathbf{u},p];[\mathbf{v},q]) := a(\mathbf{u},\mathbf{v}) - b(\mathbf{v},p) - b(\mathbf{u},q)$; *then*

$$\inf_{[\mathbf{u},p]\in ß}\sup_{[\mathbf{v},q]\in ß}\frac{Ł([\mathbf{u},p];[\mathbf{v},q])}{(|\mathbf{u}|_1+\|p\|_0)(|\mathbf{v}|_1+\|q\|_0)} \ge c > 0,$$

where c *is a constant independent of* $\mathbf{u}$, $\mathbf{v}$, p, *and* q.

Furthermore, the following a priori estimates are well known (see [18], for example).

LEMMA 2.2. *Assume that* Ω *is convex. Let* (Ψ,ρ) *be the solution of the following equation:*

$$\begin{aligned} a(\Psi,\mathbf{w}) \pm b(\mathbf{w},\rho) &= (\Phi,w) \qquad \forall\mathbf{w}\in(H_0^1(\Omega))^2,\\ b(\Psi,q) &= 0 \qquad \forall q\in L_0^2(\Omega), \end{aligned} \tag{2.3}$$

where $a(\cdot,\cdot)$ *and* $b(\cdot,\cdot)$ *are defined as above. Then*

$$\|\Psi\|_{2,\Omega} + \|\rho\|_{1,\Omega} \le C\|\Phi\|_{0,\Omega}.$$

Using the weak formula, our control problem can be restated as the following (SCP):

$$\begin{aligned} &\min_{\mathbf{u}\in\mathbf{K}\subset\mathbf{U}}\{g(\mathbf{y})+h(\mathbf{u})\},\\ a(\mathbf{y}(\mathbf{u}),\mathbf{w}) - b(\mathbf{w},r(\mathbf{u})) &= (\mathbf{f}+B\mathbf{u},\mathbf{w}) \qquad \forall\mathbf{w}\in\mathbf{Y},\\ b(\mathbf{y}(\mathbf{u}),\phi) &= 0 \qquad \forall\phi\in Q. \end{aligned} \tag{2.4}$$

It is well known (see, e.g., [31]) that the control problem (SCP) has a unique solution $(\mathbf{y}^*, r^*, \mathbf{u}^*)$ and that $(\mathbf{y}^*, r^*, \mathbf{u}^*)$ is the solution of (SCP) if and only if there is a co-state $(\mathbf{p}^*, s^*) \in \mathbf{Y} \times Q$ such that $(\mathbf{y}^*, r^*, \mathbf{p}^*, s^*, \mathbf{u}^*)$ satisfies the following optimality conditions (SCP–OPT):

$$\begin{aligned} a(\mathbf{y}^*, \mathbf{w}) - b(\mathbf{w}, r^*) &= (\mathbf{f} + B\mathbf{u}^*, \mathbf{w}) && \forall \mathbf{w} \in \mathbf{Y}, \\ b(\mathbf{y}^*, \phi) &= 0 && \forall \phi \in Q, \\ a(\mathbf{q}, \mathbf{p}^*) + b(\mathbf{q}, s^*) &= (g'(\mathbf{y}^*), \mathbf{q}) && \forall \mathbf{q} \in \mathbf{Y}, \\ b(\mathbf{p}^*, \psi) &= 0 && \forall \psi \in Q, \\ (h'(\mathbf{u}^*) + B^*\mathbf{p}^*, \mathbf{v} - \mathbf{u}^*)_{\mathbf{U}} &\geq 0 && \forall \mathbf{v} \in \mathbf{K} \subset \mathbf{U}, \end{aligned} \tag{2.5}$$

where B^* is the adjoint operator of B, and $(\cdot, \cdot)_{\mathbf{U}}$ is the inner product of $\mathbf{U}$.

We note that for any $(\mathbf{y}, \mathbf{u}) \in \mathbf{V} \times \mathbf{U}$, $g'(\mathbf{y})$ and $h'(\mathbf{u})$ are in $\mathbf{H} = \mathbf{H}' = (L^2(\Omega))^2$ and $\mathbf{U}' = \mathbf{U} = (L^2(\Omega_{\mathbf{U}}))^2$, respectively. Therefore they can be viewed as functions in $(L^2(\Omega))^2$ and $(L^2(\Omega_{\mathbf{U}}))^2$, respectively, from the well-known representation theorem in a Hilbert space.

Let us consider finite element approximation of the control problem (SCP). Here we consider only triangular elements as they are among the most widely used ones. Also, we consider only conforming elements for the state and co-state equations.

Let Ω^h be a polygonal approximation of Ω with boundary $\partial\Omega^h$. Let T^h be a partitioning of Ω^h into disjoint regular triangular τ, so that $\bar{\Omega}^h = \bigcup_{\tau \in T^h} \bar{\tau}$. Each element has at most one edge on $\partial\Omega^h$, and $\bar{\tau}$ and $\bar{\tau}'$ have either only one common vertex or a whole edge if τ and $\tau' \in T^h$. We further require that $P_i \in \partial\Omega^h \Rightarrow P_i \in \partial\Omega$, where $\{P_i\}$ $(i = 1 \dots J)$ is the vertex set associated with the triangulation T^h. For ease of exposition, we will assume that $\Omega^h = \Omega$, though all the results can be extended to the more general case where $\Omega^h \subset \Omega$.

Associated with T^h is a finite dimensional subspace $\mathbf{Y}^h \times Q^h$ of $(H_0^1(\Omega^h))^2 \times L_0^2(\Omega^h)$. In this paper, we assume that $\mathbf{Y}^h$ and Q^h contain the functions which are piecewise polynomials of degree at least 1 and 0, respectively, and of degree at most l and m, respectively.

Then the discretized weak formula of the state equations reads as

$$a(\mathbf{y}_h, \mathbf{w}_h) - b(\mathbf{w}_h, r_h) = (\mathbf{f} + B\mathbf{u}_h, \mathbf{w}_h) \qquad \forall \mathbf{w}_h \in \mathbf{Y}^h \subset \mathbf{Y}, \tag{2.6}$$

$$b(\mathbf{y}_h, \phi_h) = 0 \qquad \forall \phi_h \in Q^h \subset Q. \tag{2.7}$$

In order to guarantee that the above problem is well-posed, we assume that the spaces $\mathbf{Y}^h$ and Q^h satisfy the well-known Babuska–Brezzi conditions: There is constant $c > 0$, independent of h, such that

$$\inf_{r_h \in Q^h} \sup_{\mathbf{y}_h \in \mathbf{Y}^h} (r_h, \nabla \cdot \mathbf{y}_h) / \|r_h\|_{0,\Omega} \|\mathbf{y}_h\|_{1,\Omega} \geq c.$$

These above assumptions are satisfied by many finite elements, e.g., the Taylor–Hood elements and the mini elements; see, [3], [15], [18], and [50] for the details.

Let $\Omega_{\mathbf{U}}^h$ be a polygonal approximation to Ω_U with boundary $\partial\Omega_{\mathbf{U}}^h$. Let $T_{\mathbf{U}}^h$ be a partitioning of $\Omega_{\mathbf{U}}^h$ into disjoint regular triangular $\tau_{\mathbf{U}}$ so that $\bar{\Omega}_{\mathbf{U}}^h = \bigcup_{\tau_{\mathbf{U}} \in T_{\mathbf{U}}^h} \bar{\tau}_{\mathbf{U}}$. Each element has at most one edge on $\partial\Omega_{\mathbf{U}}^h$, and $\bar{\tau}_{\mathbf{U}}$ and $\bar{\tau}'_{\mathbf{U}}$ have either only one common vertex or a whole edge if $\tau_{\mathbf{U}}$ and $\tau'_{\mathbf{U}} \in T_{\mathbf{U}}^h$. We further require that $P_i \in \partial\Omega_{\mathbf{U}}^h \Rightarrow P_i \in \partial\Omega_{\mathbf{U}}$, where $\{P_i\}$ $(i = 1 \dots J)$ is the vertex set associated with the triangulation $T_{\mathbf{U}}^h$.

Associated with $T_{\mathbf{U}}^h$ is another finite dimensional subspace $W_{\mathbf{U}}^h$ of $L^2(\Omega_{\mathbf{U}}^h)$, such that $\chi_{|\tau_{\mathbf{U}}}$ is a polynomial of k-order ($k \geq 0$) for all $\chi \in W_{\mathbf{U}}^h$ and $\tau \in T_{\mathbf{U}}^h$. Here there is no requirement for the continuity or boundary conditions. Due to the limited regularity of the optimal control $\mathbf{u}$ (which is at most in $(H^1(\Omega_{\mathbf{U}}))^2$ in general), we consider only the piecewise constant elements, that is, $k = 0$. For ease of exposition, we again assume that $\Omega_{\mathbf{U}}^h = \Omega_{\mathbf{U}}$.

Let $\mathbf{U}^h = (W_{\mathbf{U}}^h)^2$. It is easy to see that $\mathbf{U}^h \subset \mathbf{U}$. Let h_τ ($h_{\tau_{\mathbf{U}}}$) denote the maximum diameter of the element τ ($\tau_{\mathbf{U}}$) in T^h ($T_{\mathbf{U}}^h$); let ρ_τ ($\rho_{\tau_{\mathbf{U}}}$) denote the diameter of the largest ball contained in τ ($\tau_{\mathbf{U}}$). Assume that there is a regularity constant R such that $1 \leq \max_{\tau \in T^h}(h_\tau/\rho_\tau) \leq R$ ($1 \leq \max_{\tau_{\mathbf{U}} \in T_{\mathbf{U}}^h}(h_{\tau_{\mathbf{U}}}/\rho_{\tau_{\mathbf{U}}}) \leq R$). Let $h = \max_{\tau \in T^h} h_\tau$ ($h_{\mathbf{U}} = \max_{\tau_{\mathbf{U}} \in T_{\mathbf{U}}^h} h_{\tau_{\mathbf{U}}}$).

Then a possible finite element approximation of (SCP) is the control problem $(\mathrm{SCP})^h$:

$$\min_{\mathbf{u_h} \in \mathbf{K^h} \subset \mathbf{U_h}} \{g(\mathbf{y}_h) + h(\mathbf{u}_h)\},$$

$$\begin{aligned} a(\mathbf{y}_h, \mathbf{w}_h) - b(\mathbf{w}_h, r_h) &= (\mathbf{f} + B\mathbf{u}_h, \mathbf{w}_h) \qquad \forall \mathbf{w}_h \in \mathbf{Y}^h \subset \mathbf{Y}, \\ b(\mathbf{y}_h, \phi_h) &= 0 \qquad \forall \phi_h \in Q^h \subset Q, \end{aligned} \tag{2.8}$$

where $\mathbf{K}^h$ is a closed convex set in $\mathbf{U}^h$, an approximation of $\mathbf{K}$.

It follows that the control problem $(\mathrm{SCP})^h$ has a unique solution $(\mathbf{y}_h^*, r_h^*, \mathbf{u}_h^*)$ and that $(\mathbf{y}_h^*, r_h^*, \mathbf{u}_h^*) \in \mathbf{Y}^h \times Q^h \times \mathbf{U}^h$ is the solution of $(\mathrm{SCP})^h$ if and only if there is a co-state $(\mathbf{p}_h^*, s_h^*) \in \mathbf{Y}^h \times Q^h$ such that $(\mathbf{y}_h^*, r_h^*, \mathbf{p}_h^*, s_h^*, \mathbf{u}_h^*)$ satisfies the following optimality conditions $(\mathrm{SCP\text{–}OPT})^h$:

$$\begin{aligned} a(\mathbf{y}_h^*, \mathbf{w}_h) - b(\mathbf{w}_h, r_h^*) &= (\mathbf{f} + B\mathbf{u}_h^*, \mathbf{w}_h) \qquad \forall \mathbf{w}_h \in \mathbf{Y}^h \subset \mathbf{Y}, \\ b(\mathbf{y}_h^*, \phi_h) &= 0 \qquad \forall \phi_h \in Q^h \subset Q, \\ a(\mathbf{q}_h, \mathbf{p}_h^*) + b(\mathbf{q}_h, s_h^*) &= (g'(\mathbf{y}_h^*), \mathbf{q}_h) \qquad \forall \mathbf{q}_h \in \mathbf{Y}^h \subset \mathbf{Y}, \\ b(\mathbf{p}_h^*, \psi_h) &= 0 \qquad \forall \psi_h \in Q^h \subset Q, \\ (h'(\mathbf{u}_h^*) + B^*\mathbf{p}_h^*, \mathbf{v}_h - \mathbf{u}_h^*)_{\mathbf{U}} &\geq 0 \qquad \forall \mathbf{v}_h \in \mathbf{K}^h \subset \mathbf{U}^h \subset \mathbf{U}. \end{aligned} \tag{2.9}$$

To derive a posteriori error estimates, we need some stable interpolators from $\mathbf{Y}$ to $\mathbf{Y}^h$, and from Q to Q^h, respectively. These are given in the following lemmas, which are important in deriving residual-type a posteriori error estimates. These lemmas are well known, and can be found in, e.g., [11], [29], and [46].

LEMMA 2.3 (see [11]). *Let π_h be the standard piecewise linear Lagrange interpolation operator. Let $\bar{\pi}_h$ be such that*

$$\bar{\pi}_h p|_\tau = \int_\tau p/|\tau| \qquad \forall p \in L^2(\tau),\ \tau \in T^h,$$

where $|\tau|$ is the measure of τ. Then for $1 \leq q \leq \infty$ and $v \in W^{2,q}(\tau)$, $p \in W^{1,q}(\tau)$,

$$|v - \pi_h v|_{W^{m,q}(\tau)} \leq C h_\tau^{2-m} |v|_{W^{2,q}(\tau)}, \qquad m = 0, 1,$$

and

$$\|p - \bar{\pi}_h p\|_{L^q(\tau)} \leq C h_\tau |p|_{W^{1,q}(\tau)}.$$

LEMMA 2.4 (see [46]). *Let $\hat{\pi}_h$ be the average interpolation operator defined in* [46]. *For $m = 0$ or 1, $1 \leq q \leq \infty$, and $v \in W^{1,q}(\Omega)$,*

$$|v - \hat{\pi}_h v|_{W^{m,q}(\tau)} \leq \sum_{\bar{\tau}' \cap \bar{\tau} \neq \emptyset} C h_\tau^{1-m} |v|_{W^{1,q}(\tau')}.$$

LEMMA 2.5 (see [29]). *For $v \in W^{1,q}(\tau)$, $1 \le q < \infty$,*

$$\|v\|_{W^{0,q}(\partial\tau)} \le C(h_\tau^{-\frac{1}{q}}\|v\|_{W^{0,q}(\tau)} + h_\tau^{1-\frac{1}{q}}|v|_{W^{1,q}(\tau)}).$$

3. A posteriori error estimators for Stokes flow control. In order to obtain a numerical solution of acceptable accuracy for the optimal control problem, the finite element meshes have to be refined or adjusted according to a mesh refinement scheme. A widely used approach in engineering is adaptive finite element approximation. At the heart of any adaptive finite element method is an a posteriori error indicator. Adaptive finite element approximation refines or adjusts only the area where the error indicator is large so that a high density of nodes is distributed over the area where the solution is difficult to approximate. Therefore, the error indicator has to reflect reliably the approximation error distribution over the computational domain.

In this section we derive a posteriori error estimators for the optimal control of Stokes flows. It is fairly clear that this is not an easy task, as the solution of the control problem satisfies a complicated coupled variational inequality system (SCP–OPT). There seems to be no existing work in the literature on a posteriori error bounds for a system of such a type. Very recently in [34], [35], and [36] a posteriori error estimates were derived for the finite element approximation of constrained elliptic optimal control problems with distributed or boundary control. In general, the finite element approximation of a flow control problem is more complicated to analyze than that of an elliptic control problem. For example, one has to use the Babuska–Brezzi conditions to handle the mixed finite element formulations. Furthermore, we shall derive sharper a posteriori error estimates for the control approximation and error estimates in the L^2 norm for the state and co-state, since for the control problems of this type, one normally would be more interested in the approximation error of the values of the state and co-state than in their gradient. To this end, we have to modify some techniques from the adjoint approach. Although some of the ideas and techniques used in [34] and [36] are adopted here, we believe that there are substantial differences between the two problems in both the estimates obtained and the methods used.

We first derive a posteriori error bounds for the approximation of the optimal control of (SCP).

We shall first assume that the discretized constraint set $\mathbf{K}^h$ is such that $\mathbf{K}^h \subset \mathbf{K}$. For many cases, it is not difficult to construct such an approximation of $\mathbf{K}$. We shall briefly examine how to relax this condition at the end of this section. We shall further assume that there is a constant $c > 0$ such that

$$(h'(\mathbf{u}) - h'(\mathbf{v}), \mathbf{u} - \mathbf{v})_{\mathbf{U}} \ge c\|\mathbf{u} - \mathbf{v}\|_{\mathbf{U}}^2.$$

We first establish an important lemma for the approximation of the optimal control of (SCP).

LEMMA 3.1. *Let $(\mathbf{y}^*, r^*, \mathbf{p}^*, s^*, \mathbf{u}^*)$ and $(\mathbf{y}_h^*, r_h^*, \mathbf{p}_h^*, s_h^*, \mathbf{u}_h^*)$ be the solutions of (2.5) and (2.9), respectively. Assume that $(B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*))|_{\tau_{\mathbf{U}}} \in (H^1(\tau_{\mathbf{U}}))^2$ and that there is a $\mathbf{v}_h \in \mathbf{K}^h$ such that*

$$|(B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*), \mathbf{v}_h - \mathbf{u}^*)_{\mathbf{U}}| \le C\sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}|B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_{\mathbf{U}}}\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\tau_{\mathbf{U}}}. \tag{3.1}$$

Then we have

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}^2 \le C\sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2|B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_{\mathbf{U}}}^2 + C\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2, \tag{3.2}$$

where $\mathbf{p}(\mathbf{u}_h^*)$ *is the solution of the following system:*

$$\begin{aligned} a(\mathbf{y}(\mathbf{u}_h^*),\mathbf{w}) - b(\mathbf{w}, r(\mathbf{u}_h^*)) &= (\mathbf{f} + B\mathbf{u}_h^*, \mathbf{w}) && \forall \mathbf{w} \in \mathbf{Y}, \\ b(\mathbf{y}(\mathbf{u}_h^*), \phi) &= 0 && \forall \phi \in Q, \\ a(\mathbf{q}, \mathbf{p}(\mathbf{u}_h^*)) + b(\mathbf{q}, s(\mathbf{u}_h^*)) &= (g'(\mathbf{y}(\mathbf{u}_h^*)), \mathbf{q}) && \forall \mathbf{q} \in \mathbf{Y}, \\ b(\mathbf{p}(\mathbf{u}_h^*), \psi) &= 0 && \forall \psi \in Q. \end{aligned} \tag{3.3}$$

Proof. It follows from (SCP–OPT)–(SCP–OPT)^h that for all $\mathbf{v}_h \in \mathbf{K}^h \subset \mathbf{K}$,

$$\begin{aligned} c\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}^2 &\le (h'(\mathbf{u}^*), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} - (h'(\mathbf{u}_h^*), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} \\ &\le -(B^*\mathbf{p}^*, \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} - (h'(\mathbf{u}_h^*), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} \\ &= -(B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} + (B^*(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} \\ &\quad + (B^*(\mathbf{p}(\mathbf{u}_h^*) - \mathbf{p}^*), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} \\ &= -(B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*), \mathbf{u}^* - \mathbf{v}_h)_\mathbf{U} + (B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*), \mathbf{u}_h^* - \mathbf{v}_h)_\mathbf{U} \\ &\quad + (B^*(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U} + (g'(\mathbf{y}(\mathbf{u}_h^*)) - g'(\mathbf{y}^*), \mathbf{y}^* - \mathbf{y}(\mathbf{u}_h^*)) \\ &\le (B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*), \mathbf{v}_h - \mathbf{u}^*)_\mathbf{U} + (B^*(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \mathbf{u}^* - \mathbf{u}_h^*)_\mathbf{U}. \end{aligned}$$

Hence,

$$c\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}^2 \le C\sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}}^2 |B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_\mathbf{U}}^2 + C\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2 + \frac{c}{2}\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}^2.$$

This proves the lemma.

In the following we first derive a posteriori error estimates using the energy norms. We bound up $\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}$ with the energy norms and then obtain the following.

THEOREM 3.1. *Let* $(\mathbf{y}^*, r^*, \mathbf{p}^*, s^*, \mathbf{u}^*)$ *and* $(\mathbf{y}_h^*, r_h^*, \mathbf{p}_h^*, s_h^*, \mathbf{u}_h^*)$ *be the solutions of* (2.5) *and* (2.9)*, respectively. Assume that all the conditions in Lemma* 3.1 *hold. Then*

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_\mathbf{U}^2 + \|\mathbf{y}^* - \mathbf{y}_h^*\|_\mathbf{Y}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_\mathbf{Y}^2 \le C\hat{\eta}^2, \tag{3.4}$$

where

$$\hat{\eta}^2 = \sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}}^2 |B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_\mathbf{U}}^2 + \hat{\eta}_1^2,$$

where

$$\begin{aligned} \hat{\eta}_1^2 = &\sum_\tau h_\tau^2 \int_\tau (\Delta\mathbf{p}_h^* + \nabla s_h^* + g'(\mathbf{y}_h^*))^2 + \sum_l h_l \int_l [\mathbf{A}_l]^2 + \|\mathrm{div}\mathbf{p}_h^*\|_{0,\Omega}^2 \\ &+ \sum_\tau h_\tau^2 \int_\tau (\Delta\mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + B\mathbf{u}_h^*)^2 + \sum_l h_l \int_l [\mathbf{D}_l]^2 + \|\mathrm{div}\mathbf{y}_h^*\|_{0,\Omega}^2, \end{aligned}$$

where h_l *is the size of the edge* l*;* $[\mathbf{A}_l]$ *and* $[\mathbf{D}_l]$ *are jumps on the edge* $l = \bar{\tau}_l^1 \cap \bar{\tau}_l^2$ *defined by*

$$[\mathbf{D}_l] = ((\nabla\mathbf{y}_h^*)_{\tau_l^1} - (\nabla\mathbf{y}_h^*)_{\tau_l^2}) \cdot \mathbf{n} - (r_h^*|_{\tau_l^1} - r_h^*|_{\tau_l^2})\mathbf{n},$$

$$[\mathbf{A}_l] = ((\nabla\mathbf{p}_h^*)_{\tau_l^1} - (\nabla\mathbf{p}_h^*)_{\tau_l^2}) \cdot \mathbf{n} + (s_h^*|_{\tau_l^1} - s_h^*|_{\tau_l^2})\mathbf{n},$$

where $\mathbf{n}$ *is the outer normal direction of* τ_l^1*. For ease of exposition, we let* $[\mathbf{A}_l] = [\mathbf{D}_l] = 0$ *when* $l \subset \partial\Omega$*.*

Proof. We need only to estimate $\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2$. Let $\mathbf{E_p} = \mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)$, and $E_s = s_h^* - s(\mathbf{u}_h^*)$. By Lemma 2.1, there exist $\mathbf{q}$, $\psi \in \mathbf{Y} \times Q$ such that

$$\begin{aligned} &c(\|\mathbf{E_p}\|_{1,\Omega} + \|E_s\|_{0,\Omega})(\|\mathbf{q}\|_{1,\Omega} + \|\psi\|_{0,\Omega}) \\ &\le (\nabla\mathbf{q}, \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}\mathbf{q}, s_h^* - s(\mathbf{u}_h^*)) + (\mathrm{div}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \psi). \end{aligned}$$

Let $\mathbf{q}_I = \hat{\pi}_h\mathbf{q} \in \mathbf{Y}^h$ be the interpolations of $\mathbf{q}$ defined in Lemma 2.4. By (2.9), (3.3) and Lemmas 2.4 and 2.5,

$$\begin{aligned} &c(\|\mathbf{E_p}\|_{1,\Omega} + \|E_s\|_{0,\Omega})(\|\mathbf{q}\|_{1,\Omega} + \|\psi\|_{0,\Omega}) \\ &\le (\nabla\mathbf{q}, \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}\mathbf{q}, s_h^* - s(\mathbf{u}_h^*)) + (\mathrm{div}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \psi) \\ &= (\nabla(\mathbf{q} - \mathbf{q}_I), \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}(\mathbf{q} - \mathbf{q}_I), s_h^* - s(\mathbf{u}_h^*)) + (\nabla\mathbf{q}_I, \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) \\ &\quad + (\mathrm{div}\mathbf{q}_I, s_h^* - s(\mathbf{u}_h^*)) + (\mathrm{div}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \psi) \\ &= -\sum_\tau \int_\tau \Delta(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))(\mathbf{q} - \mathbf{q}_I) + \sum_\tau \int_{\partial\tau} \frac{\partial}{\partial n}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))(\mathbf{q} - \mathbf{q}_I) \\ &\quad - \sum_\tau \int_\tau \nabla(s_h^* - s(\mathbf{u}_h^*))(\mathbf{q} - \mathbf{q}_I) + \sum_\tau \int_{\partial\tau} (s_h^* - s(\mathbf{u}_h^*))(\mathbf{q} - \mathbf{q}_I) \cdot \mathbf{n} \\ &\quad + (g'(\mathbf{y}_h^*) - g'(\mathbf{y}(\mathbf{u}_h^*)), \mathbf{q}_I) + (\mathrm{div}\mathbf{p}_h^*, \psi) \\ &= \sum_\tau \int_\tau (-\Delta\mathbf{p}_h^* - \nabla s_h^* - g'(\mathbf{y}(\mathbf{u}_h^*)))(\mathbf{q} - \mathbf{q}_I) + \sum_l \int_l [\mathbf{A}_l](\mathbf{q} - \mathbf{q}_I) \\ &\quad + (g'(\mathbf{y}_h^*) - g'(\mathbf{y}(\mathbf{u}_h^*)), \mathbf{q}_I) + (\mathrm{div}\mathbf{p}_h^*, \psi) \\ &= \sum_\tau \int_\tau (-\Delta\mathbf{p}_h^* - \nabla s_h^* - g'(\mathbf{y}_h^*))(\mathbf{q} - \mathbf{q}_I) + (g'(\mathbf{y}_h^*) - g'(\mathbf{y}(\mathbf{u}_h^*)), \mathbf{q} - \mathbf{q}_I) \\ &\quad + \sum_l \int_l [\mathbf{A}_l](\mathbf{q} - \mathbf{q}_I) + (g'(\mathbf{y}_h^*) - g'(\mathbf{y}(\mathbf{u}_h^*)), \mathbf{q}_I) + (\mathrm{div}\mathbf{p}_h^*, \psi) \\ &\le C\bigg(\sum_\tau h_\tau^2 \int_\tau (\Delta\mathbf{p}_h^* + g'(\mathbf{y}_h^*) + \nabla s_h^*)^2 + \sum_l h_l \int_l [\mathbf{A}_l]^2 + \|\mathrm{div}\mathbf{p}_h^*\|_{0,\Omega}^2 \\ &\quad + \|\mathbf{y}(\mathbf{u}_h^*) - \mathbf{y}_h^*\|_{0,\Omega}^2\bigg)^{\frac{1}{2}} (\|\mathbf{q}\|_{1,\Omega} + \|\psi\|_{0,\Omega}). \end{aligned}$$

Then, it follows that

$$\begin{aligned} \|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{1,\Omega} \le C\bigg(&\sum_\tau h_\tau^2 \int_\tau (\Delta\mathbf{p}_h^* + g'(\mathbf{y}_h^*) + \nabla s_h^*)^2 + \sum_l h_l \int_l [\mathbf{A}_l]^2 \\ &+ \|\mathrm{div}\mathbf{p}_h^*\|_{0,\Omega}^2 + \|\mathbf{y}(\mathbf{u}_h^*) - \mathbf{y}_h^*\|_{0,\Omega}^2\bigg)^{\frac{1}{2}}. \end{aligned} \tag{3.5}$$

Similarly, let $\mathbf{E_y} = \mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)$, $E_r = r_h^* - r(\mathbf{u}_h^*)$. Let $\mathbf{w}_I = \hat{\pi}_h\mathbf{w} \in \mathbf{Y}^h$ be the interpolation of $\mathbf{w} \in \mathbf{Y}$ defined in Lemma 2.4. Then, we can prove that

$$\begin{aligned} &c(\|\mathbf{E_y}\|_{1,\Omega} + \|E_r\|_{0,\Omega})(\|\mathbf{w}\|_{1,\Omega} + \|\phi\|_{0,\Omega}) \\ &\le (\nabla(\mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)), \nabla\mathbf{w}) - (\mathrm{div}\mathbf{w}, r_h^* - r(\mathbf{u}_h^*)) - (\mathrm{div}(\mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)), \phi) \\ &= (\nabla(\mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)), \nabla(\mathbf{w} - \mathbf{w}_I)) - (\mathrm{div}(\mathbf{w} - \mathbf{w}_I), r_h^* - r(\mathbf{u}_h^*)) - (\mathrm{div}(\mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)), \phi) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{\tau}\int_{\tau}\Delta(\mathbf{y}_h^*-\mathbf{y}(u_h^*))(\mathbf{w}-\mathbf{w}_I)+\sum_{\tau}\int_{\partial\tau}\frac{\partial}{\partial n}(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*))(\mathbf{w}-\mathbf{w}_I)\\
&\quad+\sum_{\tau}\int_{\tau}\nabla(r_h^*-r(\mathbf{u}_h^*))(\mathbf{w}-\mathbf{w}_I)-\sum_{\tau}\int_{\partial\tau}(r_h^*-r(\mathbf{u}_h^*))(\mathbf{w}-\mathbf{w}_I)\cdot\mathbf{n}-(\mathrm{div}\mathbf{y}_h^*,\phi)\\
&=\sum_{\tau}\int_{\tau}(-\Delta\mathbf{y}_h^*+\nabla r_h^*-\mathbf{f}-B\mathbf{u}_h^*)(\mathbf{w}-\mathbf{w}_I)+\sum_{l}\int_{l}[\mathbf{D}_l](\mathbf{w}-\mathbf{w}_I)-(\mathrm{div}\mathbf{y}_h^*,\phi)\\
&\le C\Bigg(\sum_{\tau}h_\tau^2\int_{\tau}(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*)^2\\
&\quad+\sum_{l}h_l\int_{l}[\mathbf{D}_l]^2+\|\mathrm{div}\mathbf{y}_h^*\|_{0,\Omega}^2\Bigg)^{\frac{1}{2}}(\|\mathbf{w}\|_{1,\Omega}+\|\phi\|_{0,\Omega}).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\|\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)\|_{1,\Omega}^2\\
&\qquad\le C\Bigg(\sum_{\tau}h_\tau^2\int_{\tau}(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*)^2+\sum_{l}h_l\int_{l}[\mathbf{D}_l]^2+\|\mathrm{div}\mathbf{y}_h^*\|_{0,\Omega}^2\Bigg).
\end{aligned}
$$

Then, it follows from (3.5) that

$$\|\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)\|_{1,\Omega}^2\le C\hat{\eta}_1^2.$$

Then by Lemma 3.1,

$$\|\mathbf{u}_h^*-\mathbf{u}^*\|_{0,\Omega_{\mathbf{U}}}^2\le C\hat{\eta}^2. \tag{3.6}$$

Note that

$$\|\mathbf{y}^*-\mathbf{y}(\mathbf{u}_h^*)\|_{1,\Omega}\le C\|\mathbf{u}_h^*-\mathbf{u}^*\|_{0,\Omega_{\mathbf{U}}},$$

and

$$\|\mathbf{p}^*-\mathbf{p}(\mathbf{u}_h^*)\|_{1,\Omega}\le C\|\mathbf{y}^*-\mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega}\le C\|\mathbf{u}^*-\mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}.$$

We have

$$\|\mathbf{y}^*-\mathbf{y}_h^*\|_{1,\Omega}^2+\|\mathbf{p}^*-\mathbf{p}_h^*\|_{1,\Omega}^2\le C\hat{\eta}^2. \tag{3.7}$$

Then, the theorem follows from (3.6) and (3.7).

Remark 3.1. It is clear that the a posteriori error estimator $\hat{\eta}$ consists of two parts. The part $\hat{\eta}_1$ is contributed from the approximation error of the state and co-state equations, and the other part results from the approximation error of the variational inequality. Among them, $\hat{\eta}_1$ mainly indicates the approximation error for the state and co-state, and the term $\sum_{\tau_{\mathbf{U}}\subset T_{\mathbf{U}}^h}h_{\tau_{\mathbf{U}}}^2|h'(\mathbf{u}_h^*)+B^*\mathbf{p}_h^*|_{H^1(\tau_{\mathbf{U}})}^2$ mainly reflects the approximation error for the control. There does not have to exist a relationship between the computational meshes for the state and those for the control. Clearly, the most suitable implementation and thus the optimal mesh refinements will much depend on what is the most important quantity to be computed in a particular control problem. It also depends on the structure of the meshes used in the computations. Further investigation is still much needed.

The part $\hat{\eta}_1$ can be further divided into two parts: one from the approximation error of the state equation and the other from that of the co-state equation. Clearly, a posteriori error estimators obtained solely from the state equation may fail to reflect the main approximation error of the optimal control problem and thus fail to yield efficient mesh refinements. This will become even more clear after Theorem 3.2.

It follows from the above proof that when Ω is convex, one can bound up $\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}$ with the L^2 norm, using the dual equations, and thus derive the sharper estimates for $\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}$. This is what we are going to do in the following theorem. Furthermore we derive a posteriori error estimates for the control problem using the L^2 norms. If one is more interested in the control and the *values* of the state, then the following result may be very useful.

THEOREM 3.2. *Let* $(\mathbf{y}^*, r^*, \mathbf{p}^*, s^*, \mathbf{u}^*)$ *and* $(\mathbf{y}_h^*, r_h^*, \mathbf{p}_h^*, s_h^*, \mathbf{u}_h^*)$ *be the solutions of* (2.5) *and* (2.9)*, respectively. Assume that all the conditions in Lemma* 2.2 *and Lemma* 3.1 *hold. Then*

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}^2 + \|\mathbf{y}^* - \mathbf{y}_h^*\|_{0,\Omega}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_{0,\Omega}^2 \le C\eta^2, \tag{3.8}$$

where

$$\eta^2 = \sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}}^2 |B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_\mathbf{U}}^2 + \eta_1^2,$$

where

$$\begin{aligned}\eta_1^2 = {} & \sum_\tau h_\tau^4 \int_\tau (\Delta\mathbf{p}_h^* + \nabla s_h^* + g'(\mathbf{y}_h^*))^2 + \sum_l h_l^3 \int_l [\mathbf{A}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{p}_h^*)^2 \\ & + \sum_\tau h_\tau^4 \int_\tau (\Delta\mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + B\mathbf{u}_h^*)^2 + \sum_l h_l^3 \int_l [\mathbf{D}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{y}_h^*)^2,\end{aligned}$$

where h_l *is the size of the edge* l*;* $[\mathbf{A}_l]$ *and* $[\mathbf{D}_l]$ *are jumps on the edge* $l = \bar{\tau}_l^1 \cap \bar{\tau}_l^2$ *defined by*

$$[\mathbf{D}_l] = ((\nabla\mathbf{y}_h^*)_{\tau_l^1} - (\nabla\mathbf{y}_h^*)_{\tau_l^2}) \cdot \mathbf{n} - (r_h^*|_{\tau_l^1} - r_h^*|_{\tau_l^2})\mathbf{n},$$

$$[\mathbf{A}_l] = ((\nabla\mathbf{p}_h^*)_{\tau_l^1} - (\nabla\mathbf{p}_h^*)_{\tau_l^2}) \cdot \mathbf{n} + (s_h^*|_{\tau_l^1} - s_h^*|_{\tau_l^2})\mathbf{n},$$

where $\mathbf{n}$ *is the outer normal direction of* τ_l^1*. For ease of exposition, we let* $[\mathbf{A}_l] = [\mathbf{D}_l] = 0$ *when* $l \subset \partial\Omega$*.*

Proof. Let (Ψ, ρ) be the solution of (2.3) with $\Phi = \mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)$ and + sign. Let $\Psi_I = \pi_h\Psi \in \mathbf{Y}^h$ and $\rho_I = \bar{\pi}_h\rho \in Q^h$ be the interpolations of Ψ and ρ defined in Lemma 2.3. By (2.9), (3.3) and Lemmas 2.2, 2.3, and 2.5,

$$\begin{aligned}& \|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2 = (\Phi, \mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)) \\ & = (\nabla\Psi, \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \rho) + (\mathrm{div}\Psi, s_h^* - s(\mathbf{u}_h^*)) \\ & = (\nabla(\Psi - \Psi_I), \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}(\Psi - \Psi_I), s_h^* - s(\mathbf{u}_h^*)) \\ & \quad + (\nabla\Psi_I, \nabla(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))) + (\mathrm{div}\Psi_I, s_h^* - s(\mathbf{u}_h^*)) + (\mathrm{div}\mathbf{p}_h^*, \rho - \rho_I) \\ & = -\sum_\tau \int_\tau \Delta(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))(\Psi - \Psi_I) + \sum_\tau \int_{\partial\tau} \frac{\partial}{\partial n}(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))(\Psi - \Psi_I) \\ & \quad - \sum_\tau \int_\tau \nabla(s_h^* - s(\mathbf{u}_h^*))(\Psi - \Psi_I) + \sum_\tau \int_{\partial\tau} (s_h^* - s(\mathbf{u}_h^*))(\Psi - \Psi_I) \cdot \mathbf{n}\end{aligned}$$

$$
\begin{aligned}
&+(g'(\mathbf{y}_h^*)-g'(\mathbf{y}(\mathbf{u}_h^*)),\Psi_I)+(\mathrm{div}\mathbf{p}_h^*,\rho-\rho_I)\\
&\le C\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{p}_h^*+\nabla s_h^*+g'(\mathbf{y}_h^*))^2+C\sum_l h_l^3\int_l[\mathbf{A}_l]^2+C\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{p}_h^*)^2\\
&\quad+C\|\mathbf{y}(\mathbf{u}_h^*)-\mathbf{y}_h^*\|_{0,\Omega}^2+C\delta(\|\Psi\|_{2,\Omega}^2+\|\rho\|_{1,\Omega}^2)\\
&\le C\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{p}_h^*+\nabla s_h^*+g'(\mathbf{y}_h^*))^2+C\sum_l h_l^3\int_l[\mathbf{A}_l]^2+C\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{p}_h^*)^2\\
&\quad+C\|\mathbf{y}(\mathbf{u}_h^*)-\mathbf{y}_h^*\|_{0,\Omega}^2+C\delta\|\Phi\|_{0,\Omega}^2\\
&=C\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{p}_h^*+\nabla s_h^*+g'(\mathbf{y}_h^*))^2+C\sum_l h_l^3\int_l[\mathbf{A}_l]^2+C\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{p}_h^*)^2\\
&\quad+C\|\mathbf{y}(\mathbf{u}_h^*)-\mathbf{y}_h^*\|_{0,\Omega}^2+C\delta\|\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2.
\end{aligned}
$$

Then, let $\delta=\frac{1}{2C}$; it follows that

$$
\begin{aligned}
\|\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2&\le C\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{p}_h^*+\nabla s_h^*+g'(\mathbf{y}_h^*))^2+C\sum_l h_l^3\int_l[\mathbf{A}_l]^2\\
&\quad+C\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{p}_h^*)^2+C\|\mathbf{y}(\mathbf{u}_h^*)-\mathbf{y}_h^*\|_{0,\Omega}^2.
\end{aligned}
\tag{3.9}
$$

Similarly, let (Ψ,ρ) be the solution of (2.3) with $\Phi=\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)$ and $-$ sign. Let $\Psi_I=\pi_h\Psi\in\mathbf{Y}^h$ and $\rho_I=\bar{\pi}_h\rho\in Q^h$ be the interpolations of Ψ and ρ defined in Lemma 2.3. Again, we can prove that

$$
\begin{aligned}
&\|\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega}^2=(\Phi,\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*))\\
&=(\nabla(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)),\nabla\Psi)-(\mathrm{div}\Psi,r_h^*-r(\mathbf{u}_h^*))-(\mathrm{div}(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)),\rho)\\
&=(\nabla(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)),\nabla(\Psi-\Psi_I))-(\mathrm{div}(\Psi-\Psi_I),r_h^*-r(\mathbf{u}_h^*))-(\mathrm{div}\mathbf{y}_h^*,\rho-\rho_I)\\
&=-\sum_\tau\int_\tau\Delta(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*))(\Psi-\Psi_I)+\sum_\tau\int_{\partial\tau}\frac{\partial}{\partial n}(\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*))(\Psi-\Psi_I)\\
&\quad+\sum_\tau\int_\tau\nabla(r_h^*-r(\mathbf{u}_h^*))(\Psi-\Psi_I)-\sum_\tau\int_{\partial\tau}(r_h^*-r(\mathbf{u}_h^*))(\Psi-\Psi_I)\cdot\mathbf{n}\\
&\quad-(\mathrm{div}\mathbf{y}_h^*,\rho-\rho_I)\\
&=\sum_\tau\int_\tau(-\Delta\mathbf{y}_h^*+\nabla r_h^*-\mathbf{f}-B\mathbf{u}_h^*)(\Psi-\Psi_I)+\sum_l\int_l[\mathbf{D}_l](\Psi-\Psi_I)-(\mathrm{div}\mathbf{y}_h^*,\rho-\rho_I)\\
&\le C\Big(\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*)^2+\sum_l h_l^3\int_l[\mathbf{D}_l]^2+\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{y}_h^*)^2\Big)\\
&\quad+C\delta(\|\Psi\|_{2,\Omega}^2+\|\rho\|_{1,\Omega}^2)\\
&\le C\Big(\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*)^2\\
&\quad+\sum_l h_l^3\int_l[\mathbf{D}_l]^2+\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{y}_h^*)^2\Big)+C\delta\|\Phi\|_{0,\Omega}^2\\
&=C\Big(\sum_\tau h_\tau^4\int_\tau(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*)^2+\sum_l h_l^3\int_l[\mathbf{D}_l]^2+\sum_\tau h_\tau^2\int_\tau(\mathrm{div}\mathbf{y}_h^*)^2\Big)\\
&\quad+C\delta\|\mathbf{y}_h^*-\mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega}^2.
\end{aligned}
$$

It follows that

$$\|\mathbf{y}_h^* - \mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega}^2 \leq C\Big(\sum_\tau h_\tau^4 \int_\tau (\Delta \mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + B\mathbf{u}_h^*)^2 + \sum_l h_l^3 \int_l [\mathbf{D}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{y}_h^*)^2\Big).$$

Then, it follows from (3.9) that

$$\|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2 \leq C\eta_1^2.$$

Then by Lemma 3.1,

$$\|\mathbf{u}_h^* - \mathbf{u}^*\|_{0,\Omega_\mathbf{U}}^2 \leq C\eta^2. \tag{3.10}$$

Note that

$$\|\mathbf{y}^* - \mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega} \leq C\|\mathbf{u}_h^* - \mathbf{u}^*\|_{0,\Omega_\mathbf{U}},$$

and

$$\|\mathbf{p}^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega} \leq C\|\mathbf{y}^* - \mathbf{y}(\mathbf{u}_h^*)\|_{0,\Omega} \leq C\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}.$$

We have

$$\|\mathbf{y}^* - \mathbf{y}_h^*\|_{0,\Omega}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_{0,\Omega}^2 \leq C\eta^2. \tag{3.11}$$

Then, the theorem follows from (3.10) and (3.11).

It follows from Theorem 3.2 that when using the same meshes for the state and the control, the estimator

$$\sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}}^2 |B^*\mathbf{p}_h^* + h'(\mathbf{u}_h^*)|_{1,\tau_\mathbf{U}}^2$$

is *dominant*! This fact could be useful in simplifying implementation of the resulting error estimates in computations and have impacts on developing implementation techniques. Furthermore, it is not hard to see that one could use much coarser meshes for the state and co-state approximation, since η_1 has a different approximation order. Thus, computational efficiency could be further improved if we use different meshes for the state and the control. It is clear that much more research is needed to investigate possible implementations of these error indicators with multiset meshes.

Remark 3.2. The techniques developed in this paper can be extended to derive similar a posteriori error estimates for the optimal control problem governed by the Navier–Stokes equations (at least when the Reynolds number is small):

$$\begin{aligned} &\min_{\mathbf{u}\in\mathbf{K}\subset\mathbf{U}} \{g(\mathbf{y}) + h(\mathbf{u})\}, \\ &-\Delta\mathbf{y} + Re(\mathbf{y}\cdot\nabla)\mathbf{y} + \nabla r = \mathbf{f} + B\mathbf{u} \qquad \text{in } \Omega, \\ &\mathrm{div}\mathbf{y} = 0 \text{ in } \Omega, \ \ \mathbf{y} = 0 \ \text{ on } \ \partial\Omega, \quad \int_\Omega r = 0, \end{aligned}$$

where Re is the Reynolds number. It is well known that $(\mathbf{y}^*, r^*, \mathbf{u}^*)$ is the solution of above problem only if there is a co-state $(\mathbf{p}^*, s^*) \in \mathbf{Y}\times Q$ such that $(\mathbf{y}^*, r^*, \mathbf{p}^*, s^*, \mathbf{u}^*)$

satisfies the following optimality conditions:

$$
\begin{aligned}
(3.12)\quad a(\mathbf{y}^*,\mathbf{w})+(Re(\mathbf{y}^*\cdot\nabla)\mathbf{y}^*,\mathbf{w})-b(\mathbf{w},r^*)&=(\mathbf{f}+B\mathbf{u}^*,\mathbf{w}) && \forall \mathbf{w}\in\mathbf{Y},\\
b(\mathbf{y}^*,\phi)&=0 && \forall \phi\in Q,\\
a(\mathbf{q},\mathbf{p}^*)-Re((\mathbf{y}^*\cdot\nabla)\mathbf{p}^*,\mathbf{q})&\\
+Re((\nabla\mathbf{y}^*)^T\mathbf{p}^*,\mathbf{q})+b(\mathbf{q},s^*)&=(g'(\mathbf{y}^*),\mathbf{q}) && \forall \mathbf{q}\in\mathbf{Y},\\
b(\mathbf{p}^*,\psi)&=0 && \forall \psi\in Q,\\
(h'(\mathbf{u}^*)+B^*\mathbf{p}^*,\mathbf{v}-\mathbf{u}^*)_{\mathbf{U}}&\geq 0 && \forall \mathbf{v}\in\mathbf{K}\subset\mathbf{U},
\end{aligned}
$$

where B^* is the adjoint operator of B, $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ are defined in section 2, and $(\cdot,\cdot)_{\mathbf{U}}$ is the inner product of $\mathbf{U}$.

Although rigorous proof details have yet to be worked out, one already should be in the position of seeing similar a posteriori error estimates for the above control problem with smaller Re by using the same ideas and techniques. For example, let $(\mathbf{y}^*,r^*,\mathbf{p}^*,s^*,\mathbf{u}^*)$ be the solutions of (3.12) and $(\mathbf{y}_h^*,r_h^*,\mathbf{p}_h^*,s_h^*,\mathbf{u}_h^*)$ be the finite element approximations of $(\mathbf{y}^*,r^*,\mathbf{p}^*,s^*,\mathbf{u}^*)$. Assume that all the conditions in Lemma 3.1 hold. Then

$$
(3.13)\qquad \|\mathbf{u}^*-\mathbf{u}_h^*\|_{\mathbf{U}}^2+\|\mathbf{y}^*-\mathbf{y}_h^*\|_{\mathbf{Y}}^2+\|\mathbf{p}^*-\mathbf{p}_h^*\|_{\mathbf{Y}}^2\leq C\tilde{\eta}^2,
$$

where

$$
\tilde{\eta}^2=\sum_{\tau_{\mathbf{U}}}h_{\tau_{\mathbf{U}}}^2|B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*)|_{1,\tau_{\mathbf{U}}}^2+\tilde{\eta}_1^2,
$$

where

$$
\begin{aligned}
\tilde{\eta}_1^2=&\sum_{\tau}h_\tau^2\int_\tau(\Delta\mathbf{p}_h^*+\nabla s_h^*+g'(\mathbf{y}_h^*)+Re(\mathbf{y}_h^*\cdot\nabla)\mathbf{p}_h^*-Re(\nabla\mathbf{y}_h^*)^T\mathbf{p}_h^*)^2\\
&+\sum_l h_l\int_l[\mathbf{A}_l]^2+\|\mathrm{div}\mathbf{p}_h^*\|_{0,\Omega}^2\\
&+\sum_\tau h_\tau^2\int_\tau(\Delta\mathbf{y}_h^*-\nabla r_h^*+\mathbf{f}+B\mathbf{u}_h^*-Re(\mathbf{y}_h^*\cdot\nabla)\mathbf{y}_h^*)^2+\sum_l h_l\int_l[\mathbf{D}_l]^2+\|\mathrm{div}\mathbf{y}_h^*\|_{0,\Omega}^2.
\end{aligned}
$$

One can also write down the error estimates in the L^2 norm.

In the following, we extend Lemma 3.1 to the cases where $\mathbf{K}^h\not\subset\mathbf{K}$. Consequently Theorems 3.1 and 3.2 can also be extended to the cases using the same methods used in the proofs of Theorems 3.1 and 3.2.

Remark 3.3. Assume that all the conditions in Lemma 3.1 hold except the condition that $\mathbf{K}^h\subset\mathbf{K}$. Use the same notation in Lemma 3.1. Assume that $(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*))|_{\tau_{\mathbf{U}}}\in(H^1(\tau_{\mathbf{U}}))^2$ and that there is a $\mathbf{v}_h\in\mathbf{K}^h$ such that

$$
\begin{aligned}
(3.14)\qquad &|(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*),\mathbf{v}_h-\mathbf{u}^*)_{\mathbf{U}}|\\
&\quad\leq C\sum_{\tau_{\mathbf{U}}\subset T_{\mathbf{U}}^h}h_{\tau_{\mathbf{U}}}|B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*)|_{1,\tau_{\mathbf{U}}}\|\mathbf{u}^*-\mathbf{u}_h^*\|_{0,\tau_{\mathbf{U}}}.
\end{aligned}
$$

Then for all $\mathbf{v}\in\mathbf{K}$ we have

$$
\|\mathbf{u}^*-\mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}^2\leq C\Bigg(\sum_{\tau_{\mathbf{U}}\subset T_{\mathbf{U}}^h}h_{\tau_{\mathbf{U}}}^2|B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*)|_{1,\tau_{\mathbf{U}}}^2+\|\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2
$$

$$+|(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|+|(B^*(\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}| \tag{3.15}$$
$$+|(B^*(\mathbf{p}(\mathbf{u}_h^*)-\mathbf{p}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|+|(h'(\mathbf{u}_h^*)-h'(\mathbf{u}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|\Big),$$

where $\mathbf{p}(\mathbf{u}_h^*)$ is defined in Lemma 3.1.

The proof of this estimate is similar to that of Lemma 3.1. First we have

$$(h'(\mathbf{u}^*),\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}\le -(B^*\mathbf{p}^*,\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}\qquad\forall\mathbf{v}\in\mathbf{K},$$

$$(h'(\mathbf{u}_h^*),\mathbf{u}_h^*-\mathbf{v}_h)_{\mathbf{U}}+(B^*\mathbf{p}_h^*,\mathbf{u}_h^*-\mathbf{v}_h)_{\mathbf{U}}\le 0\qquad\forall\mathbf{v}_h\in\mathbf{K}^h.$$

Then by applying the same techniques used in the proof of Lemma 3.1, we have for any $\mathbf{v}_h\in\mathbf{K}^h$ and $\mathbf{v}\in\mathbf{K}$ that

$$c\|\mathbf{u}^*-\mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}^2\le(h'(\mathbf{u}^*),\mathbf{u}^*-\mathbf{u}_h^*)_{\mathbf{U}}-(h'(\mathbf{u}_h^*),\mathbf{u}^*-\mathbf{u}_h^*)_{\mathbf{U}}$$
$$=(h'(\mathbf{u}^*),\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}+(h'(\mathbf{u}^*),\mathbf{v}-\mathbf{u}_h^*)_{\mathbf{U}}-(h'(\mathbf{u}_h^*),\mathbf{u}^*-\mathbf{u}_h^*)_{\mathbf{U}}$$
$$\le-(B^*\mathbf{p}^*,\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}+(h'(\mathbf{u}^*),\mathbf{v}-\mathbf{u}_h^*)_{\mathbf{U}}-(h'(\mathbf{u}_h^*),\mathbf{u}^*-\mathbf{u}_h^*)_{U}$$
$$=-(B^*\mathbf{p}_h^*,\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}+(B^*(\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)),\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}+(B^*(\mathbf{p}(\mathbf{u}_h^*)-\mathbf{p}^*),\mathbf{u}^*-\mathbf{u}_h^*)_{\mathbf{U}}$$
$$+(B^*(\mathbf{p}(\mathbf{u}_h^*)-\mathbf{p}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}+(h'(\mathbf{u}^*),\mathbf{v}-\mathbf{u}_h^*)_{\mathbf{U}}-(h'(\mathbf{u}_h^*),\mathbf{u}^*-\mathbf{v}+\mathbf{v}-\mathbf{u}_h^*)_{\mathbf{U}}$$
$$=(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*),(\mathbf{v}-\mathbf{u}_h^*)+(\mathbf{u}_h^*-\mathbf{v}_h)+(\mathbf{v}_h-\mathbf{u}^*))_{\mathbf{U}}$$
$$+(B^*(\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)),\mathbf{u}^*-\mathbf{v})_{\mathbf{U}}-(g'(\mathbf{y}^*)-g'(\mathbf{y}(\mathbf{u}_h^*)),\mathbf{y}^*-\mathbf{y}(\mathbf{u}_h^*))$$
$$+(B^*(\mathbf{p}(\mathbf{u}_h^*)-\mathbf{p}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}+(h'(\mathbf{u}^*)-h'(\mathbf{u}_h^*),\mathbf{v}-\mathbf{u}_h^*)_{\mathbf{U}}$$
$$\le|(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*),\mathbf{v}_h-\mathbf{u}^*)_{\mathbf{U}}|+|(B^*\mathbf{p}_h^*+h'(\mathbf{u}_h^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|$$
$$+|(B^*(\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)),\mathbf{u}^*-\mathbf{u}_h^*)_{\mathbf{U}}|+|(B^*(\mathbf{p}_h^*-\mathbf{p}(\mathbf{u}_h^*)),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|$$
$$+|(B^*(\mathbf{p}(\mathbf{u}_h^*)-\mathbf{p}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|+|(h'(\mathbf{u}_h^*)-h'(\mathbf{u}^*),\mathbf{u}_h^*-\mathbf{v})_{\mathbf{U}}|.$$

Then the desired result (3.15) follows.

4. Some examples. In this section we consider some applications of the results in section 3. First let $\Omega=[-1,1]\times[-1,1]$. Let $\mathbf{Y}=(H_0^1(\Omega))^2$, $\mathbf{U}=(L^2(\Omega))^2$, $\mathbf{H}=(L^2(\Omega))^2$, and $Q=L_0^2(\Omega)$. For any $\mathbf{u}\in\mathbf{U}$, let $B\mathbf{u}=\mathbf{u}$. Now let us consider the following well-known quadratic control problem (see [31]), (EXP):

$$\min_{\mathbf{u}\in\mathbf{K}\subset\mathbf{U}}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{z}_d\|_{\mathbf{H}}^2+\frac{1}{2}\|\mathbf{u}\|_{\mathbf{U}}^2\right\},$$
$$-\Delta\mathbf{y}+\nabla r=\mathbf{f}+\mathbf{u}\text{ in }\Omega,$$
$$\text{div}\mathbf{y}=0\text{ in }\Omega,$$
$$\mathbf{y}=0\ \text{ on }\ \partial\Omega,\quad\int_\Omega r=0,$$

where $\mathbf{f}\in\mathbf{L}=(L^2(\Omega))^2$, $\mathbf{z}_d\in\mathbf{H}$, and $\mathbf{K}$ is a closed convex set in $\mathbf{U}$.

Example 4.1. Let $\mathbf{K}_1 = \mathbf{U}$, $\mathbf{K}_2 = \{\mathbf{u} \in \mathbf{U} : \mathbf{u} \geq 0\}$, $\mathbf{K}_3 = \{\mathbf{u} \in \mathbf{U} : |\mathbf{u}| \leq 1\}$, and $\mathbf{K}_4 = \{\mathbf{u} \in \mathbf{U} : \int_\Omega \mathbf{u} \geq 0\}$. Here the notations $\mathbf{u} \geq 0$, $|\mathbf{u}| \leq 1$, $\int_\Omega \mathbf{u} \geq 0$ mean that $u_i \geq 0$, $|u_i| \leq 1$, $\int_\Omega u_i \geq 0$, with $i = 1, 2$, and $\mathbf{u} = (u_1, u_2)$. Construct the finite element approximation as in section 2. Especially, we take $\mathbf{U}^h = \{\mathbf{u} \in \mathbf{U} : \mathbf{u}|_{\tau_\mathbf{U}} \in (P_0)^2\}$, where P_0 is the 0-order polynomial space, and take $\mathbf{K}_1^h = \mathbf{U}^h$, $\mathbf{K}_2^h = \{\mathbf{u}_h \in \mathbf{U}^h : \mathbf{u}_h \geq 0\}$, $\mathbf{K}_3^h = \{\mathbf{u}_h \in \mathbf{U}^h : |\mathbf{u}_h| \leq 1\}$, and $\mathbf{K}_4^h = \{\mathbf{u}_h \in \mathbf{U}^h : \int_\Omega \mathbf{u}_h \geq 0\}$. For any $\mathbf{u} \in \mathbf{K}$, define $(\pi_h^c \mathbf{u})|_{\tau_\mathbf{U}} = \bar{\mathbf{u}}|_{\tau_\mathbf{U}}$, where $\bar{\mathbf{u}}|_{\tau_\mathbf{U}}$ is the integral average of $\mathbf{u}$ on the element $\tau_\mathbf{U}$. Then, it is easy to see that $\mathbf{K}_i^h \subset \mathbf{K}_i$, and $\pi_h^c \mathbf{u} \in \mathbf{K}_i^h$ if $\mathbf{u} \in \mathbf{K}_i$, $i = 1, 2, 3, 4$.

Let $\mathbf{u}^*$ be the solution of (EXP) and $\mathbf{u}_h^*$ be the solution of the finite element approximation of (EXP). It follows that $(\mathbf{p}_h^* + \mathbf{u}_h^*)|_{\tau_\mathbf{U}} \in (H^1(\tau_\mathbf{U}))^2$. Then

$$\begin{aligned}|(\mathbf{p}_h^* + \mathbf{u}_h^*, \pi_h^c \mathbf{u}^* - \mathbf{u}^*)_\mathbf{U}| &= |(\mathbf{p}_h^* + \mathbf{u}_h^* - \pi_h^c(\mathbf{p}_h^* + \mathbf{u}_h^*), \pi_h^c(\mathbf{u}^* - \mathbf{u}_h^*) - (\mathbf{u}^* - \mathbf{u}_h^*))_\mathbf{U}| \\ &\leq C \sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}} |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_\mathbf{U}} \|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\tau_\mathbf{U}}.\end{aligned}$$

Therefore, Theorems 3.1 and 3.2 hold:

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_\mathbf{U}^2 + \|\mathbf{y}^* - \mathbf{y}_h^*\|_\mathbf{Y}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_\mathbf{Y}^2 \leq C\hat{\eta}^2, \tag{4.1}$$

where

$$\hat{\eta}^2 = \sum_{\tau_\mathbf{U} \subset T_\mathbf{U}^h} h_{\tau_\mathbf{U}}^2 |\mathbf{u}_h^* + \mathbf{p}_h^*|_{1,\tau_\mathbf{U}}^2 + \hat{\eta}_1^2,$$

$$\begin{aligned}\hat{\eta}_1^2 = &\sum_\tau h_\tau^2 \int_\tau (\Delta \mathbf{p}_h^* + \nabla s_h^* + \mathbf{y}_h^* - \mathbf{z}_d)^2 + \sum_l h_l \int_l [\mathbf{A}_l]^2 + \|\mathrm{div}\mathbf{p}_h^*\|_{0,\Omega}^2 \\ &+ \sum_\tau h_\tau^2 \int_\tau (\Delta \mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + \mathbf{u}_h^*)^2 + \sum_l h_l \int_l [\mathbf{D}_l]^2 + \|\mathrm{div}\mathbf{y}_h^*\|_{0,\Omega}^2.\end{aligned}$$

Also,

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_\mathbf{U}}^2 + \|\mathbf{y}^* - \mathbf{y}_h^*\|_{0,\Omega}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_{0,\Omega}^2 \leq C\eta^2, \tag{4.2}$$

where

$$\eta^2 = \sum_{\tau_\mathbf{U} \subset T_\mathbf{U}^h} h_{\tau_\mathbf{U}}^2 |\mathbf{u}_h^* + \mathbf{p}_h^*|_{1,\tau_\mathbf{U}}^2 + \eta_1^2,$$

$$\begin{aligned}\eta_1^2 = &\sum_\tau h_\tau^4 \int_\tau (\Delta \mathbf{p}_h^* + \nabla s_h^* + \mathbf{y}_h^* - \mathbf{z}_d)^2 + \sum_l h_l^3 \int_l [\mathbf{A}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{p}_h^*)^2 \\ &+ \sum_\tau h_\tau^4 \int_\tau (\Delta \mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + \mathbf{u}_h^*)^2 + \sum_l h_l^3 \int_l [\mathbf{D}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{y}_h^*)^2.\end{aligned}$$

One can use $\sum_{\tau_\mathbf{U}} h_{\tau_\mathbf{U}}^2 |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_\mathbf{U}}^2$ as an error indicator for the control approximation mesh refinement and use η_1^2 (or $\hat{\eta}_1^2$) for the state and co-state mesh refinement.

The (single) obstacle problem case where the obstacle is not constant can be easily dealt with, for instance, by using the transformation $\mathbf{u}_{new} = \mathbf{u}_{old} - \Phi$, where Φ is the obstacle. Now let us consider the two side obstacle case.

Example 4.2. Let $K = \{\mathbf{u} \in \mathbf{U} : \mathbf{g}_0 \le \mathbf{u} \le \mathbf{g}_1\}$, where $\mathbf{g}_0, \mathbf{g}_1 \in (H^1(\Omega))^2$. Let $\mathbf{U}^h$ be the piecewise constant space. Let $\mathbf{K}^h = \{\mathbf{u}_h \in \mathbf{U}^h : \pi_h^c \mathbf{g}_0 \le \mathbf{u}_h \le \pi_h^c \mathbf{g}_1\}$. It is clear that $\mathbf{K}$ may not contain $\mathbf{K}^h$ this time. We then apply the estimates in Remark 3.2. Let $\mathbf{u}^*$ be the solution of (EXP) and $\mathbf{u}_h^*$ be the finite element approximation of the problem. It follows from [13] that one can construct $P(\mathbf{u}_h^*) \in \mathbf{K}$ such that

$$\pi_h^c(P(\mathbf{u}_h^*)) = \pi_h^c(\mathbf{u}_h^*), \qquad \|\mathbf{u}_h^* - P(\mathbf{u}_h^*)\|_{0,\Omega} \le C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}(|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}}).$$

First, as in Example 4.1, let $\mathbf{v}_h = \pi_h^c \mathbf{u}^*$,

$$|(\mathbf{p}_h^* + \mathbf{u}_h^*, \pi_h^c \mathbf{u}^* - \mathbf{u}^*)_{\mathbf{U}}| \le C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}} |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_{\mathbf{U}}} \|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\tau_{\mathbf{U}}}.$$

Let $\mathbf{v} = P(\mathbf{u}_h^*)$ in the estimate (3.15) of Remark 3.2. Then, first,

$$|(\mathbf{p}_h^* + \mathbf{u}_h^*, \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}| = |((\mathbf{p}_h^* + \mathbf{u}_h^* - \pi_h^c(\mathbf{p}_h^* + \mathbf{u}_h^*)), \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}|$$

$$\le C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_{\mathbf{U}}} (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}}),$$

since $\pi_h^c(P(\mathbf{u}_h^*)) = \pi_h^c(\mathbf{u}_h^*)$. Therefore

$$|(\mathbf{p}_h^* + \mathbf{u}_h^*, \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}| \le C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_{\mathbf{U}}}^2 + C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}})^2.$$

Second,

$$|((\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)), \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}| \le C \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}} \|(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))\|_{0,\tau_{\mathbf{U}}} (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}})$$
$$\le C \|(\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*))\|_{0,\Omega}^2 + C \sum_{\tau_U} h_{\tau_{\mathbf{U}}}^2 (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}})^2.$$

Third, from the proof of Theorem 3.1,

$$|((\mathbf{p}(\mathbf{u}_h^*) - \mathbf{p}^*), \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}| \le C \|\mathbf{p}(\mathbf{u}_h^*) - \mathbf{p}(\mathbf{u}^*)\|_{0,\Omega} \|\mathbf{u}_h^* - \mathbf{v}\|_{\mathbf{U}}$$

$$\le C \|\mathbf{u}_h^* - \mathbf{u}^*\|_{\mathbf{U}} \|\mathbf{u}_h^* - \mathbf{v}\|_{\mathbf{U}} \le \epsilon \|\mathbf{u}_h^* - \mathbf{u}^*\|_{\mathbf{U}}^2 + C \|\mathbf{u}_h^* - \mathbf{v}\|_{\mathbf{U}}^2,$$

where ϵ is positive but can be made very small. Finally,

$$|(\mathbf{u}_h^* - \mathbf{u}^*, \mathbf{u}_h^* - \mathbf{v})_{\mathbf{U}}| \le \epsilon \|\mathbf{u}_h^* - \mathbf{u}^*\|_{\mathbf{U}}^2 + C \|\mathbf{u}_h^* - \mathbf{v}\|_{\mathbf{U}}^2.$$

Therefore, we have

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}^2$$
$$\le C \bigg(\sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_{\mathbf{U}}}^2 + \|\mathbf{p}_h^* - \mathbf{p}(\mathbf{u}_h^*)\|_{0,\Omega}^2 + \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}})^2 \bigg).$$

Then it follows from the proof of Theorem 3.2; we finally have

$$\|\mathbf{u}^* - \mathbf{u}_h^*\|_{0,\Omega_{\mathbf{U}}}^2 + \|\mathbf{y}^* - \mathbf{y}_h^*\|_{0,\Omega}^2 + \|\mathbf{p}^* - \mathbf{p}_h^*\|_{0,\Omega}^2 \le C\eta_3^2,$$

with

$$\eta_3^2 = \eta_2^2 + \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 (|\mathbf{g}_0|_{1,\tau_{\mathbf{U}}} + |\mathbf{g}_1|_{1,\tau_{\mathbf{U}}})^2,$$
$$\eta_2^2 = \sum_{\tau_{\mathbf{U}}} h_{\tau_{\mathbf{U}}}^2 |\mathbf{p}_h^* + \mathbf{u}_h^*|_{1,\tau_{\mathbf{U}}}^2 + \eta_1^2,$$

and

$$\eta_1^2 = \sum_{\tau} h_\tau^4 \int_\tau (\Delta \mathbf{p}_h^* + \nabla s_h^* + \mathbf{y}_h^* - \mathbf{z}_d)^2 + \sum_l h_l^3 \int_l [\mathbf{A}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{p}_h^*)^2$$
$$+ \sum_{\tau} h_\tau^4 \int_\tau (\Delta \mathbf{y}_h^* - \nabla r_h^* + \mathbf{f} + \mathbf{u}_h^*)^2 + \sum_l h_l^3 \int_l [\mathbf{D}_l]^2 + \sum_\tau h_\tau^2 \int_\tau (\mathrm{div}\mathbf{y}_h^*)^2.$$

REFERENCES

[1] F. Abergel and T. Teman, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990), pp. 305–325.

[2] P. Alotto, P. Girdinio, O. Horigami, S. Ito, K. Iwanaga, K. Kato, T. Kuriyama, and H. Maed, *Mesh adaption and optimisation techniques in magnet design*, IEEE Transactions on Magnetics, 32 (1996).

[3] D.N. Arnold, F. Brezzi, and M. Fortin, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[4] N.V. Banichuk, F.J. Barthold, A. Falk, and E. Stein, *Mesh refinement for shape optimisation*, Structural Optimisation, 9 (1995), pp. 45–51.

[5] R.E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[6] J. Baranger and H.E. Amri, *A posteriori error estimators in finite element approximation of quasi-Newtonian flows*, Math. Model. Numer. Anal., 25 (1991), pp. 31–48.

[7] I. Babŭska and W.C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[8] R. Becker and H. Happ, *Optimization in PDE Models with Adaptive Finite Element Discretization*, Report 98-20 (SFB 359), University of Heidelberg, Heidelberg, Germany, 1998.

[9] R. Becker, H. Kapp, and R. Rannacher, *Adaptive Finite Element Methods for Optimal Control of Partial Differential Equations: Basic Concept*, SFB 359, University of Heidelberg, Heidelberg, Germany, 1998.

[10] C. Bernardi and G. Raugel, *Analysis of some finite elements for the Stokes problem*, Math. Comp., 44 (1985), pp. 71–79.

[11] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] Z. Ding, L. Ji, and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations*, SIAM J. Control Optim., 34 (1996), pp. 264–294.

[13] F.S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[14] A. Fursikov, *Control problems and theorems concerning the unique solvability of a mixed boundary value problem for three-dimensional Navier-Stokes and Euler equations*, Mat. Sb. (N.S.), 115 (1981), pp. 281–306, 320.

[15] M. Fortin, *Old and new finite elements for incompressible flows*, Internat. J. Numer. Methods Fluids, 1 (1981), pp. 347–364.

[16] M.B. Giles, M.G. Larson, J.M. Levenstam, and E. Suli, *Adaptive Error Control for Finite Element Approximation of the Lift and Drag Coefficients in Viscous Flow*, Report NA-97/06, Oxford University Computing Laboratory, Oxford, UK, 1997.

[17] M.B. Giles and N.A. Pierce, *An introduction to the adjoint approach to design*, Flow Turbul. Combust., 65 (2000), pp. 393–415.

[18] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer. Ser. Comput. Math. 5, Springer–Verlag, Berlin, Heidelberg, New York, 1986.

[19] M.D. Gunzburger, ed., *Flow Control*, Springer–Verlag, New York, 1995.
[20] M.D. Gunzburger and L.S. Hou, *Finite-dimensional approximation of a class of constrained nonlinear optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1001–1043.
[21] M.D. Gunzburger, L. Hou, and Th. Svobodny, *Analysis and finite element approximation of optimal control problems for stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.
[22] M.D. Gunzburger, L. Hou, and T.P. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.
[23] M.D. Gunzburger, L. Hou, and Th. Svobodny, *Optimal control and optimization of viscous, incompressible flows*, in Incompressible Computational Fluid Dynamics, M.D. Gunzburger and R.A. Nicolaides, eds., Cambridge University Press, Cambridge, UK, 1993, pp. 109–150.
[24] J. Haslinger and P. Neittaanmaki, *Finite Element Approximation for Optimal Shape Design*, John Wiley, Chichester, UK, 1989.
[25] L. Hou and J.C. Turner, *Analysis and finite element approximation of an optimal control problem in electrochemistry with current density controls*, Numer. Math., 71 (1995), pp. 289–315.
[26] A. Jameson, *Aerodynamic design via control theory*, J. Sci. Comput., 3 (1988), pp. 233–260.
[27] A. Jameson, L. Martinelli, and N.A. Pierce, *Optimum aerodynamic design using the Navier-Stokes equations*, Theoret. Comput. Fluid Dynamics, 10 (1998), pp. 213–237.
[28] I. Kevrekidis and S. Shvartsman, *Nonlinear model reduction for control of distributed parameter systems*, AIChE J., 44 (1998), pp. 1579.
[29] A. Kufner, O. John, and S. Fucik, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.
[30] I. Lasiecka, *Ritz–Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions*, SIAM J. Control Optim., 22 (1984), pp. 477–500.
[31] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer–Verlag, Berlin, 1971.
[32] J.L. Lions, *Control of Distributed Singular Systems*, Bordas, Paris, 1985.
[33] W.B. Liu and D. Tiba, *Error estimates for the finite element approximation of a class of nonlinear optimal control problems*, Numer. Funct. Anal. Optim., 22 (2001), pp. 953–972.
[34] W.B. Liu and N. Yan, *A posteriori error estimates for a model boundary optimal control problem*, J. Comput. Appl. Math., 120 (2000), pp. 159–173.
[35] W.B. Liu and N. Yan, *A posteriori error estimates for a nonlinear control problem*, in EUNMA'99 Proceedings, World Scientific, Singapore, 2001, pp. 146–152.
[36] W. Liu and N. Yan, *A posteriori error estimates for convex boundary control problems*, SIAM J. Numer. Anal., 39 (2001), pp. 73–99.
[37] W. Liu and N. Yan, *Quasi-norm local error estimates for p-Laplacian*, SIAM J. Numer. Anal., 39 (2001), pp. 100–127.
[38] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control constrained, optimal control systems*, Appl. Math. Optim., 8 (1982), pp. 69–95.
[39] K. Maute, S. Schwarz, and E. Ramm, *Adaptive topology optimization of elastoplastic structures*, Structural Optimization, 15 (1998), pp. 81–91.
[40] Moffatt, *Magnetic Field Generation in Electrically Conducting Fluids*, Cambridge University Press, New York, 1978.
[41] B. Mohammadi, J.I. Molho, and J.G. Santiago, *Design of minimal dispersion fluidic channels in a CAD-free framework*, in Proceedings of the Summer Program 2000, Center for Turbulence Research, 2000, pp. 49–62.
[42] B. Mohammadi and O. Pironneau, *Applied Shape Optimal Design*, Oxford University Press, Oxford, UK, 2001.
[43] P. Neittaanmaki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms and Applications*, Marcel Dekker, New York, 1994.
[44] N.A. Pierce and M.B. Giles, *Adjoint recovery of superconvergent functionals from PDE approximations*, SIAM Rev., 42 (2000), pp. 247–264.
[45] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.
[46] L.R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[47] A. Schleupen, K. Maute, and E. Ramm, *Adaptive FE-procedures in shape optimization*, Structural and Multidisciplinary Optimization, 19 (2000), pp. 282–302.
[48] D. Tiba and F. Troltzsch, *Error estimates for the discretization of state constrained convex control problems*, Numer. Funct. Anal. Optim., 17 (1996), pp. 1005–1028.

[49] R. Verfürth, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.
[50] R. Verfürth, *A posteriori error estimators and adaptive mesh-refinement techniques for the Navier-Stokes equations*, in Incompressible Computational Fluid Dynamics, M.D. Gunzburger and R.A. Nicolaides, eds., Cambridge University Press, Cambridge, UK, 1993, pp. 428–447.
[51] R. Verfürth, *A Review of Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley-Teubner, New York, 1996.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1870–1895

ON QUASI-NORM INTERPOLATION ERROR ESTIMATION AND A POSTERIORI ERROR ESTIMATES FOR p-LAPLACIAN*

WENBIN LIU† AND NINGNING YAN‡

Abstract. In this paper, we establish a series of interpolation error estimates for several widely used averaging interpolators in some quasi norms. These estimates are among the key ingredients in our improved a posteriori error analysis for the p-Laplacian. The quasi-norm interpolation error estimates are used to derive a posteriori error estimators for some conforming and nonconforming finite element approximation of the p-Laplacian, which are shown to provide upper and lower bounds for the discretization error.

Key words. finite element approximation, p-Laplacian, a posteriori error estimators, quasi norms, interpolation error estimates

AMS subject classifications. 65N30, 49J40

PII. S0036142901393589

1. Introduction.

In this work we derive a series of new interpolation error estimates for some well-known interpolators. These new estimates are among the key tools in our improved a posteriori error analysis for the finite element approximation of the p-Laplacian with Dirichlet data:

$$\begin{aligned} -\operatorname{div}(|\nabla u|^{p-2}\nabla u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where $1 < p < \infty$ and Ω is a bounded open subset of R^2 with a Lipschitz boundary $\partial\Omega$. This equation is viewed as one of the typical examples of a large class of nonlinear problems—degenerate nonlinear systems. Indeed it is believed that this equation contains most of the essential difficulties in studies of finite element approximations for this class of degenerate nonlinear systems, where many existing techniques (such as the linearization or deformation procedure) in the finite element method do not seem to work well.

Finite element approximations of the p-Laplacian have been extensively studied in the literature, and one can find some previous work, for example, in [10], [11], and [13], and overviews of some recent work in [3], [4], [5], [16], [17], [18], and [19], where among others, the quasi-norm approach is developed. This approach has proved quite successful in deriving sharp a priori error bounds for the finite element approximation of the degenerate systems. Some accounts of very recent work on the p-Laplacian can be found in the papers [18] and [19] as well. Another important area is a posteriori error estimation of the p-Laplacian. The work in this area seems to date back to [20], and some of the recent work can be found in [2], [4], [8], [21], [23], and [25], where among other things, a posteriori error estimates on the conforming and nonconforming

*Received by the editors August 14, 2001; accepted for publication (in revised form) June 3, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/sinum/40-5/39358.html
†CBS & Institute of Mathematics and Statistics, University of Kent, Canterbury, CT2 7NF, United Kingdom (w.b.liu@ukc.ac.uk).
‡Institute of System Sciences, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, China (yan@staff.iss.ac.cn). The research of this author was supported by EPSRC research grant GR/L67387.

discretization errors are derived with both upper and lower bounds. However, in all the cases there are *gaps* in the power between the existing upper and lower estimates.

Very recently in [18] and [19], we extended the quasi-norm techniques and developed some improved a posteriori error estimators of residual type for the conforming and nonconforming finite element approximation of the p-Laplacian. The structure of the new estimators seems very different from all the existing ones. The estimators are in fact shown to be equivalent up to higher order terms in the quasi norms. Initial analysis and numerical tests indicate that the new estimators are sharper than the existing ones, and indeed lead to more efficient computational meshes in some important cases. However, the error estimates in [18] and [19] are not fully a posteriori in the sense that some of the higher order terms in the error bounds contain a priori information of the solutions and therefore are not directly computable. This drawback is due to the lack of suitable interpolation error estimates in the quasi norms for the interpolators used in the papers. It is well known that suitable interpolation error estimation for averaging interpolators is one of the key ingredients in deriving full a posteriori error estimates of residual type for the finite element approximation of elliptic, and indeed more general, systems. Unfortunately such results in the quasi norms have not been available for the averaging interpolators used in the literature, and this leads to those a priori terms in the error bounds in [18] and [19]. It is far from straightforward to generalize the existing estimates on the interpolation error to the quasi norms, since some unique properties of a norm are essentially needed in all the existing proofs of the estimates.

In this work we address this very important issue. In particular, we establish a series of new interpolation error estimates for some of the most widely used interpolators in the literature. The results obtained are of their own importance. It is found that there sometimes exist essential differences between the quasi-norm estimates and the existing ones. We then utilized these results in deriving further improved new a posteriori error estimates for the finite element approximation of the p-Laplacian; see a further explanation at the end of section 2. The results and the methods developed in this paper are applicable to more general degenerate systems.

The plan of this paper is as follows: In section 2 we state the weak formulation of the p-Laplacian and its finite element approximation. Some important inequalities, the definition of the quasi norm, and related results are also presented in section 2. These inequalities are among the key ingredients of the quasi-norm a priori and a posteriori error analysis for the finite element approximation of the p-Laplacian. In section 3 we derive some quasi-norm interpolation error estimates, which can be viewed as the quasi-norm counterparts of some well-known results. In section 4 we use the interpolation error estimates obtained in section 3 to derive a posteriori error bounds for conforming and nonconforming finite element approximations of the p-Laplacian.

2. Preliminaries. Let Ω be a bounded open set in R^2 with a Lipschitz boundary $\partial\Omega$. In this paper we adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with norm $\|\cdot\|_{m,q,\Omega}$ and seminorm $|\cdot|_{m,q,\Omega}$. We set $W_0^{m,q}(\Omega) \equiv \{w \in W^{m,q}(\Omega) : w|_{\partial\Omega} = 0\}$. We denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ with norm $\|\cdot\|_{m,\Omega}$ and seminorm $|\cdot|_{m,\Omega}$.

Consider the p-Laplacian with zero Dirichlet data:

$$\begin{aligned} -\mathrm{div}(|\nabla u|^{p-2}\nabla u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $1 < p < \infty$ and $f \in L^2(\Omega)$. The weak formulation of (2.1) is as follows (WP):

Find $u \in W_0^{1,p}(\Omega)$ such that

$$a(u,v) = (f,v) \quad \forall v \in W_0^{1,p}(\Omega), \tag{2.2}$$

where

$$a(u,v) = \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v$$

and

$$(f,v) = \int_\Omega fv.$$

It is a simple matter to show that there exists a unique solution to (WP); see [11]. We note that for sufficiently regular data, global $C^{1,\alpha}$ regularity is established in [15].

Let Ω^h be a polygonal approximation to Ω with the boundary $\partial\Omega^h$. Let T^h be a partitioning of Ω^h into disjoint open regular simplices K, so that $\bar{\Omega}^h = \bigcup_{K\in T^h} \bar{K}$. $\bar{K}$ and $\bar{K}'$ have either an empty or only one common vertex, or a whole edge if K and $K' \in T^h$. We further require that $x_i \in \partial\Omega^h \Rightarrow x_i \in \partial\Omega$, where $\{x_i\}$ $(i = 1 \dots J)$ is the vertex set associated with the partitioning T^h. Let h_K denote the maximum diameter of the element K in T^h and let ρ_K denote the diameter of the largest ball contained in K. We assume that there is a regularity constant R of T^h such that $1 \le \max_{K\in T^h}(h_K/\rho_K) \le R$. For ease of exposition we will assume that $\Omega^h = \Omega$.

Due to limited regularity for the solution of the p-Laplacian, we shall only discuss the conforming piecewise linear elements and a simple nonconforming element—Crouzeix–Raviart—in this paper.

Conforming element. Associated with T^h is a finite dimensional subspace V^h of $C^0(\bar{\Omega}^h)$ such that $\chi|_K \in P_1$ for all $\chi \in V^h$ and $K \in T^h$, where P_1 is the space of polynomials of first degree. Let

$$V_0^h = \{\chi \in V^h : \chi(x_i) = 0 \ \ \forall x_i \in \partial\Omega^h\}.$$

Then the finite element approximation of (WP) is as follows $(WP)^h$: Find $u_h \in V_0^h$ such that

$$a(u_h, v_h) = (f, v_h) \qquad \forall v_h \in V_0^h, \tag{2.3}$$

where

$$a(u_h, v_h) = \int_{\Omega^h} |\nabla u_h|^{p-2}\nabla u_h \cdot \nabla v_h,$$

$$(f, v_h) = \int_{\Omega^h} f v_h.$$

It is a simple matter to show that $(WP)^h$ has a unique solution u_h; see [11] again. Analysis of the finite element approximation of (WP) and a priori error bounds for this approximation were first discussed in [11] and [13]. Some of the best results are obtained in [3].

Nonconforming element. Associated with T^h is the Crouzeix–Raviart-type finite dimensional subspace $\tilde{V}^h$ of $L^2(\Omega^h)$:

$$\tilde{V}^h = \{v \in L^2(\Omega^h) : v|_K \in P_1 \ \forall K \in T^h, v \text{ is continuous on the midpoints of edges}\}.$$

Let

$$\tilde{V}_0^h = \{v \in \tilde{V}^h : v = 0 \text{ on the midpoints of edges on } \partial\Omega\}.$$

Note that $\tilde{V}_0^h \bar{\in} W_0^{1,p}(\Omega)$. The finite element approximation of (WP) is as follows $(WP)_n^h$: Find $u_h \in \tilde{V}_0^h$ such that

$$a_h(u_h, v_h) = (f, v_h) \qquad \forall v_h \in \tilde{V}_0^h, \tag{2.4}$$

where

$$a_h(u_h, v_h) = \sum_K \int_K |\nabla u_h|^{p-2} \nabla u_h \cdot \nabla v_h,$$

$$(f, v_h) = \int_{\Omega^h} f v_h.$$

It also is a simple matter to show that $(WP)_n^h$ has a unique solution u_h. Analysis of the finite element approximation of (WP) and a priori error bounds for this approximation are discussed in [19]. In particular, the optimal a priori error bounds have been established there.

In the following we state some inequalities which play an essential role in our error analysis. In all these lemmas, C is a positive constant which depends only on p. The first two lemmas have been used in our work on a priori quasi-norm error bounds for the finite element approximation of degenerate nonlinear PDEs; see, e.g., [3], [16]. The proofs of Lemma 2.1 and 2.2 can be found in [5] and [18].

LEMMA 2.1. *For all $p > 1$, ξ, $\eta \in R^n$,*

$$| \, |\xi|^{p-2}\xi - |\eta|^{p-2}\eta \, | \le C|\xi - \eta|(|\xi| + |\eta|)^{p-2}, \tag{2.5}$$

$$(|\xi|^{p-2}\xi - |\eta|^{p-2}\eta, \xi - \eta) \ge C|\xi - \eta|^2(|\xi| + |\eta|)^{p-2}. \tag{2.6}$$

LEMMA 2.2. *For all a, σ_1, $\sigma_2 \ge 0$, $p > 1$, $\theta > 0$,*

$$(a + \sigma_1)^{p-2}\sigma_1\sigma_2 \le \theta^{-\gamma}(a + \sigma_1)^{p-2}\sigma_1^2 + \theta(a + \sigma_2)^{p-2}\sigma_2^2,$$

where

$$\gamma = \begin{cases} 1, & 1 < p \le 2, \ \theta \in [1, \infty) \quad \textit{or} \quad 2 < p < \infty, \ \theta \in (0, 1), \\ \dfrac{1}{p-1}, & 1 < p \le 2, \ \theta \in (0, 1) \quad \textit{or} \quad 2 < p < \infty, \ \theta \in [1, \infty). \end{cases}$$

The following lemma is essential for estimating a bilinear form via the quasi norms. It can be viewed as a generalization of the Young inequality where $a = 0$. A proof can be found in [18].

LEMMA 2.3. *For all $a, \sigma_1, \sigma_2 \ge 0$, $p > 1$, and $\delta > 0$,*

$$\sigma_1\sigma_2 \le \delta^{-\beta}(a^{p-1} + \sigma_1)^{p'-2}\sigma_1^2 + \delta(a + \sigma_2)^{p-2}\sigma_2^2,$$

where β is such that $\delta^{-\beta} = \max\{\delta^{-1}, \delta^{-\frac{1}{p-1}}\}$, and p' is such that $\frac{1}{p} + \frac{1}{p'} = 1$.

The following two lemmas will be used to prove some triangle inequality-like results for a power of the quasi norms. They have been proved in [18].

LEMMA 2.4. *For all* $1 < p < \infty$, $\sigma_1, \sigma_2 \in R^n$, *and* $a \geq 0$,

$$(a + |\sigma_1 + \sigma_2|)^{p-2}|\sigma_1 + \sigma_2|^2 \leq C(a + |\sigma_1|)^{p-2}|\sigma_1|^2 + C(a + |\sigma_2|)^{p-2}|\sigma_2|^2.$$

LEMMA 2.5. *For all* $1 < p < \infty$ *and* $\sigma, \sigma_1, \sigma_2 \in R^n$,

$$(|\sigma_1| + |\sigma_2|)^{p-2}|\sigma_1 - \sigma_2|^2 \leq C(|\sigma| + |\sigma - \sigma_1|)^{p-2}|\sigma - \sigma_1|^2 + C(|\sigma| + |\sigma - \sigma_2|)^{p-2}|\sigma - \sigma_2|^2.$$

One of the key ideas in our approach is to introduce some quasi norms to handle the possible degeneracy of the p-Laplacian in order to obtain sharp error bounds. Let us try to explain the main motivation of the method. A norm may be considered to be a "straight ruler" or measure in a linear space. It can handle the smooth nonlinear problems that can be linearized. However, a standard Sobolev norm may not be able to give sharp estimation for the errors of the finite element approximation of some nonlinear equations where, for instance, the locations of degeneracy points of the equations are solution dependent, since it has to assume the worst scenario that there exists the degeneracy everywhere. The type of the equation studied here, which may be degenerate and highly nonlinear, falls into this category. Naturally some "curved" solution dependent rulers or distances may have to be introduced to handle the degeneracy *adaptively* in order to present more accurate error analysis for the finite element approximation of the equation. Furthermore, these "rulers" should be consistent with the special structure of the equation studied, in particular, both monotone and continuous, see (2.11) and (2.12), and thus equivalent to the energy difference. In the case of the p-Laplacian, these rulers (or some of their powers) happen to be quasi norms. This idea has been widely used in our work; see [16], [18], and [19]. As seen later, this idea is also essential in deriving improved a posteriori error estimates for the p-Laplacian. In the following we very briefly introduce such a quasi-norm and some relations between it and the standard Sobolev norms. Let $W^{1,p}(\Omega, T^h) \equiv \{v \in L^p(\Omega) : v|_K \in W^{1,p}(K)\}\ \forall K \in T^h\}$. Let

$$\|v\|_{1,p,\Omega,T^h} = \left(\sum_{K \in T^h} \|v\|^p_{1,p,K}\right)^{1/p}$$

and

$$|v|_{1,p,\Omega,T^h} = \left(\sum_{K \in T^h} |v|^p_{1,p,K}\right)^{1/p}.$$

Let $w \in W^{1,p}(\Omega, T^h)$. Similarly as in [19], we define for any $v \in W^{1,p}(\Omega, T^h)$

$$|v|^2_{(w,p)} \equiv \sum_{K \in T^h} \int_K |\nabla v|^2 (|\nabla w| + |\nabla v|)^{(p-2)}. \tag{2.7}$$

Remark 2.1. It is easy to see that when $v \in W^{1,p}(\Omega)$, $\|v\|_{1,p,\Omega,T^h} = \|v\|_{1,p,\Omega}$, $|v|_{1,p,\Omega,T^h} = |v|_{1,p,\Omega}$, and

$$\sum_{K \in T^h} \int_K (|\nabla w| + |\nabla v|)^{p-2}|\nabla v|^2 = \int_{\Omega^h} (|\nabla w| + |\nabla v|)^{p-2}|\nabla v|^2$$

if $w, v \in W^{1,p}(\Omega)$. The quasi norm on the right side has been introduced in [3] and [18] to study $(WP)^h$.

We have the following proposition.

PROPOSITION 2.6. (i) $|v|_{(w,p)} \geq 0$, *and when* $v \in W_0^{1,p}(\Omega)$, $|v|_{(w,p)} = 0$ *if and only if* $v = 0$. (ii) $|v_1 + v_2|_{(w,p)} \leq C(|v_1|_{(w,p)} + |v_2|_{(w,p)})$ *for any* $v_1, v_2 \in W^{1,p}(\Omega, T^h)$, *where* C *depends only on* p.

Furthermore for $1 < p \leq 2$, *we have*

$$|v|_{1,p,\Omega,T^h} \leq C(1 + |w|_{1,p,\Omega,T^h} + |v|_{1,p,\Omega,T^h})^{(2-p)/2}|v|_{(w,p)}, \tag{2.8}$$

$$|v|^2_{(w,p)} \leq \sum_{K \in T^h} |v|^p_{1,p,K} = |v|^p_{1,p,\Omega,T^h}, \tag{2.9}$$

and for $2 \leq p < \infty$,

$$|v|^p_{1,p,\Omega,T^h} \leq \|v\|^2_{(w,p)} \leq C(1 + |w|_{1,r,\Omega,T^h} + |v|_{1,r,\Omega,T^h})^{(p-2)/2}|v|^2_{1,s,\Omega,T^h}, \tag{2.10}$$

where $s \in [2, p]$, $r = s(2-p)/(2-s)$.

The conclusion (ii) can be easily proved by Lemma 2.4. The rest of the proposition can be shown similarly as in [5].

Although in a priori error analysis we almost always take w to be u, the solution of (WP), it is sometimes desirable to replace it with u_h (see [18]) in a posteriori error analysis, where u_h is the finite element approximation of u, in order to make the quasi-norm "computable." It follows from the triangle inequality that there are constants $c, C > 0$, independent of h such that

$$c|u - u_h|_{(u,p)} \leq |u - u_h|_{(u_h,p)} \leq C|u - u_h|_{(u,p)}.$$

We shall simply write $|\cdot|_{(u,p)}$ as $|\cdot|_{(p)}$ when no confusion is likely caused.

The essential relations between the quasi norm and the equation are reflected in the following inequalities, which follow from Lemmas 2.1 and 2.2 (see [4] for a proof). Let u be the solution of (WP). For any $v \in W^{1,p}(\Omega)$,

$$a(u, u - v) - a(v, u - v) \geq C|u - v|^2_{(u,p)}, \tag{2.11}$$

and for any $\theta > 0$,

$$|a(u, w) - a(v, w)| \leq C\theta^{-\gamma}|u - v|^2_{(u,p)} + C\theta|w|^2_{(u,p)} \qquad \forall v, w \in W^{1,p}(\Omega), \tag{2.12}$$

where the constant $\gamma > 0$ is defined in Lemma 2.2. It follows from (2.11) and (2.12) that there are constants $c, C > 0$ such that for any $u, v \in W^{1,p}(\Omega)$

$$c(a(u, v - u) - a(v, u - v)) \leq |u - v|^2_{(u,p)} \leq C(a(u, u - v) - a(v, u - v)).$$

Thus the quasi norm is naturally related to the total energy difference $a(u, u - v) - a(v, u - v)$.

The relations (2.11) and (2.12) are important to prove the optimal a priori error bound in the quasi norm. For instance, for the conforming piecewise linear finite element solution u_h of the p-Laplacian, we have (see [3] and [16])

$$|u - u_h|^2_{(p)} \leq C \min_{v_h \in V_0^h} |u - v_h|^2_{(p)} \leq Ch^2$$

provided u is smooth enough. Thus when $1 < p \le 2$, for instance, one has the optimal a priori error bound in the $W^{1,p}$ norm:

$$\|u - u_h\|_{1,p,\Omega} \le C|u - u_h|_{(p)} \le Ch,$$

when u is smooth enough.

In [18] and [19], the quasi norm has been used to obtain improved a posteriori error estimators for the p-Laplacian. For instance, let u_h be the piecewise linear conforming finite element approximation of the p-Laplacian. Then

$$c(\eta_1^2 + \eta_2^2) + c\epsilon_1 \le |u - u_h|_{(p)}^2 \le C(\eta_1^2 + \eta_2^2) + C\epsilon_2,$$

where η_1, η_2 are defined in section 4, and ϵ_1, ϵ_2 are some higher order terms for smooth f and the smooth solutions, though ϵ_2 is not directly computable. The readers are referred to [18] and [19] for the details. However, it is not clear if ϵ_2 is always negligible. Thus the above estimators may not be fully reliable. This is an annoying theoretical problem, as logically one wishes to be assured that one's estimators are always reliable and at least efficient for the smooth data. In this paper, we shall derive new a posteriori error estimates, which are fully reliable. We will discuss and interpret these estimates in Remark 4.1 and Remark 4.3.

In the next section, we shall derive some important interpolation error estimates in the quasi norm for some well-known averaging interpolators, which are among the keys to our improved estimates.

3. Interpolation error estimates. In this section, we shall establish some interpolation error estimates in the quasi norm, which are essential in our a posteriori error analysis for the p-Laplacian. Some of the relevant ideas can be found in [9], [18], and [19]. First, let us prove a lemma which is a quasi-norm version of the inequalities of the Poincare type. The following proof was adopted from [9].

LEMMA 3.1. *Let Ω be a nonempty bounded connected open set in R^2. Let $1 < p < \infty$ and $f \in (W^{1,p}(\Omega))^*$ with $R \cap \mathrm{Ker}(f) = \{0\}$. Then there exists a constant $C = C(f, p, \Omega)$ such that, for all $a \in R$, $a \ge 0$, and $v \in W^{1,p}(\Omega)$,*

$$\int_\Omega (a + |v|)^{p-2}|v|^2 dx \le C(a + |f(v)|)^{p-2}|f(v)|^2 + C\int_\Omega (a + |\nabla v|)^{p-2}|\nabla v|^2\, dx.$$

Proof. We first introduce the following notation to simplify the proof. For $x, y \ge 0$, $1 < p < \infty$, let

$$G(x, y) := \begin{cases} y^2(x + y)^{p-2} & \text{if } x + y > 0, \\ 0 & \text{if } x = y = 0. \end{cases} \tag{3.1}$$

Then the lemma states that there exists a constant $C = C(f, p, \Omega)$ such that, for all $a \ge 0$ and for all $v \in W^{1,p}(\Omega)$,

$$\int_\Omega G(a, |v|)\, dx \le C\, G(a, |f(v)|) + C\int_\Omega G(a, |\nabla v|)\, dx. \tag{3.2}$$

Note that $G(x, y)$ is monotone increasing and convex with respect to the variable y.

We argue by contradiction and suppose that (3.2) is false. Then there would exist a sequence v_j in $W^{1,p}(\Omega)$ with $\delta_j := \|v_j\|_{1,q,\Omega} > 0$, $q = \min\{2, p\}$, and a sequence a_j of nonnegative real numbers such that

$$G(a_j, |f(v_j)|) + \int_\Omega G(a_j, |\nabla v_j|)\, dx \le \frac{1}{j}\int_\Omega G(a_j, |v_j|)\, dx \tag{3.3}$$

for all $j \in \mathcal{N}$. We observe in any case that there exists a $u \in W^{1,q}(\Omega)$ with

$$u_j := \frac{v_j}{\delta_j} \text{ satisfies } \|u_j\|_{1,q,\Omega} = 1, \qquad u_j \rightharpoonup u \text{ in } W^{1,q}(\Omega). \tag{3.4}$$

Here we have chosen a weak convergent subsequence with the Banach–Alaoglu theorem. In the first case we suppose that there exists a constant γ, $0 < \gamma < \infty$, with

$$a_j \le \gamma\, \delta_j \quad \text{for } j = 1, 2, 3, \ldots. \tag{3.5}$$

At least we suppose (3.5) for a subsequence we have not relabelled. If $1 < p \le 2$, then $G(a, x) \le x^p$ for all $x \ge 0$. Therefore,

$$\int_\Omega G\left(\frac{a_j}{\delta_j}, |u_j|\right) dx \le \|u_j\|^p_{0,p,\Omega} \le 1$$

even without (3.5). If $2 \le p$, then $G(\cdot, |u_j|)$ is monotone increasing. Hence, (3.4)–(3.5) yield

$$\begin{aligned}\int_\Omega G\left(\frac{a_j}{\delta_j}, |u_j|\right) dx &\le \int_\Omega (\gamma + |u_j|)^{p-2} |u_j|^2\, dx \\ &\le \|\gamma + |u_j|\|^p_{0,p,\Omega} \le (1 + \gamma |\Omega|^{\frac{1}{p}})^p.\end{aligned}$$

Hence, for all $1 < p < \infty$, $\int_\Omega G(a_j/\delta_j, |u_j|)\, dx$ is bounded. A scaling of (3.3) then shows

$$\lim_{j\to\infty} \int_\Omega G\left(\frac{a_j}{\delta_j}, |\nabla u_j|\right) dx = \lim_{j\to\infty} G\left(\frac{a_j}{\delta_j}, |f(u_j)|\right) = 0. \tag{3.6}$$

If $1 < p \le 2$, a Hölder inequality with exponents $2/p$ and $2/(2-p)$ leads to

$$\begin{aligned}\|\nabla u_j\|^p_{0,p,\Omega} &= \int_\Omega |\nabla u_j|^p \left(\frac{a_j}{\delta_j} + |\nabla u_j|\right)^{\frac{p(p-2)}{2}} \left(\frac{a_j}{\delta_j} + |\nabla u_j|\right)^{\frac{p(2-p)}{2}} dx \\ &\le \left(\int_\Omega G\left(\frac{a_j}{\delta_j}, |\nabla u_j|\right) dx\right)^{\frac{p}{2}} \left(\int_\Omega \left(\frac{a_j}{\delta_j} + |\nabla u_j|\right)^p dx\right)^{1-\frac{p}{2}}. \end{aligned} \tag{3.7}$$

The last factor is bounded as $j \to \infty$ by (3.4)–(3.5) and the second last tends to zero by (3.6). Again, for $1 < p \le 2$ (when $G(\cdot, |f(u_j)|)$ is monotone decreasing), (3.6) shows that $G(\gamma, |f(u_j)|)$ tends to zero and, hence, so does $|f(u_j)|$. Consequently,

$$\lim_{j\to\infty} \|\nabla u_j\|_{0,q,\Omega} = \lim_{j\to\infty} |f(u_j)| = 0. \tag{3.8}$$

So far we established (3.8) for $1 < p \le 2$. For $2 < p < \infty$, $|\nabla u_j|^p \le G(a_j/\delta_j, |\nabla u_j|)$ and $|f(u_j)|^p \le G(a_j/\delta_j, |f(u_j)|)$, and so (3.6) implies (3.8) directly. From (3.8) we deduce a contradiction to (3.4): Since $W^{1,q}(\Omega)$ is compactly embedded in $L^q(\Omega)$, we have $u_j \to u$ in $L^q(\Omega)$. With (3.8), $u_j \to u$ in $W^{1,q}(\Omega)$ and so $\|u\|_{1,q,\Omega} = 1$. Conversely, u is constant (as $\nabla u_j \to 0$ in $L^q(\Omega)$). Since f is a bounded linear form, $f(u_j) \to f(u)$ and $f(u) = 0$. Since $u \in R \cap \mathrm{Ker} f$, we have $u = 0$. This contradiction with $\|u\|_{1,q,\Omega} = 1$ concludes (3.2) in case (3.5).

In the remaining second case we suppose that a_j/δ_j is not bounded (not even for a subsequence). Hence, $\lim_{j\to\infty}(a_j/\delta_j) = +\infty$. One can assume that

$$\delta_j \le \gamma\, a_j \text{ for } q = \min\{2, p\} \text{ and for } j = 1, 2, 3, \ldots \tag{3.9}$$

for a constant γ (and at least for sufficiently large j which we have not relabelled). If $1 < p \le 2$, we use $(1 + (\delta_j/a_j)|u_j|)^{p-2} \le 1$. If $2 \le p < \infty$, we use $\delta_j/a_j \le \gamma$. This leads to

$$\begin{aligned}(3.10)\qquad &\frac{1}{j}\int_\Omega \left(1+\frac{\delta_j}{a_j}|u_j|\right)^{p-2}|u_j|^2\,dx \\ &\le \begin{cases} \dfrac{1}{j}\|u_j\|^2_{0,2,\Omega} & \text{if } 1<p\le 2,\\[2ex] \dfrac{1}{j}\,\gamma^{p-2}\left(\dfrac{1}{\gamma}\,|\Omega|^{1/p}+\|u_j\|_{0,p,\Omega}\right)^p & \text{if } 2\le p<\infty. \end{cases}\end{aligned}$$

Since $q = \min\{2, p\}$ and $\|u_j\|_{1,q,\Omega} = 1$, we conclude that (3.10) tends to zero as $j \to \infty$ from embedding. A scaling of (3.3) therefore yields

$$\begin{aligned}(3.11)\qquad &\lim_{j\to\infty}\int_\Omega \left(1+\frac{\delta_j}{a_j}\,|\nabla u_j|\right)^{p-2}|\nabla u_j|^2\,dx \\ &= \lim_{j\to\infty}\left(1+\frac{\delta_j}{a_j}\,|f(u_j)|\right)^{p-2}|f(u_j)|^2 = 0.\end{aligned}$$

If $2 \le p < \infty$, we directly deduce (3.8) for $q = 2$ and finish the proof of (3.2) as in the first case since $\|u_j\|_{1,2,\Omega} = 1$. If $1 < p \le 2$, we argue with a Hölder inequality analogy to (3.7) and infer

$$\|\nabla u_j\|^2_{0,p,\Omega} \le \int_\Omega \left(1+\frac{\delta_j}{a_j}\,|\nabla u_j|\right)^{p-2}|\nabla u_j|^2\,dx\,\left(\int_\Omega\left(1+\frac{\delta_j}{a_j}\,|\nabla u_j|\right)^p dx\right)^{\frac{2-p}{p}}.$$

The last factor is bounded according to (3.9) and $\|u_j\|_{1,p,\Omega} = 1$. This and (3.11) show (3.8) with $p = q \le 2$. (3.2) then follows as in the first case. Hence, Lemma 3.1 follows.

COROLLARY 3.2. *Let Ω be a nonempty bounded connected open set in R^2. Then, for any $a \ge 0$, $v \in W^{1,p}(\Omega)$, and $p > 1$, there exists a constant C, which is only dependent on Ω and p, such that*

$$\inf_{q\in R}\int_\Omega (a+|v+q|)^{p-2}|v+q|^2 \le C\int_\Omega (a+|\nabla v|)^{p-2}|\nabla v|^2.$$

Proof. It follows from (3.2) that for given $p > 1$ and $f \in (W^{1,p}(\Omega))^*$ with $R \cap \mathrm{Ker}(f) = \{0\}$ there exists a constant $C = C(f, p, \Omega)$ such that for all $v \in W^{1,p}(\Omega)$, $q \in R$,

$$\begin{aligned}&\int_\Omega (a+|v+q|)^{p-2}|v+q|^2 + \int_\Omega (a+|\nabla(v+q)|)^{p-2}|\nabla(v+q)|^2 \\ &\le C\left(\int_\Omega (a+|\nabla(v+q)|)^{p-2}|\nabla(v+q)|^2 + (a+|f(v+q)|)^{p-2}|f(v+q)|^2\right).\end{aligned}$$

Let us take a fixed linear functional f (say, $f(v) = \int_\Omega v$). Note that there exists a $q^*(v) \in R$ such that $f(v + q^*) = 0$. Then, we have

$$\inf_{q\in R}\left\{\int_\Omega (a+|v+q|)^{p-2}|v+q|^2 + \int_\Omega (a+|\nabla(v+q)|)^{p-2}|\nabla(v+q)|^2\right\}$$

$$\le \int_\Omega (a+|v+q^*|)^{p-2}|v+q^*|^2 + \int_\Omega (a+|\nabla(v+q^*)|)^{p-2}|\nabla(v+q^*)|^2$$
$$\le C\bigg(\int_\Omega (a+|\nabla(v+q^*)|)^{p-2}|\nabla(v+q^*)|^2 + (a+|f(v+q^*)|)^{p-2}|f(v+q^*)|^2\bigg)$$
$$= C\int_\Omega (a+|\nabla(v+q^*)|)^{p-2}|\nabla(v+q^*)|^2 = C\int_\Omega (a+|\nabla v|)^{p-2}|\nabla v|^2.$$

Hence, Corollary 3.2 follows.

Thanks to the above lemma and its corollary, now we can prove new quasi-norm interpolation error estimates for some finite element spaces.

First, let us consider the nonconforming finite element—Crouzeix–Raviart element—space $\tilde{V}^h$. An interpolation operator $\tilde{\pi}_h$ from $W^{1,1}(K)$ to $\tilde{V}^h$ is defined (see [12] and [24] for examples) such that for any $w \in W^{1,1}(K)$, $\tilde{\pi}_h w|_K \in P_1$ and

$$\int_{l_i} \tilde{\pi}_h w = \int_{l_i} w, \qquad i = 1,2,3,$$

where l_i are the edges of the element K. For the above interpolation, we have the following interpolation error estimate.

THEOREM 3.3. *Let $\tilde{\pi}_h$ be the interpolation operator defined as above. For all $K \in T^h$, $v \in W^{1,p}(K)$, and $p > 1$, one has*

$$\int_K (|\nabla u_h| + h_K^{-1}|v-\tilde{\pi}_h v|)^{p-2} h_K^{-2}|v-\tilde{\pi}_h v|^2 \le C\int_K (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2 \tag{3.12}$$

and

$$\int_K (|\nabla u_h| + |\nabla(v-\tilde{\pi}_h v)|)^{p-2}|\nabla(v-\tilde{\pi}_h v)|^2 \le C\int_K (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2\,. \tag{3.13}$$

Proof. Let $\hat{K}$ be a reference element of K; it is easy to see that

$$\int_K (|\nabla u_h| + h_K^{-1}|v-\tilde{\pi}_h v|)^{p-2} h_K^{-2}|v-\tilde{\pi}_h v|^2 \le C h_K^{2-p}\int_{\hat{K}} (|\nabla u_h| + |v-\tilde{\pi}_h v|)^{p-2}|v-\tilde{\pi}_h v|^2.$$

Note that for all $q \in R$, $\tilde{\pi}_h q = q$. It follows from Lemma 2.4 that

$$\begin{aligned}
&\int_K (|\nabla u_h| + h_K^{-1}|v-\tilde{\pi}_h v|)^{p-2} h_K^{-2}|v-\tilde{\pi}_h v|^2 \\
&\le C h_K^{2-p}\inf_{q\in R}\bigg\{\int_{\hat{K}} (|\nabla u_h| + |(I-\tilde{\pi}_h)(v+q)|)^{p-2}|(I-\tilde{\pi}_h)(v+q)|^2\bigg\} \\
&\le C h_K^{2-p}\inf_{q\in R}\bigg\{\int_{\hat{K}} (|\nabla u_h| + |v+q|)^{p-2}|v+q|^2 \\
&\qquad\qquad + \int_{\hat{K}} (|\nabla u_h| + |\tilde{\pi}_h(v+q)|)^{p-2}|\tilde{\pi}_h(v+q)|^2\bigg\}.
\end{aligned} \tag{3.14}$$

Note that by the definition of $\tilde{\pi}_h$ and a trace theorem,

$$\int_{\hat{K}} (|\nabla u_h| + |\tilde{\pi}_h(v+q)|)^{p-2}|\tilde{\pi}_h(v+q)|^2$$
$$\le \sum_{i=1}^3 C_i \int_{\hat{K}} \bigg(|\nabla u_h| + \int_{l_i}|v+q|\bigg)^{p-2}\bigg(\int_{l_i}|v+q|\bigg)^2$$

$$\le \sum_{i=1}^{3} C_i \int_{\hat{K}} \left(|\nabla u_h| + \int_{\hat{K}} |v+q| + \int_{\hat{K}} |\nabla(v+q)| \right)^{p-2} \left(\int_{\hat{K}} |v+q| + \int_{\hat{K}} |\nabla(v+q)| \right)^2$$
$$\le C \Bigg(\int_{\hat{K}} \left(|\nabla u_h| + \int_{\hat{K}} |v+q| \right)^{p-2} \left(\int_{\hat{K}} |v+q| \right)^2 + \int_{\hat{K}} \left(|\nabla u_h| + \int_{\hat{K}} |\nabla(v+q)| \right)^{p-2} \left(\int_{\hat{K}} |\nabla(v+q)| \right)^2 \Bigg).$$

Let $s(x) = (a+x)^{p-2}x^2$. Then it can be shown that $s(x)$ is increasing on $[0,\infty)$ for any $a \ge 0$ and is further a convex continuous function on $[0,\infty)$. Therefore it follows from the integral form of the Jensen inequality that

$$meas(\hat{K}) s \left(\int_{\hat{K}} r \right) \le \int_{\hat{K}} s(meas(\hat{K}) r). \tag{3.15}$$

Note that there exist constants c and C such that $c \le meas(\hat{K}) \le C$. Then applying the Jensen inequality, we have that

$$\int_{\hat{K}} (|\nabla u_h| + |\tilde{\pi}_h(v+q)|)^{p-2} |\tilde{\pi}_h(v+q)|^2$$
$$\le C \Bigg(\int_{\hat{K}} (|\nabla u_h| + |v+q|)^{p-2} |v+q|^2 + \int_{\hat{K}} (|\nabla u_h| + |\nabla(v+q)|)^{p-2} |\nabla(v+q)|^2 \Bigg).$$

Hence by (3.14) and Corollary 3.2, we have that

$$\int_K (|\nabla u_h| + h_K^{-1}|v - \tilde{\pi}_h v|)^{p-2} h_K^{-2} |v - \tilde{\pi}_h v|^2$$
$$\le C h_K^{2-p} \inf_{q \in R} \Bigg\{ \int_{\hat{K}} (|\nabla u_h| + |v+q|)^{p-2} |v+q|^2 + \int_{\hat{K}} (|\nabla u_h| + |\nabla(v+q)|)^{p-2} |\nabla(v+q)|^2 \Bigg\}$$
$$\le C h_K^{2-p} \int_{\hat{K}} (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2 \le C h_K^{2-p} h_K^{-2} h_K^p \int_K (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2$$
$$= C \int_K (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2 .$$

This proves (3.12). The estimate (3.13) can be proved similarly.

Now let us consider the conforming finite element space. Let π_h be the average interpolator defined in [22] from $H^1(\Omega)$ to V^h (or from $H_0^1(\Omega)$ to V_0^h); see [22] for details. For this average interpolator, we have the following quasi-norm interpolation error estimates.

THEOREM 3.4. *Let π_h be the average interpolation operator defined in* [22]. *Then, for all $p > 1$, constant $a \ge 0$, $K \in T^h$, and $v \in W^{1,p}(\Omega)$, one has*

$$\int_K (a + h_K^{-1}|v - \pi_h v|)^{p-2} h_K^{-2} |v - \pi_h v|^2 \le C \sum_{\bar{K}' \cap \bar{K} \ne \emptyset} \int_{K'} (a + |\nabla v|)^{p-2} |\nabla v|^2 \tag{3.16}$$

and

$$\int_K (a + |\nabla(v - \pi_h v)|)^{p-2} |\nabla(v - \pi_h v)|^2 \le C \sum_{\bar{K}' \cap \bar{K} \ne \emptyset} \int_{K'} (a + |\nabla v|)^{p-2} |\nabla v|^2 , \tag{3.17}$$

where the constant C depends only on Ω and p.

Proof. Let

$$S_i = \{\cup K' : K' \in T^h, \bar{K}' \cap x_i \neq \emptyset\},$$

where x_i, $i = 1, 2, 3$, are the vertices of the element K. Let $\hat{S}_i$ and $\hat{K}$ be reference elements of S_i and K, respectively, $q \in R$. By the definition of interpolation π_h and Lemma 2.4, we have that

$$\begin{aligned}
&\int_{\hat{K}} (a + |\pi_h(v+q)|)^{p-2} |\pi_h(v+q)|^2 \\
&\leq \sum_{i=1}^{3} C_i \int_{\hat{K}} \left(a + \int_{\hat{S}_i} |v+q| + \int_{\hat{S}_i} |\nabla(v+q)| \right)^{p-2} \left(\int_{\hat{S}_i} |v+q| + \int_{\hat{S}_i} |\nabla(v+q)| \right)^2 \\
&\leq C \sum_{i=1}^{3} \int_{\hat{K}} \left(a + \int_{\hat{S}_i} |v+q| \right)^{p-2} \left(\int_{\hat{S}_i} |v+q| \right)^2 \\
&\qquad + C \sum_{i=1}^{3} \int_{\hat{K}} \left(a + \int_{\hat{S}_i} |\nabla(v+q)| \right)^{p+2} \left(\int_{\hat{S}_i} |\nabla(v+q)| \right)^2.
\end{aligned}$$

Using the Jensen inequality (3.15), we have that

$$\int_{\hat{K}} \left(a + \int_{\hat{S}_i} |v+q| \right)^{p-2} \left(\int_{\hat{S}_i} |v+q| \right)^2 \leq C \int_{\hat{S}_i} (a + |v+q|)^{p+2} |v+q|^2$$

and

$$\int_{\hat{K}} \left(a + \int_{\hat{S}_i} |\nabla(v+q)| \right)^{p-2} \left(\int_{\hat{S}_i} |\nabla(v+q)| \right)^2 \leq C \int_{\hat{S}_i} (a + |\nabla(v+q)|)^{p+2} |\nabla(v+q)|^2.$$

Hence,

$$\begin{aligned}
&\int_{\hat{K}} (a + |\pi_h(v+q)|)^{p-2} |\pi_h(v+q)|^2 \\
&\leq C \sum_{i=1}^{3} \left(\int_{\hat{S}_i} (a + |v+q|)^{p-2} |v+q|^2 + \int_{\hat{S}_i} (a + |\nabla(v+q)|)^{p-2} |\nabla(v+q)|^2 \right).
\end{aligned}$$

Note that for all $q \in R$, $\pi_h q = q$ and $\cup_{i=1}^{3} S_i$ is a bounded open set and $meas(K) = O(h_K^2) meas(\hat{K})$. Then, similarly as in the proof of Theorem 3.3, we have that (using Corollary 3.2)

$$\begin{aligned}
&\int_K (a + h_K^{-1} |v - \pi_h v|)^{p-2} h_K^{-2} |v - \pi_h v|^2 dx \\
&\leq C \int_{\hat{K}} (a + h_K^{-1} |v - \pi_h v|)^{p-2} |v - \pi_h v|^2 d\hat{x} \\
&\leq C h_K^{2-p} \int_{\hat{K}} (h_K a + |v - \pi_h v|)^{p-2} |v - \pi_h v|^2 \\
&\leq C h_K^{2-p} \inf_{q \in R} \left\{ \int_{\hat{K}} (h_K a + |(I - \pi_h)(v+q)|)^{p-2} |(I - \pi_h)(v+q)|^2 \right\}
\end{aligned}$$

$$
\begin{aligned}
&\le Ch_K^{2-p}\inf_{q\in R}\Big\{\int_{\hat K}(h_Ka+|v+q|)^{p-2}|v+q|^2\\
&\qquad\qquad+\int_{\hat K}(h_Ka+|\pi_h(v+q)|)^{p-2}|\pi_h(v+q)|^2\Big\}\\
&\le Ch_K^{2-p}\inf_{q\in R}\Big\{\sum_{i=0}^{3}\int_{\hat S_i}(h_Ka+|v+q|)^{p-2}|v+q|^2\\
&\qquad\qquad+\int_{\hat S_i}(h_Ka+|\nabla(v+q)|)^{p-2}|\nabla(v+q)|^2)\Big\}\\
&\le Ch_K^{2-p}\inf_{q\in R}\Big\{\sum_{\bar K\cap\bar K'\neq\emptyset}\int_{\hat K'}(h_Ka+|v+q|)^{p-2}|v+q|^2\\
&\qquad\qquad+\int_{\hat K'}(h_Ka+|\nabla(v+q)|)^{p-2}|\nabla(v+q)|^2)\Big\}\\
&\le Ch_K^{2-p}\sum_{\bar K\cap\bar K'\neq\emptyset}\int_{\hat K'}(h_Ka+|\nabla v|)^{p-2}|\nabla v|^2\\
&\le Ch_K^{2-p}h_K^{-2}h_K^p\sum_{\bar K\cap\bar K'\neq\emptyset}\int_{K'}(a+|\nabla v|)^{p-2}|\nabla v|^2\\
&\le C\sum_{\bar K'\cap\bar K\neq\emptyset}\int_{K'}(a+|\nabla v|)^{p-2}|\nabla v|^2.
\end{aligned}
$$

This proves (3.16). The estimate (3.17) can be proved similarly.

However, in deriving a posteriori error estimates for the conforming finite element approximation of the p-Laplacian, a $(=|\nabla u_h|)$ is generally different on different elements. Thus Theorem 3.4 has to be generalized to accommodate the interactions across the elements.

To this end, we first introduce some notations, as in [18] and [19]. Let $u_h\in V^h$. Let l be an edge of an element $K\in T^h$. If l is on the boundary of Ω^h, then we define the element $K^l_{max}=K^l_{min}=K$. Otherwise let $l=\bar K^1_l\cap\bar K^2_l$, where K^1_l, K^2_l are the two elements sharing the common face l. Then we define the element $K^l_{max}(K^l_{min})=K^i_l$ $(i=1$ or $2)$ such that

$$|\nabla u_h|^{p-2}_{K^l_{max}}=\max_{i=1,2}\{|\nabla u_h|^{p-2}_{K^i_l}\},$$

$$|\nabla u_h|^{p-2}_{K^l_{min}}=\min_{i=1,2}\{|\nabla u_h|^{p-2}_{K^i_l}\}.$$

We will take $K^l_{max}=K^l_{min}=K^1_l$ just for fixing the idea if $|\nabla u_h|^{p-2}_{K^1_l}=|\nabla u_h|^{p-2}_{K^2_l}$. Let $[w]_l=w|_{K^1_l}-w|_{K^2_l}$. The purpose of introducing K_{min} and K_{max} is to make some estimators (like η below) sharper. We shall come back to this point at the end of section 4. Using Theorem 3.4, we shall derive two important interpolation error estimates for the conforming piecewise linear finite element which are stated in Theorems 3.3 and 3.4 and will be used in the proofs of Theorems 4.1 and 4.2. We first need a simple proposition, which was adopted from [9].

PROPOSITION 3.5. *Let $G(x,y)$ be defined in* (3.1). *For all $a_1,a_2,\dots,a_n\in R^2$,*

where n is a given positive integer, there exists a constant $C = C(p,n)$ such that

$$\sum_{j=1}^{n}\sum_{k=1}^{j-1} G(|a_j|,|a_j-a_k|) \le C\sum_{\ell=1}^{n-1}\min_{m=1,\dots,n} G(|a_m|,|a_{\ell+1}-a_\ell|). \tag{3.18}$$

Proof. Let $\alpha := (a_1+\cdots+a_n)/n \in R^2$ and $b_j := a_j-\alpha \in R^2$ so that $b_1+\cdots+b_n = 0$. Define

$$f(\alpha;b_1,\dots,b_n) := \sum_{j=1}^{n}\sum_{k=1}^{j-1} G(|\alpha+b_j|,|b_j-b_k|),$$
$$g(\alpha;b_1,\dots,b_n) := \sum_{\ell=1}^{n-1}\min_{m=1,\dots,n} G(|\alpha+b_m|,|b_{\ell+1}-b_\ell|).$$

Observe that $g(\alpha,\cdot)$ is positive for nonzero arguments on

$$X := \{(b_1,\dots,b_n)\in R^{2\times n} :\ b_1+\cdots+b_n = 0\}$$

since $g(\alpha;b_1,\dots,b_n) = 0$ implies $b_1 = b_2 = \ldots = b_n$. Let $B := \{(b_1,\dots,b_n)\in X :\ |b_1|^2+\cdots+|b_n|^2 = 1\}$ denote the unit ball surface in X. Then, for any $\beta\in R^2$,

$$c(\beta) := \max_{(b_1,\dots,b_n)\in B} f(\beta;b_1,\dots,b_n)/g(\beta;b_1,\dots,b_n) < \infty$$

since the denominator is positive and $f(\alpha;\cdot), g(\alpha;\cdot)$ are continuous on the compact set B. The same argument shows

$$c_\infty := \max_{(b_1,\dots,b_n)\in X\setminus\{0\}} \sum_{j=1}^{n}\sum_{k=1}^{j-1}|b_j-b_k|^2 / \sum_{\ell=1}^{n-1}|b_{\ell+1}-b_\ell|^2 < \infty.$$

Note that $\limsup_{|\beta|\to\infty} c(\beta) \le c_\infty < \infty$, and so $c(\beta)$ is a bounded continuous function in $\beta\in R^2$. For all $a_1,\dots,a_n\in R^2$, we have $\alpha\in R^2$, and $(b_1,\dots,b_n)\in X$ as above. Since f and g are positively homogeneous functions, we have, for $\lambda := (|b_1|^2+\cdots+|b_n|^2)^{1/2} > 0$,

$$\begin{aligned} f(\alpha;b_1,\dots,b_n) &= \lambda^p f(\alpha/\lambda;b_1/\lambda,\dots,b_n/\lambda)\\ &\le C\lambda^p g(\alpha/\lambda,b_1/\lambda,\dots,b_n/\lambda) = Cg(\alpha;b_1,\dots,b_n),\end{aligned}$$

where C is only dependent on p and n. This proves our conclusion.

THEOREM 3.6. *Let $u_h\in V^h$. Under the conditions of Theorem* 3.2*,*

$$\begin{aligned}&\sum_K\int_K (|\nabla u_h| + h_K^{-1}|v-\pi v|)^{p-2}h_K^{-2}|v-\pi v|^2\\ &\qquad\le C\sum_K\int_K (|\nabla u_h|+|\nabla v|)^{p-2}|\nabla v|^2 + C\eta^2,\end{aligned} \tag{3.19}$$

$$\begin{aligned}&\sum_K\int_K (|\nabla u_h| + |\nabla(v-\pi v)|)^{p-2}|\nabla(v-\pi v)|^2\\ &\qquad\le C\sum_K\int_K (|\nabla u_h|+|\nabla v|)^{p-2}|\nabla v|^2 + C\eta^2,\end{aligned} \tag{3.20}$$

where $\frac{\partial u_h}{\partial n}$ *is the normal derivative of* u_h*, and*

$$\eta^2 = \sum_l \int_{K^l_{min}} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 .$$

Proof. Note that for all $a \geq 0$, $p > 1$, $(a+x)^{p-2}x^2$ is a convex function, and for any real number a, b,

$$\frac{1}{2}(|a| + |b|) \leq |a| + |a+b| \leq 2(|a| + |b|). \tag{3.21}$$

It follows from Theorem 3.4, Lemma 2.4, and (3.21) that

(3.22)

$$\begin{aligned}
&\sum_K \int_K (|\nabla u_h| + h_K^{-1}|v - \pi v|)^{p-2} h_K^{-2} |v - \pi v|^2 \\
&\leq C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h|_K + |\nabla v|)^{p-2} |\nabla v|^2 \\
&\leq C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h|_K + |\nabla v| + |\nabla (u_h - u_h|_K)|)^{p-2} (|\nabla v| + |\nabla (u_h - u_h|_K)|)^2 \\
&\leq C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla v| + |\nabla (u_h - u_h|_K)|)^{p-2} (|\nabla v| + |\nabla (u_h - u_h|_K)|)^2 \\
&\leq C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2 \\
&\quad + C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla (u_h - u_h|_K)|)^{p-2} |\nabla (u_h - u_h|_K)|^2 \\
&\leq C \sum_K \int_K (|\nabla u_h| + |\nabla v|)^{p-2} (|\nabla v|)^2 \\
&\quad + C \sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla (u_h - u_h|_K)|)^{p-2} |\nabla (u_h - u_h|_K)|^2 .
\end{aligned}$$

We now apply Proposition 3.5. Let $a_j = \nabla u_h|_{K_j}$ with $K_1 \cup K_2 \cup \cdots \cup K_n = \{K' \in T^h : \ \bar{K}' \cap \bar{K} \neq \emptyset\}$. Then n is finite because that T^h is regular. Therefore, it follows from (3.18) that

$$\begin{aligned}
&\sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla (u_h - u_h|_K)|)^{p-2} |\nabla (u_h - u_h|_K)|^2 \\
&\leq C \sum_l \int_{K^l_{min}} (|\nabla u_h| + |[\nabla u_h]_l|)^{p-2} |[\nabla u_h]_l|^2 .
\end{aligned}$$

Note that $\frac{\partial u_h}{\partial t}|_l$ is continuous, where t is the tangent direction of l. We have that $[\frac{\partial u_h}{\partial n}]|_l = [\nabla u_h]|_l$. Hence,

$$\begin{aligned}
&\sum_K \sum_{\bar{K} \cap \bar{K}' \neq \emptyset} \int_{K'} (|\nabla u_h| + |\nabla (u_h - u_h|_K)|)^{p-2} |\nabla (u_h - u_h|_K)|^2 \\
&\leq C \sum_l \int_{K^l_{min}} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 = C\eta^2 .
\end{aligned} \tag{3.23}$$

Then (3.19) follows from (3.22) and (3.23). The estimate (3.20) can be proved similarly.

The following result is the dual version of Theorem 3.6.

THEOREM 3.7. *Let $u_h \in V^h$. Under the conditions of Theorem* 3.4*,*

$$\begin{aligned}&\sum_K \int_K (|\nabla u_h|^{p-1} + h_K^{-1}|v-\pi v|)^{p'-2} h_K^{-2}|v-\pi v|^2 \\ &\le C\sum_K \int_K (|\nabla u_h|^{p-1} + |\nabla v|)^{p'-2}|\nabla v|^2 + C\tilde{\eta}^2, \end{aligned} \tag{3.24}$$

$$\begin{aligned}&\sum_K \int_K (|\nabla u_h|^{p-1} + |\nabla (v-\pi v)|)^{p'-2}|\nabla (v-\pi v)|^2 \\ &\le C\sum_K \int_K (|\nabla u_h|^{p-1} + |\nabla v|)^{p'-2}|\nabla v|^2 + C\tilde{\eta}^2, \end{aligned} \tag{3.25}$$

where

$$\tilde{\eta}^2 = \sum_l \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_l|^2.$$

Proof. Similarly as in the proof Theorem 3.6, it follows from Theorem 3.4 that

$$\begin{aligned}&\sum_K \int_K (|\nabla u_h|^{p-1} + h_K^{-1}|v-\pi v|)^{p'-2} h_K^{-2}|v-\pi v|^2 \\ &\le C\sum_K \sum_{\bar{K}\cap\bar{K}'\neq\emptyset} \int_{K'} (|\nabla u_h|_K^{p-1} + |\nabla v|)^{p'-2}|\nabla v|^2 \\ &\le C\sum_K \int_K (|\nabla u_h|^{p-1} + |\nabla v|)^{p'-2}(|\nabla v|)^2 \\ &\quad + C\sum_K \sum_{\bar{K}\cap\bar{K}'\neq\emptyset} \int_{K'} (|\nabla u_h|^{p-1} + |\,|\nabla u_h|_{K'}^{p-1} - |\nabla u_h|_K^{p-1}|)^{p'-2} \\ &\qquad\qquad \times |\,|\nabla u_h|_{K'}^{p-1} - |\nabla u_h|_K^{p-1}|^2. \end{aligned} \tag{3.26}$$

Let $a_j = |\nabla u_h|^{p-2}\nabla u_h|_{K_j}$. Using Proposition 3.5 and

$$|\,|\nabla u_h|_{K'}^{p-1} - |\nabla u_h|_K^{p-1}| \le |\,|\nabla u_h|^{p-2}\nabla u_h|_{K'} - |\nabla u_h|^{p-2}\nabla u_h|_K|,$$

we have that

$$\begin{aligned}&\int_{K'} (|\nabla u_h|^{p-1} + |\,|\nabla u_h|_{K'}^{p-1} - |\nabla u_h|_K^{p-1}|)^{p'-2}|\,|\nabla u_h|_{K'}^{p-1} - |\nabla u_h|_K^{p-1}|^2 \\ &\le \int_{K'} (|\nabla u_h|^{p-1} + |\,|\nabla u_h|^{p-2}\nabla u_h|_{K'} - |\nabla u_h|^{p-2}\nabla u_h|_K|)^{p'-2} \\ &\quad \times |\,|\nabla u_h|^{p-2}\nabla u_h|_{K'} - |\nabla u_h|^{p-2}\nabla u_h|_K|^2 \\ &\le C\sum_{i=1}^m \int_{K^{l_i}_{max}} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_{l_i}|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_{l_i}|^2 = \tilde{\eta}^2. \end{aligned} \tag{3.27}$$

Then (3.24) follows from (3.26) and (3.27). The estimate (3.25) can be proved similarly.

Next is a well-known trace theorem.

LEMMA 3.8 (see [14]). *For all $v \in W^{1,q}(K)$, $1 \le q < \infty$,*

$$\|v\|_{0,q,\partial K} \le C(h_K^{-\frac{1}{q}}\|v\|_{0,q,K} + h_K^{1-\frac{1}{q}}|v|_{1,q,K}). \tag{3.28}$$

We need another lemma which is the quasi-norm version trace theorem for piecewise polynomials.

LEMMA 3.9. *Let $K \in T^h$ and v be an s-degree polynomial with $s \le k$, where k is a fixed nonnegative integer. Then*

$$h_K \int_{\partial K} (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2 \le C \int_K (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2, \tag{3.29}$$

where the constant C depends only on k, the maximum degree of the polynomials, and the regularity constant of T^h.

Proof. Because v is a polynomial in K, by an inverse inequality, we have that for any $x \in K$,

$$|\nabla v(x)| \le |v|_{1,\infty,K} \le Ch_K^{-2} \int_K |\nabla v| \, .$$

Therefore, it follows from the Jensen inequality (3.15) that

$$\begin{aligned} & h_K \int_{\partial K} (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2 \\ & \le Ch_K^2 \left(|\nabla u_h| + \int_K Ch_K^{-2}|\nabla v|\right)^{p-2} \left(\int_K Ch_K^{-2}|\nabla v|\right)^2 \\ & \le C \int_K (|\nabla u_h| + |\nabla v|)^{p-2}|\nabla v|^2 \, . \end{aligned}$$

This proves (3.29).

4. Applications for a posteriori error estimates. As applications of Theorems 3.3–3.7, we derive a posteriori error estimates for the finite element approximations of the p-Laplacian.

4.1. Conforming finite element. Let the finite element space V_0^h be the standard conforming piecewise linear triangular element defined in section 2. Using the quasi-norm interpolation error estimates in section 3, we can prove the following a posteriori error bounds.

THEOREM 4.1. *Let u and u_h be the solutions of (2.2) and (2.3), respectively. Let $p > 1$ and $p' > 1$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Assume that $f \in L^{p'}(\Omega)$. Then, there is a $\delta_0 > 0$ for all $0 < \delta \le \delta_0$ such that*

$$|u - u_h|_{(p)}^2 \le C(\delta)(\eta_1^2 + \eta_2^2) + C\delta\eta^2, \tag{4.1}$$

$$\eta_1^2 + \eta_2^2 \le C|u - u_h|_{(p)}^2 + C\epsilon^2, \tag{4.2}$$

with

$$|u - u_h|_{(p)}^2 = \int_\Omega (|\nabla u_h| + |\nabla(u - u_h)|)^{p-2}|\nabla(u - u_h)|^2,$$

$$\eta_1^2 = \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2,$$

$$\eta_2^2 = \sum_l \int_{K_{max}^l} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2} A_l^2,$$

$$\eta^2 = \sum_l \int_{K_{min}^l} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2,$$

$$\epsilon^2 = \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f - \bar{f}|)^{p'-2} h_K^2 |f - \bar{f}|^2,$$

where $\bar{f}|_K = \int_K f/|K|$, *and* A_l *is the jump of the p-normal derivative of* u_h *over the interior edge* $l = \bar{K}_l^1 \cap \bar{K}_l^2$, *defined by*

$$A_l = \left[|\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} \right]_l = ((|\nabla u_h|^{p-2} \nabla u_h)_{K_l^1} - (|\nabla u_h|^{p-2} \nabla u_h)_{K_l^2}) \cdot n,$$

where n *is the unit normal vector of the face* l *outwards* K_l^1. *For later convenience, we shall define the jump* A_l *to be zero when an edge* l *is on the boundary.*

Proof. Let $e = u - u_h$. Then, it follows from Lemma 2.1, (2.2), and (2.3) that

$$\begin{aligned}
|u - u_h|_{(p)}^2 &= \int_\Omega (|\nabla u| + |\nabla(u - u_h)|)^{p-2} |\nabla(u - u_h)|^2 \\
&\le C \int_\Omega (|\nabla u|^{p-2} \nabla u - |\nabla u_h|^{p-2} \nabla u_h) \nabla(u - u_h) \\
&= C \int_\Omega (|\nabla u|^{p-2} \nabla u - |\nabla u_h|^{p-2} \nabla u_h) \nabla e \\
&= C \int_\Omega (|\nabla u|^{p-2} \nabla u - |\nabla u_h|^{p-2} \nabla u_h) \nabla(e - \pi_h e) \qquad (4.3) \\
&= -C \sum_K \int_K \operatorname{div}(|\nabla u|^{p-2} \nabla u - |\nabla u_h|^{p-2} \nabla u_h)(e - \pi_h e) \\
&\quad + C \sum_K \int_{\partial K} \left(|\nabla u|^{p-2} \frac{\partial u}{\partial n} - |\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} \right) (e - \pi_h e) \\
&= C \sum_K \int_K f(e - \pi_h e) - C \sum_l \int_l \left[|\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} \right] (e - \pi_h e) \\
&= I_1 + I_2,
\end{aligned}$$

where the average interpolator π_h is defined in section 3, and we have used the fact that $e \in W_0^{1,p}(\Omega)$ and $\pi_h e \in V_0^h$. By Lemma 2.3 and Theorem 3.6, for any $\delta_1 > 0$

$$\begin{aligned}
I_1 &= C \sum_K \int_K h_K f h_K^{-1} (e - \pi_h e) \\
&\le C \delta_1^{-\beta} \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2
\end{aligned}$$

$$
\begin{aligned}
&+ C\delta_1 \sum_K \int_K (|\nabla u_h| + h_K^{-1}|e-\pi_h e|)^{p-2} h_K^{-2} |e-\pi_h e|^2 \\
&\le C\delta_1^{-\beta}\eta_1^2 + C\delta_1 \sum_K \int_K (|\nabla u_h| + |\nabla e|)^{p-2} |\nabla e|^2 \\
&\quad + C\delta_1 \sum_l \int_{K^l_{min}} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 \\
&= C\delta_1^{-\beta}\eta_1^2 + C\delta_1 |u-u_h|^2_{(p)} + C\delta_1 \eta^2,
\end{aligned} \tag{4.4}
$$

where β is defined in Lemma 2.3. Similarly, it follows from Lemma 2.3, Lemma 3.8, and Theorem 3.6 that for any $\delta_2 > 0$

$$
\begin{aligned}
I_2 &= -C\sum_l \int_l \left[|\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} \right] (e-\pi_h e) \le C \sum_l \int_{\partial K^l_{max}} |A_l| \, |e-\pi_h e| \\
&\le C \sum_l \int_{K^l_{max}} |A_l| \, (h^{-1}_{K^l_{max}} |e-\pi_h e| + |\nabla (e-\pi_h e)|) \\
&\le C\delta_2^{-\beta} \sum_l \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2} A_l^2 \\
&\quad + C\delta_2 \sum_l \int_{K^l_{max}} (|\nabla u_h| + h^{-1}_{K^l_{max}} |e-\pi_h e|)^{p-2} h^{-2}_{K^l_{max}} |e-\pi_h e|^2 \\
&\quad + C\delta_2 \sum_l \int_{K^l_{max}} (|\nabla u_h| + |\nabla (e-\pi_h e)|)^{p-2} |\nabla (e-\pi_h e)|^2 \\
&\le C\delta_2^{-\beta}\eta_2^2 + C\delta_2 \sum_K \int_K (|\nabla u_h| + |\nabla e|)^{p-2} |\nabla e|^2 \\
&\quad + C\delta_2 \sum_l \int_{K^l_{min}} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 \\
&= C\delta_2^{-\beta}\eta_2^2 + C\delta_2 |u-u_h|^2_{(p)} + C\delta_2 \eta^2.
\end{aligned} \tag{4.5}
$$

Then (4.1) follows from (4.3)–(4.5) if $\delta_0 \le \frac{1}{4C}$. The estimate (4.2) has been proved in [18].

Remark 4.1. Let us examine the structure of the estimators in Theorem 4.1. As in the linear case, the idea is to bound the residual

$$
R(u_h) = (f, u-\pi_h u) - a(u_h, u-\pi_h u) = \sum_K \int_K h_K f h_K^{-1} (e-\pi_h e) - \int_l A_l (e-\pi_h e).
$$

Thus the building blocks $h_K f$ and A_l enter the estimators η_1 and η_2 as usual. The reason why η_1 and η_2 have the unconventional formats is due to the fact we have to estimate the above bilinear terms in the quasi norm, and this needs to apply the generalized Young inequality in Lemma 2.3.

The term η comes from the interpolation error estimates in Theorem 3.6—due to the gradient jumps $|\nabla u_h|$ over the elements—and it seems to be indispensable in the estimates. The relative contribution of η may be controlled by δ to a certain degree; see below.

To see η more clearly, let

$$V_h^k = \{v \in C^1(\bar{\Omega}) : v|_K \in P_k \ \forall K \in T^h\},$$

where P_k is the space of k-degree polynomials, $k \geq 1$. It follows from Lemma 3.9 and Lemma 2.5 that for all $v_h^k \in V_h^k$,

$$\begin{aligned}
\eta^2 &= \sum_l \int_{K_{min}^l} \left(|\nabla u_h| + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 \\
&\leq C \sum_l h_{K_{min}^l} \int_l \left(|\nabla u_h|_{K_{min}^l} + \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} \right]_l \right|^2 \\
&= C \sum_l h_{K_{min}^l} \int_l \left(|\nabla u_h|_{K_{min}^l} + \left| \left[\frac{\partial u_h}{\partial n} - \frac{\partial v_h^k}{\partial n} \right]_l \right| \right)^{p-2} \left| \left[\frac{\partial u_h}{\partial n} - \frac{\partial v_h^k}{\partial n} \right]_l \right|^2 \\
&\leq C \sum_l \int_{K_l^1 \cup K_l^2} (|\nabla u_h|_{K_{min}^l} + |\nabla (u_h - v_h^k)|)^{p-2} |\nabla (u_h - v_h^k)|^2 \\
&\leq C \sum_K \int_K (|\nabla u_h| + |\nabla (u_h - v_h^k)|)^{p-2} |\nabla (u_h - v_h^k)|^2 \\
&\leq C |u - u_h|_{(p)}^2 + C |u - v_h^k|_{(p)}^2.
\end{aligned}$$

Then immediately we can obtain the upper error estimates derived in [18] by letting δ be small enough.

Furthermore it can be seen (ignoring ϵ for the time being) that

$$cE^2 - c\delta \inf_{v_h^k \in V_h^k} |u_h - v_h^k|_{(u_h,p)}^2 \leq |u - u_h|_{(p)}^2 \leq C(\delta) E^2,$$

where $E^2 = \eta_1^2 + \eta_2^2 + \delta \eta^2$ with $C(\delta)\delta \to 0$ as $\delta \to 0$. Thus

$$cE^2 - c\delta \inf_{v_h^k \in V_h^k} |u - v_h^k|_{(p)}^2 \leq |u - u_h|_{(p)}^2 \leq C(\delta) E^2.$$

A posteriori error estimates of this type have also been obtained recently for some well-known a posteriori error estimators based on gradient recovery, like the Z-Z estimator; see [6], [7] for instance. Clearly E is always reliable, and if u is smooth enough, then the estimator E is both reliable and efficient. It appears that our result is slightly stronger, as here δ can be made small. We tested the model problems Examples 5.3 and 5.4 in [3] for a range of $\delta \in [0.01, 0.1]$ by using the Polak–Ribiere conjugate gradient method. The implementation settings are similar to those in [3] and [18]. We found that the estimators $\eta_1^2 + \eta_2^2$ and E^2 behave quite similarly, at least for the tests problems. Thus it seems that $\delta \eta^2$ is serviced only as a deterrence for most applications.

4.2. Nonconforming finite element. Next, we shall discuss the nonconforming finite element approximation of the p-Laplacian stated in section 2. Thanks to Theorems 3.3–3.7, we can prove the following a posteriori error estimates.

THEOREM 4.2. *Let u and u_h be the solutions of* (2.2) *and* (2.4), *respectively. Let $p > 1$ and $p' > 1$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Assume that $f \in L^{p'}(\Omega)$. Then there is a $\delta_0 > 0$ for all $0 < \delta \leq \delta_0$ such that*

$$|u - u_h|_{(p)}^2 \leq C(\delta)(\tilde{\eta}_1^2 + \tilde{\eta}_2^2) + C\delta \tilde{\eta}^2, \tag{4.6}$$

$$\tilde{\eta}_1^2 + \tilde{\eta}_2^2 \le C|u-u_h|_{(p)}^2 + C\epsilon^2, \tag{4.7}$$

with

$$|u-u_h|_{(p)}^2 = \sum_K \int_K (|\nabla u_h| + |\nabla(u-u_h)|)^{p-2}|\nabla(u-u_h)|^2,$$

$$\tilde{\eta}_1^2 = \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2,$$

$$\tilde{\eta}_2^2 = \sum_l \int_{K^l_{min}} (|\nabla u_h| + |B_l|)^{p-2} B_l^2,$$

$$\tilde{\eta}^2 = \sum_l \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2} |[|\nabla u_h|^{p-2}\nabla u_h]_l|^2,$$

$$\epsilon^2 = \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f-\bar{f}|)^{p'-2} h_K^2 |f-\bar{f}|^2,$$

where B_l is the jump of the tangent derivative of u_h on edge $l = \bar{K}_l^1 \cap \bar{K}_l^2$:

$$B_l = \frac{\partial u_h}{\partial t}|_{K_l^1} - \frac{\partial u_h}{\partial t}|_{K_l^2},$$

t is the tangent direction of l. If $l \subset \partial\Omega$ and l is the edge of K^l, we denote $B_l = \frac{\partial u_h}{\partial t}|_{K^l}$.

Proof. Let $\nabla u_h|_K = \nabla(u_h|_K)$. Let us introduce an auxiliary function (see [21] for the details) $\phi \in W_0^{1,p}(\Omega)$ satisfying

$$\int_\Omega |\nabla\phi|^{p-2}\nabla\phi\nabla v = \int_\Omega |\nabla u_h|^{p-2}\nabla u_h \nabla v \quad \forall v \in W_0^{1,p}(\Omega). \tag{4.8}$$

It follows from Lemma 2.1 and Lemma 2.5 that

$$\begin{aligned} |u-u_h|_{(p)}^2 &\le C\sum_K \int_K (|\nabla\phi| + |\nabla(\phi-u)|)^{p-2}|\nabla(\phi-u)|^2 \\ &+ C\sum_K \int_K (|\nabla\phi| + |\nabla(\phi-u_h)|)^{p-2}|\nabla(\phi-u_h)|^2 = I_1 + I_2. \end{aligned} \tag{4.9}$$

Let $\tilde{\pi}_h$ be the interpolation operator defined in section 3. Note that $\tilde{\pi}_h w \in \tilde{V}_0^h$ if $w \in W_0^{1,p}(\Omega)$, and

$$\int_l \tilde{\pi}_h w = \int_l w,$$

where l are the edges of an element. It follows from (2.2), (2.4), (4.8), and Lemma 2.1 that

$$I_1 \le C\sum_K \int_K (|\nabla u|^{p-2}\nabla u - |\nabla\phi|^{p-2}\nabla\phi)\nabla(u-\phi)$$

$$= C\left(\sum_K \int_K f(u-\phi) - \sum_K \int_K |\nabla u_h|^{p-2}\nabla u_h \nabla(u-\phi)\right)$$
$$= C\left(\sum_K \int_K f((u-\phi)-\tilde{\pi}_h(u-\phi)) - \sum_K \int_K |\nabla u_h|^{p-2}\nabla u_h \nabla((u-\phi)-\tilde{\pi}_h(u-\phi))\right)$$
$$= C\sum_K \int_K f((u-\phi)-\tilde{\pi}_h(u-\phi)).$$

Let $\tilde{e} = u - \phi$. It follows from Lemmas 2.3–2.4 and Theorem 3.3 that

$$\begin{aligned} I_1 &= C\sum_K \int_K f((u-\phi)-\tilde{\pi}_h(u-\phi)) = C\sum_K \int_K h_K f h_K^{-1}(\tilde{e}-\tilde{\pi}_h\tilde{e}) \\ &\le C\delta_1^{-\beta}\sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2 \\ &\quad + C\delta_1 \sum_K \int_K (|\nabla u_h| + h_K^{-1}|\tilde{e}-\tilde{\pi}_h\tilde{e}|)^{p-2} h_K^{-2}|\tilde{e}-\tilde{\pi}_h\tilde{e}|^2 \\ &\le C\delta_1^{-\beta}\tilde{\eta}_1^2 + C\delta_1 \sum_K \int_K (|\nabla u_h| + |\nabla \tilde{e}|)^{p-2}|\nabla \tilde{e}|^2 \qquad (4.10) \\ &\le C\delta_1^{-\beta}\tilde{\eta}_1^2 + C\delta_1 \sum_K \int_K (|\nabla u_h| + |\nabla(u-u_h)|)^{p-2}|\nabla(u-u_h)|^2 \\ &\quad + C\delta_1 \sum_K \int_K (|\nabla u_h| + |\nabla(\phi-u_h)|)^{p-2}|\nabla(\phi-u_h)|^2 \\ &\le C\delta_1^{-\beta}\tilde{\eta}_1^2 + C\delta_1(|u-u_h|_{(p)}^2 + I_2). \end{aligned}$$

Note the well-known fact (see [21]) that since $\mathrm{div}(|\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h) = 0$, there exists a function $\psi \in W^{1,p'}(\Omega)$ such that

$$|\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h = curl\,\psi. \qquad (4.11)$$

Moreover, note that for any continuous piecewise linear function $v_h \in V^h$,

$$\sum_K \int_K \nabla u_h curl\ v_h = 0.$$

Then it follows from Lemma 2.1, Lemma 3.8, (4.8), and integrating by parts that

$$\begin{aligned} I_2 &\le C\sum_K \int_K (|\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h)\nabla(\phi-u_h) \\ &= -C\sum_K \int_K (|\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h)\nabla u_h \\ &= -C\sum_K \int_K curl\ \psi \nabla u_h = -C\sum_K \int_K curl\ (\psi - \pi_h\psi)\nabla u_h \\ &= -C\sum_K \int_{\partial K} (\psi-\pi_h\psi)\frac{\partial u_h}{\partial t} \le C\sum_l \int_l |\psi-\pi_h\psi|\ \left|\left[\frac{\partial u_h}{\partial t}\right]\right| \\ &= C\sum_l \int_l |\psi-\pi_h\psi|\ |B_l| \le \sum_l \int_{\partial K_{min}^l} |B_l|\ |\psi-\pi_h\psi| \end{aligned}$$

$$\leq C\sum_l \int_{K^l_{min}} |B_l|\ (h^{-1}_{K^l_l}|\psi-\pi_h\psi| + |\nabla(\psi-\pi_h\psi)|),$$

where we have used the average interpolation operator π_h defined in section 3 and the fact that $\pi_h\psi \in V^h$ is a continuous piecewise linear function. It follows from Lemma 2.3 and Theorem 3.7 that

$$\begin{aligned}
I_2 &\leq C\sum_l \int_{K^l_{min}} |B_l|(h_K^{-1}|\psi-\pi_h\psi| + |\nabla(\psi-\pi_h\psi)|)\\
&\leq C\delta_2 \sum_l \int_{K^l_{min}} (|\nabla u_h| + |B_l|)^{p-2}|B_l|^2\\
&\quad + C\delta_2^{-\beta}\sum_l \int_{K^l_{min}} (|\nabla u_h|^{p-1} + h^{-1}_{K^l_{min}}|\psi-\pi_h\psi|)^{p'-2}\\
&\quad \times h^{-2}_{K^l_{min}}|\psi-\pi_h\psi|^2 + C\delta_2^{-\beta}\sum_l \int_{K^l_{min}} (|\nabla u_h|^{p-1} + |\nabla(\psi-\pi_h\psi)|)^{p'-2}|\nabla(\psi-\pi_h\psi)|^2\\
&\leq C\delta_2\tilde{\eta}_2^2 + C\delta_2^{-\beta}\sum_K \int_K (|\nabla u_h|^{p-1} + |\nabla\psi|)^{p'-2}|\nabla\psi|^2\\
&\quad + C\delta_2^{-\beta}\sum_l \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_l|^2\\
&\leq C\delta_2\tilde{\eta}_2^2 + C\delta_2^{-\beta}\sum_K \int_K (|\nabla u_h|^{p-1} + |curl\psi|)^{p'-2}|curl\psi|^2 + C\delta_2^{-\beta}\tilde{\eta}^2.
\end{aligned}$$

It follows from (4.11) and Lemma 2.1 that for all $K \in T^h$

$$\begin{aligned}
&\int_K (|\nabla u_h|^{p-1} + |curl\psi|)^{p'-2}|curl\psi|^2\\
&= \int_K (|\nabla u_h|^{p-1} + |\ |\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h|)^{p'-2}|\ |\nabla\phi|^{p-2}\nabla\phi - |\nabla u_h|^{p-2}\nabla u_h|^2\\
&\leq \int_K (|\nabla u_h|^{p-1} + (|\nabla u_h| + |\nabla(u_h-\phi)|)^{p-2}|\nabla(u_h-\phi)|)^{p'-2}\\
&\quad \times (|\nabla u_h| + |\nabla(u_h-\phi)|)^{2(p-2)}|\nabla(u_h-\phi)|^2\\
&= \int_K Q(|\nabla u_h| + |\nabla(u_h-\phi)|)^{p-2}|\nabla(u_h-\phi)|^2,
\end{aligned}$$

where

$$Q = \frac{[|\nabla u_h|^{p-1} + (|\nabla u_h| + |\nabla(u_h-\phi)|)^{p-2}|\nabla(u_h-\phi)|]^{p'-2}}{(|\nabla u_h| + |\nabla(u_h-\phi)|)^{2-p}}.$$

When $1 < p \leq 2$, for all $x \in K$, if $|\nabla u_h|^{p-1} \leq (|\nabla u_h| + |\nabla(u_h-\phi)|)^{p-2}|\nabla(u_h-\phi)|$,

$$Q \leq \frac{2^{p'-2}|\nabla(u_h-\phi)|^{(p-1)(p'-2)}}{|\nabla(u_h-\phi)|^{2-p}} = 2^{p'-2}.$$

If $(|\nabla u_h| + |\nabla(u_h-\phi)|)^{p-2}|\nabla(u_h-\phi)| < |\nabla u_h|^{p-1}$,

$$Q \leq \frac{2^{p'-2}|\nabla u_h|^{(p-1)(p'-2)}}{|\nabla u_h|^{2-p}} = 2^{p'-2}.$$

When $p > 2$, for all $x \in K$, if $|\nabla u_h| \le |\nabla(u_h - \phi)|$,

$$Q \le \frac{2^{p-2}|\nabla(u_h - \phi)|^{p-2}}{|\nabla(u_h - \phi)|^{(p-1)(2-p')}} = 2^{p-2}.$$

If $|\nabla(u_h - \phi)| < |\nabla u_h|$,

$$Q \le \frac{2^{p-2}|\nabla u_h|^{p-2}}{|\nabla u_h|^{(p-1)(2-p')}} = 2^{p-2}.$$

Hence, we have that for all $x \in K$, $Q \le C$, where C is only dependent on p. It follows that for all $K \in T^h$,

$$\begin{aligned}\int_K &(|\nabla u_h|^{p-1} + |curl\psi|)^{p'-2}|curl\psi|^2 \\ &\le C\int_K (|\nabla u_h| + |\nabla(u_h - \phi)|)^{p-2}|\nabla(u_h - \phi)|^2 = CI_2.\end{aligned}$$

Therefore,

$$I_2 \le C\delta_2\tilde{\eta}_2^2 + C\delta_2^{-\beta}I_2 + C\delta_2^{-\beta}\tilde{\eta}^2.$$

Then, we have

$$I_2 \le C\delta_2\tilde{\eta}_2^2 + C\delta_2^{-\beta}\tilde{\eta}^2\ . \tag{4.12}$$

By (4.10) and (4.12),

$$I_1 \le C(\delta)(\tilde{\eta}_1^2 + \tilde{\eta}_2^2) + \frac{1}{2}|u - u_h|_{(p)}^2 + C\delta\tilde{\eta}^2\ . \tag{4.13}$$

Hence, (4.6) follows from (4.9), (4.13), and (4.12). The estimate (4.7) has been proved in [19].

Remark 4.2. It can be seen that the purpose of introducing K_{min} and K_{max} is to make our estimators smaller (sharper). That is why we use either K_{max} or K_{min}, depending on whether the power in the estimators is $p' - 2$ or $p - 2$. Furthermore, it can be shown that the bubble function Lemma 3.1 in [18], which is needed to prove the lower bounds there, does not hold if K_{max} is not used. In this sense introducing K_{min} and K_{max} seems to be necessary.

Remark 4.3. Similar interpretations made in Remark 4.1 apply here. In particular, it follows from Lemma 2.1 that

$$|[|\nabla u_h|^{p-2}\nabla u_h]_l| \le C(|\nabla u_h|_{K_{min}^l} + |[\nabla u_h]_l|)^{p-2}|[\nabla u_h]_l|.$$

Then,

$$\begin{aligned}\tilde{\eta}^2 &= \sum_l \int_{K_{max}^l} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_l|^2 \\ &\le C\sum_l h_{K_{max}^l}\int_l (|\nabla u_h|_{K_{max}^l}^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_l|^2 \\ &\le C\sum_l h_{K_{max}^l}\int_l (|\nabla u_h|_{K_{max}^l}^{p-1} + (|\nabla u_h|_{K_{min}^l} + |[\nabla u_h]_l|)^{p-2}|[\nabla u_h]_l|)^{p'-2}\end{aligned}$$

$$\begin{aligned}
&\times (|\nabla u_h|_{K^l_{min}} + |[\nabla u_h]_l|)^{2(p-2)}|[\nabla u_h]_l|^2 \\
&= C\sum_l h_{K^l_{max}} \int_l Q(|\nabla u_h|_{K^l_{min}} + |[\nabla u_h]_l|)^{p-2}|[\nabla u_h]_l|^2 \\
&\le C\sum_l \int_{K^l_{min}} Q(|\nabla u_h| + |[\nabla u_h]_l|)^{p-2}|[\nabla u_h]_l|^2,
\end{aligned}$$

with

$$Q = \frac{(|\nabla u_h|^{p-1}_{K^l_{max}} + (|\nabla u_h|_{K^l_{min}} + |[\nabla u_h]_l|)^{p-2}|[\nabla u_h]_l|)^{p'-2}}{(|\nabla u_h|_{K^l_{min}} + |[\nabla u_h]_l|)^{2-p}} \le C,$$

where C is a constant only dependent on p. Hence, similarly as in Remark 4.1, it can be proved that for all $v_h^k \in V_h^k$,

$$\tilde{\eta}^2 \le C|u - u_h|^2_{(p)} + C|u - v_h^k|^2_{(p)}.$$

Thus the estimates in Theorem 4.2 seem to be sharp as well.

Remark 4.4. Note that

$$\begin{aligned}
\tilde{\eta}^2 &= \sum_l \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |[|\nabla u_h|^{p-2}\nabla u_h]_l|)^{p'-2}|[|\nabla u_h|^{p-2}\nabla u_h]_l|^2 \\
&\le \sum_l \int_{K^l_{max}} \left(|\nabla u_h|^{p-1} + \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial t}\right]_l\right|\right)^{p'-2} \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial t}\right]_l\right|^2 \\
&\quad + \sum_l \int_{K^l_{max}} \left(|\nabla u_h|^{p-1} + \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial n}\right]_l\right|\right)^{p'-2} \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial n}\right]_l\right|^2,
\end{aligned}$$

where t and n are the tangent and normal directions of l, respectively. Moreover, it has been proved that (see Theorem 4.1 and [18] for the details)

$$\sum_l \int_{K^l_{max}} \left(|\nabla u_h|^{p-1} + \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial n}\right]_l\right|\right)^{p'-2} \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial n}\right]_l\right|^2 \le C|u-u_h|^2_{(p)} + C\epsilon^2,$$

where ϵ is defined in Theorem 4.1. Then, it can be proved that for all $\delta > 0$

$$|u - u_h|^2_{(p)} \le C(\delta)(\tilde{\eta}_1^2 + \tilde{\eta}_2^2) + C\delta\hat{\eta}^2 + C\delta\epsilon^2,$$

where $\tilde{\eta}_1^2$ and $\tilde{\eta}_2^2$ are defined in Theorem 4.2, ϵ is defined in Theorem 4.1, and

$$\hat{\eta}^2 = \sum_l \int_{K^l_{max}} \left(|\nabla u_h|^{p-1} + \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial t}\right]_l\right|\right)^{p'-2} \left|\left[|\nabla u_h|^{p-2}\frac{\partial u_h}{\partial t}\right]_l\right|^2.$$

5. Conclusions. We have established new interpolation error estimates for some well-known averaging interpolators. These estimates have proved essential in deriving improved a posteriori error estimates in the quasi norms.

Acknowledgments. The authors wish to express their sincere thanks to Mr. R. Klose for his computational work in Remark 4.1.

REFERENCES

[1] I. Babuška, R. Durán, and R. Rodríguez, *Analysis of the efficiency of an a posteriori error estimator for linear triangular finite elements*, SIAM J. Numer. Anal., 29 (1992), pp. 947–964.

[2] J. Baranger and H. El. Amri, *Estimateurs a posteriori d'erreur pour le calcul adaptatif d'ecoulements quasi-Newtoniens*, R.A.I.R.O. Modél Math. Anal. Numér, 25 (1991), pp. 31–48.

[3] J.W. Barrett and W.B. Liu, *Finite element approximation of the p-Laplacian*, Math. Comp., 61 (1993), pp. 523–537.

[4] J.W. Barrett and W.B. Liu, *Finite element approximation of some degenerate quasi-linear problems*, in Numerical Analysis 1993, Lecture Notes in Math., 303 (1994), pp. 1–16.

[5] J.W. Barrett and W.B. Liu, *Quasi-norm error bounds for finite element approximation of quasi-Newtonian flows*, Numer. Math., 68 (1994), pp. 437–456.

[6] C. Carstensen and S.A. Funken, *Averaging technique for FE-a posteriori error control in elasticity. Part* I: *Conforming FEM*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2483–2498.

[7] C. Carstensen and S.A. Funken, *Averaging technique for FE-a posteriori error control in elasticity. Part* II: λ*-independent estimates*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4663–4675.

[8] C. Carstensen and R. Klose, *Guaranteed a posteriori finite element error control for p-Laplacian*, SIAM J. Sci. Comput., submitted.

[9] C. Carstensen, W.B. Liu, and N. Yan, *A posteriori error estimators based on gradient recovery for finite element approximation of p-Laplacian*, SIAM J. Numer. Anal., submitted.

[10] S.S. Chow, *Finite element error estimates for non-linear elliptic equations of monitone type*, Numer. Math., 54 (1988), pp. 373–393.

[11] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] M. Crouzeix and P.A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–76.

[13] R. Glowinski and A. Marrocco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problémes de Dirichlet non linéaires (French), RAIRO Anal. Numer., 9 (1975), pp. 41–76.

[14] A. Kufner, O. John, and S. Fucik, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.

[15] G.M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.

[16] W.B. Liu and J.W. Barrett, *Finite element approximation of some degenerate monotone quasilinear elliptic systems*, SIAM J. Numer. Anal., 33 (1996), pp. 88–106.

[17] W.B. Liu, *Degenerate quasilinear elliptic equations arising from bimaterial problems in elastic-plastic mechanics* II, Numer. Math., 86 (2000), pp. 491–506.

[18] W.B. Liu and N. Yan, *Quasi-norm local error estimator for p-Laplacian*, SIAM. J. Numer. Anal., 39 (2001), pp. 100–127.

[19] WB Liu and N.N. Yan, *Quasi-norm a posteriori error estimates for non-conforming finite element approximation of p-Laplacian*, Numer. Math., 89 (2001), pp. 341–378.

[20] J.T. Oden, L. Demkowicz, T. Strouboulis, and Ph. Devloo, *Adaptive methods for problems in solid and fluid mechanics*, in Accuracy Estimates and Adaptive Refinements in Finite Element Computations, I. Babuška, O.C. Zienkiewicz, J. Gago, and E.R. de A. Oliveira, eds., Wiley, Chichester, 1986, pp. 249–280.

[21] C. Padra, *A posteriori error estimators for nonconforming approximation of some quasi-Newtonian flows*, SIAM J. Numer. Anal., 34 (1997), pp. 1600–1615.

[22] L.R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[23] G. Simms, *Finite element approximation of some nonlinear elliptic and parabolic problems*, Thesis, Imperial College, University of London, London, 1995.

[24] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[25] R. Verfürth, *A posteriori error estimates for non-linear problems*, Math. Comp., 62 (1994), pp. 445–475.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1896–1939

A CONSTRAINT SATISFACTION APPROACH FOR ENCLOSING SOLUTIONS TO PARAMETRIC ORDINARY DIFFERENTIAL EQUATIONS*

MICHA JANSSEN†, PASCAL VAN HENTENRYCK‡, AND YVES DEVILLE†

Abstract. This paper considers initial value problems for ordinary differential equations (ODEs), where some of the data is uncertain and given by intervals as is the case in many areas of science and engineering. Interval methods provide a way to approach these problems, but they raise fundamental challenges in obtaining high accuracy and low computation costs. This work introduces a constraint satisfaction approach to these problems which enhances traditional interval methods with a pruning step based on a global relaxation of the ODE. The relaxation uses Hermite interpolation polynomials and enclosures of their error terms to approximate the ODE. Our work also shows how to find an evaluation time for the relaxation that minimizes its local error. Theoretical and experimental results show that the approach produces significant improvements in accuracy over the best interval methods for the same computation costs. The results also indicate that the new algorithm should be significantly faster when the ODE contains many operations.

Key words. ordinary differential equation, interval methods, constraint satisfaction polynomial

AMS subject classifications. 65L05, 65G20

PII. S0036142901392316

1. Introduction.

Initial value problems (IVPs) for ordinary differential equations (ODEs) arise naturally in many applications in science and engineering, including chemistry, physics, molecular biology, and mechanics to name only a few. An ODE $\mathbb{O}$ is a system of the form

$$\begin{aligned} u_1'(t) &= f_1(u_1(t), \ldots, u_n(t)), \\ &\vdots \\ u_n'(t) &= f_n(u_1(t), \ldots, u_n(t)) \end{aligned}$$

often denoted in vector notation by $u'(t) = f(u(t))$ or $u' = f(u)$.[1] An IVP is an ODE with an initial condition $u(t_0) = u_0$. It is often the case that the parameters and/or the initial values are not known with certainty but are given as intervals. Hence, traditional methods may not be the simplest way to approach the resulting parametric ODEs since, in essence, they would have to solve infinitely many systems. *Interval methods*, pioneered by Moore [21], provide an approach to tackle parametric ODEs. They return enclosures of exact solutions at different points in time; i.e., for a given IVP, they are guaranteed to return intervals containing the exact solution. In addition, they inherently accommodate uncertainty in the parameters or initial values by using intervals instead of floating-point numbers. In this paper, we talk about ODEs to denote both traditional and parametric ODEs.

*Received by the editors July 12, 2001; accepted for publication (in revised form) April 3, 2002; published electronically December 3, 2002. This research was partially supported by the actions de recherche concertée ARC/95/00-187 and an NSF NYI award.

http://www.siam.org/journals/sinum/40-5/39231.html

†UCL, 2 Place Sainte Barbe, B-1348 Louvain-La-Neuve, Belgium (mja@info.ucl.ac.be, yde@info.ucl.ac.be).

‡Brown University, Box 1910, Providence, RI 02912 (pvh@cs.brown.edu).

[1]Only autonomous systems are considered in this paper, but it is not difficult to generalize the results to nonautonomous systems.

Traditional interval methods usually consist of two processes applied at each integration step: (1) a *bounding box* process that proves existence and uniqueness of the solution and computes a rough enclosure (called a *bounding box*) of the solution over a time interval $[t_0, t_1]$; (2) a *forward* process that computes an enclosure of the solution at t_1. The bounding box process, which is specific to interval methods, is necessary to bound the error terms in the forward process. The forward process is generally realized by applying a one-step Taylor interval method and making extensive use of automatic differentiation [27] to obtain the Taylor coefficients [8, 16, 21, 22]. However, the major problem of such methods is the explosion of the size of the boxes at successive points as they often accumulate errors from point to point and lose accuracy by enclosing the solution by a box. (This is called the *wrapping effect*.) Lohner's AWA system [20] was an important step in interval methods which features efficient coordinate transformations to tackle the wrapping effect. More recently, Nedialkov and Jackson's interval Hermite-Obreschkoff method [24] improved on AWA by extending a Hermite–Obreschkoff's approach (which can be viewed as a generalized Taylor method) to intervals. Another recent approach, the Taylor models, was proposed by Berz and Makino [4] for reducing the wrapping effect. Their scheme validates existence and uniqueness and also computes tight enclosures of the solution in one process, contrary to the other methods mentioned above.

The research described in this work takes a constraint satisfaction approach to ODEs. Its basic idea [7, 13, 14] is to view the solving of ODEs as the iteration of three processes: (1) a *bounding box* process, (2) a *predictor* process that computes initial enclosures at given times from enclosures at previous times and bounding boxes, and (3) a *pruning* process that reduces the initial enclosures without removing solutions.[2] The real novelty in our approach is the pruning component. It is based on the construction of a nontrivial constraint from a *relaxation* of the ODE, a key concept in constraint satisfaction [32]. This constraint can then be used to prune the solution space at the various integration points.

The main contribution of this work is to show that an effective pruning technique can be derived from a relaxation of the ODE, importing a fundamental principle from constraint satisfaction into the field of validated differential equations. Four main steps are necessary to derive an effective pruning algorithm.

1. The first step consists of obtaining a relaxation of the ODE by safely approximating its solution using Hermite interpolation polynomials.
2. The second step consists of using the mean-value form of this relaxation for more accuracy and efficiency. Unfortunately, these two steps, which were sketched in [13], are not sufficient, and the resulting pruning algorithm still suffers from traditional problems of interval methods.
3. The third fundamental step [14] consists of globalizing the pruning by considering several successive relaxations together. This idea of generating a global constraint from a set of more primitive constraints is also at the heart of constraint satisfaction. It makes it possible, in this new context, to address the problem of dependencies (and hence the accumulation of errors) and the wrapping effect simultaneously.[3]
4. The fourth and final step consists of finding an evaluation time for the relax-

[2] Observe that interval extensions of predictor/corrector methods (e.g., [24]) can also be viewed as the composition of a predictor and a pruning step.

[3] Global constraints in ODEs have also been found useful in [6]. The problem and the techniques in [6] are, however, fundamentally different.

ation which minimizes the local error of the relaxation. Indeed, the global constraint generated in the third step, being a relaxation of the ODE, is parametrized by an evaluation time. Interestingly, for global filters based on Hermite interpolation polynomials, the (asymptotically) optimal evaluation time is independent from the ODE and induces negligible overhead on the computational cost of the methods.

Theoretical and experimental results show the benefits of the approach. From a theoretical standpoint, the constraint satisfaction approach provides a quadratic improvement in accuracy (asymptotically) over the best interval method we know of for the same computation costs. The theoretical results also show that our approach should be significantly faster for a given precision when the ODE contains many operations. Experimental results, obtained from an object-oriented implementation of our algorithms, confirm the theory. They show that the constraint satisfaction approach often produces significant improvements in accuracy over existing methods for the same computation costs and should produce significant gain in computation times when the ODE contains many operations. Of particular interest is the versatility of the approach which can be tailored to the problem at hand.

The rest of the paper is organized as follows. Section 2 introduces the main definitions and notations. Section 3 gives a high-level overview of the constraint satisfaction approach to parametric ODEs. The next four sections are the core of the paper. Section 4 introduces multistep filters, section 5 presents multistep Hermite filters as a special case of multistep filters, section 6 describes how to choose an evaluation time to minimize the local error of a multistep Hermite filter, and section 7 presents the overall algorithm. Sections 8 and 9 report the theoretical and experimental analyses, and section 10 concludes the paper.

2. Background and definitions.

2.1. Basic notational conventions. Small letters denote real values, vectors, and functions of real values. Capital letters denote matrices, sets, intervals, vectors, and functions of intervals. A vector of intervals $D \in \mathbb{IR}^n$ is called a *box*. If $A \subseteq \mathbb{R}^n$, then $\Box A$ denotes the smallest box $D \in \mathbb{IR}^n$ such that $A \subseteq D$, and $g(A)$ denotes the set $\{g(x) \mid x \in A\}$. If M is a regular (point or interval) matrix, then M^{-1} denotes an *enclosure*[4] of the inverse of M. A relation is a function $r : \mathbb{R}^n \to \mathit{Bool}$, where *Bool* denotes the booleans. We also assume that t_i, t_e, and t are reals, u_i is in $\mathbb{R}^n$, and D_i and B_i are in $\mathbb{IR}^n$ ($i \in \mathbb{N}$). We use $m(D)$ to denote the midpoint of D and $s(D)$ to denote $D - m(D)$. Observe that $m(D) + s(D) = D$. We use $\omega(D)$ to denote the width of a box. More precisely, $\omega([a,b]) = b - a$ and $\omega((I_1, \ldots, I_n)) = (\omega(I_1), \ldots, \omega(I_n))$ if $I_i \in \mathbb{IR}$. If $g : \mathbb{R}^m \to \mathbb{R}^n$, $x = (x_1, \ldots, x_m)$ and $\tilde{x} = (x_{i_1}, \ldots, x_{i_p})$ with $i_1, \ldots, i_p \in 1..m$, then $\mathcal{J}_{\tilde{x}} g(x)$ denotes the Jacobian matrix

$$\begin{bmatrix} \frac{\partial g_1}{\partial x_{i_1}}(x) & \cdots & \frac{\partial g_1}{\partial x_{i_p}}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_{i_1}}(x) & \cdots & \frac{\partial g_n}{\partial x_{i_p}}(x) \end{bmatrix}.$$

In particular, we write $\mathcal{J} g(x) = \mathcal{J}_x g(x)$ (differentiation w.r.t. all variables of g). If not specified, n denotes the dimension of the ODE (i.e., the number of scalar equations), $h > 0$ denotes the step size of the integration, and k denotes the number of previous

[4] By *enclosure* of a set A, we mean a set containing A.

values of the solution at times $t_0, \dots, t_{k-1}$ used to compute the new value at time t_k (k-step approach).

NOTATION 1 (boldface notations). *Let A be a set and $a_i \in A$, where $i \in \mathbb{N}$. We use the following boldface notations.*

$$\begin{array}{lcl} \mathbf{a} & = & (a_0, \dots, a_k) \in A^{k+1}, \\ \mathbf{a}_i & = & (a_{ik}, \dots, a_{(i+1)k-1}) \in A^k, \\ \mathbf{a}_{i..i+j} & = & (a_i, \dots, a_{i+j}) \in A^{j+1}. \end{array}$$

Observe that $\mathbf{a}_0 = (a_0, \dots, a_{k-1})$, $\mathbf{a}_1 = (a_k, \dots, a_{2k-1})$, and $\mathbf{a} = (a_0, \dots, a_k)$. The following asymptotical notations are standard.

NOTATION 2 (asymptotical notations). *Consider two functions $f, g : \mathbb{R} \to \mathbb{R}$ and let $x > 0$. We use the following standard notations:*

$$f(x) = \begin{cases} O(g(x)) & if \quad \exists c > 0, \exists \varepsilon > 0 : x \geq \varepsilon \Rightarrow |f(x)| \leq c|g(x)|, \\ \mathcal{O}(g(x)) & if \quad \exists c > 0, \exists \varepsilon > 0 : x \leq \varepsilon \Rightarrow |f(x)| \leq c|g(x)|, \\ \Omega(g(x)) & if \quad \exists c > 0, \exists \varepsilon > 0 : x \leq \varepsilon \Rightarrow |f(x)| \geq c|g(x)|, \\ \Theta(g(x)) & if \quad f(x) = \mathcal{O}(g(x)) \quad and \quad f(x) = \Omega(g(x)). \end{cases}$$

The notations extend componentwise for vectors and matrices of functions.

Finally we assume that the underlying interval arithmetic is exact for the theoretical parts of this work (i.e., there are no rounding errors). The implementation of course uses outwardly directed rounding.

2.2. Basic definitions. As is traditional, when we consider an ODE $u' = f(u)$ and an interval of integration T, we assume $f \in C^r(\Omega)$, where r is sufficiently large and Ω is an open set such that $T \times \Omega$ contains the trajectories of the solutions on T.[5] In addition, we restrict our attention to ODEs that have a unique solution for a given initial value. Techniques to verify this hypothesis numerically are well known [25, 21, 22, 5, 23]. In order to make the dependence on the initial condition (t_0, u_0) explicit, we introduce the following definition of the solution to an ODE.

DEFINITION 1 (solution of an ODE). *Let $\Lambda \subseteq \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ be an open set. The* solution *of an ODE $u' = f(u)$ is the function $s : \Lambda \to \mathbb{R}^n$ such that*

$$\forall (t_0, u_0, t) \in \Lambda : \begin{cases} \frac{\partial s}{\partial t}(t_0, u_0, t) = f(s(t_0, u_0, t)), \\ s(t_0, u_0, t_0) = u_0. \end{cases}$$

Observe that, since we restrict attention to autonomous systems in this work, we can write

$$s(t_0, x, t) = s(0, x, \tau),$$

where $\tau = t - t_0$, and thus

$$\frac{\partial^j s}{\partial t^j}(t_0, x, t) = \frac{\partial^j s}{\partial \tau^j}(0, x, \tau).$$

In particular, when $t = t_0$, the function

$$\frac{\partial^j s}{\partial t^j}(t_0, x, t)\bigg|_{(t_0, x, t_0)} = \frac{\partial^j s}{\partial \tau^j}(t_0, x, \tau)\bigg|_{(0, x, 0)}$$

[5] The standard mathematical symbol $C^r(\Omega)$ denotes the set of all functions whose rth derivative exists and is continuous on Ω.

depends *only* on x. This justifies the following notation, which captures the notions of real and interval Taylor coefficients of the solution of an ODE as well as their Jacobians.

NOTATION 3 (Taylor coefficients and Jacobians). *Let s be the solution of an ODE $\mathbb{O}$, $x \in \mathbb{R}^n$, $D \in \mathbb{IR}^n$, and let t_0 be any real number. Then*

1. $(x)_j = \frac{1}{j!}\frac{\partial^j s}{\partial t^j}(t_0, x, t)\Big|_{(t_0,x,t_0)}$;
2. $\{(x)_j \mid x \in D\} \subseteq (D)_j \in \mathbb{IR}^n$;
3. $\mathcal{J}(x)_j = \mathcal{J}_x \frac{1}{j!}\frac{\partial^j s}{\partial t^j}(t_0, x, t)\Big|_{(t_0,x,t_0)}$;
4. $\{\mathcal{J}(x)_j \mid x \in D\} \subseteq \mathcal{J}(D)_j \in \mathbb{IR}^{n\times n}$;
5. $(x)_{j,l}$, $(D)_{j,l}$, $\mathcal{J}(x)_{j,l}$, *and* $\mathcal{J}(D)_{j,l}$ *denote, respectively, the lth component of* $(x)_j$, $(D)_j$, $\mathcal{J}(x)_j$, *and* $\mathcal{J}(D)_j$.

In the context of our multistep approach (to be presented in section 3), it is useful to generalize Definition 1 in order to make the dependence on the last $k+1$ redundant conditions $(t_0, u_0), \ldots, (t_k, u_k)$ explicit.

DEFINITION 2 (multistep solution of an ODE). *Let s be the solution of an ODE $\mathbb{O}$. The* multistep solution *of $\mathbb{O}$ is the* partial *function $ms : A \subseteq \mathbb{R}^{k+1} \times (\mathbb{R}^n)^{k+1} \times \mathbb{R} \to \mathbb{R}^n$:*

$$ms(\mathbf{t}, \mathbf{u}, t) = \begin{cases} s(t_0, u_0, t) \text{ if } u_i = s(t_0, u_0, t_i),\ 1 \le i \le k, \\ \textit{undefined otherwise.} \end{cases}$$

Since we are dealing with interval methods, we need to introduce the notions of interval extensions of a function and a relation. These notions were introduced in [31]. However, because the techniques proposed in this work use multistep solutions, which are *partial* functions, it is necessary to generalize the notion of interval extension to partial functions and relations.

DEFINITION 3 (interval extension of a partial function). *The interval function $G : \mathbb{IR}^n \to \mathbb{IR}^m$ is an* interval extension *of the partial function $g : E \subseteq \mathbb{R}^n \to \mathbb{R}^m$ if*

$$\forall D \in \mathbb{IR}^n : g(E \cap D) \subseteq G(D).$$

DEFINITION 4 (interval extension of a partial relation). *The interval relation $R : \mathbb{IR}^n \to Bool$ is an* interval extension *of the partial relation $r : E \subseteq \mathbb{R}^n \to Bool$ if*

$$\forall D \in \mathbb{IR}^n : (\exists x \in E \cap D : r(x)) \Rightarrow R(D).$$

Finally, we generalize the concept of bounding boxes, a fundamental concept in interval methods for ODEs, to multistep methods. Intuitively, a bounding box encloses all solutions of an ODE going through certain boxes at given times over a given time interval. Bounding boxes are needed to enclose error terms in validated methods for ODEs (see section 5).

DEFINITION 5 (bounding box). *Let $\mathbb{O}$ be an ODE system, ms be the multistep solution of $\mathbb{O}$, and $\{t_0, \ldots, t_k\} \subseteq T \in \mathbb{IR}$. A box B is a* bounding box *of $\mathbb{O}$ over T wrt $(\mathbf{t}, \mathbf{D})$ if, for all $t \in T$, $ms(\mathbf{t}, \mathbf{D}, t) \subseteq B$.*

2.3. The midpoint technique. The midpoint technique is a standard tool in interval computation. It consists of decomposing a matrix A as the sum of its midpoint matrix and the remainder matrix composed of symmetric intervals:

$$A = m(A) + s(A).$$

In this paper, the midpoint technique is used in the following two cases:

1. enclosing a set of real matrix-matrix-vector products (see sections 4.4 and 4.5);
2. converting an implicit interval linear system into an explicit one by matrix inversion (see section 4.2).

Assume that we are interested in enclosing the set

$$P = \{\mathcal{A}\mathcal{B}d \mid \mathcal{A} \in A,\ \mathcal{B} \in B,\ d \in D\},$$

where A, B are interval matrices and D is an interval vector. Assume also that $\omega(A)$ is small and that the wrapping effect in the product CD, where $C = AB$, is small. A straightforward and cheap way to enclose the set P consists of computing the product $A(BD)$. In general, this product does not yield accurate results because of the wrapping effect which occurs in the product $E = BD$ and in the product AE. Another straightforward way of enclosing the set P is to compute the product $(AB)D$. By hypothesis, the wrapping effect is small in this case, and the product is an accurate enclosure of P. However, the multiplication of the two interval matrices A and B is a costly process (due to costly sign tests and rounding mode switches in modern RISC architectures; see [15] for more details). In order to avoid this product, we apply the midpoint technique on A. By distribution and rearrangement of the parentheses, we can write

$$P \subseteq Q = (m(A)B)D + s(A)(BD). \tag{1}$$

It is interesting to observe that no multiplication between two interval matrices occurs in Q. (Note the importance of the parentheses!) From an accuracy standpoint, the wrapping effect in $(m(A)B)D$ is small (by hypothesis) and the remainder term $s(A)(BD)$ is small (because $\omega(A)$ is small). Hence, Q is an accurate enclosure of the set P which avoids the costly multiplication of two interval matrices.

Now consider the implicit interval linear system

$$\begin{aligned} &A_0X_0 + A_1X_1 = B, \\ &X_0 \subseteq D_0,\ X_1 \subseteq D_1, \end{aligned} \tag{2}$$

where A_0, A_1 are interval matrices and B, D_0, D_1 are interval vectors. We assume that A_0 contains no singular point matrix. The exact solution set to this system is given by

$$S = \{(x_0, x_1) \in (D_0, D_1) \mid \exists \mathcal{A}_0 \in A_0,\ \exists \mathcal{A}_1 \in A_1,\ \exists b \in B : \mathcal{A}_0x_0 + \mathcal{A}_1x_1 = b\}.$$

We are interested in converting the system (2) into a system

$$X_0 = CX_1 + E,$$

which is explicit in the variable X_0 and such that

$$S \subseteq \{(x_0, x_1) \in (D_0, D_1) \mid \exists \mathcal{C} \in C,\ \exists e \in E : x_0 = \mathcal{C}x_1 + e\}.$$

A straightforward solution consists of computing an enclosure A_0^{-1} of the inverse of A_0, multipling both sides of (2) by A_0^{-1}, and rearranging the parentheses:

$$X_0 = -(A_0^{-1}A_1)X_1 + A_0^{-1}B. \tag{3}$$

However, the system (3) suffers from two drawbacks:

- We have to invert the interval matrix A_0. Computing an accurate enclosure of the inverse of an interval matrix is a costly process [23]).
- We have to multiply the two interval matrices A_0^{-1} and A_1.

To eliminate these operations, we apply the midpoint technique both on A_0 and A_1 in (2). By distributivity, we have

$$m(A_0)X_0 = -m(A_1)X_1 + B - s(A_0)X_0 - s(A_1)X_1. \tag{4}$$

Since $X_0 \subseteq D_0$ and $X_1 \subseteq D_1$, we can replace X_0 by D_0 in the term involving $s(A_0)$ and X_1 by D_1 in the term involving $s(A_1)$:

$$m(A_0)X_0 = -m(A_1)X_1 + B - s(A_0)D_0 - s(A_1)D_1. \tag{5}$$

Note that it is important to have precise enclosures D_1 and D_2. To obtain a system which is explicit in the variable X_0, we compute an *enclosure* $m(A_0)^{-1}$ of the inverse of the *point* matrix $m(A_0)$, we multiply both sides of (5) by $m(A_0)^{-1}$, and we rearrange the parentheses:[6]

$$X_0 = -(m(A_0)^{-1}m(A_1))X_1 + m(A_0)^{-1}(B - s(A_0)D_0 - s(A_1)D_1).$$

Observe that, in this last system, there is no interval matrix inversion and no product of two interval matrices.

3. The constraint satisfaction approach. The constraint satisfaction approach followed in this work was first presented in [7]. It consists of a generic algorithm for ODEs that iterates three processes:

1. a *bounding box* process that computes bounding boxes for the current step and proves (numerically) the existence and uniqueness of the solution;
2. a *predictor* process that computes initial enclosures at given times from enclosures at previous times and bounding boxes;
3. a *pruning* process that reduces the initial enclosures without removing solutions.

The intuition of the successive steps is illustrated in Figure 1. Bounding box and predictor components are standard in interval methods for ODEs. This paper thus focuses on the pruning process, the main novelty of the approach. *Our pruning component is based on relaxations of the ODE, a fundamental concept in the field of constraint satisfaction.* To our knowledge, no other approach uses relaxations of the ODE to derive pruning operators, and the only other approaches using a pruning component [24, 28] were developed independently. Note also that, in the following, predicted boxes are generally superscripted with the symbol $-$ (e.g., D_1^-), while pruned boxes are generally superscripted with the symbol $*$ (e.g., D_1^*).

The pruning component uses *safe approximations* of the ODE to shrink the boxes computed by the predictor process. To understand this idea, it is useful to contrast the constraint satisfaction to nonlinear programming [30, 31] and to ODEs. In nonlinear programming, a constraint $c(x_1, \ldots, x_n)$ can be used almost directly for pruning the search space (i.e., the Cartesian product of the intervals I_i associated with the variables x_i). It suffices to take an interval extension $C(X_1, \ldots, X_n)$ of the constraint. Now if $C(I_1', \ldots, I_n')$ does not hold, it follows, by the definition of interval extensions, that no solution of c lies in $I_1' \times \cdots \times I_n'$. The interval extension can be seen as a filter

[6] Note that, even though $m(A_0)$ is a *point* matrix, the enclosure $m(A_0)^{-1}$ of its inverse is generally *not* a point matrix, because of rounding errors.

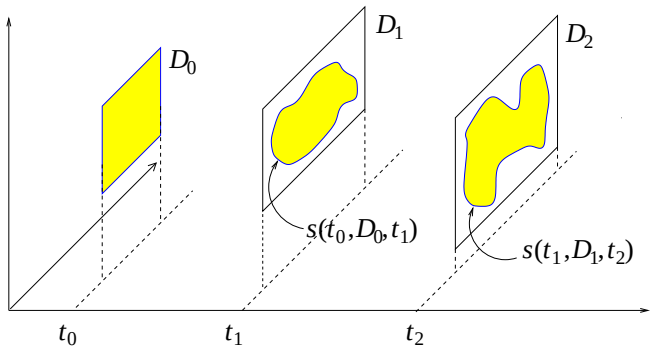


FIG. 1. *Successive integration steps.*

that can be used for pruning the search space in many ways. For instance, Numerica uses box(k)-consistency on these interval constraints [31]. ODEs raise new challenges. In an ODE for all $t:\ u' = f(u)$, functions u and u' are, of course, unknown. Hence, it is not obvious how to obtain a filter to prune boxes.

One of the main contributions of our approach is to show how to derive effective pruning operators for parametric ODEs. The first step consists of rewriting the ODE for all $t:\ u' = f(u)$ in terms of its multistep solution ms to obtain

$$\forall\, t:\ \frac{\partial ms}{\partial t}(\mathbf{t}, \mathbf{u}, t) = f(ms(\mathbf{t}, \mathbf{u}, t)). \tag{6}$$

Let us denote this relation for all $t : fl(\mathbf{t}, \mathbf{u}, t)$. This rewriting may not appear useful since ms is still an unknown function. However, it suggests a way to approximate the ODE. Indeed, we show in section 5 how to obtain interval extensions of ms and $\frac{\partial ms}{\partial t}$ by using Hermite polynomial interpolations together with their error terms. This simply requires a bounding box for the considered time interval and safe approximations of ms at successive times, both of which are available from the bounding box and predictor processes. Once these interval extensions are available, it is possible to obtain an interval relation of the form

$$\forall\, t : FL(\mathbf{t}, \mathbf{D}, t), \tag{7}$$

which approximates the original ODE *safely*; i.e., if $FL(\mathbf{t}, \mathbf{D}, t)$ does not hold for a time t, it follows that no solution of the ODE can go through boxes $D_0, \ldots, D_k$ at times $t_0, \ldots, t_k$. Relation (7) is still not ready to be used as a filter because t is universally quantified. The solution here is simpler and consists of restricting attention to a finite set T of times (possibly a singleton) to obtain the relation

$$\forall\, t \ \in\ T : FL(\mathbf{t}, \mathbf{D}, t),$$

which produces a computable filter. The relation FL is a *relaxation* of the ODE (6) in a constraint satisfaction sense [32]; i.e., given a time t, it produces a relation that can be used to prune the domain of the variables. The so-obtained relation is in fact a conservative approximation of the actual ODE at the given time. The following definition and proposition capture these concepts more formally.

DEFINITION 6 (multistep filter). *Let $\mathbb{O}$ be an ODE and s its solution. A* multistep filter *for $\mathbb{O}$ is an interval relation $FL : \mathbb{R}^{k+1} \times (\mathbb{IR}^n)^{k+1} \times \mathbb{R} \to Bool$ satisfying*

$$\left.\begin{array}{c} u_i \ \in\ D_i \\ s(t_0, u_0, t_i) = u_i\ (0 \le i \le k) \end{array}\right\} \Rightarrow\ \ \forall t:\ FL(\mathbf{t}, \mathbf{D}, t).$$

The variable t is called the evaluation time *of the multistep filter.*

PROPOSITION 1 (soundness of multistep filters). *Let $\mathbb{O}$ be an ODE, and let FL be a multistep filter for $\mathbb{O}$. If $FL(\mathbf{t}, \mathbf{D}, t)$ does not hold for some t, then there exists no solution of $\mathbb{O}$ going through $\mathbf{D}$ at times $\mathbf{t}$.*

How can we use this filter to obtain tighter enclosures of the solution? A simple technique consists of pruning the last box computed by the predictor process. Assume that D_i^* is a box enclosing the solution at time t_i $(0 \leq i < k)$ and that we are interested in pruning the last predicted box D_k^-. A subbox $D \subseteq D_k^-$ can be pruned away if the condition

$$FL(\mathbf{t}, (D_0^*, \ldots, D_{k-1}^*, D), t_e)$$

does not hold for some evaluation point t_e. Let us explain briefly the geometric intuition behind this relation by considering what we call *natural filters*. Given interval extensions *MS*, *DMS*, and F, respectively, of ms, $\frac{\partial ms}{\partial t}$, and f, it is possible to approximate the ODE $u' = f(u)$ by the relation

$$DMS(\mathbf{t}, \mathbf{D}, t) = F(MS(\mathbf{t}, \mathbf{D}, t)).$$

In this relation, the left-hand side of the equation represents *the approximation of the slope of u* while the right-hand side represents *the slope of the approximation of u*. Since the approximations are conservative, these two sides must intersect on boxes containing a solution. Hence an empty intersection means that the boxes used in the relation do not contain the solution to the ODE system. Figure 2 illustrates the intuition. It is generated from an actual ODE, considers only points instead of intervals, uses an interpolation polynomial as an approximation of u, and ignores error terms for simplicity. It illustrates how this technique can prune away a value as a potential solution at a given time. In the figure, we consider the solution to the equation that evaluates to u_0 and u_1 at t_0 and t_1, respectively. Two possible points u_2 and u_2' are then considered as possible values at t_2. The curve marked `KO` describes an interpolation polynomial going through u_0, u_1, u_2' at times t_0, t_1, t_2. To determine if u_2' is the value of the solution at time t_2, the idea is to test if the equation is satisfied at time t_e. (We will say more about how to choose t_e later in this paper.) As can be easily seen, the slope of the interpolation polynomial is different from the slope specified by f at time t_e, and hence u_2' cannot be the value of the solution at t_2 since we assume that the values u_0 and u_1 were correct at t_0 and t_1. The curve marked `OK` describes an interpolation polynomial going through u_0, u_1, u_2 at times t_0, t_1, t_2. In this case, the equation is satisfied at time t_e, which means that u_2 cannot be pruned away.

The filter proposed earlier generalizes this intuition to boxes. Both the left- and right-hand sides represent sets of slopes, and the filter fails when their intersection is empty. Traditional consistency techniques and algorithms based on this filter can now be applied. For instance, one may be interested in updating the last box computed by the predictor process using the operator

$$D_k^* = \Box\{r \in D_k^- \mid FL(\mathbf{t}, (D_0^*, \ldots, D_{k-1}^*, r), t_e)\},$$

which is defined in terms of an evaluation time t_e. *One of the main results of this paper consists of showing that t_e can be chosen optimally (in an asymptotic sense) to maximize pruning.* The following definition is a novel notion of consistency for ODEs to capture pruning of the last r boxes.[7]

[7] We will give an explicit form for D_k^* later in the paper.

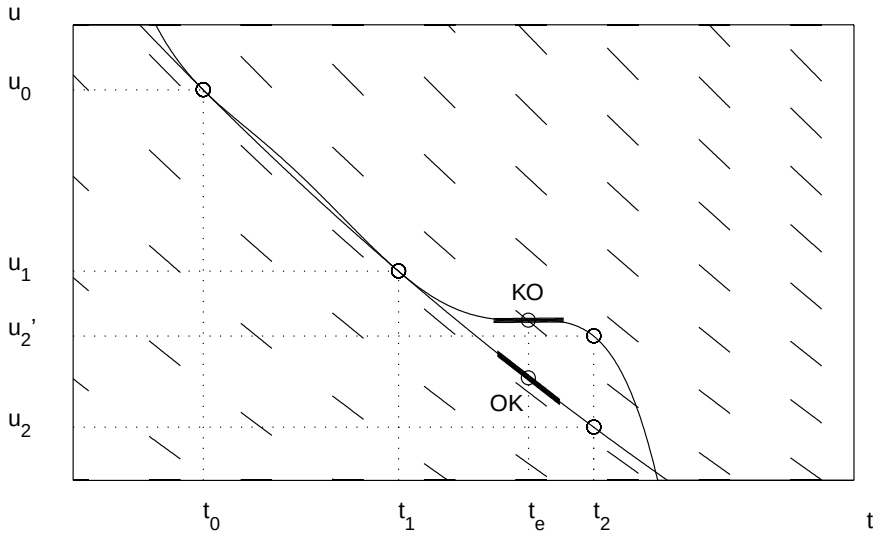


FIG. 2. *Geometric intuition of the multistep filter.*

DEFINITION 7 (backward consistency of multistep filters). *A multistep filter FL is* backward-consistent *in* $(\mathbf{t}, \mathbf{D})$ *for time* e *if*

$$D_k = \square \{u_k \in D_k \mid \exists \mathbf{u}_0 \in \mathbf{D}_0 : FL(\mathbf{t}, \mathbf{u}, e)\}.$$

A system of r *successive multistep filters* $\{FL_i\}_{0 \le i < r}$ *is* backward(r)-consistent *in* $(\mathbf{t}_{0..k+r-1}, \mathbf{D}_{0..k+r-1})$ *for time vector* $(e_0, \ldots, e_{r-1})$ *if*

$$\begin{aligned} \mathbf{D}_{k..k+r-1} = \square\{&\mathbf{u}_{k..k+r-1} \in \mathbf{D}_{k..k+r-1} \mid \exists \mathbf{u}_0 \in \mathbf{D}_0 : \\ &\forall\ 0 \le i < r : FL_i(\mathbf{t}_{i..k+i}, \mathbf{u}_{i..k+i}, e_i)\}. \end{aligned} \tag{8}$$

Informally speaking, the parameter r in the definition determines the strength of the consistency, i.e., the number of backward variables each variable depends on. The following proposition is an immediate consequence of Definition 7. It states that the strength of the consistency increases with parameter r.

PROPOSITION 2 (property of backward consistency). *If a system of* $r+1$ $(r > 0)$ *successive multistep filters* $\{FL_i\}_{0 \le i \le r}$ *is backward*$(r+1)$*-consistent in* $(\mathbf{t}_{0..k+r}, \mathbf{D}_{0..k+r})$ *for time vector* $(e_0, \ldots, e_r)$*, then the system*

1. $\{FL_i\}_{0 \le i < r}$ *is backward*(r)*-consistent in* $(\mathbf{t}_{0..k+r-1}, \mathbf{D}_{0..k+r-1})$ *for time vector* $(e_0, \ldots, e_{r-1})$*;*
2. $\{FL_i\}_{1 \le i \le r}$ *is backward*(r)*-consistent in* $(\mathbf{t}_{1..k+r}, \mathbf{D}_{1..k+r})$ *for time vector* $(e_1, \ldots, e_r)$*.*

In the next section, we introduce coordinate transformations in multistep filters to represent the sets of solutions compactly, i.e., to handle the wrapping effect (see section 4.5). It is thus useful to generalize the above definition by introducing affine transformations.

DEFINITION 8 (generalized backward consistency). *Let* $Y_i \in \mathbb{IR}^n$ $(i \in \mathbb{N})$*. A multistep filter FL is* backward-consistent *in* $(\mathbf{t}, \mathbf{Y})$ *for time* e *if there exists an invertible*

affine transformation $\mathbf{a} : \mathbb{R}^{n(k+1)} \rightarrow \mathbb{R}^{n(k+1)}$ *such that*

$$Y_k = \Box\{y_k \in Y_k \mid \exists \mathbf{y}_0 \in \mathbf{Y}_0 : FL(\mathbf{t}, \mathbf{a}(\mathbf{y}), e)\}.$$

A system of r *successive multistep filters* $\{FL_i\}_{0 \leq i < r}$ *is* backward(r)-consistent *in* $(\mathbf{t}_{0..k+r-1}, \mathbf{Y}_{0..k+r-1})$ *for time vector* $(e_0, \ldots, e_{r-1})$ *if there exists an invertible affine transformation* $\mathbf{a}_{0..k+r-1} : \mathbb{R}^{n(k+r)} \rightarrow \mathbb{R}^{n(k+r)}$ *such that*

$$\begin{aligned} \mathbf{Y}_{k..k+r-1} = \Box\{&\mathbf{y}_{k..k+r-1} \in \mathbf{Y}_{k..k+r-1} \mid \exists \mathbf{y}_0 \in \mathbf{Y}_0 : \\ &\forall 0 \leq i < r : FL_i(\mathbf{t}_{i..k+i}, \mathbf{a}_{i..k+i}(\mathbf{y}_{0..k+r-1}), e_i)\}. \end{aligned} \tag{9}$$

Note that Proposition 2 also holds for generalized backward consistency. In the rest of this paper, we use "backward consistency" instead of "generalized backward consistency" for simplicity. The algorithm used in our computational results enforces backward(k)-consistency of a system of k filters we now describe.

4. Multistep filters. Filters rely on interval extensions of the multistep solution and of its derivative w.r.t. t. These extensions are, in general, based on decomposing the (unknown) multistep solution into the sum of a computable approximation p and an (unknown) error term e, i.e.,

$$ms(\mathbf{t}, \mathbf{u}, t) = p(\mathbf{t}, \mathbf{u}, t) + e(\mathbf{t}, \mathbf{u}, t). \tag{10}$$

There exist standard techniques to build p and $\frac{\partial p}{\partial t}$ and to bound e and $\frac{\partial e}{\partial t}$. Section 5 reviews how they can be derived from Hermite interpolation polynomials. Here we simply assume that they are available, and we show how to use them to build filters.

4.1. Natural filters. Section 3 explained how natural multistep filters can be obtained by simply replacing the multistep solution *ms*, its derivative $\frac{\partial ms}{\partial t}$, and the function f by their interval extensions *MS*, *DMS*, and F to obtain

$$DMS(\mathbf{t}, \mathbf{D}, t) = F(MS(\mathbf{t}, \mathbf{D}, t)).$$

It is not easy, however, to enforce backward consistency on a natural filter since the variables may occur in complex nonlinear expressions. This problem is addressed by mean-value filters that we now study.

4.2. Mean-value filters.

Mean-value forms. Mean-value forms (MVFs) play a fundamental role in interval computations and are derived from the mean-value theorem. They correspond to problem linearizations around a point and result in filters that are systems of linear equations with interval coefficients and whose solutions can be enclosed reasonably efficiently. MVFs are effective when the sizes of the boxes are sufficiently small, which is the case in ODEs. In addition, being linear equations, they allow for an easier treatment of the so-called *wrapping effect*, a crucial problem in interval methods for ODEs to be discussed in sections 4.3 and 4.5. As a consequence, MVFs are especially appropriate in our context and will produce filters which are efficiently amenable to backward consistency. The rest of this section describes how to obtain mean-value filters.

Implicit mean-value filters. Consider the function

$$\delta(\mathbf{t},\mathbf{u},e,de,t)=\frac{\partial p}{\partial t}(\mathbf{t},\mathbf{u},t)+de-f(p(\mathbf{t},\mathbf{u},t)+e).$$

If the multistep solution ms is defined at $(\mathbf{t},\mathbf{u})$, i.e., the ODE has a solution going through $u_0,\dots,u_k$ at $t_0,\dots,t_k$, then, by (10), we have the relation

$$\delta\left(\mathbf{t},\mathbf{u},e(\mathbf{t},\mathbf{u},t),\frac{\partial e}{\partial t}(\mathbf{t},\mathbf{u},t),t\right)=0.$$

Let $\mathbf{u}^*,\mathbf{u}\in\mathbf{D}^0\in\mathbb{IR}^{n(k+1)}$, $e^*,e\in E\in\mathbb{IR}^n$, and $de^*,de\in DE\in\mathbb{IR}^n$. By the mean-value theorem, we can write $(1\le i\le n)$

$$\begin{aligned}\delta_i(\mathbf{t},\mathbf{u},e,de,t) &= \delta_i(\mathbf{t},\mathbf{u}^*,e^*,de^*,t)\\ &\quad+\mathcal{J}_{(\mathbf{u},e,de)}\delta_i(\mathbf{t},\mu_i,\xi_i,\zeta_i,t)\ (\mathbf{u}-\mathbf{u}^*,e-e^*,de-de^*)\\ &= \delta_i(\mathbf{t},\mathbf{u}^*,e^*,de^*,t)+\phi_i(\mathbf{t},\mu_i,\xi_i,t)(\mathbf{u}-\mathbf{u}^*)\\ &\quad+\psi_i(\mathbf{t},\mu_i,\xi_i,t)(e^*-e)+de_i-de_i^*,\end{aligned}$$

where

$$\begin{aligned}\phi_i(\mathbf{t},\mu_i,\xi_i,t)&=\mathcal{J}_{\mathbf{u}}\tfrac{\partial p_i}{\partial t}(\mathbf{t},\mu_i,t)-\mathcal{J}f_i(p(\mathbf{t},\mu_i,t)+\xi_i)\mathcal{J}_{\mathbf{u}}p(\mathbf{t},\mu_i,t),\\ \psi_i(\mathbf{t},\mu_i,\xi_i,t)&=\mathcal{J}f_i(p(\mathbf{t},\mu_i,t)+\xi_i)\end{aligned}$$

for some $\mu_i\in\mathbf{D}^0$, $\xi_i\in E$, and $\zeta_i\in DE$. This allows us to define a new multistep filter, which we will call an *implicit mean-value filter*. Such a filter is parametrized by the initial domain $\mathbf{D}^0$ of the variable $\mathbf{u}$.

Definition 9 (implicit mean-value filter). *An* implicit mean-value filter *for ODE* $u'=f(u)$ *in* $\mathbf{D}^0\in\mathbb{IR}^{n(k+1)}$ *is an interval relation*

(11)
$$\begin{gathered}FL(\mathbf{t},\mathbf{D},t)\Leftrightarrow\\ \delta(\mathbf{t},\mathbf{m}^0,m_e,m_{de},t)+\Delta(\mathbf{t},\mathbf{D}^0,E(\mathbf{t},\mathbf{D}^0,t),DE(\mathbf{t},\mathbf{D}^0,t),t)\ (\mathbf{X},E_m,DE_m)=0,\end{gathered}$$

where

(12)
$$\begin{aligned}&\Delta \text{ is an interval extension of the function } \mathcal{J}_{(\mathbf{u},e,de)}\delta,\\ &E \text{ and } DE \text{ are interval extensions, respectively, of } e \text{ and } \tfrac{\partial e}{\partial t},\\ &\mathbf{D}\subseteq\mathbf{D}^0,\\ &\mathbf{X}=\mathbf{D}-\mathbf{m}^0, E_m=E(\mathbf{t},\mathbf{D}^0,t)-m_e, DE_m=DE(\mathbf{t},\mathbf{D}^0,t)-m_{de},\\ &\mathbf{m}^0=m(\mathbf{D}^0), m_e=m(E(\mathbf{t},\mathbf{D}^0,t)), m_{de}=m(DE(\mathbf{t},\mathbf{D}^0,t)).\end{aligned}$$

Formula (11) is called implicit because $\mathbf{D}$ appears implicitly. The Jacobians in (12) can be computed by means of automatic differentiation tools (see, e.g., [27]). The following proposition states that an implicit mean-value filter does not eliminate any solution of the ODE. It is a direct consequence of the mean-value theorem.

Proposition 3. *An implicit mean-value filter for ODE* $\mathcal{O}$ *is a multistep filter for* $\mathcal{O}$.

Explicit mean-value filters. In general, for IVPs, we will be interested in pruning the last predicted box D_k^-. Hence, it is convenient to derive a mean-value filter which is explicit in D_k. Let $\mathbf{D}^-\in\mathbb{IR}^{n(k+1)}$ be the predicted box of variable $\mathbf{u}$ and define $\mathbf{X}$ as $\mathbf{D}-m(\mathbf{D}^-)$. An implicit mean-value filter is an interval constraint of the form

$$\Phi(t)\mathbf{X}=\Gamma(t),$$

where $\Phi(t) \in \mathbb{IR}^{n \times n(k+1)}$ and $\Gamma(t) \in \mathbb{IR}^n$. Let us apply the midpoint technique (see point 2 of section 2.3) on the matrix $\Phi(t)$. We can write $\Phi(t) = m(\Phi(t)) + s(\Phi(t))$, and

$$m(\Phi(t))\mathbf{X} = \Gamma(t) - s(\Phi(t))\mathbf{X}. \tag{13}$$

The term $s(\Phi(t))\mathbf{X}$ is normally small (of size $\mathcal{O}(\|\omega(\mathbf{D}^-)\|^2)$), and we can substitute $\mathbf{X}$ on the right side of (13) for $s(\mathbf{D}^-)$, since $\mathbf{X} = \mathbf{D} - m(\mathbf{D}^-)$ and we are looking for a pruned box $\mathbf{D}^* \subseteq \mathbf{D}^-$. We obtain the system

$$m(\Phi(t))\mathbf{X} = \Gamma(t) - s(\Phi(t))s(\mathbf{D}^-). \tag{14}$$

Equation (14) can be rewritten as

$$\sum_{i=0}^{k} A_i(t)X_i = K(t),$$

where $A_i(t) \in \mathbb{R}^{n \times n}$, $i = 0, \ldots, k$, and $K(t) \in \mathbb{IR}^n$. Let us isolate the term involving X_k:

$$A_k(t)X_k = K(t) - \sum_{i=0}^{k-1} A_i(t)X_i. \tag{15}$$

Multiplying both sides of (15) by $A_k(t)^{-1}$ (recall that $A_k(t)^{-1}$ denotes an enclosure of the inverse of $A_k(t)$) gives

$$X_k = A_k(t)^{-1}K(t) - \sum_{i=0}^{k-1} \left(A_k(t)^{-1}A_i(t)\right) X_i.$$

We are now in position to define explicit mean-value filters which play a fundamental role in our approach.

Definition 10 (explicit mean-value filter). *An* explicit mean-value filter *for ODE $\mathcal{O}$ in $\mathbf{D}^0 \in \mathbb{IR}^{n(k+1)}$ is an interval relation*

$$FL(\mathbf{t}, \mathbf{D}, t) \;\Leftrightarrow\; X_k = A_k(t)^{-1}K(t) - \sum_{i=0}^{k-1} \left(A_k(t)^{-1}A_i(t)\right) X_i,$$

where

$\mathbf{X} = \mathbf{D} - m(\mathbf{D}^0)$,
$\mathbf{D} \subseteq \mathbf{D}^0$,
$(A_0(t) \cdots A_k(t)) = m(\Phi(t)) \in \mathbb{R}^{n \times n(k+1)}$,
$K(t) = \Gamma(t) - s(\Phi(t))s(\mathbf{D}^0) \in \mathbb{IR}^n$,
the relation $\Phi(t)\mathbf{X} = \Gamma(t)$ is an implicit mean-value filter for $\mathcal{O}$ in $\mathbf{D}^0$.

Proposition 4. *An explicit mean-value filter for ODE $\mathcal{O}$ is a multistep filter for $\mathcal{O}$.*

It is easy to use an explicit mean-value filter to prune the predicted box D_k^- at time t_k given the boxes $D_0^*, \ldots, D_{k-1}^*$ from the previous integration steps, since X_k (and thus D_k) has been isolated. The filter simply becomes

$$D_k = m(D_k^-) + A_k(t)^{-1}K(t) - \sum_{i=0}^{k-1} \left(A_k(t)^{-1}A_i(t)\right) (D_i^* - m(D_i^*)), \tag{16}$$

and the pruned box D_k^* at time t_k is given by

$$D_k^* = D_k \cap D_k^-.$$

It follows directly that the explicit mean-value filter is backward-consistent in $\mathbf{D}^*$.

4.3. Problems in mean-value filters. Mean-value filters often produce significant pruning of the boxes computed by the predictor process. However, they suffer from two limitations: the *wrapping effect* which is inherent in interval analysis and a *variable dependency* problem induced by the use of a multistep method. We review both of these before describing how to address them.

Wrapping effect. The wrapping effect is the name given to the overestimation that arises from enclosing a set by a box. In the context of ODEs, the set of solutions at each integration step is overapproximated by a box. These overapproximations accumulate step after step and may result in an explosion in the sizes of the computed boxes. The standard solution used in interval methods for ODEs to obtain tighter solution bounds is to choose, at each step, an appropriate local coordinate system to represent the solutions compactly (see [20, 24]). How does the wrapping effect occur in our context? Let us rewrite an explicit mean-value filter from (16) as

$$X_k = K(t) + \sum_{i=0}^{k-1} A_i(t) X_i,$$

and let us assume that $A_0(t), \ldots, A_{k-1}(t)$ are point matrices and that $K(t)$ is a point vector. Given the boxes $X_0, \ldots, X_{k-1}$ computed at the previous steps, the exact solution set to be enclosed by X_k is

$$Z = \left\{ K(t) + \sum_{i=0}^{k-1} A_i(t) x_i \mid (x_0, \ldots, x_{k-1}) \in (X_0, \ldots, X_{k-1}) \right\}.$$

The set Z is called a *zonotope*[8] (i.e., a generalization of a parallelepiped). Figure 3(a) illustrates a zonotope in $\mathbb{R}^2$ (for $k = 3$) and its smallest enclosing box. As can be seen, the box significantly overestimates the zonotope. Figure 3(b) shows that the zonotope can be enclosed much more tightly by using a coordinate transformation. It should be mentioned, however, that finding a good coordinate system is not necessarily a trivial task (e.g., one idea is to find approximations of the main directions of the zonotope) and may not be sufficient because of the variable dependency problem that we now discuss.

Variable dependencies in explicit filters. Consider the application of an explicit mean-value filter at two successive time steps with respective evaluation times e_0 and e_1. We obtain equations of the form

$$\begin{aligned} X_k &= K_0(e_0) + A_{0,0}(e_0) X_0 + \cdots + A_{0,k-1}(e_0) X_{k-1}, \\ X_{k+1} &= K_1(e_1) + A_{1,0}(e_1) X_1 + \cdots + A_{1,k-1}(e_1) X_k. \end{aligned}$$

The second equation computes the box X_{k+1} assuming that $X_1, \ldots, X_k$ are independent, which is not the case because of the first equation. Hence, the dependencies between $X_1, \ldots, X_k$ are lost when moving from the first to the second time step.

[8] Note that Kühn uses zonotopes in another context, i.e., as compact enclosures of solutions [17, 18, 19].

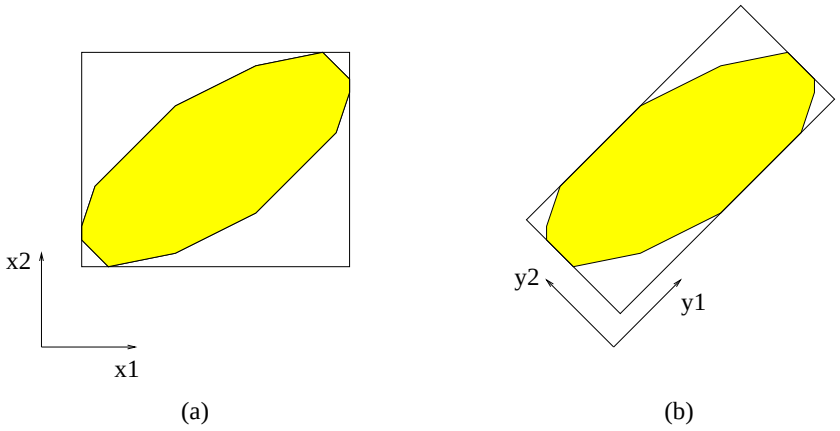


FIG. 3. (a) *A zonotope in* $\mathbb{R}^2$ *and the smallest enclosing box.* (b) *Coordinate transformation where the enclosing box better fits the zonotope.*

The variable dependency problem arises because successive explicit mean-value filters overlap; i.e., each computed box X_i is used in k successive filters. One-step methods do not encounter this problem because each computed box X_i is used only at one time step to compute the following box: X_{i+1}. *Global filters*, which are presented in the next section, avoid this variable dependency problem and make it possible to apply standard techniques for the wrapping effect.

4.4. Global filters. The main idea underlying global filters is to cluster several mean-value filters together so that they do not overlap. The intuition is illustrated in Figure 4 for $k = 3$. It can be seen that the global filter prunes the three predicted boxes D_3^-, D_4^-, and D_5^- for times t_3, t_4, and t_5 using the boxes D_0^*, D_1^*, and D_2^* computed for times t_0, t_1, and t_2. Observe also that global filters do not overlap; i.e., the boxes D_0^*, D_1^*, and D_2^* are not used in subsequent filters. More precisely, a global filter is a system of k successive explicit mean-value filters.

DEFINITION 11 (global filter). *A* global filter *for ODE* $\mathcal{O}$ *in* $\mathbf{D}^0_{0..2k-1}$ *is a system* $\{FL_i(\mathbf{t}_{i..k+i}, \mathbf{D}_{i..k+i}, e_i)\}_{0\le i<k}$ *of* k *successive explicit mean-value filters for* $\mathcal{O}$ *in* $\mathbf{D}^0_{0..k}, \ldots, \mathbf{D}^0_{k-1..2k-1}$ *respectively given as*

(17)
$$\begin{cases} X_k &= K_0(e_0) + A_{0,0}(e_0)X_0 + \cdots + A_{0,k-1}(e_0)X_{k-1}, \\ X_{k+1} &= K_1(e_1) + A_{1,0}(e_1)X_1 + \cdots + A_{1,k-1}(e_1)X_k, \\ &\vdots \\ X_{2k-1} &= K_{k-1}(e_{k-1}) + A_{k-1,0}(e_{k-1})X_{k-1} + \cdots + A_{k-1,k-1}(e_{k-1})X_{2k-2}, \end{cases}$$

where $\mathbf{X}_{0..2k-1} = \mathbf{D}_{0..2k-1} - m(\mathbf{D}^0_{0..2k-1})$.

The key idea to remove the variable dependency problem is to solve (17) globally by transforming the global filter into an explicit form

$$\begin{bmatrix} X_k \\ \vdots \\ X_{2k-1} \end{bmatrix} = C(\mathbf{e}_0) \begin{bmatrix} X_0 \\ \vdots \\ X_{k-1} \end{bmatrix} + R(\mathbf{e}_0)$$

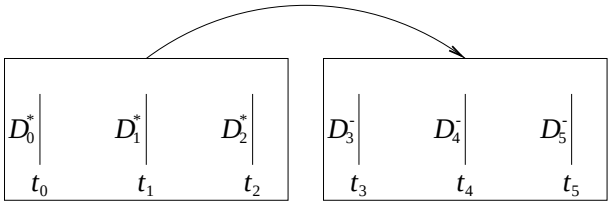


FIG. 4. *Intuition of the globalization process ($k = 3$): Predicted boxes D_3^-, D_4^-, and D_5^- for times t_3, t_4, and t_5 are pruned globally using boxes D_0^*, D_1^*, and D_2^* computed for times t_0, t_1, and t_2.*

or, more concisely,

$$\mathbf{X}_1 = C(\mathbf{e}_0)\mathbf{X}_0 + R(\mathbf{e}_0), \tag{18}$$

where $C(\mathbf{e}_0) \in \mathbb{IR}^{nk \times nk}$ and $R(\mathbf{e}_0) \in \mathbb{IR}^{nk}$.

An interesting property of global filters is that each pruned box at times t_3, t_4, or t_5 can be computed only in terms of the predicted boxes and the boxes at times t_0, t_1, and t_2 by using Gaussian elimination. Hence, it removes the dependencies introduced in D_3^- and D_4^-. Consider a system with $k = 3$:

$$\begin{cases} X_3 = A_{00}X_0 + A_{01}X_1 + A_{02}X_2 + K_0, \\ X_4 = A_{10}X_1 + A_{11}X_2 + A_{12}X_3 + K_1, \\ X_5 = A_{20}X_2 + A_{21}X_3 + A_{22}X_4 + K_2. \end{cases}$$

Variable X_4 can be eliminated from the last equation to obtain

$$X_5 = A_{20}X_2 + A_{21}X_3 + A_{22}(A_{10}X_1 + A_{11}X_2 + A_{12}X_3 + K_1) + K_2.$$

To avoid multiplying interval matrices (e.g., $A_{22}A_{10}$), we can apply the midpoint technique (see point 1 of section 2.3) to obtain

$$\begin{aligned} X_5 = {} & A_{20}X_2 + A_{21}X_3 + m(A_{22})(A_{10}X_1 + A_{11}X_2 + A_{12}X_3 + K_1) \\ & + K_2 + s(A_{22})s(D_4^-). \end{aligned} \tag{19}$$

By distribution and rearrangement of the parentheses, we can rewrite (19) as

$$\begin{aligned} X_5 = {} & (m(A_{22})A_{10})X_1 + (A_{20} + m(A_{22})A_{11})X_2 + (A_{21} + m(A_{22})A_{12})X_3 \\ & + m(A_{22})K_1 + K_2 + s(A_{22})s(D_4^-). \end{aligned}$$

Variable X_3 can be eliminated from this equation in a similar fashion to obtain a filter involving only X_5, X_0, X_1, and X_2. Similarly, variable X_3 can be eliminated from the second equation to obtain a filter involving only X_4, X_0, X_1, and X_2.[9]

A generic algorithm for computing an explicit global filter is given in Figure 5. It receives as input the ODE system $\mathbb{O}$, the previous integration times $\mathbf{t}_0$, the pruned boxes $\mathbf{D}_0^0$, and the bounding boxes $\mathbf{B}_{1..k-1}$, the new integration points $\mathbf{t}_1$, the predicted boxes $\mathbf{D}_1^0$ for these integration points, the bounding boxes $\mathbf{B}_1$ for the new integration points, and the evaluation times for the filters. It generates the matrix and vectors of the explicit global filter which can be used to compute the pruned boxes. The resulting filter is backward(k)-consistent w.r.t. the resulting boxes. Its precise specification is as follows.

[9] As observed by one of the reviewers, there are still some dependencies, but these are very small.

```
function ExplicitGlobalFilter(𝕆, t_0, D_0^0, B_{1..k-1}, t_1, D_1^0, B_1, e_0)
  begin
1   for i := 0 to k − 1 do
2     ⟨K_i, A_{i,0}, ..., A_{i,k-1}⟩ := EMVFL(𝕆, t_{i..i+k}, D^0_{i..i+k}, B_{i+1..i+k}, e_i);
3   endfor
4   for i := k − 1 downto 0 do
5     R_i := K_i;
6     for l := i downto 1 do
7       A* := m(A_{i,k-1});
8       R_i := R_i + A*K_{l-1} + s(A_{i,k-1})s(D^0_{k+l-1});
9       for j := k − 1 downto 1 do
10        A_{i,j} := A_{i,j-1} + A*A_{l-1,j}
11      endfor
12      A_{i,0} := A*A_{l-1,0}
13    endfor
14  endfor
15  return ⟨(A_{i,j})_{0≤i≤k-1, 0≤j≤k-1}, (R_i)_{0≤i≤k-1}⟩
  end
```

FIG. 5. *An algorithm for computing an explicit global filter.*

SPECIFICATION 1 (EXPLICITGLOBALFILTER). *Let B_i be a bounding box of ODE $\mathbb{O}$ over $[t_{i-1}, t_i]$ w.r.t. (t_0, D_0) for $1 \le i \le 2k-1$. Let*

$$\langle C(\mathbf{e}_0), R(\mathbf{e}_0)\rangle = \text{EXPLICITGLOBALFILTER}(\mathbb{O}, \mathbf{t}_0, \mathbf{D}_0^0, \mathbf{B}_{1..k-1}, \mathbf{t}_1, \mathbf{D}_1^0, \mathbf{B}_1, \mathbf{e}_0),$$

$\mathbf{X}_0 = \mathbf{D}_0 - m(\mathbf{D}_0^0)$, and $\mathbf{X}_1 = \mathbf{D}_1 - m(\mathbf{D}_1^0)$. Then the system $\mathcal{S} : \mathbf{X}_1 = C(\mathbf{e}_0)\mathbf{X}_0 + R(\mathbf{e}_0)$ is a global filter for $\mathbb{O}$ in $(\mathbf{D}_0^0, \mathbf{D}_1^0)$.

The algorithm is generic in the sense that it uses the function EMVFL to generate an explicit mean-value filter. How to generate such a filter is discussed in section 5, but its specification is given as follows.

SPECIFICATION 2 (EMVFL). *Let B_i be a bounding box of ODE $\mathbb{O}$ over $[t_{i-1}, t_i]$ w.r.t. (t_0, D_0) for $1 \le i \le k$. Let*

$$\langle K(t), A_0(t), \ldots, A_{k-1}(t)\rangle = \text{EMVFL}(\mathbb{O}, \mathbf{t}, \mathbf{D}^0, (B_1, \ldots, B_k), t).$$

Then the interval relation

$$FL(\mathbf{t}, \mathbf{D}, t) \Leftrightarrow X_k = K(t) + \sum_{i=0}^{k-1} A_i(t) X_i,$$

where $\mathbf{X} = \mathbf{D} - m(\mathbf{D}^0)$ and $\mathbf{D} \subseteq \mathbf{D}^0$ is an explicit mean-value filter for $\mathbb{O}$ in $\mathbf{D}^0$.

Finally, observe that global filters not only remove the variable dependency problem by globalizing the pruning process, but they also have the advantage of producing square systems which makes it possible to apply standard techniques to address the wrapping effect. The next section discusses the wrapping effect in detail.

4.5. The wrapping effect in global filters. The wrapping effect in global filters arises when multiplying a $nk \times nk$ matrix and a box of nk elements. Fortunately, since the matrices in global filters are square, the wrapping effect can be handled as

in one-step methods by using local coordinate transformations and QR factorizations [20]. We now explain this process in detail. Initially, starting from the previous boxes $\mathbf{D}_0^*$ and predicted boxes $\mathbf{D}_1^-$, we need to solve the system

$$\mathbf{D}_1 - m(\mathbf{D}_1^-) = C_1(\mathbf{e}_0)(\mathbf{D}_0^* - m(\mathbf{D}_0^*)) + R_1(\mathbf{e}_0)$$

or, equivalently, the system

$$\mathbf{X}_1 = C_1(\mathbf{e}_0)\mathbf{X}_0 + R_1(\mathbf{e}_0),$$

where $\mathbf{X}_1 = \mathbf{D}_1 - m(\mathbf{D}_1^-)$ and $\mathbf{X}_0 = \mathbf{D}_0^* - m(\mathbf{D}_0^*)$. The pruned boxes are then obtained by

$$\mathbf{D}_1^* = \mathbf{D}_1^- \cap (\mathbf{X}_1 + m(\mathbf{D}_1^-)).$$

The key idea in tackling the wrapping effect is to find a good coordinate system to represent the solution $\mathbf{X}_1$ compactly so that errors will not accumulate drastically in subsequent integration steps. For this purpose, we introduce a coordinate transformation

$$M_1\mathbf{y}_1 = \mathbf{u}_1 - m(\mathbf{D}_1^*)$$

which can be reexpressed in terms of the $\mathbf{x}$ variables as

$$M_1\mathbf{y}_1 = \mathbf{x}_1 + m(\mathbf{D}_1^-) - m(\mathbf{D}_1^*).$$

We then solve the system

$$M_1\mathbf{Y}_1 = C_1(\mathbf{e}_0)\mathbf{X}_0 + R_1(\mathbf{e}_0) + m(\mathbf{D}_1^-) - m(\mathbf{D}_1^*)$$

by inverting the matrix M_1:

$$\mathbf{Y}_1 = (M_1^{-1}C_1(\mathbf{e}_0))\mathbf{X}_0 + M_1^{-1}(R_1(\mathbf{e}_0) + m(\mathbf{D}_1^-) - m(\mathbf{D}_1^*)).$$

The matrix M_1 and the boxes $\mathbf{Y}_1$ and $\mathbf{D}_1^*$ are then sent to the next integration step. Observe that $\mathbf{Y}_1$ is a compact representation of $\mathbf{D}_1^*$ in the local coordinate system.

In the next integration step, the boxes $\mathbf{D}_1^*$ are used (together with other data) to compute new predicted boxes $\mathbf{D}_2^-$ as well as the new global filter

$$\mathbf{D}_2 - m(\mathbf{D}_2^-) = C_2(\mathbf{e}_1)(\mathbf{D}_1^* - m(\mathbf{D}_1^*)) + R_2(\mathbf{e}_1)).$$

Since $M_1\mathbf{y}_1 = \mathbf{u}_1 - m(\mathbf{D}_1^*)$ by the coordinate transformation, the above filter can be rewritten as

$$\mathbf{X}_2 = (C_2(\mathbf{e}_1)M_1)\mathbf{Y}_1 + R_2(\mathbf{e}_1),$$

where $\mathbf{X}_2 = \mathbf{D}_2 - m(\mathbf{D}_2^-)$. Observe the associativity of the multiplication which is critical in reducing the wrapping effect. The new boxes are computed as

$$\mathbf{D}_2^* = \mathbf{D}_2^- \cap (\mathbf{X}_2 + m(\mathbf{D}_2^-)).$$

Once again, we would like to represent the set of solutions $\mathbf{X}_2$ compactly, and we use a local coordinate transformation

$$M_2\mathbf{y}_2 = \mathbf{u}_2 - m(\mathbf{D}_2^*)$$

to obtain the system

$$M_2\mathbf{Y}_2 = (C_2(\mathbf{e}_1)M_1)\mathbf{Y}_1 + R_2(\mathbf{e}_1) + m(\mathbf{D}_2^-) - m(\mathbf{D}_2^*).$$

This equation system can be solved by inverting M_2:

$$\mathbf{Y}_2 = (M_2^{-1}(C_2(\mathbf{e}_1)M_1))\mathbf{Y}_1 + M_2^{-1}(R_2(\mathbf{e}_1) + m(\mathbf{D}_2^-) - m(\mathbf{D}_2^*)).$$

Once again, observe the associativity in the multiplication to tackle the wrapping effect. The hope is that the matrix $M_2^{-1}(C_2(\mathbf{e}_1)M_1)$ is diagonally dominant or triangular. Also, M_2, $\mathbf{Y}_2$, and $\mathbf{D}_2^*$ will be sent to the next integration step. As a consequence, at integration step i, we solve

$$\mathbf{X}_i = (C_i(\mathbf{e}_{i-1})M_{i-1})\mathbf{Y}_{i-1} + R_i(\mathbf{e}_{i-1}),$$

where $\mathbf{X}_i = \mathbf{D}_i - m(\mathbf{D}_i^-)$, and the new boxes are obtained by

$$\mathbf{D}_i^* = \mathbf{D}_i^- \cap (\mathbf{X}_i + m(\mathbf{D}_i^-)).$$

The local coordinate transformation

$$M_i\mathbf{y}_i = \mathbf{u}_i - m(\mathbf{D}_i^*)$$

is used to compute the new $\mathbf{Y}_i$ which is given by

$$\mathbf{Y}_i = (M_i^{-1}(C_i(\mathbf{e}_{i-1})M_{i-1}))\mathbf{Y}_{i-1} + M_i^{-1}(R_i(\mathbf{e}_{i-1}) + m(\mathbf{D}_i^-) - m(\mathbf{D}_i^*)).$$

In addition, in order to avoid the costly (see [15]) product of the two interval matrices M_i^{-1} and $C_i(\mathbf{e}_{i-1})M_{i-1}$, we use the standard midpoint technique (see point 1 of section 2.3) to obtain

$$\begin{aligned}\mathbf{Y}_i = {} & (m(M_i^{-1})(C_i(\mathbf{e}_{i-1})M_{i-1}))\mathbf{Y}_{i-1} + m(M_i^{-1})(R_i(\mathbf{e}_{i-1}) + \mathbf{d}_i) \\ & + s(M_i^{-1})((C_i(\mathbf{e}_{i-1})M_{i-1})\mathbf{Y}_{i-1} + R_i(\mathbf{e}_{i-1}) + \mathbf{d}_i),\end{aligned}$$

where $\mathbf{d}_i = m(\mathbf{D}_i^-) - m(\mathbf{D}_i^*)$. This last system can be rewritten as

$$\begin{aligned}\mathbf{Y}_i = {} & (m(M_i^{-1})(C_i(\mathbf{e}_{i-1})M_{i-1}))\mathbf{Y}_{i-1} + m(M_i^{-1})(R_i(\mathbf{e}_{i-1}) + \mathbf{d}_i) \\ & + s(M_i^{-1})(\mathbf{X}_i + \mathbf{d}_i)\end{aligned}$$

by the definition of $\mathbf{X}_i$. In this process, the choice of an appropriate matrix M_i is, of course, crucial. Lohner's QR factorization technique [20] is a very successful scheme to obtain such a matrix.

4.6. A pruning algorithm based on global filters. We are now in position to present a pruning algorithm based on global filters. The pruning algorithm enforces backward(k)-consistency on a global filter composed of k mean-value filters. The algorithm is shown in Figure 6, and its specification is as follows.

SPECIFICATION 3 (PRUNE). *Let ms be the multistep solution of ODE* $\mathbb{O}$ *and* B_i *a bounding box of* $\mathbb{O}$ *over* $[t_{i-1}, t_i]$ *w.r.t.* (t_0, D_0) *for* $1 \le i \le 2k-1$. *Let*

$$\langle \mathbf{D}_1^*, \mathbf{Y}_1, M_1 \rangle = \text{PRUNE}(\mathbb{O}, \mathbf{t}_0, \mathbf{D}_0^*, \mathbf{B}_{1..k-1}, \mathbf{Y}_0, M_0, \mathbf{t}_1, \mathbf{D}_1^-, \mathbf{B}_1),$$

$\mathcal{A}_0 = \{M_0\mathbf{y}_0 + m(\mathbf{D}_0^*) \mid \mathbf{y}_0 \in \mathbf{Y}_0\} \cap \mathbf{D}_0^*$, *and* $\mathcal{A}_1 = \{M_1\mathbf{y}_1 + m(\mathbf{D}_1^*) \mid \mathbf{y}_1 \in \mathbf{Y}_1\} \cap \mathbf{D}_1^*$. *Then*

function PRUNE($\mathbb{O}, \mathbf{t}_0, \mathbf{D}_0^*, \mathbf{B}_{1..k-1}, \mathbf{Y}_0, M_0, \mathbf{t}_1, \mathbf{D}_1^-, \mathbf{B}_1$)
begin
1 $\langle C_1, R_1\rangle := \text{EXPLICITGLOBALFILTER}(\mathbb{O}, \mathbf{t}_0, \mathbf{D}_0^*, \mathbf{B}_{1..k-1}, \mathbf{t}_1, \mathbf{D}_1^-, \mathbf{B}_1, \mathbf{e}_0)$;
2 $C_1^* = C_1 M_0$;
3 $\mathbf{X}_1 := C_1^* \mathbf{Y}_0 + R_1$;
4 $\mathbf{D}_1^* := (\mathbf{X}_1 + m(\mathbf{D}_0^-)) \cap (\mathbf{D}_0^-)$;
5 $M_1 := \text{COORDTRANSFO}(C_1^*, \mathbf{Y}_0)$;
6 $\mathbf{d}_1 := m(\mathbf{D}_1^-) - m(\mathbf{D}_1^*)$;
7 $\mathbf{Y}_1 := (m(M_1^{-1})C_1^*)\mathbf{Y}_0 + m(M_1^{-1})(R_1 + \mathbf{d}_1) + s(M_1^{-1})(\mathbf{X}_1 + \mathbf{d}_1)$;
8 **return** $\langle \mathbf{D}_1^*, \mathbf{Y}_1, M_1\rangle$
end

FIG. 6. *The pruning algorithm on global filters.*

1. $ms((\mathbf{t}_0, \mathbf{t}_1), (\mathcal{A}_0, \mathbf{D}_1^-), t_i) \subseteq ms((\mathbf{t}_0, \mathbf{t}_1), (\mathcal{A}_0, \mathcal{A}_1), t_i)$ *for* $k \le i \le 2k-1$;
2. $\mathbf{D}_1^* \subseteq \mathbf{D}_1^-$;
3. *there exists a global filter which is backward(k)-consistent in* $((\mathbf{t}_0, \mathbf{t}_1), (\mathbf{Y}_0, \mathbf{D}_1^*))$ *and in* $((\mathbf{t}_0, \mathbf{t}_1), (\mathbf{Y}_0, \mathbf{Y}_1))$ *for a given time vector.*

The algorithm receives as input the ODE $\mathcal{O}$, the previous integration times $\mathbf{t}_0$, the pruned boxes $\mathbf{D}_0^*$ computed at times $\mathbf{t}_0$, the bounding boxes $\mathbf{B}_{1..k-1}$ for all previous integration steps, the boxes $\mathbf{Y}_0$ and matrix M_0 from the previous integration step as well as the new integration times $\mathbf{t}_1$, the predicted boxes $\mathbf{D}_1^-$, and the bounding boxes $\mathbf{B}_1$ for these integration times. It returns the pruned boxes $\mathbf{D}_1^*$ for integration steps $\mathbf{t}_1$ as well as the new boxes $\mathbf{Y}_1$ and the new matrix M_1 to be used in the next integration step. The algorithm itself follows the same steps as outlined in the preceding section. It computes the explicit form of the global filter (line 1), the new boxes $\mathbf{X}_1$ (line 3), and the pruned boxes $\mathbf{D}_1^*$ (line 4). It then computes the new matrix M_1 (line 5) and the new boxes $\mathbf{Y}_1$ (line 7).

5. Hermite filters. In the previous section, we assumed the existence of interval extensions of p and $\partial p/\partial t$, and we assumed that we could bound the error terms e and $\partial e/\partial t$. We now show how to use Hermite interpolation polynomials for this purpose. Informally speaking, a Hermite interpolation polynomial approximates a function $g \in C^r$ (for sufficiently large r) which is known implicitly by its values and the values of its successive derivatives at various points. A Hermite interpolation polynomial is specified by imposing that its values and the values of its successive derivatives at some given points be equal to the values of g and of its derivatives at the same points. Note that the number of conditions (i.e., the number of successive derivatives that are considered) may vary at the different points [29, 1].

DEFINITION 12 (Hermite(σ) interpolation polynomial). *Consider the ODE* $u' = f(u)$ *and let* $\sigma = (\sigma_0, \ldots, \sigma_k) \in \mathbb{N}^{k+1}$, $\sigma_i \neq 0$ $(0 \le i \le k)$, *and* $\sigma_s = \sum_{i=0}^{k} \sigma_i$. *The* Hermite($\sigma$) interpolation polynomial *w.r.t.* f *and* $(\mathbf{t}, \mathbf{u})$ *is the unique polynomial* q *of degree* $\le \sigma_s - 1$ *satisfying*

$$q^{(j)}(t_i) = j!(u_i)_j \quad (0 \le j \le \sigma_i - 1,\ 0 \le i \le k). \tag{20}$$

PROPOSITION 5 (Hermite(σ) interpolation polynomial). *The polynomial* q *satisfying the conditions* (20) *is given by*

$$q(t) = \sum_{i=0}^{k} \sum_{j=0}^{\sigma_i - 1} j!(u_i)_j \varphi_{ij}(t), \tag{21}$$

where

$$\begin{aligned}
&\varphi_{i,\sigma_i-1}(t) = l_{i,\sigma_i-1}(t), \quad i = 0, \ldots, k, \\
&\varphi_{ij}(t) = l_{ij}(t) - \sum_{\nu=j+1}^{\sigma_i-1} l_{ij}^{(\nu)}(t_i)\varphi_{i\nu}(t), \quad i = 0, \ldots, k,\ j = 0, \ldots, \sigma_i - 2, \\
&l_{ij}(t) = \frac{(t-t_i)^j}{j!} \prod_{\substack{\nu=0 \\ \nu \neq i}}^{k} \left(\frac{t - t_\nu}{t_i - t_\nu} \right)^{\sigma_\nu}, \quad i = 0, \ldots, k,\ j = 0, \ldots, \sigma_i - 1.
\end{aligned} \tag{22}$$

It is easy to take interval extensions of a Hermite interpolation polynomial and of its derivative. The Taylor coefficients $(D_i)_j$ of the solution specifying the derivative conditions at the various interpolation points, as well as their Jacobians $\mathcal{J}(D_i)_j$ needed in the mean-value Hermite filters, can be computed by automatic differentiation techniques (see, e.g., [21, 22, 27]). The only remaining issue is to bound the error terms. The following standard theorem (e.g., [29, 1]) provides the necessary theoretical basis.

THEOREM 1 (Hermite error term). *Let $p(\mathbf{t}, \mathbf{u}, t)$ be the Hermite(σ) interpolation polynomial in t w.r.t. f and $(\mathbf{t}, \mathbf{u})$. Let $u(t) = ms(\mathbf{t}, \mathbf{u}, t)$, $ms(\mathbf{t}, \mathbf{u}, t) = p(\mathbf{t}, \mathbf{u}, t) + e(\mathbf{t}, \mathbf{u}, t)$, $T = \square\{t_0, \ldots, t_k, t\}$, $\sigma_s = \sum_{i=0}^{k} \sigma_i$, and $w(t) = \prod_{i=0}^{k}(t - t_i)^{\sigma_i}$. We have $(1 \leq i \leq n)$*

1. $\exists\ \xi_i \in T : e_i(\mathbf{t}, \mathbf{u}, t) = \frac{1}{\sigma_s!} u_i^{(\sigma_s)}(\xi_i) w(t)$;
2. $\exists\ \xi_{1,i}, \xi_{2,i} \in T : \frac{\partial e_i}{\partial t}(\mathbf{t}, \mathbf{u}, t) = \frac{1}{\sigma_s!} u_i^{(\sigma_s)}(\xi_{1,i}) w'(t) + \frac{1}{(\sigma_s+1)!} u_i^{(\sigma_s+1)}(\xi_{2,i}) w(t)$.

How do we use this theorem to bound the error terms? If B is a bounding box (produced by the bounding box process) for the ODE over $T = \square\{t_0, \ldots, t_k, t\}$ w.r.t. $(\mathbf{t}_0, \mathbf{u}_0)$, it suffices to compute two boxes $(B)_{\sigma_s}$ and $(B)_{\sigma_s+1}$ by automatic differentiation. We then obtain

$$\begin{aligned}
e(\mathbf{t}, \mathbf{u}, t) &\in (B)_{\sigma_s} w(t); \\
\frac{\partial e}{\partial t}(\mathbf{t}, \mathbf{u}, t) &\in (B)_{\sigma_s} w'(t) + (B)_{\sigma_s+1} w(t).
\end{aligned}$$

As a consequence, we can compute an effective relaxation of the ODE by specializing global filters with a Hermite interpolation polynomial and its error bound. In the following, filters based on Hermite(σ) interpolation are called *Hermite(σ) filters*, and a global Hermite(σ) filter is denoted by GHF(σ). Reference [12] discusses how to evaluate Hermite polynomials accurately.

6. Optimal Hermite filters. Let us summarize what we have achieved so far. The basic idea of our approach is to approximate the ODE for all $t:\ u' = f(u)$ by a filter

$$\forall\ t : FL(\mathbf{t}, \mathbf{D}, t).$$

We have shown that a global filter which prunes the last k boxes by using k successive mean-value filters addresses the wrapping effect and the variable dependency problem.

We have also shown that a global filter can be obtained by using Hermite interpolation polynomials together with their error bounds. As a consequence, we obtain a filter

$$\forall\ \mathbf{e}_0 : GHF(\sigma)(\mathbf{t}, \mathbf{D}, \mathbf{e}_0)$$

which can be used to prune the last k predicted boxes. The main remaining issue is to find an *evaluation time vector* $\mathbf{e}_0$ which miminizes the sizes of the solution boxes in

$$GHF(\sigma)(\mathbf{t}, \mathbf{D}, \mathbf{e}_0).$$

The purpose of this section is to show that there exists an optimal evaluation time vector (in a precise sense that we will define) and that it can be approximated or computed efficiently.

6.1. Preview of the approach. Our goal is to find an evaluation time vector $\mathbf{e}_0$ which miminizes the sizes of the solution boxes in a global Hermite filter. However, this is a difficult problem in general. We will thus solve a simpler problem, which consists of choosing an evaluation time that minimizes the *local error* of an individual filter, i.e., the size of the enclosure of $ms(\mathbf{t}_0, \mathbf{u}_0, t_k)$ produced by the filter, assuming that the *point* values $u_0, \ldots, u_{k-1}$ are given (and, of course, that $ms(\mathbf{t}_0, \mathbf{u}_0, t_k)$ is defined).[10]

DEFINITION 13 (local error of a filter). *Let FL be a filter for ODE $u' = f(u)$. The local error of FL w.r.t. $(\mathbf{t}_0, \mathbf{u}_0, t)$, denoted by $e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t)$, is defined as*

$$e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t) = \omega\left(\square\{u_k \in \mathbb{R}^n \mid FL(\mathbf{t}, \mathbf{u}, t)\}\right).$$

Since in a global filter we compute k boxes in one step, the step size is defined as $h = t_k - t_0$. Our analysis is based on the assumption that the step size h is sufficiently small. When we talk about an optimal evaluation time, the term *optimal* is thus to be understood in an *asymptotic* sense.

In the following, we restrict our attention to Hermite filters which satisfy a certain hypothesis (section 6.2). To find an optimal evaluation time, we first derive the local error (section 6.3). From the local error, we can then characterize the optimal evaluation time (section 6.4). Two of the main results of this section are as follows:

1. For a sufficiently small step size h, the relative distance $(t_e - t_k)/h$ between the optimal evaluation time t_e and the point t_k in a Hermite(σ) filter depends only on the relative distances $(t_{i+1} - t_i)/h$ $(i = 0, \ldots, k-1)$ between the interpolation points $t_0, \ldots, t_k$ and on σ.[11] In particular, it does not depend on the ODE itself.
2. From a practical standpoint, the computation of the optimal evaluation time induces a negligible overhead of the method. In particular, if we assume $t_{i+1} - t_i = h/k$ $(i \in \mathbb{N})$, the relative distance between the optimal evaluation time and t_k can be precomputed once for all for given k and σ.

The third main result is concerned with the order of a Hermite($(\sigma_0, \ldots, \sigma_k)$) filter which is shown to be $\mathcal{O}(h^{\sigma_s+1})$, where $\sigma_s = \sum_{i=0}^{k} \sigma_i$ when the evaluation point is chosen carefully.

[10] As observed by one of the reviewers, the local error may be called more appropriately excess-width, since the enclosure contains the exact solution. We kept the term "local error" because of the analogy with traditional methods.

[11] Note that $h = t_k - t_0$ (and not $h = t_k - t_{k-1}$) because of the globalization process.

6.2. Assumptions and notations. The following assumptions are used in this section. We assume that the integration times are increasing, i.e., $t_0 < \cdots < t_k$, and that $t - t_k = \mathcal{O}(h)$. We also assume that the function f satisfies a Lipschitz condition on $\Omega \subseteq \mathbb{R}^n$:

$$\exists c \in \mathbb{R} \quad \forall u, v \in \Omega : \|f(u) - f(v)\| \leq c\|u - v\|. \tag{23}$$

Note that (23) holds if we assume $f \in C^1(\Omega)$. We further assume that the interval extension F of function f satisfies ($D \subseteq \Omega$)

$$\omega(F(D)) = \mathcal{O}(\omega(D)). \tag{24}$$

For instance, (24) holds if F is the natural interval extension of f and (23) holds. We also assume that B is a bounding box of $u' = f(u)$ over $T = \square\{t_0, \ldots, t_k, t\}$ w.r.t. $(\mathbf{t}_0, \mathbf{u}_0)$ and that (see [23])

$$\omega\left((B)_j\right) = \Theta(\omega(B)) = \Theta(h), \qquad j \in \mathbb{N}. \tag{25}$$

From (23), the condition (25) holds if $(B)_j$ is a sufficiently tight enclosure of the set $\{(x)_j \mid x \in B\}$. In addition, we assume that the multistep solution *ms* is defined at $(\mathbf{t}_0, \mathbf{u}_0)$ or, in other words, that the ODE has a solution going through $u_0, \ldots, u_{k-1}$ at times $t_0, \ldots, t_{k-1}$. We also use the notations $\sigma = (\sigma_0, \ldots, \sigma_k)$, $\sigma_s = \sum_{i=0}^{k} \sigma_i$, and $w(t) = \prod_{i=0}^{k} (t - t_i)^{\sigma_i}$. Since we are interested in computing an enclosure of $ms(\mathbf{t}_0, \mathbf{u}_0, t_k)$ from the *point* values $u_0, \ldots, u_{k-1}$, we will consider a Hermite filter *FL* satisfying

$$FL(\mathbf{t}, (\mathbf{u}_0, v), t) \Rightarrow \frac{\partial p}{\partial t}(\mathbf{t}, (\mathbf{u}_0, v), t) + DE(t) - F(p(\mathbf{t}, (\mathbf{u}_0, v), t) + E(t)) = 0, \tag{26}$$

where

- F is an interval extension of f;
- $E(t) = (B)_{\sigma_s} w(t)$;
- $DE(t) = (B)_{\sigma_s} w'(t) + (B)_{\sigma_s + 1} w(t)$;
- $p(\mathbf{t}, (\mathbf{u}_0, v), t)$ is the Hermite(σ) interpolation polynomial in t w.r.t. f and $(\mathbf{t}, (\mathbf{u}_0, v))$.

Let us introduce the function

$$\delta(\mathbf{t}, (\mathbf{u}_0, v), t) = \frac{\partial p}{\partial t}(\mathbf{t}, (\mathbf{u}_0, v), t) - f(p(\mathbf{t}, (\mathbf{u}_0, v), t) + m_e(t)),$$

where $m_e(t) = m(E(t))$. From the hypothesis (24), the condition (26) can be rewritten as

$$FL(\mathbf{t}, (\mathbf{u}_0, v), t) \Rightarrow \delta(\mathbf{t}, (\mathbf{u}_0, v), t) = -DE(t) + \mathcal{O}(\omega(E(t))). \tag{27}$$

In case (24), the condition (27) is satisfied for natural Hermite filters (see section 4.1), provided that the interval extensions *MS* and *DMS* of *ms* and $\frac{\partial ms}{\partial t}$ yield point values when evaluated at point arguments. (Recall that we assume exact interval arithmetic for the theoretical parts of this paper.) If we assume that the interval extension of the Jacobian of f satisfies the same condition as F, i.e., $\omega(\mathcal{J}(D)_0) = \mathcal{O}(\omega(D))$, then (27) is satisfied for *implicit mean-value* Hermite filters. It is also a good approximation

for *explicit mean-value* Hermite filters if the matrix inversion is accurate (see section 4.2). We will also denote the Jacobian of δ w.r.t. variable v by

$$\begin{aligned}\Phi(t,v) &= \mathcal{J}_v\delta(\mathbf{t},(\mathbf{u}_0,v),t)\\ &= \mathcal{J}_v\frac{\partial p}{\partial t}(\mathbf{t},(\mathbf{u}_0,v),t) - \mathcal{J}f(p(\mathbf{t},(\mathbf{u}_0,v),t)+m_e(t))\mathcal{J}_v p(\mathbf{t},(\mathbf{u}_0,v),t).\end{aligned}$$

Finally, we introduce the following functions:

$$\begin{aligned}&\lambda(t) = \left(\left(\sum_{j=0}^{\sigma_k-2}\beta_{j+1}\tfrac{(t-t_k)^j}{j!}\right)+\left(\sum_{j=0}^{\sigma_k-1}\beta_j\tfrac{(t-t_k)^j}{j!}\right)\sum_{\nu=0}^{k-1}\tfrac{\sigma_\nu}{t-t_\nu}\right)\pi(t);\\ &\beta_0=1,\beta_j=-\pi^{(j)}(t_k),\ j=1,\dots,\sigma_k-1;\\ &\pi(t)=\prod_{\nu=0}^{k-1}\left(\tfrac{t-t_\nu}{t_k-t_\nu}\right)^{\sigma_\nu};\\ &\gamma(t)=\sum_{i=0}^{k}\tfrac{\sigma_i}{t-t_i}.\end{aligned}$$

6.3. Local error of a natural Hermite filter. To characterize the local error of a Hermite filter, we first need a technical lemma which characterizes the behavior of the derivatives of the filter.

LEMMA 1. *We have*

1. $\Phi(t,v) = I\lambda(t)+\mathcal{O}(1)$;
2. $\lambda(t)=\mathcal{O}(h^{-1})$;
3. $\lambda(t)=\Theta(h^{-1})$ *for* $t_{k-1}<t<t_k$.

This lemma shows that $\Phi(t,v)$ is a $\Theta(h^{-1})$ asymptotically diagonal matrix for $t_{k-1}<t<t_k$. Its proof is given in [12]. We are now in position to characterize the local error of a Hermite filter.

THEOREM 2 (local error of a Hermite filter). *Let FL be a Hermite(σ) filter for* $u'=f(u)$ *satisfying* (27). *We have*

1. $e_{loc}(FL,\mathbf{t}_0,\mathbf{u}_0,t) = |(I\lambda(t)+\mathcal{O}(1))^{-1}|\Theta(\omega(B))\left(|w'(t)|+|w(t)|\right)$;
2. $e_{loc}(FL,\mathbf{t}_0,\mathbf{u}_0,t) = \Omega(h^2)\left(|w'(t)|+|w(t)|\right)$;
3. *if* $t_{k-1}<t<t_k$, *then* $e_{loc}(FL,\mathbf{t}_0,\mathbf{u}_0,t)=\Theta(h^2)\left(|w'(t)|+|w(t)|\right)$.

Proof. Consider two arbitrary vectors $v_1,v_2\in\mathbb{R}^n$ such that

$$FL(\mathbf{t},(\mathbf{u}_0,v_1),t) \quad\text{and}\quad FL(\mathbf{t},(\mathbf{u}_0,v_2),t).$$

By the mean-value theorem, we can write

$$\delta(\mathbf{t},(\mathbf{u}_0,v_2),t)-\delta(\mathbf{t},(\mathbf{u}_0,v_1),t)=\Phi(t,\nu)(v_2-v_1),$$

where ν is on the straight line between v_1 and v_2. When the matrix $\Phi(t,\nu)$ is regular, we can write by Lemma 1 and (27)

$$\begin{aligned}v_2-v_1 &= \Phi^{-1}(t,\nu)\left(\delta(\mathbf{t},(\mathbf{u}_0,v_2),t)-\delta(\mathbf{t},(\mathbf{u}_0,v_1),t)\right)\\ &= \left(I\lambda(t)+\mathcal{O}(1)\right)^{-1}\left(DE(t)-DE(t)+\mathcal{O}(\omega(E(t)))\right).\end{aligned}$$

Since the two vectors v_1 and v_2 are chosen arbitrarily, it follows from (25) that

$$\begin{aligned}e_{loc}(FL,\mathbf{t}_0,\mathbf{u}_0,t) &= |(I\lambda(t)+\mathcal{O}(1))^{-1}|\left(\omega(DE(t))+\mathcal{O}(\omega(E(t)))\right)\\ &= |(I\lambda(t)+\mathcal{O}(1))^{-1}|\Theta(\omega(B))\left(|w'(t)|+|w(t)|\right),\end{aligned}$$

which proves point 1. Points 2 and 3 are now direct consequences of Lemma 1 and (25). ∎

We are now ready to show how to find an optimal evaluation time for Hermite filters.

6.4. Optimal evaluation time for a natural Hermite filter. Our first result characterizes the order of a Hermite filter. It also hints on how to obtain an optimal evaluation time. Recall that the order of a method (or of a filter) is the order of the local error minus 1.

THEOREM 3 (order of a Hermite filter). *Let FL be a Hermite(σ) filter for $u' = f(u)$ satisfying (27). Then*

1. *there exists t such that $t_{k-1} < t < t_k$, and $w'(t) = 0$;*
2. *if $t_{k-1} < t < t_k$ and $w'(t) = 0$, then $e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t) = \mathcal{O}(h^{\sigma_s+2})$;*
3. *if $w'(t) \neq 0$, then $e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t) = \Omega(h^{\sigma_s+1})$.*

Proof. Consider an evaluation time t such that $t - t_k = \mathcal{O}(h)$. We have $w(t) = \mathcal{O}(h^{\sigma_s})$ and $w'(t) = \mathcal{O}(h^{\sigma_s-1})$. First assume that $t_{k-1} < t < t_k$ and $w'(t) = 0$. By Rolle's theorem, since $w(t_{k-1}) = w(t_k) = 0$, there exists such an evaluation time t. By Theorem 2, $e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t) = \mathcal{O}(h^{\sigma_s+2})$. Now assume that $w'(t) \neq 0$. By Theorem 2, $e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t) = \Omega(h^{\sigma_s+1})$. ∎

Theorem 3 indicates that a better order for Hermite filters is obtained when we choose an evaluation time t that is a root of the polynomial w'. This is the basis of our next result which describes a necessary condition for optimality.

THEOREM 4 (necessary condition for optimal Hermite filters). *Let FL be a Hermite(σ) filter for $u' = f(u)$ satisfying (27), and let $t_e \in \mathbb{R}$ be such that*

$$e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t_e) = \min_{t - t_k = \mathcal{O}(h)} \{e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t)\}$$

for h sufficiently small. Then t_e is a zero of the function γ.

Proof. Assume that $t - t_k = \mathcal{O}(h)$ and that h is sufficiently small. By Theorem 3, $w'(t_e)$ must be zero to minimize the local error. Note that $FL(\mathbf{t}, (\mathbf{u}_0, v), t_i)$ holds for *any* $v \in \mathbb{R}^n$ if $w'(t_i) = 0$ $(0 \leq i \leq k)$. Thus $t_e \notin \{t_0, \ldots, t_k\}$ and $w(t_e) \neq 0$. Since $w'(t) = w(t)\gamma(t)$, we conclude that $\gamma(t_e) = 0$. ∎

Our next result specifies the number of zeros of the function γ as well as their locations.

PROPOSITION 6. *The function γ in Theorem 4 has exactly k zeros $s_0, \ldots, s_{k-1}$ such that $t_i < s_i < t_{i+1}$ $(0 \leq i < k)$.*

Proof. We have $w'(t) = w(t)\gamma(t)$. By Rolle's theorem, as $w(t_i) = w(t_{i+1}) = 0$, w' has a root s_i with $t_i < s_i < t_{i+1}$ and $w(s_i) \neq 0$ $(0 \leq i < k)$. Furthermore, the roots of w' are in $\{s_0, \ldots, s_{k-1}, t_0, \ldots, t_k\}$ because t_i is a root of multiplicity $\sigma_i - 1$ $(0 \leq i \leq k)$ and w' is of degree $\sigma_s - 1$, i.e., $k + \sum_{i=0}^{k}(\sigma_i - 1) = \sigma_s - 1$. Since γ is not defined at $t_0, \ldots, t_k$, its zeros are in $\{s_0, \ldots, s_{k-1}\}$. ∎

We are now ready to characterize precisely the optimal evaluation time for a Hermite filter.

THEOREM 5 (optimal evaluation time). *Let FL be a Hermite(σ) filter for $u' = f(u)$ satisfying (27), let $s_0 < \cdots < s_{k-1}$ be the zeros of γ, and let $t_e \in \mathbb{R}$ such that*

$$e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t_e) = \min_{t - t_k = \mathcal{O}(h)} \{e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, t)\}.$$

Then, for h sufficiently small,

$$|(w/\lambda)(t_e)| = \min_{s \in \{s_0, \ldots, s_{k-1}\}} \{|(w/\lambda)(s)|\}.$$

TABLE 1
Relative distance between the rightmost zero t_e of γ and t_k when $\sigma_0 = \cdots = \sigma_k$.

k	1	2	3	4	5	6
$(t_e - t_k)/h$	-0.5000	-0.2113	-0.1273	-0.0889	-0.0673	-0.0537

Proof. Let us assume that h is sufficiently small. From Theorem 4, we know that $t_e \in \{s_0, \ldots, s_{k-1}\}$. By definition, for $i = 0, \ldots, k-1$, $w'(s_i) = w(s_i)\gamma(s_i) = 0$ and, from Theorem 2,

$$e_{loc}(FL, \mathbf{t}_0, \mathbf{u}_0, s_i) = |(I\lambda(s_i) + \mathcal{O}(1))^{-1}|\Theta(\omega(B))|w(s_i)|.$$

From Proposition 6, if $t = s_i$ $(i = 0, \ldots, k-1)$, B is a bounding box over $T = \square\{t_0, \ldots, t_k, s_i\} = [t_0, t_k]$ w.r.t. $(\mathbf{t}_0, \mathbf{u}_0)$ and the factor $\Theta(\omega(B))$ does not depend on $t = s_i$. We have thus to minimize the function

$$\rho(t) = |(I\lambda(t) + \mathcal{O}(1))^{-1}||w(t)|$$

for $t \in \{s_0, \ldots, s_{k-1}\}$. By Lemma 1, $\lambda(s_{k-1}) = \Theta(h^{-1})$. Therefore, we must have $\lambda(t_e) = \Theta(h^{-1})$ and $\rho(t_e) \approx |(w/\lambda)(t_e)|$. Let us now assume that there exists $i \in 0..k-1$ such that $|(w/\lambda)(s_i)| < |(w/\lambda)(t_e)|$. We can write

$$\begin{aligned} |(w/\lambda)(s_i)| < |(w/\lambda)(t_e)| &\Rightarrow \lambda(s_i) = \Theta(h^{-1}) \\ &\Rightarrow \rho(s_i) \approx |(w/\lambda)(s_i)| \\ &\Rightarrow \rho(s_i) < \rho(t_e), \end{aligned}$$

which is a contradiction. ∎

6.5. Discussion. It is important to discuss the consequences of Theorem 4 in some detail. First observe that the *relative distance* $(t_e - t_k)/h$ between the optimal evaluation time t_e and the point t_k depends *only* on the *relative distances* $(t_{i+1} - t_i)/h$ $(i = 0, \ldots, k-1)$ between the interpolation points $t_0, \ldots, t_k$ and on the vector σ. In particular, it is independent from the ODE itself. For instance, for $k = 1$, we have $\gamma(t) = \frac{\sigma_0}{t - t_0} + \frac{\sigma_1}{t - t_1}$, and γ has a single zero given by $t_e = \frac{\sigma_1 t_0 + \sigma_0 t_1}{\sigma_0 + \sigma_1}$. In addition, if $\sigma_0 = \cdots = \sigma_k$, then the zeros of γ are independent from σ. In particular, for $k = 1$, we have $t_e = (t_0 + t_1)/2$. From a practical standpoint, the computation of the optimal evaluation time induces a negligible overhead of the method. In particular, if we assume $t_{i+1} - t_i = h/k$ $(i \in \mathbb{N})$, then the relative distance between t_k and the optimal evaluation time can be precomputed and stored for a variety of values of k and σ. Finally, it is worth stressing that any zero of function γ gives an $\mathcal{O}(h^{\sigma_s+1})$ order for the Hermite filter provided that $\lambda(t) = \Theta(h^{-1})$ at that zero. Hence any such zero is in fact a potential candidate for the optimal evaluation time. In our experiments (see the next section), the rightmost zero was always the optimal evaluation time when $\sigma_0 = \cdots = \sigma_k$, although we have not been able to prove this result.

6.6. Illustration. We now illustrate the theoretical results presented in this section. Table 1 gives approximative values of the relative distance $(t_e - t_k)/h$ between the rightmost zero t_e of the function γ and the point t_k $(1 \le k \le 6)$ for $\sigma_0 = \cdots = \sigma_k$ and $t_{i+1} - t_i = h/k$ $(i = 0, \ldots, k-1)$. For two interpolation points, t_e is in the middle of t_0 and t_1. It then moves closer and closer to t_k for larger values of k.

Figure 7 illustrates the functions γ, w, w', λ, and w/λ for $k = 4$ and $\sigma = (2, 2, 2, 2, 2)$. The top left figure shows the function w' and γ, as well as the zeros of

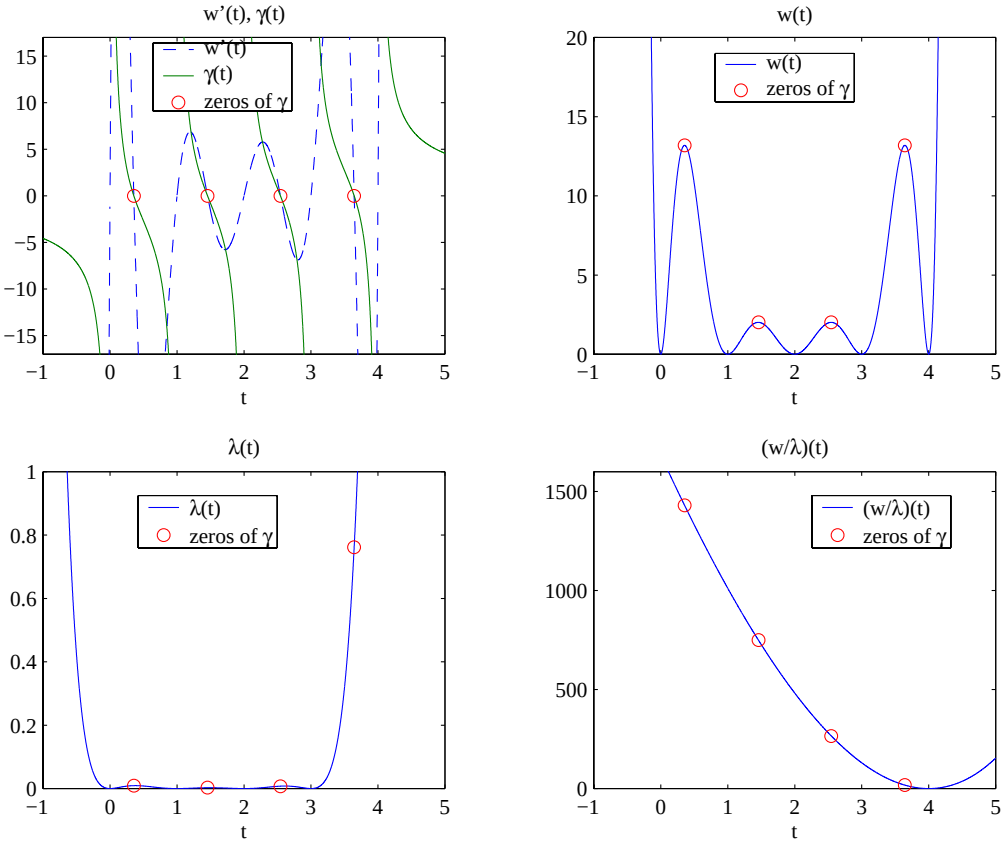


FIG. 7. *The functions* γ, w, w', λ *and* w/λ *for the case* $k = 4, \sigma = (2, 2, 2, 2, 2)$.

γ. The top right figure shows the function w with the zeros of γ in superposition. The bottom left figure shows function λ with the zeros of γ in superposition. The bottom right picture shows the function w/λ and the zeros of γ. It can be seen that the rightmost zero minimizes the local error in this example.

6.7. Validity of the asymptotic assumption. Our analysis is based on the assumption that the step size h is sufficiently small. But *how small is sufficiently small?* According to our experiments, the actual step sizes are generally small enough so that the *asymptotically* optimal evaluation times produced by the above theory are good approximations of the *real* optima. There are two reasons for these small actual step sizes:

1. the need to bound the local error, which limits the stability of validated methods and makes stiff problems more challenging;
2. the existing bounding box process, which often impose the strongest restriction on the step size, especially for stiff problems.

Figure 8 illustrates our theoretical results experimentally on a specific ODE. It plots the local error of several global Hermite filters (GHFs) as a function of the evaluation time for the Lorenz system (e.g., [10]). It is assumed that $t_{i+1} - t_i$ is constant $(0 \le i \le 2k - 2)$. In addition, we assume that, in each mean-value filter composing GHF, the distance between the evaluation time and the rightmost interpolation point

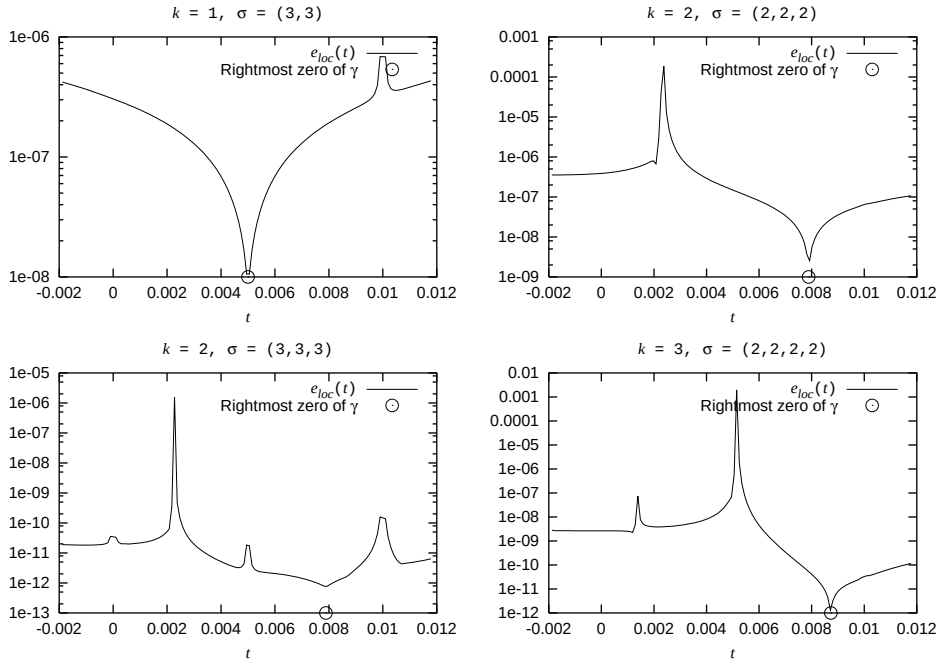


FIG. 8. *Local error of GHFs as a function of the evaluation time for the Lorentz system.*

is constant. In the graphs, $[t_0, t_k] = [0, 0.01]$ and $h = t_k - t_0 = 0.01$. The figure also shows the rightmost zero of the function γ as obtained from Table 1. As we can see, the rightmost zero of γ is a very good approximation of the optimal evaluation time of the filter for all the cases displayed.

7. The algorithm. We are now in position to present our algorithm for enclosing solutions of IVPs for parametric ODEs. The algorithm is presented in Figure 9, and Figure 10 gives the specification of the functions not covered so far. The first two lines initialize the integration process and compute the initial bounding boxes, pruned domains, and the boxes and matrices needed for the wrapping effect. The main step of the integration are lines 4-6. Line 4 computes the new bounding boxes, line 5 uses them to compute the new predicted boxes, and line 6 applies the pruning step to compute the new pruned boxes.

8. Theoretical analysis. This section presents theoretical results on the efficiency of our method and compares it to the best interval methods we are aware of.

8.1. Overview of the methods. We analyze the cost of our SOLVE algorithm based on the GHF method and compare it to Nedialkov's interval Hermite–Obreschkoff (IHO) method [24], the best interval method we know of. Indeed, the IHO method outperforms interval Taylor series methods such as Lohner's method [20]. Here are the various methods used in the theoretical and experimental comparisons.

The GHF method. In the GHF method, each iteration in the loop of function SOLVE is called a *step* of the integration. The (constant) step size in GHF is given by $h = t_k - t_0$. Assuming that $\sigma_m = \max(\sigma)$ and $\sigma_s = \sigma_0 + \cdots + \sigma_k$, the remaining components of GHF are specified as follows:

function SOLVE($\mathbb{O}, D_0, \mathbf{t}_{0..mk-1}$)
begin
1 $\mathbf{B}_0 := \text{BOUNDINGBOX}(\mathbb{O}, t_0, D_0, \mathbf{t}_0)$;
2 $\langle \mathbf{D}_0^*, \mathbf{Y}_0, M_0 \rangle := \text{INITIALIZEMULTISTEP}(\mathbb{O}, \mathbf{t}_0, D_0, \mathbf{B}_{1..k-1})$;
3 **for** $i := 1$ **to** $m-1$ **do**
4 $\mathbf{B}_i := \text{BOUNDINGBOX}(\mathbb{O}, t_{ik-1}, D_{ik-1}^*, \mathbf{t}_i)$;
5 $\mathbf{D}_i^- := \text{PREDICTOR}(\mathbb{O}, t_{ik-1}, D_{ik-1}^*, \mathbf{t}_i, \mathbf{B}_i)$;
6 $\langle \mathbf{D}_i^*, \mathbf{Y}_i, M_i \rangle := \text{PRUNE}(\mathbb{O}, \mathbf{t}_{i-1}, \mathbf{D}_{i-1}^*, \mathbf{B}_{(i-1)k+1..ik-1}, \mathbf{Y}_{i-1}, M_{i-1}, \mathbf{t}_i, \mathbf{D}_i^-, \mathbf{B}_i)$;
7 **endfor**
8 **return** $\mathbf{D}_{1..mk-1}^*$;
end

FIG. 9. *The constraint satisfaction algorithm for IVPs for parametric ODEs.*

SPECIFICATION 4 (SOLVE). *Let s be the solution of ODE $\mathbb{O}$ and $\mathbf{D}_{1..mk-1} = \text{SOLVE}(\mathbb{O}, D_0, \mathbf{t}_{0...mk-1})$. Then, for $1 \le i \le mk-1$, $s(t_0, D_0, t_i) \subseteq D_i$.*
SPECIFICATION 5 (BOUNDINGBOX). *Let $\mathbf{B}_{1..k} = \text{BOUNDINGBOX}(\mathbb{O}, t_0, D_0, \mathbf{t}_{1..k})$. Then, for $1 \le i \le k$, B_i is a bounding box of $\mathbb{O}$ over $[t_{i-1}, t_i]$ w.r.t. (t_0, D_0).*
SPECIFICATION 6 (INITIALIZEMULTISTEP). *Let ms be the multistep solution of ODE $\mathbb{O}$ and B_i be a bounding box of $\mathbb{O}$ over $[t_{i-1}, t_i]$ w.r.t. (t_0, D_0) for $1 \le i \le k-1$. Let*

$$\langle \mathbf{D}_0, \mathbf{Y}_0, M \rangle = \text{INITIALIZEMULTISTEP}(\mathbb{O}, \mathbf{t}_0, D_0, \mathbf{B}_{1..k-1})$$

and $\mathcal{A} = \{M\mathbf{y}_0 + m(\mathbf{D}_0) \mid \mathbf{y}_0 \in \mathbf{Y}_0\} \cap \mathbf{D}_0$. Then, for $0 \le i \le k-1$, $ms(t_0, D_0, t_i) \subseteq ms(\mathbf{t}_0, \mathcal{A}, t_i)$.
SPECIFICATION 7 (PREDICTOR). *Let s be the solution of ODE $\mathbb{O}$ and B_i a bounding box of $\mathbb{O}$ over $[t_{i-1}, t_i]$ w.r.t. (t_0, D_0) for $1 \le i \le k$. Let*

$$\mathbf{D}_{1..k} = \text{PREDICTOR}(\mathbb{O}, t_0, D_0, \mathbf{t}_{1..k}, \mathbf{B}_{1..k}).$$

Then, for $1 \le i \le k$, $s(t_0, D_0, t_i) \subseteq D_i$.

FIG. 10. *The specification of the main functions.*

1. The BOUNDINGBOX function in GHF uses a Taylor series method [21, 5, 25] of order $p+q+1$ to compute $\mathbf{B}_i$. Moreover, we assume that $B_{ik} = \cdots = B_{(i+1)k-1}$; i.e., the function computes a single bounding box over $[t_{ik-1}, t_{(i+1)k-1}]$ $(i \ge 1)$.
2. The PREDICTOR function uses Moore's Taylor method [21] of order $q+1$ to compute the boxes $\mathbf{D}_i^-$. Note that we compute the Taylor coefficients of f only once at (t_{ik-1}, D_{ik-1}^*).
3. The evaluation point in Hermite filters (i.e., in function EMVFL) is the rightmost zero of function γ (see section 6 and Table 1). GHF(σ) is thus a method of order $\sigma_s + 1$.
4. Function EXPLICITGLOBALFILTER needs $\sigma_m - 1$ Jacobians (i.e., $\mathcal{J}(D_j)_1, \ldots, \mathcal{J}(D_j)_{\sigma_m - 1}$) at each interpolation point t_j for $(i-1)k \le j \le (i+1)k - 1$ to compute the k explicit mean-value Hermite filters in EMVFL. GHF computes only Jacobians at predicted boxes and not at pruned boxes. More precisely, it computes only $k(\sigma_m - 1)$ Jacobians at $(\mathbf{t}_i, \mathbf{D}_i^-)$ and reuses the

$k(\sigma_m - 1)$ Jacobians at $(\mathbf{t}_{i-1}, \mathbf{D}_{i-1}^-)$ which were computed during the previous step $i-1$.

5. The function CoordTransfo uses Lohner's QR factorization technique (see [20]).
6. The function InitializeMultistep uses a one-step mean-value Taylor method.

The IHO method. The IHO method is implemented exactly as described in [24]. Its step size is h as in the GHF method. Besides the pruning, there are some interesting differences between GHF and IHO. First, the predictor function in IHO uses a mean-value Taylor method of order $q+1$. Second, the Jacobians in IHO are recomputed at pruned boxes. IHO uses a Taylor series method of order $p+q+1$ to compute a bounding box as in GHF.

The IHO* method. To obtain experimental results as informative as possible, we introduce IHO*, a variant of IHO that is closer to GHF. In particular, the predictor in IHO* uses Moore's Taylor method of order $q+1$ instead of the mean-value Taylor method of the same order. Also, IHO* does not recompute the Jacobians at pruned boxes; it reuses the Jacobians at predicted boxes instead as in GHF. IHO* and GHF differ only in the pruning step. Interestingly, IHO* is extremely close in precision to IHO on almost all benchmarks for a given step size. There are a few benchmarks where the loss of precision is significant or where a smaller step size must be used. Of course, IHO* is faster than IHO for a given step size.

8.2. Comparison hypotheses. We make the following assumptions and conventions for simplicity. Consider the ODE $u' = f(u)$. We assume that (the natural encoding of) function f contains only arithmetic operations. We denote by N_1 the number of $*, /$ operations in f, by N_2 the number of $\pm$ operations, and by N the sum $N_1 + N_2$. We also assume that the cost of evaluating $\mathcal{J}(D_i)_j$ is n times the cost of evaluating $(D_i)_j$. We report only the *main* operations of the methods, i.e., (1) products of a real and an interval matrix which arise in the pruning step and (2) the generation of Jacobians.[12] These are the main operations for problems of sufficiently high dimension where f contains sufficiently many operations. Note that products of a real and an interval matrix can be optimized to substantially reduce the number of sign tests and rounding mode switches, which are costly tasks (see [15]). As a consequence, the cost *per* interval arithmetic operation in a real-interval matrix product is less than the cost of an operation on two intervals in a Jacobian computation. We thus report separately the number of interval arithmetic operations involved in products of a real and an interval matrix in the pruning step (Cost-1) and the generation of Jacobians (Cost-2). Note that Cost-1 is a fixed cost in the sense that it is independent from the ODE. Cost-2 is a variable cost which increases as the expression of f contains more operations.

8.3. Methods of the same order. We first compare the costs of GHF(σ) and IHO$^{(*)}(p,q)$ for $p+q=\sigma_s$ and $q \in \{p, p+1\}$. The methods are thus of order $\sigma_s + 1$. Table 2 reports the main cost of a step in IHO, IHO*, and GHF. It also shows the complexity of two particular cases of GHF: GHF-1 is an implementation with only two interpolation points ($k=1$) and $|\sigma_1 - \sigma_0| \leq 1$, while GHF-2 is an implementation with two conditions on every interpolation points ($\sigma_0 = \cdots = \sigma_k = 2$).

The first main result is that GHF-1 is always cheaper than IHO$^{(*)}$. *Hence a GHF method with only two interpolation points is guaranteed to run faster than IHO$^{(*)}$.* The next section shows that an improvement in accuracy is also obtained in this case.

[12] Matrix inversions and the QR factorization in CoordTransfo are not counted here.

TABLE 2
Cost analysis: Methods of the same order.

	Cost-1	Cost-2
IHO	–	$2\lceil\frac{\sigma_s}{2}\rceil^2 nN_1 + O(\sigma_s nN_2)$
IHO*	–	$\lceil\frac{\sigma_s}{2}\rceil^2 nN_1 + O(\sigma_s nN_2)$
GHF	$7k^3n^3$	$((\sigma_m-1)^2+1)knN_1 + \sigma_m knN_2$
GHF-1	–	$(\lfloor\frac{\sigma_s-1}{2}\rfloor^2+1)nN_1 + O(\sigma_s nN_2)$
GHF-2	$(\frac{7}{8}\sigma_s - \frac{21}{4})\sigma_s^2 n^3$	$(\sigma_s-2)nN$

TABLE 3
Cost analysis: Methods of different orders but of similar cost.

	Cost-2
IHO	$2\lfloor\frac{\sigma_s-1}{2}\rfloor^2 nN_1 + O(\sigma_s nN_2)$
IHO*	$\lfloor\frac{\sigma_s-1}{2}\rfloor^2 nN_1 + O(\sigma_s nN_2)$
GHF-1	$(\lfloor\frac{\sigma_s-1}{2}\rfloor^2+1)nN_1 + O(\sigma_s nN_2)$

Observe that Cost-2 in IHO* is approximately half as much as in IHO because the Jacobians are not computed at pruned boxes in IHO*. Note also that Cost-2 is smaller in GHF-1 than in IHO* because IHO* evaluates one more Jacobian, i.e., $\mathcal{J}(D_i)_q$.

GHF-2 is more expensive than GHF-1 and IHO$^{(*)}$ when f contains few operations because the Jacobians are cheap to compute in this case and the fixed cost Cost-1 becomes large w.r.t. Cost-2. However, when f contains many $*, /$ operations (which is the case in many practical applications), GHF-2 becomes substantially faster because Cost-1 in GHF-2 is independent of f and Cost-2 is substantially smaller in GHF-2 than in GHF-1 and IHO$^{(*)}$. *This result shows the versatility of the approach that can be tailored to the application at hand.*

8.4. One-step methods of different orders but of similar cost. *We now show that GHF methods can be tailored to be asymptotically more precise than IHO methods for a similar cost.* Consider the costs of the IHO$^{(*)}(p,q)$ and GHF-1 methods when we assume that $p+q = \sigma_s - 2$ and $q \in \{p, p+1\}$. Under these conditions, IHO$^{(*)}$ is a method of order $\sigma_s - 1$, while GHF-1 is a method of order $\sigma_s + 1$. Table 3 reports the main cost of a step in IHO, IHO*, and GHF-1. Cost-2 is similar in GHF-1 and IHO* (and about twice as much in IHO). The GHF-1 method is thus asymptotically more precise (by two orders of magnitude) than IHO* for a similar cost.

9. Experimental analysis. We now report experimental results of a C++ implementation[13] of our SOLVE algorithm based on the GHF method GHF(σ). We performed our tests on a Sun Ultra 10 workstation with a 333 MHz UltraSparc CPU. The underlying interval arithmetic and automatic differentiation packages are PROFIL/BIAS [15] and FADBAD/TADIFF [3, 2].

The benchmarks. Many of the benchmarks are standard. They come from various domains, including chemistry, biology, mechanics, physics, and electricity. The equation, initial conditions, and interval of integration for each IVP are given in [12]. Note that the comparisons uses only point initial conditions; they could easily be generalized to interval conditions. The "full Brusselator" (BRUS), the "Oregonator" (OREG), and HIRES all model famous chemical reactions. Both OREG and HIRES are stiff problems. The Lorenz system (LOR) exemplifies the so-called strange at-

[13]The code is available at http://www.info.ucl.ac.be.

tractors. The two-body problem (2BP) comes from mechanics, and the van der Pol (VDP) equation describes an electrical circuit. All these problems are described in detail in [10, 11]. We also consider a problem from molecular biology (BIO) and the Stiff DETEST problem D1 [9]. Finally, we consider four dynamical systems (LIEN, P1, P2, P3), where the function f contains more operations. LIEN, P2, and P3 are taken from [26].

Overview of the experiments. The experimental results obey the same assumptions as the theoretical analysis. They include three types of comparisons:

1. one-step methods of the same order;
2. one-step methods of different orders but of similar cost;
3. multistep versus one-step methods of the same order.

The tables report, for a given step size, the global error, the error ratio (an error ratio higher than 1 means that GHF is more precise), the execution time of both methods (in seconds), and the time ratio (a time ratio higher than 1 means that GHF is faster). They also report the execution time of IHO* between parentheses. As mentioned, we observed small precision loss in IHO* over IHO and only for the larger step sizes. Since this was not very significant, we assume that the error values in IHO* are nearly the same as in IHO. A "-" symbol in the tables means that the method failed to integrate the ODE for the corresponding step size. Finally, note that the global error at point t_i is given by the infinity norm of the width of the enclosure D_i at t_i, i.e., the quantity $\|\omega(D_i)\|_\infty$ at the end of the interval of integration.

9.1. One-step methods.

Same order. Table 4 reports the experimental results for the IHO$^{(*)}(p,p)$ and GHF(p,p) methods of order $2p+1$ on several benchmarks, orders, and step sizes. In general, for a given step size, GHF and IHO* have a similar accuracy and execution time. GHF is usually slightly faster as predicted by the theoretical results. The difference should be larger for higher dimensional problems where f contains many operations. IHO is slower than GHF and IHO*. For a given problem and given order, the error ratio is generally constant w.r.t. the step size, confirming that GHF and IHO$^{(*)}$ are methods of the same order.

Different orders. The theoretical results indicated that, given a step size, the GHF method can always be tailored to be asymptotically more precise than IHO* for a similar computation cost. We now validate this claim experimentally. Table 5 compares IHO(p,p) (order $2p+1$) and GHF$(p+1,p+1)$ (order $2p+3$). On the benchmarks, GHF is always faster than IHO, and it produces significant improvements in accuracy. As expected, the gain in precision increases when the step size decreases, confirming that GHF is a method of higher order than IHO. GHF is slightly slower than IHO*, but, of course, it produces significant improvement in accuracy. GHF and IHO* should have a similar execution time for higher dimensional problems where f contains many operations, as predicted by the theoretical analysis.

Error w.r.t. time. It is interesting to compare the various methods by plotting the error as a function of the execution time. Figure 11 plots IHO$^{(*)}(p,p)$, GHF(p,p), and GHF$(p+1,p+1)$ using the results in Tables 4 and 5. We take $p=8$ for D1 and HIRES and $p=3$ for the other problems. The curve of IHO* is always slightly above the curve of GHF(p,p) (except for D1). GHF$(p+1,p+1)$ is almost always below the other curves, and IHO is always above the other curves. These results confirm the theoretical results and indicate that GHF$(p+1,p+1)$ is superior to the other methods.

TABLE 4
One-step methods of the same order.

IVP	IHO	GHF	h	Error			Time		
	p, q	σ		IHO	GHF	Ratio	IHO	GHF	Ratio
BRUS	3,3	(3,3)	1E-1	2.3E-3	1.2E-3	1.9			
			7.5E-2	4.5E-5	2.4E-5	1.9			
			5E-2	9.7E-7	4.9E-7	2.0			
			2.5E-2	5.2E-9	2.7E-9	1.9			
			1.25E-2	3.2E-11	1.7E-11	1.9			
			1E-2	6.5E-12	3.5E-12	1.9	5.1 (4.0)	3.9	1.3
	4,4	(4,4)	1E-1	1.7E-4	9.9E-5	1.7			
			7.5E-2	2.0E-6	1.1E-6	1.8			
			5E-2	1.0E-8	5.0E-9	2.0			
			2.5E-2	7.4E-12	3.2E-12	2.3	2.8 (2.1)	2.0	1.4
	5,5	(5,5)	1E-1	2.4E-5	1.6E-5	1.5			
			7.5E-2	1.2E-7	7.6E-8	1.6			
			5E-2	1.6E-10	9.4E-11	1.7	1.9 (1.4)	1.3	1.5
	7,7	(7,7)	1E-1	7.6E-7	5.2E-7	1.5			
			7.5E-2	6.6E-10	4.7E-10	1.4	1.9 (1.4)	1.3	1.5
	8,8	(8,8)	1E-1	1.5E-7	1.1E-7	1.4			
			7.5E-2	5.4E-11	4.0E-11	1.4	2.2 (1.6)	1.5	1.5
LOR	3,3	(3,3)	1.25E-2	4.8E-1	3.2E-1	1.5			
			1E-2	6.7E-2	4.5E-2	1.5			
			7.5E-3	7.7E-3	4.9E-3	1.6			
			5E-3	4.3E-4	2.6E-4	1.7			
			2.5E-3	3.1E-6	2.0E-6	1.6	11 (8)	8	1.4
	4,4	(4,4)	2E-2	1.5E-1	1.0E-1	1.5			
			1.75E-2	2.7E-2	1.8E-2	1.5			
			1.5E-2	5.0E-3	3.0E-3	1.7			
			1.25E-2	8.0E-4	4.6E-4	1.7			
			1E-2	9.0E-5	5.0E-5	1.8			
			7.5E-3	6.0E-6	3.1E-6	1.9	4.7 (3.6)	3.6	1.3
	7,7	(7,7)	3E-2	3.0E-3	2.4E-3	1.2			
			2.75E-2	4.5E-4	3.6E-4	1.2			
			2.5E-2	6.6E-5	5.3E-5	1.2			
			2.25E-2	7.7E-6	6.2E-6	1.2	3.0 (2.3)	2.2	1.4
2BP	3,3	(3,3)	1E-1	4.5E-3	7.6E-4	6.0			
			7.5E-2	1.1E-4	3.7E-5	3.0			
			5E-2	3.3E-6	1.2E-6	2.7			
			2.5E-2	1.5E-8	4.5E-9	3.3	3.6 (2.9)	2.6	1.4
	4,4	(4,4)	1.25E-1	2.9E-4	7.4E-5	3.9			
			1E-1	1.2E-5	3.0E-6	4.0			
			7.5E-2	3.4E-7	8.5E-8	4.0			
			5E-2	3.4E-9	9.2E-10	3.7	2.5 (1.9)	1.7	1.5
	7,7	(7,7)	1.5E-1	1.1E-6	5.6E-7	2.0			
			1.25E-1	2.3E-9	9.7E-10	2.4	2.0 (1.5)	1.3	1.5
VDP	3,3	(3,3)	4E-2	1.5E-2	5.8E-3	2.6			
			3E-2	5.9E-5	3.8E-5	1.6			
			2E-2	1.7E-6	9.6E-7	1.8			
			1E-2	1.0E-8	5.3E-9	1.9			
			5E-3	7.4E-11	3.8E-11	1.9			
			2.5E-3	4.7E-13	2.6E-13	1.8	14 (11.2)	11.6	1.2
	4,4	(4,4)	4E-2	4.7E-5	4.0E-5	1.2			
			3E-2	8.4E-7	5.1E-7	1.6			
			2E-2	9.0E-9	4.5E-9	2.0			
			1E-2	1.1E-11	4.7E-12	2.3	4.5 (3.7)	3.8	1.2
	5,5	(5,5)	4E-2	2.6E-6	2.1E-6	1.2			
			3E-2	2.3E-8	1.6E-8	1.4			
			2E-2	6.7E-11	3.9E-11	1.7	2.9 (2.3)	2.4	1.3
BIO	3,3	(3,3)	7.5E-3	4.6E-6	2.0E-6	2.3			
			5E-3	8.2E-9	3.4E-9	2.4			
			2.5E-3	2.2E-11	9.2E-12	2.4	7.0 (5.4)	5.1	1.4
	4,4	(4,4)	7.5E-3	1.3E-6	7.6E-7	1.7			
			5E-3	2.9E-10	1.3E-10	2.2			
			2.5E-3	9.7E-14	3.3E-14	2.9	10 (7.5)	7.0	1.4
OREG	3,3	(3,3)	1.5E-2	1.5E-4	2.2E-4	0.7			
			1E-2	8.0E-6	1.1E-5	0.7			
			7.5E-3	1.0E-6	1.4E-6	0.7			
			5E-3	6.0E-8	7.9E-8	0.8	9.6 (7.7)	7.5	1.3
	4,4	(4,4)	2.5E-2	2.4E-4	3.4E-4	0.7			
			2E-2	1.2E-5	1.6E-5	0.7			
			1.5E-2	6.1E-7	7.6E-7	0.8			
			1E-2	1.5E-8	1.9E-8	0.8			
			7.5E-3	1.1E-9	1.4E-9	0.8	8.2 (6.5)	6.4	1.3
D1	8,8	(8,8)	1.1E-1	1.1E-6	1.3E-6	0.8			
			1E-1	1.3E-7	1.4E-7	0.9			
			9E-2	1.5E-8	1.7E-8	0.9			
			8E-2	1.5E-9	1.7E-9	0.9			
			7E-2	1.3E-10	1.4E-10	0.9			
			6E-2	7.3E-12	8.3E-12	0.9			
			5E-2	2.8E-13	3.1E-13	0.9	2.4 (1.8)	1.9	1.3
HIRES	4,4	(4,4)	2.5E-1	3.2E-7	6.1E-7	0.5			
			2E-1	2.4E-8	4.3E-8	0.6			
			1.5E-1	1.1E-9	2.6E-9	0.4			
			1E-1	2.8E-11	5.0E-11	0.6			
			5E-2	4.8E-14	6.9E-14	0.7	23 (17)	16	1.4
	8,8	(8,8)	4E-1	2.9E-6	1.2E-5	0.2			
			3.5E-1	4.9E-8	3.9E-8	1.3			
			3E-1	8.0E-10	6.2E-10	1.3			
			2.5E-1	7.7E-12	6.0E-12	1.3			
			2E-1	3.4E-14	2.8E-14	1.2	10.9 (7.4)	7.2	1.5

TABLE 5
One-step methods of different orders.

IVP	IHO p,q	GHF σ	h	Error IHO	Error GHF	Error Ratio	Time IHO	Time GHF	Time Ratio
BRUS	3,3	(4,4)	1E-1	2.3E-3	1.0E-3	2.3			
			7.5E-2	4.5E-5	1.3E-5	3.5			
			5E-2	9.7E-7	1.2E-7	8.1			
			2.5E-2	5.2E-9	9.5E-11	55			
			1.25E-2	3.2E-11	2.0E-13	160	4.0 (3.2)	3.6	1.1
	4,4	(5,5)	1E-1	1.7E-4	1.0E-4	1.7			
			7.5E-2	2.0E-6	9.9E-7	2.0			
			5E-2	1.0E-8	3.2E-9	3.1			
			2.5E-2	7.4E-12	6.4E-13	12	2.8 (2.1)	2.4	1.2
LOR	3,3	(4,4)	1.25E-2	4.8E-1	1.3E-2	1.5			
			1E-2	6.7E-2	1.2E-3	56			
			7.5E-3	7.7E-3	5.7E-5	135			
			5E-3	4.3E-4	9.7E-7	443	5.4 (4.0)	4.9	1.1
	4,4	(5,5)	2E-2	1.5E-1	6.2E-2	2.4			
			1.75E-2	2.7E-2	9.0E-3	3.0			
			1.5E-2	5.0E-3	1.2E-3	4.2			
			1.25E-2	8.0E-4	1.2E-4	6.7			
			1E-2	9.0E-5	7.2E-6	13			
			7.5E-3	6.0E-6	2.6E-7	23	4.7 (3.6)	4.1	1.1
2BP	3,3	(4,4)	1E-1	4.5E-3	2.5E-5	180			
			7.5E-2	1.1E-4	7.6E-7	145			
			5E-2	3.3E-6	8.9E-9	371			
			2.5E-2	1.5E-8	4.1E-11	366	3.6 (2.9)	3.0	1.2
	4,4	(5,5)	1.25E-1	2.9E-4	1.1E-5	26			
			1E-1	1.2E-5	3.6E-7	33			
			7.5E-2	3.4E-7	5.6E-9	61			
			5E-2	3.4E-9	5.5E-11	62	2.5 (1.9)	2.0	1.3
VDP	3,3	(4,4)	4E-2	1.5E-2	2.5E-3	6.0			
			3E-2	5.9E-5	9.7E-6	6.1			
			2E-2	1.7E-6	8.8E-8	19			
			1E-2	1.0E-8	6.2E-11	161			
			5E-3	7.4E-11	9.0E-14	822	7.4 (5.6)	7.2	1.0
	4,4	(5,5)	4E-2	4.7E-5	3.6E-5	1.3			
			3E-2	8.4E-7	3.6E-7	2.3			
			2E-2	9.0E-9	1.6E-9	5.6			
			1E-2	1.1E-11	2.8E-13	39	4.5 (3.7)	4.2	1.1
BIO	3,3	(4,4)	7.5E-3	4.6E-6	1.7E-6	2.7			
			5E-3	8.2E-9	1.2E-9	6.8			
			2.5E-3	2.2E-11	4.8E-13	46	7.0 (5.4)	6.2	1.1
	4,4	(5,5)	7.5E-3	1.3E-6	7.7E-7	1.7			
			5E-3	2.9E-10	9.3E-11	3.1			
			2.5E-3	9.7E-14	1.0E-14	9.7	10 (7.5)	8.4	1.2
OREG	3,3	(4,4)	2E-2	2.6E-3	7.0E-5	37			
			1.5E-2	1.5E-4	1.1E-6	136			
			1E-2	8.0E-6	2.2E-8	364			
			7.5E-3	1.0E-6	1.5E-9	667			
			5E-3	6.0E-8	4.6E-11	1304	9.6 (7.7)	8.6	1.1
	4,4	(5,5)	2.5E-2	2.4E-4	1.4E-4	1.7			
			2E-2	1.2E-5	3.9E-6	3.1			
			1.5E-2	6.1E-7	1.6E-8	38			
			1E-2	1.5E-8	6.3E-11	238	6.2 (4.9)	5.3	1.2
D1	8,8	(9,9)	1.1E-1	1.1E-6	3.9E-8	28			
			1E-1	1.3E-7	3.6E-9	36			
			9E-2	1.5E-8	3.5E-10	43			
			8E-2	1.5E-9	2.9E-11	53			
			7E-2	1.3E-10	1.8E-12	72			
			6E-2	7.3E-12	7.8E-14	94	2.0 (1.5)	1.8	1.1
HIRES	4,4	(5,5)	3E-1	1.3E-5	1.9E-6	6.8			
			2.5E-1	3.2E-7	6.0E-8	5.3			
			2E-1	2.4E-8	2.4E-9	10			
			1.5E-1	1.1E-9	4.6E-11	24			
			1E-1	2.8E-11	3.2E-13	88	12 (8.5)	9.3	1.3
	8,8	(9,9)	4E-1	2.9E-6	2.5E-5	0.1			
			3.5E-1	4.9E-8	4.1E-8	1.2			
			3E-1	8.0E-10	6.5E-10	1.2			
			2.5E-1	7.7E-12	6.2E-12	1.2			
			2E-1	3.4E-14	2.9E-14	1.2	10.9 (7.4)	7.9	1.4

9.2. Multistep versus one-step methods. We now compare multistep GHF methods versus IHO$^{(*)}$ and the one-step GHF method of the same order. We restrict our attention to problems where the function f contains more operations. Tables 6, 7, 8, and 9 report the results, respectively, for the four tested examples and for several orders and step sizes.[14] For a given step size, multistep GHF methods usually produce much more precise results than one-step methods (especially for large step sizes); they also allow for larger step sizes. Multistep GHF methods are generally as fast as the one-step GHF method and IHO*; they are faster when f has many operations, as is

[14]Note that in the LIEN problem, we used a bounding box computation method of order 13 for $\sigma_s \geq 12$.

the case in LIEN (which contains many multiplications). The tables also show that, for a given step size, the one-step GHF method is slightly more precise and faster than IHO* and that IHO is slower.

Figures 12, 13, 14, and 15 plot the error as a function of the execution time. The main result is that multistep GHF methods perform better than one-step methods on these problems. In general, multistep methods produce several orders of magnitude improvements in precision for a fixed execution time. The one-step GHF method performs slightly better than IHO*. Note that, for the LIEN problem, GHF methods with many interpolation points are more efficient and allow for smaller execution times.

9.3. Discussion. Before concluding this section, it is important to make a number of remarks.

In GHF, the enhancement in precision obtained by recomputing the Jacobians at pruned boxes is insignificant in all problems we tested. Instead, this recomputation increases the computational cost. Our experimental results showed that this also holds for the IHO method in general.

As pointed out by Nedialkov [23], the stability of interval methods depends not only on the stability of the underlying approximation formula (as in standard numerical methods) but also on the corresponding formula for the truncation error. Hence, interval extensions of standard numerical methods designed for stiff problems may need smaller step sizes. Another restriction on the step size in interval methods comes from the bounding box process, whose current implementations require very small step sizes to be able to compute bounding boxes in the case of stiff problems. This explains why the differences in efficiency between interval methods are not as sharp as for traditional methods.

In our experiments, we always chose $\sigma_0 = \cdots = \sigma_k$. Indeed, the main cost of the method is determined by $\max_{0 \leq i \leq k}\{\sigma_i\}$, and the order of the method is maximized when $\sigma_0 = \cdots = \sigma_k$. Since the actual step sizes are sufficiently small, this choice is thus always better. If we could use larger step sizes (e.g., by improving the bounding box process), then stability requirements might make other choices preferable.

The results close to machine precision are not very significant since rounding errors, not the actual method, are determining the accuracy. This explains why the curves in the figures tend to join for high precisions in some cases (e.g., in LIEN, P1, and P2).

9.4. Summary. We now summarize our experimental results. The main conclusions are as follows:

1. The one-step GHF method is almost always better than existing (one-step) interval methods.
2. When f contains few operations, the one-step GHF method outperforms multistep GHF methods (and other existing methods).
3. When f contains many operations, multistep GHF methods outperform the one-step GHF method (and other existing methods).
4. GHF methods are very versatile and can be tailored to the application at hand.
5. The experimental results confirm the theoretical analysis.

In particular, the one-step GHF method performs generally better than the IHO* method, a variant of Nedialkov's IHO method we proposed and which performed better than the original method on almost all our benchmarks. For low dimensional problems or when f contains few operations, the one-step GHF method is only slightly

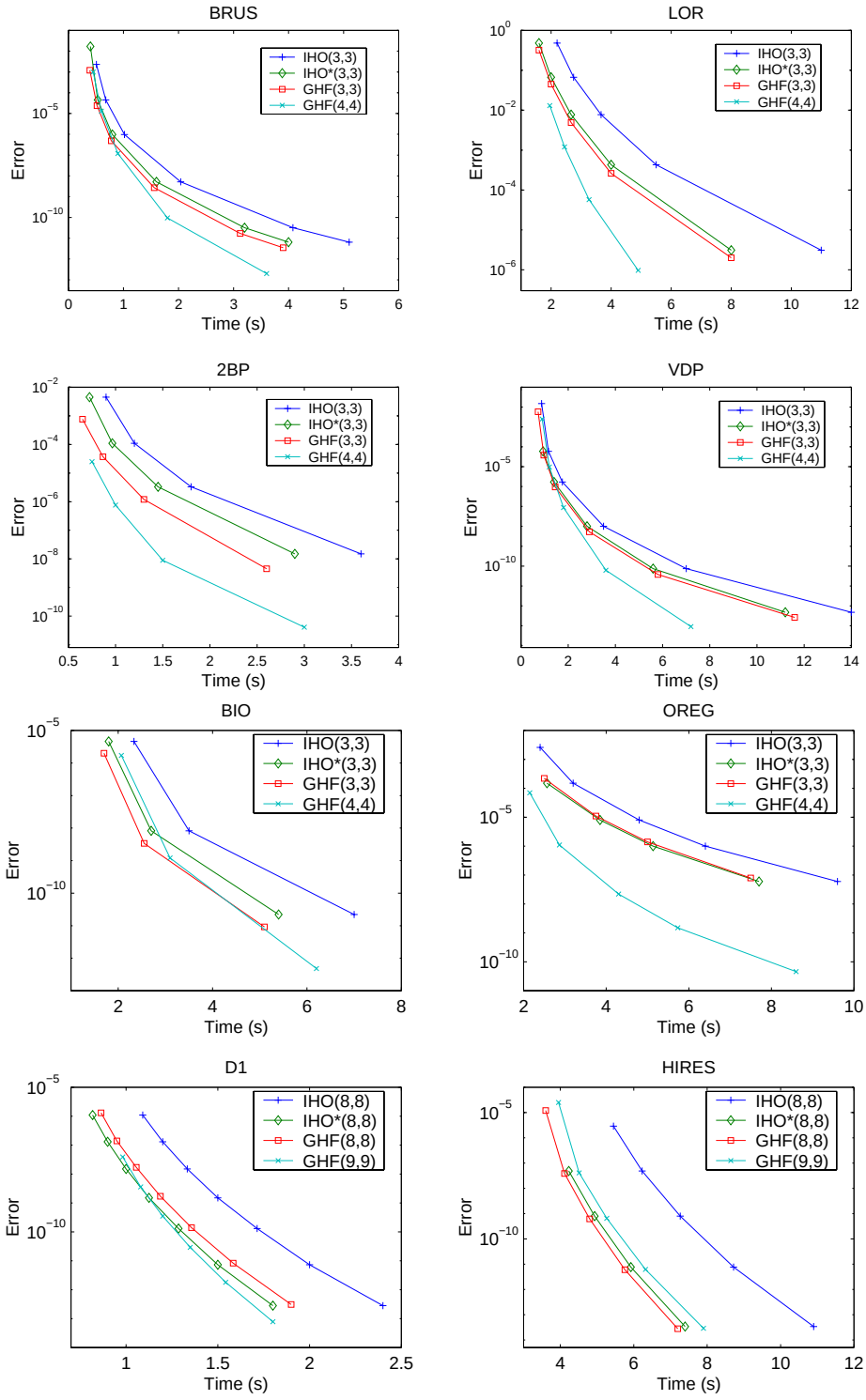


FIG. 11. *Comparison of the methods* $IHO^{(*)}(p,p)$, $GHF(p,p)$ *and* $GHF(p+1,p+1)$ *for the problems BRUS, LOR, 2BP, VDP, BIO, OREG, D1, and HIRES.*

TABLE 6
Multistep versus one-step methods: The LIEN problem.

IVP	IHO	GHF	h	Error			Time		
	p, q	σ		IHO	GHF	Ratio	IHO	GHF	Ratio
LIEN	3,3	(3,3)	5E-1	8.8E-7	7.2E-7	1.2			
			4E-1	1.4E-8	1.1E-8	1.3			
			3E-1	8.4E-10	4.7E-9	0.2			
			2E-1	2.3E-11	5.4E-11	0.4			
			1E-1	1.3E-13	1.3E-13	1.0			
			5E-2	8.7E-16	8.8E-16	1.0	8.3 (6.7)	6.1	1.4
	3,3	(2,2,2)	5.5E-1	-	2.1E-6	-			
			5E-1	8.8E-7	2.5E-7	3.5			
			4E-1	1.4E-8	3.0E-9	4.7			
			3E-1	8.4E-10	1.9E-10	4.4			
			2E-1	2.3E-11	6.9E-12	3.3			
			1E-1	1.3E-13	3.7E-14	3.5			
			5E-2	8.7E-16	2.6E-16	3.3	8.3 (6.7)	6.3	1.3
	4,4	(4,4)	5E-1	2.5E-7	2.0E-7	1.3			
			4E-1	1.0E-9	7.6E-10	1.3			
			3E-1	1.9E-11	1.4E-11	1.4			
			2E-1	1.1E-13	8.3E-14	1.3			
			1E-1	8.3E-17	6.5E-17	3.5	6.1 (4.8)	4.4	1.4
	4,4	(2,2,2,2)	5.8E-1	-	6.7E-8	-			
			5.5E-1	-	8.5E-9	-			
			5E-1	2.5E-7	7.2E-9	35			
			4E-1	1.0E-9	5.0E-11	20			
			3E-1	1.9E-11	1.1E-12	17			
			2E-1	1.1E-13	9.9E-15	11			
			1E-1	8.3E-17	3.8E-17	2.2	6.1 (4.8)	4.6	1.3
	5,5	(5,5)	5E-1	1.2E-7	9.4E-8	1.3			
			4E-1	1.2E-10	9.1E-11	1.3			
			3E-1	7.4E-13	5.7E-13	1.3			
			2E-1	9.9E-16	7.2E-16	1.4	4.2 (3.3)	3.0	1.4
	5,5	(2,2,2,2,2)	5.8E-1	-	6.3E-9	-			
			5.5E-1	-	8.2E-10	-			
			5E-1	1.2E-7	9.3E-11	1290			
			4E-1	1.2E-10	2.0E-12	60			
			3E-1	7.4E-13	2.4E-14	31			
			2E-1	9.9E-16	1.0E-16	10	4.2 (3.3)	3.1	1.3
	6,6	(6,6)	5E-1	7.2E-8	6.0E-8	1.2			
			4.5E-1	3.5E-10	2.9E-10	1.2			
			4E-1	1.7E-11	1.4E-11	1.2			
			3.5E-1	9.1E-13	7.4E-13	1.2			
			3E-1	4.0E-14	3.3E-14	1.2	3.7 (2.9)	2.7	1.4
	6,6	(4,4,4)	5.5E-1	-	1.2E-7	-			
			5E-1	7.2E-8	2.0E-10	360			
			4.5E-1	3.5E-10	1.8E-11	19			
			4E-1	1.7E-11	1.3E-12	13			
			3.5E-1	9.1E-13	9.0E-14	10			
			3E-1	4.0E-14	3.9E-15	10	3.7 (2.9)	2.7	1.4
	6,6	(3,3,3,3)	5.8E-1	-	3.8E-8	-			
			5.5E-1	-	8.6E-10	-			
			5E-1	7.2E-8	3.9E-10	185			
			4.5E-1	3.5E-10	3.0E-11	12			
			4E-1	1.7E-11	6.0E-13	28			
			3.5E-1	9.1E-13	4.8E-14	19			
			3E-1	4.0E-14	2.3E-15	17	3.7 (2.9)	2.6	1.4
	6,6	(2,2,2,2,2,2)	6E-1	-	1.5E-8	-			
			5.5E-1	-	1.7E-10	-			
			5E-1	7.2E-8	1.4E-11	5143			
			4.5E-1	3.5E-10	1.6E-12	219			
			4E-1	1.7E-11	1.4E-13	121			
			3.5E-1	9.1E-13	1.3E-14	70			
			3E-1	4.0E-14	9.0E-16	44	3.7 (2.9)	2.8	1.3
	8,8	(8,8)	5E-1	2.5E-8	2.1E-8	1.2			
			4.5E-1	2.5E-11	2.1E-11	1.2			
			4E-1	5.1E-13	4.3E-13	1.2			
			3.5E-1	1.1E-14	9.7E-15	1.2			
			3E-1	2.0E-16	1.7E-16	1.2	5.1 (3.7)	3.4	1.5
	8,8	(4,4,4,4)	5.8E-1	-	1.2E-8	-			
			5.5E-1	-	8.9E-11	-			
			5E-1	2.5E-8	1.7E-11	1471			
			4.5E-1	2.5E-11	6.7E-13	37			
			4E-1	5.1E-13	6.8E-15	75			
			3.5E-1	1.1E-14	4.1E-16	27	4.4 (3.2)	2.8	1.6
	9,9	(9,9)	5E-1	1.5E-8	1.3E-8	1.2			
			4.5E-1	7.2E-12	6.2E-12	1.2			
			4E-1	9.7E-14	8.3E-14	1.2			
			3.5E-1	1.4E-15	1.2E-15	1.2	5.1 (3.7)	3.4	1.5
	9,9	(6,6,6)	5.5E-1	-	1.9E-8	-			
			5E-1	1.5E-8	5.3E-12	2830			
			4.5E-1	7.2E-12	1.6E-13	45			
			4E-1	9.7E-14	4.0E-15	24			
			3.5E-1	1.4E-15	1.3E-16	11	5.1 (3.7)	3.2	1.6
	9,9	(3,3,3,3,3,3)	6E-1	-	3.5E-7	-			
			5.5E-1	-	2.5E-11	-			
			5E-1	1.5E-8	7.5E-13	20000			
			4.5E-1	7.2E-12	2.2E-13	33			
			4E-1	9.7E-14	2.0E-14	4.8			
			3.5E-1	1.4E-15	2.5E-14	0.06	5.1 (3.7)	3.1	1.6

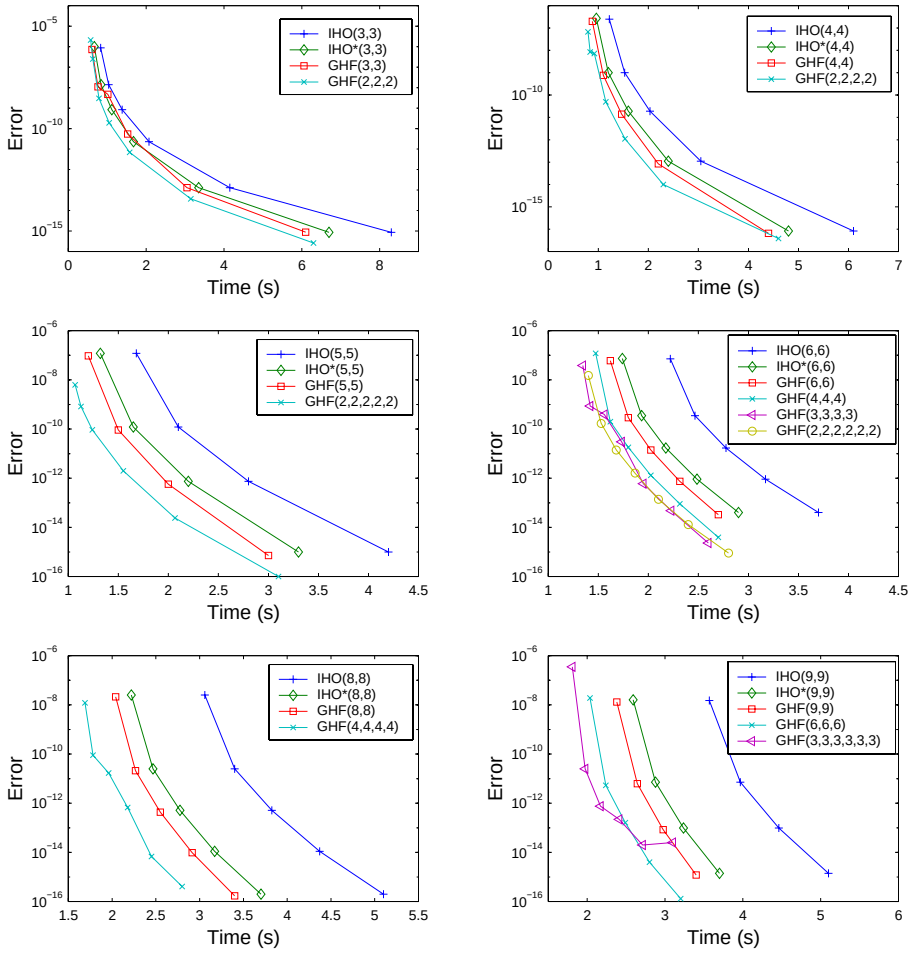


FIG. 12. *Multistep versus one-step methods: The LIEN problem.*

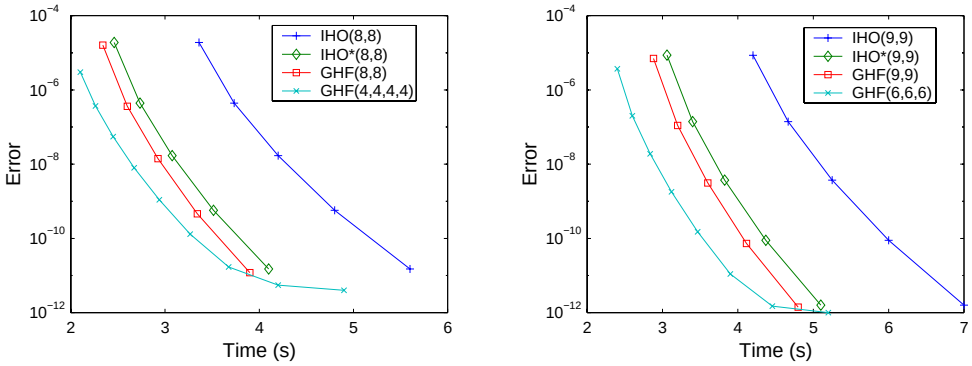


FIG. 13. *Multistep versus one-step methods: The P1 problem.*

TABLE 7
Multistep versus one-step methods: The P1 problem.

IVP	IHO p, q	GHF σ	h	Error IHO	Error GHF	Error Ratio	Time IHO	Time GHF	Time Ratio
P1	3,3	(3,3)	5E-2	8.1E-5	5.2E-5	1.6			
			4E-2	4.0E-6	2.3E-6	1.7			
			3E-2	2.0E-7	1.1E-7	1.8			
			2E-2	6.1E-9	2.9E-9	2.1			
			1E-2	3.3E-11	1.4E-11	2.4			
			5E-3	2.4E-13	9.7E-14	2.5	27 (22)	21	1.3
	3,3	(2,2,2)	6.5E-2	-	1.7E-4	-			
			6E-2	-	3.2E-5	-			
			5E-2	8.1E-5	2.9E-6	28			
			4E-2	4.0E-6	2.8E-7	14			
			3E-2	2.0E-7	2.1E-8	9.5			
			2E-2	6.1E-9	7.9E-10	7.8			
			1E-2	3.3E-11	4.0E-12	8.4			
			5E-3	2.4E-13	3.3E-14	7.3	27 (22)	23	1.2
	6,6	(6,6)	5E-2	6.3E-7	4.8E-7	1.3			
			4E-2	4.8E-9	3.7E-9	1.3			
			3E-2	1.7E-11	1.3E-11	1.3			
			2E-2	1.2E-14	9.3E-15	1.3	19 (13.4)	12.8	1.5
	6,6	(4,4,4)	7E-2	-	3.6E-5	-			
			6E-2	-	3.2E-7	-			
			5E-2	6.3E-7	8.5E-9	74			
			4E-2	4.8E-9	1.4E-10	34			
			3E-2	1.7E-11	9.7E-13	18			
			2E-2	1.2E-14	7.0E-15	1.7	19 (13.4)	13.9	1.4
	6,6	(3,3,3,3)	7.5E-2	-	5.9E-5	-			
			7E-2	-	3.1E-6	-			
			6E-2	-	9.7E-8	-			
			5E-2	6.3E-7	3.2E-9	197			
			4E-2	4.8E-9	6.2E-11	78			
			3E-2	1.7E-11	4.9E-13	35			
			2E-2	1.2E-14	2.9E-14	0.4	19 (13.4)	15.4	1.2
	8,8	(8,8)	6E-2	1.2E-4	9.9E-5	1.2			
			5.5E-2	6.8E-7	5.4E-7	1.3			
			5E-2	3.5E-8	2.8E-8	1.3			
			4.5E-2	1.9E-9	1.5E-9	1.3			
			4E-2	8.1E-11	6.4E-11	1.3			
			3.5E-2	2.7E-12	2.2E-12	1.2			
			3E-2	6.6E-14	5.4E-14	1.2	19 (13.5)	12.8	1.5
	8,8	(4,4,4,4)	7.5E-2	-	3.7E-6	-			
			7E-2	-	1.8E-7	-			
			6.5E-2	-	2.3E-8	-			
			6E-2	1.2E-4	3.2E-9	37500			
			5.5E-2	6.8E-7	4.1E-10	1659			
			5E-2	3.5E-8	4.8E-11	729			
			4.5E-2	1.9E-9	4.6E-12	413			
			4E-2	8.1E-11	3.7E-13	219			
			3.5E-2	2.7E-12	5.4E-14	50			
			3E-2	6.6E-14	3.5E-14	1.9	19 (13.5)	14	1.4
	9,9	(9,9)	6E-2	4.5E-5	3.7E-5	1.2			
			5.5E-2	2.1E-7	1.7E-7	1.2			
			5E-2	8.6E-9	6.9E-9	1.2			
			4.5E-2	3.4E-10	2.7E-10	1.3			
			4E-2	1.1E-11	8.6E-12	1.3			
			3.5E-2	2.6E-13	2.1E-13	1.2	19 (13.9)	13.4	1.4
	9,9	(6,6,6)	7E-2	-	1.3E-6	-			
			6.5E-2	-	6.0E-8	-			
			6E-2	4.5E-5	5.6E-9	8393			
			5.5E-2	2.1E-7	5.1E-10	412			
			5E-2	8.6E-9	4.1E-11	210			
			4.5E-2	3.4E-10	2.6E-12	131			
			4E-2	1.1E-11	1.5E-13	73			
			3.5E-2	2.6E-13	1.3E-14	20	19 (13.9)	13.6	1.4

better than IHO*. For higher dimensional problems where f contains many operations, the one-step GHF method is asymptotically more precise (by two orders of magnitude) than IHO* for the same cost. When f contains few operations, the one-step GHF method is more effective than multistep GHF methods which have a relatively high fixed cost. When f contains many operations, multistep GHF methods perform better than one-step methods. They may produce orders of magnitude improvements in accuracy for a given execution time. Alternatively, they may reduce computation times substantially for a given precision since they avoid expensive Jacobian computations. Finally note that, although our implementation used a constant order and step size, it can be easily enhanced to incorporate standard order and step size control strategies, e.g., Eijgenraam's [8] or Nedialkov's [23] techniques.

10. Conclusion. This paper described a constraint satisfaction approach to IVPs for parametric ODEs (i.e., ODEs where some data or initial conditions are

TABLE 8
Multistep versus one-step methods: The P2 problem.

IVP	IHO	GHF	h	Error			Time		
	p, q	σ		IHO	GHF	Ratio	IHO	GHF	Ratio
P2	8,8	(8,8)	1E-1	1.9E-5	1.6E-5	1.2			
			9E-2	4.4E-7	3.6E-7	1.2			
			8E-2	1.7E-8	1.4E-8	1.2			
			7E-2	5.7E-10	4.6E-10	1.2			
			6E-2	1.5E-11	1.2E-11	1.2	5.6 (4.1)	3.9	1.4
	8,8	(4,4,4,4)	1.4E-1	-	3.0E-6	-			
			1.3E-1	-	3.7E-7	-			
			1.2E-1	-	5.5E-8	-			
			1.1E-1	-	8.0E-9	-			
			1E-1	1.9E-5	1.1E-9	17273			
			9E-2	4.4E-7	1.3E-10	3385			
			8E-2	1.7E-8	1.7E-11	1000			
			7E-2	5.7E-10	5.5E-12	104			
			6E-2	1.5E-11	4.0E-12	3.7	5.6 (4.1)	4.9	1.1
	9,9	(9,9)	1E-1	8.5E-6	7.0E-6	1.2			
			9E-2	1.4E-7	1.1E-7	1.3			
			8E-2	3.7E-9	3.1E-9	1.2			
			7E-2	8.9E-11	7.3E-11	1.2			
			6E-2	1.6E-12	1.4E-12	1.1	7.0 (5.1)	4.8	1.5
	9,9	(6,6,6)	1.3E-1	-	3.7E-6	-			
			1.2E-1	-	2.0E-7	-			
			1.1E-1	-	1.9E-8	-			
			1E-1	8.5E-6	1.8E-9	4722			
			9E-2	1.4E-7	1.5E-10	933			
			8E-2	3.7E-9	1.1E-11	336			
			7E-2	8.9E-11	1.5E-12	59			
			6E-2	1.6E-12	1.0E-12	1.6	7.0 (5.1)	5.2	1.3

TABLE 9
Multistep versus one-step methods: The P3 problem.

IVP	IHO	GHF	h	Error			Time		
	p, q	σ		IHO	GHF	Ratio	IHO	GHF	Ratio
P3	4,4	(4,4)	5E-1	1.9E-3	1.4E-3	1.4			
			4E-1	4.0E-6	2.7E-6	1.5			
			3E-1	6.2E-8	3.9E-8	1.6			
			2E-1	3.4E-10	2.0E-10	1.7			
			1E-1	2.7E-13	9.4E-14	2.9	3.5 (2.7)	2.5	1.4
	4,4	(2,2,2,2)	6.5E-1	-	9.1E-5	-			
			6E-1	-	1.1E-5	-			
			5E-1	1.9E-3	7.0E-7	2714			
			4E-1	4.0E-6	4.3E-8	93			
			3E-1	6.2E-8	1.5E-9	43			
			2E-1	3.4E-10	1.5E-11	23			
			1E-1	2.7E-13	1.6E-14	17	3.5 (2.7)	3.3	1.1
	8,8	(8,8)	5E-1	2.6E-5	2.1E-5	1.2			
			4.5E-1	1.5E-7	1.2E-7	1.2			
			4E-1	5.4E-9	4.4E-9	1.2			
			3.5E-1	1.7E-10	1.4E-10	1.2			
			3E-1	3.7E-12	3.0E-12	1.2	3.3 (2.4)	2.2	1.5
	8,8	(4,4,4,4)	6.8E-1	-	8.9E-5	-			
			6.5E-1	-	8.3E-7	-			
			6E-1	-	4.8E-8	-			
			5.5E-1	-	6.4E-9	-			
			5E-1	2.6E-5	7.6E-10	34211			
			4.5E-1	1.5E-7	8.8E-11	1705			
			4E-1	5.4E-9	7.9E-12	684			
			3.5E-1	1.7E-10	5.4E-13	315			
			3E-1	3.7E-12	5.1E-14	73	3.3 (2.4)	2.5	1.3
	9,9	(9,9)	5E-1	1.0E-5	8.2E-6	1.2			
			4.5E-1	4.3E-8	3.5E-8	1.2			
			4E-1	1.1E-9	9.2E-10	1.2			
			3.5E-1	2.4E-11	2.0E-11	1.2			
			3E-1	3.5E-13	2.9E-13	1.2	3.9 (2.9)	2.7	1.4
	9,9	(6,6,6)	6E-1	-	6.8E-7	-			
			5.5E-1	-	2.0E-8	-			
			5E-1	1.0E-5	1.6E-9	6250			
			4.5E-1	4.3E-8	1.2E-10	358			
			4E-1	1.1E-9	7.1E-12	155			
			3.5E-1	2.4E-11	3.0E-13	80			
			3E-1	3.5E-13	1.4E-14	25	3.9 (2.9)	2.9	1.3

uncertain and given by intervals). *The main novelty of the constraint satisfaction approach is to introduce, inside traditional interval methods, a pruning component which reduces the size of the predicted boxes by using relaxations of the ODE (also called filters).* Then we presented an effective pruning algorithm which uses (1) relaxations of the ODE based Hermite interpolation polynomials and enclosures of their error terms; (2) a globalization process to reduce variable dependency problems and evaluation points that minimize the local error of the relaxations. The pruning component

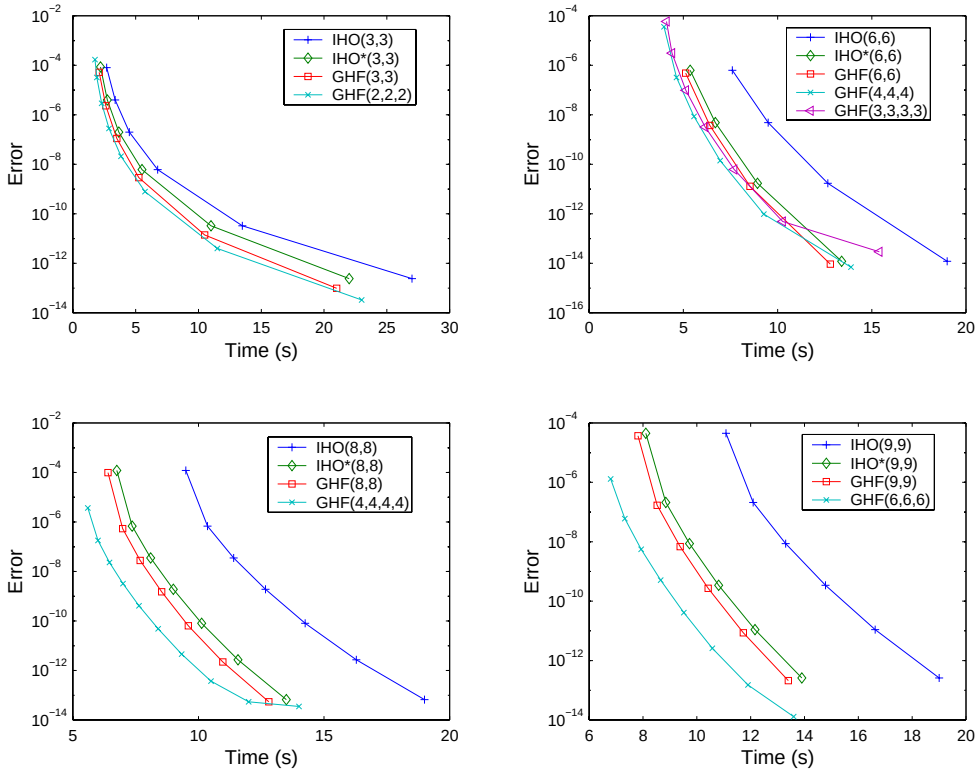


FIG. 14. *Multistep versus one-step methods: The P2 problem.*

was integrated in an integration algorithm which also uses traditional techniques to handle the wrapping effect.

The novel integration algorithm was analyzed both theoretically and experimentally. The theoretical results indicate that, for the same computation costs, our algorithm provides quadratic (asymptotic) improvement in accuracy over the best interval method we know of. They also show that our algorithm is significantly faster when the ODE contains many operations. Experimental results on a variety of standard and new benchmarks validated the theoretical results. The algorithm shows significant gains in accuracy, while not degrading computational performance. The experimental results also illustrate that the approach could produce significant gain in computation time when the ODE contains many operations.

It is also important to stress the versatility of our algorithm and of our approach. On the one hand, GHFs can be tailored to the problem at hand by choosing the number of interpolation points as well as the number of derivative conditions imposed at each interpolation point. On the other hand, the pruning algorithm itself is generic, and new pruning techniques may easily be incorporated.

There are a wealth of topics for further research:

1. The current algorithm can be enhanced in many ways to include, for instance, order and step size control strategies, and the automatic selection of the number of interpolation points and the number of derivative conditions imposed at each interpolation point.
2. The constraint satisfaction approach is clearly in its infancy and new relax-

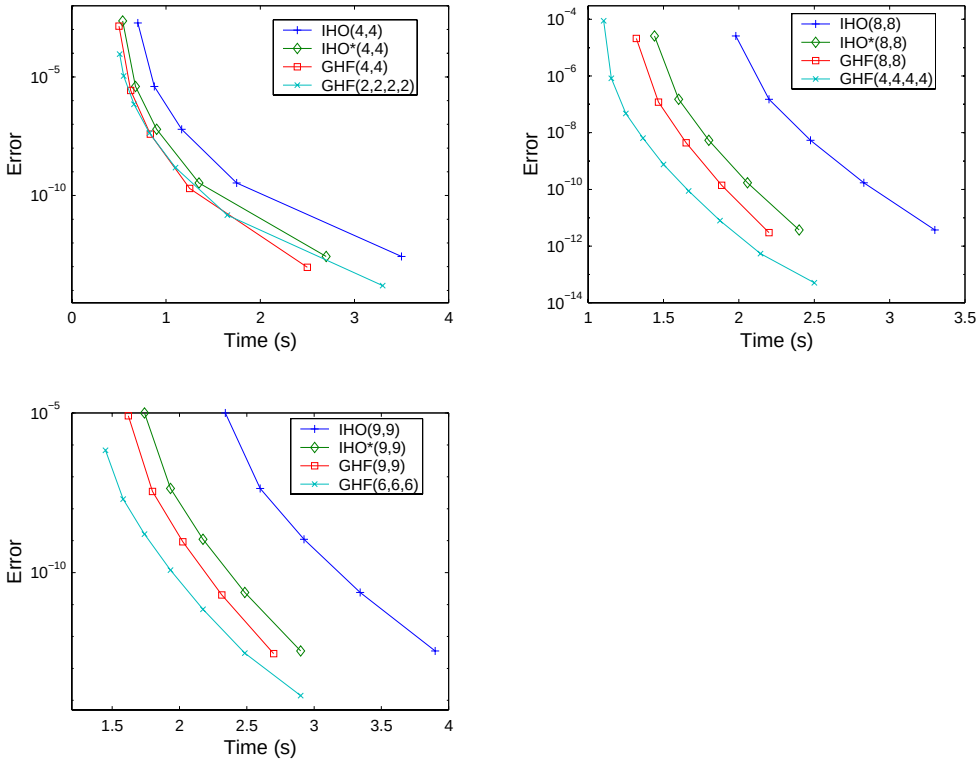


FIG. 15. *Multistep versus one-step methods: The P*3 *problem.*

ations (e.g., using splines, trigonometric interpolation, Legendre, Chebyshev, and Laguerre polynomials) should be investigated.

3. Compared to standard numerical methods, validated methods generally use smaller step sizes, and stiff problems are particularly challenging. The main factors that limit the step size are the need to enclose error terms and the bounding box process. Finding efficient bounding box techniques is probably the main bottleneck at this point, and it would be interesting to study how pruning techniques could help in this respect. Once we will be able to increase the step size, it will be important to analyze the stability of our approach and to compare it to the stability of other validated methods. The choice of many of the parameters mentioned in point (1) will be guided by stability requirements in the case of stiff problems. Furthermore, our asymptotic theory for choosing an optimal evaluation time may not be valid anymore, and we may have to find new techniques for choosing a good evaluation time.
4. A possible alternative to validated methods consists of dropping the enclosures of the error terms and the bounding box process in the interval method. We can thus keep the parametric aspect of the ODEs, but we lose the validated aspect of the method. However, the advantage is that larger step sizes can be used in this case. From our experimental results, we can expect a higher gain in performance of our GHF method over the IHO$^{(*)}$ method for those larger step sizes. In addition, if we consider an ODE for which it is not possible to compute the Taylor coefficients $(u)_2, (u)_3, \ldots$ of the solution,

a multistep GHF(σ) method with $\sigma_i \leq 2,\ i = 0, \ldots, k$, is the *only* interval method (we know of) which is able to integrate the ODE, since it does not need any Taylor coefficient.

5. A very promising direction of further research is the application of our approach to standard numerical methods for ODEs. Indeed, to our knowledge, the idea of evaluating a Hermite filter at a point which is different from the point at which the current value is computed is completely new. We can apply our asymptotic theory for the choice of an optimal evaluation time in the case of nonstiff problems. For stiff problems, the choice of a good evaluation time will be guided by stability requirements. Note that when $\sigma = (1, \ldots, 1)$, i.e., the Hermite interpolation polynomial reduces to a Lagrange interpolation polynomial, we can apply the classical linear stability theory to our approach.
6. Finally, it would be interesting to apply the constraint satisfaction approach to boundary value problems, where pruning arises naturally.

In summary, the constraint satisfaction approach should be a valuable addition to existing methods for the reliable solutions of differential equations, and there is considerable room for further research in this area.

Acknowledgments. We give special thanks to Philippe Delsarte for interesting discussions and for his detailed comments. We also would like to express our gratitude to the reviewers for their detailed suggestions.

REFERENCES

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & Sons, New York, 1988.
[2] C. Bendsten and O. Stauning, *TADIFF, a Flexible C++ Package for Automatic Differentiation Using Taylor Series*, Technical report 1997-x5-94, Technical University of Denmark, Kgs. Lyngby, Denmark, 1997.
[3] C. Bendsten and O. Stauning, *FADBAD, a Flexible C++ Package for Automatic Differentiation Using the Forward and Backward Methods*, Technical report 1996-x5-94, Technical University of Denmark, Kgs. Lyngby, Denmark, 1996.
[4] M. Berz and K. Makino, *Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models*, Reliab. Comput., 4 (1998), pp. 361–369.
[5] G. F. Corliss and R. Rihm, *Validating an a priori enclosure using high-order Taylor series*, in Scientific Computing, Computer Arithmetic, and Validated Numerics, Akademie Verlag, Berlin, 1996, pp. 228–238.
[6] J. Cruz and P. Barahona, *An interval constraint approach to handle parametric ordinary differential equations for decision support*, in Proceedings of EKBD-99, 1999, pp. 93–108.
[7] Y. Deville, M. Janssen, and P. Van Hentenryck, *Consistency techniques in ordinary differential equations*, in Proceedings of the Fourth International Conference on Principles and Practice of Constraint Programming, Pisa, Italy, 1998.
[8] P. Eijgenraam, *The Solution of Initial Value Problems Using Interval Arithmetic*, Math. Centre Tracts 144, Stichting Mathematisch Centrum, Amsterdam, The Netherlands, 1981.
[9] W. H. Enright, T. E. Hull, and B. Lindberg, *Comparing numerical methods for stiff systems of ODEs*, BIT, 15 (1975), pp. 10–48.
[10] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.
[11] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.
[12] M. Janssen, *A Constraint Satisfaction Approach for Enclosing Solutions to Parametric Ordinary Differential Equations*, Ph.D. thesis, Department of Computer Science, UCL, Louvain, Belgium, 2001; also available online at http://www.info.ucl.ac.be.
[13] M. Janssen, Y. Deville, and P. Van Hentenryck, *Multistep filtering operators for ordinary differential equations*, in Proceedings of the Fifth International Conference on Principles and Practice of Constraint Programming, Alexandria, VA, 1999.

[14] M. Janssen, P. Van Hentenryck, and Y. Deville, *A constraint satisfaction approach to parametric differential equations*, in Proceedings of the Joint International Conference on Artificial Intelligence, Seattle, WA, 2001.
[15] O. Knüppel, *PROFIL/BIAS — a fast interval library*, Computing, 53 (1994), pp. 277–287.
[16] F. Krueckeberg, *Ordinary differential equations*, in Topics in Interval Analysis, E. Hansen, ed., Clarendon Press, Oxford, UK 1969, pp. 91–97.
[17] W. Kühn, *Rigorously computed orbits of dynamical systems without the wrapping effect*, Computing, 61 (1998), pp. 47–67.
[18] W. Kühn, *Zonotope dynamics in numerical quality control*, in Mathematical Visualization. Algorithms, Applications, and Numerics, H.-Ch. Hege and K. Polthier, eds., Springer-Verlag, Berlin, 1998, pp. 125–134.
[19] W. Kühn, *Towards an optimal control of the wrapping effect*, in Developments in Reliable Computing, T. Csendes, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands 1999, pp. 43–51.
[20] R. J. Lohner, *Enclosing the solutions of ordinary initial and boundary value problems*, in Computer Arithmetic: Scientific Computation and Programming Languages, Teubner, Stuttgart, 1987.
[21] R. E. Moore, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[22] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM Stud. Appl. Math. 2, SIAM, Philadelphia, 1979.
[23] N. S. Nedialkov, *Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential Equation*, Ph.D. thesis, Computer Science Department, University of Toronto, Toronto, ON, Canada, 1999.
[24] N. S. Nedialkov and K. R. Jackson, *An interval Hermite-Obreschkoff method for computing rigorous bounds on the solution of an initial value problem for an ODE*, in Developments in Reliable Computing, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 289–310.
[25] N. S. Nedialkov, K. R. Jackson, and G. F. Corliss, *Validated solutions of initial value problems for ordinary differential equations*, Appl. Math. Comput., 105 (1999), pp. 21–68.
[26] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 2000.
[27] L. B. Rall, *Automatic Differentiation: Techniques and Applications*, Lecture Notes in Comput. Sci. 120, Springer-Verlag, Berlin, 1981.
[28] R. Rihm, *Implicit methods for enclosing solutions of ODEs*, J. UCS, 4 (1998).
[29] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.
[30] P. Van Hentenryck, D. McAllester, and D. Kapur, *Solving polynomial systems using a branch and prune approach*, SIAM J. Numer. Anal., 34 (1997), pp. 797–827.
[31] P. Van Hentenryck, L. Michel, and Y. Deville, *Numerica: A Modeling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.
[32] P. Van Hentenryck, *A gentle introduction to Numerica*, Artificial Intelligence, 103 (1998), pp. 209–235.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1940–1967

DIFFERENCE APPROXIMATIONS FOR THE SECOND ORDER WAVE EQUATION*

HEINZ-OTTO KREISS†, N. ANDERS PETERSSON‡, AND JACOB YSTRÖM§

Abstract. Difference approximations are derived for the second order wave equation in one and two space dimensions, without first writing it as a first order system. Both the Dirichlet and the Neumann problems are treated for the one-dimensional case. Relations between the boundary error and the interior phase error are derived for a fully second order accurate discretization as well as a scheme that is fourth order accurate in the interior and second order accurate at the boundary. General two-dimensional domains are considered for the Dirichlet problem where the domain is embedded in a Cartesian grid and the boundary conditions are approximated by interpolation. A stable conservative scheme is derived where the time step is determined only by the interior discretization formula. Discretization cells cut by the boundary are treated implicitly, but the resulting scheme becomes explicit because the implicit dependence only is pointwise. Numerical examples are provided to verify the stability and accuracy of the proposed method.

Key words. wave equation, stability, accuracy, embedded boundary

AMS subject classifications. 65M06, 65M12

PII. S0036142901397435

1. Introduction.

The theory of difference approximations for first order strongly hyperbolic systems is by now very well developed. However, in many applications like seismology, acoustics, and general relativity the underlying differential equations are systems of second order hyperbolic partial differential equations. It is surprising that in this case the corresponding theory is much less developed. Instead, one often rewrites the equations as a first order system and then uses methods developed for such systems. While these methods provide the most natural way to solve problems that come as first order systems, we will argue that there can be drawbacks with rewriting second order systems into this form before they are discretized. Instead, we propose a numerical method that directly discretizes the second order system.

Consider, for example, the wave equation

$$u_{tt} = u_{xx} \tag{1.1}$$

in the strip $0 \le x \le 1,\ t \ge 0$. Thus we have to give initial conditions

$$u(x,0) = f(x), \quad u_t(x,0) = g(x), \tag{1.2}$$

and boundary conditions, for example,

$$u(0,t) = h_0(t), \quad u(1,t) = h_1(t). \tag{1.3}$$

*Received by the editors November 2, 2001; accepted for publication (in revised form) April 30, 2002; published electronically December 3, 2002. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

http://www.siam.org/journals/sinum/40-5/39743.html

†Department of Mathematics, University of California, Los Angeles, CA 90024 (kreiss@math.ucla.edu).

‡Center for Applied Scientific Computing, Lawrence Livermore National Lab, Livermore, CA 94551 (andersp@llnl.gov).

§Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden (yxan@nada.kth.se).

To solve the problem numerically, we introduce a grid by

$$t_n = nk, \quad k > 0, \quad n = 0, 1, 2, \ldots, \quad x_\nu = \nu h, \quad h = 1/N, \quad \nu = 0, 1, 2, \ldots, N\,,$$

and approximate (1.1)–(1.3) by a completely centered approximation

$$\begin{aligned} D_+^t D_-^t v(x_\nu, t_n) &= D_+^x D_-^x v(x_\nu, t_n), \quad \nu = 1, 2, \ldots, N-1\,, \\ v(x_\nu, 0) &= f(x_\nu), \quad v(x_\nu, k) = f(x_\nu) + kg(x_\nu) + \frac{k^2}{2} D_+^x D_-^x f(x_\nu), \\ v(0, t_n) &= h_0(t_n), \quad v(1, t_n) = h_1(t_n). \end{aligned} \tag{1.4}$$

Here

$$\begin{aligned} hD_+^x v(x_\nu, t) &= v(x_{\nu+1}, t) - v(x_\nu, t), \\ hD_-^x v(x_\nu, t) &= v(x_\nu, t) - v(x_{\nu-1}, t), \\ D_0^x &= \frac{1}{2}(D_+^x + D_-^x) \end{aligned}$$

denote the usual forward, backward, and centered difference operators. As we will see, this approximation and its generalization to more space dimensions work very well. There are no difficulties with the boundary conditions; i.e., we do not need to supply any extrapolation conditions.

One can write (1.1) as a first order system

$$\mathbf{u}_t = A\mathbf{u}_x, \quad \mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{1.5}$$

with initial conditions

$$\mathbf{u}(x, 0) = \mathbf{f}, \quad \mathbf{f} = \left(f, \int_0^x g(\tilde{x}) d\tilde{x} \right)^T,$$

and boundary conditions (1.3). The leap-frog scheme is often used to solve wave propagation problems. For (1.5) it is, in its simplest form, given by

$$\begin{aligned} D_0^t \mathbf{u}(x_\nu, t_n) &= AD_0^x \mathbf{u}(x_\nu, t_n), \quad \nu = 1, 2, \ldots, N-1\,, \\ \mathbf{u}(x_\nu, 0) &= \mathbf{f}(x_\nu), \quad \mathbf{u}(x_\nu, k) = \mathbf{f}(x_\nu) + kAD_0^x \mathbf{u}(x_\nu, 0). \end{aligned} \tag{1.6}$$

There are a number of drawbacks with this procedure:

1. One needs to calculate two variables. (More variables are needed in several space dimensions.)
2. To obtain the same accuracy as (1.4), one needs to double the number of grid points in space and time.
3. Since there are no boundary conditions for v, one has to supply extrapolation conditions to obtain $v(0, t)$ and $v(1, t)$. This can be done, but one has to be careful not to introduce instabilities; see [5].
4. If the solution is not properly resolved, i.e., if one does not use enough points/wavelength, then one creates spurious waves which travel in the wrong direction; see [1].

To avoid the three latter difficulties, one may introduce a so-called staggered grid. However, this amounts to nothing else but solving (1.4) in disguise. In more space dimensions, staggered grids can lead to complications at the boundaries.

In the present paper, we directly approach the wave equation as a second order system. The equations are discretized on a Cartesian grid that covers the domain of interest, and the spatial derivatives are approximated by finite differences. On the boundary, which is embedded in the Cartesian grid, we use interpolation to approximate the boundary condition. This procedure results in a closed second order system of ordinary differential equations, and we derive an appropriate time-integration method for this system.

Numerical methods for first order systems are by now well developed, and many useful techniques have been established, such as higher order accurate boundary conditions [9], accurate treatment of discontinuous coefficients [4], and nonreflecting boundary conditions for external domains. The method presented here currently lacks these refinements, so a direct comparison on a realistic problem is hard to make. Instead, this work should be seen as a starting point for the development of a numerical technique that directly approaches second order hyperbolic systems.

Embedded boundary techniques for discretizing Laplace's equation subject to Dirichlet boundary conditions date back to Weller and Shortley [10], who used a finite-volume method (perhaps before that term was coined) to set up first order accurate difference approximations near the boundary. Collatz [2] derived higher order difference methods for both the Neumann and Dirichlet problems. More recently, several embedded boundary methods have been presented for various types of partial differential equations. For example, Pember et al. [8] used a Cartesian grid method for solving the time-dependent equations of gas dynamics. Zhang and LeVeque [11] solved the acoustic wave equation with discontinuous coefficients written as a first order system. They derived special difference stencils that satisfy the jump conditions at the interior interfaces, where the coefficients are discontinuous. A staggered grid method was used by Ditkowski, Dridi, and Hesthaven [3] for solving Maxwell's equations on a Cartesian grid. The methods described in these papers all solve first order systems (in time). Johansen and Colella [6] derived a finite-volume scheme for solving Poisson's equation with Dirichlet boundary conditions using an embedded boundary technique. Away from the boundary, the truncation error for the Laplace operator is $O(h^2)$, but in cells cut by the boundary it becomes $O(h/\Gamma)$. (Here h is the mesh spacing and Γ is the area fraction of the cut cell.) A potential theoretic argument is used to show that the solution of Poisson's equation is still second order accurate, even as $\Gamma \to 0$. However, the large truncation error in cells cut by the boundary makes this method unsuitable for solving the wave equation, where the truncation error for the Laplace operator also needs to be small near the boundary.

We shall now summarize the remainder of the paper. In section 2 we discuss a second order accurate time-integration method to solve the system of ordinary differential equations that arises after the wave equation is discretized in space. In particular, we derive a semi-implicit approach to avoid the severe time-step restriction that otherwise can occur from small cells cut by the boundary. When certain symmetry conditions are satisfied, the time-integration method is shown to be stable and conservative.

Section 3 contains a discussion of the stability and accuracy of semidiscrete approximations in one space dimension both for Dirichlet and Neumann boundary conditions. We consider methods that are second order accurate overall and methods

that are fourth order accurate in the interior and second order accurate at the boundary. Analytically, we derive the relation between the boundary error and the interior phase error. The analysis clearly shows that for the second order accurate method, the phase error dominates if we integrate over long distances. By using the fourth order method in the interior, we show that the phase error is greatly reduced.

The Dirichlet problem for the wave equation in general two-dimensional domains is treated in section 4. We show that we can construct stable energy conserving schemes that are either second order accurate overall or second order accurate at the boundary and fourth order accurate in the interior. We show that the scheme can be derived "dimension by dimension", essentially by employing the one-dimensional scheme in each direction. Since the semi-implicit treatment of the cut cells at the boundary is pointwise, the resulting scheme is fully explicit and therefore highly effective.

Numerical examples are provided in section 5, where we solve the two-dimensional Dirichlet problem to demonstrate the accuracy and stability of the proposed method. Future research is outlined in section 6.

2. Ordinary differential equations. Consider the initial value problem for the scalar equation

$$u_{tt} = \lambda u + F(t) \tag{2.1}$$

with initial conditions

$$u(0) = u_0, \quad u_t(0) = u_1. \tag{2.2}$$

Here $F(t)$ is a smooth function and $\lambda < 0$ is a negative constant.

The usual way to solve (2.1), (2.2) numerically is to rewrite the equation as a first order system and then apply any of the standard schemes. In this paper we solve the equation directly. Let k be the time step, $t_n = nk$, and denote the discrete approximation $v^n \approx u(t_n)$. We use two different second order accurate schemes:

1. If $\lambda \sim -1$, we use

$$v^{n+1} - 2v^n + v^{n-1} = k^2(\lambda v^n + F(t_n)). \tag{2.3}$$

2. If $\lambda \ll -1$, then we use instead

$$v^{n+1} - 2v^n + v^{n-1} = k^2 \left(\frac{\lambda}{2}(v^{n+1} + v^{n-1}) + F(t_n) \right).$$

The last method can also be written as

$$\left(1 - \frac{\lambda k^2}{2}\right)(v^{n+1} - 2v^n + v^{n-1}) = \lambda k^2 v^n + k^2 F(t_n). \tag{2.4}$$

We initialize the schemes by

$$v_0 = u_0, \quad v_1 = u(0) + k u_t(0) + \frac{k^2}{2} u_{tt}(0) = u_0 + k u_1 + \frac{k^2}{2}(\lambda u_0 + F(0)).$$

The characteristic equations for (2.3) and (2.4) are given by

$$(\kappa - 1)^2 - k^2 \lambda \kappa = 0, \quad \left(1 - \frac{\lambda k^2}{2}\right)(\kappa - 1)^2 - \lambda k^2 \kappa = 0,$$

respectively. Thus,

$$\kappa = 1 + \frac{1}{2}\zeta \pm \sqrt{\zeta + \frac{1}{4}\zeta^2},$$

where

$$\zeta = \lambda k^2 \quad \text{for (2.3)} \quad \text{and} \quad \zeta = \frac{\lambda k^2}{1 - \lambda k^2/2} \quad \text{for (2.4)}.$$

Thus, $|\kappa_1| = |\kappa_2| = 1,\ \kappa_1 \neq \kappa_2$, for $-4 < \zeta < 0$. Hence, the approximation (2.3) is stable for $k < 2/\sqrt{-\lambda}$, while the scheme (2.4) is unconditionally stable.

Now we consider systems

$$\begin{aligned} \boldsymbol{u}_{tt} &= A\boldsymbol{u} + F(t), \\ \boldsymbol{u}(0) = \boldsymbol{u}_0, &\quad \boldsymbol{u}_t(0) = \boldsymbol{u}_1, \end{aligned} \tag{2.5}$$

and the corresponding homogeneous problem

$$\begin{aligned} \boldsymbol{v}_{tt} &= A\boldsymbol{v}, \\ \boldsymbol{v}(0) = \boldsymbol{v}_0, &\quad \boldsymbol{v}_t(0) = \boldsymbol{v}_1. \end{aligned} \tag{2.6}$$

LEMMA 2.1. *The solutions of* (2.6) *are uniformly bounded in time if and only if the eigenvalues of A are real and negative and there is a complete system of eigenvectors.*

Proof. Let λ be an eigenvalue of A and φ_0 the corresponding eigenvector. Then, for any constants σ_1, σ_2,

$$\boldsymbol{v}_\lambda = \begin{cases} \left(\sigma_1 e^{\sqrt{\lambda}t} + \sigma_2 e^{-\sqrt{\lambda}t}\right)\varphi_0 & \text{if } \lambda \neq 0, \\ (\sigma_1 + \sigma_2 t)\varphi_0 & \text{if } \lambda = 0 \end{cases}$$

is a solution to $\boldsymbol{v}_{tt} = A\boldsymbol{v}$. Thus, $\boldsymbol{v}_\lambda$ is uniformly bounded in time if and only if λ is real and negative. If there is a complete eigensystem, then we can write the solutions to (2.6) as a sum of eigensolutions. The solutions are therefore uniformly bounded if and only if all the eigenvalues are real and negative.

An easy calculation shows that the solutions of

$$\boldsymbol{v}_{tt} = J\boldsymbol{v},$$

where J is a Jordan block, are not uniformly bounded. Therefore, if the eigensystem is incomplete, the solutions of (2.6) are not uniformly bounded. This proves the lemma.

If $A = A^* < 0$ is a negative definite symmetric matrix, then all conditions of the above lemma are satisfied and the solutions are uniformly bounded. We can also prove this by an energy estimate. We have

$$\begin{aligned} \frac{\partial}{\partial t}|\boldsymbol{v}_t|^2 &= \langle \boldsymbol{v}_t, \boldsymbol{v}_{tt}\rangle + \langle \boldsymbol{v}_t, \boldsymbol{v}_{tt}\rangle \\ &= \langle \boldsymbol{v}_t, A\boldsymbol{v}\rangle + \langle A\boldsymbol{v}, \boldsymbol{v}_t\rangle = \frac{\partial}{\partial t}\langle \boldsymbol{v}, A\boldsymbol{v}\rangle, \end{aligned}$$

i.e.,

$$\frac{\partial}{\partial t}\left(|\boldsymbol{v}_t|^2 + \langle \boldsymbol{v}, (-A)\boldsymbol{v}\rangle\right) = 0. \tag{2.7}$$

Since $-A$ is positive definite, boundedness follows.

We approximate the system (2.6) by

$$t_n = nk, \quad \boldsymbol{v}^n \approx \boldsymbol{v}(t_n),$$

and

$$\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1} = k^2 A\boldsymbol{v}^n \quad \text{if } |A| \sim 1, \tag{2.8}$$

$$\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1} = \frac{k^2}{2} A\left(\boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1}\right) \quad \text{if } |A| \gg 1. \tag{2.9}$$

We can write (2.9) in the form

$$\left(I - \frac{k^2}{2}A\right)\left(\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1}\right) = k^2 A\boldsymbol{v}^n.$$

Since we can reduce the system to scalar equations, the difference approximation corresponding to (2.3) is stable if

$$\max_j |\lambda_j| k^2 < 4.$$

The approximation (2.9) is unconditionally stable.

We proceed by using energy methods to derive the discrete counterpart of (2.7) to show that the discrete energy is conserved by the scheme (2.8). We write (2.8) in the form

$$\boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1} = (2I + k^2 A)\boldsymbol{v}^n.$$

Therefore,

$$\langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^{n-1}, \boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1}\rangle = \langle \boldsymbol{v}^{n+1}, (2I + k^2 A)\boldsymbol{v}^n\rangle - \langle \boldsymbol{v}^{n-1}, (2I + k^2 A)\boldsymbol{v}^n\rangle.$$

Assuming that $\boldsymbol{v}^n, A$ are real, we obtain

$$\begin{aligned} L(t_{n+1}, A) &=: |\boldsymbol{v}^{n+1}|^2 + |\boldsymbol{v}^n|^2 - \langle \boldsymbol{v}^{n+1}, (2I + k^2 A)\boldsymbol{v}^n\rangle \\ &= |\boldsymbol{v}^n|^2 + |\boldsymbol{v}^{n-1}|^2 - \langle \boldsymbol{v}^n, (2I + k^2 A)\boldsymbol{v}^{n-1}\rangle \\ &= L(t_n, A). \end{aligned}$$

Thus, we obtain an energy estimate if L is positive definite. We have

$$L(t_{n+1}, A) = \langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^n, \boldsymbol{v}^{n+1} - \boldsymbol{v}^n\rangle - k^2\langle \boldsymbol{v}^{n+1}, A\boldsymbol{v}^n\rangle.$$

Since $A = A^*$ is symmetric,

$$\langle \boldsymbol{v}^{n+1}, A\boldsymbol{v}^n\rangle = \frac{1}{4}\langle \boldsymbol{v}^{n+1} + \boldsymbol{v}^n, A(\boldsymbol{v}^{n+1} + \boldsymbol{v}^n)\rangle - \frac{1}{4}\langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^n, A(\boldsymbol{v}^{n+1} - \boldsymbol{v}^n)\rangle.$$

Hence,

$$\begin{aligned} L(t_{n+1}, A) = &\left\langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^n, \left(I + \frac{k^2}{4}A\right)(\boldsymbol{v}^{n+1} - \boldsymbol{v}^n)\right\rangle \\ &- \frac{k^2}{4}\langle \boldsymbol{v}^{n+1} + \boldsymbol{v}^n, A(\boldsymbol{v}^{n+1} + \boldsymbol{v}^n)\rangle. \end{aligned} \tag{2.10}$$

Let $\lambda_j < 0$ be the eigenvalues of A. For $4 - k^2 \max_j |\lambda_j| \geq k^2 \min_j |\lambda_j|$,

$$\left\langle \boldsymbol{v}, \left(I + \frac{k^2}{4} A \right) \boldsymbol{v} \right\rangle \geq \frac{k^2}{4} \min_j |\lambda_j| |\boldsymbol{v}|^2$$

and

$$L(t_{n+1}, A) \geq \frac{k^2}{4} \min_j |\lambda_j| \left(|\boldsymbol{v}^{n+1} - \boldsymbol{v}^n|^2 + |\boldsymbol{v}^{n+1} + \boldsymbol{v}^n|^2 \right).$$

Thus, $L(t_n, A)$ is positive definite if the time step satisfies

$$(\max_j |\lambda_j| + \min_j |\lambda_j|) k^2 \leq 4$$

and, essentially, we recover the previous time-step restriction.

By comparing (2.11) and (2.7) we note that $L(t_{n+1}, A)/k^2$ is a second order accurate approximation of the energy $|\boldsymbol{v}_t|^2 - \langle \boldsymbol{v}, A\boldsymbol{v} \rangle$, evaluated at time $t_n + k/2$.

Often there are only relatively few elements of A which are large, and we can write

$$A = A_1 + A_2, \quad A_1 = A_1^* \leq 0, \quad |A_1| \gg 1, \quad A_2 = A_2^* < 0, \quad |A_2| \sim 1.$$

To avoid severe restrictions of the step size, we can use the second order approximation

$$\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1} = \frac{k^2}{2} A_1 \left(\boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1} \right) + k^2 A_2 \boldsymbol{v}^n, \tag{2.11}$$

which we write as

$$\left(I - \frac{k^2}{2} A_1 \right) (\boldsymbol{v}^{n+1} - 2\boldsymbol{v}^n + \boldsymbol{v}^{n-1}) = k^2 (A_1 + A_2) \boldsymbol{v}^n$$

or

$$\left(I - \frac{k^2}{2} A_1 \right) (\boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1}) = \left(2I + k^2 A_2 \right) \boldsymbol{v}^n.$$

Thus,

$$\left\langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^{n-1}, \left(I - \frac{k^2}{2} A_1 \right) (\boldsymbol{v}^{n+1} + \boldsymbol{v}^{n-1}) \right\rangle = \langle \boldsymbol{v}^{n+1} - \boldsymbol{v}^{n-1}, \left(2I + k^2 A_2 \right) \boldsymbol{v}^n \rangle.$$

Similar to the scheme (2.8), we can derive a discrete energy that is conserved. Assuming that $\boldsymbol{v}^n$, A_1, A_2 are real, we have

$$\begin{aligned} L_1(t_{n+1}, A_1, A_2) =: & \left\langle \boldsymbol{v}^{n+1}, \left(I - \frac{k^2}{2} A_1 \right) \boldsymbol{v}^{n+1} \right\rangle + \left\langle \boldsymbol{v}^n, \left(I - \frac{k^2}{2} A_1 \right) \boldsymbol{v}^n \right\rangle \\ & - \langle \boldsymbol{v}^{n+1}, (2I + k^2 A_2) \boldsymbol{v}^n \rangle \\ = & \left\langle \boldsymbol{v}^n, \left(I - \frac{k^2}{2} A_1 \right) \boldsymbol{v}^n \right\rangle + \left\langle \boldsymbol{v}^{n-1}, \left(I - \frac{k^2}{2} A_1 \right) \boldsymbol{v}^{n-1} \right\rangle \\ & - \langle \boldsymbol{v}^n, (2I + k^2 A_2) \boldsymbol{v}^{n-1} \rangle \\ = & \; L_1(t_n, A_1, A_2). \end{aligned}$$

We have now to show that $L_1(t_n, A_1, A_2)$ is positive definite. Since $-A_1$ is positive semidefinite,

$$\left\langle \boldsymbol{v}^{n+1}, -\frac{k^2}{2} A_1 \boldsymbol{v}^{n+1} \right\rangle + \left\langle \boldsymbol{v}^n, -\frac{k^2}{2} A_1 \boldsymbol{v}^n \right\rangle \geq 0,$$

and it follows that

$$L_1(t_n, A_1, A_2) \geq L(t_n, A_2).$$

Thus $L_1(t_n, A_1, A_2)$ is positive definite since $L(t_n, A_2)$ is positive definite. Hence, the previous time-step restriction applies with A replaced by A_2. We summarize our results in the following lemma.

LEMMA 2.2. *The time-integration scheme* (2.11) *is stable and the discrete energy* $L_1(t_n, A_1, A_2)$ *is conserved if* $A_1 = A_1^* \leq 0$, $A_2 = A_2^* < 0$, *and the time step satisfies*

$$\left(\max_j |\lambda_j| + \min_j |\lambda_j| \right) k^2 < 4,$$

where λ_j *are the eigenvalues of* A_2.

All our results are also valid for systems

$$B\boldsymbol{u}_{tt} = A\boldsymbol{u}, \quad A = A^* < 0, \quad B = B^* > 0,$$

because the change of variables $B^{1/2}\boldsymbol{u} = \tilde{\boldsymbol{u}}$ gives us

$$\tilde{\boldsymbol{u}}_{tt} = \tilde{A}\tilde{\boldsymbol{u}}, \quad \tilde{A} = \tilde{A}^* = B^{-1/2} A B^{-1/2}.$$

3. The wave equation in one space dimension. We consider the wave equation

$$u_{tt} = u_{xx} \tag{3.1}$$

for $x \geq l,\ t \geq 0$. Here $l \geq 0$ is a small number. At $t = 0$ we give initial conditions

$$u(x, 0) = f(x), \quad u_t(x, 0) = g(x). \tag{3.2}$$

Here f is a smooth function with compact support.

3.1. Second order methods. At $x = l$ we give boundary conditions and we start by analyzing Dirichlet conditions

$$u(l, t) = 0. \tag{3.3}$$

Here we discuss second order difference approximations. We only discretize space. In the time direction we use the approximation discussed in the previous section. Let $h > 0$ be a step size. We assume that $l = \alpha h,\ 0 \leq \alpha < 1$. Grid points are given by $x_\nu = \nu h$ and grid functions by $w(x_\nu, t) = w_\nu(t)$. We approximate (3.1), (3.2) by

$$\begin{aligned} w_{\nu tt} &= D_+ D_- w_\nu, \\ w_\nu(0) &= f_\nu, \quad w_{\nu t}(0) = g_\nu, \quad \nu = 1, 2, \ldots. \end{aligned} \tag{3.4}$$

We shall use the simplest second order accurate boundary condition given by the interpolation condition

$$\alpha w_1 + (1 - \alpha) w_0 = 0. \tag{3.5}$$

We can express w_0 in terms of w_1 and eliminate w_0 from (3.4). Then the differential equation for w_1 becomes

$$w_{1tt} = \frac{w_2 - 2w_1 + w_0}{h^2} = \frac{1}{h^2}(aw_1 + bw_2),$$

where

$$a = -\left(2 + \frac{\alpha}{1-\alpha}\right), \quad b = 1.$$

In matrix form (3.4) can be written as

$$\mathbf{w}_{tt} = \frac{1}{h^2} \begin{pmatrix} a & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & \\ 0 & 1 & -2 & 1 & & \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & & & & & \end{pmatrix} \mathbf{w}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ \vdots \end{pmatrix}.$$

The matrix is symmetric and negative definite. Since $|a|$ becomes large as $\alpha \to 1$, we split the matrix and use the scheme (2.11) with

$$A_1 = \frac{1}{h^2} \begin{pmatrix} -\frac{\alpha}{1-\alpha} & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

We now consider Neumann boundary conditions. At $x = \alpha h$ we have

$$u_x(\alpha h) = u_x(0) + \alpha h u_{xx}(0) + \mathcal{O}(h^2).$$

Also,

$$D_+ u_0 = u_x(0) + \frac{h}{2} u_{xx}(0) + \mathcal{O}(h^2).$$

Therefore,

$$u_x(\alpha h) = D_+ u_0 + h\left(\alpha - \frac{1}{2}\right) D_+^2 u_0 + \mathcal{O}(h^2).$$

Thus, we approximate the boundary condition

$$u_x(\alpha h) = 0$$

by

$$D_+ w_0 + h\left(\alpha - \frac{1}{2}\right) D_+^2 w_0 = 0, \tag{3.6}$$

i.e.,

$$\left(\frac{3}{2}-\alpha\right)w_0=\left(\alpha-\frac{1}{2}\right)w_2+(2-2\alpha)w_1.$$

We eliminate w_0 from the differential equations (3.4) and obtain

$$\begin{aligned}w_{1tt}&=\frac{1}{h^2}(w_2-2w_1+w_0)=\frac{1}{h^2}\left(\left(1+\frac{\alpha-\frac{1}{2}}{\frac{3}{2}-\alpha}\right)w_2-\left(2-\frac{2-2\alpha}{\frac{3}{2}-\alpha}\right)w_1\right)\\&=\frac{1}{ah^2}(w_2-w_1),\quad a=\frac{3}{2}-\alpha.\end{aligned}$$

Thus, (3.4) can be written as

(3.7)

$$B\mathbf{w}_{tt}=:\begin{pmatrix}a&&&0\\&1&&\\&&\ddots&\\0&&&1\end{pmatrix}\mathbf{w}_{tt}=\frac{1}{h^2}\begin{pmatrix}-1&1&0&0&\cdots&0\\1&-2&1&0&\cdots&\\0&1&-2&1&&\\\vdots&&\ddots&\ddots&\ddots&\\0&&&&&\end{pmatrix}\mathbf{w}:=A\mathbf{w}.$$

The matrix A is symmetric and negative definite on the space of grid functions with bounded discrete l_2-norm. In this way we exclude solutions which are constant in x.

Note that $1/2\le a\le 3/2$, so the system does not become stiff for $0\le\alpha\le 1$. We can therefore apply the scheme (2.8) to integrate in time.

3.2. Higher order methods. It is well known that for the Cauchy problem fourth order methods are much more effective than second order methods when solving wave propagation problems. The number of points/wavelength analysis tells us that the phase error is very much decreased. However, for problems in bounded domains, it is often difficult to construct stable fourth order accurate approximations of the boundary conditions. We want to show that a method that is fourth order accurate in the interior but only second order accurate at the boundary is an acceptable compromise. In this way we control the phase error.

We consider the half-plane problem for the wave equation

$$u_{tt}=u_{xx}+F(x,t),\quad x\ge\alpha h,\ t\ge 0, \tag{3.8}$$

$$u(x,0)=f^{(1)}(x),\quad u_t(x,0)=f^{(2)}(x), \tag{3.9}$$

with Dirichlet boundary conditions

$$u(\alpha h,t)=g(t). \tag{3.10}$$

We approximate (3.8)–(3.10) by

$$v_{\nu tt}=D_+D_-v_\nu,\quad v_\nu=v(x_\nu,t),\ x_\nu=\nu h,\ \nu=0,1,2,\dots, \tag{3.11}$$

$$v_\nu(0)=f_\nu^{(1)},\quad v_{\nu t}(0)=f_\nu^{(2)}, \tag{3.12}$$

with boundary condition

$$\alpha v_0+(1-\alpha)v_1=g(t). \tag{3.13}$$

Here $F, f^{(j)}, g \in C_0^\infty$. Without restriction we can assume that $F \equiv 0, f^{(j)} \equiv 0$, because we can extend $F, f^{(j)}$ to the whole space, solve the Cauchy problem, and subtract its solution from u. For the Cauchy problem we know that the fourth order method is much more accurate.

Under the above assumption, we solve the above problems by Laplace transform. The transformed problems are

$$s^2\hat{u} = \hat{u}_{xx}, \quad \hat{u}(\alpha h, s) = \hat{g}, \quad s = i\xi + \eta, \quad \eta > 0, \tag{3.14}$$

$$s^2\hat{v}_\nu = D_+D_-\hat{v}_\nu, \quad \alpha v_1 + (1-\alpha)v_0 = \hat{g}, \quad \nu = 1, 2, \ldots, \tag{3.15}$$

and their solutions are given by

$$\hat{u}(x,t) = e^{-sx}e^{\alpha hs}\hat{g}, \tag{3.16}$$

$$\hat{v}_\nu = \kappa^\nu \hat{v}_0, \quad \big(\alpha\kappa + (1-\alpha)\big)\hat{v}_0 = \hat{g}, \tag{3.17}$$

respectively. Here κ with $|\kappa| < 1$ is the solution of the characteristic equation

$$\frac{(\kappa-1)^2}{\kappa} = s^2h^2.$$

For the discussion of accuracy, we can assume that $|sh| \le \delta \ll 1$. We obtain

$$\kappa = 1 + \frac{s^2h^2}{2} - \sqrt{s^2h^2 + \frac{s^4h^4}{4}} \sim 1 - sh + \frac{s^2h^2}{2} - \frac{s^3h^3}{8} \sim e^{-sh(1-\frac{s^2h^2}{24})}.$$

By (3.15)

$$v_0 = \frac{\hat{g}}{\alpha e^{-sh} + 1 - \alpha} \sim \hat{g}e^{\alpha hs}\left(1 - \frac{\alpha - \alpha^2}{2}s^2h^2\right).$$

Thus, for $x = x_\nu$,

$$\hat{v}(x,s) \sim e^{-sx(1-\frac{s^2h^2}{24})}e^{\alpha hs}\hat{g}\left(1 - \frac{\alpha-\alpha^2}{2}s^2h^2\right)$$

and

$$|\hat{u}(x,s) - \hat{v}(x,s)| \le |\hat{g}e^{\alpha hs}|\left\{\left|\frac{\alpha-\alpha^2}{2}s^2h^2\right| + |e^{-sx}|\left(1 - e^{\frac{sxh^2s^2}{24}}\right)\right\} \tag{3.18}$$

$$\sim |\hat{g}|\left(\frac{|sh|^2}{8} + \frac{|sx|}{24}|sh|^2\right) = \frac{|\hat{g}|\,|sh|^2}{8}\left(1 + \frac{|sx|}{3}\right). \tag{3.19}$$

We can invert the Laplace transform on the imaginary axis $s = i\xi$. Therefore, we can consider $u(x,t),\ v(x,t)$ as a superposition of waves which travel into the region. The error consists of the phase error $\frac{|sx|}{24}(sh)^2$ and the boundary error $\frac{|sh|^2}{8}$, due to the interpolation on the boundary. It shows that the phase error dominates the boundary error if $|sx| > 3$.

We now consider the fourth order method

$$v_{\nu tt} = D_+D_-v_\nu - \frac{h^2}{12}D_+^2D_-^2v_\nu, \quad \nu = 1, 2, \ldots, \tag{3.20}$$

with boundary condition

$$\alpha v_1 + (1-\alpha)v_0 = g. \tag{3.21}$$

Now the stencil depends also on v_{-1}. Therefore, we have to supply an extra boundary such that we can eliminate v_{-1}. This condition can have different forms, depending on our requirements.

3.2.1. A method that is fourth order accurate in the interior and on the boundary. Using the differential equation, the boundary condition

$$u(x,0) = g(t)$$

implies

$$u_{xx}(x,0) = u_{tt}(x,0) = g_{tt}.$$

Therefore, we obtain a method that is fourth order accurate overall if we add the condition

$$\alpha D_+D_-v_1 + (1-\alpha)D_+D_-v_0 = g_{tt}. \tag{3.22}$$

Another advantage of (3.22) is that in matrix form we obtain a symmetric system. We write (3.20) and (3.21), (3.22) in the form

$$\begin{aligned} &v_{\nu tt} = D_+D_-v_\nu - \frac{h^2}{12}D_+D_-w_\nu, \quad w_\nu = D_+D_-v_\nu, \quad \nu = 0,1,2,\dots,\\ &\alpha v_1 + (1-\alpha)v_0 = g, \quad \alpha w_1 + (1-\alpha)w_0 = g. \end{aligned}$$

Then we can eliminate v_0 and w_0 and obtain

$$\mathbf{v}_{tt} = A\mathbf{v} - \frac{h^2}{12}A\mathbf{w} + F = A\mathbf{v} - \frac{h^2}{12}A^2\mathbf{v} + F.$$

Since $A - \frac{h^2}{12}A^2$ is negative definite, we can apply our previous results and obtain a stable scheme. Unfortunately, this technique cannot easily be generalized to more space dimensions.

3.2.2. Methods that are fourth order accurate in the interior but only second order accurate on the boundary. As we have seen earlier, the error at the boundary is often much smaller than the phase error in the interior. Therefore, it is reasonable to use a method that is fourth order accurate in the interior and second order accurate at the boundary.

The simplest way to achieve this it to calculate the fourth order term only if its stencil does not depend on boundary or exterior points. The resulting scheme is not symmetric and, unfortunately, it is slightly unstable.

We can also replace (3.20) by

$$v_{\nu tt} = D_+D_-v_\nu - \frac{h^2}{12}D_+D_-(\gamma_\nu D_+D_-v_\nu), \quad \nu = 1,2,\dots, \tag{3.23}$$

with $\gamma_0 = \gamma_1 = 0,\ \gamma_2 = \gamma_3 = \cdots = 1$. Since

$$h^2D_+D_-(\gamma_\nu D_+D_-v_\nu) = \gamma_{\nu+1}D_+D_-v_{\nu+1} - 2\gamma_\nu D_+D_-v_\nu + \gamma_{\nu-1}D_+D_-v_{\nu-1}, \tag{3.24}$$

the matrix for the semidiscrete problem is again symmetric and negative definite and we obtain an energy estimate. We shall now discuss the accuracy of the new method. Since the fourth order terms do not depend on v_{-1} and v_0, we do not need to specify an extra boundary condition. Thus, we consider (3.23) with boundary condition (3.21) and solve the problem by Laplace transform.

Since $\gamma_\nu = 1$ for $\nu \geq 2$, the Laplace transformed problem becomes, using (3.24),

$$s^2\hat{v}_\nu = D_+D_-\hat{v}_\nu - \frac{h^2}{12}D_+^2D_-^2\hat{v}_\nu, \quad \mathrm{Re}(s) > 0, \quad \nu = 3, 4, \ldots, \tag{3.25}$$

$$s^2\hat{v}_2 = D_+D_-\hat{v}_2 - \frac{1}{12}D_+D_-\hat{v}_3 + \frac{1}{6}D_+D_-\hat{v}_2, \tag{3.26}$$

$$s^2\hat{v}_1 = D_+D_-\hat{v}_1 - \frac{1}{12}D_+D_-\hat{v}_2, \tag{3.27}$$

$$\alpha\hat{v}_1 + (1-\alpha)\hat{v}_0 = \hat{g}. \tag{3.28}$$

We can eliminate $\hat{v}_0$ by writing (3.27) and (3.28) in the form

$$-(1-\alpha)\left(h^2s^2\hat{v}_1 - h^2D_+D_-\hat{v}_1 + \frac{h^2}{12}D_+D_-\hat{v}_2\right) = \alpha\hat{v}_1 + (1-\alpha)\hat{v}_0 - \hat{g}. \tag{3.29}$$

We now solve (3.25), (3.26), and (3.29). The general solution of (3.25) is

$$\hat{v}_\nu = \sigma_1\kappa_1^\nu + \sigma_2\kappa_2^\nu, \tag{3.30}$$

where κ with $|\kappa| < 1$ are solutions of

$$s^2h^2 = \mu - \frac{1}{12}\mu^2, \quad \mu = \frac{(\kappa-1)^2}{\kappa}.$$

For small $|sh|$, the solutions of

$$\mu^2 - 12\mu + 12s^2h^2 = 0$$

are

$$\begin{aligned}
\mu_1 &= 6 + \sqrt{36 - 12s^2h^2} \sim 12,\\
\mu_2 &= 6 - \sqrt{36 - 12s^2h^2}\\
&= 6\left(1 - \sqrt{1 - \frac{1}{3}s^2h^2}\right)\\
&\sim 6\left(1 - \left(1 - \frac{1}{6}s^2h^2 - \frac{1}{72}s^4h^4 - \frac{1}{3^3\cdot 16}s^6h^6\right)\right)\\
&\sim s^2h^2 + \frac{1}{12}s^4h^4 + \frac{1}{72}s^6h^6.
\end{aligned}$$

The corresponding κ are solutions of

$$\kappa^2 - (2+\mu)\kappa + 1 = 0.$$

After some painful computations,

$$\kappa_1 \sim \frac{1}{12}, \tag{3.31}$$

$$\kappa_2 = 1 + \frac{\mu_2}{2} - \sqrt{\mu_2 + \frac{\mu_2^2}{4}} \sim e^{-sh(1+\gamma(sh)^4)}, \quad \gamma \sim \frac{1}{120}. \tag{3.32}$$

Inserting the ansatz (3.30) into (3.26),

$$\sigma_1\kappa_1^2\left(s^2h^2-\frac{7}{6}\frac{(\kappa_1-1)^2}{\kappa_1}+\frac{1}{12}(\kappa_1-1)^2\right)$$
$$+\ \sigma_2\kappa_2^2\left(s^2h^2-\frac{7}{6}\frac{(\kappa_2-1)^2}{\kappa_2}+\frac{1}{12}(\kappa_2-1)^2\right)=0,$$

and entering the above values of κ_1 and κ_2 yields

$$\frac{\sigma_1}{12^2}\left(\frac{11^2}{12^2}\left(-14+\frac{1}{12}\right)+s^2h^2\right)-\sigma_2\left(\frac{s^2h^2}{12}+O((sh)^3)\right)=0. \tag{3.33}$$

Thus we can neglect σ_1 in (3.29) and commit an error $O(h^2s^2)$ if we replace (3.29) by (3.28). The result is a solution with essentially the same amplitude as in the second order case but with a much improved phase error.

Similar arguments can be used for Neumann boundary conditions. In one space dimension we can obtain an approximation which is fourth order overall by using

$$u_x=g,\quad u_{xtt}=u_{xxx}=g_{tt},$$

as boundary conditions. That the approximation is second order accurate for Dirichlet conditions at the boundary depends crucially on the relation (3.33), which holds only because the boundary condition (3.28) has no influence on (3.26). Second order accurate discrete Neumann conditions depend on v_0, v_1, and v_2. Therefore, we will obtain (3.33) only if we replace (3.23) by

$$\gamma_0=\gamma_1=\gamma_2=0,\quad \gamma_\nu=1\quad\text{for}\quad \nu\geq 3. \tag{3.34}$$

In one space dimension this new approximation is stable. We will investigate the two-dimensional Neumann problem in a forthcoming paper.

4. The wave equation in two space dimensions with Dirichlet boundary condition. In this section we consider the scalar wave equation in the bounded domain $\Omega\subset\Re^2$, subject to Dirichlet conditions on the boundary Γ,

$$\begin{aligned} u_{tt}&=\Delta u,\quad \boldsymbol{x}\in\Omega,\quad t>0,\\ u(\boldsymbol{x},t)&=f(\boldsymbol{x},t),\quad \boldsymbol{x}\in\Gamma,\quad t>0,\\ u(\boldsymbol{x},0)&=u_0(\boldsymbol{x}),\quad u_t(\boldsymbol{x},0)=u_1(\boldsymbol{x}),\quad \boldsymbol{x}\in\Omega. \end{aligned} \tag{4.1}$$

4.1. Algorithm. We cover Ω by a Cartesian grid with step size h; see Figure 1. The grid points are given by $\boldsymbol{x}_{i,j}=(x_i,y_j)^T$,

$$x_i=x^{(0)}+(i-1)h,\quad i=1,2,\ldots,N,$$
$$y_j=y^{(0)}+(j-1)h,\quad j=1,2,\ldots,M,$$

where $h=(x^{(1)}-x^{(0)})/(N-1)\equiv(y^{(1)}-y^{(0)})/(M-1)$. Let all points $\boldsymbol{x}=(x,y)\in\Omega$ satisfy $x_{\min}\leq x\leq x_{\max}$, $y_{\min}\leq y\leq y_{\max}$. To make sure the grid covers Ω, we require $x^{(0)}\leq x_{\min}-h$, $x^{(1)}\geq x_{\max}+h$ and $y^{(0)}\leq y_{\min}-h$, $y^{(1)}\geq y_{\max}+h$.

Before we can discretize the problem, we need to classify each grid point. We denote the classification by $m_{i,j}$,

$$m_{i,j}=\begin{cases} 0, & \boldsymbol{x}_{i,j}\text{ outside of }\Omega,\\ 1, & \boldsymbol{x}_{i,j}\in\Omega,\boldsymbol{x}_{i\pm1,j}\in\Omega\text{ and }\boldsymbol{x}_{i,j\pm1}\in\Omega,\\ -1 & \text{otherwise.}\end{cases}$$

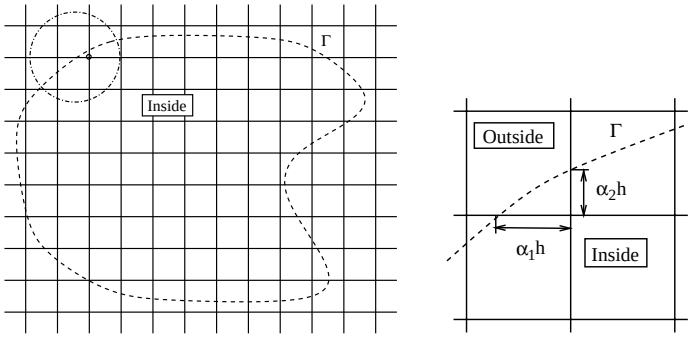


FIG. 1. *The computational grid and the embedded boundary (left) and a close-up of one boundary point (right).*

Hence, interior points have $m_{i,j} = 1$, exterior points have $m_{i,j} = 0$, and boundary points have $m_{i,j} = -1$. Let $u^n_{i,j}$ denote the difference approximation to $u(x_i, y_j, t_n)$. At interior points, we use a centered approximation both in space and time,

$$\frac{u^{n+1}_{i,j} - 2u^n_{i,j} + u^{n-1}_{i,j}}{k^2} = (D^x_+ D^x_- + D^y_+ D^y_-)u^n_{i,j} =: \Delta_h u^n_{i,j}, \quad m_{i,j} = 1. \tag{4.2}$$

At boundary points, the Dirichlet boundary condition is used to eliminate the exterior points from the centered difference formula. For example, let $\boldsymbol{x}_{i,j}$ be a boundary point and let the points $(x_i - \alpha_1 h, y_j)^T$ and $(x_i, y_j + \alpha_2 h)^T$ be on the boundary, as is shown in Figure 1. The Dirichlet conditions are approximated by linear interpolation,

$$\begin{aligned}(1-\alpha_1)u^n_{i,j} + \alpha_1 u^n_{i-1,j} &= f(x_i - \alpha_1 h, y_j, t_n),\\ (1-\alpha_2)u^n_{i,j} + \alpha_2 u^n_{i,j+1} &= f(x_i, y_j + \alpha_2 h, t_n).\end{aligned}$$

Assuming that $\alpha_1 > 0$, $\alpha_2 > 0$,

$$\begin{aligned}u^n_{i-1,j} &= -\frac{1-\alpha_1}{\alpha_1} u^n_{i,j} + \frac{1}{\alpha_1} f(x_i - \alpha_1 h, y_j, t_n),\\ u^n_{i,j+1} &= -\frac{1-\alpha_2}{\alpha_2} u^n_{i,j} + \frac{1}{\alpha_2} f(x_i, y_j + \alpha_2 h, t_n).\end{aligned}$$

Eliminating $u^n_{i-1,j}$ and $u^n_{i,j+1}$ from (4.2) results in

$$\frac{u^{n+1}_{i,j} - 2u^n_{i,j} + u^{n-1}_{i,j}}{k^2} = \frac{1}{h^2}\left(-(4 + d_{i,j})u^n_{i,j} + u^n_{i+1,j} + u^n_{i,j-1}\right) + \frac{\tilde{f}^n_{i,j}}{h^2} \tag{4.3}$$

$$=: \tilde{\Delta}_h u^n_{i,j} + \frac{\tilde{f}^n_{i,j}}{h^2}, \tag{4.4}$$

where $d_{i,j} = (1-\alpha_1)/\alpha_1 + (1-\alpha_2)/\alpha_2 > 0$ and $\tilde{f}^n_{i,j} = f(x_i - \alpha_1 h, y_j, t_n)/\alpha_1 + f(x_i, y_j + \alpha_2 h, t_n)/\alpha_2$. Since α_1 and α_2 can be arbitrarily close to zero, d can be very large, which makes the time-step restriction for (4.4) severe. To avoid this problem, we use the splitting discussed in section 2,

$$\begin{aligned}\frac{u^{n+1}_{i,j} - 2u^n_{i,j} + u^{n-1}_{i,j}}{k^2} &= \frac{1}{h^2}\left(-4u^n_{i,j} + u^n_{i+1,j} + u^n_{i,j-1}\right)\\ &\quad - \frac{d_{i,j}}{2h^2}\left(u^{n+1}_{i,j} + u^{n-1}_{i,j}\right) + \frac{\tilde{f}^n_{i,j}}{h^2}, \quad m_{i,j} = -1.\end{aligned}$$

In the general case, $d_{i,j}$ and $\tilde{f}^n_{i,j}$ get contributions from all exterior nearest neighbors, and the first term on the right-hand side includes all interior or boundary nearest neighbors,

$$\frac{u^{n+1}_{i,j} - 2u^n_{i,j} + u^{n-1}_{i,j}}{k^2} = \frac{1}{h^2}\Big(-4u^n_{i,j} + \sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}| u^n_{\boldsymbol{j}} \Big) - \frac{d_{i,j}}{2h^2}\left(u^{n+1}_{i,j} + u^{n-1}_{i,j}\right) + \frac{\tilde{f}^n_{i,j}}{h^2}, \quad m_{i,j} = -1. \tag{4.5}$$

Here $\boldsymbol{j}$ is a multi-index and the sum extends over all nearest neighbors $\boldsymbol{j} = (i+1,j),(i-1,j),(i,j+1),(i,j-1)$.

4.2. Stability. By letting $\boldsymbol{u}^n$ denote the vector containing the solution $u^n_{i,j}$ at all interior and boundary points (using line ordering, for example), we can write the time-integration scheme (4.2), (4.5) in the form (2.11). The matrix A_1, given by

$$A_1\boldsymbol{u} = \begin{cases} 0, & m_{i,j} = 1, \\ -\frac{d_{i,j}}{h^2}, & m_{i,j} = -1, \end{cases}$$

is negative semidefinite since it is diagonal and $-d_{i,j} < 0$. The matrix A_2 is defined by

$$A_2\boldsymbol{u} = \begin{cases} \frac{1}{h^2}\left(-4u_{i,j} + \sum_{\boldsymbol{j}} u_{\boldsymbol{j}}\right), & m_{i,j} = 1, \\ \frac{1}{h^2}\left(-4u_{i,j} + \sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}| u_{\boldsymbol{j}}\right), & m_{i,j} = -1. \end{cases}$$

Again, the sum extends over all nearest neighbors $\boldsymbol{j} = (i+1,j),(i-1,j),(i,j+1),(i,j-1)$.

To study the symmetry of A_2, we note that if (i,j) is an interior point, the corresponding row in $A_2\boldsymbol{u}$ contains the four off-diagonal terms $u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}$. On the other hand, the row in $A_2\boldsymbol{u}$ corresponding to point $(i+1,j)$ includes the term $u_{i,j}$, whether $(i+1,j)$ is an interior or a boundary point. The same argument applies to the other three off-diagonal terms. When (i,j) is a boundary point, the corresponding row in $A_2\boldsymbol{u}$ contains at most three off-diagonal terms. For example, let one of these terms be $u_{i+1,j}$. Again, the row in $A_2\boldsymbol{u}$ corresponding to point $(i+1,j)$ includes the term $u_{i,j}$. We conclude that the matrix A_2 is symmetric.

Since the sum of all elements on each row of A_2 is less than or equal to zero, the Gershgorin circle theorem implies that A_2 is negative semidefinite. We proceed by showing that A_2 is negative definite. We begin with the case when Γ is convex; see Figure 2. Let us define the inner product between two real-valued grid functions $\boldsymbol{u}$ and $\boldsymbol{v}$ by

$$(\boldsymbol{u},\boldsymbol{v})_h = \sum_{i=1}^{N}\sum_{j=js(i)}^{je(i)} u_{i,j}v_{i,j} \equiv \sum_{j=1}^{M}\sum_{i=is(j)}^{ie(j)} u_{i,j}v_{i,j}.$$

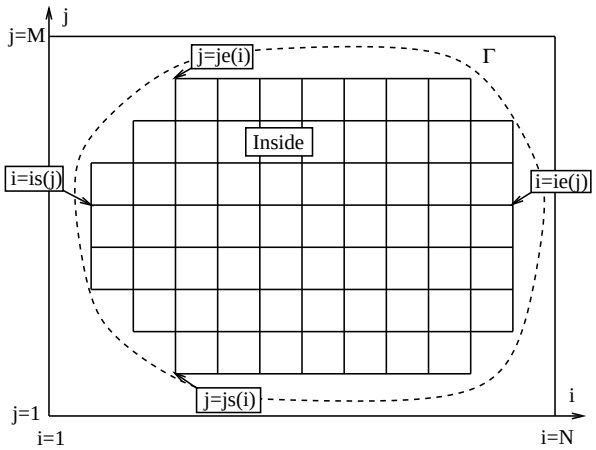


FIG. 2. *Notation used to show that A_2 is negative definite.*

It is convenient to treat the terms in $(\boldsymbol{u}, A_2\boldsymbol{u})_h$ "dimension by dimension." We have

$$
\begin{aligned}
(\boldsymbol{u}, A_2\boldsymbol{u})_h &= \sum_{i=1}^{N} S_1(i) + \sum_{j=1}^{M} S_2(j),\\
S_1(i) &= \sum_{j=js(i)+1}^{je(i)-1} u_{i,j} D_+^y D_-^y u_{i,j} + \frac{1}{h^2} u_{i,js(i)}(-2u_{i,js(i)} + u_{i,js(i)+1})\\
&\quad + \frac{1}{h^2} u_{i,je(i)}(-2u_{i,je(i)} + u_{i,je(i)-1}),\\
S_2(j) &= \sum_{i=is(j)+1}^{ie(j)-1} u_{i,j} D_+^x D_-^x u_{i,j} + \frac{1}{h^2} u_{is(j),j}(-2u_{is(j),j} + u_{is(j)+1,j})\\
&\quad + \frac{1}{h^2} u_{ie(j),j}(-2u_{ie(j),j} + u_{ie(j)-1,j}).
\end{aligned}
$$

The sum in $S_1(i)$ satisfies

$$
\begin{aligned}
\sum_{j=js(i)+1}^{je(i)-1} u_{i,j} D_+^y D_-^y u_{i,j} &= \sum_{j=js(i)+1}^{je(i)-1} u_{i,j} \frac{1}{h}(D_-^y u_{i,j+1} - D_-^y u_{i,j})\\
&= -\sum_{j=js(i)+1}^{je(i)-1} (D_-^y u_{i,j})^2 - \frac{u_{i,js(i)}}{h} D_-^y u_{i,js(i)+1} + \frac{u_{i,je(i)-1}}{h} D_-^y u_{i,je(i)}.
\end{aligned}
$$

Now, the two last terms in $S_1(i)$ satisfy

$$
\begin{aligned}
&\frac{u_{i,js(i)}}{h^2}(-2u_{i,js(i)} + u_{i,js(i)+1}) + \frac{u_{i,je(i)}}{h^2}(-2u_{i,je(i)} + u_{i,je(i)-1})\\
&\qquad = -\frac{u_{i,js(i)}^2}{h^2} + \frac{u_{i,js(i)}}{h} D_-^y u_{i,js(i)+1} - \frac{u_{i,je(i)}^2}{h^2} - \frac{u_{i,je(i)}}{h} D_-^y u_{i,je(i)}.
\end{aligned}
$$

Therefore,

$$S_1(i) = -\sum_{j=js(i)+1}^{je(i)} (D_-^y u_{i,j})^2 - \frac{u_{i,js(i)}^2}{h^2} - \frac{u_{i,je(i)}^2}{h^2}.$$

In the same way,

$$S_2(j) = -\sum_{i=is(j)+1}^{ie(j)} (D_-^x u_{i,j})^2 - \frac{u_{is(j),j}^2}{h^2} - \frac{u_{ie(j),j}^2}{h^2}.$$

Since all terms in $(\boldsymbol{u}, A_2\boldsymbol{u})_h$ are less than or equal to zero, $(\boldsymbol{u}, A_2\boldsymbol{u})_h$ can only be zero if all terms are zero. For example, $S_1(i) = 0$ if and only if

$$\begin{aligned} D_-^y u_{i,j} &= 0, \quad js(i)+1 \le j \le je(i), \\ u_{i,js(i)} &= 0, \\ u_{i,je(i)} &= 0, \end{aligned}$$

which implies $u_{i,j} \equiv 0$, $js(i) \le j \le je(i)$. In the same way, $S_2(j) = 0$ if and only if $u_{i,j} \equiv 0$, $is(j) \le i \le ie(j)$. This proves that $(\boldsymbol{u}, A_2\boldsymbol{u})_h = 0$ if and only if $\boldsymbol{u} \equiv 0$. The nonconvex case can be handled in the same way, except that the terms S_1 and S_2 must be divided into several parts, corresponding to the number of times the boundary splits the same grid line. Hence, the matrix A_2 is negative definite for both convex and nonconvex boundaries.

We have shown that both A_1 and A_2 are symmetric, A_1 is negative semidefinite, and A_2 is negative definite. Hence, Lemma 2.2 applies, so the time-step restriction of (4.2), (4.5) is determined only by A_2, i.e., the interior formula, and the solution is uniformly bounded in time. Moreover, the solution at the new time level, $\boldsymbol{u}^{n+1}$, can be computed pointwise since the matrix A_1 is diagonal.

4.3. Accuracy. We have constructed our difference approximation "dimension by dimension." If we write the approximation in matrix form, the coefficient matrix is symmetric and negative definite. Therefore, there are no stability problems. However, it is not obvious that the approximation is second order accurate. The reason is that an exterior point P can have two interior points P_1, P_2 as neighbors, P_1 in the x-direction and P_2 in the y-direction; see Figure 3. In this case the value of u at P is not unique, since it depends on which interpolation direction we use. For our scheme, this does not matter because we eliminate $u(P)$ both from $D_+^x D_-^x u(P_1)$ and $D_+^y D_-^y u(P_2)$, using the corresponding interpolation formula. However, the usual truncation error analysis fails.

We shall now use a more refined argument to show that the approximation is second order accurate. For simplicity, we study only the semidiscrete problem where time is left continuous,

$$\frac{\partial^2 u_h}{\partial t^2} = \tilde{\Delta}_h u_h + \frac{\tilde{f}}{h^2}, \tag{4.6}$$

$$u_h(\boldsymbol{x_j}, 0) = u_0(\boldsymbol{x_j}), \quad \frac{\partial u_h}{\partial t}(\boldsymbol{x_j}, 0) = u_1(\boldsymbol{x_j}).$$

We assume that the solution of the continuous wave equation (4.1) is smooth and can be extended smoothly from Ω to a larger region Ω_1 which contains all external points

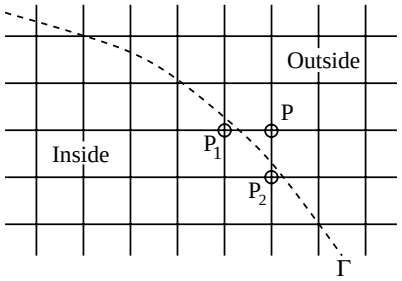


FIG. 3. *The outside point P is used when discretizing the Laplacian at both of the inside points P_1 and P_2.*

with boundary points as neighbors. Let $u(x, y, t)$ be the solution of (4.1). It solves the inhomogeneous difference equation

$$u_{tt} = \Delta_h u + h^2 G. \tag{4.7}$$

Here $h^2 G$ represents the truncation error. Let $P_1 = (x, y)$ be a boundary point and $(x + h, y)$ an exterior point. The extended solution satisfies an inhomogeneous interpolation formula:

$$(1 - \alpha_1)u(x, y, t) + \alpha_1 u(x + h, y, t) = f(x + \alpha_1 h, y, t) + h^2 g_1. \tag{4.8}$$

If also $(x, y + h)$ is an exterior point, then there is another interpolation formula:

$$(1 - \alpha_2)u(x, y, t) + \alpha_2 u(x, y + h, t) = f(x, y + \alpha_2 h, t) + h^2 g_2. \tag{4.9}$$

We use (4.8) and (4.9) to eliminate $u(x+h, y, t)$ and $u(x, y+h, t)$ from $D_+^x D_-^x u(x, y, t)$ and $D_+^y D_-^y u(x, y, t)$, respectively. After we have eliminated all exterior points, we obtain

$$u_{tt} = \tilde{\Delta}_h u + h^2 G + g + \frac{\tilde{f}}{h^2} \quad \text{for } \boldsymbol{x} = \boldsymbol{x_j},\ |m_{\boldsymbol{j}}| = 1,\ t > 0. \tag{4.10}$$

Note that $g \neq 0$ and $\tilde{f} \neq 0$ only at boundary points and $\tilde{\Delta}_h u = \Delta_h u$ at all interior points. At boundary points (x, y) (which have at least one neighboring exterior point), we obtain a q-point formula with $q \leq 4$ of the form

$$h^2 \tilde{\Delta}_h u(\boldsymbol{x}_{i,j}, t) = -(4 + d_{i,j})u(\boldsymbol{x}_{i,j}, t) + \sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}| u(\boldsymbol{x_j}, t), \tag{4.11}$$

where the sum extends over the nearest neighbors $(x_i \pm h, y_j)$, $(x_i, y_j \pm h)$, and $d_{i,j} > 0$.

Subtracting the difference approximation (4.6) from (4.10) gives us, for the error $e = u - u_h$,

$$\begin{aligned} e_{tt} &= \tilde{\Delta}_h e + h^2 G + g, \\ e(\cdot, 0) &= e_t(\cdot, 0) = 0. \end{aligned}$$

To analyze the effect the term g has on the error, we study

$$\tilde{\Delta}_h \varphi = g.$$

Away from the boundary, $g = 0$ and $\tilde{\Delta}_h = \Delta_h$. It is easy to show that the solution of

$$\begin{aligned} \Delta_h \phi_{i,j} &= 0, \qquad m_{i,j} = 1, \\ \phi_{i,j} &= \gamma_{i,j}, \quad m_{i,j} = -1, \end{aligned}$$

takes its maximum at a boundary point (where $m_{i,j} = -1$). Let (x_i, y_j) be the boundary point where $|\varphi(x_i, y_j)| = |\varphi|_\infty$. From (4.11) we have

$$(4 + d_{i,j})|\varphi|_\infty \le \sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}||\varphi|_\infty + h^2 |g|_\infty.$$

Hence,

$$|\varphi|_\infty \le \frac{h^2}{4 + d_{i,j} - \sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}|} |g|_\infty.$$

Since $d_{i,j} > 0$ and $\sum_{\boldsymbol{j}} |m_{\boldsymbol{j}}| \le 3$,

$$|\varphi|_\infty \le h^2 |g|_\infty.$$

From the definition of g, it is a smooth function of t. Therefore, $\tilde{e} = e + \varphi$ solves

$$\tilde{e}_{tt} = \tilde{\Delta}_h \tilde{e} + h^2 G + \varphi_{tt}, \quad |\varphi_{tt}|_\infty \le h^2 |g_{tt}|_\infty,$$

i.e.,

$$|e|_\infty = \mathcal{O}(h^2).$$

This shows that the approximation is second order accurate.

5. Numerical examples. In this section we consider (4.1) with a forcing function,

$$\begin{aligned} u_{tt} &= \Delta u + F(\boldsymbol{x}, t), && \boldsymbol{x} \in \Omega, \quad t > 0, \\ u(\boldsymbol{x}, t) &= f(\boldsymbol{x}, t), && \boldsymbol{x} \in \Gamma, \quad t > 0, \\ u(\boldsymbol{x}, 0) &= u_0(\boldsymbol{x}), \quad u_t(\boldsymbol{x}, 0) = u_1(\boldsymbol{x}), && \boldsymbol{x} \in \Omega. \end{aligned} \tag{5.1}$$

This problem will be solved by both a fully second and an internally fourth order accurate method, and the forcing functions will be chosen such that an exact solution is known. The second order scheme is given by (4.2) and (4.5), and the internally fourth order scheme is obtained by adding the correction term

$$\Delta_{h,4} v^n_{i,j} = -\frac{h^2}{12} \left(D^x_+ D^x_- \gamma_{i,j} D^x_+ D^x_- + D^y_+ D^y_- \gamma_{i,j} D^y_+ D^y_- \right) v^n_{i,j}$$

to the right-hand sides of (4.2) and (4.5), respectively. Here

$$\gamma_{i,j} = \begin{cases} 1, & m_{i,j} = 1, \\ 0 & \text{otherwise}, \end{cases}$$

where $m_{i,j}$ is defined in section 4.1.

The correction term $\Delta_{h,4} v^n_{i,j}$ gives a symmetric negative semidefinite contribution to the matrix representation of the scheme that does not involve the boundary.

TABLE 1
Grid refinement study for the Dirichlet problem for a trigonometric exact solution; CFL = 0.5.

	2'nd order error			4'th order error		
t	$N = 100$	$N = 200$	ratio	$N = 100$	$N = 200$	ratio
0.330	2.54e-02	5.77e-03	4.4	1.27e-02	3.58e-03	3.5
1.980	2.30e-02	5.47e-03	4.2	1.46e-02	3.97e-03	3.7

Hence, the internally fourth order scheme is stable. However, the internally fourth order scheme can be expected to be only second order accurate near the boundary, since a second order approximation of the Dirichlet boundary condition is used. For simplicity, the internally fourth order scheme will henceforth be called the fourth order scheme.

We start the numerical integration at $n = 0$. For the cases where an analytical solution is known, we use this solution to initialize the computation at time levels t_{-1} and t_0. For the cases where an analytical solution is not known we use the initialization

$$v_{i,j}^{-1} = u_0(x_i, y_j) - k u_1(x_i, y_j) + \frac{k^2}{2}(D_+^x D_-^x + D_+^y D_-^y) u_0(x_i, y_j),$$

and $v_{i,j}^0 = u_0(x_i, y_j)$.

We will denote the CFL-number by $\mathrm{CFL} \equiv k/h$. Note that for a two-dimensional periodic domain, the scheme (4.2) is stable for $\mathrm{CFL} \leq 1/\sqrt{2}$. Also note that all errors are measured in the max-norm.

5.1. Convergence study for a trigonometric exact solution. Let us choose the forcing function F and boundary data f such that the exact solution is the trigonometric traveling wave

$$u(x, y, t) = \sin(\omega(x - t)) \sin(\omega y), \quad \omega = 4\pi. \tag{5.2}$$

The domain Ω is taken to be an ellipse centered at the origin with semiaxes $x_s = 1$ and $y_s = 0.75$. The Cartesian grid covers the rectangle $-1.1 \leq x \leq 1.1$, $-0.85 \leq y \leq 0.85$. In Table 1, we present a grid refinement study for the two schemes with CFL= 0.5. In Table 2 we present the same study with CFL=0.1. Note that the error in the second order scheme is not improved by decreasing the CFL number, indicating that the error is dominated by spatial discretization errors. The error for the fourth order scheme is improved somewhat by decreasing the CFL number, implying that temporal discretization errors cannot be neglected when CFL= 0.5. For both CFL-numbers, the error is smaller for the fourth order scheme than for the second order scheme. However, the order of convergence is only around two for the fourth order scheme and CFL= 0.1. In section 3.2 the spatial discretization error is shown to consist of a second order amplitude error arising from the boundary discretization and a fourth order phase error originating from the discretization in the interior. It is therefore likely that the dominant errors for the fourth order scheme and CFL= 0.1 are generated at the boundary.

5.2. Convergence study for an inwards traveling wave solution in a circle. To illustrate the benefits of using a fourth order scheme away from the boundary, we select the forcing function F and boundary data f such that the exact solution

TABLE 2
Grid refinement study for the Dirichlet problem for a trigonometric exact solution; CFL = 0.1.

	2'nd order error			4'th order error		
t	$N = 100$	$N = 200$	ratio	$N = 100$	$N = 200$	ratio
0.330	2.31e-02	5.77e-03	4.0	1.06e-02	2.23e-03	4.8
1.980	2.40e-02	5.92e-03	4.1	7.55e-03	2.16e-03	3.5

TABLE 3
Grid refinement study for the Dirichlet problem for an inwards traveling wave solution. Here, CFL = 0.5.

	2'nd order error			4'th order error
t	$N = 400$	$N = 800$	ratio	$N = 400$
0.315	1.51e-02	3.66e-03	4.12	6.51e-03
0.525	3.29e-02	8.79e-03	3.75	1.15e-02
1.155	8.44e-02	2.67e-02	3.17	3.29e-02
1.365	1.03e-01	3.40e-02	3.03	4.16e-02

(in polar coordinates) is a spatially localized traveling wave,

$$u(r,t) = \phi(r+t), \quad \phi(\xi) = \frac{1}{4}\left(1 + \tanh\frac{\xi-\xi_0}{\epsilon}\right)\left(1 - \tanh\frac{\xi-\xi_1}{\epsilon}\right). \tag{5.3}$$

Note that such waves are exact solutions to the unforced wave equation in one and three space dimensions, but they are not in the two-dimensional case. The domain Ω is taken to be the circle, $|r| \le 2$, and the Cartesian grid covers the square $-2.1 \le x \le 2.1$, $-2.1 \le y \le 2.1$.

The parameters in the exact solution are chosen so that initially the wave is essentially outside the domain and enters the region through the boundary after some time. To make the problem challenging to solve numerically, we make the transitions around $\xi = \xi_0$ and $\xi = \xi_1$ rapid and close together by choosing

$$\xi_0 = 2.2, \quad \xi_1 = 2.4, \quad \epsilon = 0.035.$$

The maximum of the wave reaches the boundary at $t = 0.3$ and has passed through the boundary after $t \approx 0.4 - 0.5$. In Table 3 we present a study of the errors for the two schemes with CFL= 0.5 both when the wave has reached the boundary and after the wave has passed the boundary. In Table 4 we present a similar study but with CFL= 0.1. Note first that the error in the fourth order scheme is improved by decreasing the CFL-number for $N = 400$, indicating again that time discretization errors are not small when CFL= 0.5. Furthermore, note that the fourth order method with CFL= 0.1 gives close to fourth order accuracy after the wave has passed the boundary, implying that the phase error dominates; see the discussions in sections 5.1 and 3.2. Also, the error in the second order scheme does not improve when the CFL number is decreased, indicating again that the second order scheme is dominated by spatial discretization errors. When 400 grid points are used in each direction, the second order scheme produces a solution that has many spurious oscillations, while the fourth order scheme gives a much cleaner result; see Figure 4.

TABLE 4
Grid refinement study for the Dirichlet problem for an inwards traveling wave solution. Here, CFL = 0.1.

	2'nd order error	4'th order error		
t	$N = 400$	$N = 400$	$N = 800$	ratio
0.315	1.73e-02	4.39e-03	8.37e-04	5.24
0.525	3.98e-02	6.55e-03	8.30e-04	7.89
1.155	9.95e-02	1.09e-02	8.12e-04	13.44
1.365	1.20e-01	1.31e-02	9.64e-04	13.62

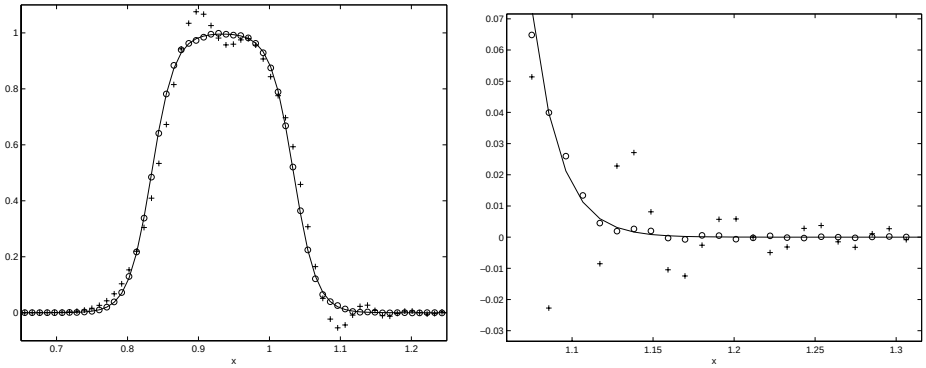


FIG. 4. *The inwards traveling wave solution for the Dirichlet problem along the line* $y = -h/2$ *at time* $t = 1.365$. *The right figure shows a close-up centered around* $x = 1.2$. *The exact solution is a solid line, the computed solution with the second order scheme is denoted by* (+), *and the fourth order scheme is marked with* (o). *In these computations,* $N = 400$ *and* CFL $= 0.1$.

5.3. Convergence study for an outwards traveling wave solution in a circle. To further study the benefits of using a fourth order scheme away from the boundary, we let the exact solution be an outwards traveling wave,

$$u(r,t) = \phi(r-t), \tag{5.4}$$

where ϕ is given by (5.3). The domain Ω is in this case taken to be the unit circle, $|r| \le 1$, and the Cartesian grid covers the square $-1.1 \le x \le 1.1$, $-1.1 \le y \le 1.1$. Here, the parameters in ϕ are taken to be

$$\xi_0 = 0.2, \quad \xi_1 = 0.4, \quad \epsilon = 0.035.$$

The wave reaches the boundary at $t \approx 0.5 - 0.6$ and has passed through the boundary after $t \approx 0.8 - 0.9$. In Table 5 we present a grid convergence study for the two schemes with CFL= 0.5 before and after the wave has passed the boundary. In Table 6 we present the same study but with CFL= 0.1. Note that the fourth order method gives fourth order convergence for the smaller CFL-number before the wave reaches the boundary and gives second order convergence after the wave has reached the boundary. When 200 grid points are used in each direction, the second order scheme produces a solution that has many spurious oscillations, while the fourth order scheme gives a much cleaner result; see Figure 5.

TABLE 5
Grid refinement study for the Dirichlet problem for an outwards traveling wave solution; CFL = 0.5.

	2'nd order error			4'th order error		
t	$N = 200$	$N = 400$	ratio	$N = 200$	$N = 400$	ratio
0.22	1.99e-02	5.44e-03	3.7	6.70e-03	1.79e-03	3.7
0.33	2.78e-02	7.82e-03	3.6	9.64e-03	2.60e-03	3.7
0.66	5.35e-02	1.40e-02	3.8	1.61e-02	4.47e-03	3.6
0.77	5.77e-02	1.76e-02	3.3	2.30e-02	6.11e-03	3.8

TABLE 6
Grid refinement study for the Dirichlet problem for an outwards traveling wave solution; CFL = 0.1.

	2'nd order error			4'th order error		
t	$N = 200$	$N = 400$	ratio	$N = 200$	$N = 400$	ratio
0.22	2.53e-02	7.08e-03	3.6	2.37e-03	1.40e-04	17.0
0.33	3.50e-02	1.01e-02	3.5	3.33e-03	2.03e-04	16.5
0.66	5.73e-02	1.71e-02	3.3	8.50e-03	7.34e-04	11.6
0.77	6.46e-02	2.23e-02	2.9	8.26e-03	1.17e-03	7.1

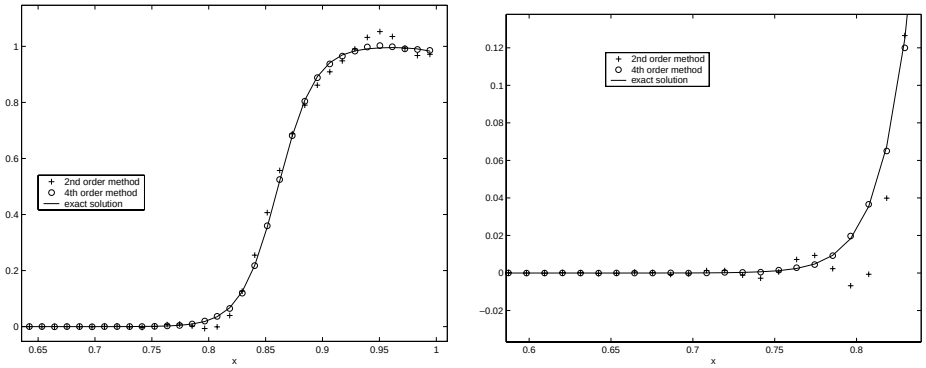


FIG. 5. *The outwards traveling wave solution for the Dirichlet problem along the line* $y = -h/2$ *at time* $t = 0.66$. *The right figure shows a close-up centered around* $x = 0.7$. *The exact solution is a solid line, the computed solution with the second order scheme is denoted by* (+), *and the fourth order scheme is marked with* (o). *Here* $N = 200$ *and* CFL= 0.1.

5.4. Bouncing wave in an ellipse. Here we study the homogeneous problem

$$F(\boldsymbol{x}, t) \equiv 0, \quad f(\boldsymbol{x}, t) \equiv 0,$$

in a domain bounded by an ellipse centered at the origin, with semiaxes $x_s = 2.0$ and $y_s = 2.54$. The Cartesian grid covers the square $-2.1 \le x \le 2.1$, $-2.64 \le y \le 2.64$. We take initial data to be

$$u_0(x, y) = \phi(r),$$

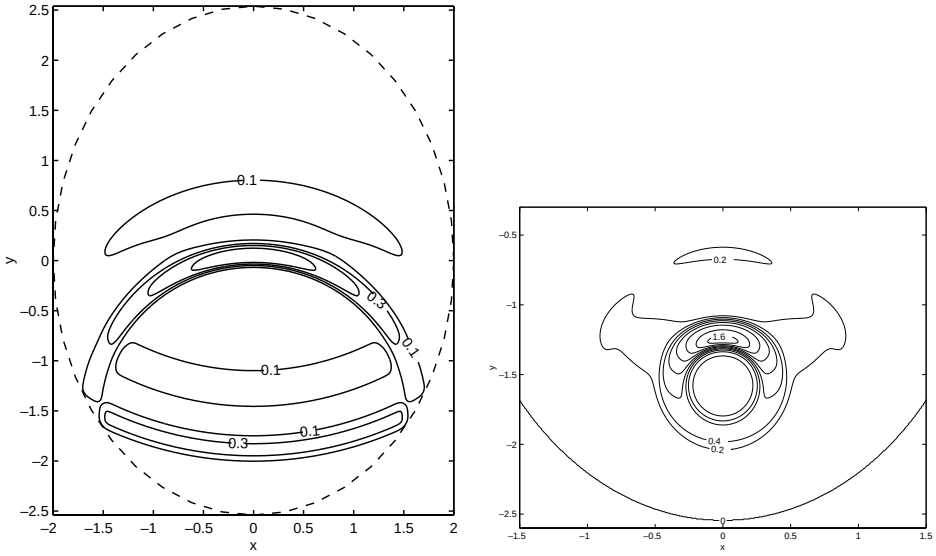


FIG. 6. *The bouncing wave reference solution for the Dirichlet problem with the fourth order scheme at* $t = 3.15$ *(left) and* $t = 4.41$ *(right). Here* $N = 800$, CFL $= 0.1$. *The contour spacing is* 0.2*, and the dashed curve in the left plot indicates the boundary.*

where $\phi(r)$ is given by (5.3) and $r = \sqrt{x^2 + (y - y_F)^2}$. The upper focal point is located at $y_F = \sqrt{y_s^2 - x_s^2} \approx 1.56$ and

$$u_1(\boldsymbol{x}) = u_1(r) = -\phi'(r).$$

The parameters in $\phi(r)$ are

$$\xi_0 = 0.2, \quad \xi_1 = 0.4, \quad \epsilon = 0.035.$$

Note that the initial data is chosen such that the wave is essentially traveling in the positive r-direction out from the focal point $(0, y_F)$. By making a ray-tracing argument, we see that a high frequency wave should reflect the boundary and refocus at the other focal point $(0, -y_F)$. To verify this behavior, we make a reference calculation using a fine grid with $N = 800$ and the fourth order method with CFL$=$ 0.1. We then make calculations with $N = 400$ and $N = 100$ for the fourth and the second order methods. These solutions are compared to the reference calculation at the time $t = 4.41$, just before the solution is refocused at the other focal point. In Figure 6, we show contour plots of the reference calculation at $t = 3.15$ and $t = 4.41$. In Figure 7, we display contour plots of the solutions using the fourth and second order schemes for $N = 400$ at $t = 4.41$. We also plot the numerical solutions along the line $x = 0$, where the deviation from the reference solution is the largest; see Figure 8. In Figures 9 and 10, the corresponding calculations are presented for $N = 100$. Clearly, the fourth order method gives the best result.

6. Conclusions. Numerical methods have been proposed and analyzed for solving a second order wave equation without first writing it as a first order system. Both the Dirichlet and the Neumann problems were treated for the one-dimensional case. The Dirichlet problem was analyzed in detail for general two-dimensional domains, and we proved that the proposed scheme is stable and conservative both for second and

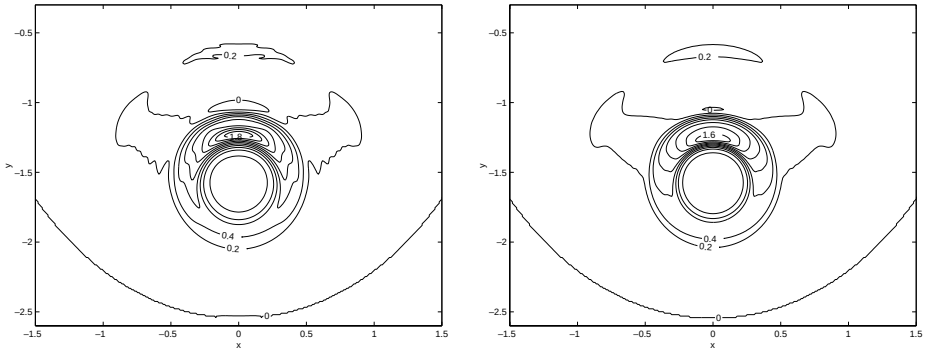


FIG. 7. *The bouncing wave solutions for the Dirichlet problem at* $t = 4.41$. *Left: the second order scheme with* CFL $= 0.5, N = 400$. *Right: the fourth order scheme with* CFL $= 0.1$, $N = 400$. *The contour spacing is* 0.2.

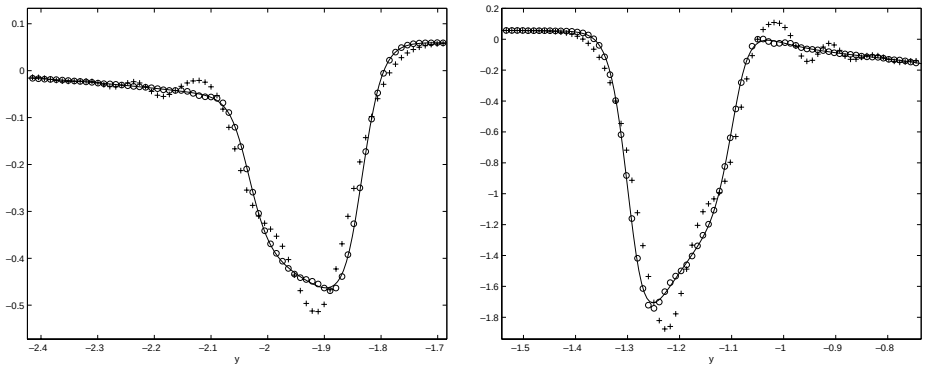


FIG. 8. *The bouncing wave solutions on a fine grid for the Dirichlet problem along the line* $x = 0$ *at* $t = 4.41$, *centered around* $y = -2$ *(left) and* $y = -1.2$ *(right). Solid: the reference solution, "+": the second order scheme with* CFL $= 0.5$, $N = 400$. *"o": the fourth order scheme with* CFL $= 0.1, N = 400$. *Note the over- and undershoots obtained with the second order scheme.*

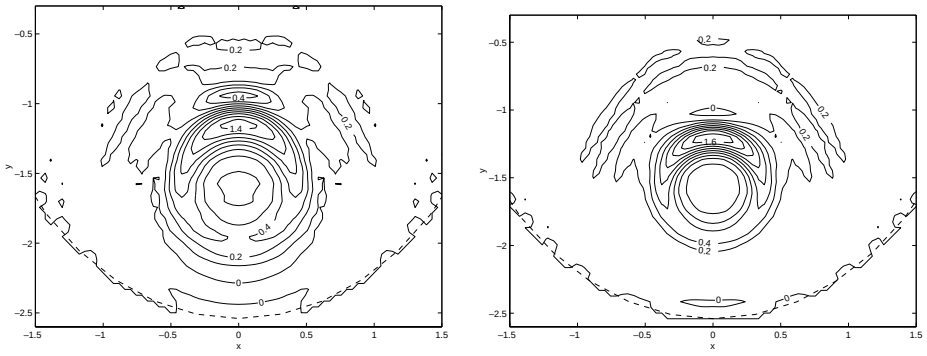


FIG. 9. *The bouncing wave solutions on a coarse grid for the Dirichlet problem at* $t = 4.41$. *Left: the second order scheme with* CFL $= 0.5$, $N = 100$. *Right: the fourth order scheme with* CFL $= 0.1$, $N = 100$. *The contour spacing is* 0.2, *and the dashed curve represents the boundary.*

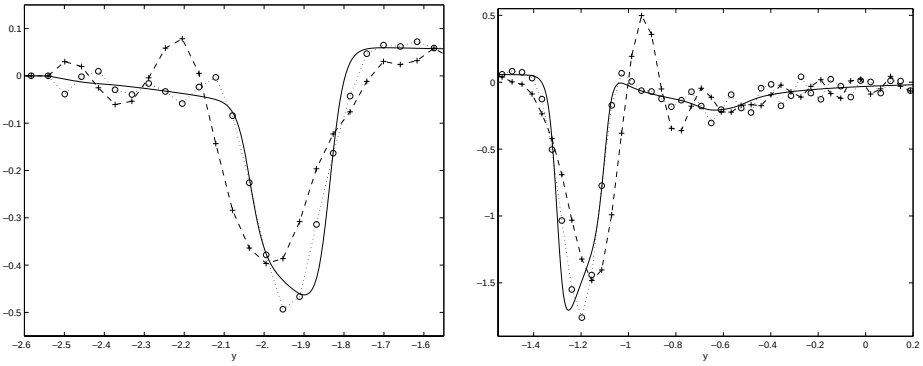


FIG. 10. *The bouncing wave solutions for the Dirichlet problem along the line* $x = 0$ *at* $t = 4.41$*, centered around* $y = -2.1$ *(left) and* $y = -0.75$ *(right). The solid line is the reference solution, and the dashed line with "+" is the second order scheme with* $\mathrm{CFL} = 0.5$, $N = 100$. *The dotted line with "o" marks the fourth order scheme with* $\mathrm{CFL} = 0.1$, $N = 100$. *Note that the spurious oscillations are much larger with the second order scheme.*

fourth order spatial discretizations. We are currently working on the two-dimensional Neumann problem [7], and we plan to extend the approach to three space dimensions.

For the fourth order spatial discretization, we have seen that the second order temporal discretization error cannot be neglected unless the CFL-number is reduced substantially below the stability limit. In future work we intend to develop a higher order time integration method where the accuracy better matches that of the fourth order spatial discretization.

In many applications the wave propagation speed varies in space. We see no difficulty modifying our approach to handle smoothly varying coefficients, but additional work will be required to treat discontinuous coefficients. Systems of second order wave equations also occur in applications, for example in general relativity. Another example is Maxwell's equations for electromagnetic wave propagation, which often is given as a first order system, but also can be written as a second order system. We believe that generalizing the proposed method to systems will provide an accurate and straightforward technique for analyzing these types of problems.

REFERENCES

[1] G. BROWNING, H.-O. KREISS, AND J. OLIGER, *Mesh refinement*, Math. Comp., 27 (1973), pp. 29–39.

[2] L. COLLATZ, *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1960.

[3] A. DITKOWSKI, K. DRIDI, AND J. S. HESTHAVEN, *Convergent Cartesian grid methods for Maxwell's equations in complex geometries*, J. Comput. Phys., 170 (2001), pp. 39–80.

[4] T. A. DRISCOLL AND B. FORNBERG, *Block pseudospectral methods for Maxwell's equations* II*: Two-dimensional, discontinuous-coefficient case*, SIAM J. Sci. Comput., 21 (1999), pp. 1146–1167.

[5] B. GUSTAFSSON, H.-O. KREISS, AND J. OLIGER, *Time Dependent Problems and Difference Methods*, Wiley–Interscience, New York, 1995.

[6] H. JOHANSEN AND P. COLELLA, *A Cartesian grid embedded boundary method for Poisson's equation on irregular domains*, J. Comput. Phys., 147 (1998), pp. 60–85.

[7] H.-O. KREISS, N. A. PETERSSON, AND J. YSTRÖM, *Difference Approximations for the Second Order Wave Equation*, Technical report UCRL-JC-145614, Center for Applied Scientific Computing, Lawrence Livermore National Lab, Livermore, CA, 2001.

[8] R. B. Pember, J. B. Bell, P. Collella, W. Y. Crutchfield, and M. Welcome, *An adaptive Cartesian grid method for unsteady compressible flow in irregular regions*, J. Comput. Phys., 120 (1995), p. 278–304.

[9] B. Strand, *Simulations of acoustic wave phenomena using high-order finite difference approximations*, SIAM J. Sci. Comput., 20 (1999), pp. 1585–1604.

[10] R. Weller and G. H. Shortley, *Calculation of stresses within the boundary of photoelastic models*, J. Appl. Mech., 6 (1939), pp. A71–A78.

[11] C. Zhang and R. LeVeque, *The immersed interface method for acoustic wave equations with discontinuous coefficients*, Wave Motion, 25 (1997), pp. 237–263.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1968–1992

FULLY DISCRETE, ENTROPY CONSERVATIVE SCHEMES OF ARBITRARY ORDER*

P. G. LEFLOCH†, J. M. MERCIER‡, AND C. ROHDE§

Abstract. We consider weak solutions of (hyperbolic or hyperbolic-elliptic) systems of conservation laws in one-space dimension and their approximation by finite difference schemes in conservative form. The systems under consideration are endowed with an entropy-entropy flux pair. We introduce a general approach to construct *second and third order accurate, fully discrete* (in both space and time) *entropy conservative schemes.* In general, these schemes are fully nonlinear implicit, but in some important cases can be explicit or linear implicit. Furthermore, semidiscrete entropy conservative schemes of *arbitrary order* are presented. The entropy conservative schemes are used to construct a numerical method for the computation of weak solutions containing nonclassical regularization-sensitive shock waves. Finally, specific examples are investigated and tested numerically. Our approach extends the results and techniques by Tadmor [in *Numerical Methods for Compressible Flows—Finite Difference, Element and Volume Techniques,* ASME, New York, 1986, pp. 149–158], LeFloch and Rohde [*SIAM J. Numer. Anal.,* 37 (2000), pp. 2023–2060].

Key words. hyperbolic conservation law, finite difference scheme, entropy conservative scheme, system of mixed type, diffusion, dispersion, nonclassical shock

AMS subject classifications. 65M06, 35L65, 35M10

PII. S003614290240069X

1. Introduction.

In this paper, we are interested in the numerical approximation of discontinuous solutions of general systems of conservation laws of the form

$$\partial_t u + \partial_x f(u) = 0, \qquad u = u(x,t) \in \mathbb{R}^N, \qquad x \in \mathbb{R},\ t > 0, \tag{1.1}$$

endowed with a smooth entropy-entropy flux pair $(U, F) : \mathbb{R}^N \to \mathbb{R}^2$. In (1.1), the flux-function $f : \mathbb{R}^N \to \mathbb{R}^N$ is a smooth given mapping. As is well known, we should seek solutions satisfying the entropy inequality

$$\partial_t U(u) + \partial_x F(u) \leq 0 \tag{1.2}$$

understood in the sense of distributions.

From the numerical standpoint, following Lax and Wendroff [12], it is natural to search for (fully discrete in space and time) conservative schemes associated with (1.1) which, furthermore, satisfy a discrete version of the inequality (1.2). Whenever the Cauchy problem for (1.1)–(1.2) is well-posed (for instance, when (1.1) is a scalar conservation law with convex flux) such a scheme can converge only to the (so-called) entropy solution of interest.

Weak (entropy) solutions of (1.1) can be considered as limits of solutions of higher order systems with vanishing regularization terms. The physical meaning of these terms comes from viscosity, heat conduction, or capillarity usually leading to a smooth

*Received by the editors January 8, 2002; accepted for publication May 30, 2002; published electronically December 3, 2002. The authors were supported by the European Training, Mobility, and Research Grant HCL ERBFMRXCT96033.

http://www.siam.org/journals/sinum/40-5/40069.html

†Centre de Mathématiques Appliquées, Centre National de la Recherche Scientifique, U.M.R. 7641, Ecole Polytechnique, 91128 Palaiseau Cedex, France (lefloch@cmap.polytechnique.fr).

‡SOPHIS Technology, 34 rue Boissy d'Anglas, 75008 Paris, France (jeanmarc.mercier@free.fr).

§Institut für Angewandte Mathematik, Albert-Ludwigs-Universität Freiburg, Hermann-Herder-Str. 10, D-79104 Freiburg im Breisgau, Germany (chris@mathematik.uni-freiburg.de).

solution that satisfies (1.2) in the pointwise sense. In some situations it is necessary to control explicitly the rate of dissipation that one introduces (in the continuous as well as in the discrete setting).

In this context it has been suggested that the numerical approximation of (1.1) should be based on schemes satisfying (1.2) as an *equality* (cf. [10]), that is

$$\partial_t U(u) + \partial_x F(u) = 0. \tag{1.3}$$

High order terms such as viscosity, heat conduction, capillarity, etc., should then be added to such an *entropy conservative scheme* in a way to get an *entropy dissipative scheme*, i.e., satisfying a discrete (consistent) version of (1.2). The notion of *entropy conservative* schemes for conservation laws was introduced first and investigated in a pioneering work by Tadmor [24, 25] when constructing semidiscrete difference schemes satisfying a discrete form of (1.2). For another approach we refer to [21]. In a close context, linear implicit, fully discrete, *energy conservative* schemes were designed in Aregba-Driollet and Mercier [4] (in the spirit of a fully nonlinear scheme introduced by Strauss and Vasquez [22]) to study solutions of semilinear hyperbolic systems satisfying an *energy conservation*, i.e., satisfying (1.3) for a (possibly nonconvex) energy U.

In the light of the above work, attention in the present paper is focused precisely on constructing fully discrete, conservative, and entropy conservative schemes for conservation laws, consistent with both (1.1) and (1.3).

The investigation of semidiscrete schemes (keeping the time variable continuous) was completed only recently. A second order entropy conservative scheme was discovered by Tadmor [24, 25] who introduced this notion in order to construct schemes satisfying a discrete form of (1.2). Next, the notion was further investigated by LeFloch and Rohde [16], who discovered a class of third order entropy conservative schemes.

The study of fully discrete schemes for diffusive-dispersive conservation laws was initiated by Chalons and LeFloch [5]. The authors made a direct use of the semidiscrete numerical fluxes proposed in the earlier papers. By enforcing a suitable CFL stability condition, the entropy inequality (1.2) holds, provided diffusive terms are taken into account in the right-hand side of (1.1).

Our motivation to construct entropy conservative schemes was to study systems of conservation laws that either have nonconvex modes or are of hyperbolic-elliptic type. In this paper we will focus on two representative examples: the first is the cubic scalar conservation law, a nonconvex hyperbolic equation, for which dynamics is well understood and which is used as a test model. The second is a p-system that models adiabatic phase transition dynamics, a hyperbolic-elliptic system; see Truskinovsky [26], Abeyaratne and Knowles [2, 3], and LeFloch [13] for related results in the linearly degenerate case, and see Mercier and Piccoli [18] and references therein for the genuinely nonlinear case. The main difficulty of a nonconvex hyperbolic or hyperbolic-elliptic system of conservation laws is that the single entropy inequality (1.2) does not characterize a unique solution of the system and further selection mechanisms must be added, specifically the so-called *kinetic relation.* For general nonconvex systems, we refer to Hayes and LeFloch [9], LeFloch and Thanh [17], and LeFloch [14].

Kinetic relations can be determined in several situations from physics. From the mathematical point of view they can be exhibited from regularization terms. Kinetic regularizations associated with difference schemes were numerically determined and

compared with analytical kinetic relations in [10]. The dependence of the kinetic relation upon physical and numerical parameters was discussed therein.

An important point is that capillarity terms require high order schemes (at least third order). Thus our first aim is to derive a general approach to construct finite difference schemes for systems of conservation laws that are

(1) fully discrete in space and time,
(2) conservative in the sense of Lax and Wendroff [12],
(3) entropy conservative in the sense of Tadmor [23, 24],
(4) and high order accurate (at least third order).

This program will be carried out in sections 2 and 3. First, we propose a general approach for the construction of such schemes in section 2. Next, in section 3, several classes of second and third order schemes are identified, which can be fully implicit, linear implicit, or explicit methods. This is certainly not a straightforward task. Recall that, for nonaffine f, there are no two time-level, fully discrete, *explicit,* and conservative schemes with smooth numerical flux satisfying a discrete version of the entropy equality; see [16].

In section 4 we return to the investigation of semidiscrete schemes. We will present entropy conservative schemes of *arbitrarily high order*. This can be transferred to the fully discrete case, however, only for a weaker form of entropy conservation.

Finally in section 5, adding appropriate dissipative terms, we will obtain schemes for the above mentioned model problems. Numerical experiments presented in particular in section 6 underline their good performance.

We emphasize that the techniques developed in this paper also apply to other types of evolution equations for which an energy conservation or dissipation is available, such as the heat, Schrödinger, or wave equations. A first result in this direction is given in the second part of subsection 5.2 (Theorem 5.2). Furthermore, these techniques, considered in the one dimensional case, apply straightforwardly to higher dimensions when using Cartesian grids.

2. A general approach to construct entropy conservative schemes. In this section we propose a general method to construct fully discrete, conservative, and entropy conservative schemes.

We follow the notation in Tadmor [24] and LeFloch and Rohde [16]. Call $v(u) = \nabla U(u)$ the *entropy variable* associated with the given entropy U. When the entropy is strictly convex, $v \mapsto v(u)$ is a one-to-one mapping. This can be used as a change of variable (Friedrichs and Lax [7]); that is, we can set

$$g(v) := f(u), \quad G(v) := F(u), \quad B(v) := Dg(v). \tag{2.1}$$

The matrix $B(v)$ is symmetric since $Dg(v) = Df(u)D^2U(u)^{-1}$ is symmetric matrix for U being a strictly convex entropy. It follows that there exists a scalar-valued function $\psi = \psi(v)$ such that $g = \nabla\psi$; in fact

$$\psi(v) = v \cdot g(v) - G(v), \tag{2.2}$$

uniquely defined up to a constant.

Furthermore, to deal with examples when U is not globally convex, the following assumption on the flux-function of (1.1) is made:

(2.3) $f(u)$ and $F(u)$ can be expressed as functions of the entropy variable v;

that is, (2.1) holds for some functions g and G. Then, again, ψ can be defined by (2.2). The assumption (2.3), which we make from now on, is motivated by several examples of interest; see [16] and section 5 below. We stress that (2.3) holds in $\mathbb{R}^N$ when U is strictly convex.

For mesh parameters $h, \tau > 0$, let $x_j = j\,h$, $j \in \mathbb{Z}$, and $t_n = n\,\tau$, $n \in \mathbb{N}_0$. We set $\lambda \equiv \tau/h$ and start discussing the (multilevel) time discretization. For $q \in \mathbb{N}$, choose a locally Lipschitz continuous mapping

$$u^* : \left(u^{-q+1}, \ldots, u^0\right) \in \mathbb{R}^{qN} \mapsto u^*\left(u^{-q+1}, \ldots, u^0\right) \in \mathbb{R}^N$$

consistent with the conservative variable u in the sense that

$$u^*(u, \ldots, u) = u, \quad u \in \mathbb{R}^N.$$

It will be called the *discrete conservative variable* in what follows. The integer q indicates the number of time-levels used by the scheme and is related to the order of accuracy in time. Setting $u_j^{*n} = u^*(u_j^{n-q+1}, \ldots, u_j^n)$, we approximate the continuous derivative $\partial_t u$ in (1.1) by the following discrete derivative:

$$\frac{u_j^{*n+1} - u_j^{*n}}{\tau}.$$

To guarantee that the difference equation is solvable in terms of the conservative variable u_j^{n+1}, we assume that

$$\text{(2.4)} \qquad \begin{array}{c} \text{the mapping } u \mapsto u^*\left(u^{n-q+1}, \ldots, u^n, u\right) \text{ is smoothly invertible} \\ \text{for all } u^{n-q+1}, \ldots, u^n \in \mathbb{R}^N. \end{array}$$

Next, choose some locally Lipschitz continuous mapping

$$U^* : (u^{-q+1}, \ldots, u^0) \in \mathbb{R}^{qN} \mapsto U^*\left(u^{-q+1}, \ldots, u^0\right) \in \mathbb{R}$$

consistent with the continuous entropy; i.e.,

$$U^*(u, \ldots, u) = U(u), \quad u \in \mathbb{R}^N.$$

It will be called the *discrete entropy function*. Also set $U_j^{*n} = U^*(u_j^{n-q+1}, \ldots, u_j^n)$.

As we will see below the two functions u^* and U^* cannot be chosen arbitrarily from each other. We make the following assumption.

Assumption 2.1. There exists a continuous mapping $v^* : \mathbb{R}^{(N+1)q} \to \mathbb{R}^N$ with the properties

$$\text{(2.5)} \qquad \begin{array}{ll} \text{(i)} & v^*(u, \ldots, u) = v(u) \qquad (v \in \mathbb{R}^N), \\ \text{(ii)} & U^*\left(u^{-q+1}, \ldots, u^0\right) - U^*\left(u^{-q}, \ldots, u^{-1}\right) \\ & \quad = \left(u^*\left(u^{-q+1}, \ldots, u^0\right) - u^*\left(u^{-q}, \ldots, u^{-1}\right)\right) \cdot v^*\left(u^{-q}, \ldots, u^0\right). \end{array}$$

v^* is called a discrete entropy variable.

Finally, we also set

$$v_j^{*n+1} = v^*\left(u_j^{n-q+1}, \ldots, u_j^{n+1}\right).$$

The validity of Assumption 2.1 will be discussed later on for specific examples.

We now turn to discuss the space discretization, based on a discrete flux

$$g^* : (v_{-p+1}, \dots, v_p) \in \mathbb{R}^{2pN} \mapsto g^*(v_{-p+1}, \dots, v_p) \in \mathbb{R}^N,$$

consistent with the continuous flux-function $g(v)$; i.e.,

$$g^*(v, \dots, v) = g(v), \quad v \in \mathbb{R}^N.$$

Observe that now we rely directly on the entropy variable v. Here the integer p indicates that the scheme uses $2p+1$ space-levels and is related to the order of accuracy in space: setting

$$g^{*n+1}_{j+1/2} = g^*\left(v^{*n+1}_{j-p+1}, \dots, v^{*n+1}_{j+p}\right),$$

we are led to a space discretization by replacing the continuous derivative $\partial_x g(v) = \partial_x f(u)$ in (1.1) with

$$\frac{g^{*n+1}_{j+1/2} - g^{*n+1}_{j-1/2}}{h}.$$

Our approach relies on entropy conservative discrete fluxes. Recall from [25] that a discrete flux g^* (expressed in the entropy variable v) is *entropy conservative* if there exists a *discrete entropy flux*

$$G^* : (v_{-p+1}, \dots, v_p) \in \mathbb{R}^{2pN} \mapsto G^*(v_{-p+1}, \dots, v_p) \in \mathbb{R}$$

consistent with the entropy flux $G(v)$ such that

$$\begin{aligned} v_0 \cdot \Big(g^*(v_{-p+1}, \dots, v_p) - g^*(v_{-p}, \dots, v_{p-1}) \Big) \\ = \quad G^*(v_{-p+1}, \dots, v_p) - G^*(v_{-p}, \dots, v_{p-1}). \end{aligned} \tag{2.6}$$

Finally, also set

$$G^{*n+1}_{j+1/2} = G^*\left(v^{*n+1}_{j-p+1}, \dots, v^{*n+1}_{j+p}\right).$$

The existence of such entropy conservative fluxes will be discussed below. First we state the central result of this section, providing a general approach to construct classes of fully discrete schemes.

THEOREM 2.2. *Consider a hyperbolic or hyperbolic-elliptic system of conservation laws* (1.1) *endowed with an entropy-entropy flux pair* (U, F) *satisfying condition* (2.3). *Consider a discrete conservative variable* u^* *and a discrete entropy function* U^* *such that Assumption* 2.1 *holds. For* $n \in \mathbb{N}$ *fixed, let the sequence* $\{u^n_j\}_{j\in\mathbb{Z}}$ *in* $\mathbb{R}^N$ *be given.*

Then, for any entropy conservative discrete flux g^* *and* $0 < \lambda << 1$, *the* $(q+1) \times (2p+1)$*-point difference equation*

$$u^{*n+1}_j - u^{*n}_j + \lambda \left(g^{*n+1}_{j+1/2} - g^{*n+1}_{j-1/2} \right) = 0 \qquad (j \in \mathbb{Z}) \tag{2.7}$$

has a unique solution $u^{n+1}_j \in \mathbb{R}^N$.

The associated scheme is entropy conservative with respect to U^* *in the sense that*

$$U^{*n+1}_j - U^{*n}_j + \lambda \left(G^{*n+1}_{j+1/2} - G^{*n+1}_{j-1/2} \right) = 0 \qquad (n \in \mathbb{N}, j \in \mathbb{Z}). \tag{2.8}$$

Proof. The result follows from the discussion preceding the theorem. Indeed, in view of (2.4) and (2.5) and by applying the inverse function theorem, there exists $0 < \lambda << 1$ for which (2.7) determines a unique solution u_j^{n+1}. Next, multiplying (2.7) by v_j^{*n+1}, we obtain

$$\left(u_j^{*n+1} - u_j^{*n}\right) \cdot v_j^{*n+1} = \lambda \left(g_{j+1/2}^{*n+1} - g_{j-1/2}^{*n+1}\right) \cdot v_j^{*n+1}.$$

The conservative form (2.8) is a direct consequence of the definitions of discrete entropy variable (2.5) and entropy conservative discrete flux (2.6). ∎

It is the main goal of the following sections to show precisely that the framework in Theorem 2.2 covers a variety of situations of practical interest.

NOTE 2.3.

(1) *The key points for using Theorem* 2.2 *are an appropriate choice of the functions* u^*, U^* *such that Assumption* 2.1 *holds and entropy conservative discrete fluxes* g^* *exist. The choice of the functions* u^*, U^* *will be discussed in section* 3. *A second order entropy conservative discrete flux has been designed in* [25]. *Third order entropy conservative fluxes have been derived in* [16]. *In section* 4 *below, we will return to constructing even higher order entropy conservative fluxes.*

(2) *The entropy equality* (2.8) *implies the following nonlinear stability property:*

$$\sum_{j=-\infty}^{\infty} U_j^{*n} = \text{const.}$$

Depending on the properties of the discrete entropy U^*, *this may provide us with some a priori bound on the discrete solution. For instance, if* U^* *is strictly convex, essentially we recover the* L^2*-stability of the scheme. In general,* (2.7) *is a fully nonlinear implicit scheme. As we have an implicit scheme we may expect to have stability for large CFL-numbers. However, convergence of an iterative method for solving the nonlinear system might enforce a stricter CFL-like condition.*

(3) *In general, the schemes* (2.7) *are fully implicit. However, in some situations of interest we obtain linear implicit or even explicit schemes (section* 3*).*

3. Two and three time-level entropy conservative schemes. In this section we give first applications of Theorem 2.2. We start investigating the simplest case of a two time-level discretization. We will see that such schemes are always fully nonlinear, except in the case of linear systems of conservation laws. Next we investigate three time-level schemes, for which there exists more freedom in choosing a convenient discretization of the entropy. We use this freedom to construct explicit or linear implicit schemes of third order.

We will rely on the consistent entropy conservative numerical flux-function g_2^* that has been constructed by Tadmor [25]; i.e.,

$$g_2^*(v_0, v_1) = \int_0^1 g(v_0 + s\,(v_1 - v_0))\, ds, \qquad v_0, v_1 \in \mathbb{R}^N. \tag{3.1}$$

The associated numerical entropy flux reads as

$$\begin{aligned} G_2^*(v_0, v_1) &= \frac{G(v_0) + G(v_1)}{2} + \frac{(v_0 + v_1)}{2} \cdot g_2^*(v_0, v_1) \\ &\quad - \frac{1}{2}\left(v_0 \cdot g(v_0) + v_1 \cdot g(v_1)\right). \end{aligned} \tag{3.2}$$

3.1. A class of two time-level entropy conservative schemes. We first consider schemes based on two time-levels only and on two-point discrete fluxes. Consider the following discretization

$$u_j^{n+1} = u_j^n - \lambda \left(g_{j+1/2}^{*n+1} - g_{j-1/2}^{*n+1} \right) \tag{3.3}$$

corresponding to the simple choice $q = 1$:

$$u^* \left(u^0 \right) = u^0.$$

For schemes with $q = 1$ the only consistent entropy is $U^*(u) = U(u)$. The only two-point entropy conservative flux is the one proposed by Tadmor. We get the following result from Theorem 2.2.

THEOREM 3.1. *Let $u^*(u) = u$, $U^*(u) = U(u)$. Let Assumption* 2.1 *be valid. Then the scheme* (3.3) *considered with Tadmor flux* (3.1) *is entropy conservative with respect to the entropy U^*. Furthermore, this scheme is second order accurate in space and time in the sense that its equivalent equation is*

$$\partial_t u(x_j, t^{n+1/2}) + \partial_x g(v(x_j, t^{n+1/2})) = \mathcal{O}(h^2).$$

To satisfy Assumption 2.1 we can choose v^* to be

$$v^*(u^0, u^1) = \int_0^1 v(su^*(u^1) + (1-s)u^*(u^0))\, ds. \tag{3.4}$$

Note that—at least in the linear case and with $U(u) = u^2/2$—the time discretization in (3.3) is exactly the Crank–Nicholson time discretization.

In general, (3.3) with (3.4) is fully nonlinear in u_j^{n+1}. To obtain an at least linear implicit scheme, g_2^* has to be linear, and $v^* = v^*(u^0, u^1)$ has to be linear with respect to u^1. The latter is true if and only if U is quadratic. By definition, the Tadmor flux g_2^* is linear if and only if g is linear. With U to be quadratic we obtain that the flux f has to be linear. In the next section we will provide explicit and linear implicit entropy conservative schemes.

Example 3.2. For the sake of illustration of the scheme (3.3) we present a numerical experiment. We consider the scalar case

$$f(u) = U(u) = \frac{u^2}{2}.$$

This leads to the scheme

$$\begin{aligned} u_j^{n+1} = u_j^n - \frac{\lambda}{24} \Big(& (u_j^n + u_j^{n+1})(u_{j+1}^n + u_{j+1}^{n+1} - u_{j-1}^n - u_{j-1}^{n+1}) \\ & + (u_{j+1}^n + u_{j+1}^{n+1})^2 - (u_{j-1}^n + u_{j-1}^{n+1})^2 \Big). \end{aligned} \tag{3.5}$$

For each time step, the nonlinear difference equation (3.5) is solved by a fixed-point iteration method which is stopped if the L^1-relative difference between two succeeding approximate solutions is less than a threshold. This fixed-point iteration approach will be used throughout this paper for all numerical experiments. Results for initial data $u_0(x) = \sin(2\pi x) + 1$ at different times are shown in the left picture of Figure 3.1. The computational domain is $[0, 1]$ with periodic boundary conditions. Here we chose 250 cells, and the CFL-number to be 0.25. As expected for a central scheme, the method leads to a highly oscillating wave pattern after formation of the shock wave, indicating that the method will not converge in any strong topology when refining the grid. We note that by adding artificial dissipation the oscillations can be suppressed (cf. section 5 for examples with nonclassical shocks).

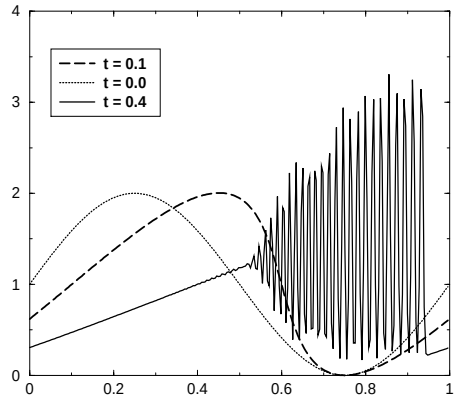


FIG. 3.1. *Numerical approximation of weak solutions of Burgers' equation with a two time-level entropy conservative scheme.*

3.2. A class of three time-level entropy conservative schemes. We consider three time-level schemes of the type

$$u_j^{*n+1} = u_j^{*n} - \lambda \left(g_{j+1/2}^{*n+1} - g_{j-1/2}^{*n+1} \right), \tag{3.6}$$

where the discrete conservative variable u^* is defined by

$$u^* \left(u^0, u^1 \right) = \alpha u^0 + (1-\alpha)\, u^1 \qquad (\alpha \in \mathbb{R}). \tag{3.7}$$

Straightforwardly we get the following theorem.

THEOREM 3.3. *Let $U^* \left(u^0, u^1 \right)$ be chosen such that Assumption* 2.1 *is satisfied for some entropy variable $v^* = v^* \left(u^{-1}, u^0, u^1 \right)$.*

Then the scheme (3.6) *considered with Tadmor flux* (3.1) *is a three time-level entropy conservative scheme with respect to the entropy $U^* \left(u^0, u^1 \right)$.*

To satisfy Assumption 2.1 we can always choose U^* and v^* as

$$\begin{aligned} U^*(u^0, u^1) &= U(u^*(u^0, u^1)), \\ v^*(u^{-1}, u^0, u^1) &= \int_0^1 v(su^*(u^0, u^1) + (1-s)u^*(u^{-1}, u^0))\, ds. \end{aligned} \tag{3.8}$$

If the entropy U is nonnegative, another possible choice for the discrete entropy is $U^* \left(u^0, u^1 \right) = \sqrt{U\left(u^1\right) U\left(u^0\right)}$, together with the entropy variable v^* as above.

3.3. Explicit three time-level schemes for quadratic entropies. Symmetric systems yield a general class of hyperbolic systems. For these systems, we can design three time-level explicit entropy conservative schemes. Let B be any constant positive symmetric matrix. For symmetric systems, the function

$$U(u) = u \cdot B\, u$$

is a strictly convex entropy for which the entropy variable is $v\left(u\right) = Bu$.

Let u^* be given by (3.7) for the special choice $\alpha = 1/2$, and choose the discrete entropy function

$$U^* \left(u^0, u^1 \right) = \frac{1}{2}\, u^0 \cdot B\, u^1.$$

To satisfy Assumption 2.1 define

$$v^*\left(u^{-1},u^0,u^1\right)=B\,u^0. \tag{3.9}$$

The Tadmor flux gives an *explicit* scheme.

PROPOSITION 3.4. *Suppose that Df is symmetric. Choose $U^*\left(u^0,u^1\right)=\frac{1}{2}\,u^0\cdot B\,u^1$ and Tadmor's flux* (3.1). *With v^* from* (3.9) *the scheme* (3.6) *is an explicit scheme, entropy conservative with respect to the entropy U^*.*

3.4. Linear implicit three time-level schemes. As pointed out in section 3.2, the three-point conservative scheme (3.6) allows different choices for the entropy U^*. Here we consider scalar conservation laws and highlight a choice of U^* that leads to a *linear implicit* scheme.

Consider the case $N=1$ with the flux $f(u)=u^3$ and entropy

$$U(u)=\int_0^u f(s)\,ds=\frac{u^4}{4}.$$

The flux written in the entropy variable is $g(v)=v$. Consider the discrete entropy

$$U_j^{*n}=U^{*n}\left(u_j^n,u_j^{n-1}\right)=\frac{1}{4}\left(u_j^n u_j^{n-1}\right)^2.$$

Assumption 2.1 is satisfied if the discrete entropy variable v^* is defined to be

$$v_j^{*n+1}=v^*(u_j^{n-1},u_j^n,u_j^{n+1})=\frac{1}{2}(u_j^n)^2(u_j^{n+1}+u_j^{n-1}).$$

For the flux we take

$$g_{j+1/2}^{*,n+1}=g_2^*\left(v_j^{*n+1},v_j^{*n+1}\right)=\frac{1}{4}\Big((u_j^n)^2(u_j^{n-1}+u_j^{n+1})+(u_{j+1}^n)^2(u_{j+1}^{n-1}+u_{j+1}^{n+1})\Big).$$

The resulting three time-level scheme (3.6) is linear implicit:

$$u_j^{n+1}=u_j^{n-1}-\frac{\lambda}{2}\left(\left(u_{j+1}^{n+1}+u_{j+1}^{n-1}\right)\left(u_{j+1}^n\right)^2-\left(u_{j-1}^{n+1}+u_{j-1}^{n-1}\right)\left(u_{j-1}^n\right)^2\right). \tag{3.10}$$

Example 3.5. We present a numerical experiment for scheme (3.10). Consider the cubic scalar conservation law for $u_0(x)=\sin(2x/\pi)$ on $[0,1]$ with periodic boundaries. The results for 250 cells and the CFL-number 0.25 are displayed in the right picture of Figure 3.2. Again we stress the fact that these schemes produce extreme oscillations after the shock has formed. When supplementing regularizing terms this effect will disappear.

3.5. Third order, three time-level entropy conservative schemes. Consider the following choice for the discrete entropy variable:

$$u^*\left(u^0,u^1\right)=\left(\frac{1}{2}-\frac{1}{\sqrt{2}}\right)u^0+\left(\frac{1}{2}+\frac{1}{\sqrt{2}}\right)u^1. \tag{3.11}$$

THEOREM 3.6. *Consider a hyperbolic or hyperbolic-elliptic system of conservation laws* (1.1) *endowed with an entropy-entropy flux pair (U,F) satisfying condition* (2.3). *Consider the discrete conservative variable u^* from* (3.11), *$U^*\left(u^0,u^1\right)=U\left(u^*\left(u^0,u^1\right)\right)$, and v^* to be*

$$v^*\left(u^{-1},u^0,u^1\right)=\int_0^1 v\left(su^*\left(u^0,u^1\right)+(1-s)\,u^*\left(u^{-1},u^0\right)\right)ds.$$

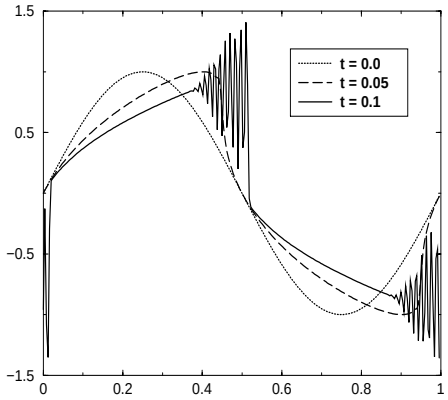


FIG. 3.2. *Numerical approximation of the cubic scalar conservation law with a three time-level entropy conservative scheme.*

For an entropy conservative flux g^* of order $2p$, $p \in \mathbb{N}$, the scheme

$$u_j^{*n+1} - u_j^{*n} + \lambda \left(g_{j+1/2}^{*n+1} - g_{j-1/2}^{*n+1} \right) = 0$$

has a unique solution u_j^{n+1} if and only if λ is small enough. The scheme is entropy conservative and third order accurate in time; i.e., its equivalent equation is

$$\partial_t u \left(x_j, t^n + \frac{\tau}{\sqrt{2}} \right) + \partial_x g \left(v \left(x_j, t^n + \frac{\tau}{\sqrt{2}} \right) \right) = O\left(\tau^3\right) + O\left(h^{2p}\right).$$

The order of accuracy of these scheme can be checked easily. See also section 4.2 for a constructive demonstration.

NOTE 3.7. *Using the Tadmor flux leads to a second order in space accurate conservative scheme ($p = 1$ in the previous formula). The next section provides the explicit construction of conservative fluxes of arbitrary higher order. A numerical example of this section is considered in section* 5.

4. Entropy conservative schemes of arbitrary order.

4.1. Semidiscrete entropy conservative schemes of arbitrary order. Consider the conservation law (1.1) with entropy-entropy flux pair (U, F). Let $x_j = jh$, $j \in \mathbb{Z}$, be a regular mesh, h denoting the grid point distance. For $v_j = \nabla U(u_j)$, we consider $(2p+1)$-point semidiscrete schemes of type

$$\begin{aligned} u_j'(t) &= -\frac{1}{h}\left(g_{2p,j+1/2}^* - g_{2p,j-1/2}^* \right) \\ &= -\frac{1}{h}\left(g_{2p}^*(v_{j-p+1}, \ldots, v_{j+p}) - g_{2p}^*(v_{j-p}, \ldots, v_{j+p-1}) \right), \end{aligned} \tag{4.1}$$

where $u_j(t)$ approximates the solution u of (1.1) in (x_j, t) and $'$ denotes time derivation. In this section we show that there exist smooth numerical fluxes $g_{2p}^* : \mathbb{R}^{2pN} \to \mathbb{R}^N$ satisfying the following conditions for all $j \in \mathbb{Z}$, $p \in \mathbb{N}$ and all smooth enough

functions $v = \nabla U(u)$ (denoting $v_j = v(x_j, t)$):

(i) $g^*(v_j, \dots, v_j) = g(v_j)$.

(ii) $$\frac{g^*_{2p}(v_{j-p+1}, \dots, v_{j+p}) - g^*_{2p}(v_{j-p}, \dots, v_{j+p-1})}{h} = \partial_x g(v_j) + \mathcal{O}(h^{2p}).$$

(iii) There is a function $G^*_{2p} : \mathbb{R}^{2pN} \to \mathbb{R}$ consistent with G such that

$$U(u_j(t))' = -\frac{1}{h}\Big(G^*_{2p}(v_{j-p+1}, \dots, v_{j+p}) - G^*_{2p}(v_{j-p}, \dots, v_{j+p-1})\Big).$$

In other words, we will show that there exist consistent semidiscrete entropy conservative schemes (4.1) of arbitrary order. So far, only fluxes of order two [23] or three [16] have been available.

For $\alpha_{1,p}, \dots, \alpha_{p,p} \in \mathbb{R}$, we make an ansatz for g^*_{2p} as a linear combination of Tadmor's flux g^*_2 (cf. (3.1)):

$$g^*_{2p}(v_{-p+1}, \dots, v_p) = \sum_{i=1}^{p} \alpha_{i,p}\Big(g^*_2(v_0, v_i) + \cdots + g^*_2(v_{-i+1}, v_1)\Big). \tag{4.2}$$

So the flux difference is given by

$$g^*_{2p}(v_{-p+1}, \dots, v_p) - g^*_{2p}(v_{-p}, \dots, v_{p-1}) = \sum_{i=1}^{p} \alpha_{i,p}\Big(g^*_2(v_0, v_i) - g^*_2(v_{-i}, v_0)\Big). \tag{4.3}$$

NOTE 4.1 (linear entropy flux). *Assume that the function g can be written as an affine function, say $g(v) = Av + b$, $A \in \mathbb{R}^{N\times N}, b \in \mathbb{R}^N$ (cf. sections* 3.4, 3.5*). Then the Tadmor flux difference is simply the centered difference $A(v_1 - v_{-1})/2$, and we get in our case*

$$g^*_{2p}(v_{-p+1}, \dots, v_p) - g^*_{2p}(v_{-p}, \dots, v_{p-1}) = A\sum_{i=1}^{p} \alpha_{i,p}(v_i - v_{-i}). \tag{4.4}$$

We show first that the general ansatz (4.2) leads to a scheme satisfying (i), (iii).

PROPOSITION 4.2. *Let $p \in \mathbb{N}$. Consider the scheme* (4.1) *for g^*_{2p} from* (4.2) *and $\alpha_{1,p}, \dots, \alpha_{p,p} \in \mathbb{R}$ satisfying*

$$2\sum_{i=1}^{p} i\alpha_{i,p} = 1. \tag{4.5}$$

Then (i) *and* (iii) *are satisfied for*

$$G^*_{2p} = G^*_{2p}(v_{-p+1}, \dots, v_p) = \sum_{i=1}^{p} \alpha_{i,p}\Big(G^*_2(v_0, v_i) + \cdots + G^*_2(v_{-i+1}, v_1)\Big), \tag{4.6}$$

where G^*_2 is given by (3.2).

Proof. Using (4.3) and Tadmor entropy fluxes $G^*_2(v_0, v_1)$, we get

$$\begin{aligned} & v_0 \cdot \Big(g^*_{2p}(v_{-p+1}, \dots, v_p) - g^*_{2p}(v_{-p}, \dots, v_{p-1})\Big) \\ &= v_0 \cdot \sum_{i=1}^{p} \alpha_{i,p}\Big(g^*_2(v_0, v_i) - g^*_2(v_{-i}, v_0)\Big) \\ &= \sum_{i=1}^{p} \alpha_{i,p}\Big(G^*_2(v_0, v_i) - G^*_2(v_{-i}, v_0)\Big). \end{aligned}$$

The last line equals $G^*_{2p}(v_{-p+1},\dots,v_p) - G^*_{2p}(v_{-p},\dots,v_{p-1})$, which proves (iii). The consistency of g^*_{2p}, G^*_{2p} with g, G follows from (4.5). ∎

Next, we fix the up-to-now free coefficients $\alpha_{1,p},\dots,\alpha_{p,p}$ to provide a high-order scheme.

PROPOSITION 4.3. *For $p \in \mathbb{N}$, assume that $\alpha_{1,p},\dots,\alpha_{p,p}$ solve the p linear equations*

$$2\sum_{i=1}^{p} i\alpha_{i,p} = 1, \quad \sum_{i=1}^{p} i^{2s-1}\alpha_{i,p} = 0 \quad (s = 2,\dots,p). \tag{4.7}$$

Then the flux g^*_{2p} given by formula (4.2) *satisfies* (ii)*; i.e., for smooth enough function v we have*

$$\frac{g^*_{2p}(v_{j-p+1},\dots,v_{j+p}) - g^*_{2p}(v_{j-p},\dots,v_{j+p-1})}{h} = \partial_x g(v_j) + \mathcal{O}(h^{2p}). \tag{4.8}$$

Here we used $C_{2p} = \sum_{i=1}^{p} \frac{\alpha_{i,p} i^{2p+1}}{(2p+1)!}$ and $v_j = v(x_j)$ for $j \in \mathbb{Z}$.

Proof. By Taylor expansion around x_0 we obtain for $i = 1,\dots,p$

$$g^*_2(v_0, v_i) - g^*_2(v_{-i}, v_0) = 2\sum_{k=0}^{p} \frac{(ih)^{2k+1}}{(2k+1)!}\partial_x^{(2k+1)} g(v_0) + \mathcal{O}(h^{2p+2}).$$

This leads by (4.3) to the expression

$$\begin{aligned} & g^*_{2p}(v_{-p+1},\dots,v_p) - g^*_{2p}(v_{-p},\dots,v_{p-1}) \\ &= 2\sum_{i=1}^{p} \alpha_{i,p} \left(\sum_{k=0}^{p} \frac{(ih)^{2k+1}}{(2k+1)!}\partial_x^{(2k+1)} g(v_0) \right) + \mathcal{O}(h^{2p+2}). \end{aligned}$$

The definition of $\alpha_{1,p},\dots,\alpha_{p,p}$ in (4.7) gives the statement of the proposition. ∎

Note that the first equation in (4.7) equals (4.5) and ensures consistency. We summarize Proposition 4.2 and 4.3 in the following theorem.

THEOREM 4.4. *Consider a hyperbolic or hyperbolic-elliptic system of conservation laws* (1.1) *with an entropy-entropy flux pair (U, F). Assume that $\alpha_{i,1},\dots,\alpha_{i,p}$ solve* (4.7).

Then the flux g^*_{2p} given by formula (4.2) *satisfies the conditions* (i), (ii), (iii).

The scheme (4.1) *is an entropy conservative semidiscrete scheme with respect to U which is of order $2p$.*

4.2. Fully discrete entropy conservative schemes of arbitrary order. In this section we present fully discrete schemes of arbitrary order verifying a weaker form of entropy conservation. For an integer $q \geq 1$, the schemes will use $q+1$ time-levels and be of order $q+1$ in time.

Let $j \in \mathbb{Z}$ and $n \in \mathbb{N}$. We approximate the continuous derivative $\partial_t u$ in (1.1) by

$$\frac{u_j^{*n} - u_j^{*n-1}}{\tau} := \sum_{i=0}^{q} \beta^t_{i,q} u_j^{n-q+i}. \tag{4.9}$$

In the formula above, $\beta^t_{0,q},\dots,\beta^t_{q,q} \in \mathbb{R}$ are parameters that have to be chosen according to the desired $(q+1)$st order of accuracy; i.e., for smooth enough function u, we

have the following expansion around a time $\bar{t}^q > 0$ to be determined:

$$\sum_{i=0}^{q} \beta_{i,q}^t u_j^{n-q+i} = \partial_t u\left(x_j, \bar{t}^n\right) + O\left(\tau^{q+1}\right). \tag{4.10}$$

Consider the $q+1$ linear equations

$$\sum_{i=0}^{q} (t^{n-q+i} - \bar{t}^{\,n})\beta_{i,q}^t = 1, \quad \sum_{i=0}^{q} (t^{n-q+i} - \bar{t}^{\,n})^s \beta_{i,q}^t = 0 \quad (s = 0, \dots, q,\ s \neq 1). \tag{4.11}$$

This last system is a Vandermonde system. If $\bar{t}^n$ does not belong to the set of time grid points $\{t^n\}_{n\geq 0}$, it also is nondegenerate. In this last case, the unique solutions $\beta_{0,q}^t, \dots, \beta_{q,q}^t \in \mathbb{R}$ of (4.11) provides via (4.9) an approximation of $\partial_t u(x_j, \bar{t}^n)$ with at least order q, as a straightforward Taylor expansion of $u(x_j, t^{n-q}), \dots, u(x_j, t^n)$ around $(x_j, \bar{t}^n)$ shows. Note that we are left with one degree of freedom, namely to choose $\bar{t}^n$. There exists a choice that allows us to gain one order of accuracy in time and obtain (4.10): choose $\bar{t}^n$ satisfying (4.11) and

$$\sum_{i=0}^{q} (t^{n-q+i} - \bar{t}^{\,n})^{s+1} \beta_{i,q}^t = 0. \tag{4.12}$$

To prove the existence of such an intermediate solution, we introduce the following polynomial

$$P(t) = \sum_{i=0}^{q} \frac{(-1)^i \prod_{\substack{0\leq l\leq q \\ l\neq i}} \left(t^{n-l} - t\right)^2}{\prod_{\substack{0\leq l\leq i \\ l\neq i}} |t^{n-i} - t^{n-l}|}.$$

One can check, using the explicit solution of the Vandermonde system (4.11), that any root of the previous polynomial provides a solution of (4.11), (4.12). It can be easily seen that this polynomial has q solutions, the ith solution ($i = 0, \dots, q-1$) lying in $\left[t^{n-i}, t^{n-i-1}\right]$. For stability reasons we always take the solution in $[t^{n-1}, t^n]$. For instance, in the case $q = 1$ we get $\bar{t}^n = \frac{t^{n-1}+t^n}{2}$, that is the Crank–Nicholson choice. For $q = 2$, we get $\bar{t}^n = t^n - \frac{\tau}{\sqrt{2}}$, that is, a third order scheme as considered in section 3.5. For $q > 2$, we compute numerically the solution that belongs to $[t^{n-1}, t^n]$.

In a similar way, define the coefficients $\beta_{0,q}^u, \dots, \beta_{q,q}^u \in \mathbb{R}$ to be such that the expansion

$$\sum_{i=0}^{q} \beta_{i,q}^u u_j^{n-q+i} = u\left(x_j, \bar{t}^q\right) + \mathcal{O}\left(\tau^{q+1}\right) \tag{4.13}$$

holds, that is, solving the equations

$$\sum_{i=0}^{q} \beta_{i,q}^u = 1, \quad \sum_{i=0}^{q} (t^{n-q+i} - \bar{t}^{\,n})^s \beta_{i,q}^u = 0 \quad (s = 1, \dots, q). \tag{4.14}$$

The entropy variable being $v(u)$, define the discrete entropy variable $v_{q+1}^{*n} : \mathbb{R}^{(q+1)N} \to \mathbb{R}^N$ to be

$$v_{q+1}^{*n}\left(u_j^{n-q}, \dots, u_j^n\right) = v\left(\sum_{i=0}^{q} \beta_{i,q}^u u_j^{n-q+i}\right), \tag{4.15}$$

and denote $v^{*n+1}_{q+1,j} := v^{*n+1}_{q+1}\left(u^{n-q}_j, \ldots, u^n_j\right)$. Now using the high order entropy conservative numerical fluxes constructed in the preceding section, we obtain arbitrarily high order fully discrete schemes, however, only satisfying a weaker form of entropy conservation.

THEOREM 4.5. *Consider a hyperbolic or hyperbolic-elliptic system of conservation laws* (1.1) *endowed with an entropy-entropy flux pair* (U, F) *satisfying condition* (2.3). *Let* u^{*n}_{q+1} *be a discrete conservative variable defined with* (4.9)–(4.10) *and a discrete entropy variable* v^*_{q+1} *satisfying* (4.13). *For a* $2p$*-point numerical flux* g^*_{2p} *from Theorem* 4.4, *consider the following* $(2p+1)\times(q+1)$*-point scheme:*

$$u^{*n+1}_j = u^{*n}_j - \lambda\left(g^{*n+1}_{j+1/2} - g^{*n+1}_{j-1/2}\right).$$

Then, for λ *small enough, there exists an unique solution* u^{n+1}_j. *The scheme is entropy conservative in the sense*

$$\left(u^{*n+1}_j - u^{*n}_j\right) v^{*n+1}_{q+1,j} + \lambda\left(G^{*n+1}_{j+1/2} - G^{*n+1}_{j-1/2}\right) = 0. \tag{4.16}$$

Furthermore, it is of order $(q+1)$ *in time and* $2p$ *in space in the sense that its equivalent equation is*

$$\partial_t u\left(x_j, \bar{t}^n\right) + \partial_x g\left(v\left(u\left(x_j, \bar{t}^n\right)\right)\right) = \mathcal{O}\left(h^{2p}\right) + \mathcal{O}\left(\tau^{q+1}\right).$$

Proof. The weak entropy conservation (4.16) follows from multiplying the scheme difference equation by $v^{*n+1}_{q+1}\left(u^{n-q}_j, \ldots, u^n_j\right)$ and using property (2.6) for g^*_{2p}.

The equivalent equation comes from (4.13) and Theorem 4.4. ∎

NOTE 4.6. *For* $q = 1$ *(Crank–Nicholson choice) and* $q = 2$, *the discrete entropy variable constructed above, i.e., satisfying* (4.13), *also verifies Assumption* 2.1 *for* U. *It follows that these schemes are entropy conservative in the sense*

$$U^{*n+1}_j - U^{*n}_j + \lambda\left(G^{*n+1}_{j+1/2} - G^{*n+1}_{j-1/2}\right) = 0 \qquad (n \in \mathbb{N}, j \in \mathbb{Z})$$

with $U^{*n+1}_j = U\left(u^{*n+1}_j\right)$.

We illustrate this section with a fully discrete, fourth order accurate entropy scheme for the system of nonlinear elasticity.

For a stress-strain function $w \mapsto \sigma(w)$, consider the system

$$\partial_t w - \partial_x V = 0, \qquad \partial_t V - \partial_x \sigma(w) = 0. \tag{4.17}$$

Here V is the particle velocity and w is the stress, collected in $u := (w, V)$. The mathematical entropy pair is

$$(U(u), F(u)) = \left(\int_0^w \sigma(s)\, ds + \frac{V^2}{2},\ \sigma(w)V\right).$$

We choose the stress-strain function σ given by

$$\sigma(w) = w^3 - w.$$

Then (4.17) represents a model for phase transitions in shape memory alloys. Note that, for $w \in [-1/\sqrt{3}, 1/\sqrt{3}]$, the problem is elliptic, and hyperbolic outside this interval. The flux in (4.17) can be written in terms of the entropy variable $v =$

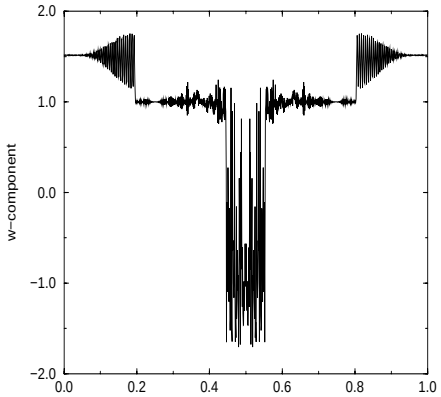


FIG. 4.1. *A fourth order in time, fourth order in space conservative scheme for the p-system (w-component).*

$(v_1, v_2)^T = (\sigma, V)^T$ and—as in the scalar case of section 3.4—is a linear function: $g(v_1, v_2) = -(v_2, v_1)^T$.

To discretize this system we design a four time-level scheme using the construction given in Theorem 3.6. We compute the values of the parameters $\beta^{t/u}_{0,q}, \dots, \beta^{t/u}_{q,q}$ and $\overline{t}^n$ as described above. Define

$$v_j^{*n+1} = \left(v_{1,j}^{*n+1}, v_{2,j}^{*n+1}\right)^T = \left(\sum_{i=0}^{3} \beta^u_{i,3} V_j^{n-3+i}, \sigma\left(\sum_{i=0}^{3} \beta^u_{i,3} w_j^{n-3+i}\right)\right)^T.$$

Consider now the fourth order conservative flux (cf. (4.2))

$$g_{j+1/2}^{*n+1} = g\left(\frac{2}{3}\left(v_j^{*n+1} + v_{j+1}^{*n+1}\right) - \frac{1}{12}\left(v_{j-1}^{*n+1} + \cdots + v_{j+2}^{*n+1}\right)\right).$$

The resulting scheme is, denoting componentwise $g_{j+1/2}^{*n+1} = (g_{1,j+1/2}^{*n+1}, g_{2,j+1/2}^{*n+1})^T$,

$$\left\{\begin{array}{rcl} w_j^{*n+1} - w_j^{*n} + \lambda\left(g_{1,j+1/2}^{*n+1} - g_{1,j-1/2}^{*n+1}\right) &=& 0, \\ V_j^{*n+1} - V_j^{*n} + \lambda\left(g_{2,j+1/2}^{*n+1} - g_{2,j-1/2}^{*n+1}\right) &=& 0 \end{array}\right. \qquad (j \in \mathbb{Z}). \tag{4.18}$$

Such a scheme is a fully nonlinear fourth order scheme. The numerical experiment takes place in the interval $[0, 5]$ with periodic boundaries. Choose initial data

$$u_0(x) = \left\{\begin{array}{l} (1,-1)^T : x \in [0, 2.5), \\ (1,1)^T \ : x \in [2.5, 5). \end{array}\right. \tag{4.19}$$

For such Riemann initial data, an intermediate middle state lying in the opposite phase, i.e., $\{w \in \mathbb{R} \,|\, w \le -1/\sqrt{3}\}$, must evolve for positive time [18]. The results for 1000 cells and the CFL-number 0.25 at time 0.1 are displayed in Figure 4.1.

5. Computation of regularization-sensitive weak solutions.

5.1. Analytical background and the basic numerical scheme. In the physical context the conservation law (1.1) is embedded into a higher order regularized but singularly perturbed model. For a small perturbation parameter $\varepsilon > 0$ and $D_2, D_3 : \mathbb{R}^N \to \mathbb{R}^{N\times N}$, let us consider systems of equations involving spatial derivatives up to order three:

$$\partial_t u^\varepsilon + \partial_x f(u^\varepsilon) = \varepsilon \partial_x \Big(D_2(u^\varepsilon)\partial_x u^\varepsilon \Big) + \varepsilon^2 \partial_x \Big(D_3(u^\varepsilon)\partial_{xx} u^\varepsilon \Big). \tag{5.1}$$

We are interested in weak solutions u of (1.1) that arise as limits of a sequence of smooth solutions $\{u^\varepsilon\}_{\varepsilon>0}$ of (5.1) for vanishing regularization parameter ε. While the second order derivatives in (5.1) correspond to physical effects like fluid viscosity or heat conduction, the third order term models capillarity phenomena [11, 15, 26].

A very interesting property of these *viscosity-capillarity* approximations u^ε is the fact that the limit solution u can contain undercompressive regularization-sensitive shock waves. Changing D_2, D_3 can produce a different weak solution; in other words, the limit function depends crucially on the entropy dissipation.

The numerical approximation of such weak solutions is a big challenge since also for the discrete counterpart the numerical entropy dissipation has to be tuned exactly. To overcome these difficulties Hayes and Lefloch suggested using *entropy conservative* numerical fluxes as a building block for finite difference schemes. To approximate the weak solution $u = \lim_{\varepsilon\to 0} u^\varepsilon$ of (1.1) they consider the following class of schemes (written down in the semidiscrete version, for simplicity):

$$\begin{aligned} u_j'(t) &= -\frac{1}{h}\left(\tilde g^*_{2p,j+1/2} - \tilde g^*_{2p,j+1/2} \right), \\ \tilde g^*_{2p,j+1/2} &:= g^*_{2p,j+1/2} - f^{2*}_{j+1/2} - f^{3*}_{j+1/2}. \end{aligned} \tag{5.2}$$

Here g^*_{2p} is the smooth entropy conservative numerical flux from (4.2), and

$$f^{2/3*}_{j+1/2} = f^{2/3*}(u_{j-r+1}, \ldots, u_{j+r}) \quad (r \in \mathbb{N}),$$

where $f^{2/3*} : \mathbb{R}^{2rN} \to \mathbb{R}^N$ are smooth and satisfy for all smooth enough functions u (denoting $u_j = u(x_j, t)$)

$$\begin{aligned} \frac{f^{2*}(u_{j-r+1}, \ldots, u_{j+r}) - f^{2*}(u_{j-r}, \ldots, u_{j+r-1})}{h} &= h\partial_x \Big(D_2(u_j)\partial_x u_j \Big) + \mathcal{O}(h^3), \\ \frac{f^{3*}(u_{j-r+1}, \ldots, u_{j+r}) - f^{3*}(u_{j-r}, \ldots, u_{j+r-1})}{h} &= h^2\partial_x \Big(D_3(u_j)\partial_{xx} u_j \Big) + \mathcal{O}(h^3). \end{aligned}$$

Then we obtain the following equivalent equation for the scheme (5.2):

$$\begin{aligned} &u_j'(t) + \partial_x f(u_j(t)) \\ &\quad = h\partial_x \Big(D_2(u_j(t))\partial_x u_j(t) \Big) + h^2 \partial_x \Big(D_3(u_j(t))\partial_{xx} u_j(t) \Big) + \mathcal{O}(h^{2p}) + \mathcal{O}(h^3). \end{aligned} \tag{5.3}$$

We observe that the equivalent equation mimics (5.1) provided we have $p \geq 2$. This is precisely the motivation for considering (5.2) with *high order* fluxes. While in

[9, 16] only semidiscrete entropy conservative schemes were available, here we have constructed fully discrete high order entropy conservative schemes.

In what follows we will consider numerical experiments. Furthermore, in a special case this construction allows us to consider a discrete counterpart for the entropy inequality.

5.2. Regularizations that are linear in the entropy variable. In this section we consider in this section a regularization mechanism of (1.1) in which the dissipative terms are *linear* functions of the entropy variable v:

$$\partial_t u^\varepsilon + \partial_x f(u^\varepsilon) = \varepsilon B_2 \partial_{xx} v + \varepsilon^2 B_3 \partial_{xxx} v^\varepsilon. \tag{5.4}$$

Here we assume that

$$B_2, B_3 \text{ are } (N \times N) \text{ constant matrices,} \tag{5.5}$$

and we make the hypothesis

$$B_2 \text{ is positive definite and } B_3 \text{ is symmetric.} \tag{5.6}$$

The advantage of this particular choice is the following. Multiplying (5.4) by v^ε and performing integration by parts, the hypothesis (5.6) leads immediately to the entropy stability estimate

$$\frac{d}{dt} \int_{\mathbb{R}} U\left(u^\varepsilon(x)\right) dx \leq 0. \tag{5.7}$$

In what follows we assume that there is a classical solution of the Cauchy problem for (5.4) and a weak solution u of the Cauchy problem for (1.1) such that $\lim_{\varepsilon \to 0} u^\varepsilon = u$. As we are interested in the numerical approximation of the function u we consider on the (semi)discrete level the scheme (5.2) together with smooth fluxes $f^{2/3*}_{j+1/2} : \mathbb{R}^{2rN} \to \mathbb{R}^N$, $r \in \mathbb{N}$, that are *linear* in v and satisfy for all smooth enough functions u,

$$\begin{aligned} \frac{f^{2*}(u_{j-r+1}, \ldots, u_{j+r}) - f^{2*}(u_{j-r}, \ldots, u_{j+r-1})}{h} &= hB_2 \partial_x^2 v_j + \mathcal{O}(h^3), \\ \frac{f^{3*}(u_{j-r+1}, \ldots, u_{j+r}) - f^{3*}(u_{j-r}, \ldots, u_{j+r-1})}{h} &= h^2 B_3 \partial_{xxx} v_j + \mathcal{O}(h^3). \end{aligned}$$

NOTE 5.1. *Since the estimate* (5.7) *is the immediate and natural a priori bound for* u^ε*, it should be also possible to prove a discrete entropy dissipation property for the approximate solution. For a result in a particular case we refer to* [16, Theorem 5.1] *and* [6].

Independent of these analytical issues, the numerical experiments in section 6.2 *below clearly demonstrate the benefits of the linear regularization.*

In the rest of this subsection we will focus on a somewhat different but strongly related issue: We consider special high order discretizations for smooth solutions of (5.4) (and not for weak solutions of (1.1) that arise as vanishing dissipation limits of (5.4)).

We introduce a discrete version of (5.7): a form $\tilde{g}^*(v_{-p+1}, \ldots, v_p)$ is entropy dissipative if, for any compactly supported sequence $(v_j)_{j \in \mathbb{Z}}$ in $\mathbb{R}^N$,

$$\sum_{j \in \mathbb{Z}} v_j \cdot \left(g^*(v_{-p+j+1}, \ldots, v_{p+j}) - g^*(v_{-p+j}, \ldots, v_{p+j-1})\right) \leq 0. \tag{5.8}$$

Note that a conservative form $\tilde{g}^* = \tilde{g}^*(v_{-p+j+1}, \ldots, v_{p+j})$ (i.e., a form satisfying (2.6)) verifies (5.8) as an equality.

To provide a $2p$ order discretization of the capillarity term $\partial_x^3 v$, let us introduce the coefficients $\alpha_{i,1}^{(3)}, \ldots, \alpha_{i,p}^{(3)}$ as the solutions of the p linear equations:

$$2\sum_{i=1}^{p} i^3 \alpha_{i,p} = 1, \quad \sum_{i=1}^{p} i^{2s-1} \alpha_{i,p}^{(3)} = 0 \quad (s = 1, \ldots, p, s \neq 2). \tag{5.9}$$

As for (4.7), the previous system is a Vandermonde system and thus has an unique solution. Let us introduce the form $v_{2p}^{(3)*}$, defined by

$$v_{2p}^{(3)*}(v_{-p+1}, \ldots, v_p) = \sum_{i=1}^{p} \alpha_{i,p}^{(3)} (v_i + v_{i-1} + \cdots + v_{-i+1}), \tag{5.10}$$

Here v_i stands for $v(x_i)$, v being any smooth enough vector-valued function $v(x) \in \mathbb{R}^N$. As for (4.7), the difference

$$v_{2p}^{(3)*}(v_{-p+1}, \ldots, v_p) - v_{2p}^{(3)*}(v_{-p}, \ldots, v_{p-1}) = \sum_{i=1}^{p} \alpha_{i,p}^{(3)} (v_i - v_{-i}) \tag{5.11}$$

provides a formula of order $2p$ for $\partial_x^3 v_0$. This is straightforward from Taylor expansions of order $2p$ around v_0. Also note that such a form is conservative in the sense of (2.6), because the structure exhibited in (5.11) corresponds to the special form exhibited in (4.4).

Now we turn to a $2p$ order discretization of the viscous term $\partial_x^2 v$. Let us introduce the coefficients $\alpha_{i,1}^{(2)}, \ldots, \alpha_{i,p}^{(2)}$ as the solutions of the p linear equations

$$\sum_{i=1}^{p} \alpha_{i,p}^{(2)} = 1, \quad \sum_{i=1}^{p} i^{2s} \alpha_{i,p}^{(2)} = 0 \quad (s = 1, \ldots, p-1). \tag{5.12}$$

We also introduce the form $v_{2p}^{(2)}$ defined by

$$v_{2p}^{(2)*}(v_{-p+1}, \ldots, v_p) = \sum_{i=1}^{p} \alpha_{i,p}^{(2)} (v_i + \cdots + v_1 - v_0 - \cdots - v_{-i+1}). \tag{5.13}$$

Straightforwardly from Taylor expansions around v_0, the difference

$$v_{2p}^{(2)*}(v_{-p+1}, \ldots, v_p) - v_{2p}^{(2)*}(v_{-p}, \ldots, v_{p-1}) = \sum_{i=1}^{p} \alpha_{i,p}^{(2)} (v_i + v_{-i} - 2v_0) \tag{5.14}$$

provides a $2p$ order discretization of $\partial_x^2 v_0$.

To provide a discretization for the whole equation (5.4), denote

$$\widetilde{g}_{2p}^* = g_{2p}^* - v_{2p}^{(2)*} - v_{2p}^{(3)*}, \tag{5.15}$$

where g_{2p}^* is defined in the previous section (see formula (4.2)). Set $\widetilde{g}_{2p,j+1/2}^{*n+1} = \widetilde{g}_{2p}^* \left(v_{j-p+1}^{*n+1}, \ldots, v_{j+p}^{*n+1}\right)$. The main theorem of this section follows.

THEOREM 5.2. *Consider the system of conservation laws* (5.4) *together with an entropy pair* (U, F) *and the compatibility conditions* (5.6). *Let* $p > 1$ *and consider the semidiscrete scheme*

$$u_j'(t) = -\frac{1}{h}\Big(\widetilde{g}^*_{2p,j+1/2} - \widetilde{g}^*_{2p,j-1/2}\Big), \qquad t > 0.$$

The equivalent equation of this scheme is the system (5.4) *evaluated in* (x_j, t) *up to a term of order* $2p$ *in space.*

Assume that $u_j(t)$ *vanishes for* $|j|$ *big enough for all* $t > 0$. *Then the scheme is entropy decreasing:*

$$\sum_{j\in\mathbb{Z}} U'(u_j(t)) \le 0, \qquad t > 0. \tag{5.16}$$

NOTE 5.3. *We could also have stated a fully discrete version of the previous theorem using the time discretization exhibited in Theorem* 4.5*: let* $q + 1$ *as defined in Theorem* 4.5 *be the number of time-levels used by the scheme. Then we are able to construct a fully discrete scheme of order* $q + 1$ *in time,* $2p$ *in space with respect to* (5.4) *It satisfies the entropy dissipation property*

$$\sum_{j\in\mathbb{Z}} \left(u_j^{*n+1} - u_j^{*n+1}\right) v_j^{*n+1} \le 0.$$

We notice also that, using an entropy variable satisfying Assumption 2.1, *we are led to a scheme verifying the strongest entropy dissipation property*

$$\sum_{j\in\mathbb{Z}} U\left(u_j^{*n+1}\right) \le 0. \tag{5.17}$$

In particular, consider the third order accurate conservative scheme described in section 3.5. *Following the guidelines described above, we are able to construct fully discrete schemes of accuracy order* 3 *in time,* $2p$ *in space with respect to* (5.4), *satisfying the entropy dissipation property* (5.17).

Proof of Theorem 5.2. It is enough to prove the dissipation property.

$$\sum_{i\in\mathbb{Z}} \Big(\widetilde{g}^*_{2p,j+1/2} - \widetilde{g}^*_{2p,j-1/2}\Big) v_j \le 0.$$

Since g^*_{2p} and $v^{(3)*}_{2p}$ are entropy conservative fluxes (i.e., they satisfy a stronger version of (5.8)), the only point is to show the statement for $v^{(2)*}_{2p}$.

Note that the elementary forms $(v_{-i} + v_i - 2v_0)$ are the building block of (5.14). We compute

$$(v_{-i} + v_i - 2v_0)\, v_0 = -\left(v_i - v_0\right)^2 + v_i^2 - v_i v_0 - \left(v_0^2 - v_0 v_{-i}\right).$$

Denoting $G_2^{(2)*}(v_0, v_i) = v_i^2 - v_i v_0$ and $G^{(2)*}_{2p,j+1/2} = \sum_{l=0}^{j-1} G_2^{(2)*}(v_{-l}, v_{j-l})$, we have

$$\begin{aligned}
(v_{-i} + v_i - 2v_0)\, v_0 &= -\left(v_i - v_0\right)^2 + G_2^{(2)*}(v_0, v_i) - G_2^{(2)*}(v_{-i}, v_0) \\
&= -\left(v_i - v_0\right)^2 + \sum_{l=0}^{i-1} G_2^{(2)*}(v_{-l}, v_{i-l}) - \sum_{l=0}^{i-1} G_2^{(2)*}(v_{-l-1}, v_{i-l-1}) \\
&= -\left(v_i - v_0\right)^2 + \Big(G^{(2)*}_{2p,i+1/2} - G^{(2)*}_{2p,i-1/2}\Big).
\end{aligned}$$

This proves that

$$\sum_{j\in\mathbb{Z}}\Big(v_j\Big(v^{(2)}_{2p,j+1/2}-v^{(2)}_{2p,j-1/2}\Big)\Big)+\sum_{i=1}^{p}\alpha^{(2)}_{i,p}\sum_{j\in\mathbb{Z}}\frac{\left(v_{i+j}-v_j\right)^2}{2}=0.$$

The last sum in the last equation can be estimated from below by a sum independent of i. Therefore $\sum_{i=1,..,p}\alpha^{(2)}_{i,p}=1$ from (5.12) shows that $v^{(2)}_{2p}$ is entropy decreasing. ∎

Further results on the discrete Laplace operator in this context can be found in [1], for instance.

6. Numerical experiments.

6.1. A shock-capturing method for the scalar cubic problem. For $\gamma>0$ fixed and some initial data $u_0:\mathbb{R}\to\mathbb{R}$, consider as a model problem the (regularized) scalar Cauchy problem

$$\begin{aligned} u^{\gamma,\epsilon}_t+\left((u^{\gamma,\epsilon})^3\right)_x &= \epsilon\, u^{\gamma,\epsilon}_{xx}+\gamma\epsilon^2\, u^{\gamma,\epsilon}_{xxx},\\ u^{\gamma,\epsilon}(.,0) &= u_0 \end{aligned} \tag{6.1}$$

corresponding to (5.1).

It is well known [20] that there exists a weak solution u^γ of the hyperbolic conservation law, i.e., (6.1) with $\epsilon=0$, which is the L^1-limit of a sequence of solutions $\{u^{\gamma,\epsilon}\}_{\epsilon>0}$ for vanishing ϵ. In particular for Riemann problem initial data u_0, the function u^γ might contain undercompressive shock waves which depend on u_0 and the coefficient γ [11, 8].

Following subsection 5.1, we choose our viscosity and capillarity fluxes according to

$$\begin{aligned} f^{2*}(u_{j-1},\ldots,u_{j+2}) &= \frac{\beta}{2}\Big(u_{j+1}-u_j\Big),\\ f^{3*}(u_{j-1},\ldots,u_{j+2}) &= \frac{\delta}{6}\Big(u_{j+2}-u_{j+1}-u_j+u_{j-1}\Big). \end{aligned}$$

To satisfy (5.3) assume $\delta/\beta^2=3\gamma/4$ for $\beta,\delta>0$. With the entropy of choice $U(u)=u^4/4$ the basic entropy conservative schemes are given by either

I scheme (3.6) with $\alpha=1/2$ and $p=1$ or

II scheme (3.6) with $\alpha=1/2-1/\sqrt{2}$ and $p=2$.

In both cases we use $U^*\left(u^0,u^1\right)=U\left(u^*\left(u^0,u^1\right)\right)$ and v^* to be

$$\begin{aligned} v^*\left(u^{-1},u^0,u^1\right) &= \int_0^1 v\left(su^*\left(u^0,u^1\right)+(1-s)\,u^*\left(u^{-1},u^0\right)\right)ds\\ &= \frac{1}{4}\Big(u^*(u^0,u^1)+u^*(u^{-1},u^0)\Big)\Big(u^*(u^0,u^1)^2+u^*(u^{-1},u^0)^2\Big). \end{aligned}$$

The basic entropy conservative scheme in case I (II) is of second (third) order in space and time.

In all numerical experiments described below, the viscosity and capillarity fluxes for fluxes $f^{2/3}_{j+1/2}$ are evaluated in $u^{n+1}_{j-1},\ldots,u^{n+1}_{j+2}$, i.e., we treat them implicitly.

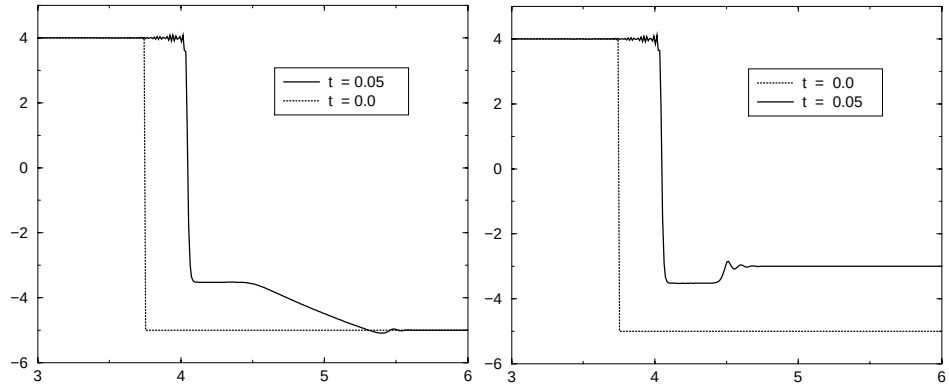


FIG. 6.1. *Typical wave patterns involving nonclassical shock waves. Results for scheme* II *are displayed.*

In Figure 6.1 we present the numerical results for two different choices of the initial data:

$$u_0^1(x) = \begin{cases} 4 : x < 0, \\ -5 : x > 0, \end{cases} \qquad u_0^2(x) = \begin{cases} 4 : x < 0, \\ -3 : x > 0. \end{cases} \tag{6.2}$$

For $\gamma = 2$ and initial data u_0^1, the weak solution u^γ consists of a slow nonclassical shock and a fast rarefaction, while u_0^2 enforces a slow nonclassical shock followed by a fast Lax shock. The numerical results have been performed with the discretization parameters

$$\beta = 5.0, \qquad \delta = 37.5, \qquad h = 0.005. \tag{6.3}$$

The figures demonstrate the ability of the scheme to reproduce nonclassical shock waves arising in Riemann problems together with shock and rarefaction waves. We approximately obtained the value -3.52 for the middle constant state in the second experiment with nonclassical and classical shock. This is better than the values obtained in [10, 16]. However, the correct value of the exact solution u^γ is $-11/3$.

To present a quantitative comparison we run the following experiment. We fix $\gamma = 2$ and choose the parameters according to (6.3). Now we compute the approximate solutions for both schemes I, II with the initial data

$$u_0(x) = \begin{cases} u_l : x < 0, \\ -\dfrac{5}{4} u_l : x > 0. \end{cases}$$

For $u_l > 1$, the exact solution u^γ consists of a nonclassical shock and a rarefaction connected by a middle state u_m as described above. In Figure 6.2 the approximate values of the middle state u_m obtained by schemes I and II are displayed for several values of $u_l \in [1, 11]$. The graphs describing the exact value $u_m = u_m(u_\gamma)$ in the cases $\gamma = 0, \gamma = 2, \gamma = \infty$ are also presented. The cases $\gamma = 0$, $\gamma = \infty$ give the exact middle value for the classical case, respectively, the extreme nonclassical case. We observe for small values of u_l a good approximation of the exact solution while bigger values of u_l lead to wrong solutions. The approximation of scheme II with the higher order

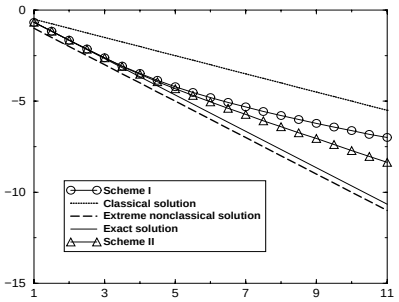


FIG. 6.2. *The (approximate) middle state u_m versus the state u_l.*

basic entropy flux is always better than the approximation by scheme I. We conclude by saying that our method seems to be reliable for computing nonclassical shocks at least for small amplitude initial data.

6.2. The "linear" shock capturing method for the scalar cubic problem. We now present numerical data for schemes approximating nonclassical weak solutions of the scalar cubic problem that are based on the regularization that is linear in the entropy variable $v = U'(u) = f(u) = u^3$ (while in subsection 6.1 the regularization was linear in the conservative variable u). Therefore, instead of (6.1), we consider

$$\begin{aligned} u_t^{\gamma,\epsilon} + f(u^{\gamma,\epsilon})_x &= \epsilon\, f(u^{\gamma,\epsilon})_{xx} + \gamma\epsilon^2\, f(u^{\gamma,\epsilon})_{xxx}, \\ u^{\gamma,\epsilon}(.,0) &= u_0. \end{aligned} \tag{6.4}$$

This leads to the following choice for the viscosity and capillarity fluxes:

$$\begin{aligned} f^{2*}(u_{j-1},\ldots,u_{j+2}) &= \frac{\beta}{2}\Big(f(u_{j+1}) - f(u_j)\Big), \\ f^{3*}(u_{j-1},\ldots,u_{j+2}) &= \frac{\delta}{6}\Big(f(u_{j+2}) - f(u_{j+1}) - f(u_j) + f(u_{j-1})\Big). \end{aligned}$$

As the basic entropy scheme we take (corresponding to scheme II in subsection 6.1) scheme (3.6) with $\alpha = 1/2 - 1/\sqrt{2}$ and $p = 2$. For the numerical parameters, let β, γ to be 37.5, respectively, 1. In Figure 6.3 we present computations for the Riemann initial data

$$u_0^1(x) = \begin{cases} 50 : x < 5, \\ -62.5 : x > 5, \end{cases} \qquad u_0^2(x) = \begin{cases} 100 : x < 5, \\ -125 : x > 5. \end{cases}$$

The calculations have been performed with discretization width $h = 0.005$. In the specific cases considered here we obtain a configuration with a slow nonclassical shock and a fast rarefaction.

Note that these type of schemes allow the stable computation of nonclassical shocks, even for *very large* amplitude data. This was not possible for the discretization based on (6.1).

6.3. The p-system with phase transition: A shape memory material. In this section, we perform long-time computations for the p-system (4.17). We consider

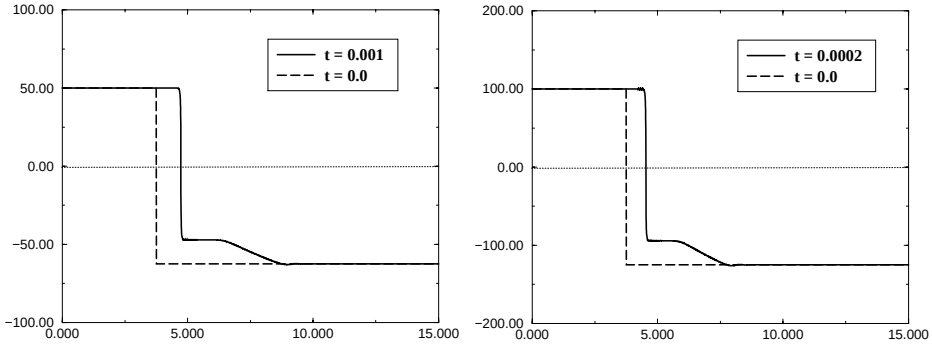


FIG. 6.3. *Stable computation of nonclassical shocks for large initial data* $u_0^{1/2}$.

the weak solution that is obtained as the limit (as $\varepsilon \to 0$) of classical solutions of the p-system with linear viscous regularization in the entropy variable:

$$\begin{aligned} \partial_t w^\varepsilon - \partial_x V^\varepsilon &= \varepsilon \partial_{xx} \sigma(w^\varepsilon), \\ \partial_t V^\varepsilon - \partial_x \sigma(w^\varepsilon) &= \varepsilon \partial_{xx} V^\varepsilon. \end{aligned} \tag{6.5}$$

The scheme that we consider here is the fourth order entropy conservative scheme (4.18) to which we add a viscous flux of fourth order. Following the notations introduced for scheme (4.18), we define the complete numerical flux by

$$\widetilde{g}_4^* = g_4^* - v_4^{(2)*},$$

where the entropy conservative flux g_4^* is given by

$$g_4^*(v_{j-1}, \ldots, v_{j+2}) = g\left(\frac{2}{3}(v_j + v_{j+1}) - \frac{1}{12}(v_{j-1} + \cdots + v_{j+2})\right),$$

whereas the viscous flux is defined by

$$v_4^{(2)*}(v_{j-1}, \ldots, v_{j+2}) = \frac{2h}{3}(v_{j+1} - v_j) - \frac{h}{24}(v_{j+1} + v_{j+2} - v_j - v_{j-1}).$$

Taylor expansion shows that this flux is of fourth order with respect to $h\partial_{xx} v$. The resulting scheme is, denoting $\widetilde{g}_{j+1/2}^{*n+1} = (\widetilde{g}_{1,j+1/2}^{*n+1}, \widetilde{g}_{2,j+1/2}^{*n+1})^T$,

$$\left\{ \begin{array}{rcl} w_j^{*n+1} - w_j^{*n} + \lambda \left(\widetilde{g}_{1,j+1/2}^{*n+1} - \widetilde{g}_{1,j-1/2}^{*n+1}\right) &=& 0, \\ V_j^{*n+1} - V_j^{*n} + \lambda \left(\widetilde{g}_{2,j+1/2}^{*n+1} - \widetilde{g}_{2,j-1/2}^{*n+1}\right) &=& 0 \end{array} \right. \qquad (j \in \mathbb{Z}).$$

We present two computations with periodic boundaries. The results have been obtained on a grid of 1000 cells and with CFL-number 0.25.

The first experiment deals with the same initial Riemann data as in the previous example (cf. (4.19)). We illustrate the effect of artificial viscous regularization on the results plotted in Figure 4.1. The numerical experiment is performed in the interval $[0, 1]$, with initial data

$$u_0(x) = \left\{ \begin{array}{l} (1,1)^T : x \in [0, 0.5), \\ (1,-1)^T : x \in [0.5, 1). \end{array} \right.$$

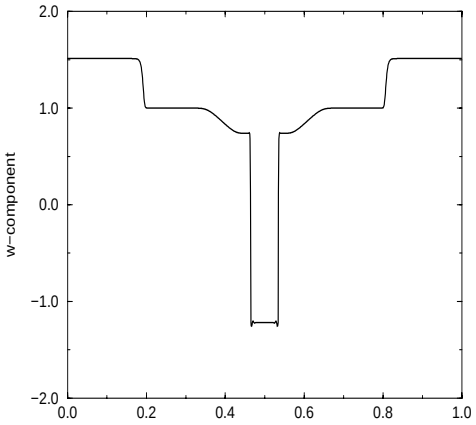


FIG. 6.4. *Numerical approximation of the p-system with artificial viscosity.*

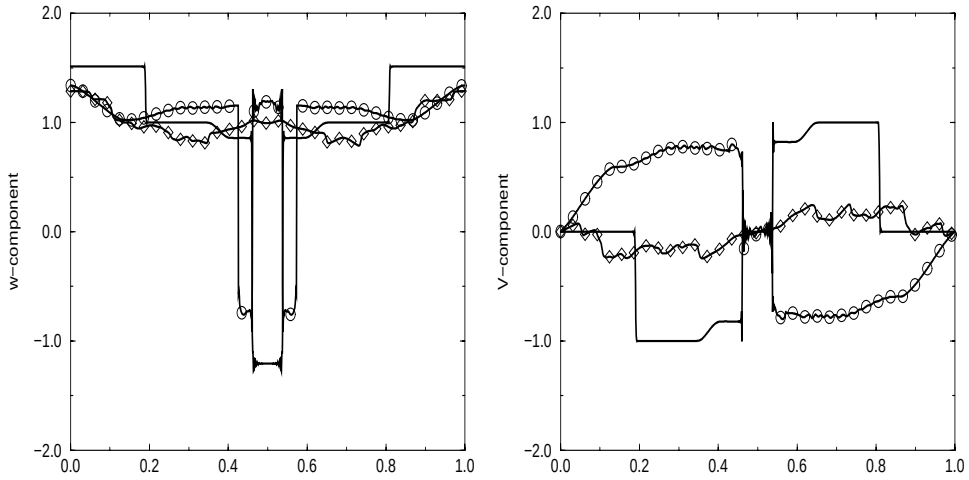


FIG. 6.5. *Time evolution of a diphasic stressed material for short time range (no symbols), intermediate time range (circles), and long time range (diamonds).*

Note that the computed solution (Figure 6.4) corresponds to the four wave "classical" pattern described by Shearer [19].

The second experiment corresponds to the same initial data, but now we performed a longer time computation. We illustrate the property of these materials to come back to their initial configuration at rest, i.e., the constant solution $(w, V) = (1, 0)$. During the computations, numerous phase transitions were created and canceled out. The evolution in time of the approximate solution is displayed in Figure 6.5 for different times. The left figure shows the w-component, and the right figure the V-component.

REFERENCES

[1] L. ANNÉ, P. JOLY, AND Q. H. TRAN, *An analysis of higher order finite difference schemes for*

the acoustic wave equation, Comput. Geosci., 4 (2000), pp. 207–249.

[2] A Abeyaratne and J.K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids,* Arch. Ration. Mech. Anal., 114 (1991), pp. 119–154.

[3] R. Abeyaratne and J.K. Knowles, *Implications of viscosity and strain-gradient effects for the kinetics of propagating phase boundaries in solids,* SIAM J. Appl. Math., 51 (1991), pp. 1205–1221.

[4] D. Aregba-Driollet and J.M. Mercier, *Convergence of non-linear algorithms for semi-linear hyperbolic systems,* Rend. Sem. Mat. Univ. Padova, 102 (1999), pp. 241–283.

[5] C. Chalons and P.G. LeFloch, *A fully discrete scheme for diffusive-dispersive conservation laws,* Numer. Math., 89 (2001), pp. 493–509.

[6] C. Chalons and P.G. LeFloch, *High-order entropy conservative schemes and kinetic relations for van der Waals fluids,* J. Comput. Phys., 167 (2001), pp. 1–23.

[7] K.O. Friedrich and P.D. Lax, *Systems of conservation laws with a convex extension,* Proc. Nat. Acad. Sci. U.S.A., 68 (1971), pp. 1686–1688.

[8] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Scalar conservation laws,* Arch. Ration. Mech. Anal., 139 (1997), pp. 1–56.

[9] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Finite difference schemes,* SIAM J. Numer. Anal., 35 (1998), pp. 2169–2194.

[10] B.T. Hayes and P.G. LeFloch, *Nonclassical shock and kinetic relations: Strictly hyperbolic schemes,* SIAM J. Math. Anal., 31 (2000), pp. 941–991.

[11] D. Jacobs, W.R. McKinney, and M. Shearer, *Traveling wave solutions of the modified Korteweg-deVries Burgers equation,* J. Differential Equations, 116 (1995), pp. 448–467.

[12] P.D. Lax and B. Wendroff, *Systems of conservation laws,* Comm. Pure Appl. Math., 13 (1960), pp. 217–237.

[13] P.G. LeFloch, *Propagating phase boundaries: Formulation of the problem and existence via the Glimm scheme,* Arch. Ration. Mech. Anal., 123 (1993), pp. 153–197.

[14] P.G. LeFloch, *Hyperbolic Systems of Conservation Laws: The Theory of Classical and Nonclassical Shock Waves*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2002.

[15] P.G. LeFloch, *An introduction to nonclassical shocks of systems of conservation laws,* Proceedings of the International School on Hyperbolic Problems, Freiburg, Germany, Oct. 97, D. Kröner, M. Ohlberger, and C. Rohde, eds., Lect. Notes Comput. Sci. Engrg. 5, Springer–Verlag, Berlin, 1998, pp. 28–72.

[16] P.G. LeFloch and C. Rohde, *High-order schemes, entropy inequalities and nonclassical shocks,* SIAM J. Numer. Anal., 37 (2000), pp. 2023–2060.

[17] P.G. LeFloch and M.D. Thanh, *Non-classical Riemann solvers and kinetic relations.* II. *An hyperbolic-elliptic model of phase transitions,* Proc. Roy. Soc. Edinburgh Sect. A, 132 (2001), pp. 181–219.

[18] J.M. Mercier and B. Piccoli, *Global continuous Riemann solver for nonlinear elasticity,* Arch. Ration. Mech. Anal., 156 (2001), pp. 89–119.

[19] M. Shearer, *The Riemann problem for a class of conservation laws of mixed type,* J. Differential Equations, 46 (1982), pp. 426–443.

[20] M.E. Schonbek, *Convergence of solutions to nonlinear dispersive equations,* Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.

[21] T. Sonar, *Entropy production in second-order three-point schemes,* Numer. Math., 62 (1992), pp. 371–390.

[22] W. Strauss and L. Vasquez, *Numerical solution of a Klein Gordon equation,* J. Comput. Phys., 28 (1978), pp. 271–278.

[23] E. Tadmor, *Numerical viscosity and the entropy condition for conservative difference schemes,* Math. Comput., 43 (1984), pp. 369–381.

[24] E. Tadmor, *Entropy conservative finite element schemes,* in Numerical Methods for Compressible Flows—Finite Difference, Element and Volume Techniques, Proceedings of the Winter Annual Meeting, AMD 78, T.E. Tezduyar and T.J.R. Hughes, eds., ASME, New York, 1986, pp. 149–158.

[25] E. Tadmor, *The numerical viscosity of entropy stable schemes for systems of conservation laws,* Math. Comput., 49 (1987), pp. 91–103.

[26] L. Truskinovsky, *Dynamics of non-equilibrium phase boundaries in a heat conducting nonlinear elastic medium,* J. Appl. Math. Mech., 51 (1987), pp. 777–784.

SIAM J. NUMER. ANAL.
Vol. 40, No. 5, pp. 1993–2006

MULTIGRID METHODS FOR ANISOTROPIC EDGE REFINEMENT*

THOMAS APEL† AND JOACHIM SCHÖBERL‡

Abstract. A finite element method with optimal convergence on nonsmooth three dimensional domains requires anisotropic mesh refinement towards the edges. Multigrid methods for anisotropic tensor product meshes are available and are based either on line (or plane) smoothers or on semicoarsening strategies. In this paper we suggest and analyze a new multigrid scheme combining semicoarsening and line smoothers to obtain a solver of optimal algorithmic complexity for anisotropic meshes along edges.

Key words. multigrid, line Jacobi smoother, edge singularity, anisotropic mesh

AMS subject classification. 65N55

PII. S0036142900375414

1. Introduction.

The finite element simulation of three dimensional problems described by partial differential equations is a challenging task. To keep the simulation time low at least two aspects have to be taken into account. First, the underlying triangulation has to be efficient for approximating the (unknown) solution, and, second, the chosen algorithm for solving the large scale system of equations should be of optimal algorithmic complexity.

For two dimensional elliptic problems optimal triangulations for low order finite elements can be achieved by isotropic mesh refinement based on a posteriori error indicators [21]. The corresponding approach in three dimensions does not, in general, lead to optimal triangulations in the sense of an energy error of order $N^{-p/3}$, p being the polynomial degree. Besides local refinement towards the corners optimal triangulations require *anisotropic* mesh refinement towards the edges of the geometry [1, 2, 3].

Multigrid methods (see [13, 7] and many references therein) are algorithms of optimal (this means linear) complexity for the solution of the systems of linear equations obtained by the finite element method. Multigrid methods have been suggested and analyzed for anisotropic problems with tensor product structure. One approach is to take care of the strong connections by properly designing line or plane smoothers [22, 14, 20, 9], another is to build up the hierarchy of triangulations by semicoarsening [24, 12, 17].

Semicoarsening and line/plane smoothing can be combined. In [5], for example, a certain class of singular perturbed problems is considered, and it is suggested to use semicoarsening with respect to the "harmless" coordinate and line relaxation in the direction of the singular perturbation. In the case of edge singularities, the edge direction could be considered as the harmless direction, but then we need a good plane smoother in the orthogonal direction. Since this strategy is not easy to implement for a hierarchical smoother, we propose using a line smoother in the edge direction

*Received by the editors July 19, 2000; accepted for publication (in revised form) May 19, 2002; published electronically December 3, 2002. This work was supported by the Austrian Science Fund Fonds zur Förderung der wissenschaftlichen Forschung, Spezialforschungsbereich F013.
http://www.siam.org/journals/sinum/40-5/37541.html
†Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (na.apel@na-net.ornl.gov, www.tu-chemnitz.de/∼tap).
‡SFB Numerical and Symbolic Scientific Computing, Johannes Kepler Universität Linz, SFB F013, Linz, Austria (joachim@sfb013.uni- linz.ac.at, www.sfb013.uni-linz.ac.at/∼joachim).

and semicoarsening in the orthogonal plane, which turns out to be easy to implement and efficient in application. In this paper we prove robust V-cycle convergence rates of the suggested scheme. The framework is due to Braess and Hackbusch [6].

We note that this multigrid method is essentially a two dimensional standard multigrid where the third dimension is treated only in the smoother. The two dimensional method with mesh refinement towards singular corners is analyzed in [23]. While in that paper regularity and interpolation results have been cited from [4, 15], we cannot use results from literature immediately. The reason is that the two dimensional plane with mesh refinement is only a trace of the three dimensional domain where the problem is posed. In order to circumvent the loss of regularity due to trace theorems, we introduce an intermediate semidiscrete space $\widetilde{V}$ (see (4.4)) and prove regularity of an auxiliary problem and interpolation results ourselves.

The rest of the paper is organized as follows. In section 2 the investigated problem is formulated. Section 3 introduces the multigrid scheme. The multigrid analysis is performed in section 4; two proofs are postponed to sections 5 and 6. In section 7 we give numerical results confirming our theory.

2. Problem formulation and discretization. Let $\Omega = G \times Z$, where $G \subset \mathbb{R}^2$ is a polygonal domain and Z is a real interval. By the local nature of corner singularities (and then edge ones for Ω), we may suppose that G has possibly one corner with interior angle $\omega > \pi$ at the origin, the other interior angles being smaller than π. The corresponding edge of Ω is part of the z-axis and will be called the singular edge of Ω. Spatial variables are written as $(x, z) = (x_1, x_2, z)$ with $x \in G$ and $z \in Z$. Accordingly, the gradient is split into partial derivatives as $\nabla = (\partial_x, \partial_z)$. Let $V := H_0^1(\Omega)$ be the usual Sobolev space. We consider the Poisson equation with Dirichlet boundary conditions whose variational form is as follows: Find $u \in V$ such that

$$A(u, v) = f(v) \qquad \forall\, v \in V \tag{2.1}$$

with the symmetric, continuous, and elliptic bilinear form $A(.,.)$ and the continuous linear form $f(.)$ on V, namely,

$$A(u, v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x \qquad \text{and} \qquad f(v) := \int_\Omega f v \,\mathrm{d}x.$$

The energy norm is defined as $\|u\|_A := A(u, u)^{1/2}$.

The domain Ω is covered by a tensor product triangulation $\mathcal{T} = \mathcal{T}_x \otimes \mathcal{T}_z$, where $\mathcal{T}_x$ and $\mathcal{T}_z$ are conforming triangulations of G and Z, respectively [10]. The two dimensional triangulation $\mathcal{T}_x$ is assumed to fulfill the bounded minimal angle condition. The triangulation $\mathcal{T}_z$ is arbitrary. We define the mesh size functions

$$h_{L,x} = h_{L,x}(x, z) = \operatorname{diam} T_x \qquad \text{for } x \in T_x \in \mathcal{T}_x,\ z \in Z, \tag{2.2}$$

$$h_{L,z} = h_{L,z}(x, z) = \operatorname{diam} T_z \qquad \text{for } x \in G,\ z \in T_z \in \mathcal{T}_z \tag{2.3}$$

for plane and edge directions. The positive integer L denotes the final refinement level of the multigrid hierarchy defined below. We do not assume relations between $h_{L,x}$ and $h_{L,z}$, and thus anisotropic triangulations are included.

We introduce the piecewise affine finite element spaces

$$\mathcal{M}_0^1(\mathcal{T}_x) = \{u \in C^0(G) : u|_{\partial G} = 0, u|_{T_x} \in \mathcal{P}^1\ \forall\, T_x \in \mathcal{T}_x\},$$
$$\mathcal{M}_0^1(\mathcal{T}_z) = \{u \in C^0(Z) : u|_{\partial Z} = 0, u|_{T_z} \in \mathcal{P}^1\ \forall\, T_z \in \mathcal{T}_z\}$$

with the nodal bases $\{\varphi^i_{L,x}\}_{i=1}^{N_{L,x}}$ and $\{\varphi^i_{L,z}\}_{i=1}^{N_{L,z}}$ and space dimensions $N_{L,x}$ and $N_{L,z}$. Then the tensor product bilinear finite element space is defined by

$$V_L := \mathcal{M}^1_0(\mathcal{T}_x) \otimes \mathcal{M}^1_0(\mathcal{T}_z) = \left\{ u = \sum_{i,j} u_{i,j} \varphi^i_{L,x}(x) \varphi^j_{L,z}(z) \right\}.$$

The finite element approximation $u_L \in V_L$ of the variational problem (2.1) is defined by Galerkin projection:

$$A(u_L, v_L) = f(v_L) \qquad \forall\, v_L \in V_L. \tag{2.4}$$

Finally, we define the distance to the singular edge of Ω (the singular point of G, respectively) by

$$r = r(x,z) = r(x) = |x|. \tag{2.5}$$

For the following a priori estimate we refer to [1, 2].

THEOREM 2.1 (a priori estimate). *Let (x_T, z_T) denote the center of the element $T \in \mathcal{T}$. Assume that the mesh sizes fulfill*

$$\begin{array}{lcll} h_{L,x}(x,z) & \simeq & h_L\, r(x_T)^\beta & \forall\,(x,z) \in T \in \mathcal{T}, \\ h_{L,z}(x,z) & \simeq & h_L & \forall\,(x,z) \in T \in \mathcal{T} \end{array} \tag{2.6}$$

with the global (positive) mesh size parameter h_L. The grading parameter β is fixed and is assumed to fulfill

$$1 - \frac{\pi}{\omega} \cdot \frac{p}{2p-2} < \beta < 1. \tag{2.7}$$

Then there holds the a priori error estimate

$$\|u - u_h\|_1 \preceq h_L \|f\|_{0,p} \tag{2.8}$$

for $p > 2$. The number of elements is of optimal order h_L^{-3}.

The condition (2.7) shortens for $p = 2$ to the slightly weaker assumption

$$1 - \frac{\pi}{\omega} < \beta < 1, \tag{2.9}$$

but the estimate (2.8) has been proved in this case in [1] for certain mixed boundary conditions only. For the Dirichlet problem, only the result as stated in the theorem has been obtained yet. We underline that our multigrid theory is also valid under the weaker assumption (2.9).

3. Multigrid algorithm. The multigrid algorithm requires a sequence of triangulations $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_L$. We may and will assume that the triangulations and the generated finite element spaces are nested. The proposed refinement strategy is to perform first full refinement in the z-direction and then generate the hierarchy of meshes by refinement in the x-plane. Each triangulation in the hierarchy has the tensor product structure

$$\mathcal{T}_l = \mathcal{T}_{l,x} \otimes \mathcal{T}_z, \qquad 1 \le l \le L.$$

This means that there is the full refinement in z-direction for all levels l, $1 \le l \le L$. We define the mesh size functions $h_{l,x}$ and $h_{l,z} = h_z$ as analogies to (2.2) and (2.3). We assume that the triangulations $\mathcal{T}_{l,x}$ fulfill the bounded minimal angle condition. Further, the grading of the meshes fulfills

$$h_{l,x} = h_{l,x}(x,z) \simeq h_l\, r(x_T)^\beta \qquad \forall\, (x,z) \in T \in \mathcal{T}_l \tag{3.1}$$

with the global mesh size parameter h_l of level l and β from (2.9). The ratio of successive parameters h_{l-1}/h_l is assumed to be bounded.

We mention two methods to generate the sequence of meshes fulfilling (3.1). The first one is to split each triangle of $\mathcal{T}_{l-1,x}$ into four triangles, and move only new nodes next to the singular corner towards the corner. Another possibility, the so-called *dyadic partitioning*, is to use local mesh refinement of $\mathcal{T}_x$, where the elements with $h_{l,x} > C_0 h_l r(x_T)^\beta$ (with a suitably defined constant C_0) are marked for refinement. Both methods have advantages. The first one enables a more efficient data structure; the second one is related to a posteriori mesh size control.

We define the sequence of nested finite element spaces by

$$V_l := \mathcal{M}_0^1(\mathcal{T}_{l,x}) \otimes \mathcal{M}_0^1(\mathcal{T}_z)$$

and the linear operator $A_l : V_l \to V_l$ by

$$(A_l u_l, v_l)_0 = A(u_l, v_l) \qquad \forall\, u_l, v_l \in V_l,$$

$l = 1, 2, \ldots, L$. Additionally, we define for $u_L \in V_L$ the L_2 and energy projections $Q_l : V_L \to V_l$ and $P_l : V_L \to V_l$ by

$$\begin{aligned} (Q_l u_L, v_l)_0 &= (u_L, v_l)_0 \quad \forall v_l \in V_l, \\ A(P_l u_L, v_l) &= A(u_L, v_l) \quad \forall v_l \in V_l, \end{aligned}$$

$l = 1, \ldots, L$. For $1 \le l \le k \le L$ there holds the equation

$$P_l = A_l^{-1} Q_l A_k \qquad \text{on } V_k.$$

The smoother of the considered multigrid scheme is a line Jacobi or symmetric line Gauss–Seidel iteration along mesh lines in the z-direction. Let $\{\varphi_{l,i}^x\}_{i=1}^{N_{l,x}}$ be the nodal basis of $\mathcal{M}_0^1(\mathcal{T}_x)$ and $N_{l,x} = \dim \mathcal{M}_0^1(\mathcal{T}_x)$. We define on each level l the subspaces

$$V_{l,i} := \operatorname{span}\{\varphi_{l,i}^x\} \otimes \mathcal{M}_0^1(\mathcal{T}_z), \quad i = 1, \ldots, N_{l,x},$$

and the corresponding energy projections $P_{l,i} : V_l \to V_{l,i}$, determined for $u_l \in V_l$ by

$$A(P_{l,i} u_l, v_{l,i}) = A(u_l, v_{l,i}) \qquad \forall v_{l,i} \in V_{l,i}.$$

Then the (damped) line Jacobi smoother

$$S_l := I - \tau \sum_{i=1}^{N_{l,x}} P_{l,i}$$

with a suitable damping parameter $\tau \simeq 1$ can be written as

$$S_l = I - \tau D_l^{-1} A_l.$$

The operator D_l^{-1} is indeed the inverse of a self-adjoint and positive definite operator $D_l : V_l \to V_l$. It leads to the inner product

$$D_l(u_l, v_l) := (D_l u_l, v_l)_0 \qquad \forall\, u_l, v_l \in V_l$$

and the associated norm $\|u_l\|_{D_l} := D_l(u_l, u_l)^{1/2}$. By the technique of [8] the analysis of the present paper applies also to the symmetric multiplicative counterpart. The multiplicative version does not need damping at all.

Since the spaces are nested, the grid transfer operators are canonically defined by embedding. As usual, we define the V-cycle multigrid preconditioning operators $C_l^{-1} : V_l \to V_l$ by induction beginning with $C_1 = A_1$. For $l > 1$ and $f \in V_l$ we define $C_l^{-1} f = x_{2m+1}$ where $x_0 = 0$,

$$\left.\begin{aligned} x_i &= x_{i-1} + \tau D_l^{-1}(f - A_l x_{i-1}), \qquad i = 1, 2, \ldots, m, \\ x_{m+1} &= x_m + C_{l-1}^{-1} Q_{l-1}(f - A_l x_m), \\ x_i &= x_{i-1} + \tau D_l^{-1}(f - A_l x_{i-1}), \qquad i = m+2, m+3, \ldots, 2m+1. \end{aligned}\right\} \tag{3.2}$$

First, m steps of presmoothing are performed; then, the coarse grid correction takes place; finally, m steps of postsmoothing are applied. The self-adjoint operator C_L^{-1} can be used in the multigrid iteration with iteration matrix $I - C_L^{-1} A_L$ or as a preconditioner in the conjugate gradient iteration.

4. Multigrid analysis. In this section we analyze the convergence of the multigrid scheme formulated above. In order to apply the multigrid framework of Braess and Hackbusch [6] (see also Theorem 3.6 in [7]), we first need to verify the approximation property

$$\|u_l - P_{l-1} u_l\|_{D_l}^2 \le C \|u_l\|_A^2 \qquad \forall\, u_l \in V_l,\ l = 2, 3, \ldots L; \tag{4.1}$$

see Theorem 4.4. The V-cycle convergence rate estimate is then a corollary. We start with three lemmata.

LEMMA 4.1 (representation of D_l-norm). *For the norm induced by the line Jacobi preconditioner D_l, there holds the following equivalence:*

$$\|u_l\|_{D_l}^2 \simeq \|h_{l,x}^{-1} u_l\|_0^2 + \|\partial_z u_l\|_0^2 \qquad \forall\, u_l \in V_l. \tag{4.2}$$

Proof. Let $u_l \in V_l$. The decomposition $u_l = \sum_{i=1}^{N_{l,x}} u_{l,i}$ with $u_{l,i} \in V_{l,i}$ is unique. By the additive Schwarz method [11] (using the most similar notation) we obtain

$$\|u_l\|_{D_l}^2 = \sum_{i=1}^{N_{l,x}} \|u_{l,i}\|_A^2.$$

Inverse inequalities applied to the basis functions $\varphi_{l,x}^i$ give

$$\begin{aligned} \|u_l\|_{D_l}^2 &= \sum_{i=1}^{N_{l,x}} \left(\|\partial_x u_{l,i}\|_0^2 + \|\partial_z u_{l,i}\|_0^2 \right) \\ &\simeq \sum_{i=1}^{N_{l,x}} \left(\|h_{l,x}^{-1} u_{l,i}\|_0^2 + \|\partial_z u_{l,i}\|_0^2 \right). \end{aligned}$$

By mapping techniques one verifies the L_2 stability of the splitting (see [7, Chapter 5]):

$$\sum_{i=1}^{N_{l,x}} \|c_i \varphi_{l,i}^x\|_{0,T_x}^2 \simeq \left\| \sum_{i=1}^{N_{l,x}} c_i \varphi_{l,i}^x \right\|_{0,T_x}^2 \qquad \forall\, c \in \mathbb{R}^{N_{l,x}},\ \forall\, T_x \in \mathcal{T}_{l,x}. \tag{4.3}$$

Since the equivalence is local, we may insert the element-wise constant weight $h_{l,x}$. Summing over the elements $T_x \in \mathcal{T}_{l,x}$ gives

$$\sum_{i=1}^{N_{l,x}} \|h_{l,x}^{-1} c_i \varphi_{l,i}^x\|_{0,G}^2 \simeq \left\| \sum_{i=1}^{N_{l,x}} h_{l,x}^{-1} c_i \varphi_{l,i}^x \right\|_{0,G}^2.$$

Now set $u_{l,i} =: c_i(z)\varphi_{l,i}^x(x)$. Integration over $z \in Z$ leads to

$$\sum_{i=1}^{N_{l,x}} \|h_{l,x}^{-1} u_{l,i}\|_0^2 \simeq \|h_{l,x}^{-1} u_l\|_0^2.$$

By inserting $\partial_z u_{l,i} =: c_i(z)\varphi_{l,i}^x$ into (4.3), summing over the elements $T_x \in \mathcal{T}_{l,x}$, and integrating over $z \in Z$, we obtain

$$\sum_{i=1}^{N_{l,x}} \|\partial_z u_{l,i}\|_0^2 \simeq \|\partial_z u_l\|_0^2,$$

and the proof is complete. ∎

The sequence of nested spaces V_l is contained in the *semidiscrete* space

$$\widetilde{V} := H_0^1(G) \otimes \mathcal{M}_0^1(\mathcal{T}_z). \tag{4.4}$$

For our analysis we consider a subspace of $\widetilde{V}$,

$$V^+ := \{u \in \widetilde{V} : \|u\|_{V^+} < \infty\},$$
$$\|u\|_{V^+}^2 := \|r^\beta \partial_x \nabla u\|_0^2 + \|r^{\beta-1}\partial_x u\|_0^2,$$

with r and β defined in (2.5) and (2.9), respectively.

LEMMA 4.2 (regularity). *Let $u \in \widetilde{V}$ be the solution of the variational problem*

$$A(u,v) = (f,v)_0 \qquad \forall\, v \in \widetilde{V} \tag{4.5}$$

with f such that $r^\beta f \in L_2$, $1 - \pi/\omega < \beta < 1$. Then there holds the regularity estimate

$$\|u\|_{V^+} \preceq \|r^\beta f\|_0. \tag{4.6}$$

Note that the restriction $\beta < 1$ ensures that $f \in H^{-1}(\Omega)$ [16, Theorem 8.15], and therefore the right-hand side of (4.5) makes sense.

LEMMA 4.3 (interpolation error estimate). *There exists an interpolation operator $I_l : V^+ \to V_l$ such that the interpolation error satisfies*

$$\|u - I_l u\|_A \preceq h_l \|u\|_{V^+} \qquad \forall\, u \in V^+.$$

The proofs of Lemmata 4.2 and 4.3 are postponed to sections 5 and 6.

THEOREM 4.4 (approximation property). *The approximation property* (4.1) *is fulfilled for the considered multigrid method* (3.2).

Proof. We use the equivalence (4.2) and obtain

$$\|u_l - P_{l-1}u_l\|_{D_l}^2 \simeq \|h_{l,x}^{-1}(u_l - P_{l-1}u_l)\|_0^2 + \|\partial_z(u_l - P_{l-1}u_l)\|_0^2.$$

The second term of the right-hand side is simply estimated by

$$\|\partial_z(u_l - P_{l-1}u_l)\|_0 \le \|u_l - P_{l-1}u_l\|_A \le \|u_l\|_A.$$

It remains to show

$$\|h_{l,x}^{-1}(u_l - P_{l-1}u_l)\|_0 \preceq \|u_l\|_A. \tag{4.7}$$

As usual we formulate a dual problem. Since $V_l \subset \widetilde{V}$ for all l, we define $w \in \widetilde{V}$ by

$$A(w,v) = (h_{l,x}^{-2}(u_l - P_{l-1}u_l), v)_0 \qquad \forall\, v \in \widetilde{V}.$$

Lemma 4.2 and assumption (3.1) on $h_{l,x}$ yield

$$\|w\|_{V^+} \preceq \|r^\beta h_{l,x}^{-2}(u_l - P_{l-1}u_l)\|_0 \preceq h_l^{-1}\|h_{l,x}^{-1}(u_l - P_{l-1}u_l)\|_0.$$

Here, no special consideration of the origin is necessary. We continue with Galerkin orthogonality, approximation, and regularity:

$$\begin{aligned}
\|h_{l,x}^{-1}(u_l - P_{l-1}u_l)\|_0^2 &= A(w, u_l - P_{l-1}u_l)\\
&= A(w - I_{l-1}w, u_l - P_{l-1}u_l)\\
&\le \|w - I_{l-1}w\|_A \, \|u_l - P_{l-1}u_l\|_A\\
&\preceq h_l \|w\|_{V^+} \, \|u_l\|_A\\
&\preceq \|h_{l,x}^{-1}(u_l - P_{l-1}u_l)\|_0 \, \|u_l\|_A.
\end{aligned}$$

Dividing by one factor gives (4.7) and thus the desired approximation property. ∎

THEOREM 4.5 (convergence rate estimate). *For the V-cycle multigrid algorithm* (3.2) *with m presmoothing and m postsmoothing steps there holds the convergence rate estimate*

$$\|I - C_L^{-1}A_L\|_A \le \frac{C}{C+2m}. \tag{4.8}$$

Proof. The result follows from the general multigrid theory of Braess and Hackbusch [6] (see also Theorem 3.6 in [7]) by using the approximation property (4.1) which is proved in Theorem 4.4. ∎

5. Regularity.

Proof of Lemma 4.2. First, we use Fourier decomposition in the z-direction to transform the three dimensional problem into a sequence of two dimensional problems. For that, let $\{e_i\}_{i=1}^{N_z}$ be the Fourier basis in $\mathcal{M}_0^1(\mathcal{T}_z)$; that means $e_i = e_i(z)$ are the eigenvectors of

$$(e_i', v')_{0,Z} = \lambda_i^2 (e_i, v)_{0,Z} \qquad \forall\, v \in \mathcal{M}_0^1(\mathcal{T}_z)$$

with $\lambda_i > 0$, $(e_i, e_j)_{0,Z} = \delta_{ij}$, and $(e_i', e_j')_{0,Z} = \lambda_i^2 \delta_{ij}$. Inserting $u = \sum_{i=1}^{N_z} u_i(x) e_i(z)$ into (4.5) yields that $u_i(x)$ solves

$$(\partial_x u_i, \partial_x v)_{0,G} + \lambda_i^2 (u_i, v)_{0,G} = (f_i, v)_{0,G} \qquad \forall\, v \in H_0^1(G), \tag{5.1}$$

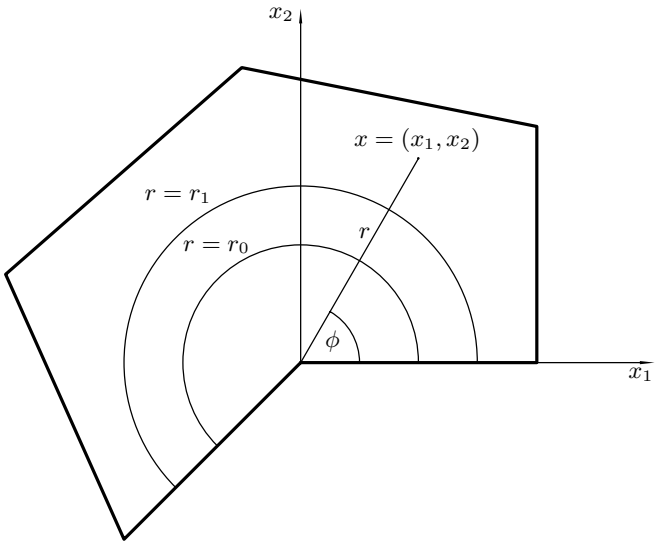


FIG. 5.1. *Illustration of the notation.*

with $f_i = (f, e_i)_{0,Z}$ and $r^\beta f_i \in L_2(G)$. Since

$$\begin{aligned}
\|u\|^2_{V^+} &= \|r^\beta \partial_x^2 u\|_0^2 + \|r^\beta \partial_x \partial_z u\|_0^2 + \|r^{\beta-1}\partial_x u\|_0^2 \\
&= \sum_{i=1}^{N_z} \left\{ \|r^\beta \partial_x^2 u_i\|_{0,G}^2 + \lambda_i^2 \|r^\beta \partial_x u_i\|_{0,G}^2 + \|r^{\beta-1}\partial_x u_i\|_{0,G}^2 \right\}, \\
\|r^\beta f\|_0^2 &= \sum_{i=1}^{N_z} \|r^\beta f_i\|_{0,G}^2,
\end{aligned}$$

the lemma follows after proving the regularity estimate

$$\|r^\beta \partial_x^2 u_i\|_{0,G}^2 + \lambda_i^2 \|r^\beta \partial_x u_i\|_{0,G}^2 + \|r^{\beta-1}\partial_x u_i\|_{0,G}^2 \le C \, \|r^\beta f_i\|_{0,G}^2, \tag{5.2}$$

for the family of two dimensional problems (5.1). The constant C does not depend on λ_i. In the following we will skip the subscript G.

Introduce now a cut-off function $\xi \in C^\infty(\mathbb{R}_+)$, $\xi(r) \in [0,1]$, $\xi(r) = 1$ for $r \le r_0$, $\xi(r) = 0$ for $r \ge r_1 > r_0$; see Figure 5.1 for an illustration. The regular part $u_i^R = (1-\xi)u_i$ satisfies

$$(\partial_x u_i^R, \partial_x v)_0 + \lambda_i^2 (u_i^R, v)_0 = (f_i^R, v)_0 \qquad \forall\, v \in H_0^1(G), \tag{5.3}$$

with $f_i^R = -\Delta[(1-\xi)u_i] + \lambda_i^2(1-\xi)u_i = (1-\xi)f_i - 2(\partial_x u_i)\partial_x(1-\xi) - u_i\Delta(1-\xi)$. Observe that $u_i^R = f_i^R = 0$ for $r \le r_0$. From

$$\|u_i\|_1 \preceq \|f_i\|_{-1} = \sup_{w \in H_0^1(G)} \frac{(f_i, w)}{\|w\|_1} \le \|r^\beta f_i\|_0 \sup_{w \in H_0^1(G)} \frac{\|r^{-\beta} w\|_0}{\|w\|_1} \preceq \|r^\beta f_i\|_0$$

($\|r^{-\beta} w\|_0 \preceq \sum_{j=0}^1 \|r^{1-\beta} D^j w\|_0 \preceq \|w\|_1$ for $\beta < 1$, see [16, Theorem 8.15], there follows

$$\|f_i^R\|_0 \preceq \|r^\beta f_i\|_0. \tag{5.4}$$

Inserting $v = u_i^R$ into (5.3), applying the Cauchy–Schwarz inequality, and dividing one factor $\|u_i^R\|_0$ leads to $\|u_i^R\|_0 \preceq \lambda_i^{-2}\|f_i^R\|_0$ and, consequently, to

$$\lambda_i \|\partial_x u_i^R\|_0 \preceq \|f_i^R\|_0. \tag{5.5}$$

Since u_i^R and f_i^R vanish in an r_0-neighborhood of the corner, a smoothed domain provides the same solution and the full regularity estimate (from the Poisson problem)

$$\|u_i^R\|_2 \preceq \|f_i^R - \lambda_i^2 u_i^R\|_0 \preceq \|f_i^R\|_0.$$

Thus, by using (5.4) and (5.5), the estimate

$$\|r^\beta \partial_x^2 u_i^R\|_0^2 + \lambda_i^2 \|r^\beta \partial_x u_i^R\|_0^2 + \|r^{\beta-1}\partial_x u_i^R\|_0^2 \preceq \|r^\beta f_i\|_0^2$$

is established, and we are left to prove the corresponding inequality for the singular part $u_i^S = \xi u_i$.

Since u_i^S and $f_i^S := f_i - f_i^R$ vanish for $r \geq r_1$, we can extend both by 0 onto the infinite cone $K := \{(r\cos\phi, r\sin\phi) \in \mathbb{R} : 0 < r < \infty, 0 < \phi < \omega\}$, where they fulfill the variational problem

$$(\partial_x u_i^S, \partial_x v)_{0,K} + \lambda_i^2 (u_i^S, v)_{0,K} = (f_i^S, v)_{0,K} \qquad \forall\, v \in H_0^1(K).$$

After the change of variables $\hat{x} = \lambda_i x$, we obtain that $\hat{u}_i^S(\hat{x}) = u_i^S(x)$ is the solution of the following problem with right-hand side $\hat{f}_i^S(\hat{x}) := f_i^S(x)$:

$$(\partial_{\hat{x}} \hat{u}_i^S, \partial_{\hat{x}} \hat{v})_{0,K} + (\hat{u}_i^S, \hat{v})_{0,K} = (\lambda_i^{-2} \hat{f}_i^S, \hat{v})_{0,K} \qquad \forall\, \hat{v} \in H_0^1(K).$$

We can now use the regularity result in Proposition 1.1 of [18, p. 385],

$$\|\hat{u}_i^S\|_{E^2_\beta(K)} \preceq \|\lambda_i^{-2}\hat{f}_i^S\|_{E^0_\beta(K)} \qquad \text{if } |\beta - 1| < \frac{\pi}{\omega},$$

where the space $E^\ell_\beta(K)$ is the completion of $C_0^\infty(\overline{K} \setminus 0)$ with respect to the norm

$$\|v\|^2_{E^\ell_\beta(K)} := \sum_{|\alpha| \leq \ell} \int_K \hat{r}^{2\beta}(1 + \hat{r}^{|\alpha|-\ell})^2 |\hat{D}^\alpha \hat{v}|^2 \, \mathrm{d}\hat{x}$$

[18, p. 300], $\hat{r} := |\hat{x}|$. By transforming the norms one obtains

$$\begin{aligned}
&\int_G \left(r^{2\beta}|\partial_x^2 u_i^S|^2 + (r^{2\beta}\lambda_i^2 + r^{2\beta-2})|\partial_x u_i^S|^2\right) \mathrm{d}x \\
&= \lambda_i^{-2\beta+2} \int_K \hat{r}^{2\beta} \left(|\partial_{\hat{x}}^2 \hat{u}_i^S|^2 + (1 + \hat{r}^{-2})|\partial_{\hat{x}} \hat{u}_i^S|^2\right) \mathrm{d}\hat{x} \\
&\preceq \lambda_i^{-2\beta+2} \int_K \hat{r}^{2\beta} |\lambda_i^{-2} \hat{f}_i^S|^2 \, \mathrm{d}\hat{x} \\
&= \int_G r^{2\beta} |f_i^S|^2 \, \mathrm{d}x.
\end{aligned}$$

Estimate (5.4) implies $\|r^\beta f_i^S\|_0 \preceq \|r^\beta f_i\|_0$, and the desired result is proved. ∎

6. Interpolation.

Proof of Lemma 4.3. Let $\mathrm{Z}_h : H^1(G) \to \mathcal{M}_0^1(\mathcal{T}_{l,x})$ be the Scott–Zhang interpolation operator [19]. For an arbitrary triangle $T_x \in \mathcal{T}_{l,x}$ and for $m = 0, 1$, $\ell = 1, 2$, $p \in [1, \infty]$, the error estimate

$$|u - \mathrm{Z}_h u|_{m,T_x} \preceq |T_x|^{1/2-1/p} h_{l,x}^{\ell-m} |u|_{\ell,p,\tilde{T}_x} \tag{6.1}$$

is satisfied [19], with $\tilde{T}_x$ being the union of T_x and the triangles adjacent to T_x.

Denote by $\{\varphi_i\}_{i=1}^{N_z}$ the nodal basis in $\mathcal{M}_0^1(\mathcal{T}_z)$ and split u with respect to this basis,

$$u = \sum_{i=1}^{N_z} u_i(x)\varphi_i(z).$$

Note that the u_i here are different from those used in section 5. Then we define the interpolation operator $\mathrm{I}_l : \widetilde{V} \to V_l$ by

$$\mathrm{I}_l u = \sum_{i=1}^{N_z} (\mathrm{Z}_h u_i)(x)\varphi_i(z).$$

For an arbitrary element $T = T_x \times (z_j, z_{j+1})$, $T_x \in \mathcal{T}_{l,x}$, $(z_j, z_{j+1}) \in \mathcal{T}_z$, introduce $\tilde{T} := \tilde{T}_x \times (z_j, z_{j+1})$. Now divide the set $\mathcal{T}_l$ into two subsets, $\mathcal{T}_l = \mathcal{T}_{l,R} \cup \mathcal{T}_{l,S}$,

$$\mathcal{T}_{l,R} := \left\{ T \in \mathcal{T}_l : \inf_{(x,z)\in\tilde{T}} |x| > 0 \right\},$$
$$\mathcal{T}_{l,S} := \left\{ T \in \mathcal{T}_l : \inf_{(x,z)\in\tilde{T}} |x| = 0 \right\},$$

namely (regular) elements away from the edge and (singular) elements close to the edge. For elements $T \in \mathcal{T}_{l,R}$ we obtain from (6.1) the estimates

$$\begin{aligned}
\|\partial_x(u - \mathrm{I}_l u)\|_{0,T} &\simeq \sum_{i=j}^{j+1} h_z^{1/2} \|\partial_x(u_i - \mathrm{Z}_h u_i)\|_{0,T_x} \\
&\preceq \sum_{i=j}^{j+1} h_z^{1/2} h_{l,x} \|\partial_x^2 u_i\|_{0,\tilde{T}_x} \\
&\simeq h_{l,x} \|\partial_x^2 u\|_{0,\tilde{T}}, \\
\|\partial_z(u - \mathrm{I}_l u)\|_{0,T} &= h_z^{-1/2} \|(u_{j+1} - u_j) - \mathrm{Z}_h(u_{j+1} - u_j)\|_{0,T_x} \\
&\preceq h_z^{-1/2} h_{l,x} \|\partial_x(u_{j+1} - u_j)\|_{0,\tilde{T}_x} \\
&= h_{l,x} \|\partial_x \partial_z u\|_{0,\tilde{T}}.
\end{aligned}$$

That means, by using (3.1) and $r(x_T) \simeq r(x)$ for $x \in \tilde{T}$,

$$\sum_{T\in\mathcal{T}_{l,R}} \|u - \mathrm{I}_l u\|_A^2 \preceq \sum_{T\mathcal{T}_{l,R}} h_{l,x}^2 \|\nabla\partial_x u\|_{0,\tilde{T}}^2 \preceq h_l^2 \|u\|_{\widetilde{V}}^2, \tag{6.2}$$

where we have also used that only a finite number of $\tilde{T}$ overlap in any point.

For elements $T \in \mathcal{T}_{l,S}$ we derive estimates in a weighted space, namely

$$\begin{aligned}\|\partial_x(u - \mathrm{I}_l u)\|_{0,T} &\simeq \sum_{i=j}^{j+1} h_z^{1/2} \|\partial_x(u_i - \mathrm{Z}_h u_i)\|_{0,T_x} \\ &\preceq \sum_{i=j}^{j+1} h_z^{1/2} \|\partial_x u_i\|_{0,\tilde{T}_x} \\ &\preceq \|\partial_x u\|_{0,\tilde{T}} \\ &\preceq h_{l,x}^{1-\beta} \|r^{\beta-1}\partial_x u\|_{0,\tilde{T}_x} \\ &\simeq h_l \|r^{\beta-1}\partial_x u\|_{0,\tilde{T}_x},\end{aligned}$$

which is valid due to $r \preceq h_{l,x},\ r(x_T) \simeq h_{l,x}$ (thus $h_{l,x}^{1-\beta} \simeq h_l$), and $\beta \leq 1$. Moreover, we get

$$\begin{aligned}\|\partial_z(u - \mathrm{I}_l u)\|_{0,T} &\simeq h_z^{-1/2} \|(u_{j+1} - u_j) - \mathrm{Z}_h(u_{j+1} - u_j)\|_{0,T_x} \\ &\preceq h_z^{-1/2} |T_x|^{-1/2} h_{l,x} \|\partial_x(u_{j+1} - u_j)\|_{0,1,\tilde{T}_x} \\ &\preceq h_z^{-1/2} |T_x|^{-1/2} h_{l,x} \|r^{-\beta}\|_{0,\tilde{T}_x} \|r^{\beta}\partial_x(u_{j+1} - u_j)\|_{0,\tilde{T}_x} \\ &\preceq h_{l,x}^{1-\beta} \|r^{\beta}\partial_x\partial_z u\|_{0,\tilde{T}} \\ &\simeq h_l \|r^{\beta}\partial_x\partial_z u\|_{0,\tilde{T}}.\end{aligned}$$

Consequently, we get

$$\sum_{T \in \mathcal{T}_{l,S}} \|u - \mathrm{I}_l u\|_A^2 \preceq h_{l,x}^{1-\beta} \|u\|_{\widetilde{V}}^2. \tag{6.3}$$

With (6.2) and (6.3) the lemma is proved. ∎

7. Numerical results. For verification of the analysis and to demonstrate the performance of the method, we present the following numerical results. We consider the three dimensional L-shaped domain

$$\Omega = G \times (0,1) \qquad \text{with} \qquad G = (-1,1)^2 \setminus [0,1]^2.$$

An initial triangulation was generated with sixteen nodes and six prismatic elements. For the first tests (Tables 7.1 and 7.2), the elements were successively bisected in the vertical direction until the triangulation $\mathcal{T}_1$ was obtained. For further tests (Table 7.3), we first split the prisms at $z = 0.05$ and proceeded with bisecting as before. The hierarchy of triangulations was obtained by bisecting the whole stack of elements based on a priori element markers. All those elements $T \in \mathcal{T}_l$ were refined for which

$$h_{T,x} r_T^{-\beta} \geq 0.3 \max_{T' \in \mathcal{T}_l} \{h_{T',x} r_{T'}^{-\beta}\}$$

holds. We chose the refinement factor $\beta = 1/2$, which fulfills the condition $\beta > 1/3$ to ensure an asymptotically optimal discretization error. Pictures of the meshes are shown in Figure 7.1 and Figure 7.2.

For preconditioning the resulting finite element system, the multigrid scheme (3.2) was applied with one multiplicative presmoothing and one reverse-order multiplicative

TABLE 7.1
Results for uniform refinement in the z-direction, 16 layers of elements.

	Point smoother			Line smoother		
Nodes	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]
136	1	2		1	2	
289	7.0	18		1.1	5	
816	4.8	17	0.1	1.3	7	0.1
1717	2.7	13	0.2	1.5	8	0.2
3536	2.3	12	0.4	1.6	9	0.6
7480	2.1	11	1.0	1.7	10	1.5
20060	2.1	12	3.1	1.9	11	4.3
40766	2.0	12	7.2	1.9	11	9.9
111027	2.5	13	20.2	1.9	11	26.1
241536	3.2	14	50.3	2.2	11	61.9
320093	2.8	13	78.6	2.0	11	104.6

TABLE 7.2
Results for uniform refinement in the z-direction, 64 layers of elements.

	Point smoother			Line smoother		
Nodes	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]
520	1	2		1	2	
1105	100.8	63	0.4	1.1	5	0.2
3120	65.6	63	1.4	1.3	8	0.6
6565	29.9	47	3.2	1.5	8	1.4
13520	19.9	40	6.7	1.6	9	3.4
28600	14.6	32	12.4	1.7	10	8.3
76700	12.2	29	29.9	1.9	11	22.8
155870	10.9	27	62.4	1.9	11	50.6
424515	10.3	26	155.7	1.9	11	129.1
923520	9.9	25	346.6	2.2	11	298.3

TABLE 7.3
Results for mesh with nonuniform refinement in the z-direction (boundary layer at $z = 0$), 64 layers of elements.

	Point smoother			Line smoother		
Nodes	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]	$\kappa\{C_L^{-1}A_L\}$	CG its.	Time [sec]
520	1	2		1	2	
1105	261.2	84	0.6	1.1	5	0.2
3120	237.7	139	3.0	1.3	8	0.6
6565	169.5	122	8.1	1.5	8	1.4
13520	149.1	110	18.3	1.6	9	3.4
28600	114.8	96	36.6	1.7	10	8.3
76700	87.9	81	82.1	1.9	11	22.8
155870	62.6	64	145.2	1.9	11	50.7
424515	36.1	50	293.4	1.9	11	128.6
923520	30.6	45	610.7	2.2	11	297.7

postsmoothing step. For comparison, we did all computations also with a multigrid method with the standard point smoother on the same hierarchy of meshes.

First we computed the condition numbers $\kappa\{C_L^{-1}A_L\}$ of the preconditioned matrix. In addition, we solved the Poisson problem

$$-\Delta u = 1 \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega,$$

and used the multigrid preconditioner in the CG method to reduce the residual error (measured by $\sqrt{r^T C_L^{-1} r}$) by a factor of 10^{-8}. Tables 7.1, 7.2, and 7.3 show the results

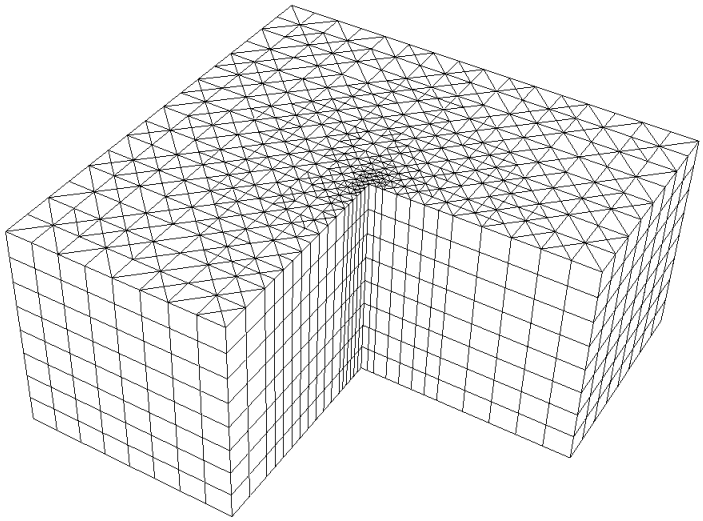


FIG. 7.1. *Mesh with* 8 *uniform layers and* 5013 *nodes.*

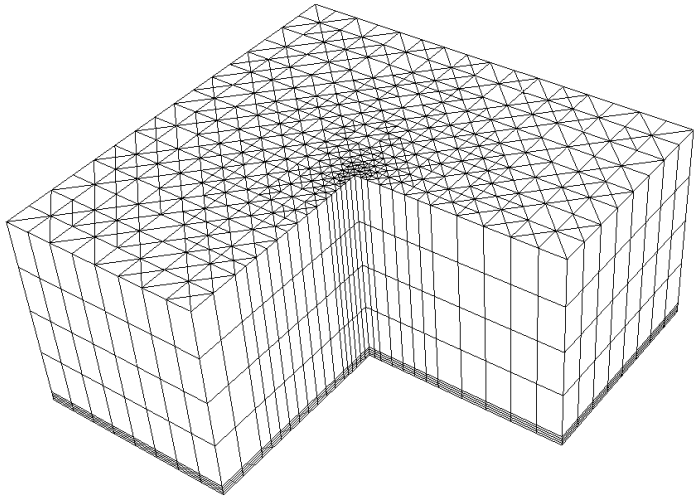


FIG. 7.2. *Mesh with boundary layer at* $z = 0.05$ *and* 5013 *nodes.*

for various numbers of layers in the vertical direction and numbers of nodes. Processor time refers to an SGI Octane R 10000, 250 MHz.

The tests show the robust performance of our multigrid method. The iteration numbers are independent of the refinement depth, and they are also independent of the mesh in edge direction. In comparison, the point smoother has problems with strongly anisotropic meshes, expressed through a large condition number and a large number of CG iterations. The ratio of CPU times is less since our implementation of the line smoother is about twice as expensive as the point smoother. Finally, we mention that the proposed multigrid algorithm can be extended to complicated geometries by macro element techniques.

Acknowledgment. The authors want to thank Prof. Serge Nicaise, Univ. Valenciennes, for his contribution to Lemma 4.2.

REFERENCES

[1] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.
[2] T. Apel and S. Nicaise, *The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges*, Math. Methods Appl. Sci., 21 (1998), pp. 519–549.
[3] T. Apel, S. Nicaise, and J. Schöberl, *A non-conforming finite element method with anisotropic mesh grading for the Stokes problem in domains with edges*, IMA J. Numer. Anal., 21 (2001), pp. 843–856.
[4] I. Babuška, R. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.
[5] S. Börm and R. Hiptmair, *Analysis of tensor product multigrid*, Numer. Algorithms, 26 (2001), pp. 219–234.
[6] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967–975.
[7] J. H. Bramble, *Multigrid Methods*, Longman Scientific and Technical, Essex, UK, 1993.
[8] J. H. Bramble and J. E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.
[9] J. H. Bramble and X. Zhang, *Uniform convergence of the multigrid V-cycle for an anisotropic problem*, Math. Comp., 70 (2001), pp. 453–470.
[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[11] M. Griebel and P. Oswald, *On the abstract theory of additive and multiplicative schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.
[12] M. Griebel and P. Oswald, *Tensor product type subspace splittings and multi-level iterative methods for anisotropic problems*, Adv. Comput. Math., 4 (1995), pp. 171–206.
[13] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, 1985.
[14] W. Hackbusch, *The frequency decomposition multi-grid method. Part* I: *Application to anisotropic equations*, Numer. Math., 56 (1989), pp. 229–245.
[15] V. A. Kondrat'ev, *Boundary value problems for elliptic equations on domains with conical or angular points*, Trudy Moskov. Mat. Obshch., 16 (1967), pp. 209–292 (in Russian).
[16] A. Kufner, *Weighted Sobolev Spaces*, Teubner, Leipzig, Germany, 1980.
[17] S. Margenov, L. Xanthis, and L. Zikatanov, *On the optimality of the semi-coarsening AMLI algorithm*, in Iterative Methods in Linear Algebra, II, IMACS, New Brunswick, NJ, 1995, pp. 254–269.
[18] S. A. Nazarov and B. A. Plamenevsky, *Elliptic problems in domains with piecewise smooth boundaries*, de Gruyter Exp. Math. 13, Walter de Gruyter, Berlin, 1994.
[19] L. R. Scott and S. Zhang, *Finite element interpolation of non-smooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[20] R. Stevenson, *Robustness of multi-grid applied to anisotropic equations on convex domains with reentrant corners*, Numer. Math., 66 (1993), pp. 373–398.
[21] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley–Teubner, Chichester, UK, Stuttgart, Germany, 1996.
[22] G. Wittum, *On the robustness of ILU smoothing*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 699–717.
[23] H. Yserentant, *The convergence of multi-level methods for solving finite-element equations in the presence of singularities*, Math. Comp., 47 (1986), pp. 399–409.
[24] S. Zhang, *Optimal-order nonnested multigrid methods for solving finite element equations* III: *On degenerate meshes*, Math. Comp., 64 (1995), pp. 23–49.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2007–2025

ANALYSIS OF A SEMIDISCRETE VERSION OF THE WIGNER EQUATION*

THIERRY GOUDON†

Abstract. We introduce a semidiscretized version of the Wigner equation—discretization concerning the velocity variable. We show that the corresponding discrete velocity problem is well-posed and permits us to approach the solution of the continuous problem when the mesh size of the discretization vanishes. The approximation shows spectral accuracy because the rate of convergence corresponds to the (Sobolev) regularity of the solution of the continuous problem. We also discuss the behavior of the solution with respect to the Planck constant.

Key words. Wigner equation, discrete velocity models, semiclassical limit.

AMS subject classifications. 35Q40, 65N35

PII. S0036142901388366

1. Introduction.

We are concerned with the following Wigner (or "quantum Liouville") equation:

$$\partial_t f + \xi \cdot \nabla_x f = \Theta(V)(f) \qquad \text{in } \mathbb{R}_t^+ \times \mathbb{R}_x^N \times \mathbb{R}_\xi^N, \tag{1.1}$$

where the right-hand side is defined by the pseudodifferential operator

$$\Theta(V)(f) = i\mathcal{F}_{y\to\xi}^{-1}\big((V(x+y/2) - V(x-y/2))\widehat{f}(t,x,y)\big). \tag{1.2}$$

Throughout the paper we indifferently denote by $\mathcal{F}(f)$ or by $\widehat{f}$ the Fourier transform given, under suitable integrability conditions on f, by

$$\widehat{f}(y) = \int_{\mathbb{R}^N} e^{-iy\cdot\xi} f(\xi)\, d\xi,$$

and $\mathcal{F}^{-1}(f)(\xi) = \int e^{+iy\cdot\xi} f(y)\, dy/(2\pi)^N$.

This equation is intended to model the quantum transport of electrons in a semiconductor device. The unknown $f(t,x,\xi)$ is real valued, but, contrary to what happens usually in kinetic theory, it is not naturally nonnegative. Usually, the electric potential V is obtained in a self-consistent way through the Poisson equation $\Delta V(t,x) = \int f\, d\xi - D(x)$, D being a given doping profile. This relationship describes the Coulombian interactions between the electrons. Note that this definition is not obvious at all since the natural framework for f is the space $L^2(\mathbb{R}^N \times \mathbb{R}^N)$ so that it is not clear how the integral of f can make sense. However, here and below, we restrict ourselves to a linear situation where the potential V is given; it lies at least in $L^\infty(\mathbb{R}^N)$. Consequently, the operator $\Theta(V)$ is bounded in $L^2(\mathbb{R}^N)$ (see Lemma 3.1).

Equation (1.1) may be seen as the quantum equivalent of the classical Vlasov equation

$$\partial_t f + \xi \cdot \nabla_x f - \nabla_x V \cdot \nabla_\xi f = 0,$$

*Received by the editors April 24, 2001; accepted for publication (in revised form) May 21, 2002; published electronically December 13, 2002. Part of this work was performed during a stay at the Erwin Schrödinger Institute in Vienna. This work was supported by the Austrian START project "Nonlinear Schrödinger and quantum Boltzmann equations."

http://www.siam.org/journals/sinum/40-6/38836.html

†Laboratoire J.A. Dieudonné, UMR 6621, Université Nice-Sophia Antipolis, Parc Valrose F-06108 Nice cedex 02, France (goudon@math.unice.fr).

where we consider that electrons are moving along the trajectories of the Hamiltonian $\xi^2/2 + V$. Actually, the integro-differential equation (1.1) can be obtained by considering the Wigner transform, introduced in [15], of the solution of the Schrödinger equation with potential V. On the other hand, the Vlasov equation can be recovered as the Planck constant goes to 0 when dealing with suitably rescaled Wigner transforms. The mathematical analysis of these questions has been performed by Lions and Paul [11]. We also refer for further details and results on the model to the works of Markowich [6], Markowich and Ringhofer [9], Degond and Markowich [3], and the classical treatise of Markowich, Ringhofer, and Schmeiser [10]. We also quote for an extensive study and application of the Wigner transform the recent paper of Gérard, Markowich, Mauser, and Poupaud [4, 5].

In this work, we wish to introduce a semidiscrete version of the Wigner equation (1.1) and to discuss some properties of the approximation obtained in this way. Discretization here is performed with respect to the ξ variable and we search for a discrete model that looks like (1.1):

$$\partial_t w_n + \xi_n \cdot \nabla_x w_n = \big(\Theta_d(V)(w)\big)_n \qquad \text{in } \mathbb{R}_t^+ \times \mathbb{R}_x^N \times \mathcal{Z}, \tag{1.3}$$

where $\mathcal{Z}$ is some subset of $\mathbb{Z}^N$, possibly $\mathbb{Z}^N$ itself, and ξ_n belongs to a discrete set of velocities parametrized by $n \in \mathcal{Z}$. Therefore, our aim is two fold:

(1) On one hand, the problem (1.3) should be well-posed; see Theorem 3.4.

(2) On the other hand, we expect that (1.3) "approaches" (1.1) in the sense that we may construct from the w_n's a function $f^h(t, x, \xi)$ which converges to f, solution of (1.1), as the parameter h, related to some "mesh size," goes to 0; see Theorem 4.4.

This kind of question has been addressed, for instance, by Ringhofer, with a slightly different approach, in [13], [14] and in Arnold and Ringhofer [2]; more generally, we refer to the recent and complete review on computational methods for semiconductor models [12]. We also mention the work of Arnold, Lange, and Zweifel in [1] in a monodimensional and stationary framework, with a discussion of relevant inflow boundary conditions. Here, our approach is rather close to the analysis of finite difference schemes by Markowich and Poupaud [8]. Also notice the use of Wigner transform techniques to analyze numerical schemes by Markowich, Pietra, and Pohl [7]. Our paper is organized as follows. The next section will introduce the main ideas to achieve the program: it contains basic preliminaries and notations. Section 3 is devoted to the discrete problem: having defined the discrete operator Θ_d, we analyze well-posedness of the corresponding problem (1.3). Roughly speaking, we use the (formal) interpretation of Θ as a convolution to define the operator Θ_d as a discrete convolution. Then, in section 4, we define the approximation f^h from the discrete solution and show a convergence result to the solution of (1.1). Our method has infinite order: the rate of convergence is given by the degree of regularity on the Sobolev scale with respect to the ξ variable of the solution of the continuous problem. Eventually, section 5 is concerned with the semiclassical limit: we investigate the behavior of the solutions as the (scaled) Planck constant goes to 0, and we show convergence to the solution of the Vlasov equation.

2. Preliminaries. Roughly speaking, the idea consists of neglecting the large values of the dual variable, which is denoted $y \in \mathbb{R}^N$, of ξ: we restrict the band length. Then, our construction relies on the following simple observation. Let $g(\xi)$ be a smooth function, in the sense that its Fourier transform has a compact support, i.e., $\mathrm{supp}(\widehat{g}) \subset B(0, R)$. Let $Y = [-L/2, +L/2]^N$ be a box in $\mathbb{R}^N$ with length $L/2$ larger than R. We denote by $L^2_\#$ the set of square integrable functions on Y, which

are Y-periodically extended on $\mathbb{R}^N$. We associate with such a function g a function $G \in L^2_{\#}$ as follows:

$$\left\{ \begin{array}{l} Y = [-L/2, +L/2]^N, \qquad 0 < R < L/2, \\ G(y) = \widehat{g}(y) \quad \text{on } B(0,R), \quad G(y) = 0 \quad \text{on } Y \backslash B(0,R), \\ y \longmapsto G(y) \quad \text{is } Y\text{-periodic.} \end{array} \right. \tag{2.1}$$

A simple computation shows that the Fourier coefficients of G are given by discrete equidistributed values of g; namely, we have

$$\begin{aligned} \widehat{G}(n) &= L^{-N} \int_Y e^{i2\pi y \cdot n/L} G(y)\, dy \\ &= L^{-N} \int_{\mathbb{R}^N} e^{i2\pi y \cdot n/L} \widehat{g}(y)\, dy = (2\pi/L)^N g(2\pi n/L) \end{aligned}$$

by the inversion formula. Therefore, we naturally set $g_M(\xi) = \mathcal{F}^{-1}_{y \to \xi}(\widehat{g_M})(\xi)$ with $\widehat{g_M}(y) = \sum_{|n| \le M} \widehat{G}(n) e^{-i2\pi n \cdot y/L} \chi_{|y| \le R}(y)$. Here and below $|y|$ and $B(0,R)$ refer to the infinite norm on $\mathbb{R}^N$ and the corresponding ball, while, for a multi-integer $n \in \mathbb{Z}^N$, $|n|$ stands for the length $|n| = |n_1| + \cdots + |n_N|$. This actually means that

$$\begin{aligned} g_M(\xi) &= \sum_{|n| \le M} \int_{\mathbb{R}^N} e^{iy \cdot \xi} e^{i2\pi n \cdot y/L} \widehat{G}(n) \chi_{|y| \le R}(y)\, dy/(2\pi)^N \\ &= \sum_{|n| \le M} (2\pi/L)^N \; g(2\pi n/L) \int_{B(0,R)} \exp(iy \cdot (\xi - 2\pi n/L))\, dy/(2\pi)^N \\ &= \sum_{|n| \le M} \left(L^{-N} \; g(2\pi n/L) \prod_{j=1}^N \frac{2\sin(R(\xi_j - 2\pi n_j/L))}{\xi_j - 2\pi n_j/L} \right). \end{aligned}$$

Since, by definition of the Fourier series,

$$\sum_{|n| \le M} \widehat{G}(n) e^{-i2\pi n \cdot y/L} \xrightarrow[M \to \infty]{} G \qquad \text{in } L^2_{\#},$$

it follows that g_M tends to g in $L^2(\mathbb{R}^N)$ as M goes to ∞. Indeed, we have

$$\begin{aligned} \|g - g_M\|^2_{L^2(\mathbb{R}^N)} &= (2\pi)^{-N} \|\widehat{g} - \widehat{g_M}\|^2_{L^2(\mathbb{R}^N)} \\ &= (2\pi)^{-N} \int_{B(0,R)} \Big| \widehat{g}(y) - \sum_{|n| \le M} \widehat{G}(n) e^{-i2\pi n \cdot y/L} \Big|^2 dy \\ &\le (2\pi)^{-N} \int_Y \Big| \widehat{g}(y) - \sum_{|n| \le M} \widehat{G}(n) e^{-i2\pi n \cdot y/L} \Big|^2 dy \\ &\le (L/(2\pi))^N \; \Big\| G - \sum_{|n| \le M} \widehat{G}(n) e^{-i2\pi n \cdot y/L} \Big\|^2_{L^2_{\#}(Y)}. \end{aligned}$$

Let us summarize these simple facts in the following claim (which is known as the Shannon sampling theorem in signal processing).

LEMMA 2.1. *Let $g(\xi)$ be a smooth function whose Fourier transform has a compact support* $\operatorname{supp}(\widehat{g}) \subset B(0,R)$. *Let $Y = [-L/2, +L/2]^N$ with $L/2 > R$ and set*

$$g_M(\xi) = \sum_{|n| \le M} \left(L^{-N} \; g(2\pi n/L) \prod_{j=1}^N \frac{2\sin(R(\xi_j - 2\pi n_j/L))}{\xi_j - 2\pi n_j/L} \right).$$

Then, one has $\lim_{M\to\infty}\|g-g_M\|_2=0$.

Remark 1. In (1.1) the unknown f is obtained as the Wigner transform of some density matrix; in turn, the density $\rho(t,x)=\int f\,d\xi$ makes sense and remains nonnegative; see [11]. It is worth remarking that if g is given through a truncated Wigner transform, namely,

$$g(x,\xi)=\int_{\mathbb{R}^N}\psi(x+y/2)\,\overline{\psi(x-y/2)}\,\chi_{|y|\le R}\,e^{iy\cdot\xi}\,dy/(2\pi)^N$$

with, say, $\psi\in L^2(\mathbb{R}^N)$ (for the sake of simplicity), then the associated discrete density remains nonnegative. This means that

$$\begin{aligned}\sum_{n\in\mathbb{Z}^N}\widehat{G}(x,n)&=\sum_{n\in\mathbb{Z}^N}\widehat{G}(x,n)e^{in\cdot 0}=G(0)\\&=\widehat{g}(x,0)=\Big(\psi(x+y/2)\,\overline{\psi(x-y/2)}\,\chi_{|y|\le R}\Big)\Big|_{y=0}=\psi(x)\,\overline{\psi(x)}\ge 0.\end{aligned}$$

For future use, let us set up here the notations and classical formulae. For f,g in $L^2_\#$, we set

$$\begin{cases}\big(f,g\big)_{L^2_\#}=L^{-N}\displaystyle\int_Y f(y)\overline{g(y)}\,dy, & \|f\|^2_{L^2_\#}=L^{-N}\displaystyle\int_Y|f(y)|^2\,dy,\\ \widehat{f}(n)=\big(f,e^{-i2\pi y\cdot n/L}\big)_{L^2_\#}=L^{-N}\displaystyle\int_Y f(y)e^{+i2\pi y\cdot n/L}\,dy.\end{cases}$$

Note that in this definition the norm on $L^2_\#$ depends on the size of the box. For (complex valued) sequences u,v in $\ell^2=\ell^2(\mathbb{Z}^N)$, we set

$$\big(u,v\big)_{\ell^2}=\sum_{n\in\mathbb{Z}^N}u_n\overline{v_n}.$$

Of course, we have the Parseval formula

$$\big(f,g\big)_{L^2_\#}=\sum_{n\in\mathbb{Z}^N}\widehat{f}(n)\overline{\widehat{g}(n)}=\big(\widehat{f},\widehat{g}\big)_{\ell^2},\tag{2.2}$$

which holds for functions f,g in $L^2_\#$. We will also use the notation

$$(u*v)_n=\sum_{k\in\mathbb{Z}^N}u_kv_{n-k},$$

and we recall that, when u or v belongs to ℓ^1, we have

$$\sum_{n\in\mathbb{Z}^N}(u*v)_n=\sum_{n\in\mathbb{Z}^N}u_n\sum_{n\in\mathbb{Z}^N}v_n.\tag{2.3}$$

3. The discrete problem. In this section, we define the discrete operator by means of discrete convolution. Then, we also discuss its fundamental properties and the well-posedness of the semidiscrete evolution equation.

3.1. The discrete operator. We recall that the Vlasov–Wigner operator (1.2) reads as

$$\Theta(V)(f)(\xi) = i\mathcal{F}^{-1}_{y\to\xi}(D_V\widehat{f})(\xi) = i\int_{\mathbb{R}^N} e^{i\xi\cdot y} D_V(x,y)\widehat{f}(y)\,dy/(2\pi)^N,$$

where we note that

$$D_V(x,y) = V(x+y/2) - V(x-y/2).$$

Then, the Parseval equality gives

$$\begin{aligned}\|\Theta(V)(f)\|_2 &= (2\pi)^{-N/2}\|\mathcal{F}\big(\Theta(V)(f)\big)\|_2 = (2\pi)^{-N/2}\|D_V\mathcal{F}(f)\|_2\\ &\le 2(2\pi)^{-N/2}\|V\|_\infty\|\mathcal{F}(f)\|_2 = 2\|V\|_\infty\|f\|_2.\end{aligned}$$

As mentioned in the introduction, we deduce the following statement.

LEMMA 3.1. *Let V belong to $L^\infty(\mathbb{R}^N)$. The operator $\Theta(V)$ is continuous on $L^2(\mathbb{R}^N)$ with norm $|\!|\!|\Theta|\!|\!|_{\mathcal{L}(L^2(\mathbb{R}^N))} \le 2\|V\|_\infty$.*

Let us consider the action of this operator on the smooth function g studied in the previous section. The support property implies that

$$\Theta(V)(g)(\xi) = i\int_{B(0,R)} e^{i\xi\cdot y} D_V(x,y)\widehat{g}(y)\,dy/(2\pi)^N.$$

Since $\Theta(V)$ is continuous on $L^2(\mathbb{R}^N)$, by Lemma 2.1 $\Theta(V)(g_M)$ converges to $\Theta(V)(g)$ in $L^2(\mathbb{R}^N)$ as $M\to\infty$. However, we get

$$\begin{aligned}\Theta(V)(g_M)(\xi) &= i\int_{B(0,R)} D_V(x,y)\widehat{g_M}(y)e^{i\xi\cdot y}\,dy/(2\pi)^N\\ &= i\sum_{|n|\le M}\widehat{G}(n)\int_{B(0,R)} D_V(x,y)e^{i(\xi-2\pi n/L)\cdot y}\,dy/(2\pi)^N\\ &= \sum_{|n|\le M} g(2\pi n/L)\ a(\xi-2\pi n/L),\end{aligned}\tag{3.1}$$

where

$$a(\zeta) = i\ L^{-N}\ \int_{B(0,R)} D_V(x,y)e^{i\zeta\cdot y}\,dy.\tag{3.2}$$

These formulae will be our guide to write a relevant discrete operator.

In order to construct a discrete model, we set for $n\in\mathbb{Z}^N$ and a mesh size $h>0$,

$$\xi_n = 2\pi nh,$$

and we interpret the w_n's that will be naturally searched for in ℓ^2 as the Fourier coefficients of a function $G\in L^2_\#$ with an h-dependent box $Y^h = [-1/(2h),+1/(2h)]^N$:

$$G(y) = \sum_{n\in\mathbb{Z}^N} w_n e^{-i2\pi hy\cdot n} \in L^2_\#(Y^h).$$

This idea is reminiscient of the deep interpretation of finite difference schemes by the pseudodifferential formalism given by Markowich and Poupaud [8]. According to

(3.1)–(3.2), we define the discrete operator by convolution:

$$\left\{\begin{array}{l} \Big(\Theta_d(V)(w)\Big)_n = \displaystyle\sum_k w_k a_{n-k} = (a*w)_n, \\ a_n = i\ h^N \displaystyle\int_{\mathbb{R}^N} D_V(x,y)\chi_{|y|\leq R^h}(y)\ e^{i2\pi hn\cdot y}\,dy. \end{array}\right. \tag{3.3}$$

Now notice that the truncation acts on the potential with a radius $R^h < (2h)^{-1}$, where R^h is a sequence of positive numbers which tends to ∞ as $h \to 0$: we truncate the frequencies at a level less than the inverse of twice the mesh size. Following (2.1), we introduce

$$\left\{\begin{array}{l} \mathcal{V}(y) = D_V(x,y)\chi_{|y|\leq R^h}(y) \text{ on } B(0,R^h), \qquad \mathcal{V}(y) = 0 \text{ on } Y^h\backslash B(0,R^h), \\ y \longmapsto \mathcal{V}(y) \quad \text{is } Y^h\text{-periodic} \end{array}\right. \tag{3.4}$$

so that $\frac{1}{i}\ a_n$ appears as the nth Fourier coefficient of the Y^h-periodic function $\mathcal{V}$. We also remark that $\mathcal{V} \in L^\infty_\#$. Therefore, we are concerned with the following discrete problem:

$$\left\{\begin{array}{l} \partial_t w_n + \xi_n \cdot \nabla_x w_n = \Big(\Theta_d(V)(w)\Big)_n \qquad \text{in } \mathbb{R}^+_t \times \mathbb{R}^N_x \times \mathbb{Z}^N_n, \\ w_n(t=0,x) = w^0_n(x). \end{array}\right. \tag{3.5}$$

It is not clear at all that (3.3) defines a bounded operator on the natural functional space ℓ^2 for the w_n's since nothing, up to additional regularity assumptions on V, ensures that a belongs to ℓ^1. (See [1] for some comments on this difficulty.) However, we are able to establish this property, which is the discrete analogue of Lemma 3.1.

PROPOSITION 3.2. *The operator Θ_d is a bounded linear operator on ℓ^2, and its norm is estimated uniformly with respect to h by $|\!|\!|\Theta_d|\!|\!|_{\mathcal{L}(\ell^2)} \leq 2\|V\|_\infty$. Furthermore, the coefficients a_n which define the operator Θ_d are real and even; in turn, the operator is skew-symmetric.*

Proof. Let f and g be two functions in $L^2_\#$. In particular, we notice that the product fg lies in $L^1_\#$, and consequently the Fourier coefficients $\widehat{fg}$ belong to ℓ^∞. We shall use the following relation:

$$\widehat{fg}(n) = \widehat{f} * \widehat{g}(n). \tag{3.6}$$

Also note that if $f \in L^2_\#$ and $g \in L^\infty_\#$, we have $fg \in L^2_\#$; therefore, the corresponding coefficients $\widehat{fg}(n) = \widehat{f} * \widehat{g}(n)$ also belong to ℓ^2.

Here, the operator Θ_d is defined by

$$\Big[\Theta_d(v)w\Big](n) = a * w(n) = i\Big(\widehat{\mathcal{V}} * \widehat{G}\Big)(n) = i\widehat{\mathcal{V}G}(n)$$

by using the definitions (3.3) and (3.4), while G is the function in $L^2_\#$ having the w_n's as Fourier coefficients. Then it yields (recall that the norm $\|\cdot\|_{L^2_\#}$ depends on the size of the box Y^h)

$$\|\Theta_d(V)w\|_{\ell^2} = \|\widehat{\mathcal{V}G}\|_{\ell^2} = \|\mathcal{V}G\|_{L^2_\#} \leq \|\mathcal{V}\|_{L^\infty_\#}\|G\|_{L^2_\#} \leq 2\|V\|_{L^\infty}\|w\|_{\ell^2},$$

which proves the boundedness of Θ_d.

Let us go back to (3.6). Relation (2.2) yields

$$\big(fg,\varphi\big)_{L^2_\#} = \sum_{n\in\mathbb{Z}^N} \widehat{fg}(n)\overline{\widehat{\varphi}(n)} = \big(f,\overline{g}\varphi\big)_{L^2_\#} = \sum_{n\in\mathbb{Z}^N} \widehat{f}(n)\overline{\widehat{\overline{g}\varphi}(n)}, \tag{3.7}$$

where φ is a regular enough test function (so that $\widehat{\varphi}\in\ell^1$ for instance). We can write, by using (2.3),

$$\overline{g}\varphi(y) = \sum_{n\in\mathbb{Z}^N} \widehat{\overline{g}\varphi}(n)e^{-i2\pi n\cdot y/L} = \sum_{n\in\mathbb{Z}^N}\left(\sum_{k\in\mathbb{Z}^N}\widehat{\varphi}(k)\widehat{\overline{g}}(n-k)\right)e^{-i2\pi n\cdot y/L},$$

which leads to

$$\widehat{\overline{g}\varphi}(n) = \sum_{k\in\mathbb{Z}^N}\widehat{\varphi}(k)\widehat{\overline{g}}(n-k) = \widehat{\varphi}*\widehat{\overline{g}}(n).$$

Notice that

$$\widehat{\overline{g}}(n) = L^{-N}\int_Y e^{+in\cdot z}\overline{g(z)}\,dz = \overline{L^{-N}\int_Y e^{-in\cdot z}g(z)\,dz} = \overline{\widehat{g}(-n)}.$$

Let us temporarily assume that f and g also are regular functions. We can perform the following computations from (3.7):

$$\begin{aligned}\big(fg,\varphi\big)_{L^2_\#} &= \sum_{n\in\mathbb{Z}^N}\widehat{f}(n)\left(\overline{\sum_{k\in\mathbb{Z}^N}\widehat{\varphi}(k)\overline{\widehat{g}(k-n)}}\right)\\ &= \sum_{n\in\mathbb{Z}^N}\widehat{f}(n)\left(\sum_{k\in\mathbb{Z}^N}\overline{\widehat{\varphi}(k)}\widehat{g}(k-n)\right)\\ &= \sum_{k\in\mathbb{Z}^N}\left(\sum_{n\in\mathbb{Z}^N}\widehat{f}(n)\widehat{g}(k-n)\right)\overline{\widehat{\varphi}(k)}\end{aligned}$$

by using Fubini's theorem (which requires the regularity of the functions). Identifying with the second term in (3.7), we are led to

$$\big(fg,\varphi\big)_{L^2_\#} = \sum_{k\in\mathbb{Z}^N}\Big(\widehat{f}*\widehat{g}\Big)(k)\overline{\widehat{\varphi}(k)} = \sum_{k\in\mathbb{Z}^N}\widehat{fg}(k)\overline{\widehat{\varphi}(k)}.$$

Since this holds for any test function φ, we actually have

$$\widehat{fg}(k) = \widehat{f}*\widehat{g}(k).$$

By a suitable regularization of f and g, but keeping φ regular, we justify this equality that holds in ℓ^∞ for f and g in $L^2_\#$.

Finally, let us investigate the other properties of the operator Θ_d. We recall that

$$a_n(x) = i\ h^N\ \int_{\mathbb{R}^N} D_V(x,y)e^{i2\pi hn\cdot y}\chi_{|y|\le R^h}\,dy.$$

Therefore, we get (with $z=-y$ and using $D_V(x,y)=-D_V(x,-y)$)

$$\begin{aligned}\overline{a_n} &= \overline{i\ h^N \int_{\mathbb{R}^N} D_V(x,y) e^{i2\pi hn\cdot y}\chi_{|y|\leq R^h}\,dy}\\ &= -i\ h^N \int_{\mathbb{R}^N} D_V(x,y) e^{-i2\pi hn\cdot y}\chi_{|y|\leq R^h}\,dy\\ &= i\ h^N \int_{\mathbb{R}^N} D_V(x,z) e^{i2\pi hn\cdot z}\chi_{|z|\leq R^h}\,dz = a_n\\ &= -i\ h^N \int_{\mathbb{R}^N} D_V(x,y) e^{i2\pi h(-n)\cdot y}\chi_{|y|\leq R^h}\,dy = -a_{-n},\end{aligned}$$

which proves that the coefficients a_n are real and even. In turn, we readily check that Θ_d is skew-symmetric. □

Let us conclude with some remarks. Since the discrete operator (3.3) is essentially obtained by truncating the potential, it will be convenient to discuss some properties of the approximate operator

$$\Theta^h(V)(f) = \mathcal{F}^{-1}_{y\to\xi}\big(D_V\ \chi_{|y|\leq R^h}\ \widehat{f}(y)\big)$$

defined for any $f\in L^2(\mathbb{R}^N)$. The proof of the following claim is quite obvious.

LEMMA 3.3. *The approximated operator Θ^h fulfills the following properties:*

(i) *if* $\mathrm{supp}(\widehat{f})\subset B(0,R)$ *with* $0<R\leq R^h$, *then* $\Theta^h(f)=\Theta(f)$;

(ii) $\Theta^h(V)(f)$ *converges to* $\Theta(f)$ *in* $L^2(\mathbb{R}^N)$ *as* $h\to 0$. *If, furthermore,* f *lies in the Sobolev space* $H^s(\mathbb{R}^N)$, $s>0$, *we get*

$$\|\Theta^h(V)(f)-\Theta(V)(f)\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} \leq \frac{4\|V\|^2_\infty}{(2\pi)^N}\ \frac{\|f\|^2_{H^s(\mathbb{R}^N)}}{(1+(R^h)^2)^s}.$$

As an example, it is maybe worth writing the discrete problem in a usual matrix form, when considering the mono-dimensional situation and restricting to $2M+1$ unknowns $(w_{-M},\dots,w_M)=W$. We get

$$\partial_t W + \Lambda\partial_x W = AW,$$

where Λ is the $(2M+1)\times(2M+1)$ diagonal matrix $\Lambda=\mathrm{diag}(\xi_{-M},\dots,\xi_M)$, while the coefficients of $A\in\mathcal{M}_{(2M+1)\times(2M+1)}$ are defined by

$$A_{nk} = a_{n-k}(x) = i\ h\ \int_{-R^h}^{+R^h} D_V(x,y) e^{i2\pi h(n-k)y}\,dy.$$

A short computation leads to

$$A_{nk} = 4h\ \int_{x-R^h/2}^{x+R^h/2} V(z)\sin(4\pi h(n-k)(x-z))\,dz.$$

We see easily on this formula that A is skew-symmetric.

3.2. Analysis of the discrete problem. Now it is quite easy to prove existence-uniqueness for the discrete problem (3.5). It is convenient here to interpret $\ell^2(\mathbb{Z}^N)=L^2(\mathbb{Z}^N,dn)$, where dn stands for the counting measure on $\mathbb{Z}^N$, and to introduce

$$L^2(\mathbb{R}^N_x\times\mathbb{Z}^N_n) = \Big\{w:(x,n)\in\mathbb{R}^N_x\times\mathbb{Z}^N_n \longmapsto w(x,n)\in\mathbb{C},$$

$$\text{such that } \int_{\mathbb{R}^N_x\times\mathbb{Z}^N_n}|w(x,n)|^2\,dx\,dn = \sum_{n\in\mathbb{Z}^N}\int_{\mathbb{R}^N_x}|w(x,n)|^2\,dx<\infty\Big\}.$$

Accordingly, in what follows, it is convenient to denote by $u(n)$ instead of u_n the nth term of a sequence u parametrized by $n \in \mathbb{Z}^N$. When no confusion can arise we will also drop the variables t, x.

THEOREM 3.4. *Let $w^0 \in L^2(\mathbb{R}^N_x \times \mathbb{Z}^N_n)$, and $V \in L^\infty(\mathbb{R}^N)$. The problem (3.5) has a unique mild solution $w \in C^0([0,T]; L^2(\mathbb{R}^N_x \times \mathbb{Z}^N_n))$. Furthermore, if the initial data w^0 is real valued (i.e., $w^0(x,n) = \overline{w^0(x,n)}$ for all $n \in \mathbb{Z}^N$ and a.a. $x \in \mathbb{R}^N$), then the solution w remains real valued and the following relation*

$$\|w(t)\|_{L^2(\mathbb{R}^N \times \mathbb{Z}^N)} = \|w^0\|_{L^2(\mathbb{R}^N \times \mathbb{Z}^N)} \tag{3.8}$$

holds.

Proof. The proof is quite classical, and the argument is the same as for the continuous problem in [3], [9]. The key point is to interpret Θ_d as a bounded perturbation of the semigroup generator $-\xi_n \cdot \nabla_x$. Integrating along the characteristic lines $x + t\xi_n$ leads to

$$w(t,x,n) = w^0(x - t\xi_n, n) + \int_0^t \Theta_d(V) w(\tau, x - (t-\tau)\xi_n, n)\, d\tau.$$

Then, we show the existence-uniqueness of the solution by a classical contraction argument.

Since the $a(n)$'s are real, now we deduce that

$$\overline{\Theta_d(w)(n)} = \overline{\sum_{k \in \mathbb{Z}^N} a(n-k) w(k)} = \sum_{k \in \mathbb{Z}^N} a(n-k) \overline{w(k)} = \Theta_d(\overline{w})(n).$$

It follows that $\overline{w(n)}$ satisfies the same equation as $w(n)$; hence, if $w^0(n) = \overline{w^0(n)}$, we still have $w(n) = \overline{w(n)}$. In other words, w is real valued.

Finally, we are left with the task of proving the relation (3.8) in $L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^N \times \mathbb{Z}^N,\, dx \otimes dn))$. Indeed, multiplying (3.5) by $\overline{w^h_n}$ and integrating yield

$$\|w^h(t)\|^2_{L^2(\mathbb{R}^N \times \mathbb{Z}^N)} + \int_0^t \int \big(\Theta_d(w), w\big)_{\ell^2}(s,x)\, dx\, ds = \|w^h(0)\|^2_{L^2(\mathbb{R}^N \times \mathbb{Z}^N)}.$$

Since $\big(\Theta_d(w), w\big)_{\ell^2} \in i\mathbb{R}$, we immediately obtain (3.8). ∎

4. Approximation and convergence analysis. In this section, we investigate the limit of vanishing mesh size $h \to 0$. We thus show that the discrete problem approaches the continuous one.

4.1. Construction of the approximation. In the previous section, we have obtained a sequence, parametrized by $t \in \mathbb{R}^+$ and $x \in \mathbb{R}^N$, $\big(w(t,x,n)\big)_{n \in \mathbb{Z}^N} \in \ell^2$, a solution of (3.5). Of course, this sequence depends on the parameter h through the truncation acting on the potential and the definition of the discrete velocities $\xi_n = 2\pi n h$. We emphasize this dependence from now on by denoting the solution $w^h(t,x,n)$.

According to the strategy described in section 2, let us introduce now for $(t,x,\xi) \in \mathbb{R}^+ \times \mathbb{R}^N \times \mathbb{R}^N$ the function

$$f^h(t,x,\xi) = \mathcal{F}^{-1}_{y \to \xi} \left(\sum_{n \in \mathbb{Z}^N} w^h(t,x,n)\; e^{-i2\pi h y \cdot n}\; \chi_{|y| \le R^h} \right).$$

It is maybe more explicit to write, as in Lemma 2.1,

$$f^h(t,x,\xi) = \sum_{n\in\mathbb{Z}^N} \left(w^h(t,x,n) \prod_{j=1}^N \frac{\sin(R^h(\xi_j - 2\pi h n_j))}{\pi(\xi_j - 2\pi h n_j)} \right).$$

Since we naturally choose $R^h \sim (2h)^{-1}$ for h small, this indicates that $f(\xi_n)$ is approximated by $(2\pi h)^{-N} w_n^h$.

Our final aim is to show that f^h approaches the solution f of (1.1) as the mesh size h goes to 0. Notice that here f is approximated by a formula looking like $\sum_{n\in\mathbb{Z}^N} w^h(t,x,n) S^h(\xi - \xi_n)$. In comparison to [13], [14], instead of approaching f by periodic functions, we use an expansion on cardinal sinus functions. In [13], [14] difficulties arise since we use a basis of functions that are not elements of the same space as the exact solution; in turn, our proof of convergence becomes very simple.

To this end, we naturally have to prepare in a suitable way the initial data $w^{h,0}$ for the discrete problem (3.5). They obtained by regularizing the initial data f^0 of the continuous problem. Precisely, we set, still using the arguments of section 2,

$$\begin{cases} f^{h,0}(x,\xi) = \mathcal{F}^{-1}_{y\to\xi}\big(\widehat{f^0}(x,y)\ \chi_{|y|\le R^h}\big), \\ G^{h,0}(x,y) = \widehat{f^0}(x,y) \quad \text{on } B(0,R^h), \qquad G^{h,0}(x,y) = 0 \quad \text{on } Y^h\backslash B(0,R^h), \\ G^{h,0}(x,y) = \quad \text{is } Y^h\text{-periodic}, \\ w^{h,0}(x,n) = \widehat{G^{h,0}(x,\cdot)}(n) = f^{h,0}(x,2\pi n h)(2\pi h)^N, \end{cases}$$

where f^0 is real valued and belongs to $L^2(\mathbb{R}^N\times\mathbb{R}^N)$. This construction leads to the following behavior of the discrete solution with respect to h.

LEMMA 4.1. *One has the estimate* $\|w^{h,0}\|^2_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)} \le (2\pi h)^N \ \|f^0\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}$. *Consequently, one deduces that* $\|w^h(t)\|^2_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)} \le (2\pi h)^N \ \|f^0\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}$.

Proof. The estimate on the solution w^h is an immediate consequence of (3.8) combined with the estimate on the initial data. By construction, one has

$$\begin{aligned} \|w^{h,0}\|^2_{\ell^2} &= \|\widehat{G^{h,0}}\|^2_{\ell^2} = \|G^{h,0}\|^2_{L^2_\#} \\ &= h^N \int_{Y^h} |G^{h,0}(y)|^2\,dy = h^N \int_{\mathbb{R}^N} |\widehat{f^{h,0}}(y)|^2\,dy \\ &= h^N \int_{\mathbb{R}^N} |\widehat{f^0}(y)\chi_{|y|\le R^h}|^2\,dy \le h^N\ (2\pi)^N\ \|f^0\|^2_{L^2(\mathbb{R}^N)}, \end{aligned}$$

which establishes Lemma 4.1. ∎

One deduces that f^h is a bounded sequence in the natural functional space.

COROLLARY 4.2. *The sequence* $\big(f^h\big)_{h>0}$ *is bounded in* $L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^N\times\mathbb{R}^N))$. *Furthermore,* f^h *is real valued.*

Proof. This is a consequence of Lemma 2.1 and Lemma 4.1. It is convenient to introduce the Y^h-periodic function G^h defined by the coefficients w_n^h:

$$G^h(t,x,y) = \sum_{n\in\mathbb{Z}^N} w^h(t,x,n)\ e^{-i2\pi h y\cdot n} \quad \in L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^N; L^2_\#(Y^h))).$$

We have (omiting t, x variables for the sake of clarity)

$$\begin{aligned}
\|f^h\|^2_{L^2(\mathbb{R}^N_\xi)} &= (2\pi)^{-N}\ \|\widehat{f^h}\|^2_{L^2(\mathbb{R}^N_y)} \\
&= (2\pi)^{-N}\ \int_{\mathbb{R}^N} \left| \sum_{n\in\mathbb{Z}^N} w^h(n)\ e^{-i2\pi hy\cdot n} \right|^2 \ \chi_{|y|\le R^h}\, dy \\
&= (2\pi)^{-N}\ \int_{B(0,R^h)} \left| \sum_{n\in\mathbb{Z}^N} w^h(n)\ e^{-i2\pi hy\cdot n} \right|^2 \ dy \\
&\le (2\pi h)^{-N}\ h^N \int_{Y^h} \left| \sum_{n\in\mathbb{Z}^N} w^h(n)\ e^{-i2\pi hy\cdot n} \right|^2 \ dy \\
&\le (2\pi h)^{-N}\ h^N \int_{Y^h} |G^h(y)|^2\, dy \\
&\le (2\pi h)^{-N}\ \|G^h\|^2_{L^2_\#(Y^h)} = (2\pi h)^{-N}\ \|w^h\|^2_{\ell^2} \le \|f^0\|^2_{L^2(\mathbb{R}^N)},
\end{aligned}$$

which proves the asserted bound.

It remains to prove that $f^h \in \mathbb{R}$. Conjugating $\widehat{f^h}$ we get

$$\begin{aligned}
\overline{\widehat{f^h}(y)} &= \sum_{n\in\mathbb{Z}^N} \overline{w^h(n)}\ e^{+i2\pi hy\cdot n} \chi_{|y|\le R^h} \\
&= \sum_{n\in\mathbb{Z}^N} w^h(n)\ e^{-i2\pi h(-y)\cdot n} \chi_{|-y|\le R^h} = \widehat{f^h}(-y)
\end{aligned}$$

since, by Theorem 3.4, the w^h_n's are real valued. This ends the proof. □

Now let us show that f^h is not far from satisfying (1.1).

LEMMA 4.3. *The function f^h satisfies a approximate form of* (1.1) *in the sense that*

$$\langle (\partial_t + \xi\cdot\nabla_x) f^h - \Theta(V)(f^h), \varphi\rangle_{\mathcal{S}';\mathcal{S}} = 0$$

holds for all φ in $\mathcal{S}((0,T)\times\mathbb{R}^N\times\mathbb{R}^N)$ with $\mathrm{supp}(\widehat{\varphi}) \subset (0,T)\times\mathbb{R}^N\times B(0,R)$, $0 < R < R^h$.

Proof. Let us set

$$T^h(t,x,n) = (\partial_t + \xi_n\cdot\nabla_x) w^h(t,x,n) - \Theta_d(w^h(t,x,\cdot))(n).$$

We have to compute the action on φ of

$$\mathcal{F}^{-1}_{y\to\xi}\left(\sum_{n\in\mathbb{Z}^N} e^{-i2\pi hy\cdot n}\ T^h(t,x,n)\ \chi_{|y|\le R^h} \right).$$

Obviously, there is no difficulty with the time derivative, which commutes with the other operations. It leads to $\partial_t f^h$. Let us consider the Vlasov term. Since, by its definition, f^h has a Fourier transform compactly supported (with respect to the variable y) in $B(0,R^h)$, we have, by Lemma 3.3(i), $\Theta^h(f^h) = \Theta(f^h)$, with

$$\widehat{\Theta(f^h)}(t,x,y) = i\ D_V(x,y)\ \chi_{|y|\le R^h} \sum_{n\in\mathbb{Z}^N} w^h(t,x,n)\ e^{-i2\pi hy\cdot n}.$$

Let us denote this quantity by $\gamma^h(y)$. It corresponds to

$$\left\{\begin{array}{l}\Gamma^h(t,x,y)=\gamma^h(y)=i\ \mathcal{V}^h(x,y)\ G^h(t,x,y)\quad \text{on } y\in B(0,R^h),\\ \Gamma^h(t,x,y)=0\quad \text{on } y\in Y^h\backslash B(0,R^h),\\ \Gamma^h(t,x,y)\quad \text{is } Y^h\text{-periodic (with respect to } y).\end{array}\right.$$

The corresponding Fourier coefficients, therefore, are given by

$$\widehat{\Gamma^h}(n)=i\ \widehat{\mathcal{V}^h\ G^h}(n)=i\ \widehat{\mathcal{V}^h}*\widehat{G^h}(n)=a^h*w^h(n)=\Theta_d(w^h)(n)$$

by using the definitions of $a^h,\mathcal{V}^h$ and w^h,G^h.

Applying Lemma 2.1 yields the following equality:

$$\begin{aligned}\Theta(f^h)=\mathcal{F}^{-1}_{y\to\xi}(\gamma^h)&=\mathcal{F}^{-1}_{y\to\xi}\left(\sum_{n\in\mathbb{Z}^N}\widehat{\Gamma^h}(t,x,n)\ e^{-i2\pi hy\cdot n}\ \chi_{|y|\le R^h}\right)\\&=\mathcal{F}^{-1}_{y\to\xi}\left(\sum_{n\in\mathbb{Z}^N}\Theta_d(w^h)(t,x,n)\ e^{-i2\pi hy\cdot n}\ \chi_{|y|\le R^h}\right),\end{aligned}$$

which is exactly what we need when treating the right-hand side of (3.5). It remains to deal with the convective term. Difficulties could arise due to the truncation $\chi_{|y|\le R^h}$ in the Fourier variable; however, we have chosen a test function φ that does not feel the action of this operation. Set

$$\begin{aligned}&C^h(t,x,y)\\&=\sum_{n\in\mathbb{Z}^N}e^{-i2\pi hy\cdot n}\ \xi_n\cdot\nabla_x w^h(t,x,n)\ \chi_{|y|\le R^h}\\&=(-i)^{-1}\sum_{n\in\mathbb{Z}^N}\nabla_y\nabla_x\big(e^{-i2\pi hy\cdot n}\ w^h(t,x,n)\big)\ \chi_{|y|\le R^h}\in\mathcal{S}'(\mathbb{R}^+\times\mathbb{R}^N\times\mathbb{R}^N).\end{aligned}$$

We compute, by using the exchange formula,

$$\begin{aligned}&\langle\mathcal{F}^{-1}_{y\to\xi}(C^h(t,x,\cdot))(\xi),\varphi\rangle_{\mathcal{S}',\mathcal{S}}\\&=(2\pi)^{-N}\langle\overline{\mathcal{F}(\overline{C^h})},\varphi\rangle\\&=(2\pi)^{-N}\overline{\left\langle\overline{C^h},\mathcal{F}(\overline{\varphi})\right\rangle}\\&=(2\pi)^{-N}\overline{\left\langle\overline{(-i)^{-1}\nabla_y\nabla_x\big(e^{-i2\pi hy\cdot n}\ w^h(t,x,n)\big)},\mathcal{F}(\overline{\varphi})\right\rangle}\end{aligned}$$

since $\widehat{\varphi}\chi_{|y|\le R^h}=\widehat{\varphi}$ by using the support assumption. It follows that

$$\begin{aligned}&\langle\mathcal{F}^{-1}_{y\to\xi}(C^h(t,x,\cdot))(\xi),\varphi\rangle_{\mathcal{S}',\mathcal{S}}\\&=(2\pi)^{-N}\overline{\int\overline{(-i)^{-1}\sum_{n\in\mathbb{Z}^N}e^{-i2\pi hy\cdot n}\ w^h(t,x,n)}\ \nabla_y\nabla_x\mathcal{F}(\overline{\varphi})\,dy\,dx\,dt}.\end{aligned}$$

However, we have

$$\nabla_y\mathcal{F}(\overline{\varphi})=\nabla_y\left(\int_{\mathbb{R}^N}e^{-iy\cdot\xi}\overline{\varphi}(\xi)\,d\xi\right)=-i\mathcal{F}(\xi\overline{\varphi}),$$

which still has its support in $B(0,R^h)$ (with respect to the variable y). Hence, we obtain

$$\begin{aligned}&\langle \mathcal{F}^{-1}_{y\to\xi}(C^h(t,x,\cdot))(\xi),\varphi\rangle_{\mathcal{S}',\mathcal{S}}\\ &= (2\pi)^{-N}\overline{\int \overline{(-i)^{-1}\sum_{n\in\mathbb{Z}^N} e^{-i2\pi hy\cdot n}\ w^h(t,x,n)}\ (-i)\mathcal{F}\big(\xi\cdot\nabla_x\overline{\varphi}\big)\,dy\,dx\,dt}.\end{aligned}$$

We go back with the exchange formula to

$$\begin{aligned}&\langle \mathcal{F}^{-1}_{y\to\xi}(C^h(t,x,\cdot))(\xi),\varphi\rangle_{\mathcal{S}',\mathcal{S}}\\ &= -(2\pi)^{-N}\overline{\left\langle \mathcal{F}_{y\to\xi}\left(\overline{\sum_{n\in\mathbb{Z}^N} e^{-i2\pi hy\cdot n}w^h(t,x,n)\chi_{|y|\le R^h}}\right),\xi\cdot\nabla_x\overline{\varphi}\right\rangle}\\ &= \langle \xi\cdot\nabla_x f^h,\varphi\rangle.\end{aligned}$$

This ends the proof of Lemma 4.3. ∎

4.2. Convergence. Now we are able to establish the convergence of our scheme, which is shown to have spectral accuracy.

THEOREM 4.4. *Assume that f^0 belongs to $H^1(\mathbb{R}^N\times\mathbb{R}^N)$, and $V\in L^\infty(\mathbb{R}^N)$. Let $0<T<\infty$. Let $f\in C^0(\mathbb{R}^+;L^2(\mathbb{R}^N\times\mathbb{R}^N))\cap L^\infty(0,T;H^1(\mathbb{R}^N\times\mathbb{R}^N))$ be the associated solution of* (1.1). *Let $R^h\le(2h)^{-1}$, with $2R^h\sim 1/h$ as h goes to 0. Then, we have on the time interval $(0,T)$*

$$\|(f-f^h)(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\le C_T\ h,$$

where C_T depends on T and on these H^1 norms.

If, furthermore, $f^0\in L^2(\mathbb{R}^N_x;H^s(\mathbb{R}^N_\xi))$ with $f\in L^\infty(0,T;L^2(\mathbb{R}^N_x;H^s(\mathbb{R}^N_\xi)))$ and $\nabla_x f\in L^2((0,T)\times\mathbb{R}^N_x;H^{s-1}(\mathbb{R}^N_\xi))$ for some $s>0$, then the method becomes of order s since we have $\|(f-f^h)(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\le C_T\ h^s$.

Remark 2. In the statement, the Sobolev regularity of the solution is assumed. Then, the boundedness of the potential suffices to justify the spectral accuracy. However, regularity of the solution depends certainly on the regularity of the potential. For instance, one proves readily the H^1 regularity by assuming $V\in W^{1,\infty}(\mathbb{R}^N)$.

Proof. It would be tempting, in order to derive such an estimate, to multiply the equation satisfied by $f-f^h$ by $\overline{f-f^h}$. However, this function is not an admissible test function in the sense of Lemma 4.3 since its Fourier transform is not compactly supported. Then, let us introduce a function ζ that fulfills

$$\begin{cases}\zeta(y)=1 \quad \text{on } B(0,1/2), \qquad \zeta(y)=0 \quad \text{on } \mathbb{R}^N\backslash B(0,1),\\ 0\le\zeta(y)\le 1,\\ \zeta\in C^1(\mathbb{R}^N).\end{cases}$$

Set $\zeta^h(y)=\zeta(y/R^h)$, and $t^h(t,x,\xi)=\mathcal{F}^{-1}_{y\to\xi}(\widehat{f}(t,x,y)\ \zeta^h(y))$. Mainly, the rate of convergence is determined by the obvious estimate

$$\|f-t^h\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}=(2\pi)^{-N}\|\widehat{f}(1-\zeta^h)\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\le\frac{\|f\|^2_{L^2(\mathbb{R}^N_x;H^s(\mathbb{R}^N_\xi))}}{(2\pi)^N(1+(R^h/2)^2)^s}.$$

Furthermore, let us look at the equation satisfied by t^h: we shall show that it satisfies (1.1) up to an error term due to the truncation, which can be controlled in $\mathcal{O}(R^h)$ thanks to the H^1 regularity of the solution. From (1.1), we obtain

$$\partial_t t^h = \mathcal{F}^{-1}(\partial_t \widehat{f}\ \zeta^h) = \mathcal{F}^{-1}(\widehat{\Theta(f)}\ \zeta^h) - \mathcal{F}^{-1}(\widehat{\xi\cdot\nabla_x f}\ \zeta^h).$$

The Vlasov term simply gives

$$\begin{aligned}\mathcal{F}^{-1}(\widehat{\Theta(f)}\ \zeta^h) &= \mathcal{F}^{-1}(D_V\ \widehat{f}\ \zeta^h)\\ &= \mathcal{F}^{-1}(D_V\ \mathcal{F}\mathcal{F}^{-1}(\widehat{f}\ \zeta^h)) = \mathcal{F}^{-1}(D_V\ \mathcal{F}(t^h)) = \Theta(t^h).\end{aligned}$$

However, error terms are produced by the convective term. Indeed, we have

$$\begin{aligned}\mathcal{F}^{-1}(\widehat{\xi\cdot\nabla_x f}\ \zeta^h) &= \int_{\mathbb{R}^N} e^{iy\cdot\xi}\ \widehat{\xi\cdot\nabla_x f}(y)\ \zeta^h(y)\,dy/(2\pi)^N\\ &= \int_{\mathbb{R}^N_y} e^{iy\cdot\xi}\left(\int_{\mathbb{R}^N_\eta} e^{-iy\cdot\eta}\ \eta\cdot\nabla_x f(\eta)\,d\eta\right)\zeta^h(y)\,dy/(2\pi)^N\\ &= \int_{\mathbb{R}^N_\eta} \eta\cdot\nabla_x f(\eta)\left(\int_{\mathbb{R}^N_y} e^{iy\cdot(\xi-\eta)}\zeta^h(y)\,dy\right)\,d\eta/(2\pi)^N\\ &= \int_{\mathbb{R}^N_\eta} \eta\cdot\nabla_x f(\eta)\widehat{\zeta^h}(\eta-\xi)\,d\eta/(2\pi)^N\\ &= \left(\eta\cdot\nabla_x f(\eta) * \check{\widehat{\zeta^h}}(\eta)\right)(\xi)/(2\pi)^N.\end{aligned}$$

Let us split this expression as follows:

$$\mathcal{F}^{-1}(\widehat{\xi\cdot\nabla_x f}\ \zeta^h) = \xi\cdot\nabla_x t^h + E^h,$$

where

$$E^h = \left(\eta\cdot\nabla_x f * \check{\widehat{\zeta^h}}(\eta)\right)(\xi)/(2\pi)^N - \xi\cdot\nabla_x\big(\mathcal{F}^{-1}(\widehat{f}\ \zeta^h)\big).$$

Since

$$\xi\cdot\nabla_x\big(\mathcal{F}^{-1}(\widehat{f}\ \zeta^h)\big) = \int_{\mathbb{R}^N_y} e^{iy\cdot\xi}\zeta^h(y)\xi\cdot\left(\int_{\mathbb{R}^N_\eta} e^{-iy\cdot\eta}\nabla_x f(\eta)\,d\eta\right)\,dy/(2\pi)^N,$$

we get

$$\begin{aligned}E^h &= \int_{\mathbb{R}^N}\int_{\mathbb{R}^N} e^{iy\cdot(\xi-\eta)}(\eta-\xi)\cdot\nabla_x f(\eta)\ \zeta^h(y)\,d\eta\,dy/(2\pi)^N\\ &= \int_{\mathbb{R}^N_\eta}\nabla_x f(\eta)\left(\int_{\mathbb{R}^N_y}(\eta-\xi)e^{iy\cdot(\xi-\eta)}\zeta^h(y)\,dy\right)\,d\eta/(2\pi)^N\\ &= \int_{\mathbb{R}^N_\eta}\nabla_x f(\eta)\left(\frac{1}{i}\int_{\mathbb{R}^N_y} e^{iy\cdot(\xi-\eta)}\nabla_y(\zeta^h(y))\,dy/(2\pi)^N\right)\,d\eta\\ &= \big(\nabla_x f * \mathcal{F}^{-1}_{y\to\xi}(\nabla_y\zeta^h)/i\big)(\xi).\end{aligned}$$

This produces the estimate

$$\begin{aligned}\|E^h\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} &= (2\pi)^{-N}\|\widehat{\nabla_x f}\ \nabla_y(\zeta^h)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\\ &\le \|\nabla_x f\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\ \|\nabla_y(\zeta^h)\|_{L^\infty(\mathbb{R}^N)}\\ &\le \|\nabla_x f\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\ \|\nabla_y\zeta\|_{L^\infty(\mathbb{R}^N)}/R^h.\end{aligned}$$

Therefore, $e^h = t^h - f^h$ has a compactly supported Fourier transform in $B(0,R^h)$ and satisfies, in the sense of Lemma 4.3,

$$(\partial_t + \xi\cdot\nabla_x - \Theta^h)(e^h) = E^h.$$

Standard manipulations lead to the estimate

$$\|e^h(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} \le \|e^h(0)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} + \int_0^t \|E^h\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} \le C_T/R^h,$$

where C_T depends on T, $\|f^0\|_{H^1}$, and $\|f\|_{H^1}$. Now, we can estimate $\|f-f^h\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}$ $\le \|e^h(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} + \|f-t^h\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} \le C_T/R^h$, which concludes the proof.

Order s information on f leads to the following slight modification when estimating E^h:

$$\begin{aligned}&\|E^h\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\\ &\quad= (2\pi)^{-N}\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} |\widehat{\nabla_x f}(y)|^2\ |\nabla_y(\zeta^h)|^2\,dy\,dx\\ &\quad\le (2\pi)^{-N}\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} |\widehat{\nabla_x f}(y)|\ \chi_{|y|\ge R^h/2}\ \frac{1}{(R^h)^2}\ \|\nabla_y\zeta\|^2_{L^\infty(\mathbb{R}^N)}\,dy\,dx\\ &\quad\le \|\nabla_x f\|^2_{L^2(\mathbb{R}^N_x;H^{s-1}(\mathbb{R}^N_\xi))}\ \|\nabla_y\zeta\|^2_{L^\infty(\mathbb{R}^N)}\ \frac{2\ (2\pi)^{-N}}{(1+(R^h)^2)^{s-1}\ (R^h)^2}.\end{aligned}$$

The announced rate of convergence follows easily. ∎

5. Semiclassical limit. Let us now take into account the (scaled) Planck constant $\varepsilon > 0$. Accordingly, the Wigner operator reads as

$$\begin{aligned}\Theta^\varepsilon(V)f &= \frac{i}{\varepsilon}\mathcal{F}^{-1}_{y\to\xi}\Big(\big(V(x+\varepsilon y/2) - V(x-\varepsilon y/2)\big)\ \widehat{f}(y)\Big)\\ &= \frac{i}{(2\pi)^N}\int_{\mathbb{R}^N}\frac{V(x+\varepsilon y/2)-V(x-\varepsilon y/2)}{\varepsilon}\,\widehat{f}(y)\ e^{iy\cdot\xi}\,dy.\end{aligned}$$

It is a well-established fact that the corresponding sequence $\big(f^\varepsilon\big)_{\varepsilon>0}$ of solutions of the Wigner equation

$$\partial_t f^\varepsilon + \xi\cdot\nabla_x f^\varepsilon = \Theta^\varepsilon(V)(f^\varepsilon)$$

converges, in some weak sense, to a solution F of the Vlasov equation

$$\partial_t F + \xi\cdot\nabla_x F = \nabla_x V\cdot\nabla_\xi F \tag{5.1}$$

as $\varepsilon\to 0$. We refer to Lions and Paul [11] for precise statements and rigorous proofs; this can be formally understood from the following facts:

$$\begin{cases} i\ \dfrac{V(x+\varepsilon y/2)-V(x-\varepsilon y/2)}{\varepsilon} \longrightarrow i\nabla_x V\cdot y \quad \text{as } \varepsilon\to 0,\\ \widehat{iy} = \nabla\delta_{\xi=0}.\end{cases}$$

Therefore, we would naturally expect that our discrete scheme reproduces this behavior. Let us first set up some definitions.

Recall that the discrete operator is defined, see (3.3), by

$$\left\{\begin{array}{l} a_n^{\varepsilon,h} = i\ h^N \displaystyle\int_{\mathbb{R}^N} \frac{V(x+\varepsilon y/2)-V(x-\varepsilon y/2)}{\varepsilon}\,\chi_{|y|\leq R^h}(y)\ e^{i2\pi hn\cdot y}\,dy, \\ \Big(\Theta_d^{\varepsilon,h}(V)(w)\Big)_n = (a^{\varepsilon,h}*w)_n = i\widehat{\mathcal{V}^{\varepsilon,h}G^h}(n), \end{array}\right. \tag{5.2}$$

where $w(x,n)$ belongs to $L^2(\mathbb{R}^N\times\mathbb{Z}^N)$, and $\mathcal{V}^{\varepsilon,h}(x,\cdot)$ and $G^h(x,\cdot)$ are the Y^h-periodic functions defined on the cell by

$$\left\{\begin{array}{l} \mathcal{V}^{\varepsilon,h}(x,y) = \dfrac{V(x+\varepsilon y/2)-V(x-\varepsilon y/2)}{\varepsilon}\,\chi_{|y|\leq R^h}(y), \\ G^h(x,y) = \displaystyle\sum_{n\in\mathbb{Z}^N} w(x,n)\ e^{-i2\pi hn\cdot y}. \end{array}\right.$$

For small ε, we expect that $\Theta_d^{\varepsilon,h}$ looks like the following discrete operator:

$$\Big(\mathbb{V}^h(V)(w)\Big)_n = (u^h*w)_n = i\widehat{\mathcal{Y}^hG^h}(n), \tag{5.3}$$

where $\mathcal{Y}^h(x,\cdot)$ is the Y^h-periodic function given by

$$\mathcal{Y}^h(x,y) = \nabla_x V(x)\cdot y\ \chi_{|y|\leq R^h}(y)$$

for $x\in\mathbb{R}^N$ and $y\in Y^h$. This discrete operator corresponds to our interpretation of the Vlasov term $\nabla_x V\cdot\nabla_\xi g$ when reasoning as in section 2 for a function g with $\widehat{g}$ compactly supported in $B(0,R)$. We remark that

$$u_n^h = i\int_{\mathbb{R}^N}\nabla_x V(x)\cdot y\ \chi_{|y|\leq R^h}(y)\ e^{i2\pi hn\cdot y}\,dy = \overline{u_n^h} = -u_{-n}^h$$

so that the coefficients are real and even; in turn, $\mathbb{V}^h$ is a skew-symmetric operator of $\mathcal{L}(\ell^2)$ for $h>0$ fixed, and $V\in W^{1,\infty}(\mathbb{R}^N)$. Hence, reproducing the arguments of the proof of Theorem 3.4, we show that the corresponding discrete problem is well-posed.

THEOREM 5.1. *Let $v^{h,0}\in L^2(\mathbb{R}^N_x\times\mathbb{Z}^N_n)$ and $V\in W^{1,\infty}(\mathbb{R}^N)$. The problem*

$$\left\{\begin{array}{ll} \partial_t v_n + \xi_n\cdot\nabla_x v_n = \mathbb{V}^h(V)(v)_n & \text{in } \mathbb{R}^+_t\times\mathbb{R}^N_x\times\mathbb{Z}^N_n, \\ v_{|t=0} = v^0 & \end{array}\right. \tag{5.4}$$

has a unique mild solution $v^h\in C^0([0,T];L^2(\mathbb{R}^N_x\times\mathbb{Z}^N))$. Furthermore, if the initial data $v^{h,0}$ are real valued (i.e., $v^{h,0}(x,n)=\overline{v^{h,0}(x,n)}$ for all $n\in\mathbb{Z}^N$ and a.a. $x\in\mathbb{R}^N$), then the solution v^h remains real valued, and the relation $\|v^h(t)\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)} = \|v^{h,0}\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}$ holds.

Let us denote, according to section 4,

$$\left\{\begin{array}{l} f^{\varepsilon,h}(t,x,\xi) = \mathcal{F}^{-1}_{y\to\xi}\Big(\displaystyle\sum_{n\in\mathbb{Z}^N} w^{\varepsilon,h}(t,x,n)\ e^{-i2\pi hy\cdot n}\ \chi_{|y|\leq R^h}\Big), \\ F^h(t,x,\xi) = \mathcal{F}^{-1}_{y\to\xi}\Big(\displaystyle\sum_{n\in\mathbb{Z}^N} v^h(t,x,n)\ e^{-i2\pi hy\cdot n}\ \chi_{|y|\leq R^h}\Big), \end{array}\right.$$

where $w^{\varepsilon,h}$ and v^h are the solutions of (3.5) and (5.4), respectively, obtained from well-prepared initial data as in Lemma 4.1. We naturally assume in this construction that $f^{\varepsilon,0}(x,\xi)$, which defines $w^{\varepsilon,h}_{t=0}$, tends (strongly in $L^2(\mathbb{R}^N\times\mathbb{R}^N)$) to $F^0(x,\xi)$, associated with $v^h_{t=0}$, when $\varepsilon\to 0$. Of course, we wish to justify the following diagram:

$$\begin{array}{ccc} w^{\varepsilon,h},\ f^{\varepsilon,h} & \xrightarrow{\quad h\to 0\quad} & f^{\varepsilon} \\ \text{discrete Wigner} & & \text{continuous Wigner} \\ \Big\downarrow \varepsilon\to 0 & & \Big\downarrow \varepsilon\to 0,\ [11] \\ v^h,\ F^h & \xrightarrow[\quad h\to 0\quad]{} & F \\ \text{discrete Vlasov} & & \text{continuous Vlasov} \end{array}$$

Obviously, Lemma 4.1 adapts to the scaling.

LEMMA 5.2. *We have*

$$\|w^{\varepsilon,h}(t)\|^2_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)} \le (2\pi h)^N\ \|f^{\varepsilon,0}\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)} \le (2\pi h)^N\ C_0,$$
$$\|v^h(t)\|^2_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)} \le (2\pi h)^N\ \|F^0\|^2_{L^2(\mathbb{R}^N\times\mathbb{R}^N)},$$

and the (real valued) sequences $\left(F^h\right)_{h>0}$, $\left(f^{\varepsilon,h}\right)_{\varepsilon>0,h>0}$ *are bounded in* $L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^N\times\mathbb{R}^N))$, *uniformly with respect to* ε *and* h.

Next, the strategy developed in the proofs of Lemma 4.3 and Theorem 4.4 applies when dealing with v^h, F^h, and we are led to the following statement.

THEOREM 5.3. *Let* $V\in W^{1,\infty}(\mathbb{R}^N)$. *The function* F^h *satisfies an approximate form of* (5.1) *in the sense that*

$$\langle(\partial_t+\xi\cdot\nabla_x)F^h-\nabla_x V\cdot\nabla_\xi F^h,\varphi\rangle_{\mathcal{S}';\mathcal{S}}=0$$

holds for all φ *in* $\mathcal{S}((0,T)\times\mathbb{R}^N\times\mathbb{R}^N)$ *with* $\operatorname{supp}(\widehat{\varphi})\subset(0,T)\times\mathbb{R}^N\times B(0,R)$, $0<R<R^h$.

Assume that F^0 *belongs to* $H^1(\mathbb{R}^N\times\mathbb{R}^N)$. *Let* $0<T<\infty$. *Let* $F\in C^0(\mathbb{R}^+;L^2(\mathbb{R}^N\times\mathbb{R}^N))\cap L^\infty(0,T;H^1(\mathbb{R}^N\times\mathbb{R}^N))$ *be the associated solution of* (5.1). *Let* $R^h\le(2h)^{-1}$, *with* $2R^h\sim 1/h$ *as* h *goes to 0. Then, we have on the time interval* $(0,T)$,

$$\|(F-F^h)(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\le C_T\ h,$$

where C_T *depends on* T *and on these* H^1 *norms.*

If, furthermore, $F^0\in L^2(\mathbb{R}^N_x;H^s(\mathbb{R}^N_\xi))$ *with* $F\in L^\infty(0,T;L^2(\mathbb{R}^N_x;H^s(\mathbb{R}^N_\xi)))$, *and* $\nabla_x F\in L^2((0,T)\times\mathbb{R}^N_x;H^{s-1}(\mathbb{R}^N_\xi))$ *for some* $s>0$, *then the method becomes of order* s *since we have* $\|(F-F^h)(t)\|_{L^2(\mathbb{R}^N\times\mathbb{R}^N)}\le C_T\ h^s$.

We are left with the task of studying the behavior of $w^{\varepsilon,h}$ as $\varepsilon\to 0$. The convergence to the solution of the discrete Vlasov equation will be a consequence of the following claim.

PROPOSITION 5.4. *Let* $V\in C^1\cap W^{1,\infty}(\mathbb{R}^N)$. *The sequence of operators* $\Theta^{\varepsilon,h}_d$ *converges strongly to* $\mathbb{V}^h$: *for any* $h>0$ *and* $w\in L^2(\mathbb{R}^N\times\mathbb{Z}^N)$, *we have*

$$\lim_{\varepsilon\to 0}\|\Theta^{\varepsilon,h}_d(w)-\mathbb{V}^h(w)\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}=0.$$

Proof. We simply remark that

$$\mathcal{V}^{\varepsilon,h}(x,y)-\mathcal{Y}^h(x,y)=\int_0^1\Big(\nabla_x V(x+\varepsilon y(\theta-1/2))-\nabla_x V(x)\Big)\cdot y\ \chi_{|y|\le R^h}(y)\,d\theta.$$

Since $\nabla_x V$ is continuous and bounded, we can apply the Lebesgue theorem which yields for $h>0$, $x\in\mathbb{R}^N$, $y\in Y^h$,

$$\lim_{\varepsilon\to 0}\big(\mathcal{V}^{\varepsilon,h}(x,y)-\mathcal{Y}^h(x,y)\big)=0,$$

with the domination $|\mathcal{V}^{\varepsilon,h}(x,y)-\mathcal{Y}^h(x,y)|\leq 2\|\nabla_x V\|_{L^\infty(\mathbb{R}^N)}R^h$. Therefore, we deduce that

$$\begin{aligned}
&\|\Theta_d^{\varepsilon,h}(w)-\mathbb{V}^h(w)\|^2_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}\\
&\quad=\|(\mathcal{V}^{\varepsilon,h}-\mathcal{Y}^h)G^h\|^2_{L^2(\mathbb{R}^N;L^2(Y^h))}\\
&\quad=h^N\int_{\mathbb{R}^N}\int_{Y^h}\Big|\mathcal{V}^{\varepsilon,h}(x,y)-\mathcal{Y}^h(x,y)\Big|^2\,|G^h(x,y)|^2\,dy\,dx\\
&\quad\longrightarrow 0
\end{aligned}$$

as $\varepsilon\to 0$, still by using the Lebesgue theorem. □

THEOREM 5.5. *As ε goes to 0, the following convergences hold:*

$$\begin{cases}\displaystyle\lim_{\varepsilon\to 0}\|w^{\varepsilon,h}-v^h\|_{L^\infty(0,T;L^2(\mathbb{R}^N\times\mathbb{Z}^N))}=0,\\ \displaystyle\lim_{\varepsilon\to 0}\|f^{\varepsilon,h}-F^h\|_{L^\infty(0,T;L^2(\mathbb{R}^N\times\mathbb{R}^N))}=0.\end{cases}$$

Proof. It suffices to establish the convergence at the discrete level since the second follows easily (see the proof of Corollary 4.2). Set $e^{\varepsilon,h}(t,x,n)=w^{\varepsilon,h}(t,x,n)-v^h(t,x,n)$, which satisfies

$$\partial_t e_n^{\varepsilon,h}+\xi_n\cdot\nabla_x e_n^{\varepsilon,h}-\Theta_d^{\varepsilon,h}(V)(e^{\varepsilon,h})_n=\big[\big(\mathbb{V}^h-\Theta_d^{\varepsilon,h}(V)\big)v^h\big]_n.$$

Multiplying by $\overline{e^{\varepsilon,h}}$ and using that $\Theta_d^{\varepsilon,h}$ is skew-symmetric lead to the estimate

$$\|e^{\varepsilon,h}(t)\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}\leq\|e^{\varepsilon,h}(0)\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}+\int_0^t\|(\Theta_d^{\varepsilon,h}-\mathbb{V}^h)v^h(s)\|_{L^2(\mathbb{R}^N\times\mathbb{Z}^N)}\,ds$$

that tends to 0 when $\varepsilon\to 0$ by Proposition 5.4 and the hypothesis made on the initial data. □

As a concluding remark, let us show that the Vlasov equation can still be obtained from the discrete model when $\varepsilon,h\to 0$ together. Indeed, $f^{\varepsilon,h}$ satisfies a uniform L^2 estimate; see Lemma 5.2. Then, up to a subsequence, we can suppose that it has a weak limit

$$f^{\varepsilon,h}\rightharpoonup g\qquad\text{weakly}-*\text{ in }L^\infty(0,T;L^2(\mathbb{R}^N\times\mathbb{R}^N)).$$

Multiplying by a test function $\varphi\in\mathcal{S}(\mathbb{R}\times\mathbb{R}^N\times\mathbb{R}^N)$ with $\mathrm{supp}(\widehat{\varphi})\subset(0,T)\times\mathbb{R}^N\times B(0,R)$, $0<R<R^h$, we have the formula of Lemma 4.3:

$$\langle(\partial_t+\xi\cdot\nabla_x)f^{\varepsilon,h},\varphi\rangle_{\mathcal{S}';\mathcal{S}}=\langle f^{\varepsilon,h},\Theta^{\varepsilon,*}(V)\varphi\rangle_{\mathcal{S}';\mathcal{S}}.$$

When $\varepsilon,h\to 0$, the weak convergence of $f^{\varepsilon,h}$ combines to the strong convergence $\Theta^{\varepsilon,*}(V)\varphi\to-\nabla_x V\cdot\nabla_\xi\varphi$ in L^2, and we get

$$\langle(\partial_t+\xi\cdot\nabla_x)g-\nabla_x V\cdot\nabla_\xi g,\varphi\rangle_{\mathcal{S}';\mathcal{S}}=0$$

for any $\varphi \in \mathcal{S}$ whose Fourier transform is compactly supported. However, the set of admissible test functions is dense in $\mathcal{S}$ so that we deduce that g actually solves the Vlasov equation.

THEOREM 5.6. *When ε, h tend to 0 together, the sequence $f^{\varepsilon,h}$ converges weakly in $L^2((0,T)\times\mathbb{R}^N\times\mathbb{R}^N)$ to a solution of the Vlasov equation.*

Unfortunately, this result does not prevent oscillations that can occur in the limit $\varepsilon, h \to 0$: the convergence is obtained only in a weak sense, and we are not able to provide order of convergence for the norm of the error.

REFERENCES

[1] A. ARNOLD, H. LANGE, AND P. ZWEIFEL, *A discrete-velocity stationary Wigner equation*, J. Math. Phys., 41 (2000), pp. 7167–7180.

[2] A. ARNOLD AND C. RINGHOFER, *An operator splitting method for the Wigner–Poisson problem*, SIAM J. Numer. Anal., 33 (1996), pp. 1622–1643.

[3] P. DEGOND AND P. MARKOWICH, *A quantum-transport model for semiconductors: The Wigner–Poisson problem for a bounded Brillouin zone*, M2AN Math. Model. Numer. Anal., 24 (1990), pp. 697–710.

[4] P. GÉRARD, P. MARKOWICH, N. MAUSER, AND F. POUPAUD, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.

[5] P. GÉRARD, P. MARKOWICH, N. MAUSER, AND F. POUPAUD, *Erratum: "Homogenization limits and Wigner transforms" Comm. Pure Appl. Math.* 50 (1997), no. 4, 323–379, Comm. Pure Appl. Math., 53 (2000), pp. 280–281.

[6] P. MARKOWICH, *On the equivalence of the Schrödinger and the quantum Liouville equations*, Math. Methods Appl. Sci. 11 (1989), pp. 459–469.

[7] P. MARKOWICH, P. PIETRA, AND C. POHL, *Numerical approximation of quadratic observables of Schrödinger-type equations in the semi-classical limit*, Numer. Math., 81 (1999), pp. 595–630.

[8] P. MARKOWICH AND F. POUPAUD, *The pseudo-differential approach to finite-difference revisited*, Calcolo, 36 (1999), pp. 161–186.

[9] P. MARKOWICH AND C. RINGHOFER, *An analysis of the quantum Liouville equation*, Z. Angew. Math. Mech., 69 (1989), pp. 121–127.

[10] P. MARKOWICH, C. RINGHOFER, AND C. SCHMEISER, *Semiconductor Equations*, Springer–Verlag, Vienna, 1990.

[11] P. L. LIONS AND TH. PAUL, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana., 9 (1993), pp. 553–618.

[12] C. RINGHOFER, *Computational methods for semiclassical and quantum transport in semiconductor devices*, Acta Numer., 6 (1997), pp. 485–521.

[13] C. RINGHOFER, *A spectral collocation technique for the solution of the Wigner–Poisson problem*, SIAM J. Numer. Anal., 29 (1992), pp. 679–700.

[14] C. RINGHOFER, *On the convergence of spectral methods for the Wigner–Poisson problem*, Math. Models Methods Appl. Sci., 2, (1992), pp. 91–112.

[15] E. WIGNER, *On the quantum correction to thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 742–759.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2026–2046

SYMMETRIC FUNCTIONS APPLIED TO DECOMPOSING SOLUTION SETS OF POLYNOMIAL SYSTEMS*

ANDREW J. SOMMESE†, JAN VERSCHELDE‡, AND CHARLES W. WAMPLER§

Abstract. Many polynomial systems have solution sets comprised of multiple irreducible components, possibly of different dimensions. A fundamental problem of numerical algebraic geometry is to decompose such a solution set, using floating-point numerical processes, into its components. Prior work has shown how to generate sets of generic points guaranteed to include points from every component. Furthermore, we have shown how monodromy can be used to efficiently predict the partition of these points by membership in the components. However, confirmation of this prediction required an expensive procedure of sampling each component to find an interpolating polynomial that vanishes on it. This paper proves theoretically and demonstrates in practice that linear traces suffice for this verification step, which gives great improvement in both computational speed and numerical stability. Moreover, in the case that one may still wish to compute an interpolating polynomial, we show how to do so more efficiently by building a structured grid of samples, using divided differences, and applying symmetric functions. Several test problems illustrate the effectiveness of the new methods.

Key words. components of solutions, divided differences, embedding, generic points, traces, homotopy continuation, irreducible components, Newton identities, Newton interpolation, numerical algebraic geometry, monodromy, polynomial system, symmetric functions

AMS subject classifications. Primary, 65H10; Secondary, 13P05, 14Q99, 68W30

PII. S0036142901397101

1. Introduction.

Polynomial systems arising in scientific and engineering applications often have positive dimensional solution sets; moreover, the solution set may have components of different dimensions. For instance, in mechanical engineering, we may be given a set of rigid parts and a prescription for how they are to be connected by joints. These specifications can be formulated as a system of polynomial equations whose solution set describes the locations in space of all the parts. It may happen that some assemblies of the mechanism are rigid, whereas other assemblies of the same parts and joints allow an internal motion having one or more degrees of freedom. The notion of "degrees of freedom of motion" as used by a kinematician is thus equivalent to the "dimension of a solution set" for the polynomial system. Problems with similar characteristics arise in other disciplines.

For such polynomial systems, our task is to identify all irreducible components of the solution set, characterizing each component by its dimension and degree and providing witness points on the set. This problem is central in a developing new field: *numerical algebraic geometry*, a research program initiated in [33]. The goal is to

*Received by the editors October 28, 2001; accepted for publication (in revised form) June 4, 2002; published electronically December 13, 2002. This research was supported by the Volkswagen-Stiftung (RiP-program at Oberwolfach).

http://www.siam.org/journals/sinum/40-6/39710.html

†Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-4618 (sommese@nd.edu, http://www.nd.edu/~sommese). This material is based upon work supported by the National Science Foundation under grant 0105653 and the Duncan Chair of the University of Notre Dame.

‡Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 South Morgan (M/C 249), Chicago, IL 60607-7045 (jan@math.uic.edu, jan.verschelde@na-net.ornl.gov, http://www.math.uic.edu/~jan). This material is based upon work supported by the National Science Foundation under grant 0105739 and grant 0134611.

§General Motors Research and Development, Mail Code 480-106-359, 30500 Mound Road, Warren, MI 48090-9055 (Charles.W.Wampler@gm.com).

design numerically stable algorithms to efficiently solve polynomial systems arising in science and engineering. In the next two paragraphs we explain the relation of the current paper to previous work.

In [26], we presented a new method of embedding a polynomial system into a larger polynomial system, such that the numerical solution of a sequence of homotopies computes generic points on all solution components of the original polynomial system. These witness points form the basic data to decompose the solution sets into irreducible components, as we showed in [27]. Starting at any solution point, one may use numerical continuation to sample the component that contains it and construct an interpolation polynomial vanishing on the component. This polynomial can then be used as a filter to find all other witness points in the same component. In this way, all of the points can be sorted into components, eventually producing a list of all the components and certain properties of them, such as degree and dimension. However, the construction of the interpolating polynomial is both expensive and numerically difficult for high degree components in many variables. In [28], we reduced the number of variables by detecting the linear span of a component and reduced the degree of the interpolant by using central projections; but even so, numerically challenging cases remain.

An alternative approach to determining which witness points lie on the same component is to use monodromy to find paths connecting them [29]. In computational experiments, this approach has been found to be numerically stable on high degree components and highly successful in predicting the correct decomposition. However, it is heuristic in that connections are discovered via randomly generated monodromy loops, with no a priori way to know when all connections have been found. Thus, the prediction must still be validated by other means. In [29], this was accomplished by computing an interpolating polynomial, as before, so the problem of high degree polynomials was not eliminated.

The most significant contribution of this paper is to prove theoretically and demonstrate computationally that *linear* traces are sufficient for validating a proposed decomposition. Due to the superior numerical stability of linear systems, we are able to run our decomposition method entirely with standard machine arithmetic. For polynomial systems with coefficients given as double floating-point numbers, whose evaluation map is numerically well conditioned, and whose irreducible components have multiplicity one, our algorithm does not need multiprecision arithmetic to decompose the solution sets, even in the occurrence of high degree components. For components with multiplicity higher than one, multiprecision arithmetic is required to track the singular paths [31].

In the case that one still wishes to compute polynomials that vanish on a component, the higher order traces can be used to good effect. First, the witness points on a component can be marched forward together to provide a structured grid of sample points. Then, with a "bootstrapping" technique, we can construct the Newton form of the interpolating polynomial. The use of traces enables the direct application of Newton interpolation, eliminating the need for extra bootstrapping samples. A final improvement in efficiency is gained by using Newton identities to reduce the number of samples to the number of monomials in the interpolant, which is the minimum possible. However, this last shortcut is inadvisable for high degree components as our tests show that it is numerically less stable than using a full grid of samples.

Several test problems illustrate the effectiveness of the methods. Particularly notable are the results on a problem from mechanical engineering: a special Stewart–

Gough platform mechanism that has internal motion. For one case in which the motion is one irreducible component of degree 28, the computing time for validating the decomposition predicted by monodromy is reduced from 1.3 hours using our former methods to less than 5 seconds with the linear trace. This is now comparable to the time required for a related example in which the degree 28 component breaks up into several low degree irreducibles. Hence the running time of the algorithm is no longer sensitive to such changes in the geometry of the solution set. Moreover, it is interesting to note that the automated numerical method discovered a solution component that was missed by experts using a manual approach aided by computer symbolic processing.

In brief, the paper proceeds as follows. In the next section, we collect some results on traces, which are then applied to monodromy in section three. In sections four and five, we outline the interpolation algorithms and apply them in the last section on the cyclic 8-roots and 9-roots problems and on the mechanism problem just mentioned.

2. Traces of functions. The results in this section are quite old, e.g., Theorem 2.1 is for the most part just a statement of the constructions that go along with one of the main approaches to the Weierstrass preparation theorem [8, 9]. Since we do not know a reference for the full result, we include a proof. The statement in Corollary 2.2 is equivalent to the zero-sum relations that have been used in a similar context, e.g., [3, 6, 20, 21, 22, 23, 24].

It is natural to consider functions $f(x_1,\ldots,x_d)$ of points $(x_1,\ldots,x_d)\in\mathbb{C}^d$ which are invariant under the symmetric group S_d of permutations of the variables. To be precise, given any permutation, $\sigma(i)=j_i$ for $i=1,\ldots,d$, of the integers from 1 to d, we have a linear transformation of $\mathbb{C}^d$, which by abuse of notation we also label σ, which takes $(x_1,\ldots,x_d)$ to $(x_{j_1},\ldots,x_{j_d})$. A function on $\mathbb{C}^d$ is said to be *symmetric* if

$$f\circ\sigma = f(x_{j_1},\ldots,x_{j_d}) = f(x_1,\ldots,x_d) \tag{2.1}$$

for all $\sigma\in S_d$. It is a basic fact of invariant theory that the ring of symmetric polynomials on $\mathbb{C}^d$, denoted by $\mathbb{C}[x_1,\ldots,x_d]^{S_d}$, is abstractly isomorphic to the ring of polynomials on $\mathbb{C}^d$, i.e., there exists a ring isomorphism $\mathbb{C}[z_1,\ldots,z_d]\cong\mathbb{C}[x_1,\ldots,x_d]^{S_d}$. There are many useful choices of assignments of symmetric functions to the z_i making this isomorphism explicit. The two that we use are

1. the assignment leading to the elementary symmetric functions

$$z_i\mapsto t_i := \frac{1}{(d-i)!i!}\sum_{\sigma\in S_d} x_{\sigma(1)}\cdots x_{\sigma(i)} = \sum_{1\le j_1<\ldots<j_i\le d} x_{j_1}\ldots x_{j_i} \tag{2.2}$$

   for i from 1 to d, and
2. the assignment leading to the power sums

$$z_i\mapsto p_i := \frac{1}{(d-1)!}\sum_{\sigma\in S_d} x_{\sigma(1)}^i = \sum_{j=1}^{d} x_j^i \tag{2.3}$$

   for i from 1 to d.

The connection of t_i with roots of a polynomial of degree d is easy to see, upon noting that

$$(w-x_1)(w-x_2)\cdots(w-x_d) = w^d - t_1w^{d-1} + t_2w^{d-2} - \cdots + (-1)^d t_d. \tag{2.4}$$

If, in the right-hand side of (2.4), we (like in [16]) substitute w by x_i, for i from 1 to d, and add up these n sums, then we obtain Newton's relation:

$$p_d - t_1 p_{d-1} + t_2 p_{d-2} - \cdots + d(-1)^d t_d = 0. \tag{2.5}$$

Equation (2.5) allows us to write the elementary symmetric functions in terms of the power sums and vice versa.

The following fact is classical.

THEOREM 2.1. *Let $p : X \to Y$ be a proper, finite, d-sheeted surjective complex analytic morphism from a reduced pure n-dimensional complex analytic space X to a normal irreducible and reduced complex analytic variety Y. Given a holomorphic function $f : X \to \mathbb{C}$ and a symmetric polynomial $g : \mathbb{C}^d \to \mathbb{C}$, there is a unique holomorphic function $f_g : Y \to \mathbb{C}$ such that, for all points y in the Zariski open and dense set $U \subset Y$ such that $p : p^{-1}(U) \to U$ is an unramified cover, it follows that for the d points $\{x_1, \ldots, x_d\} = p^{-1}(y)$, we have $f_g(y) := g(f(x_1), \ldots, f(x_d))$.*

For the symmetric functions t_i we denote f_{t_i} by $\mathrm{tr}_{i,p}(f)$ and call it the ith trace. It is traditional to call $\mathrm{tr}_{1,p}(f)$, or $\frac{\mathrm{tr}_{1,p}(f)}{d}$, the trace and $\mathrm{tr}_{d,p}(f)$ the norm of f. In fact, the t_i all occur naturally as traces, e.g., letting $A : \mathbb{C}^d \to \mathbb{C}^d$ denote a matrix with eigenvalues $\{x_1, \ldots, x_d\}$, t_i is the trace of the matrix A induces on $\wedge^i \mathbb{C}^d \cong \mathbb{C}^{\binom{d}{i}}$, the ith exterior product of $\mathbb{C}^d$. The parameterized version of (2.4) is

$$f^d - \mathrm{tr}_{1,p}(f) f^{d-1} + \mathrm{tr}_{2,p}(f) f^{d-2} + \cdots + (-1)^d \mathrm{tr}_{d,p}(f) = 0. \tag{2.6}$$

This line of reasoning is used in one main approach to the Weierstrass preparation theorem [8, 9].

Proof of Theorem 2.1. The proof that f_g is holomorphic follows by a minor variant of the argument used in [8, Theorem A4]. To see this, note that there is a codimension one analytic subset $B \subset Y$ such that $X \setminus p^{-1}(B)$ and $Y \setminus B$ are smooth and such that $p_{X \setminus p^{-1}(B)} : X \setminus p^{-1}(B) \to Y \setminus B$ is a d-sheeted unramified cover. Indeed, define U' equal to Y minus the union of

1. the singular set $\mathrm{Sing}(Y)$ of Y, which is an analytic set of complex codimension 2 [8, Theorem Q12], and
2. the image under p of the singular set of X, which is an analytic set by Remmert's proper mapping theorem [8, Theorem N1].

Thus $p_{p^{-1}(U')} : p^{-1}(U') \to U'$ is a proper and finite map between complex manifolds. The set of branch points R of this map is an analytic set on U', since it is the set defined by the local condition that the determinant of the Jacobian of the mapping $p_{p^{-1}(U')}$ is zero. Define $U := U' \setminus p(R)$.

Thus for $y \in Y \setminus B$, $f_g(y) := g(f(x_1), \ldots, f(x_d))$ is a holomorphic function. Since p is proper, given any point $y \in B$, there is a relatively compact open set U containing y such that $p^{-1}(U)$ is relatively compact and f is bounded in absolute value on $p^{-1}(U)$. Thus f_g is bounded on $U \setminus (B \cap U)$. Thus by the Riemann extension theorem for bounded holomorphic functions, $f_{g|U}$ has a unique holomorphic extension to $Y \setminus \mathrm{Sing}(Y)$, where $\mathrm{Sing}(Y)$ denotes the singular set of Y. Since Y is normal, it follows from the Levi extension theorem [8, Theorem Q15i] that $f_{g|Y \setminus \mathrm{Sing}(Y)}$ has a unique holomorphic extension to Y. □

In Theorem 2.1, if in addition it is assumed that X, Y, p, and f are algebraic, then it follows that f_g is also algebraic. Rather than introduce all the needed definitions and algebraicity criteria to state the general case, we prove only a corollary that covers our needs.

COROLLARY 2.2. *Let $Z \subset \mathbb{C}^n$ be a pure k-dimensional algebraic subvariety of $\mathbb{C}^n$. Assume*

1. *that $\pi : \mathbb{C}^n \to \mathbb{C}^k$ is a generic linear projection, and*
2. *that ϕ is a linear function on $\mathbb{C}^n$, which is one-to-one on a fiber $\pi_Z^{-1}(y)$ for some $y \in \mathbb{C}^k$ with $\pi^{-1}(y)$ consisting of smooth points of Z at which the tangent space, $d\pi_Z$, has rank k.*

Then, it follows that for all j, $\mathrm{tr}_{j,\pi_Z}(\phi)$ is a polynomial on $\mathbb{C}^k$ of degree less than or equal to j. In particular, $\mathrm{tr}_{1,\pi_Z}(\phi)$ is linear or constant.

Proof. Let q denote the map $\mathbb{C}^n \to \mathbb{C}^{k+1}$ given by (ϕ, π). Let z denote the coordinate on $\mathbb{C}^{k+1}$ such that $z(q(\mathbf{x})) = \pi(\mathbf{x})$, and let L denote the projection of $\mathbb{C}^{k+1} \to \mathbb{C}^k$ such that $\pi = L \circ q$. By the Noether normalization theorem it follows that the genericity of π implies that $p := \pi_Z$ is a proper finite-to-one morphism. Since $\pi_Z = L \circ q_Z$, it follows by elementary point set topology that q_Z is proper and finite also. Moreover, by genericity it follows that the degree d of p is $\deg Z$, by the hypothesis on ϕ that q_Z maps Z generically one-to-one to $\mathbb{C}^{k+1}$, and, therefore, that $\deg q(Z) = \deg Z$. Since for a dense open set $U \subset \mathbb{C}^k$ q gives an isomorphism of $p^{-1}(U)$ with $L_{q(Z)}^{-1}(U)$, we conclude that $\mathrm{tr}_{j,p}(\phi_Z)$ and $\mathrm{tr}_{j,L_{q(Z)}}(z_{q(Z)}$ agree on U and hence on all of $\mathbb{C}^k$. Thus, with the convention that $\mathrm{tr}_{0,L_{q(Z)}}(z_{q(Z)}) = \mathrm{tr}_{0,p}(\phi_Z) = 1$, we have the equivalent relations given in (2.6) $f^d - \mathrm{tr}_{1,p}(f) f^{d-1} + \cdots + (-1)^d \mathrm{tr}_{d,p}(f) = 0$,

$$\sum_{i=0}^{d} (-1)^i \mathrm{tr}_{i,L_{q(Z)}}(z_{q(Z)}) z_{q(Z)}^{d-i} = 0, \tag{2.7}$$

$$\sum_{i=0}^{d} (-1)^i \mathrm{tr}_{i,p}(\phi_Z) {\phi_Z}^{d-i} = 0. \tag{2.8}$$

Since $q(Z)$ is a degree d hypersurface and $z^d - \mathrm{tr}_{1,p}(z_{q(Z)}) z^{d-1} + \cdots + (-1)^d \mathrm{tr}_{d,p}(z_{q(Z)})$ vanishes when restricted to it, we conclude that this must be a minimum degree defining polynomial of $q(Z)$. Thus we have proved the assertions of the corollary. ∎

Remark 2.3. Note that, assuming the genericity hypothesis on π in Corollary 2.2, the hypothesis on ϕ can be replaced by the equivalent more easily checked condition that ϕ is a linear function on $\mathbb{C}^n$, which is one-to-one on a fiber $\pi_Z^{-1}(y)$ for some $y \in \mathbb{C}^k$ with $\pi^{-1}(y)$ consisting of $\deg Z$ distinct points.

Remark 2.4. Note that, without the genericity assumption, Corollary 2.2 fails for a number of reasons. First, it might be that π_Z is not proper. In this case, the traces are only rational functions. For example, taking $Z := \{xy - 1 = 0\} \subset \mathbb{C}^2$ and $\pi : \mathbb{C}^2 \to \mathbb{C}$ given by $\pi(x, y) = x$, we get $\mathrm{tr}_{1,\pi_Z}(y) = \frac{1}{x}$. By the Noether normalization theorem, the genericity assumption about the linear projection π implies that π_Z is proper and finite, but even proper and finite, without the genericity assumption, is not enough. The key implication of genericity of the linear projection π, beyond the properness of π_Z, is the fact that for generic linear projections π, $\deg Z = \deg \pi_Z$. For example, taking $Z := \{y^2 - x^d = 0\} \subset \mathbb{C}^2$ and $\pi : \mathbb{C}^2 \to \mathbb{C}$ given by $\pi(x, y) = x$, we get $\mathrm{tr}_{2,\pi_Z}(y) = x^d$. In general, if

1. $Z \subset \mathbb{C}^n$ is a pure k-dimensional algebraic subvariety of $\mathbb{C}^n$,
2. f is linear on $\mathbb{C}^n$, and
3. $\pi : \mathbb{C}^n \to \mathbb{C}^k$ is a linear projection with π_Z proper, finite, but not necessarily satisfying $\deg Z = \deg \pi_Z$,

then it follows that $\deg \mathrm{tr}_{i,\pi_Z}(f) \leq \deg Z - \deg \pi + i$.

3. An application to monodromy. In this section we show that the linear trace gives necessary and sufficient conditions to determine the breakup of the set witness points of the algorithm of [27] into the disjoint subsets of generic points corresponding to the numerical irreducible decomposition. Usually, we use this result to give a very fast verification of the monodromy breakup of the algorithm of [29], but it is also called into play if the monodromy breakup is too fine. The algorithms will be described in the next section.

The important observation that the linear trace is sufficient is due to Rupprecht [20] in the case of curves. We note that there are two serious gaps in the argument of [20], which are filled by Theorems 3.3 and 3.4 below.

The strategy is to slice and project to reduce to the case of a curve in $\mathbb{P}^2$. To do this we need a number of lemmas on linear projections and the intersections with linear spaces that are general subject to certain constraints. Unless otherwise said, closure is in either the Zariski topology or the usual topology induced by the Euclidean metric on $\mathbb{C}^n$.

LEMMA 3.1. *Let A be a pure k-dimensional reduced algebraic subset of $\mathbb{C}^n$, with irreducible decomposition $\cup_{i=1}^r A_i$. Assume that L is an $(n-k)$-dimensional linear subspace of $\mathbb{C}^n$ meeting A in a finite set $\mathcal{A}$ consisting of $\deg A$ distinct isolated points. Then, taking closures in $\mathbb{P}^n$, $\overline{L} \cap \overline{A} = \mathcal{A}$. Moreover, if $k \geq 2$, then letting $\mathcal{L}$ be a general member of the set of $(n-k+1)$-dimensional linear subspaces of $\mathbb{C}^n$ that contain L, it follows that $\mathcal{L} \cap A_i$ is an irreducible curve for each $i = 1, \ldots, r$.*

Proof. The statement that $\overline{L} \cap \overline{A} = \mathcal{A}$ follows from any of a number of related results, e.g., [5, Example 8.4.6] or [5, Example 12.3.2].

To prove the second statement it suffices to prove the analogous result on $\mathbb{P}^n$ using the closures of the sets A_i, L, $\mathcal{L}$. Since $\overline{L} \cap \overline{A}$ is a set $\mathcal{A}$ of cardinality $\deg \overline{A}$, it follows that $\mathcal{A}$ consists of smooth points of $\overline{A}$, and that the intersection of $\overline{A}$ and $\overline{L}$ are transverse at the points of intersection. From this it follows from Bertini's theorem that the intersection with A_i of a general member of the set M of $(n-k+1)$-dimensional linear subspaces of $\mathbb{C}^n$ that contain L is smooth away from the singular locus of A_i. If $k = 2$, so that the set M consists of hyperplanes, the rest of the argument follows exactly as in [25, Theorem 3.42]. If $k \geq 2$, the result follows by a straightforward descending induction. For example, if $k = 3$, then it follows using [25, Theorem 3.42] that the intersection with A_i of a general member of the set of $(n-1)$-dimensional linear subspaces H of $\mathbb{C}^n$ that contain L is irreducible. Keeping L as it is, taking H in place of $\mathbb{C}^n$, and replacing the A_i with $A_i \cap H$, we now have $\dim A_i \cap H = 2$, i.e., we have reduced to the proven result. ∎

LEMMA 3.2. *Let n, L, $\mathcal{L}$, $\mathcal{A}$, $A = A_1 \cup \cdots \cup A_r$ be as in Lemma 3.1. Choose a general linear projection $\pi : L \to \mathbb{C}$, which is one-to-one on $\mathcal{A}$. Let $\widetilde{\pi} : \mathbb{C}^n \to \mathbb{C}^{k+1}$ be a linear map extending π, so that the fibers of $\widetilde{\pi}$ are parallel to the fibers of π. Then $\widetilde{\pi}_A$ and $\widetilde{\pi}_{A \cap \mathcal{L}}$ are generically one-to-one and proper. Moreover, $\widetilde{\pi}_A$ (respectively, $\widetilde{\pi}_{A \cap \mathcal{L}}$), maps a neighborhood of $\mathcal{A}$ in A (respectively, in $A \cap \mathcal{L}$) isomorphically onto a neighborhood of the image of $\mathcal{A}$ in A (respectively, in $A \cap \mathcal{L}$).*

Proof. We give the proof that $\widetilde{\pi}_A$ is generically one-to-one and proper, and we leave the remaining argument, which follows the same line of reasoning to the reader. We work in the projective space $\mathbb{P}^n$. As explained in [27], $\widetilde{\pi}$ corresponds to the central projection from a linear $I := \mathbb{P}^{n-k-2}$ contained in the linear $\mathbb{P}^{n-1}$ at infinity, i.e., in $\mathbb{P}^n \setminus \mathbb{C}^n$. The condition that L is mapped to a line corresponds to $I \subset \overline{L} \setminus L$. Since by Lemma 3.1 we know that $(\overline{L} \setminus L) \cap (\overline{A} \setminus A) = \emptyset$, we know, as discussed in [27], that $\widetilde{\pi}_A$ is proper and therefore finite-to-one on A. If $\widetilde{\pi}_A$ was not generically one-to-one, then

$\deg \widetilde{\pi}(A)$ would be less than $\deg A$. But this does not happen since $\widetilde{\pi}$ is one-to-one on $\mathcal{A}$, and the degree of $\widetilde{\pi}(A)$ is the cardinality of $\widetilde{\pi}(\mathcal{A})$, which is equal to the intersection of $\widetilde{\pi}(A)$ with the line $\widetilde{\pi}(L)$. ∎

THEOREM 3.3. *Let A be a reduced pure k-dimensional algebraic subset of $\mathbb{C}^n$. Let $A := A_1 \cup \cdots \cup A_r$ be the decomposition into distinct irreducible components. Let $L \subset \mathbb{C}^n$ be a general linear subspace of dimension $n-k$ meeting A transversely in the set $\mathcal{A} := A \cap L$. For all $i = 1, \ldots, r$, let $\mathcal{A}_i := A_i \cap L$, and let d_i be the cardinality of the set $\mathcal{A}_i$. Let d denote the cardinality of $\mathcal{A}$, i.e., $\sum_{i=1}^{r} d_i$. Let U denote the Zariski open set of the Grassmannian of $(n-k)$-dimensional linear subspaces of $\mathbb{C}^n$ consisting of linear spaces transverse to A. Let* $\mathrm{Sym}(\mathcal{A})$ *(respectively,* $\mathrm{Sym}(\mathcal{A}_i)$*) denote the symmetric group of all permutations of $\mathcal{A}$ (respectively, $\mathcal{A}_i$). Considering L as a basepoint of U, the image in* $\mathrm{Sym}(\mathcal{A})$ *of the natural monodromy action of $\pi_1(U, L)$ on $\mathcal{A}$ is the direct sum $\bigoplus_{i=1}^{r} \mathrm{Sym}(\mathcal{A}_i)$.*

Proof. We can assume without loss of generality that no components of A are linear, since such components do not affect the result.

Choose $\mathcal{L}$ and $\widetilde{\pi} : \mathbb{C}^n \to \mathbb{C}^{k+1}$ as in Lemma 3.2. Let U' denote the Zariski open set of the projective space of lines in $\mathbb{C}^{k+1}$, consisting of the lines transverse to $\widetilde{\pi}(A)$. Note that for $\ell \in U'$ we have that $\widetilde{\pi}^{-1}(\ell) \in U$. Thus identifying $\mathcal{A}$ with $\widetilde{\pi}(\mathcal{A})$, the homomorphism $\pi_1(U', \widetilde{\pi}(L)) \to \mathrm{Aut}(\mathcal{A})$ factors $\pi_1(U', \widetilde{\pi}(L)) \to \pi_1(U, L) \to \mathrm{Aut}(\mathcal{A})$. Thus, from here on we can assume without loss of generality that $n = k+1$. Moreover, a line in $\mathcal{L}$ transverse to $\mathcal{L} \cap A$ is transverse to A in $\mathbb{C}^n$. Thus, letting U'' denote the Zariski open set of lines in $\mathcal{L} \cong \mathbb{C}^2$ that are transverse to $\mathcal{L} \cap A$, we have a composition $\pi(U'', L) \to \pi_1(U, L) \to \mathrm{Aut}(\mathcal{A})$. Thus, without loss of generality we can assume that $n = 2$.

Thus we have $n = 2$ and $k = 1$. We have the classical fact [1, Lemma, p. 111] that the image of $\pi_1(U, L) \in \mathrm{Aut}(\mathcal{A})$ surjects onto $\mathrm{Sym}(\mathcal{A}_i)$ for each $i = 1, \ldots, r$. In particular, we can assume without loss of generality that $r \geq 2$. By elementary algebra, we see that using this surjectivity, we would be done if we showed that, for each i, there exist two distinct points $a, b \in \mathcal{A}_i$ and a $\gamma \in \pi_1(U, L)$ such that γ acts on $\mathcal{A}$ by sending $a \to b$, $b \to a$ and leaves the remaining points of $\mathcal{A}$ fixed. The classical argument for the existence of such a γ in the irreducible case [1, Lemma, p. 111] carries over with no change to the reducible case if we show that for each $i = 1, \ldots, r$, and a generic point $x \in \mathcal{A}_i$, the tangent line $\ell \subset \mathbb{C}^2$ to A_i at x is transverse to A_j for $j \neq i$. To show this, it suffices to work projectively, i.e., show the fact for the closures B_i in $\mathbb{P}^2$ of the A_i. To see this, consider the dual curves $\widehat{B}_i \subset (\mathbb{P}^2)^*$. Here $(\mathbb{P}^2)^*$ is the $\mathbb{P}^2$ whose points correspond to lines in the $\mathbb{P}^2$ that the B_i belong to, and $\widehat{B}_i$ is the closure in $(\mathbb{P}^2)^*$ of the set of points corresponding to tangent lines to smooth points of B_i. What we need is exactly that B_i and B_j for distinct i, j go to distinct curves $\widehat{B}_i$ and $\widehat{B}_j$. Noting that none of the B_i are linear, this would follow if we knew that the dual of $\widehat{B}_i$ is B_i. This is a basic fact about dual curves (and more generally varieties) [14]. ∎

To prove Corollary 3.5 below, we need the following generalization of the classical first Lefschetz theorem. This topological result is a special case of a useful general result of Goresky and MacPherson [7, Theorem, section 5.2, p. 199].

THEOREM 3.4 (Goresky–MacPherson). *Let D be an arbitrary algebraic subset of $\mathbb{C}^n$, and let $U := \mathbb{C}^n \setminus D$. Then given a general 1-dimensional linear subspace $L \subset \mathbb{C}^n$ and a point $x \in L \cap U$, it follows that we have a surjective map of fundamental groups*

$$\pi_1(L \cap U, x) \to \pi_1(U, x) \to 0.$$

Proof. Following the notation of [7, section 5.2, Theorem, p. 199], take X to be the Zariski open dense set U of $\mathbb{P}^n$ with $n := N$, π the inclusion, $c = n-1$, which gives $\phi(k) = n-1$ for $k = 0$, and $-\infty$ for $k \neq 0$, which gives $\hat{n} = 1$. □

COROLLARY 3.5. *Let A be a reduced pure k-dimensional algebraic subset of $\mathbb{C}^n$. Let $A := A_1 \cup \cdots \cup A_r$ be the decomposition into distinct irreducible components. Let $\pi : \mathbb{C}^n \to \mathbb{C}^k$ denote a generic linear projection, and let $x, y \in \mathbb{C}^k$ denote general points, with $L := \pi^{-1}(x) \subset \mathbb{C}^n$ a general linear subspace of dimension $n-k$ meeting A transversely in the set $\mathcal{A} := A \cap L$. For all $i = 1, \ldots, r$, let $\mathcal{A}_i := A_i \cap L$, and let d_i be the cardinality of the set $\mathcal{A}_i$. Let d denote the cardinality of $\mathcal{A}$, i.e., $\sum_{i=1}^r d_i$. Let U denote the Zariski open set of the line $\ell \subset \mathbb{C}^k$ containing x, y, consisting of the $u \in \ell$ such that $\pi^{-1}(u)$ is transverse to A. Let* $\mathrm{Sym}(\mathcal{A})$ *(respectively,* $\mathrm{Sym}(\mathcal{A}_i)$*) denote the symmetric group of all permutations of $\mathcal{A}$ (respectively, $\mathcal{A}_i$). Considering L as a basepoint of U, the image in* $\mathrm{Sym}(\mathcal{A})$ *of the natural monodromy action of $\pi_1(U, L)$ on $\mathcal{A}$ is the direct sum* $\bigoplus_{i=1}^r \mathrm{Sym}(\mathcal{A}_i)$.

Proof. By the same reduction as in Theorem 3.3, it can be assumed that $n = 2$, $k = 1$, and that we are working with compact curves in $\mathbb{P}^2$. Given the Zariski open set U' of points in $(\mathbb{P}^2)^*$ corresponding to lines transverse to A, and given a general line $\ell \subset (\mathbb{P}^2)^*$, with a point L on ℓ, then setting $U := \ell \cap U'$, we will be done if we show that $\pi_1(U, L) \to \pi_1(U, L) \to 0$. This is guaranteed by Theorem 3.4. □

THEOREM 3.6. *Let $A \subset \mathbb{C}^n$ be an affine algebraic set of pure dimension k. Let $A := A_1 + \cdots + A_r$ denote the irreducible decomposition of A. Let $\pi : \mathbb{C}^n \to \mathbb{C}^k$ be a generic linear projection, and let $\ell \subset \mathbb{C}^k$ be a general line. Let $L := \pi^{-1}(\mathbf{x})$ for a general point $\mathbf{x} \in \ell$. Let ϕ be a linear function on $\mathbb{C}^n$ which is one-to-one on $\mathcal{A} := \pi_A^{-1}(\mathbf{x})$. For all $i = 1, \ldots, r$, let $\mathcal{A}_i := \pi_{A_i}^{-1}(\mathbf{x})$. Let U denote the Zariski open set of the Grassmannian of $(n-k)$-dimensional linear spaces of $\mathbb{C}^n$ corresponding to $(n-k)$-dimensional linear spaces transverse to A. Let B denote a subset of $\mathcal{A}$. Then the following are equivalent:*

(1) *B is invariant under the monodromy action of $\pi_1(U, L)$ on $\mathcal{A}$;*
(2) *$B = \cup_{i \in I} \mathcal{A}_i$ for some subset $I \subset \{1, \ldots, r\}$;*
(3) *The analytic continuation of $\sum_{b \in B} \phi(b)$ as a function of $\mathbf{x}$ is linear.*

Proof. The equivalence of (1) and (2) follows from Theorem 3.3. Corollary 2.2 shows that (2) implies (3). So it remains to show that (3) implies (1).

To see this, assume that B contains a point $b \in \mathcal{A}_i$ but not a point $a \in \mathcal{A}_i$. Let $y_1, \ldots, y_N$ denote the points of $B \setminus b$. Let U' denote the Zariski open subset of ℓ consisting of the points $\mathbf{x}' \in \ell$ with $\pi^{-1}(\mathbf{x}')$ transverse to A. By Corollary 3.5, there is a $\gamma \in \pi_1(U', L)$ which takes $y_j \to y_j$ for all j and interchanges a and b. Thus, since the analytic continuation of $\sum_{b \in B} \phi(b)$ is linear in $\mathbf{x}$, we conclude that

$$\phi(a) + \sum_{j=1}^{N} \phi(y_j) = \phi(b) + \sum_{j=1}^{N} \phi(y_j), \tag{3.1}$$

and thus that $\phi(a) = \phi(b)$. But this contradicts ϕ being one-to-one on $\mathcal{A}$. Thus if $b \in \mathcal{A}_i$ for some point $b \in B$, we conclude that $\mathcal{A}_i \subset B$. □

Remark 3.7. Here is a simple example to show the sort of bad behavior that genericity rules out. Let A be the curve in $\mathbb{C}^2$ defined by $(y^2 - x)(y^2 - 4x)(y^2 - 9x) = 0$. Consider the projection $\pi : \mathbb{C}^2 \to \mathbb{C}$ given by $\pi(x, y) = x$. Note that π_A is proper and generically six-to-one with $\deg A = 6$. For the linear function ϕ on $\mathbb{C}^2$, choose y. Over $x \in \mathbb{C}$, the values of ϕ on the fiber $\pi_A^{-1}(y)$ are

$$\{\sqrt{x}, -\sqrt{x}, 2\sqrt{x}, -2\sqrt{x}, 3\sqrt{x}, -3\sqrt{x}\}. \tag{3.2}$$

The groupings corresponding to the irreducible components are $\{\sqrt{x}, -\sqrt{x}\}$, $\{2\sqrt{x}, -2\sqrt{x}\}$, and $\{3\sqrt{x}, -3\sqrt{x}\}$. Notice that the sum $(-1)\sqrt{x} + (-2)\sqrt{x} + 3\sqrt{x}$ is identically zero, and hence linear, even though the grouping $\{-\sqrt{x}, -2\sqrt{x}, 3\sqrt{x}\}$ does not correspond to a union of irreducible components of A.

Remark 3.8. In practice, when we use Theorem 3.6, it is convenient to use a generic projection $\pi : \mathbb{C}^n \to \mathbb{C}^{k+1}$. Letting $A := A_1 + \cdots + A_r$ be as in Theorem 3.6, it follows, e.g., from [27, section 5.2], that

1. the map π_A from A to its image $\pi(A)$ is proper,
2. the images of $A, A_1, \ldots, A_r$ under π are affine algebraic sets with $\pi(A)$ having the irreducible decomposition $\pi(A) := \pi(A_1) + \cdots + \pi(A_r)$,
3. $\deg \pi(A_i) = \deg A_i$ for all i, and
4. π_A is one-to-one on $\pi_A^{-1}(U)$ for some Zariski open dense set of $\pi(A)$.

Moreover, since the composition of π with the projection $\mathbb{C}^{k+1} \to \mathbb{C}^k$ given by $(z_1, \ldots, z_{k+1}) \to (z_1, \ldots, z_k)$ is generic, we can use $\pi(A)$ and this projection in place of A and the generic projection π of Theorem 3.6. Then, z_{k+1} has the properties required of ϕ in Theorem 3.6, and, since the projection $\pi : \mathbb{C}^n \to \mathbb{C}^{k+1}$ is generic, we can take one of the coordinate axes, e.g., the z_k axis, as ℓ.

It is worth noting that the defining equation of $\pi(A)$ under a generic projection will have every monomial of total degree less than or equal to the degree of A occurring, no matter how sparse the defining equations of A are.

4. Algorithms for the linear trace.

In this section we present an algorithm to verify the decomposition predicted by the monodromy algorithm. We first define a projection operator which organizes the samples in a structured grid. The main part of this section is the algorithm **Certify**, followed by comments on how to integrate this algorithm in the numerical irreducible decomposition of a pure dimensional component.

4.1. Sampling on parallel slices.

To compute the linear trace, a structured grid of sample points is useful. The same construction is used in the following section concerning higher traces. Our technique is to use random slicing hyperplanes to define the projection operator π, as follows.

DEFINITION 4.1. *Consider a k-dimensional component in $\mathbb{C}^n$ and suppose we use the k hyperplanes*

$$c_{i0} + c_{i1}x_1 + c_{i2}x_2 + \cdots + c_{in}x_n = 0, \qquad i = 1, 2, \ldots, k, \tag{4.1}$$

as slices to obtain generic points on the component. To project the generic points down to $\mathbb{C}^{k+1}$, we use the map $\pi : \mathbb{C}^n \to \mathbb{C}^{k+1}$ defined by

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \mapsto \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kn} \\ a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}, \tag{4.2}$$

where the numbers a_i, $i = 1, 2, \ldots, k$ are chosen at random.

The main property of this projection operator is highlighted in the following proposition.

PROPOSITION 4.2. *For a k-dimensional component in $\mathbb{C}^n$, let $\pi : \mathbb{C}^n \to \mathbb{C}^{k+1}$ be as in* (4.2). *Then for any generic point $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ on the slices used in the definition of π, we have*

$$\pi(\mathbf{x}) = (-c_{10}, -c_{20}, \ldots, -c_{k0}, a_1x_1 + a_2x_2 + \cdots + a_nx_n). \tag{4.3}$$

To obtained a structured grid of samples from the k-dimensional component, we let only the constant terms c_{i0} of the slicing planes vary, for $i = 1, 2, \ldots, k$. Geometrically this means we take samples on slices parallel to each other.

4.2. Certification of monodromy groupings with linear traces. Suppose we are given a set S of d generic points on random hyperplanes $L = (L_1, L_2, \ldots, L_k)$, where the points are known to be from the same k-dimensional irreducible component because the monodromy algorithm found loops connecting them. As the monodromy algorithm might miss some connections, the actual degree of the component could be higher than d. With linear traces we verify whether the degree of the component equals d, described by the **Certify** algorithm.

ALGORITHM 4.3. $b = \textbf{Certify}(f, L, S, \epsilon)$

Input: $f(\mathbf{x}) = \mathbf{0}$ is a polynomial system with $\mathbf{x} \in \mathbb{C}^n$;
$L = (L_1, L_2, \ldots, L_k)$ is a tuple of k random hyperplanes;
S is set of d generic points satisfying $f(\mathbf{x}) = \mathbf{0}$ and $L(\mathbf{x}) = \mathbf{0}$;
ϵ is tolerance to decide whether a number is close enough to zero.

Output: $b \in \{\text{false}, \text{true}\}$, b is true when S is a set of generic points on a degree d irreducible component, false otherwise.

let $S^{(0)} = S$, $c_0^{(0)} = c_0$; [*notational convenience*]
for $i = 1, 2$ do [*sample to get test points*]
  choose $c_0^{(i)} \in \mathbb{C}$ at random;
  $L_k^{(i)} := c_0^{(i)} + c_{k1}x_1 + c_{k2}x_2 + \cdots + c_{kn}x_n$; [$L_k^{(i)}$ *is parallel to* L_k]
  compute $S^{(i)}$ as solutions to $f(\mathbf{x}) = \mathbf{0}$, [*apply homotopy from* L_k *to* $L_k^{(i)}$
    and $L_1(\mathbf{x}) = L_2(\mathbf{x}) = \cdots = L_k^{(i)}(\mathbf{x}) = 0$; *using* S *as start solutions*]
end for;
use L to define $\pi : \mathbb{C}^n \to \mathbb{C}^{k+1}$ as in (4.2); [*projection operator*]
let $\phi(\mathbf{x}) = z_{k+1}$, where $z = \pi(\mathbf{x})$; [*definition of* ϕ *in Theorem* 3.6]
for $i = 0, 1, 2$ compute $s_i := \sum_{\mathbf{x} \in S^{(i)}} \phi(\mathbf{x})$; [*sum* $(k+1)$*st coordinate*]
find a, b such that $s_i = a + bc_0^{(i)}$ for $i = 0, 1$; [*linear interpolation of trace*]
return $(|s_2 - (a + bc_0^{(2)})| < \epsilon)$. [*the comparison certifies*]

The justification for this algorithm is Theorem 3.6. The first k coordinates of $\pi(\mathbf{x})$ in (4.2) are the generic projection required by the theorem and the $(k+1)$st coordinate is the linear function ϕ. We test linearity of the trace by sampling in a generic direction: the kth coordinate of $\pi(\mathbf{x})$ suffices due to the genericity of the coefficients used to define it. By the theorem, linearity implies that the set of points S is the union of witness points for irreducible sets whose degrees sum to d, while by assumption, monodromy has found that all the points are in one irreducible set. Thus, there is one set and its degree is d.

4.3. An integrated decomposition algorithm. The linear trace test can be used as a simple replacement for the filtering polynomials used in our earlier papers [27, 28, 29], but additional efficiencies can be gained by integrating the technique more deeply into the algorithms.

First, we can improve the termination condition for the monodromy method. Previously, we continued to compute monodromy loops until either all points are connected into one group or until some preset number of consecutive loops fails to find any new connections. Setting the number of these stable loops too high is costly, while too low means that some connections could be missed. With the linear trace test, one can determine when a group is complete and immediately remove it from further iterations. Once the number of uncertified groups is reduced to a small number, combinations of them can be examined to discover which ones sum to form linear traces, thereby completing the decomposition without further monodromy loops.

In this vein, it is possible to perform the decomposition only using linear traces, as is done for a single multivariate polynomial in [6, 20]. However, without monodromy, the algorithm is combinatorial and is likely to be too expensive for high degrees. The use of traces in [21] (with predecessors [22, 23, 24]) is followed by linear algebra techniques. Recently, monodromy and traces have been combined to factor a single multivariate polynomial in [3].

The factorization of a single multivariate polynomial can be regarded as a special case of the decomposition of the solution sets of polynomial systems. For this general problem, we indicate a second improvement. In implementing the monodromy algorithm in [29], it is worthwhile to compute the linear span of the components as we described in [28]. Generic points that lie in different spans lie on different irreducible components, so we have only to execute the monodromy starting at points that lie in the same linear span. Also, the restriction to the linear span will give a speedup when there is a gap between the dimension of the linear span and the dimension of the ambient space.

Finally, the main decomposition algorithm [27] requires a test to determine if a generic point obtained from the embedding algorithm at dimension k is a member of some irreducible set of dimension greater than k. Originally, we used the interpolating polynomials for these sets to determine membership, but as in [28], it is possible to use a homotopy test of membership. Using the homotopy membership test for higher dimensional sets and certifying irreducible groups by linear traces, we eliminate completely the expensive and numerically difficult step of computing interpolating polynomials, which represents a big improvement in our overall algorithm.

5. Interpolation algorithms via traces. As just mentioned, the computation of interpolating polynomials is no longer required to complete the numerical irreducible decomposition. Nevertheless, in the case that one still wishes to compute such polynomials, the higher order traces can be useful, as we show in this section. For components of low degree and span, interpolating polynomials can be competitive with a homotopy membership test.

Our techniques for computing interpolating polynomials can be briefly summarized as follows. Since the witness points for a component lie on a linear slice, they can be marched forward together to compute a structured grid of sample points. In [29], a "bootstrapping" technique was used to construct the Newton form of the interpolating polynomial. Here, by using traces, we eliminate the expense of extra samples for bootstrapping and apply Newton interpolation directly. Finally, using the Newton identities, we reduce the number of samples to the number of monomials, which is optimal.

5.1. Newton interpolation with divided differences. To interpolate a bivariate function $f(x, y)$ with a polynomial $p(x, y)$ of degree d, we need to sample

the function at points (a_i, b_j), for all i, j: $0 \leq i + j \leq d$. The Newton form of the interpolation polynomial $p(x, y)$ is classical (see, e.g., [12, 16]):

$$p(x,y) = \sum_{k=0}^{d} \sum_{l=0}^{d-k} f[a_0 \cdots a_k; b_0 \cdots b_l] \prod_{i=0}^{k-1} (x - a_i) \prod_{j=0}^{l-1} (y - b_j). \tag{5.1}$$

The coefficients $f[a_0 \cdots a_k; b_0 \cdots b_l]$ are divided differences, defined inductively. Starting with $f[a_k; b_l] = f(a_k, b_l)$, all divided differences are generated by

$$f[a_0 a_1 \cdots a_k; b_l] = \frac{f[a_0 a_1 \cdots a_{k-1}; b_l] - f[a_1 a_2 \cdots a_k; b_l]}{a_0 - a_k} \tag{5.2}$$

and

$$f[a_0 a_1 \cdots a_k; b_0 b_1 \cdots b_l]$$

$$= \frac{f[a_0 a_1 \cdots a_k; b_0 b_1 \cdots b_{l-1}] - f[a_0 a_1 \cdots a_k; b_1 b_2 \cdots b_l]}{b_0 - b_l}, \tag{5.3}$$

for $k = 0, 1, \ldots, d$ and $l = 0, 1, \ldots, d - k$. The efficient computation of divided differences is organized in a table, requiring only one vector of elements to store. Generalizing (5.1) to any number of variables is burdened only by notation.

The direct application of Newton interpolation is prevented because the interpolation points must lie on a grid structured for *all* directions. When we sample curves or surfaces we always have one last component which is different for all samples. To overcome this we may apply a "bootstrapping" technique. We explain the idea in the case of two variables. For $x = a_k$, we construct a univariate polynomial $p(y)$ interpolating through the roots. Note that at those roots, y is usually different from the chosen grid points b_l. Once we have $p(y)$, we use it to find $f(a_k, b_l) = p(b_l)$, and we have a complete structured grid on which the above formulas (5.1) apply. This construction generalizes to the Newton form of the interpolating polynomial to represent any surface of any degree and dimension. It was implemented and used in [29] to certify groupings predicted by the monodromy algorithm. We provide an alternative to the bootstrapping technique using traces, as explained next.

5.2. The trace form with a complete grid. With traces, the classical multivariate interpolation schemes with generalized divided differences are directly applicable. We will show how to interpolate with a polynomial p of degree d in three variables (x, y, z), where p is expressed like

$$p(x,y,z) = z^d - t_1(x,y) z^{d-1} + t_2(x,y) z^{d-2} - \cdots + (-1)^d t_d(x,y), \tag{5.4}$$

where $t_i(x, y)$ is the ith trace with $\deg(t_i(x, y)) = i$. To represent the polynomials $t_i(x, y)$ we use the Newton form (5.1) with coefficients constructed with divided differences, given in formulas (5.2) and (5.3).

The major cost in the construction of the interpolating polynomial is the number of required sample points. While the number of monomials grows exponentially as the degree d and dimension k of the irreducible component increases, the number of samples in a complete grid grows as $d(d+1)^k$, which is much faster than the number of monomials, as illustrated in Table 5.1.

TABLE 5.1

Number of terms for degrees d and dimensions k, respectively, increasing in the rows and columns, versus $d(d+1)^k$, which is the number of samples needed for the trace form using the full grid, without exploiting Newton identities.

Number of monomials					Number of samples				
$d\backslash k$	2	3	4	5	$d\backslash k$	2	3	4	5
1	3	4	5	6	1	2	4	8	16
2	6	10	15	21	2	6	18	54	162
3	10	20	35	56	3	12	48	192	768
4	15	35	70	126	4	20	100	500	2500
5	21	56	126	252	5	30	180	1080	6480

5.3. Using the Newton identities. Ideally, we would like to take no more samples than the number of monomials. Exploiting the Newton identities (2.5), we show how to achieve this goal on an example, the interpolation of a planar quartic. Figure 5.1 gives a schematic representation of the grid of sample points.

In the interpolation of a planar quartic, we compute the four traces consecutively. To compute the second trace t_2, we may already use t_1, and to compute t_3, we already dispose of t_1, t_2, and for t_4, we make use of t_1, t_2, and t_3. We show how this saves sample points:

1. At $x = a_2$, instead of four, we compute three samples (a_2, b_{21}), (a_2, b_{22}), (a_2, b_{23}) and compute b_{24} using the first trace t_1, evaluated at $x = a_2$:

$$b_{24} := t_1(a_2) - b_{21} - b_{22} - b_{23}. \tag{5.5}$$

2. At $x = a_3$, we know already the coefficient of the Newton forms of t_1 and t_2 and use continuation only for two samples: (a_3, b_{31}) and (a_3, b_{32}). For the values b_{33} and b_{34} we solve the system

$$\begin{cases} t_1(a_3) = b_{31} + b_{32} + b_{33} + b_{34}, \\ t_2(a_3) = b_{31}b_{32} + b_{31}b_{33} + b_{31}b_{34} + b_{32}b_{33} + b_{32}b_{34} + b_{33}b_{34}. \end{cases} \tag{5.6}$$

With the Newton identities (2.5), we compute from the values of the elementary symmetric functions $(t_1(a_3), t_2(a_3))$ the power sums for $x = a_3$: $(s_1(a_3), s_2(a_3))$. This means that the b_{ij}'s satisfy

$$\begin{cases} s_1(a_3) = b_{31} + b_{32} + b_{33} + b_{34}, \\ s_2(a_3) = b_{31}^2 + b_{32}^2 + b_{33}^2 + b_{34}^2 \end{cases} \tag{5.7}$$

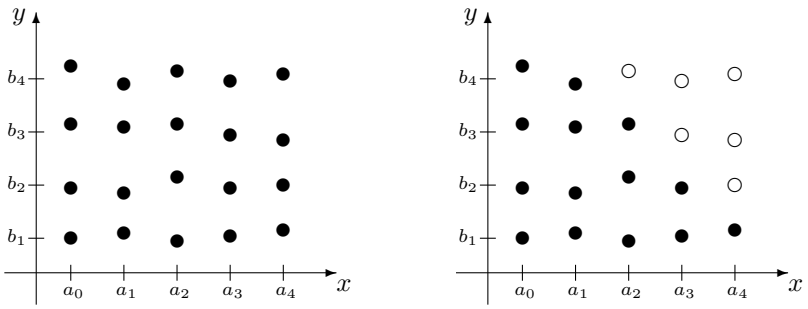


FIG. 5.1. *Two grids of sample points to interpolate a planar quartic: the grid on the left is complete, while on the right we find the white dots using Newton identities. The semiregularity of the grid (same x-value in one column, different y-values in each row) is typical.*

or

$$\begin{cases} b_{33} + b_{34} = s_1(a_3) - b_{31} - b_{32}, \\ b_{33}^2 + b_{34}^2 = s_2(a_3) - b_{31}^2 - b_{32}^2. \end{cases} \tag{5.8}$$

Let $\widetilde{s}_1(a_3) = s_1(a_3) - b_{31} - b_{32}$ and $\widetilde{s}_2(a_3) = s_2(a_3) - b_{31}^2 - b_{32}^2$. Then we invoke the Newton identities (2.5) to compute from the modified powers sums $(\widetilde{s}_1(a_3), \widetilde{s}_2(a_3))$ evaluated at a_3 into the values of the elementary symmetric functions at a_3: $(\widetilde{t}_1(a_3), \widetilde{t}_2(a_3))$. With $(\widetilde{t}_1(a_3), \widetilde{t}_2(a_3))$ we define

$$y^2 - t_1(a_3)y + t_2(a_3) = y^2 - (b_{33} + b_{34})y + b_{33}b_{34} \tag{5.9}$$

$$= (y - b_{33})(y - b_{34}) \tag{5.10}$$

$$= 0. \tag{5.11}$$

Thus, finding the missing samples (white dots on Figure 5.1 for $x = a_3$) has been reduced to solving a quadratic univariate equation.

3. At $x = a_4$ we apply the Newton identities also twice to construct a univariate equation of degree three to find the missing samples.

This procedure generalizes to any degree and any dimension, requiring only as many samples as the number of monomials.

To find the roots of univariate polynomials, we use the method of Weierstrass (also known as Durand–Kerner), described in [17] as "quite effective and increasingly popular." Convergence is global and quadratic in the limit [18]. Our implementation is basic; see [13] for algorithmic improvements to this method.

5.4. Numerical aspects and experiments. In this section we first compare our new algorithms with our approach to interpolation in [27], where we solved the linear system of interpolation conditions directly. Then we illustrate the numerical stability of the new algorithms on a test polynomial.

Compared to the direct approach of [27], we first observe an improved conditioning of the interpolation problem when a structured grid of samples is used. This improved conditioning leads to more accurate results regardless of the interpolation algorithm. In an unstructured grid, errors on the samples creep in with greater fluctuation than on a structured grid, where the ith coordinate, $i = 1, 2, \ldots, k$, is the same for all samples from a k-dimensional component. The second advantage of our new algorithms concerns time. Using divided differences to solve the linear system of N interpolation conditions requires $O(N^2)$ operations instead of $O(N^3)$ for plain Gaussian elimination or QR factorization.

The bootstrapping method for Newton interpolation we used in [29] and the basic interpolation with traces both require more samples than the number of monomials; see Table 5.1. Exploiting the Newton identities, we get the complete interpolating polynomial with an optimal number of samples. We next describe an experiment to illustrate that this exploitation is numerically stable.

We consider to interpolate the polynomial

$$p(x, y) = \sum_{0 \leq i+j \leq d} x^i y^j, \quad \text{for } d = 2, 3, \ldots, 10. \tag{5.12}$$

To this polynomial we apply our general implementation, treating $p(x, y)$ as a polynomial system. Assuming monodromy has shown the d generic points to belong to the

TABLE 5.2

Numerics on the dense polynomial with unit coefficients for degrees d from 2 up to 10, on square and triangular grids (i.e., without and with the exploitation of Newton identities) and in the last column for the linear traces. Because only magnitudes matter, 8.559E–16 is shortened to –16, and $-\infty$ stands for a zero residual. We list the accuracy of the grid (eps) and the magnitude of the highest residual after evaluation at the grid (gres) and at some test points (tres).

	Square grid			Triangular grid			Linear trace		
d	eps	gres	tres	eps	gres	tres	eps	gres	tres
2	–16	–15	–13	–15	–16	–13	–16	$-\infty$	–15
3	–15	–15	–14	–16	–15	–15	–16	$-\infty$	–15
4	–14	–14	–13	–15	–15	–14	–16	$-\infty$	–15
5	–14	–13	–12	–15	–15	–14	–16	$-\infty$	–15
6	–15	–13	–14	–16	–14	–12	–16	$-\infty$	–15
7	–15	–13	–10	–15	–14	–12	–16	$-\infty$	–15
8	–15	–12	–13	–15	–13	–11	–15	$-\infty$	–15
9	–15	–12	–14	–15	–13	–11	–16	$-\infty$	–15
10	–15	–11	–08	–14	–12	–10	–16	–16	–15

same connected component, as a test (because in this particular case there is nothing to certify) we compare three methods of certification:

1. construction of the complete trace form on a square grid;
2. construction of the complete trace form on a triangular grid, exploiting Newton identities;
3. construction of only the linear trace.

As in the general method for polynomial systems, we compute the magnitude of the highest value the interpolation polynomial returns at the grid and at some extra test points sampled from the component. These residuals would all be zero on exact data and with exact arithmetic. Due to approximate samples—accurate up to machine precision—and floating-point arithmetic, we observe errors when evaluating at the grid and at the test points; see Table 5.2.

In Table 5.2 we see an increasing loss of precision as the degree increases, for both with and without the exploitation of the Newton identities. This loss is due to the intrinsic complexity of high degree polynomials. There is no significant difference between the first two methods. From the last three columns of Table 5.2 we observe no error propagation; i.e., residuals at test points are of the same magnitude as the errors on the sample points. Compared to the complete trace form, linear traces are tolerant to approximate data and require no extra precision, at least not for the case of moderate degrees.

In Table 5.3 we list timings (on a Pentium III 800 MHz Linux machine) for the three certification methods. We see an efficiency gain with the exploitation of the Newton identities and a drastic difference when only the linear traces are computed.

In this experiment, the exploitation of the Newton identities is beneficial: fewer samples are needed and the loss of accuracy is not significantly different from the basic construction. However, for higher degrees, in situations where multiprecision arithmetic is necessary we experienced severe losses of accuracy. In particular, we applied this exploitation of Newton identities to one of the curves of degree 16 arising in the cyclic 8-roots problem (described in greater detail below). Even if the roots of the univariate polynomials were computed at full precision, the evaluation of those roots at high degree polynomials turned out to be insufficient to reach the same accuracy as without exploitation of the Newton identities. Based on these experiences, we recommend the exploitation of Newton identities only for moderate degrees.

TABLE 5.3

Timings on the dense polynomial with unit coefficients, for degrees d from 2 to 10 to construct the complete interpolator plainly on a square grid, exploiting Newton identities on a triangular grid. In the last column are timings to construct linear traces.

d	Square grid	Triangular grid	Linear traces
2	40 ms	50 ms	20 ms
3	110 ms	70 ms	30 ms
4	210 ms	170 ms	60 ms
5	420 ms	300 ms	110 ms
6	890 ms	480 ms	130 ms
7	1s 540 ms	760 ms	240 ms
8	2s 570 ms	1s 260 ms	320 ms
9	3s 730 ms	1s 800 ms	410 ms
10	4s 520 ms	3s 40 ms	600 ms

6. Applications. The algorithms have been implemented in a separate module of PHCpack [34], available online from http://www.math.uic.edu/∼jan and recently described in [30]. All reported timings are user cpu times on a Pentium III 800 Mhz Linux machine.

In the applications we consider here, the positive dimensional components are pure dimensional. Therefore, we restrict the numerical irreducible decomposition of [27] to the following three stages:

1. computation of the generic points with the embedding of [26];
2. application of monodromy [29], grouping the generic points which belong to the same irreducible component;
3. validation of the breakup predicted by monodromy by interpolation
   (a) either with the complete polynomials,
   (b) or with only the linear traces.

The methods presented in this paper only affect the last stage. We report timings for the other two stages to show the overall impact of our new approach. The experiments do not exploit the possible improvements that could result by integrating linear traces into the monodromy phase, as discussed in section 4.3.

6.1. The cyclic 8-roots and 9-roots problems. In [27] we had to limit ourselves to the *reduced* versions of those problems. With the recent advances in the decomposition algorithms we can factor the components into irreducibles *without* recourse to multiprecision arithmetic.

6.1.1. The cyclic 8-roots problem. In this section we confirm earlier results obtained in [2] by computer algebra methods. The timings for the three stages are as follows:

1. The computation of all 144 generic points on the one-dimensional components using the embedding in [26] takes 1h 12m 42s 650ms. Note that this computation also contains the calculation of the start solutions of paths leading to all isolated roots.
2. The set of 144 generic points breaks up into 8 subsets of 16 points and 8 subsets of 2 points. The monodromy breakup algorithm of [29] requires 6m 24s 930ms.
3. (a) In [29] we did the validation constructing interpolating polynomials, using standard arithmetic for the eight quadrics and using 32 decimal places for the eight curves of degree 16. This whole process took 41m 54s 780ms to complete. So stage three accounts for 35% of the total execution time.

TABLE 6.1

Numerical results of the certification of cyclic 8-roots. The columns contain the degree d, the accuracy of the samples in the grid, the largest value of the linear trace polynomial evaluated at the grid res, and the absolute value of the difference between the predicted and computed sum of the roots.

d	Accuracy of samples	Residual at grid	Difference at test pts
2	6.055E–16	1.110E–16	4.929E–14
2	4.733E–16	2.776E–16	4.308E–14
2	1.608E–15	8.882E–16	8.882E–15
2	4.143E–16	5.551E–17	5.551E–15
2	1.812E–15	1.776E–15	1.954E–14
2	1.095E–15	8.882E–16	3.642E–14
2	5.403E–16	2.220E–16	8.238E–14
2	1.815E–15	5.551E–16	2.132E–14
16	1.318E–14	6.661E–16	2.665E–14
16	6.182E–14	8.882E–16	1.199E–13
16	2.991E–14	8.882E–16	9.326E–14
16	1.239E–13	8.882E–16	9.859E–14
16	1.667E–13	8.882E–16	2.167E–13
16	8.589E–14	8.882E–16	7.372E–14
16	9.708E–15	2.220E–16	1.030E–13
16	8.168E–15	1.776E–15	5.418E–14

(b) With linear traces we need fewer samples, and expensive multiprecision arithmetic can be avoided. This interpolation takes only 27s 540ms. Compared to the time needed in stage three with the complete interpolation polynomial, this runs more than 150 times faster. The total time for the three stages reduces from about two hours to one hour and 18 minutes.

We summarize the numerical results of this calculation in Table 6.1. See [28] for the computation of the linear span of the component.

We wish to point out that the sample points are distributed widely to have a good conditioning of the interpolating polynomial. Comparing the second column with column four in Table 6.1, we observe that there is hardly any loss of accuracy for any of the roots; neither the quadrics nor the 16th degree polynomials show any significant loss.

6.1.2. The cyclic 9-roots problem. This problem has been solved with Gröbner basis methods in [4]. The timings for the three stages with our approach are as follows:

1. To compute generic points, we used the mixed-volume calculator of T.Y. Li and X. Li [15] to set up the homotopy to solve the embedding of [26]. The mixed-volume computation took 13m 4s 540ms, and the total time to compute all 20,376 paths of the embedding for two-dimensional components was 9h 11m 27s 820ms. Only 18 of these paths land on two-dimensional components: the other paths either diverge to infinity or are paths destined to lead to the isolated solutions at a later stage of the embedding technique.
2. The set of 18 generic points breaks up into 6 subsets of 3 points each. The monodromy breakup algorithm of [29] requires 2m 32s 400ms.
3. (a) In [29] we did the validation constructing interpolating polynomials, with 32 decimal places, which took 14m 56s 570ms.
   (b) The validation using only linear traces took 9s 350ms. Details are in Table 6.2.

TABLE 6.2

Numerical results of the certification of cyclic 9-roots. The columns contain the degree d, the accuracy of the samples in the grid, the largest value of the linear trace polynomial evaluated at the grid res, and the absolute value of the difference between the predicted and computed sum of the roots.

d	Accuracy of samples	Residual at grid	Difference at test pts
3	3.507E-13	0.000E+00	6.864E-14
3	5.118E-13	2.776E-17	1.456E-13
3	9.343E-13	1.388E-17	2.313E-13
3	1.529E-13	5.551E-17	3.583E-14
3	6.984E-13	0.000E+00	8.460E-14
3	1.165E-13	0.000E+00	4.080E-15

6.2. A moving Stewart–Gough platform. A generic Stewart–Gough platform mechanism has forty isolated solutions, first established by continuation [19] and later proven analytically [10, 35]. A special case of this mechanism, due to Griffis and Duffy, has a solution curve of degree forty. This means that a Griffis–Duffy platform has a one-degree-of-freedom motion, whereas a generic Stewart–Gough platform is rigid. Husty and Karger [11] pointed out this fact and also identified a more special Griffis–Duffy platform for which the solution curve breaks up into lower degree irreducible components. We treat both cases here with our numerical methods and briefly discuss some differences we found from Husty and Karger's results.

For the general Griffis–Duffy platform, which herein we call case A, the solution set consists of 12 lines and one irreducible curve of degree 28. The lines all correspond to degeneracies that do not give actual assembly configurations of the mechanism. The specialized case B also has 12 degenerate lines, but now the curve of degree 28 breaks up into lower degree irreducible components—four sextics and a quartic. The timings for the three stages in our approach are as follows:

1. To compute forty generic points using the embedding of [26] requires 52s 490ms for case A and 55s 810ms for case B.
2. For case A, the monodromy algorithm of [29] takes 33s 430ms to predict a single component of degree 28. For case B, it takes 27s 630ms for the monodromy algorithm to group the 28 generic points into five sets; four of the five have cardinality six, and one set has four points.
3. (a) For case A, the validation with Newton interpolation for the curve of degree 28 requires multiprecision (with 64 decimal places) and 812 samples (for 435 monomials), and it completes in 1h 19m 13s 110ms. For case B, using 32 decimal places in constructing the complete interpolating polynomials with divided differences takes 2m 34s 50ms.
   (b) With linear traces, case A takes 4s 750ms and case B requires 4s 320ms.

This example shows several advantages of using linear traces for validation of the monodromy breakup. Compared to the use of interpolating polynomials, the computation time is not only drastically reduced when using linear traces, but also becomes nearly identical for both cases A and B. Interpolating a degree 28 polynomial in two variables for case A is expensive and requires high precision arithmetic. In fact, because of numerical instability of traces in this case, we used the bootstrapping Newton technique as in [29] to construct the complete interpolation filter. Case B, comprised of five irreducible curves whose degrees sum to 28, is much more tractable by interpolating polynomials, but still the use of only linear traces is much superior. Table 6.3 lists numerical results of the methods, showing that the linear traces are quite stable using only double precision arithmetic. In summary, compared to inter-

TABLE 6.3

Numerical results of the certification of case A ($d = 28$) and case B ($d = 6, 6, 6, 6, 4$) for the irreducible curves occurring in moving Stewart–Gough platforms. The columns contain the accuracy of the samples in the grid, the largest value of the interpolating filter (or the linear trace) evaluated at the grid, and the residual at the test points. With linear traces, we list the difference between the predicted and computed sum at the test points.

	With complete interpolation			Using linear traces only		
d	Accuracy of samples	Residual at grid	Residual at test pts	Accuracy of samples	Residual at grid	Difference at test pts
28	1.316E–59	3.800E–37	1.107E–20	4.013E–13	1.791E–12	1.791E–12
6	3.259E–28	2.800E–27	1.020E–20	1.272E–12	2.442E–15	4.694E–13
6	5.243E–29	8.495E–28	6.416E–21	9.944E–13	1.332E–15	3.659E–13
6	1.152E–28	6.000E–30	2.502E–21	8.660E–13	2.220E–16	5.853E–13
6	4.730E–29	2.540E–28	4.936E–20	7.438E–14	2.220E–15	1.083E–11
4	4.758E–30	4.300E–31	3.357E–27	1.063E–14	2.220E–16	3.408E–14

polating polynomials, our **Certify** algorithm, based on linear traces, eliminates the large fluctuation in timings with superior numerical stability and efficiency.

We show how we can observe the propagation of roundoff errors. Compare the accuracy of the samples with the residuals at the test points in Table 6.3. For case A, the accuracy of the samples is 10^{-59}, while the residuals at the test points evaluate to 10^{-20}. During the calculation we lost about 30 decimal places. The loss in case B is more modest, between 7 and 9 decimal places, and makes the difference in exponents in the second and fourth columns of Table 6.3. With linear traces, we observe from the data in Table 6.3 that we lose at most 3 decimal places.

While the reduction in execution times may turn modest in the near future as more and faster machines will become even more widely available, the major benefit of using linear traces is that reliable results are solely obtained with standard machine arithmetic, that is, without using any multiprecision numbers. This means that errors on the coefficients of the input system that are less than the standard machine precision can be neglected and the algorithm is numerically stable.

We conclude this section with some remarks not related to numerical performance, but rather concerning the decomposition itself. The decompositions we have computed for the Griffis–Duffy platforms differ from the results obtained by Husty and Karger in [11]; one discrepancy is reconcilable, but others are not. First, for the general example, case A, we find a degree 28 curve, which at first seems to conflict with their result of a degree 20 curve. This is not, however, a contradiction, because we have analyzed the curve in the full space of rotation and translation (represented in Study coordinates). The degree falls to 20 when the curve is projected onto its rotational component only, as done by Husty and Karger. However, we also find a significant difference for the specialized case B: Husty and Karger do not list one of the five irreducible components in their analysis. Their approach, using a combination of special reasoning and computer algebra, gives some extra insight in some respects, but our automated numerical method is less subject to human error. To tackle difficult problems, it will sometimes be beneficial to use both numeric and symbolic processes. In this case, knowing about the existence of the fifth irreducible component and seeing its numerical structure, one might return to the symbolic approach to further elucidate it. (Of course, one might also pursue a completely automated symbolic approach as well, but that is another story.)

We refer to [32] for a description of this and other applications of our approach to polynomial systems in mechanism design.

7. Conclusions. In [27], we presented a numerical algorithm to decompose solution sets of polynomial systems into irreducible components of various dimensions and degrees. The main drawback of that algorithm is numerical instability on components of high degree, due to the reliance on interpolating polynomials to filter generic points into irreducible components. To deal with this difficulty, multiprecision arithmetic was used whenever high degrees were encountered, requiring a high accuracy for the input coefficients of the polynomial systems and requiring much more computer time than standard precision for each arithmetic operation. While the sequels [28] and [29] lessened the need for high precision arithmetic to some extent, it is only in this paper that we can present a numerically stable decomposition algorithm to solve the cornerstone problem in numerical algebraic geometry [33].

We summarize how standard machine arithmetic can be employed throughout the numerical irreducible decomposition algorithm. The sequence of homotopies of [26] produces generic points on every positive dimensional component, mixed with "junk": points on higher solution components. To separate those junk points from the generic points, we now propose to use the homotopy membership test of [28] instead of the filtering polynomials in [27]. Unlike with high degree polynomials, this homotopy membership test does not require multiprecision arithmetic. Also the monodromy algorithm of [29] predicts the breakup of pure dimensional components using only standard machine arithmetic. With this paper, we finally remove any need for interpolating polynomials, because linear traces suffice to certify the predicted breakup. As linear polynomials are tolerant to roundoff and efficient to interpolate, our decomposition algorithms have gained significantly in speed and robustness. Practical evidence for these claims is provided in the reports on benchmark applications.

Acknowledgment. We would like to thank the referees for their helpful remarks.

REFERENCES

[1] E. Arbarello, M. Cornalba, P.A. Griffiths, and J. Harris, *Geometry of Algebraic Curves*, Vol. I., Grundlehren Math. Wiss. 267, Springer–Verlag, New York, 1985.

[2] G. Björck and R. Fröberg, *Methods to "divide out" certain solutions from systems of algebraic equations, applied to find all cyclic* 8*-roots*, in Analysis, Algebra and Computers in Mathematical Research, Lecture Notes in Pure and Appl. Math. 156, M. Gyllenberg and L.E. Persson, eds., Dekker, New York, 1994, pp. 57–70.

[3] R.M. Corless, A. Galligo, I.S. Kotsireas, and S.M. Watt, *A geometric-numeric algorithm for factoring multivariate polynomials*, in Proceedings of the 2002 International Symposium on Symbolic and Algebraic Computation (ISSAC 2002), T. Mora, ed., ACM, New York, to appear.

[4] J.C. Faugère, *A new efficient algorithm for computing Gröbner bases (F_4)*, J. Pure Appl. Algebra, 139 (1999), pp. 61–88

[5] W. Fulton, *Intersection Theory*, Ergeb. Math. Grenzgeb. (3) 2, Springer–Verlag, Berlin, 1984.

[6] A. Galligo and D. Rupprecht, *Semi-numerical determination of irreducible branches of a reduced space curve*, in Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation (ISSAC 2001), B. Mourrain, ed., ACM, New York, 2001, pp. 137–142.

[7] M. Goresky and R. Macpherson, *Stratified Morse Theory*, Ergeb. Math. Grenzgeb. 14, Springer–Verlag, Berlin, 1988.

[8] R.C. Gunning, *Introduction to Holomorphic Functions of Several Complex Variables. Vol.* II. *Local Theory*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1990.

[9] R.C. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[10] M.L. Husty, *An algorithm for solving the direct kinematics of general Stewart-Gough platforms*, Mech. Mach. Theory, 31 (1996), pp. 365–380.

[11] M.L. Husty and A. Karger, *Self-motions of Griffis-Duffy type parallel manipulators*, in

Proceedings of the 2000 IEEE International Conference on Robotics and Automation, San Francisco, CA, CDROM, IEEE, Piscataway, NJ, 2000.

[12] E. Isaacson and H.B. Keller, *Analysis of Numerical Methods*, Dover, New York, 1994.

[13] P. Kirrinnis, *Fast numerical improvement of factors of polynomials and of partial fractions*, in Proceedings of ISSAC'98, O. Gloor, ed., ACM, New York, 1998, pp. 260–267.

[14] S.L. Kleiman, *Tangency and duality*, in Proceedings of the 1984 Vancouver Conference in Algebraic Geometry, CMS Conf. Proc. 6, J.B. Carrell, A.V. Geramita, and P. Russell, eds., AMS, Providence, RI, 1986, pp. 163–225.

[15] T.Y. Li and X. Li, *Finding mixed cells in the mixed volume computation*, Found. Comput. Math., 1 (2001), pp. 161–181.

[16] M. Mignotte and D. Ştefănescu, *Polynomials. An Algorithmic Approach*, Springer–Verlag, Singapore, 1999.

[17] V.V. Pan, *Solving a polynomial equation: Some history and recent progress*, SIAM Rev., 39 (1997), pp. 187–220.

[18] L. Pasquini and D. Trigiante, *A globally convergent method for simultaneously finding polynomial roots*, Math. Comp., 44 (1985), pp. 135–149.

[19] M. Raghavan, *The Stewart platform of general geometry has* 40 *configurations*, ASME J. Mech. Design, 115 (1993), pp. 277–282.

[20] D. Rupprecht, *Semi-numerical absolute factorization of polynomials with integer coefficients*, J. Symbolic Comput., to appear.

[21] T. Sasaki, *Approximate multivariate polynomial factorization based on zero-sum relations*, in Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation (ISSAC 2001), B. Mourrain, ed., ACM, New York, 2001, pp. 284–291.

[22] T. Sasaki, T. Saito, and T. Hilano, *Analysis of approximate factorization algorithm* I, Japan J. Indust. Appl. Math., 9 (1992), pp. 351–368.

[23] T. Sasaki and M. Sasaki, *A unified method for multivariate polynomial factorizations*, Japan J. Indust. Appl. Math., 10 (1993), pp. 21–39.

[24] T. Sasaki, M. Suzuki, M. Kolár, and M. Sasaki, *Approximate factorization of multivariate polynomials and absolute irreducibility testing*, Japan J. Indust. Appl. Math., 8 (1991), pp. 357–375.

[25] B. Shiffman and A.J. Sommese, *Vanishing Theorems on Complex Manifolds*, Prog. Math. 56, Birkhäuser, Boston, 1985.

[26] A.J. Sommese and J. Verschelde, *Numerical homotopies to compute generic points on positive dimensional algebraic sets*, J. Complexity, 16 (2000), pp. 572–602.

[27] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Numerical decomposition of the solution sets of polynomial systems into irreducible components*, SIAM J. Numer. Anal., 38 (2001), pp. 2022–2046.

[28] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Numerical irreducible decomposition using projections from points on the components*, in Symbolic Computation: Solving Equations in Algebra, Geometry, and Engineering, Contemp. Math. 286, E.L. Green, S. Hoşten, R.C. Laubenbacher, and V. Powers, eds., AMS, Providence, RI, 2001, pp. 37–51.

[29] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Using monodromy to decompose solution sets of polynomial systems into irreducible components*, in Application of Algebraic Geometry to Coding Theory, Physics and Computation, C. Ciliberto, F. Hirzebruch, R. Miranda, and M. Teicher, eds., Kluwer Academic Publishers, Norwell, MA, 2001, pp. 297–315.

[30] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Numerical irreducible decomposition using PHCpack*, in Mathematics and Visualization, M. Joswig and N. Takayama, eds., Springer–Verlag, to appear.

[31] A.J. Sommese, J. Verschelde, and C.W. Wampler, *A method for tracking singular paths with application to the numerical irreducible decomposition*, in Algebraic Geometry, a Volume in Memory of Paolo Francia, M.C. Beltrametti, F. Catanese, C. Ciliberto, A. Lanteri, C. Pedrini. and W. de Gruyter, eds., to appear.

[32] A.J. Sommese, J. Verschelde, and C.W. Wampler, *Advances in polynomial continuation for solving problems in kinematics*, in Proceedings of the ASME Design Engineering Technical Conference, Montreal, 2002, CD-ROM, ASME International, Fairfield, NJ, 2002.

[33] A.J. Sommese and C.W. Wampler, *Numerical algebraic geometry*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, J. Renegar, M. Shub, and S. Smale, eds., AMS, Providence, RI, 1996, pp. 749–763.

[34] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.

[35] C.W. Wampler, *Forward displacement analysis of general six-in-parallel SPS (Stewart) platform manipulators using soma coordinates*, Mech. Mach. Theory, 31 (1996), pp. 331–337.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2047–2068

hp-APPROXIMATION THEORY FOR BDFM AND RT FINITE ELEMENTS ON QUADRILATERALS*

MARK AINSWORTH† AND KATIA PINCHEDEZ†

Abstract. We study approximation properties of *hp*-finite element subspaces of $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$ on a polygonal domain Ω using Brezzi–Douglas–Fortin–Marini (**BDFM**) or Raviart–Thomas (**RT**) elements. Approximation theoretic results are derived for the *hp*-version finite element method on non-quasi-uniform meshes of quadrilateral elements with *hanging nodes* for functions belonging to weighted Sobolev spaces $\mathbf{H}^{s,\ell}_{\omega}(\Omega)$ and the countably normed spaces $\boldsymbol{\mathcal{B}}^{\ell}_{w}(\Omega)$. These results culminate in a proof of the characteristic exponential convergence property of the *hp*-version finite element method on suitably designed meshes under similar conditions needed for the analysis of the $\mathbf{H}^1(\Omega)$ case. By way of illustration, exponential convergence rates are deduced for mixed *hp*-approximation of flow in porous media.

Key words. mixed *hp*-finite elements, corner singularities, exponential convergence

AMS subject classifications. Primary, 65N30, 65N35, 65N15; Secondary, 73V05, 76M10, 78-08

PII. S0036142901391128

1. Introduction.

The variational formulation of several important classes of problem arising in science and engineering involves the space

$$\mathbf{H}(\mathrm{div},\Omega) \;=\; \left\{\; \boldsymbol{v}\in\mathbf{L}^2(\Omega):\; \mathrm{div}\,\boldsymbol{v}\in\mathsf{L}^2(\Omega)\right\}$$

or the related space $\mathbf{H}(\mathrm{rot},\Omega)$. We mention flow through porous media, time-harmonic Maxwell's equations and elastostatics as particular examples. The finite element approximation of such problems entails the construction of finite dimensional subspaces of $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$. Many schemes have been proposed, such as Raviart–Thomas (**RT**) elements, Brezzi–Douglas–Fortin–Marini (**BDFM**) elements, and the related Nédélec, or edge, elements, to name but a few. We refer the interested reader to the book of Brezzi and Fortin [9] for further information.

If the physical domain Ω is polygonal, then it is well known [11] that the solutions of the governing equations exhibit singularities in the neighborhood of the vertices of the domain. The lack of smoothness in the underlying solution manifests itself in the form of a degraded rate of convergence as the finite element subspace is enriched. The *h*-version of the finite element method [10] seeks convergence through reduction of the mesh-size h (either uniformly or adaptively), whereas the *p*-version of the finite element method is based on increasing the degree of the polynomial while maintaining a fixed mesh. Both versions give at best *algebraic rates of convergence* on a polygonal domain. However, a proper combination of each, the *hp*-version, may deliver *exponential rates of convergence* [12, 13]. We refer to the extensive survey article of Babuška and Suri [7], Babuška and Guo [2], or the book by Schwab [20] for more details.

Although a great deal is known of *hp*-approximation of problems posed over the Sobolev space $\mathsf{H}^1(\Omega)$, comparatively little is known in the case of $\mathbf{H}(\mathrm{div},\Omega)$ and

*Received by the editors June 20, 2001; accepted for publication (in revised form) May 9, 2002; published electronically December 13, 2002. This work was supported by the Engineering and Physical Sciences Research Council of Great Britain under grants GR/L90507 and GR/M59426.
http://www.siam.org/journals/sinum/40-6/39112.html
†Mathematics Department, Strathclyde University, 26 Richmond Street, Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk, ra.kpin@maths.strath.ac.uk).

$\mathbf{H}(\mathrm{rot},\Omega)$. Suri [22] and Milner and Suri [18] studied the properties of p-version finite element approximations of the spaces $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$ in two space dimensions with uniform order p on a regular mesh. Subsequently, the analysis was extended to the three dimensional case by Monk [19], working in the setting of Maxwell's equations again with uniform order p on a regular mesh, and then to the *hp*-version in two dimensions on a regular quasi-uniform mesh with uniform order p in [17]. One feature peculiar to the *hp*-version is the use of nonuniform polynomial degree distribution and graded, non-quasi-uniform meshes. Vardapetyan and Demkowicz [23] presented a construction allowing the generalization of certain types of Nédélec element to the *hp*-setting. These finite element subspaces of $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$ are part of a wider picture of discrete differential forms forming a bridge between the standard $\mathsf{H}^1(\Omega)$-conforming finite elements and $\mathsf{L}^2(\Omega)$-conforming elements [8, 14, 15, 16].

Despite increasing interest and use of *hp*-approximations of the spaces $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$, there has been no detailed study of the approximation properties. Specifically, there is no proof of the celebrated, characteristic, exponential convergence property of the *hp*-version. It is the aim of the present work to address this problem. Earlier work [21] dealt with *hp*-approximation on regular, globally quasi-uniform meshes based on regularity in standard Sobolev spaces. Here, we shall deal with approximation of functions belonging to the weighted Sobolev spaces $\mathbf{H}^{s,\ell}_{\omega}(\Omega)$ and the related spaces $\boldsymbol{\mathcal{B}}^{\ell}_{w}(\Omega)$, studied by Babuška and Guo [2, 3], that play a pivotal role in the analysis of the *hp*-version. We develop the approximation theoretic results needed to analyze the *hp*-version including non-quasi-uniform meshes with *hanging nodes*, culminating in a proof of exponential convergence for approximation of $\mathbf{H}(\mathrm{div},\Omega)$ and $\mathbf{H}(\mathrm{rot},\Omega)$ under the usual conditions assumed for the $\mathbf{H}^1(\Omega)$ case. By way of illustration, we deduce exponential convergence rates for mixed *hp*-finite element approximation of flow through porous media.

2. Model problem and its regularity. In the following, vector quantities shall always be represented by bold symbols and, for any real s, $[s]$ shall denote the integer part of s.

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain with vertices located at A_i, $i = 1,\ldots,M$. The Sobolev space $\mathsf{H}^r(\Omega)$, for $r \geq 0$ integer, consists of functions defined on Ω whose derivatives of order up to r are square integrable and is equipped with the usual norm and seminorm:

$$\|u\|^2_{\mathsf{H}^r(\Omega)} \quad \sum_{m=0}^{r} |u|^2_{\mathsf{H}^m(\Omega)}, \qquad |u|^2_{\mathsf{H}^r(\Omega)} \;=\; \| \,|\,\mathrm{D}^r\, u|\,\|^2_{\mathsf{L}^2(\Omega)},$$

where $|\,\mathrm{D}^r\, u|^2 = \sum_{i+j=r} |\,\mathrm{D}^{(i,j)}\, u|^2$. In particular, observe that $\mathsf{H}^0(\Omega) = \mathsf{L}^2(\Omega)$.

Next, let us introduce definitions and notations for certain weighted Sobolev spaces [12]. The distance between $\boldsymbol{x}$ and the vertex A_i is denoted by $r_i(\boldsymbol{x}) = |\boldsymbol{x} - A_i|$. Let $\omega = (\omega_1,\ldots,\omega_M)$ be an M-tuple of real numbers $0 \leq \omega_i < 1$ for $i = 1,\ldots,M$ and $\omega \neq 0$. We introduce weight functions defined by

$$\Phi_{\omega}(\boldsymbol{x}) \;=\; \textstyle\prod_{i=1}^{M} r_i^{\omega_i}(\boldsymbol{x}) \quad \text{and} \quad \Phi_{\omega\pm p}(\boldsymbol{x}) \;=\; \textstyle\prod_{i=1}^{M} r_i^{\omega_i \pm p}(\boldsymbol{x}),$$

where, for any integer p, the M-tuple $\omega \pm p$ is equal to $(\omega_1 \pm p,\ldots,\omega_M \pm p)$.

For integers $r \geq \ell \geq 0$, the weighted Sobolev space $\mathsf{H}^{r,\ell}_{\omega}(\Omega)$ denotes the completion of $\mathsf{C}^{\infty}(\overline{\Omega})$ under the norm defined by

$$\|u\|^2_{\mathsf{H}^{r,\ell}_\omega(\Omega)} \;=\; \|u\|^2_{\mathsf{H}^{\ell-1}(\Omega)} \;+\; |u|^2_{\mathsf{H}^{r,\ell}_\omega(\Omega)} \qquad \text{for } \ell \geq 1, \tag{1}$$

$$\|u\|^2_{\mathsf{H}^{r,0}_\omega(\Omega)} \;=\; \|u\|^2_{\mathsf{H}^{r}_\omega(\Omega)} \;=\; |u|^2_{\mathsf{H}^{r,0}_\omega(\Omega)} \qquad \text{for } \ell = 0, \tag{2}$$

where

$$|u|^2_{\mathsf{H}^{r,\ell}_\omega(\Omega)} \;=\; \sum_{k=\ell}^{r} \int_\Omega \left|\mathrm{D}^k\, u(\boldsymbol{x})\right|^2 \Phi^2_{\omega+k-\ell}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{3}$$

For $r = \ell = 0$ we also write $\mathsf{H}^{0,0}_\omega(\Omega) = \mathsf{L}^2_\omega(\Omega)$. The weighted Sobolev space $\mathsf{H}^{s,\ell}_\omega(\Omega)$ for a noninteger s is defined using the K-method of interpolation [20]:

$$\mathsf{H}^{s,\ell}_\omega(\Omega) \;=\; \left[\mathsf{H}^{r,\ell}_\omega(\Omega), \mathsf{H}^{r+1,\ell}_\omega(\Omega)\right]_{\theta,\infty}$$

for $s = r + \theta$, with r integer and $0 < \theta < 1$.

For $\ell \geq 0$ integer, a function u belongs to the countably normed space $\mathcal{B}^\ell_\omega(\Omega)$ if there exist constants $C > 0$ and $d \geq 1$ such that, for any integer $r \geq \ell$,

$$u \;\in\; \mathsf{H}^{r,\ell}_\omega(\Omega) \quad \text{with} \quad \left(\int_\Omega |\,\mathrm{D}^r\, u|^2 \Phi^2_{\omega+r-\ell}\, d\boldsymbol{x}\right)^{1/2} \leq C\, d^{r-\ell}\, (r-\ell)! \tag{4}$$

Furthermore, if $u \in \mathcal{B}^\ell_\omega(\Omega)$ with the corresponding constants C and d, then u belongs to the space $\mathsf{H}^{s,\ell}_\omega(\Omega)$ for any real $s \geq \ell$ and satisfies the inequality

$$\|u\|_{\mathsf{H}^{s,\ell}_\omega(\Omega)} \;\leq\; C\, d^{s-\ell}\, \sqrt{s}\;\, \Gamma(s-\ell+1), \tag{5}$$

where Γ is the gamma function.

For given data $f \in \mathsf{L}^2(\Omega)$, we shall consider the following model problem: find p such that

$$-\,\mathrm{div}\;\, A\, \mathbf{grad}\, p(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \quad \text{ in } \Omega \tag{6}$$

subject to $p = 0$ on $\partial\Omega$, where A is a 2×2 symmetric positive definite matrix. Standard arguments lead to the existence of a unique solution $p \in \mathsf{H}^1_0(\Omega)$. The following result shows that if the data f belong to the space $\mathcal{B}^0_\omega(\Omega)$, then the solution p has additional smoothness properties relative to the countably normed spaces.

THEOREM 1. *If the data $f \in \mathcal{B}^0_\omega(\Omega)$, then the solution p of* (6) *satisfies $p \in \mathcal{B}^2_\omega(\Omega)$, and the flux $\boldsymbol{u} = A\, \mathbf{grad}\, p$ satisfies $\boldsymbol{u} \in \boldsymbol{\mathcal{B}}^1_\omega(\Omega)$.*

Proof. The regularity of p is established in [12], and the regularity of $\boldsymbol{u}$ follows as an immediate consequence. ∎

3. Finite element discretization.

3.1. Partitions. The polygonal domain Ω is assumed to be partitioned into the union of finitely many nonoverlapping quadrilaterals $\mathcal{T} = \{K\}$. Each element K is assumed to be the image of the reference element $\widehat{K} = (-1,+1)^2$ under a smooth, invertible mapping $F_K : \widehat{K} \longrightarrow K$. More generally, we consider a family of such partitions, and we suppose that the elements are shape regular in the sense that the ratio of the diameter h_K of the smallest circle containing an element K to the diameter ρ_K of the largest circle contained within K is bounded uniformly over the whole family.

A standard finite element partition generally consists of elements such that the nonempty intersection of any pair of distinct elements is either a single common vertex or a single common edge. In order to facilitate local refinements, this assumption is relaxed to allow *hanging nodes* [1]. Thus, in addition to the above conditions, it is permitted for a nonempty intersection to be a complete side of at least one element and a single portion of an edge or the neighboring element obtained by bisecting the edge (see Figure 1). The node located at the midpoint is said to be *hanging*. An edge containing a hanging node is said to be *broken*. It is also possible to break edges unevenly, which is advantageous in accelerating resolution of corner singularities to be considered later.

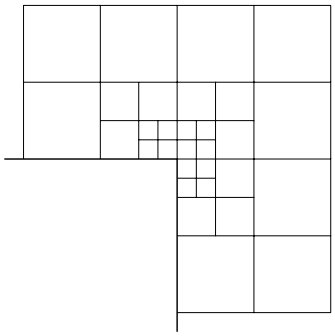


FIG. 1. *Geometric mesh with hanging nodes.*

3.2. Polynomial spaces on the reference element. For nonnegative integers k and ℓ, the space $\mathbb{Q}_{k,\ell}$ consists of polynomials of degree at most k and ℓ in the first and second variables, respectively. The notation $\mathbb{Q}_k = \mathbb{Q}_{k,k}$ is adopted in the case $k = \ell$. The space $\mathbb{P}_k$ consists of polynomials of total degree at most k.

The Brezzi–Douglas–Fortin–Marini space of order k ($\mathbf{BDFM}_k$) is defined [9, p. 123] by

$$\mathbf{BDFM}_k \;=\; \left\{\mathbb{P}_{k+1} \setminus (\eta^{k+1})\right\} \;\times\; \left\{\mathbb{P}_{k+1} \setminus (\xi^{k+1})\right\}$$

and is equipped with an interpolation operator $\widehat{\mathbf{\Pi}}_k$ satisfying $\widehat{\mathbf{\Pi}}_k \boldsymbol{u} \in \mathbf{BDFM}_k$ and

$$\text{(7)} \qquad \begin{cases} \int_{\widehat{K}} \left(\widehat{\mathbf{\Pi}}_k\, \boldsymbol{u} - \boldsymbol{u}\right) \cdot \boldsymbol{p}\, d\boldsymbol{x} \;=\; 0 \qquad \forall\, \boldsymbol{p} \;\in\; [\mathbb{P}_{k-1}]^2, \\ \int_{\gamma_m} \left(\widehat{\mathbf{\Pi}}_k\, \boldsymbol{u} - \boldsymbol{u}\right) \cdot \boldsymbol{n}\, q\, ds \;=\; 0 \quad \forall\, q \;\in\; \mathbb{R}_k(\gamma_m), \quad m = 1, 2, 3, 4, \end{cases}$$

where $\mathbb{R}_k(\gamma)$ denotes polynomials of degree at most k in the arclength s, and $\boldsymbol{n}$ is the unit outward normal on the edge. The Raviart–Thomas space of order k ($\mathbf{RT}_k$) is defined [9, p. 119] by

$$\mathbf{RT}_k \;=\; \mathbb{Q}_{k+1,k} \;\times\; \mathbb{Q}_{k,k+1}$$

with corresponding operator $\widehat{\mathbf{\Pi}}_k$ satisfying $\widehat{\mathbf{\Pi}}_k \boldsymbol{u} \in \mathbf{RT}_k$ and

$$(8)\qquad \begin{cases} \int_{\widehat{K}} \left(\widehat{\boldsymbol{\Pi}}_k\, \boldsymbol{u} - \boldsymbol{u}\right) \cdot \boldsymbol{p}\, d\boldsymbol{x} \;=\; 0 \qquad \forall\, \boldsymbol{p} \;\in\; \mathbb{Q}_{k-1,k} \times \mathbb{Q}_{k,k-1}, \\ \int_{\gamma_m} \left(\widehat{\boldsymbol{\Pi}}_k\, \boldsymbol{u} - \boldsymbol{u}\right) \cdot \boldsymbol{n}\, q\, ds \;=\; 0 \quad \forall\, q \;\in\; \mathbb{R}_k(\gamma_m), \quad m = 1,2,3,4. \end{cases}$$

The following result shows that these operators are well defined.

PROPOSITION 2. *The interpolation operators* $\widehat{\boldsymbol{\Pi}}_k$ *are unisolvent with respect to the* $\mathbf{BDFM}_k$ *and* $\mathbf{RT}_k$ *spaces. Furthermore, for sufficiently smooth functions* $\boldsymbol{u} \in \mathbf{H}(\mathrm{div}, \widehat{K})$*, the following "commuting diagram" property holds:*

$$(9)\qquad \mathrm{div}\left(\widehat{\boldsymbol{\Pi}}_k \boldsymbol{u}\right) = \widehat{\pi}_k(\mathrm{div}\, \boldsymbol{u}),$$

where $\widehat{\pi}_k$ *is the* $\mathsf{L}^2(\widehat{K})$*-orthogonal projection onto* $\mathbb{P}_k$ *(respectively,* $\mathbb{Q}_k$*) when* $\widehat{\boldsymbol{\Pi}}_k$ *is the operator associated with the space* $\mathbf{BDFM}_k$ *(respectively,* $\mathbf{RT}_k$*).*

Proof. Unisolvence follows using arguments given in [9, section III.3.2]. The commuting diagram property follows on observing that $\mathrm{div}\, \mathbf{BDFM}_k = \mathbb{P}_k$ (or $\mathrm{div}\, \mathbf{RT}_k = \mathbb{Q}_k$) and using [9, Proposition III.3.7]. ∎

3.3. Finite element spaces on a single physical element. Let $k \geq 1$ denote the polynomial degree of the space (either $\mathbf{BDFM}_k$ or $\mathbf{RT}_k$) used to construct the finite element space. The Piola transformation given by

$$(10)\qquad \mathcal{L}_K\, \widehat{\boldsymbol{u}} \;=\; \frac{1}{\det\left(J_K\right)}\, J_K\, \widehat{\boldsymbol{u}} \circ F_K^{-1}$$

defines a contravariant mapping and, in particular, preserves moments of the normal component of $\widehat{\boldsymbol{u}}$ on element edges [9, p. 97]. Here, J_K is the Jacobian matrix of the mapping F_K. The Piola transformation is used to define the local finite element spaces on the physical element K as follows:

$$\boldsymbol{\Gamma}_k^{\mathsf{BDFM}}(K) \;=\; \{\, \boldsymbol{u} \;:\quad \boldsymbol{u} \;=\; \mathcal{L}_K\, \widehat{\boldsymbol{u}}, \quad \widehat{\boldsymbol{u}} \;\in\; \mathbf{BDFM}_k \,\}$$

with $\boldsymbol{\Gamma}_k^{\mathsf{RT}}(K)$ defined in the same fashion. The space $\boldsymbol{\Gamma}_k^{\mathsf{BDFM}}(K)$ is equipped with an interpolation operator $\boldsymbol{\Pi}_k^{\mathsf{BDFM}}$ defined in terms of the local operator $\widehat{\boldsymbol{\Pi}}_k$ by the rule

$$(11)\qquad \boldsymbol{\Pi}_k \;=\; \mathcal{L}_K\, \widehat{\boldsymbol{\Pi}}_k\, \mathcal{L}_K^{-1}.$$

The global operator associated with the $\mathbf{RT}_k$ space is defined by the same expression.

3.4. Global finite element spaces. Let $\mathcal{T}$ be a partition of Ω. The global finite element spaces are defined by

$$(12)\qquad \boldsymbol{\Gamma}_k^{\mathsf{BDFM}}(\Omega) \;=\; \{\;\; \boldsymbol{u} \;\in \mathbf{H}(\mathrm{div}, \Omega) \;:\;\; \boldsymbol{u}|_K \;\in\; \boldsymbol{\Gamma}_k^{\mathsf{BDFM}}(K) \;\;\}$$

with a similar expression for $\boldsymbol{\Gamma}_k^{\mathsf{RT}}(\Omega)$. Although it is also possible to extend the analysis to the situation where combinations of the two local spaces are used in the construction of the global space, we shall not pursue this explicitly here.

If the partition $\mathcal{T}$ *contains no broken edges*, then the global interpolation operator $\boldsymbol{\Pi}$ associated with the global finite element space $\boldsymbol{\Gamma}_k(\Omega)$ may be defined in terms of the local interpolation operators in the usual way, as for example in [9, equation (3.71)]:

$$(13)\qquad (\boldsymbol{\Pi}\, \boldsymbol{u})|_K \;=\; \boldsymbol{\Pi}_{K,k}\, (\boldsymbol{u}|_K) \qquad \forall K \;\in\; \mathcal{T}.$$

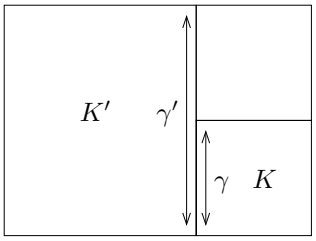


FIG. 2. *Broken edge γ' and subedge γ.*

This definition, albeit completely standard, *does not extend* to the situation where broken edges are present. For instance, consider the configuration shown in Figure 2.

The local interpolant $\boldsymbol{\Pi}_{K,k}\,\boldsymbol{u}$ on element K depends on the function $\boldsymbol{u}\cdot\boldsymbol{n}$ restricted to the edge γ. Equally well, $\boldsymbol{\Pi}_{K',k}\,\boldsymbol{u}$ depends on values of $\boldsymbol{u}\cdot\boldsymbol{n}$ on the entire edge γ'. It is not difficult to construct examples where the resulting normal components of $\boldsymbol{\Pi}_{K,k}\,\boldsymbol{u}$ and $\boldsymbol{\Pi}_{K',k}\,\boldsymbol{u}$ disagree on the common boundary γ:

$$(\boldsymbol{\Pi}_{K,k}\,\boldsymbol{u})\cdot\boldsymbol{n} \;\neq\; (\boldsymbol{\Pi}_{K',k}\,\boldsymbol{u})\cdot\boldsymbol{n} \qquad \text{on } \gamma. \tag{14}$$

This shows that it is not possible to define the global interpolant in terms of local interpolants using expression (13).

This difficulty is resolved by realizing that the degrees of freedom associated with the edges γ and γ' must be constrained to ensure $\mathbf{H}(\mathrm{div},\Omega)$ conformity holds. The degrees of freedom on the longer edge γ' take precedence over the values of the degrees of freedom on the shorter edge. Thus, for sufficiently smooth $\boldsymbol{u}\in\mathbf{H}(\mathrm{div},\Omega)$, the values of the global interpolant $\boldsymbol{\Pi u}$ on edge γ' are defined using global moment conditions on edge γ' as follows:

$$\int_{\gamma'}(\boldsymbol{\Pi u}-\boldsymbol{u})\cdot\boldsymbol{n}\;q\;ds \;=\; 0 \qquad \forall q\;\in\;\mathbb{R}_k(\gamma') \tag{15}$$

in place of (7). (Strictly speaking, q is a pull-back polynomial.) This definition coincides with (13) for a mesh containing no broken edges.

4. Approximation theory. The purpose of this section is to establish some basic approximation properties of the local interpolation operators $\widehat{\boldsymbol{\Pi}}_k$.

4.1. Explicit form for local interpolation operators. Let k specify the degree of polynomial on a reference $\widehat{K}$, as in Figure 3. First, we introduce the polynomial extension operator $\mathcal{E}_k^{\gamma_1}:\mathbb{R}_k(\gamma_1)\longrightarrow\mathbb{P}_{k+1}\setminus\{\eta^{k+1}\}$ defined by

$$(\mathcal{E}_k^{\gamma_1}\,w)\,(\xi,\eta) \;=\; \sum_{i=0}^{k} w_i\;\frac{(-1)^{k-i+1}}{2}\Big(L_{k-i+1}(\xi)-L_{k-i}(\xi)\Big)\;L_i(\eta),\quad (\xi,\eta)\in\widehat{K}, \tag{16}$$

where $w\in\mathbb{R}_k(\gamma_1)$ is of the form

$$w(\eta) \;=\; \sum_{i=0}^{k} w_i\;L_i(\eta),\quad \eta\in(-1,+1). \tag{17}$$

Here, $\{L_i\}$ are Legendre polynomials and (17) uniquely defines the coefficients w_i. Consequently, the extension is well defined by expression (16). Moreover, it possesses

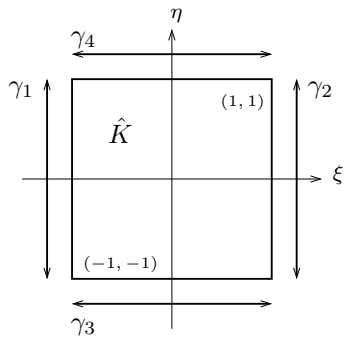


FIG. 3. *Notation for the reference element.*

the property that, for w in $\mathbb{R}_k(\gamma_1)$,

$$\mathcal{E}_k^{\gamma_1} w|_{\gamma_1} = w \qquad \text{and} \qquad \mathcal{E}_k^{\gamma_1} w|_{\gamma_2} = 0. \tag{18}$$

The extension operators $\mathcal{E}_k^{\gamma_2} : \mathbb{R}_k(\gamma_2) \longrightarrow \mathbb{P}_{k+1} \setminus \{\eta^{k+1}\}$ and $\mathcal{E}_k^{\gamma_m} : \mathbb{R}_k(\gamma_m) \longrightarrow \mathbb{P}_{k+1} \setminus \{\xi^{k+1}\}$, $m = 3, 4$, are defined in a similar way with properties analogous to (18).

The $\mathsf{L}^2(\widehat{K})$-orthogonal projection onto $\mathbb{P}_k$ is denoted by $\widehat{\pi}_k^{\mathsf{P}}$, the $\mathsf{L}^2(\gamma_m)$-orthogonal projector onto $\mathbb{R}_k(\gamma_m)$ is denoted by $\mathsf{R}_k^{\gamma_m}$, and Tr_{γ_m} denotes the trace operator. The next result provides an explicit expression for the interpolant $\widehat{\mathbf{\Pi}}_k$ associated with the space $\mathbf{BDFM}_k$.

LEMMA 3. *Let $k \geq 1$. Then the interpolation operator $\widehat{\mathbf{\Pi}}_k$ for the space $\mathbf{BDFM}_k$ is given by*

$$\begin{aligned} \widehat{\mathbf{\Pi}}_k \boldsymbol{u} \;=\; & \Big(\widehat{\pi}_{k-1}^{\mathsf{P}} u_x + \textstyle\sum_{m=1}^{2} \mathcal{E}_k^{\gamma_m} \left(\mathsf{R}_k^{\gamma_m} \mathrm{Tr}_{\gamma_m} u_x - \mathrm{Tr}_{\gamma_m} \widehat{\pi}_{k-1}^{\mathsf{P}} u_x \right), \\ & \widehat{\pi}_{k-1}^{\mathsf{P}} u_y + \textstyle\sum_{m=3}^{4} \mathcal{E}_k^{\gamma_m} \left(\mathsf{R}_k^{\gamma_m} \mathrm{Tr}_{\gamma_m} u_y - \mathrm{Tr}_{\gamma_m} \widehat{\pi}_{k-1}^{\mathsf{P}} u_y \right) \Big), \end{aligned} \tag{19}$$

where $\boldsymbol{u} = (u_x, u_y)$.

Proof. It is clear that the expression (19) defines a function belonging to $\mathbf{BDFM}_k$. It therefore suffices to verify that conditions (7) hold.

Let $\tilde{u}_x$ denote the first component of (19). Then, for $p \in \mathbb{P}_{k-1}$ we have, thanks to the orthogonality properties of Legendre polynomials,

$$\int_{\widehat{K}} \tilde{u}_x \, p \, d\xi \, d\eta = \int_{\widehat{K}} \widehat{\pi}_{k-1}^{\mathsf{P}} u_x \, p \, d\xi \, d\eta = \int_{\widehat{K}} u_x \, p \, d\xi \, d\eta.$$

Similarly, property (18) implies that, for $q \in \mathbb{R}_k(\gamma_m)$,

$$\begin{aligned} & \textstyle\int_{\gamma_m} (\mathrm{Tr}_{\gamma_m} \tilde{u}_x) \cdot q \, d\eta \\ & = \textstyle\int_{\gamma_m} \left(\mathrm{Tr}_{\gamma_m} \widehat{\pi}_{k-1}^{\mathsf{P}} u_x \right) \cdot q \, d\eta + \int_{\gamma_m} \left(\mathsf{R}_k^{\gamma_m} \mathrm{Tr}_{\gamma_m} u_x - \mathrm{Tr}_{\gamma_m} \widehat{\pi}_{k-1}^{\mathsf{P}} u_x \right) \cdot q \, d\eta \\ & = \textstyle\int_{\gamma_m} (\mathsf{R}_k^{\gamma_m} \mathrm{Tr}_{\gamma_m} u_x) \cdot q \, d\eta = \int_{\gamma_m} (\mathrm{Tr}_{\gamma_m} u_x) \cdot q \, d\eta, \end{aligned}$$

and the proof is complete. ∎

Extension operators for the $\mathbf{RT}_k$ space are defined by $\mathcal{E}_k^{\gamma_1} : \mathbb{R}_k(\gamma_1) \longrightarrow \mathbb{Q}_{k+1,k}$, where

$$\left(\mathcal{E}_k^{\gamma_1} w\right)(\xi,\eta) \;=\; \sum_{i=0}^{k} \frac{(-1)^{k+1}}{2}\; c_i\; \left(L_{k+1}(\xi) - L_k(\xi)\right) L_i(\eta)$$

with similar definitions for $\mathcal{E}_k^{\gamma_m}$, $m = 2, 3, 4$. There is no danger of confusion in using the same notation for the extension operators in both the $\mathbf{RT}$ and $\mathbf{BDFM}$ cases. The $\mathsf{L}^2(\widehat{K})$-orthogonal projection onto the space $\mathbb{Q}_{k,\ell}$ is denoted by $\widehat{\pi}_{k,\ell}^{\mathsf{Q}}$. Then, by analogy with Lemma 3, we have the following.

LEMMA 4. *Let $k \geq 1$. Then the interpolation operator $\widehat{\mathbf{\Pi}}_k$ for the space $\mathbf{RT}_k$ is given by*

$$\widehat{\mathbf{\Pi}}_k \boldsymbol{u} \;=\; \Bigg(\widehat{\pi}_{k-1,k}^{\mathsf{Q}}\, u_x + \textstyle\sum_{m=1}^{2} \mathcal{E}_k^{\gamma_m} \left(\mathsf{R}_k^{\gamma_m}\, \mathrm{Tr}_{\gamma_m}\, u_x - \mathrm{Tr}_{\gamma_m}\, \widehat{\pi}_{k-1,k}^{\mathsf{Q}}\, u_x \right), \qquad (20)$$

$$\widehat{\pi}_{k,k-1}^{\mathsf{Q}}\, u_y + \textstyle\sum_{m=3}^{4} \mathcal{E}_k^{\gamma_m} \left(\mathsf{R}_k^{\gamma_m}\, \mathrm{Tr}_{\gamma_m}\, u_y - \mathrm{Tr}_{\gamma_m}\, \widehat{\pi}_{k,k-1}^{\mathsf{Q}}\, u_y \right) \Bigg).$$

4.2. Approximation on the reference element. It is convenient to introduce the shorthand notation

$$[w]_{r,\widehat{\Omega}}^2 \;=\; \left(\frac{h_1}{2}\right)^{2r} \| \,\mathrm{D}^{(r,0)}\, w \,\|_{\mathsf{L}^2(\widehat{\Omega})}^2 \;+\; \left(\frac{h_2}{2}\right)^{2r} \| \,\mathrm{D}^{(0,r)}\, w \,\|_{\mathsf{L}^2(\widehat{\Omega})}^2$$

in the case where $\widehat{\Omega} = (0, h_1) \times (0, h_2)$ and $w \in \mathsf{H}^r(\widehat{\Omega})$, $r \in \mathbb{N}$. Furthermore, in what follows, C denotes positive constants that are independent of other quantities appearing in the same relation and whose values need not be the same in any two places.

The following result will prove useful.

LEMMA 5. *Suppose that $w \in \mathsf{H}^r(\widehat{K})$, $r \geq 0$, has Legendre series expansion*

$$w(\xi,\eta) \;=\; \sum_{i=0}^{\infty} \sum_{j=0}^{\infty}\; a_{i,j}\; L_i(\xi)\, L_j(\eta). \qquad (21)$$

Then, for all integers $0 \leq r_1 \leq p$, $0 \leq r_2 \leq q$, with $r_1 + r_2 \leq r$, there holds

$$\sum_{i=p}^{\infty} \sum_{j=q}^{\infty}\; \rho_i\, \rho_j\, a_{i,j}^2 \;\leq\; \frac{(p-r_1)!}{(p+r_1)!}\, \frac{(q-r_2)!}{(q+r_2)!}\, \| \,\mathrm{D}^{(r_1,r_2)}\, w \|_{\mathsf{L}^2(\widehat{K})}^2, \qquad (22)$$

where $\rho_i = (i + 1/2)^{-1}$.

Proof. If r_1 and r_2 satisfy the hypothesis, then it is well known (see, e.g., Lemma 4.1 [12]), that

$$\sum_{i=r_1}^{\infty} \sum_{j=r_2}^{\infty}\; \rho_i\, \rho_j\, a_{i,j}^2\, \frac{(i+r_1)!}{(i-r_1)!}\, \frac{(j+r_2)!}{(j-r_2)!} \;=\; \int_{\widehat{K}} \left(1-\xi^2\right)^{r_1} \left(1-\eta^2\right)^{r_2} \left| \,\mathrm{D}^{(r_1,r_2)}\, w \,\right|^2 d\xi\, d\eta.$$

Now, for p and q as in the statement, we have

$$\sum_{i=p}^{\infty} \sum_{j=q}^{\infty}\; \rho_i\, \rho_j\, a_{i,j}^2 \;\leq\; \left(\sum_{i=p}^{\infty} \sum_{j=q}^{\infty}\; \rho_i\, \rho_j\, a_{i,j}^2\, \frac{(i+r_1)!}{(i-r_1)!}\, \frac{(j+r_2)!}{(j-r_2)!} \right)\; \frac{(p-r_1)!}{(p+r_1)!}\, \frac{(q-r_2)!}{(q+r_2)!},$$

and the result follows at once. ∎

We begin with approximation properties for the $\mathsf{L}^2(\widehat{K})$-projections $\widehat{\pi}_k^{\mathsf{Q}}$ and $\widehat{\pi}_k^{\mathsf{P}}$.

LEMMA 6. *Let* $\widehat{\pi}_k^{\mathsf{Q}} : \mathsf{L}^2(\widehat{K}) \longrightarrow \mathbb{Q}_k$ *and* $\widehat{\pi}_k^{\mathsf{P}} : \mathsf{L}^2(\widehat{K}) \longrightarrow \mathbb{P}_k$ *denote orthogonal projections. Suppose* $w \in \mathsf{H}^r(\widehat{K})$*; then for* $0 \le r \le k+1$,

$$\|w - \widehat{\pi}_k^{\mathsf{Q}} w\|^2_{\mathsf{L}^2(\widehat{K})} \;\le\; \frac{(k+1-r)!}{(k+1+r)!}\; [w]^2_{r,\widehat{K}}, \tag{23}$$

and for $0 \le r \le [k/2]+1$,

$$\|w - \widehat{\pi}_k^{\mathsf{P}} w\|^2_{\mathsf{L}^2(\widehat{K})} \;\le\; \frac{([k/2]+1-r)!}{([k/2]+1+r)!}\; [w]^2_{r,\widehat{K}}. \tag{24}$$

Proof. By density, it suffices to consider w of the form (21) so that

$$\widehat{\pi}_k^{\mathsf{Q}} w(\xi,\eta) \;=\; \sum_{i=0}^{k}\sum_{j=0}^{k} a_{i,j}\; L_i(\xi)\; L_j(\eta).$$

Orthogonality of Legendre polynomials implies that

$$\|w - \widehat{\pi}_k^{\mathsf{Q}} w\|^2_{\mathsf{L}^2(\widehat{K})} \;\le\; \left(\sum_{i=k+1}^{\infty}\sum_{j=0}^{\infty} + \sum_{i=0}^{k}\sum_{j=k+1}^{\infty}\right) a_{i,j}^2\; \rho_i\; \rho_j. \tag{25}$$

The first term on the right-hand side of (25) is bounded using Lemma 5:

$$\sum_{i=k+1}^{\infty}\sum_{j=0}^{\infty} a_{i,j}^2\; \rho_i\; \rho_j \;\le\; \frac{(k+1-r)!}{(k+1+r)!}\, \|\,\mathrm{D}^{(r,0)}\, w\|^2_{\mathsf{L}^2(\widehat{K})}.$$

A similar bound applies for the second term and the first result follows at once.

The second bound follows trivially from the first thanks to the inclusion $\mathbb{Q}_{[k/2]} \subset \mathbb{P}_k$ and optimality of the $\mathsf{L}^2(\widehat{K})$-projection $\widehat{\pi}_k^{\mathsf{P}}$. ∎

The following technical lemma will be useful in deriving bounds for the accuracy of the interpolation operators.

LEMMA 7. *Let* $w \in \mathsf{H}^r(\widehat{K})$, $r > 1$, *be of the form* (21). *Then for fixed* $\varepsilon \in (0, r-1)$ *we have, for* $r \le k$,

$$\sum_{j=0}^{\infty} \rho_j \left(\sum_{i=k}^{\infty} a_{i,j}\; (-1)^i\right)^2 \;\le\; \frac{C\,k^2}{\varepsilon}\; \frac{(k-r)!}{(k+r)!}\; [w]^2_{r,\widehat{K}} \tag{26}$$

while, for $r \le \left[\frac{k+1}{2}\right]$,

$$\sum_{j=0}^{\infty} \rho_j \left(\sum_{i=\max(0,k-j)}^{\infty} a_{i,j}\; (-1)^i\right)^2 \;\le\; \frac{C\,k^2}{\varepsilon}\; \frac{\left(\left[\frac{k+1}{2}\right]-r\right)!}{\left(\left[\frac{k+1}{2}\right]+r\right)!}\; [w]^2_{r,\widehat{K}}. \tag{27}$$

Moreover, for $r \le k$,

$$\sum_{j=k}^{\infty} \rho_j \left(\sum_{i=0}^{\infty} a_{i,j}\; (-1)^i\right)^2 \;\le\; \frac{C\,k^2}{\varepsilon}\; \frac{(k-r)!}{(k+r)!}\; [w]^2_{r,\widehat{K}}. \tag{28}$$

In each case, $C > 0$ is independent of r and k.

Proof. Let $\varepsilon \in (0, r-1)$. Then, by the Cauchy–Schwarz inequality,

$$\sum_{j=0}^{\infty} \rho_j \left(\sum_{i=k}^{\infty} a_{i,j}\, (-1)^i \right)^2 \;\le\; \sum_{j=0}^{\infty} \sum_{i=k}^{\infty} \rho_i\, \rho_j\, a_{i,j}^2\, i^{2+2\varepsilon} \;\cdot\; \sum_{i=k}^{\infty} \rho_i^{-1}\, i^{-2-2\varepsilon}.$$

Bounding the second factor on the right-hand side by $\sum_{i=k}^{\infty} i^{-1-2\varepsilon} \;\le\; C/\varepsilon\, k^{-2\varepsilon}$, we obtain

$$\sum_{j=0}^{\infty} \rho_j \left(\sum_{i=k}^{\infty} a_{i,j}\, (-1)^i \right)^2 \;\le\; \frac{C\, k^{-2\varepsilon}}{\varepsilon} \sum_{i=k}^{\infty} \sum_{j=0}^{k} \rho_i\, \rho_j\, a_{i,j}^2\, \frac{(i+r)!}{(i-r)!}\, f(i),$$

where $f(x) = x^{2+2\varepsilon}\, (x-r)!/(x+r)!$, $x \ge k$. If $\varepsilon < r-1$, then, for k sufficiently large, f is decreasing. Hence,

$$\begin{aligned} \sum_{j=0}^{\infty}\, \rho_j \sum_{i=k}^{\infty} \Big(a_{i,j}\, (-1)^i \Big)^2 \;&\le\; \frac{C\, k^{-2\varepsilon}}{\varepsilon}\, f(k) \sum_{i=k}^{\infty} \sum_{j=0}^{\infty}\, \rho_i\, \rho_j\, a_{i,j}^2\, \frac{(i+r)!}{(i-r)!} \\ &\le\; \frac{C\, k^2}{\varepsilon}\, \frac{(k-r)!}{(k+r)!}\, \| \mathrm{D}^{(r,0)}\, w \|_{\mathsf{L}^2(\widehat{K})}^2, \end{aligned}$$

arguing as in Lemma 5. This completes the proof of the first result.

Turning to the second result, we begin by observing that

$$\begin{aligned} &\frac{1}{2} \sum_{j=0}^{\infty} \rho_j \left(\sum_{i=\max(0,k-j)}^{\infty} a_{i,j}\, (-1)^i \right)^2 \\ &\le \sum_{j=0}^{\infty}\, \rho_j \left(\sum_{i=k}^{\infty} a_{i,j}\, (-1)^i \right)^2 \;+\; \sum_{j=0}^{\infty}\, \rho_j \left(\sum_{i=\max(0,k-j)}^{k-1} a_{i,j}\, (-1)^i \right)^2. \end{aligned}$$

The first term is bounded using (26), and it suffices to consider the second term. Employing the Cauchy–Schwarz inequality once again gives

$$\sum_{j=0}^{\infty}\, \rho_j \left(\sum_{i=\max(0,k-j)}^{k-1} a_{i,j}\, (-1)^i \right)^2 \;\le\; \sum_{j=0}^{\infty} \sum_{i=\max(0,k-j)}^{k} \rho_i\, \rho_j\, a_{i,j}^2 \sum_{i=0}^{k} \rho_i^{-1}.$$

The second factor is bounded by Ck^2, where C is independent of k, whilst the first factor is bounded by observing that

$$\sum_{j=0}^{\infty} \sum_{i=\max(0,k-j)}^{k} \rho_i\, \rho_j\, a_{i,j}^2 \;\le\; \| w - \widehat{\pi}_{k-1}^{\mathsf{P}} w \|_{\mathsf{L}^2(\widehat{K})}^2 \;\le\; \frac{\big(\big[\frac{k+1}{2}\big] - r\big)!}{\big(\big[\frac{k+1}{2}\big] + r\big)!}\, [w]_{r,\widehat{K}}^2,$$

where Lemma 6 has been used. Consequently,

$$\sum_{j=0}^{\infty}\, \rho_j \left(\sum_{i=\max(0,k-j)}^{k-1} a_{i,j}\, (-1)^i \right)^2 \;\le\; C\, k^2\, \frac{\big(\big[\frac{k+1}{2}\big] - r\big)!}{\big(\big[\frac{k+1}{2}\big] + r\big)!}\, [w]_{r,\widehat{K}}^2,$$

and gathering results gives the second estimate.

In order to derive the final estimate, observe that

$$\frac{1}{2}\sum_{j=k}^{\infty}\rho_j\left(\sum_{i=0}^{\infty}a_{i,j}\,(-1)^i\right)^2 \le \sum_{j=k}^{\infty}\rho_j\left\{\left(\sum_{i=k}^{\infty}a_{i,j}\,(-1)^i\right)^2+\left(\sum_{i=0}^{k-1}a_{i,j}\,(-1)^i\right)^2\right\}.$$

The first term is bounded by

$$\sum_{j=0}^{\infty}\rho_j\left(\sum_{i=k}^{\infty}a_{i,j}\,(-1)^i\right)^2$$

and then applying the first part of Lemma 7. The second term is bounded using the Cauchy–Schwarz inequality to obtain

$$\sum_{j=k}^{\infty}\left(\sum_{i=0}^{k-1}a_{i,j}\,(-1)^i\right)^2 \le \sum_{i=0}^{k-1}\sum_{j=k}^{\infty}\rho_i\,\rho_j\,a_{i,j}^2\cdot\sum_{i=0}^{k}\rho_i^{-1}$$

$$\le C\,k^2\sum_{i=0}^{k-1}\sum_{j=k}^{\infty}\rho_i\,\rho_j\,a_{i,j}^2$$

and then using Lemma 5 to obtain, for $r\le k$,

$$\sum_{i=0}^{k-1}\sum_{j=k}^{\infty}\rho_i\,\rho_j\,a_{i,j}^2 \le \frac{(k-r)!}{(k+r)!}\,\|\,\mathrm{D}^{(0,r)}\,w\|^2_{\mathsf{L}^2(\widehat{K})}.$$

The estimate follows from these results. □

We now present bounds for the interpolation operators.

LEMMA 8. *Let $k\ge 1$. If $\boldsymbol{u}\in\mathbf{H}^r(\widehat{K})$, then, for $1<r\le k$,*

$$\|\boldsymbol{u}-\widehat{\boldsymbol{\Pi}}_k^{\mathsf{RT}}\boldsymbol{u}\|^2 \le \frac{C}{\varepsilon}\,k\,\frac{(k-r)!}{(k+r)!}\,[\boldsymbol{u}]^2_{r,\widehat{K}}, \tag{29}$$

and, for $1<r\le[k/2]$,

$$\|\boldsymbol{u}-\widehat{\boldsymbol{\Pi}}_k^{\mathsf{BDFM}}\boldsymbol{u}\|^2 \le \frac{C}{\varepsilon}\,k^2\,\frac{([k/2]-r)!}{([k/2]+r)!}\,[\boldsymbol{u}]^2_{r,\widehat{K}}, \tag{30}$$

where $\varepsilon\in(0,r-1)$.

Proof. Let $\boldsymbol{u}=(u_x,u_y)\in\mathbf{H}^r(\widehat{K})$. By a density argument, it suffices to consider u_x of the form (21). By Lemma 4,

$$\left(\boldsymbol{u}-\widehat{\boldsymbol{\Pi}}_k^{\mathsf{RT}}\boldsymbol{u}\right)_x = u_x-\widehat{\pi}^{\mathsf{Q}}_{k-1,k}u_x+\sum_{m=1}^{2}\mathcal{E}_k^{\gamma_m}\,w_m,$$

where

$$w_1(\eta) = \sum_{j=0}^{k}\left(\sum_{i=k}^{\infty}a_{i,j}\,(-1)^i\right)L_j(\eta)$$

with w_2 defined analogously. Applying the triangle inequality gives

$$\left\| \left(\boldsymbol{u} - \widehat{\boldsymbol{\Pi}}_k^{\mathsf{RT}} \boldsymbol{u} \right)_x \right\|^2_{\mathsf{L}^2(\widehat{K})} \;\leq\; \| u_x - \widehat{\pi}^{\mathsf{Q}}_{k-1,k} u_x \|^2_{\mathsf{L}^2(\widehat{K})} \;+\; \sum_{m=1}^{2} \| \mathcal{E}_k^{\gamma_m}\, w_m \|^2_{\mathsf{L}^2(\widehat{K})},$$

and the first term is bounded using Lemma 6 after observing that

$$\| u_x - \widehat{\pi}^{\mathsf{Q}}_{k-1,k} u_x \|_{\mathsf{L}^2(\widehat{K})} \;\leq\; \| u_x - \widehat{\pi}^{\mathsf{Q}}_{k-1} u_x \|_{\mathsf{L}^2(\widehat{K})}.$$

On recalling the definition of the extension operator $\mathcal{E}_k^{\gamma_1}$ associated with **RT** elements, we obtain

$$\begin{aligned} \| \mathcal{E}_k^{\gamma_1}\, w_1 \|^2_{\mathsf{L}^2(\widehat{K})} \;&=\; \frac{1}{4}\, (\rho_k + \rho_{k+1}) \sum_{j=0}^{k} \rho_j \left(\sum_{i=k}^{\infty} a_{i,j}\, (-1)^i \right)^2 \\ &\leq\; \frac{C\, k}{\varepsilon}\, \frac{(k-r)!}{(k+r)!}\, [u_x]^2_{r,\widehat{K}}, \end{aligned}$$

where the first part of Lemma 7 has been used along with $\rho_{k+1} < \rho_k \leq Ck^{-1}$. An analogous estimate holds for the remaining term. In conclusion, we obtain

$$\left\| \left(\boldsymbol{u} - \widehat{\boldsymbol{\Pi}}_k^{\mathsf{RT}} \boldsymbol{u} \right)_x \right\|^2_{\mathsf{L}^2(\widehat{K})} \;\leq\; C \left\{ \frac{(k-r)!}{(k+r)!} \;+\; \frac{k}{\varepsilon}\, \frac{(k-r)!}{(k+r)!} \right\} [u_x]^2_{r,\widehat{K}},$$

and the result follows at once using since the same estimate holds for the y-component.

The proof of the second estimate follows the same lines using instead the second parts of Lemmas 6 and 7. The only difference, compared with the **RT** case, is that an additional factor of k is present owing to bounding

$$\| \mathcal{E}_k^{\gamma_1} w_1 \|^2_{\mathsf{L}^2(\widehat{K})} \;=\; \frac{1}{4} \sum_{j=0}^{k} \rho_j\, (\rho_{k-j+1} + \rho_{k-j}) \left(\sum_{i=k-j}^{\infty} a_{i,j}\, (-1)^i \right)^2$$

by

$$\frac{1}{2}\, \rho_0 \sum_{j=0}^{k} \rho_j \left(\sum_{i=k-j}^{\infty} a_{i,j}\, (-1)^i \right)^2$$

and then using Lemma 7. ∎

4.3. Local approximation on a rectangular element. Consider the case when $K = (a,b) \times (c,d)$ is a rectangular element of size $h_K = \max(b-a, d-c)$. We shall be particularly concerned with the situation when the element is located near to a vertex of the domain Ω. For convenience, we assume the vertex is located at the origin $\mathbf{0}$ and assume that the function $\boldsymbol{u}$ belongs to a weighted space $\mathbf{H}^{s,\ell}_{\omega}(K)$ with weight function $\Phi_\omega(\boldsymbol{x}) = |\boldsymbol{x}|^\omega$.

First consider the case where the element is located away from the vertex, i.e., $\operatorname{dist}(\mathbf{0}, K) = \varrho_K > 0$. The following result then holds.

LEMMA 9. *Let K be a rectangular element as above, with degree $k \geq 2$. If $\boldsymbol{u} \in \mathbf{H}^{s,\ell}_{\omega}(K)$ with $\ell \in \{1,2\}$, then, for $2 \leq s \leq k$,*

$$\| \boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{RT}}\, \boldsymbol{u} \|^2_{\mathbf{L}^2(K)} \;\leq\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\, k\, \frac{\Gamma(k-s+1)}{\Gamma(k+s+1)}\, |\boldsymbol{u}|^2_{\mathbf{H}^{s,\ell}_{\omega}(K)}, \tag{31}$$

and, for $2 \le s \le \max(2, [k/2]) = \tilde{k}$,

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{BDFM}} \boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;\le\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\, \tilde{k}^2\, \frac{\Gamma(\tilde{k}-s+1)}{\Gamma(\tilde{k}+s+1)}\, |\boldsymbol{u}|_{\mathbf{H}_\omega^{s,\ell}(K)}^2. \tag{32}$$

Proof. Let $r \in [2, k]$ be an integer. Observe that

$$\|\,\mathrm{D}^{(r,0)}\,\boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;+\; \|\,\mathrm{D}^{(0,r)}\,\boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;\le\; C\, \varrho_K^{-2(\omega+r-\ell)}\, |\boldsymbol{u}|_{\mathbf{H}_\omega^{r,\ell}(K)}^2,$$

and consequently

$$[\boldsymbol{u}]_{r,K}^2 \;\le\; \frac{C\, h_K^{2r}}{\varrho_K^{2(\omega+r-\ell)}}\, |\boldsymbol{u}|_{\mathbf{H}_\omega^{r,\ell}(K)}^2. \tag{33}$$

Applying Lemma 8, we conclude that

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{RT}} \boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;\le\; \frac{C\, h_K^{2r}}{\varrho_K^{2(\omega+r-\ell)}}\, k\, \frac{(k-r)!}{(k+r)!}\, |\boldsymbol{u}|_{\mathbf{H}_\omega^{r,\ell}(K)}^2$$

This result is valid for integer values of r. Applying a standard interpolation argument (using the K-method) enables the result to be extended to noninteger $s \in [2, k]$. For full details, we refer to the proof of Lemma 4.48 [20].

The proof of the second result is divided into the case when $[k/2] \ge 2$, i.e., $k \ge 4$, and the case where $k = 2, 3$. The argument in the former case mirrors the one presented above. It therefore suffices to assume $k = 2$ or 3 and to show that, for $s = 2$,

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{BDFM}} \boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;\le\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\, |\boldsymbol{u}|_{\mathbf{H}_\omega^{s,\ell}(K)}^2.$$

It is shown in [21] that there exists a constant $\tilde{C}$ independent of k and h_K such that

$$\|\boldsymbol{\Pi}_k^{\mathsf{BDFM}} \boldsymbol{u}\|_{\mathbf{L}^2(K)} \;\le\; \tilde{C}\, \|\boldsymbol{u}\|_{\mathbf{H}^1(K)}.$$

Furthermore, thanks to the polynomial reproducing properties of $\widehat{\boldsymbol{\Pi}}_k$, for any first order pull-back function $\boldsymbol{u}_1$, we have

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{BDFM}} \boldsymbol{u}\|_{\mathbf{L}^2(K)} \;\le\; (1 + \tilde{C})\, \|\boldsymbol{u} - \boldsymbol{u}_1\|_{\mathbf{H}^1(K)}.$$

A routine scaling argument and use of the Bramble–Hilbert lemma reveals that, for $s = 2$,

$$\inf_{\boldsymbol{u}_1} \|\boldsymbol{u} - \boldsymbol{u}_1\|_{\mathbf{H}^1(K)} \;\le\; C\, [\boldsymbol{u}]_{s,K},$$

and applying (33) leads to

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k^{\mathsf{BDFM}} \boldsymbol{u}\|_{\mathbf{L}^2(K)}^2 \;\le\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\, |\boldsymbol{u}|_{\mathbf{H}^{s,\ell}(K)}^2$$

as claimed. ∎

Practically the same argument gives bounds for the accuracy of the L^2-projection operators.

LEMMA 10. *Let K be a rectangular element as above, with degree $k \geq 2$. If $p \in \mathsf{H}^{s,\ell}_\omega(K)$ with $\ell \in \{1,2\}$, then, for $2 \leq s \leq k+1$,*

$$\|p - \pi^{\mathsf{Q}}_k\, p\|^2_{\mathsf{L}^2(K)} \;\leq\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}} \frac{\Gamma(k-s+2)}{\Gamma(k+s+2)}\, |p|^2_{\mathsf{H}^{s,\ell}_\omega(K)}, \tag{34}$$

and, for $2 \leq s \leq \tilde{k}+1$,

$$\|p - \pi^{\mathsf{P}}_k\, p\|^2_{\mathsf{L}^2(K)} \;\leq\; \frac{C\, h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}} \frac{\Gamma(\tilde{k}-s+2)}{\Gamma(\tilde{k}+s+2)}\, |p|^2_{\mathsf{H}^{s,\ell}_\omega(K)}. \tag{35}$$

The next result deals with the situation where the element K has a vertex at corner of the domain, so that $\varrho_K = \operatorname{dist}(\mathbf{0}, K) = 0$.

LEMMA 11. *Let K be a rectangular element as above, with degree $k \geq 2$. If $\boldsymbol{u} \in \mathsf{H}^{2,\ell}_\omega(K)$ with $\ell \in \{1,2\}$, then*

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k \boldsymbol{u}\|_{\mathsf{L}^2(K)} \;\leq\; C\, h_K^{\ell-\omega}\, |\boldsymbol{u}|_{\mathsf{H}^{2,\ell}_\omega(K)}, \tag{36}$$

and if $p \in \mathsf{H}^{2,\ell}_\omega(K)$, then

$$\|p - \pi_k p\|_{\mathsf{L}^2(K)} \;\leq\; C\, h_K^{\ell-\omega}\, |p|_{\mathsf{H}^{2,\ell}_\omega(K)}, \tag{37}$$

where $\boldsymbol{\Pi}_k$ is the interpolant associated with the space BDFM_k (respectively, RT_k), and π_k is the L^2-projection associated with the space $\mathbb{P}_k$ (respectively, $\mathbb{Q}_k$).

Proof. We deal with the cases $\ell = 1$ and $\ell = 2$ separately. First, consider $\ell = 1$ and assume K is of unit size ($h_K = 1$) to begin with. By Theorem 2.1 [4], if $\boldsymbol{u} \in \mathsf{H}^{2,1}_\omega(K)$, then the trace of $\boldsymbol{u}$ on an edge $\gamma \subset \partial K$ satisfies the following properties:

(a) If $\omega \in (0, 1/2)$, then $\boldsymbol{u}|_\gamma \in \mathsf{H}^{2,1}_{\tilde{\omega}}(\gamma)$ for some $\tilde{\omega} \in (1/2, \omega + 1/2)$, and $\|\boldsymbol{u}\|_{\mathsf{H}^{2,1}_{\tilde{\omega}}(\gamma)} \leq C\, \|\boldsymbol{u}\|_{\mathsf{H}^{2,1}_\omega(K)}$.

(b) If $\omega \in (1/2, 1)$, then $\boldsymbol{u}|_\gamma \in \mathsf{H}^{2,0}_{\hat{\omega}}(\gamma)$ for some $\hat{\omega} \in (\omega - 1/2, 1/2)$, and $\|\boldsymbol{u}\|_{\mathsf{H}^{2,0}_{\hat{\omega}}(\gamma)} \leq C\, \|\boldsymbol{u}\|_{\mathsf{H}^{2,1}_\omega(K)}$.

In order to show that $\boldsymbol{\Pi}_k\, \boldsymbol{u}$ is well defined, it suffices to show that the degrees of freedom corresponding to moments on the edges

$$\boldsymbol{u} \;\longmapsto\; \int_\gamma (\boldsymbol{u} \cdot \boldsymbol{n})\; p\, ds \qquad \text{for } p \;\in\; \mathbb{R}_k(\gamma)$$

are well defined. This follows trivially in case (a) since $\boldsymbol{u}|_{\mathsf{L}^2(\gamma)}$ and we may bound the moment by $C_p\, \|\boldsymbol{u}\|_{\mathsf{H}^{2,1}_{\tilde{\omega}}(\gamma)}$. In case (b), we use the Cauchy–Schwarz inequality to deduce that

$$\left| \int_\gamma (\boldsymbol{u} \cdot \boldsymbol{n})\; p\, ds \right| \;\leq\; \| \, |s|^{\hat{\omega}}\, |\boldsymbol{u}| \, \|_{\mathsf{L}^2(\gamma)} \;\; \| \, |s|^{-\hat{\omega}}\, p \, \|_{\mathsf{L}^2(\gamma)},$$

and then, since $\hat{\omega} < 1/2$, we may bound this by $C_{p,\omega}\, \|\boldsymbol{u}\|_{\mathsf{H}^{2,0}_{\hat{\omega}}(\gamma)}$. This shows $\boldsymbol{\Pi}_k\, \boldsymbol{u}$ exists. Furthermore

$$\|\boldsymbol{\Pi}_k\, \boldsymbol{u}\|_{\mathsf{L}^2(K)} \;\leq\; C\, \left(\|\boldsymbol{u}\|_{\mathsf{L}^2(K)} + \|\boldsymbol{u}\|_{\mathsf{H}^{2,1}_\omega(K)} \right).$$

Therefore, for any constant $\boldsymbol{u}_0 \in \mathbb{R}^2$,

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k \boldsymbol{u}\|_{\mathbf{L}^2(K)} \;\leq\; (1 + C)\, \|\boldsymbol{u} - \boldsymbol{u}_0\|_{\mathbf{H}^{2,1}_{\omega}(K)}.$$

According to Lemma 4.3 [5], if v satisfies $\int_K |\boldsymbol{x}|^{2\omega} |\,\mathrm{D}^1 v|^2 d\boldsymbol{x} < \infty$, then there exists a constant $\bar{v}$ such that

$$\int_K |\boldsymbol{x}|^{2\omega-2}\, |v - \bar{v}|^2\, d\boldsymbol{x} \;\leq\; C \int_K |\boldsymbol{x}|^{2\omega}\, |\,\mathrm{D}^1 v|^2\, d\boldsymbol{x}.$$

Strictly speaking, this result was demonstrated for a triangular domain, but the argument generalizes to a rectangle K. Now, $|\boldsymbol{x}|^{-1} \geq 1/\sqrt{2}$ on K and hence

$$\|v - \bar{v}\|^2_{\mathsf{L}^2(K)} \;\leq\; C \int_K |\boldsymbol{x}|^{2\omega-2}\, |v - \bar{v}|^2\, d\boldsymbol{x} \;\leq\; C \int_K |\boldsymbol{x}|^{2\omega}\, |\,\mathrm{D}^1 v|^2\, d\boldsymbol{x}.$$

Applying this result to each component of $\boldsymbol{u}$ gives the existence of $\boldsymbol{u}_0 \in \mathbb{R}^2$ such that

$$\|\boldsymbol{u} - \boldsymbol{u}_0\|^2_{\mathbf{L}^2(K)} \;\leq\; C \,\|\, |\boldsymbol{x}|^{\omega}\, |\,\mathrm{D}^1 \boldsymbol{u}|\, \|^2_{\mathbf{L}^2(K)}.$$

Summarizing, we have shown in the case $K = (0,1)^2$ that

$$\|\boldsymbol{u} - \boldsymbol{u}_0\|^2_{\mathbf{L}^2(K)} \;\leq\; C\, |\boldsymbol{u}|^2_{\mathbf{H}^{2,1}_{\omega}(K)}.$$

The result for the general rectangular element follows from a scaling argument to obtain (36).

Next, consider the case when $\ell = 2$, and again assume K is of unit size. Arguing as in the proof of Lemma 9, we find that

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k \boldsymbol{u}\|_{\mathbf{L}^2(K)} \;\leq\; (1 + \tilde{C}) \inf_{\boldsymbol{u}_1} \|\boldsymbol{u} - \boldsymbol{u}_1\|_{\mathbf{H}^1(K)},$$

where $\boldsymbol{u}_1$ is a pull-back bilinear function. By Lemma 4.25 [20], we bound the right-hand side by

$$C \,\|\, |\boldsymbol{x}|^{\omega}\, |\,\mathrm{D}^2 \boldsymbol{u}|\, \|_{\mathbf{L}^2(K)}.$$

Hence, in the case when $K = (0,1)^2$, we have

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_k \boldsymbol{u}\|_{\mathbf{L}^2(K)} \;\leq\; C \,\|\, |\boldsymbol{x}|^{\omega}\, |\,\mathrm{D}^2 \boldsymbol{u}|\, \|_{\mathbf{L}^2(K)}.$$

The result for the general rectangular element again follows by a scaling argument. The estimate for the L^2-projection π_k is obtained using similar arguments. ∎

4.4. Approximation properties of the global interpolation operators. So far, we have restricted our attention to the approximation properties of the local interpolation operators $\boldsymbol{\Pi}_{K,k}$. In this section, these results are extended to the global interpolation operator defined in section 3.4. The main difference, it will be recalled, between the global and local interpolants concerns the values of the degrees of freedom on a broken edge, in the situation shown in Figure 2. Observe that we may use definition (10) to define a local counterpart of the global interpolant $\boldsymbol{\Pi}$ over the reference element corresponding to K by the rule

$$\widehat{\boldsymbol{\Pi}} \;=\; \mathcal{L}_K^{-1}\, \boldsymbol{\Pi}\, \mathcal{L}_K,$$

and hence, without loss of generality, we may assume K is the reference element. Equally well, we may assume K has only one edge (γ) that lies on a broken edge (γ'). For convenience, we consider the **RT** case in the first instance.

The difference between $\boldsymbol{\Pi}$ and the local interpolant $\boldsymbol{\Pi}_k$ arises from definition (15):

$$\int_{\gamma'} (\boldsymbol{\Pi}\,\boldsymbol{u})|_{\gamma'} \cdot \boldsymbol{n}\; q\; ds \;=\; \int_{\gamma'} (\boldsymbol{u}\cdot\boldsymbol{n})|_{\gamma'}\; q\; ds \qquad \forall q \in \mathbb{R}_k(\gamma').$$

The definition of the $\mathsf{L}^2(\gamma')$-projection $\mathsf{R}_k^{\gamma'} : \mathsf{L}^2(\gamma) \longrightarrow \mathbb{R}_k(\gamma')$ reveals that

$$(\boldsymbol{\Pi}\boldsymbol{u})|_{\gamma'} \cdot \boldsymbol{n} \;=\; \mathsf{R}_k^{\gamma'}\left(\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma'}\right) \quad \text{ on } \gamma'$$

while the local interpolant satisfies

$$(\boldsymbol{\Pi}_k\boldsymbol{u})|_{\gamma'} \cdot \boldsymbol{n} \;=\; \mathsf{R}_k^{\gamma}\left(\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma'}\right) \quad \text{ on } \gamma.$$

It is not difficult to verify that

$$\begin{aligned}(\boldsymbol{\Pi}_k\,\boldsymbol{u} - \boldsymbol{\Pi}\boldsymbol{u})_x &= \;\mathcal{E}_k^{\gamma}\left[\mathsf{R}_k^{\gamma}\left(\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma}\right) - \mathsf{R}_k^{\gamma'}\left(\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma'}\right)\Big|_{\gamma}\right]\\ &= \;\mathcal{E}_k^{\gamma}\;\mathsf{R}_k^{\gamma}\;\mathrm{Tr}_{\gamma}\left[\boldsymbol{u}\cdot\boldsymbol{n} - \mathsf{R}_k^{\gamma'}\left(\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma_{\prime}}\right)\right],\end{aligned}$$

and then, using the definition of the **RT**-extension $\mathcal{E}_k^{\gamma}$, we deduce that

$$\begin{aligned}\|\boldsymbol{\Pi}_k\,\boldsymbol{u} - \boldsymbol{\Pi}\,\boldsymbol{u}\|^2_{\mathsf{L}^2(K)} &\le C\;k^{-1}\;\Big\|\;\mathsf{R}_k^{\gamma}\,\mathrm{Tr}_{\gamma}\left[\boldsymbol{u}\cdot\boldsymbol{n} - \mathsf{R}_k^{\gamma'}(\,\boldsymbol{u}\cdot\boldsymbol{n}|_{\gamma'})\right]\;\Big\|^2_{\mathsf{L}^2(\gamma)}\\ &\le C\;k^{-1}\;\Big\|\;\boldsymbol{u}\cdot\boldsymbol{n} - \mathsf{R}_k^{\gamma'}\,(\boldsymbol{u}\cdot\boldsymbol{n})|_{\gamma'}\;\Big\|^2_{\mathsf{L}^2(\gamma)}\end{aligned}$$

thanks to the stability of the projection R_k^{γ}. Now, suppose that $\boldsymbol{u}|_{K'}$ is expanded as a Legendre series as in (21). Then, if $\gamma' = \gamma_1$,

$$\begin{aligned}\left\|\boldsymbol{u}\cdot\boldsymbol{n} - \mathsf{R}_k^{\gamma'}\,(\boldsymbol{u}\cdot\boldsymbol{n})|_{\gamma'}\right\|^2_{\mathsf{L}^2(\gamma)} \;&\le\; \left\|\boldsymbol{u}\cdot\boldsymbol{n} - \mathsf{R}_k^{\gamma'}(\boldsymbol{u}\cdot\boldsymbol{n})\right\|^2_{\mathsf{L}^2(\gamma')}\\ &=\; \textstyle\sum_{j=k+1}^{\infty}\rho_j\,\left(\sum_{i=0}^{\infty}a_{i,j}(-1)^i\right)^2\\ &\le\; C\;\tfrac{k^2}{\varepsilon}\;\tfrac{(k+1-r)!}{(k+1+r)!}\;[\boldsymbol{u}]^2_{r,K'} \qquad \text{for } r \le k+1,\end{aligned}$$

where the final part of Lemma 7 has been invoked. Hence,

$$\|\boldsymbol{\Pi}_k\,\boldsymbol{u} - \boldsymbol{\Pi}\,\boldsymbol{u}\|^2_{\mathsf{L}^2(K)} \;\le\; C\;\frac{k}{\varepsilon}\;\frac{(k+1-r)!}{(k+1+r)!}\;[\boldsymbol{u}]^2_{r,K'}.$$

A similar expression holds for the **BDFM** elements with an additional factor of k. Corresponding results hold on a rectangular element using a scaling argument as in Lemma 9. Consequently, using the triangle inequality

$$\|\boldsymbol{u} - \boldsymbol{\Pi}\,\boldsymbol{u}\|_{\mathsf{L}^2(K)} \;\le\; \|\boldsymbol{u} - \boldsymbol{\Pi}_k\,\boldsymbol{u}\|_{\mathsf{L}^2(K)} \;+\; \|\boldsymbol{\Pi}_k\,\boldsymbol{u} - \boldsymbol{\Pi}\,\boldsymbol{u}\|_{\mathsf{L}^2(K)},$$

we conclude the following.

LEMMA 12. *Let K be a rectangular element of size h_K with $\rho_K = \operatorname{dist}(\mathbf{0}, K)$, and degree $k \geq 2$. Let $K^\star$ denote the union of K and any elements with a broken edge neighboring K. If $\boldsymbol{u} \in \mathbf{H}^{2,\ell}(K^\star)$ with $\ell \in \{1,2\}$, then for $2 \leq s \leq k$,*

$$\|\boldsymbol{u} - \boldsymbol{\Pi}^{\mathsf{RT}}\,\boldsymbol{u}\|^2_{\mathbf{L}^2(K)} \;\leq\; \frac{C\,h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\;k\;\frac{\Gamma(k-s+1)}{\Gamma(k+s+1)}\;|\boldsymbol{u}|^2_{\mathbf{H}^{s,\ell}_\omega(K^\star)}, \tag{38}$$

and for $2 \leq s \leq \max(2, [k/2]) = \tilde{k}$,

$$\|\boldsymbol{u} - \boldsymbol{\Pi}^{\mathsf{BDFM}}\,\boldsymbol{u}\|^2_{\mathbf{L}^2(K)} \;\leq\; \frac{C\,h_K^{2s}}{\varrho_K^{2(\omega+s-\ell)}}\;\tilde{k}^2\;\frac{\Gamma(\tilde{k}-s+1)}{\Gamma(\tilde{k}+s+1)}\;|\boldsymbol{u}|^2_{\mathbf{H}^{s,\ell}_\omega(K^\star)}, \tag{39}$$

where $\boldsymbol{\Pi}$ is the global interpolation operator.

5. Finite element approximation of $\boldsymbol{\mathcal{B}^\ell_\omega(\Omega)}$.

5.1. Approximation on a square. Consider the approximation of a function $\boldsymbol{u} \in \boldsymbol{\mathcal{B}}^\ell_\omega(\widehat{\Omega})$ with $\ell \in \{1,2\}$, in the case where $\widehat{\Omega}$ is the unit square $(0,1)^2$ with the weight function given by $\Phi_\omega(\boldsymbol{x}) = |\boldsymbol{x}|^\omega$. This scenario models the typical situation where the function $\boldsymbol{u}$ is the solution of a boundary value problem in the neighborhood of a corner located at the origin. Later, we shall consider more general configurations.

The domain $\widehat{\Omega}$ is partitioned into elements as follows. Let $\sigma \in (0,1)$ and $M \in \mathbb{N}_0$ be given. The *basic geometric mesh* $\widehat{\mathcal{T}}^M_\sigma$ with $M+1$ *layers* and *grading factor* σ is defined inductively as follows. The initial mesh $\widehat{\mathcal{T}}^0_\sigma$ is taken to be the whole domain $\widehat{\Omega}$. For nonzero M, the mesh $\widehat{\mathcal{T}}^M_\sigma$ is obtained from $\widehat{\mathcal{T}}^{M-1}_\sigma$ by breaking the square element K, with a vertex at the origin, into four rectangles by subdividing its sides in the ratio $\sigma : 1-\sigma$. Figure 4 gives an example of the case of three layers with grading factor $\sigma = 1/2$. The elements forming the jth layer $\widehat{\mathcal{T}}^{M,j}_\sigma$ are numbered $\{K_{i,j}\}_{i=0}^3$ for $j > 1$, as shown in Figure 4. In general, we find that an element K in the jth layer satisfies

$$\begin{cases} \sigma^{M+2-j} \;\leq\; \varrho_K \;=\; \operatorname{dist}(\mathbf{0}, K) \;\leq\; \sqrt{2}\,\sigma^{M+2-j}, \\ \sigma^{M+1-j}\,(1-\sigma) \;\leq\; h_K \;=\; \operatorname{diam}(K) \;\leq\; \sqrt{2}\,\sigma^{M+1-j}\,(1-\sigma), \\ \frac{2}{\kappa} \;\leq\; \frac{h_K}{\varrho_K} \;\leq\; \kappa, \end{cases}$$

where $\kappa = \sqrt{2}\,(1-\sigma)/\sigma$, for $1 < j \leq M+1$. The first layer consists of the single element $K_{1,1} = (0, \sigma^M) \times (0, \sigma^M)$.

The global finite element space is based on the **RT** elements defined, as in (12), by

$$\boldsymbol{\Gamma}^{\mathsf{RT}}_N \;=\; \left\{\; \boldsymbol{u} \;\in\; \mathbf{H}(\operatorname{div}, \widehat{\Omega}) : \quad \boldsymbol{u}|_K \in \boldsymbol{\Gamma}^{\mathsf{RT}}_k \quad \forall K \in \widehat{\mathcal{T}}^M_\sigma \;\right\}$$

with the corresponding polynomial subspace of L^2 defined by

$$V^{\mathsf{Q}}_N = \left\{ q \in \mathsf{L}^2(\widehat{\Omega}) : q \circ F_K^{-1} \in \mathbb{Q}_k \quad \forall K \in \widehat{\mathcal{T}}^M_\sigma \right\}.$$

The spaces $\boldsymbol{\Gamma}^{\mathbf{BDFM}}_N$ and V^{P}_N, obtained from the combination of the $\mathbf{BDFM}_k$ elements with $\mathbb{P}_k$, are defined in the same fashion.

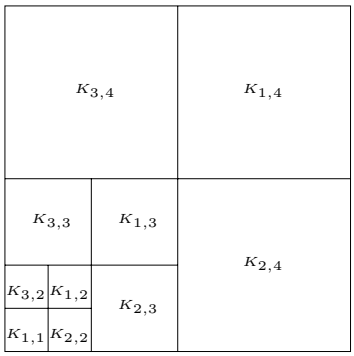


FIG. 4. *Geometric mesh $\widehat{\mathcal{T}}_\sigma^M$ for a grading factor $\sigma = 0.5$ and $M = 3$.*

The main result of this section can now be stated.

THEOREM 13. *Let $\widehat{\Omega} = (0,1)^2$, and $\widehat{\mathcal{T}}_\sigma^M$ is the basic geometric mesh with grading factor $\sigma \in (0,1)$ and $M+1$ layers. Let $\mathbf{\Gamma}_N$–V_N denote the mixed finite element spaces (based on* **RT** *or* **BDFM** *elements) with degree $k = \mu M \geq 2$, and let $\mathbf{\Pi}_N$ and π_N be the associated interpolation operators. If $\boldsymbol{u} \in \boldsymbol{\mathcal{B}}_\omega^\ell(\widehat{\Omega})$ and $p \in \mathcal{B}_\omega^\ell(\widehat{\Omega})$ with $\ell \in \{1,2\}$, then for sufficiently large $\mu > 0$,*

$$\|\boldsymbol{u} - \mathbf{\Pi}_N \boldsymbol{u}\|_{\mathbf{L}^2(\widehat{\Omega})} \;\leq\; C\, \exp(-b\, N^{1/3}),$$

and

$$\|p - \pi_N p\|_{\mathsf{L}^2(\widehat{\Omega})} \;\leq\; C\, \exp(-b\, N^{1/3}),$$

where C, b are positive constants, independent of the dimension N of the space $\mathbf{\Gamma}_N$.

Proof. Let $\boldsymbol{u} \in \boldsymbol{\mathcal{B}}_\omega^\ell(\widehat{\Omega})$. Then, there exist positive constants C and d such that for any element $K \in \widehat{\mathcal{T}}_\sigma^M$,

$$|\boldsymbol{u}|_{\mathbf{H}_\omega^{s,\ell}(K)} \;\leq\; |\boldsymbol{u}|_{\mathbf{H}_\omega^{s,\ell}(\widehat{\Omega})} \;\leq\; C\, d^{s-\ell}\, \sqrt{s}\, \Gamma(s-\ell+1).$$

In particular, applying Lemma 11, we conclude that, since $k \geq 2$,

$$\|\boldsymbol{u} - \mathbf{\Pi}_N \boldsymbol{u}\|^2_{\mathbf{L}^2(K_{1,1})} \;\leq\; Cd^{-2\ell}\sigma^{2M(\ell-\omega)}.$$

Consider the contributions from elements belonging to the jth layer $\widehat{\mathcal{T}}_\sigma^{M,j}$, $2 \leq j \leq M+1$. Recall that any element $K \in \widehat{\mathcal{T}}_\sigma^{M,j}$ has diameter $h_K \leq \kappa \varrho_K$ and that $\varrho_K \leq \sqrt{2}\,\sigma^{M+2-j}$. Inserting these quantities into the bound in Lemma 12 leads to the conclusion that for $2 \leq s \leq k$,

$$\begin{aligned}\|\boldsymbol{u} - \mathbf{\Pi}_N \boldsymbol{u}\|^2_{\mathbf{L}^2(K)} &\\ \leq\; C\, d^{-2\ell}\sigma^{2(M+2-j)(\ell-\omega)}\, &\frac{\Gamma(k-s+1)}{\Gamma(k+s+1)}\, s\, k\, \Gamma(s-\ell+1)^2\, (\kappa d)^{2s}\end{aligned}$$

for $K \in \widehat{\mathcal{T}}_\sigma^{M,j}$. Using Stirling's approximation $\Gamma(n+1) \sim \sqrt{2\pi n}\,(n/e)^n$, it is not difficult to show that if $s = \alpha k$, with constant $\alpha \in [2/k, 1)$ (to be chosen), then

$$\frac{\Gamma(k-s+1)}{\Gamma(k+s+1)}\, \Gamma(s-\ell+1)^2 (\kappa d)^{2s} \;\leq\; C\, k^{1-2\ell}\, F(2\kappa d, \alpha)^k,$$

where $F : (1,\infty) \times (0,1] \longrightarrow \mathbb{R}$ is the function

$$F(r,\alpha) \;=\; \frac{(1-\alpha)^{1-\alpha}}{(1+\alpha)^{1+\alpha}} \;\left(\frac{\alpha r}{2}\right)^{2\alpha}$$

considered in (3.3.73) [20]. In particular, the function $F(2\kappa d,\cdot)$ has minimum value $F_{\min} \in (0,1)$ at $\alpha_{\min} = (1+\kappa^2 d^2)^{-1/2}$.

Summing up the layer-wise contributions to the total error reveals that

$$\|\boldsymbol{u} - \boldsymbol{\Pi}_N\,\boldsymbol{u}\|^2_{\mathbf{L}^2(\widehat{\Omega})} \;\le\; C\;\sigma^{2M(\ell-\omega)}\left\{1 \;+\; k^{3-2\ell}F(2\kappa d,\alpha)^k \sum_{j=2}^{M+1} \sigma^{2(2-j)(\ell-\omega)}\right\}. \tag{40}$$

The second term in parentheses may be bounded by

$$C\;\sigma^{-2M(\ell-\omega)}k^{3-2\ell}F(2\kappa d,\alpha)^k, \tag{41}$$

and, inserting $k = \mu M$ and choosing $\alpha = \max(2/\mu M, \alpha_{\min}) = \alpha_{\min}$ (for M large), this is in turn bounded by

$$C\;M^{3-2\ell}\left(\frac{F_{\min}^{\mu}}{\sigma^{2(\ell-\omega)}}\right)^M. \tag{42}$$

Finally, choosing

$$\mu > \max\left\{1, \frac{2(\ell-\omega)\ln\sigma}{\ln F_{\min}}\right\}$$

ensures that the factor in parentheses has magnitude less than unity, and as a consequence, the multiplicative term in (40) is bounded above for all M. Hence,

$$\|\boldsymbol{u} - \boldsymbol{\Pi}^{\mathsf{RT}}_N\,\boldsymbol{u}\|^2_{\mathbf{L}^2(\Omega)} \;\le\; C\;\sigma^{2M(\ell-\omega)}.$$

The dimension N of the space $\boldsymbol{\Gamma}_N$ satisfies

$$N = (1+3(M-1))\dim \mathbf{RT}_{\mu M} \approx \mu^2 M^3, \tag{43}$$

and hence, since $\sigma < 1$, it follows that

$$\|\boldsymbol{u} - \boldsymbol{\Pi}^{\mathsf{RT}}_N\,\boldsymbol{u}\|^2_{\mathbf{L}^2(\Omega)} \;\le\; C\;\exp(\,-b\,N^{1/3}\,),$$

where C and b are positive constants independent of M and k. The estimate for $p \in \mathcal{B}^{\ell}_{\omega}$ follows similar lines, using the estimate for π_k in Lemma 10 in place of Lemma 12. ∎

5.2. Generalization to polygonal domains. So far, we have obtained exponential convergence for the basic square $\widehat{\Omega}$ when the function $\boldsymbol{u}$ has only one singularity. We now generalize to polygonal domains.

For a polygonal domain Ω, the geometric mesh $\mathcal{T}^M_\sigma$, with $M+1$ layers and the grading factor $0 < \sigma < 1$, is constructed by patching together affine images of basic geometric meshes $\widehat{\mathcal{T}}^M_\sigma$ near vertices in conjunction with a fixed, quasi-uniform partition of uniform polynomial order on the interior. Changes in type of boundary conditions are treated using two copies of $\widehat{\mathcal{T}}^M_\sigma$, while re-entrant corners use three (see Figure 5).

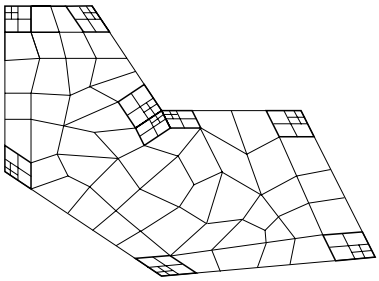


FIG. 5. *Typical mesh used on a polygonal domain.*

THEOREM 14. *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain partitioned into a geometric mesh $\mathcal{T}_\sigma^M$ of affine elements with grading factor $\sigma \in (0,1)$ and $M+1$ layers. Let $\mathbf{\Gamma}_N$–V_N denote the mixed finite element space (based on* **RT** *or* **BDFM** *elements) of degree $k = \mu M$, and let $\mathbf{\Pi}_N$ and π_N be the associated interpolation operators. Suppose $\boldsymbol{u} \in \boldsymbol{\mathcal{B}}^\ell_\omega(\Omega)$ and $p \in \mathcal{B}^\ell_\omega(\Omega)$, with $\ell \in \{1,2\}$, for $\omega = (\omega_1, \ldots, \omega_m)$, $0 < \omega_i < 1$, dependent on the angles. Then, there exists $\mu^\star > 0$ such that if $\mu > \mu^\star$, then*

$$\|\boldsymbol{u} - \mathbf{\Pi}_N \boldsymbol{u}\|_{\mathbf{L}^2(\Omega)} \;\le\; C\ \exp(-b\, N^{1/3}),$$

and

$$\|p - \pi_N p\|_{\mathsf{L}^2(\Omega)} \;\le\; C\ \exp(-b\, N^{1/3}),$$

where C, b are positive constants, independent of the dimension N of the space $\mathbf{\Gamma}_N$.

Proof. First consider the situation of a parallelogram $\tilde{\Omega}$ with a singularity at only one corner. The geometric mesh $\tilde{\mathcal{T}}_\sigma^M$ on $\tilde{\Omega}$ is then constructed using an affine image of the basic geometric mesh $\widehat{\mathcal{T}}_\sigma^M$ on the basic domain $\widehat{\Omega} = (0,1)^2$; see Figure 6.

Let $\tilde{F} : \widehat{\Omega} \longrightarrow \tilde{\Omega}$ denote the affine mapping between the basic square $\widehat{\Omega}$ and the parallelogram $\tilde{\Omega}$. Let $\tilde{\boldsymbol{u}} \in \boldsymbol{\mathcal{B}}^\ell_\omega(\tilde{\Omega})$ and denote $\widehat{\boldsymbol{u}} = \tilde{\mathcal{L}}\, \tilde{\boldsymbol{u}}$ where $\tilde{\mathcal{L}}$ is the Piola transformation. Then

$$\|\tilde{\boldsymbol{u}} - \mathbf{\Pi}_N\, \tilde{\boldsymbol{u}}\|_{\boldsymbol{L}^2(\tilde{\Omega})} \;\le\; C \|\widehat{\boldsymbol{u}} - \widehat{\mathbf{\Pi}}_N\, \widehat{\boldsymbol{u}}\|_{\boldsymbol{L}^2(\widehat{\Omega})},$$

where C is a positive constant independent of N, and $\widehat{\mathbf{\Pi}}_N$ is the interpolation operator on the domain $\widehat{\Omega}$. Since the function $\widehat{\boldsymbol{u}}$ belongs to the space $\boldsymbol{\mathcal{B}}^\ell_\omega(\widehat{\Omega})$, we may apply Theorem 13 and obtain exponential convergence on the parallelogram $\tilde{\Omega}$.

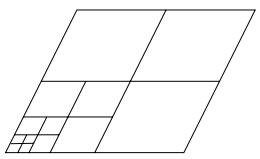


FIG. 6. *Geometric mesh $\tilde{\mathcal{T}}_\sigma^M$ on a parallelogram $\tilde{\Omega}$.*

The case of a general polygonal domain follows summing contributions to the error from the interpolant over each corner patch with the error arising from the approximation using the interpolant over the interior. Observe that the true solution is

analytic on the interior (see, for example, [20, Exercise 4.45]) so that a uniform polynomial distribution over a fixed meshing of the interior region will deliver exponential rates of convergence thanks to [21, Theorem 5.1]. ∎

6. Application to mixed finite element approximation. We illustrate the foregoing results by using them to derive exponential rates of convergence for mixed finite element approximation of problem (6). This problem may be formulated in mixed variational form as follows: find $(\boldsymbol{u}, p) \in \mathbf{H}(\mathrm{div}, \Omega) \times \mathsf{L}^2(\Omega)$ such that

$$\left\{ \begin{array}{lll} a(\boldsymbol{u}, \boldsymbol{v}) \ + \ (p, \mathrm{div}\, \boldsymbol{v}) = \ 0 & & \forall\, \boldsymbol{v} \ \in \ \mathbf{H}(\mathrm{div}, \Omega), \\ (\mathrm{div}\, \boldsymbol{u}, q) \ + \ (f, q) \ \ = \ 0 & & \forall\, q \in \mathsf{L}^2(\Omega), \end{array} \right. \tag{44}$$

where the bilinear form a is defined as follows:

$$a(\boldsymbol{u}, \boldsymbol{v}) \ = \ \int_\Omega A^{-1}\ \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x})\ d\boldsymbol{x}.$$

Problem (44) will be approximated using a pair of *hp*-finite element spaces, based on **RT** or **BDFM** elements, $\boldsymbol{\Gamma}_N \subset \mathbf{H}(\mathrm{div}, \Omega)$, and $\mathsf{V}_N \subset \mathsf{L}^2(\Omega)$, giving the following discrete problem: find $(\boldsymbol{u}_N, p_N) \in \boldsymbol{\Gamma}_N \times \mathsf{V}_N$ such that

$$\left\{ \begin{array}{lll} a(\boldsymbol{u}_N, \boldsymbol{v}_N) \ + \ (p_N, \mathrm{div}\, \boldsymbol{v}_N) = \ 0 & & \forall\, \boldsymbol{v}_N \ \in \ \boldsymbol{\Gamma}_N, \\ (\mathrm{div}\, \boldsymbol{u}_N, q_N) - (f, q_N) \ \ = \ 0 & & \forall\, q_N \ \in \ \mathsf{V}_N. \end{array} \right. \tag{45}$$

Mixed approximations of this type are studied by Brezzi and Fortin [9, pp. 137–139].

THEOREM 15. *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain partitioned into geometric meshes $\mathcal{T}_\rho^M$ with grading factor $0 < \rho < 1$ and $M + 1$ layers, as described above. Suppose that the data f are analytic and $f \in \mathcal{B}^0_\omega(\Omega)$ for $\omega = (\omega_1, \ldots, \omega_m)$, $0 < \omega_i < 1$, dependent on the interior angles. Then, there exists $\mu^\star > 0$ such that if $k = \mu M$, with $\mu > \mu^\star$, then*

$$\|\boldsymbol{u} - \boldsymbol{u}_N\|_{\mathbf{H}(\mathrm{div},\Omega)} \ + \ \|p - p_N\|_{\mathsf{L}^2(\Omega)} \ \leq \ C\ \exp(\ -b\, N^{1/3}\ ),$$

where C and b are positive constants independent of N.

Proof. First, observe that thanks to the commuting diagram property,

$$\|\,\mathrm{div}(\boldsymbol{u} - \boldsymbol{\Pi}_N \boldsymbol{u})\|_{\mathsf{L}^2(\Omega)} = \|\,\mathrm{div}\, \boldsymbol{u} - \pi_N(\mathrm{div}\, \boldsymbol{u})\|_{\mathsf{L}^2(\Omega)} = \|f - \pi_N f\|_{\mathsf{L}^2(\Omega)},$$

and then, recalling that f is analytic, the approximation properties of the L^2-projection mean that a pure p-version procedure on *any* fixed partition (and a fortiori on those partitions considered here) will deliver exponential rates of convergence for this quantity. Applying [9, Proposition IV.1.1] leads to the abstract a priori error estimate

$$\|\boldsymbol{u} - \boldsymbol{u}_N\|_{\mathbf{H}(\mathrm{div},\Omega)} \ + \ \|p - p_N\|_{\mathsf{L}^2(\Omega)} \ \leq \ C\ \left(\ \|\boldsymbol{u} - \boldsymbol{\Pi}_N \boldsymbol{u}\|_{\mathbf{H}(\mathrm{div},\Omega)} + \|p - \pi_N p\|_{\mathsf{L}^2(\Omega)}\ \right),$$

and the result follows from the above observation and Theorem 14. ∎

REFERENCES

[1] M. AINSWORTH AND B. SENIOR, *Aspects of an adaptive hp-finite element method: Adaptive strategy, conforming approximation and efficient solvers*, Comput. Methods Appl. Mech. Engrg., 150 (1997), pp. 65–87.

[2] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data. Part* I. *Boundary value problems for linear elliptic equation of second order*, SIAM J. Math. Anal., 19 (1988), pp. 172–203.

[3] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data,* II: *The trace spaces and application to the boundary value problems with nonhomogeneous boundary conditions*, SIAM J. Math. Anal., 20 (1989), pp. 763–781.

[4] I. Babuška, B. Guo, and E. Stephan, *On the exponential convergence of the hp-version for boundary element Galerkin methods on polygons*, Math. Methods Appl. Sci., 12 (1990), pp. 413–427.

[5] I. Babuška, R. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[6] I. Babuška and M. Suri, *On locking and robustness in the finite element method*, SIAM J. Numer. Anal., 29 (1992), pp. 1261–1293.

[7] I. Babuška and M. Suri, *The p and h-p versions of the finite element method, basic principles and properties*, SIAM Rev., 36 (1994), pp. 578–632.

[8] A. Bossavit, *Computational Electromagnetism. Variational Formulation, Complementarity, Edge Elements*, Academic Press, Orlando, FL, 1998.

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Elements Methods*, Springer–Verlag, New York, 1991.

[10] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[11] P. Grisvard, *Singularities in Boundary Value Problems*, Springer–Verlag, Berlin, Masson, Paris, 1992.

[12] B. Guo and I. Babuška, *The hp version of the finite element method. Part* I: *The basic approximation method*, Comput. Mech., 1 (1986), pp. 21–41.

[13] B. Guo and I. Babuška, *The hp version of the finite element method. Part* II: *General results and applications*, Comput. Mech., 1 (1986), pp. 203–220.

[14] R. Hiptmair, *Canonical construction of finite elements*, Math. Comp., 68 (1999), pp. 1325–1346.

[15] R. Hiptmair, *Higher order Whitney forms*, Progress in Electromagnetics Research, 32 (2001), pp. 271–299.

[16] R. Hiptmair, *Higher order Whitney forms–abstract*, J. Electromaget. Wave Applic., 15 (2001), pp. 341–342.

[17] M. Lee and F. Milner, *Mixed finite element methods for nonlinear elliptic problems: The hp version*, J. Comput. Appl. Math., 85 (1998), pp. 239–261.

[18] F. Milner and M. Suri, *Mixed finite element methods for quasilinear second order elliptic problems: The p-version*, RAIRO Modél. Math. Anal. Numér., 26 (1992), pp. 913–931.

[19] P. Monk, *On the p and hp extension of Nédélec's curl conforming elements*, J. Comput. Appl. Math., 53 (1994), pp. 117–137.

[20] C. Schwab, *p- and hp-Finite Element Methods, Theory and Applications in Solid and Fluid Mechanics*, Oxford University Press, New York, 1998.

[21] R. Stenberg and M. Suri, *An hp error analysis of MITC plate elements*, SIAM J. Numer. Anal., 34 (1997), pp. 544–568.

[22] M. Suri, *On the stability and convergence of higher order mixed finite element methods of second order elliptic problems*, Math. Comp., 54 (1990), pp. 1–19.

[23] L. Vardapetyan and L. Demkowicz, *hp adaptive finite elements in electromagnetics*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 331–344.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2069–2084

A CLASS OF INTRINSIC SCHEMES FOR ORTHOGONAL INTEGRATION*

ELENA CELLEDONI† AND BRYNJULF OWREN‡

Abstract. Numerical integration of ODEs on the orthogonal Stiefel manifold is considered. Points on this manifold are represented as $n \times k$ matrices with orthonormal columns, of particular interest is the case when $n \gg k$. Mainly two requirements are imposed on the integration schemes. First, they should have arithmetic complexity of order nk^2. Second, they should be intrinsic in the sense that they require only the ODE vector field to be defined on the Stiefel manifold, as opposed to, for instance, projection methods. The design of the methods makes use of retractions maps. Two algorithms are proposed, one where the retraction map is based on the QR decomposition of a matrix, and one where it is based on the polar decomposition. Numerical experiments show that the new methods are superior to standard Lie group methods with respect to arithmetic complexity, and may be more reliable than projection methods, owing to their intrinsic nature.

Key words. time integration, geometric integration, numerical integration of ordinary differential equations on manifolds, numerical analysis, Stiefel manifold, homogeneous spaces

AMS subject classification. 65L05

PII. S0036142901385143

1. Introduction.

The elements of the orthogonal Stiefel manifold are often represented as $n \times k$ matrices with orthonormal columns, where $n \geq k$,

$$\mathcal{V}_{n,k} = \{Q \in \mathbb{R}^{n\times k} : Q^T Q = I_{k\times k}\}.$$

In particular, one has $\mathcal{V}_{n,n} = \mathcal{O}(n)$, the Lie group of $n \times n$ orthogonal matrices.

Many applications involve computations with such matrices. One is the calculation of Lyapunov exponents; see [7, 14] for an overview. Another involves optimization problems in multivariate data analysis [26].

In this paper, we shall study the problem of approximating a solution of an ODE system on $\mathcal{V}_{n,k}$. We think of $\mathcal{V}_{n,k}$ as a special case of a manifold $\mathcal{M}$, and we let $T\mathcal{M}$ denote the tangent bundle of $\mathcal{M}$. A vector field on $\mathcal{M}$ is then a section $F : \mathcal{M} \to T\mathcal{M}$ which assigns to each $Q \in \mathcal{M}$ a tangent vector $F(Q) \in T_Q\mathcal{M}$. The ODE system is also allowed to be nonautonomous, thus the vector field may also depend on $t \in \mathbb{R}$,

$$\dot{Q} = F(t, Q), \quad Q(t_0) = Q_0 \in \mathcal{M}.$$

As indicated above, the orthogonal Stiefel manifold is naturally embedded in the Euclidean space of all real $n \times k$ matrices, and it is also quite common to use this representation in computer programs. The situation is similar for other manifolds $\mathcal{M}$ embedded in a Euclidean space E. Nevertheless, one should distinguish numerical methods which are intrinsic and those which are extrinsic. The latter type of methods make use of an extension of the vector field F to all of E, or at least to some

*Received by the editors February 15, 2001; accepted for publication (in revised form) May 17, 2002; published electronically December 13, 2002. This work was in part sponsored by the Norwegian Research Council under contract 111038/410, through the SYNODE project available from http://www.math.ntnu.no/num/synode/.

http://www.siam.org/journals/sinum/40-6/38514.html

†Department of Computational Engineering, SINTEF Applied Mathematics, N-7465 Trondheim, Norway (elenac@math.ntnu.no).

‡Department of Mathematical Sciences, NTNU, N-7491 Trondheim, Norway (Brynjulf.Owren@math.ntnu.no).

neighborhood of $\mathcal{M}$. Such methods include the projection methods; see, for instance, [18, 13, 20]. More recently, symmetric projection methods have been proposed in [1] to enforce conservation of energy, and later in a more general setting by [16]. Another approach is to apply direct solvers for index-2 DAEs. In this case, it is useful to formulate the problem in terms of a constrained system with Lagrange multipliers, say

$$\dot{Q} = F(t, Q) + QH, \quad Q^T Q = I_k, \quad Q \in \mathbb{R}^{n\times k}, \quad H \in \mathbb{R}^{k\times k},$$

H being a symmetric matrix. Its elements are the Lagrange multipliers which can be determined by the differentiated constraint

$$\dot{Q}Q^T + Q\dot{Q}^T = 0.$$

More recently, there has been an increased interest in designing intrinsic integration methods for manifolds in general. Examples of such methods are the Crouch–Grossman methods [6] and the RKMK methods proposed by Munthe-Kaas; see, e.g., [21]. At least when used in a naive manner, these intrinsic methods are not entirely satisfactory from the point of view of computational complexity when applied to problems on Stiefel manifolds with $n \gg k$. The reason is that they usually demand operations to be performed on $n \times n$ matrices. In particular, one typically applies operations that cost $\mathcal{O}(n^3)$ flops. In comparison, the projection of a matrix in $\mathbb{R}^{n\times k}$ onto $\mathcal{M}$ by means of the QR factorization cost only $\mathcal{O}(nk^2)$ flops. In this paper, we will consider maps known as retractions to impose local coordinates on the Stiefel manifold. The methods presented here fall in the same category as the so-called methods based on local charts, the ones proposed by Potra and Rheinboldt in [23]; see also [24] and the recent monograph by Hairer, Lubich, and Wanner [17]. Our coordinate space is always the tangent space at the initial value of the next time step. The evaluation of the retraction map and its (inverse) derivative can be shown to have complexity $\mathcal{O}(nk^2)$. Thus, if the evaluation of the ODE vector field can be done with the same complexity, we obtain an integration method which has an overall complexity of $\mathcal{O}(nk^2)$.

We present numerical results which demonstrate the low cost compared to Lie group methods implemented in the standard way. We also give numerical evidence to show that intrinsic methods in some cases are more preferable than extrinsic methods like projection methods.

2. Using retraction as a coordinate map. To begin with, let us consider in general a differentiable manifold $\mathcal{M}$ and a differential equation given by means of a vector field $F(t, y)$ such that for each t, $F(t, \cdot)$ is a vector field $\mathcal{M} \to T\mathcal{M}$, so

$$y'(t) = F(t, y(t)). \tag{1}$$

We will here use a particular choice of local charts for the manifold $\mathcal{M}$ and obtain resulting locally defined vector fields similar to what is called state-space formulations in the DAE literature.

It has been proposed by Shub [25] to use a smooth mapping $\mathcal{R} : T\mathcal{M} \to \mathcal{M}$ in the design of Newton iterations on manifolds. For the restriction $\mathcal{R}_p$ of $\mathcal{R}$ to the tangent space $T_p\mathcal{M}$ of $\mathcal{M}$ at the point $p \in \mathcal{M}$ we require the following:

1. $\mathcal{R}_p$ is defined in some open ball $B(0, r_p)$ of radius r_p about $0 \in T_p\mathcal{M}$.
2. $\mathcal{R}_p(v) = p$ if and only if $v = 0 \in T_p\mathcal{M}$.
3. $\mathcal{R}'_p\big|_0 = Id_{T_p\mathcal{M}}$.

The essence of the idea we present here lies in the fact that $\mathcal{R}_p$ serves to define local coordinates of the manifold $\mathcal{M}$ in a neighborhood of the point p. We can thus represent the solution of the differential equation in the form

$$y(t) = \mathcal{R}_p(\sigma(t)), \quad \sigma(t) \in T_p\mathcal{M}. \tag{2}$$

By differentiating with respect to t and by using (1) we get

$$y'(t) = \mathcal{R}_p'\big|_{\sigma(t)}(\sigma'(t)) = F(t, \mathcal{R}_p(\sigma(t))).$$

For $\sigma(t)$ sufficiently close to $0 \in T_p\mathcal{M}$ the map $\mathcal{R}_p'\big|_{\sigma(t)}$ is invertible, so we obtain a differential equation for $\sigma(t)$ as follows:

$$\sigma'(t) = \big(\mathcal{R}_p'\big|_{\sigma(t)}\big)^{-1}(F(t, \mathcal{R}_p(\sigma(t)))). \tag{3}$$

We may now approximate (3) by using a standard ODE solver, and we may subsequently transform the result back to $\mathcal{M}$ via (2).

This approach is intrinsic in the sense that it does not depend on whether $\mathcal{M}$ has been embedded in a bigger (say Euclidean) space with a corresponding extension of the vector field F.

2.1. Lie group methods. The procedure we have just described is very similar to the Lie group methods proposed by, for instance, Munthe-Kaas [21], Owren and Marthinsen [22], and Diele, Lopez, and Peluso [10]. It also has a lot in common with the approach based on Givens–Householder transformations proposed by Dieci and Van Vleck [9]. In fact, the coordinates based on Givens's rotations is a special case of canonical coordinates of the second kind discussed in [22].

In the case where $\mathcal{M} = G$ is a Lie matrix group, the Lie groups methods are equivalent to what we get in the approach above by setting

$$\mathcal{R}_p(v) = \Phi(v \cdot p^{-1}) \cdot p,$$

where Φ is some sufficiently smooth mapping from $\mathfrak{g}$ to G and $\mathfrak{g}$ is the Lie algebra corresponding to the Lie group G. Typically, one may use $\Phi = \exp$, the matrix exponential Munthe-Kaas used in his first papers, but other choices are possible, as discussed in the other papers cited above.

For homogeneous spaces, of which the orthogonal Stiefel manifold is an example, the methods by Munthe-Kaas apply an action by a Lie group. $\mathcal{V}_{n,k}$ is acted upon by $\mathrm{SO}(n)$, the group of $n \times n$ matrices with unit determinant. As an example of a group action, one may use left multiplication:

$$\Lambda(g, p) := \Lambda_p(g) = g \cdot p, \quad p \in \mathcal{V}_{n,k},\ g \in \mathrm{SO}(n).$$

The linear space of $n \times n$ skew-symmetric matrices, denoted $\mathfrak{so}(n)$, is mapped by the matrix exponential into $\mathrm{SO}(n)$, and by composing this with the action Λ one obtains a map $\lambda_p = \Lambda_p \circ \exp : \mathfrak{so}(n) \to \mathcal{V}_{n,k}$. This map λ_p is a smooth map from some neighborhood of $0 \in \mathfrak{so}(n)$ onto some neighborhood of $p \in \mathcal{V}_{n,k}$. We thus get a representation of the solution $y(t)$ near the point $p \in \mathcal{V}_{n,k}$ quite similar to (2):

$$y(t) = \lambda_p(\sigma(t)).$$

By differentiation, this leads again to a differential equation for $\sigma(t)$ but now on the space $\mathfrak{so}(n)$ of skew-symmetric $n \times n$ matrices. The obvious drawback with this

approach is that we have replaced a differential system on the manifold $\mathcal{V}_{n,k}$ having dimension $nk - k(k-1)/2$ by an equation on a linear space which has dimension $n(n-1)/2$. Whenever $n \gg k$ this may lead to a large increase in the number of degrees of freedom. The advantage in using the retraction approach is that we always obtain a coordinate mapping for the manifold with exactly the same number of degrees of freedom as the dimension of the manifold.

2.2. Riemannian manifolds. In the context of a Riemannian manifold, one may use the exponential mapping as defined in terms of geodesics (geodesic flow) as a retraction map. Following, for instance, Chavel [5], we define

$$\mathcal{R}_p(v) = \exp_p(v) = \gamma_v(1),$$

where $\gamma_v(t)$ is the geodesic emanating from p with $\dot\gamma(0) = v$. It is known that $\exp_p$ is defined and of maximal rank in a neighborhood of $0 \in T_p\mathcal{M}$. The derivative map of $\mathcal{R}_p$ is related to the Jacobi field Y satisfying the Jacobi equation; see [5, pp. 70—82]. We let ∇ be the Levi-Civita connection with respect to the Riemannian metric on $\mathcal{M}$ and let $\mathbf{R}$ be the corresponding curvature tensor. We consider the vector field defined along the geodesic γ, $\gamma(0) = p,\ \dot\gamma(0) = v$, satisfying the boundary value problem

$$\nabla_t^2 Y + \mathbf{R}(\dot\gamma, Y)\dot\gamma = 0, \quad Y(0) = 0, \quad Y(1) = w.$$

Then

$$\left(\mathcal{R}_p'\big|_v\right)^{-1}(w) = (\nabla_t Y)(0).$$

In practice, Riemannian manifolds are often given naturally as a submanifold of a Euclidean space, say $V = \mathbb{R}^n$. In this case one can define at each point $p \in \mathcal{M}$ the orthogonal complement to $T_p\mathcal{M}$ which we denote by $N_p\mathcal{M}$, the normal space, so $V = T_p\mathcal{M} \oplus N_p\mathcal{M}$ for every $p \in \mathcal{M}$. Similar to what has been described in [17, 23], we introduce a map $n_p : T_p\mathcal{M} \to N_p\mathcal{M}$ and define a retraction as

$$\mathcal{R}_p(v) = p + v + n_p(v),$$

where $n_p(v)$ is defined such that $R_p(v) \in \mathcal{M}$ for each v belonging to a sufficiently small neighborhood of $0 \in T_p\mathcal{M}$. One calculates the derivative

$$\mathcal{R}_p'\big|_v(w) = w + n_p'\big|_v(w),$$

so the image of the derivative in $T_{\mathcal{R}_p(v)}\mathcal{M}$ is naturally split into components in $T_p\mathcal{M}$ and $N_p\mathcal{M}$. It follows that the inverse of the derivative map is obtained by applying the orthogonal projector $\mathbb{P}_p : V \to T_p\mathcal{M}$:

$$\left(\mathcal{R}_p'\big|_v\right)^{-1}(y) = \mathbb{P}_p y,\ y \in T_{R_p(v)}\mathcal{M} \subset V. \tag{4}$$

One may further characterize the map n_p in the case that the manifold is given locally in terms of constraint functions, say $g : V \to \mathbb{R}^m$, and we characterize open sets $U \in \mathcal{M}$ as

$$U = \{x \in V : g(x) = 0\}.$$

In fact, in many interesting cases, the constraint function defines the whole manifold; for instance, for Stiefel manifolds we obviously have $m = k(k+1)/2$ with the whole manifold $\mathcal{V}_{n,k}$ characterized by g inferred from the constraints $Y^TY = I_k,\ Y \in \mathcal{V}_{n,k}$.

Denote by $g'(x)$ the Jacobian matrix of g evaluated at x and suppose that $\operatorname{rank} g'(x) = m$. Clearly, in this setting, one can take $T_x\mathcal{M} = \ker g'(x)$, and following, for instance, Potra and Yen [24], one may calculate an orthonormal basis for $T_x\mathcal{M}$ by computing a QR decomposition

$$g'(x)^T = Q_x R_x = \begin{bmatrix} V_x & U_x \end{bmatrix} \cdot \begin{bmatrix} R_0 \\ 0 \end{bmatrix}.$$

A basis for $T_x\mathcal{M}$ is then given by the first $n-m$ columns of U_x. A map $\mathcal{R}_p : T_p\mathcal{M} \to \mathcal{M}$ is constructed by setting $q = \mathcal{R}_p(v)$, where $q-(p+v) \in N_p\mathcal{M}$. Using the representation $v = U_p y,\ y \in \mathbb{R}^m$, it follows immediately that the point q is obtained by solving the equations

$$U_p^T(q-p) - y = 0, \quad g(q) = 0. \tag{5}$$

It is straightforward to check that this map is a retraction. In [24] it is proposed to solve (5) by Newton iteration, but we shall see later that when applied on Stiefel manifolds one can find retraction maps which can be computed by direct methods, i.e., methods which do not involve iteration.

3. A retraction based on the reduced QR decomposition. The tangent spaces $T_P\mathcal{V}_{n,k}$ can be characterized in various different ways, and we refer to [12] for an introduction. In the following we shall just need the following characterization:

$$T_P\mathcal{V}_{n,k} = \{v \in \mathbb{R}^{n\times k} : P^T v \in \mathfrak{so}(k)\}.$$

This fact follows easily by letting $Q(t)$ be a smooth curve on $\mathcal{V}_{n,k}$ satisfying $Q(0) = P$ and $\dot{Q}(0) = v$, and thereafter differentiating the relation $Q^TQ = I_k$ at $t = 0$. Using the Frobenius inner product $\langle P, Q\rangle = \operatorname{tr}(Q^TP)$, one has the normal space characterization

$$N_P\mathcal{V}_{n,k} = \{PS \,|\, S \ k \times k \text{ symmetric}\}.$$

The retraction in section 2.2 applied to the Stiefel manifold was discussed in [17], and one uses the map

$$R_P(v) = P + v + n_P(v), \quad n_P(v) = PS_P(v),$$

where $S_P(v) =: S$ is a symmetric matrix satisfying the Riccati equation

$$S^2 + 2S + v^Tv + v^TPS + SP^Tv = 0.$$

It is straightforward to check that this map is a retraction. The Riccati equation can be solved by iteration. The important thing to note is that the iteration involves only computations with $k \times k$ matrices once the coefficient matrices have been set up. One therefore expects a complexity of order k^3 for each subsequent iteration.

From (4) we have

$$\left(\mathcal{R}'_P|_v\right)^{-1}(w) = \mathbb{P}_P(w), \quad w \in T_{R_P(v)}\mathcal{V}_{n,k},$$

and $\mathbb{P}_P : \mathbb{R}^{n\times k} \to T_P\mathcal{V}_{n,k}$ is the orthogonal projector with respect to the Frobenius inner product on $\mathbb{R}^{n\times k}$. Setting $M = (P^Tw + w^TP)/2$, one get the simple expression

$$\mathbb{P}_P(w) = w - PM.$$

We proceed to propose a retraction based on the QR decomposition which does not involve iteration. Now let $\mathcal{S}(n,k)$ be the manifold of $n \times k$ matrices of rank k. Given any matrix $A \in \mathcal{S}(n,k)$ one can apply an orthogonalization procedure to the columns of A and obtain a decomposition of the form $A = QR$, where $Q \in \mathcal{V}_{n,k}$ and $R \in \mathcal{T}_+(k)$, i.e., R being an upper triangular $k \times k$ matrix with positive diagonal elements. The complexity of this operation is $2nk^2$ flops [15, p. 232]. The decomposition is unique as described above. We denote the QR decomposition map (coproduct) by $\mathrm{qr} : \mathcal{S}(n,k) \to \mathcal{V}_{n,k} \times \mathcal{T}_+(k)$, and we let π_1 be the projection onto the first factor. For any vector $v \in T_P\mathcal{V}_{n,k}$ we define the retraction map $\mathcal{R}_P$ relative to $P \in \mathcal{V}_{n,k}$ as

$$\mathcal{R}_P(v) = (\pi_1 \circ \mathrm{qr})(P + v).$$

In other words, calculate the QR decomposition of the matrix $P + v$ and keep the matrix Q. In addition to being well defined, we can also show by construction that the inverse of $\mathcal{R}_P$ exists in some neighborhood of P. By writing

$$P + v = QR, \tag{6}$$

and showing that for a given $Q \in \mathcal{V}_{n,k}$ sufficiently close to P in some sense to be made clear below, we can calculate v satisfying (6) by an explicit procedure. Looking at (6) columnwise, we have

$$v_j = \sum_{i=1}^{j} R_{ij} Q_i - P_j. \tag{7}$$

We take the inner product on each side by P_m, $m = 1, \dots j$, and exploit the skew-symmetry of the matrix $P^T v$:

$$\sum_{i=1}^{j} R_{ij} \langle P_m, Q_i \rangle = \delta_{mj} - \langle v_m, P_j \rangle, \quad m = 1, \dots, j, \tag{8}$$

where δ_{mj} is the Kronecker function. This is a linear system of j equations for $R_{1j}, \dots, R_{jj}$ and can be solved as long as the jth principal minor of $P^T Q$ is nonsingular. One obtains successively v_j from (7). It is a crucial observation that $\langle v_j, P_j \rangle = 0$ such that the right-hand side of (8) depends only on $v_1, \dots, v_{j-1}$. The arithmetic complexity of this algorithm is $\mathcal{O}(nk^2 + k^3)$.

For the derivative mapping $(\mathcal{R}'_P|_v)^{-1}$, we first use the chain rule to infer that

$$\left(\mathcal{R}'_P\Big|_v\right)^{-1} = \left(\mathcal{R}_P^{-1}\big|_Q\right)', \quad \text{where} \quad Q = \mathcal{R}_P(v).$$

Thus, we let $Q(t)$ be a curve in $\mathcal{V}_{n,k}$ such that $Q(0) = Q \in \mathcal{V}_{n,k}$ and $\dot{Q}(0) = w \in T_Q\mathcal{V}_{n,k}$. By setting $v(t) = \mathcal{R}_P^{-1}(Q(t))$ we get the relation

$$P + v(t) = Q(t) \cdot R(t),$$

which we differentiate with respect to t, and set $t = 0$ to obtain $\dot{v} := \dot{v}(0) = (\mathcal{R}_P^{-1}|_Q)'(w)$. Due to the triangular structure of the matrix R, it makes sense to consider the differentiation columnwise, so we get

$$\dot{v}_j = \sum_{i=1}^{j} \left(w_i R_{ij} + Q_i \dot{R}_{ij}\right). \tag{9}$$

Here $R_{ij} = R_{ij}(0)$ and $\dot{R}_{ij} = \dot{R}_{ij}(0)$. We can solve for $R_{1j}, \ldots, R_{jj}$ by the procedure described above. However, in using retractions for solving ODEs on manifolds, we will see that the map $(\mathcal{R}_P^{-1}|_Q)'$ is always evaluated just after one has computed $Q = \mathcal{R}_P(v)$, and thus we would in practice store the R-matrix obtained as a by-product from this calculation.

Now we can use the fact that $\dot{v} \in T_P\mathcal{V}_{n,k}$ and take again the inner product with P_m:

$$\sum_{i=1}^{j} \langle P_m, Q_i \rangle \dot{R}_{ij} = -\sum_{i=1}^{j} \langle P_m, w_i \rangle R_{ij} - \langle \dot{v}_m, P_j \rangle, \quad m = 1, \ldots, j.$$

Thus, we have a linear system of j equations for the unknowns $\dot{R}_{1j}, \ldots, \dot{R}_{jj}$ whose solution exists whenever the principal minors of P^TQ are nonsingular. Finally, we substitute the obtained values for $\dot{R}_{ij}$ into (9) to obtain the desired tangent matrix $\dot{v}$.

Note that the linear systems for $R_{ij},\ i = 1, \ldots, j$, and $\dot{R}_{ij},\ i = 1, \ldots, j$, are the samep; thus one may use the same LU factorization of P^TQ. Note also that when the point $Q \in \mathcal{V}_{n,k}$ is "close" to the reference point P, we will have $P^TQ \approx I$, and the LU factorization can be done without pivoting. All the k linear systems of equations can be solved by means of the same factorization.

See also Appendix A of [3] for the `Matlab` implementation of this algorithm.

Complexity. The evaluation of $\mathcal{R}_P$ involves one addition of two $n \times k$ matrices and a reduced QR factorization. Using, for instance, the modified Gram–Schmidt algorithm [15], the cost of the QR decomposition is about $2nk^2$ flops.

The computation of $\mathcal{R}_P^{-1}$ itself is not needed in our algorithm, but we count a complexity of

$$(4k^2 + k)n + \frac{4}{3}k^3 - \frac{1}{2}k^2 - \frac{5}{6}k$$

flops. Now, for the calculation of the derivative $\left(\mathcal{R}_P^{-1}|_Q\right)'$, we found that it requires

$$(7k^2 + k)n + 2k^3 + \frac{3}{2}\,k^2 + \frac{1}{2}\,k$$

flops in the case that the matrix $R \in \mathcal{T}_+(k)$ is already given where $\mathcal{R}_P^{-1}(Q) + P = QR$. This is a reasonable assumption when integrating ODEs on the Stiefel manifold, because in the use of integration methods one first applies the retraction to a vector $v \in T_P\mathcal{V}_{n,k}$ to obtain $Q \in \mathcal{V}_{n,k}$ where the vector field F is to be evaluated, and the matrix R is obtained as a by-product.

We conclude this section by noting that in using the retraction approach for solving ODEs on the Stiefel manifold it is required to compute the retraction map and its inverse derivative an equal number of times. The dominating cost of these two operations together is about $9nk^2$ when $1 \ll k \ll n$. This is comparable to using the retraction proposed by Hairer, Lubich, and Wanner in [17]. However, a possible advantage with our approach is that we do not need any iteration and need not worry about convergence, etc. In the other approach one needs to iterate to machine accuracy when solving the Riccati equation to ensure that the retraction maps an element of the tangent space into the manifold.

4. A retraction based on the reduced polar decomposition. As an alternative to QR, one can use the reduced polar decomposition where a matrix $A \in \mathcal{S}(n, k)$

is factored as

$$A \to QH, \quad Q \in \mathcal{V}_{n,k},\ H \in \mathrm{Sym}^+(k),$$

where $\mathrm{Sym}^+(k)$ are the $k \times k$ symmetric positive definite matrices. Thus, in a very similar fashion to the retraction with QR decomposition above, we now define

$$\mathcal{R}_P(v) = (\pi_1 \circ \mathrm{pol})(P + v),$$

where $\mathrm{pol} : \mathcal{S}(n,k) \to \mathcal{V}_{n,k} \times \mathrm{Sym}^+(k)$ is the polar decomposition coproduct. It is well known that for any matrix $\mathcal{S}(n,k)$ the factors Q and H above can be calculated via the reduced singular value decomposition, say $A = V\Sigma W^T$, $V \in \mathcal{V}_{n,k}$, Σ is $k \times k$ diagonal and nonsingular, and $W \in \mathcal{O}(k)$. In this case, one obtains $Q = VW^T$ and $H = W\Sigma W^T$.

The derivative map $\dot{v} = (\mathcal{R}_P^{-1}|_Q)'(w)$ is obtained in a similar way as for the QR case, and we consider the curve $P + v(t) = Q(t)H(t)$ of a continuous reduced polar decomposition. Differentiating, we get

$$\dot{v} := \dot{v}(0) = wH + Q\dot{H}, \tag{10}$$

where we have set $Q = Q(0) \in \mathcal{V}_{n,k}$, $w = \dot{Q}(0) \in T_Q\mathcal{V}_{n,k}$, $H = H(0) \in \mathrm{Sym}^+(k)$, $\dot{H} = \dot{H}(0) \in \mathrm{Sym}(k)$. Now we multiply (10) by P^T from the left and consider the symmetric part of the resulting equation. This leads to the following Lyapunov equation for $\dot{H}$:

$$M\dot{H} + \dot{H}M^T + C = 0.$$

Here $M = P^TQ$ and $C = P^TwH + Hw^TP$. It is well known that this system has a unique solution $H \in \mathrm{Sym}(k)$ if and only if the eigenvalues of the matrix $M = P^TQ$ have nonzero real parts. So $\mathcal{R}_P'|_Q$ is invertible for every Q in some neighborhood of $P \in \mathcal{V}_{n,k}$. After solving for $\dot{H}$ we obtain $\dot{v}$ by substituting it into (10). The matrix $H \in \mathrm{Sym}^+(k)$ can be obtained by solving another Lyapunov equation, but as in the QR case it is for the applications we have in mind feasible to assume that this matrix is already known whenever needed in the calculation of $(\mathcal{R}_P^{-1}|_Q)'$.

See also Appendix A of [3] for `Matlab` implementations of these algorithms.

Complexity. Using the algorithm by Golub and Reinsch, it costs approximately $14nk^2 + \frac{22}{3}k^3$ flops to form the matrices V and Σ, and an additional $2nk^2$ flops to form U and then $k^3/2$ flops to form H. In conclusion, the dominating complexity terms for calculating $\mathcal{R}_P(v)$ is

$$16\,nk^2 + 7.83\,k^3$$

flops.

To obtain the corresponding map $(\mathcal{R}_P^{-1}|_Q)'$, the dominating cost consists of calculating two products of $n \times k$ with $n \times k$ matrices, then another two products of $n \times k$ with $k \times k$, each multiplication costing $2nk^2$ flops. The Lyapunov equation is solved only for a $k \times k$ matrix, and by using, e.g., the Bartels–Stewart algorithm the cost will be approximately $25k^3$ flops. For details about this and other algorithms for solving the Lyapunov equation, see [15, p. 367], [19], and the references therein. Summing up, we get approximately

$$8\,nk^2 + 25\,k^3$$

flops for calculating $(\mathcal{R}_P^{-1}|_Q)'(w)$.

5. Runge–Kutta methods based on retractions. We now consider in more detail how to solve ODEs on manifolds by using a retraction map. We will now assume that the problem is given by the user in the form (1) meaning that there is a procedure, say `F`, available which for any $(t, y) \in \mathcal{D} \subset \mathbb{R} \times \mathcal{M}$ returns the derivative $F(t, y) \in T_y\mathcal{M}$. In addition, the user must provide an initial value $y_0 \in \mathcal{M}$ and an initial stepsize h.

Suppose a Runge–Kutta (RK) method is given, with coefficients (a_{ij}), where $i, j = 1, \dots, s$, and (b_i), $i = 1, \dots, s$. As usual we denote $c_i = \sum_j a_{ij}$. The method is generally defined over one step as follows:

$$\left.\begin{aligned} u_i &= h \textstyle\sum_j a_{ij} k_j \\ k_i &= \big(\mathcal{R}'_{y_n}|_{u_i}\big)^{-1}\big(F(t_n + hc_i, \mathcal{R}_{y_n}(u_i))\big) \end{aligned}\right\}, \qquad i = 1, \dots, s,$$
$$y_{n+1} = \mathcal{R}_{y_n}\big(h \textstyle\sum_i b_i k_i\big).$$

In particular, the method is explicit if $a_{ij} = 0,\ j \geq i$. The following algorithm generalizes an explicit RK method to an arbitrary manifold in which a retraction $\mathcal{R}$ is defined.

Algorithm 5.1.

Input y_0, stepsize h, and RK parameters a_{ij}, b_i.
for $n = 0, 1, \dots$
  for $i = 1, \dots, s$
    $u_i := h \sum_j a_{ij} k_j$
    $Y := \mathcal{R}_{y_n}(u_i)$
    $m_i := F(t_n + hc_i, Y)$
    $k_i := \big(\mathcal{R}^{-1}_{y_n}|_Y\big)'\,(m_i)$
  end for
  $v := h \sum_i b_i k_i$
  $y_{n+1} := \mathcal{R}_{y_n}(v)$
  [Update stepsize h if desired]
end for

In the approach presented here, explicit methods are recommended because implicitness may cause an even higher increase in computational complexity than is the case for standard RK methods.

Remark. It is instructive to note that the algorithm described above is intrinsic in the sense that it works entirely within the realm of the manifold $\mathcal{M}$, and no reference is made to any Euclidean space. However, if this requirement is relaxed, suppose that $\mathcal{M} \subset V$, where V is a linear space, and that we consider instead

$$\dot{y} = \bar{F}(t, y), \quad \bar{F} : \mathbb{R} \times V \to V$$

such that $\bar{F}(t, \cdot)\big|_M = F(t, \cdot)$. Given a projector $\mathcal{P} : V \to \mathcal{M}$, we could consider the method defined simply as

$$\left.\begin{aligned} Z_i &= \mathcal{P}(y_n + h \textstyle\sum a_{ij} K_j) \\ K_i &= \bar{F}(t_n + c_i h, Z_i), \end{aligned}\right\}, \quad i = 1, \dots, s, \tag{11}$$
$$y_{n+1} = \mathcal{P}(y_n + h \textstyle\sum b_i K_i).$$

Since now $Y_i \in \mathcal{M}$, the extended vector field $\bar{F}$ is never evaluated outside $\mathcal{M}$, and the perturbation introduced by the projector $\mathcal{P}$ does not affect the order of the method. We have not pursued this approach any further, since in this particular paper we are primarily interested in methods that can be phrased independently of how the manifold is represented.

However, there is an interesting connection between this method and the method based on the retraction described in section 2.2. Suppose that we define a local projector $\mathcal{P}_{y_n}$ to be used in the step from t_n to t_{n+1} such that for points $m \in V$ near y_n we have

$$\mathcal{P}_{y_n}(m) = m + \bar{n}_{y_n}(m - y_n), \quad \bar{n}_{y_n} : V \to N_{y_n},$$

where $\bar{n}_{y_n}\big|_{T_{y_n}\mathcal{M}} = n_{y_n}$. Thus, the retraction is related to the projector as $\mathcal{R}_{y_n}(v) = \mathcal{P}_{y_n}(y_n + v)$ whenever $v \in T_{y_n}\mathcal{M}$. The map n_{y_n} is extended naturally to all of V so that $\bar{n}_{y_n}(v_t + v_n) = \bar{n}_{y_n}(v_t) + \bar{n}_{y_n}(v_n) = n_{y_n}(v_t) - v_n$ when $v_t \in T_{y_n}$ and $v_n \in N_{y_n}$, and we get

$$\mathcal{P}_{y_n}(m) = y_n + (I + n_{y_n}) \circ \mathbb{P}_{y_n}(m - y_n). \tag{12}$$

Now considering Algorithm 5.1 with this retraction, we get

$$\begin{aligned} Y_i &= \mathcal{P}_{y_n}\left(y_n + h\sum a_{ij}k_j\right), \\ k_i &= \mathbb{P}_{y_n}F(t_n + c_i h, Y_i), \\ y_{n+1} &= \mathcal{P}_{y_n}\left(y_n + h\sum b_i k_i\right). \end{aligned}$$

Now consider (11) with $\mathcal{P} = \mathcal{P}_{y_n}$ where we calculate by using (12)

$$\begin{aligned} Z_i &= \mathcal{P}_{y_n}\left(y_n + h\sum a_{ij}K_j\right) = y_n + (I + n_{y_n}) \circ \mathbb{P}_{y_n}\left(h\sum a_{ij}K_j\right) \\ &= \mathcal{P}_{y_n}\left(y_n + h\sum a_{ij}\mathbb{P}_{y_n}K_j\right) \end{aligned}$$

so we see that $Y_i = Z_i$ and $k_i = \mathbb{P}_{y_n}K_i$ and the two methods are equivalent.

6. Numerical experiments. All the numerical experiments are performed in `Matlab`. We compare the methods presented in this paper, RK methods based on retractions, RKRqr (based on the qr decomposition), and RKRp (based on the polar decomposition), with the following methods:

- Projection Runge–Kutta (PRK) method: the method that performs a projection based on the QR factorization at each time step of an explicit RK method (extrinsic).
- The method recently proposed in [2] (SPRK): a splitting method applied to the perturbed problem obtained by adding the term $-\tau y(I_k - y^T y)$ to the original vector field. The parameter τ is suitably chosen in order to make $\mathcal{V}_{n,k}$ an attracting manifold. (We took $h\tau = \frac{1}{2}$, which is indicated as the optimal choice in [2].) The resulting method is a projection method (extrinsic).
- Runge–Kutta Munthe-Kaas [21] (RKMK) methods: generalization of the classical RK methods based on the use of the Lie group actions $\Lambda(g, Y_0) = g \cdot Y_0$ and $g = \exp(\sigma)$ for $\sigma \in \mathfrak{so}(n)$ and $Y_0 \in \mathcal{V}_{n,k}$ (intrinsic).

The results obtained with the RKMK methods are not reported in the plots, but we will comment on their performance in the experiments in various points of this section.

All the methods are based on the classical explicit RK method of order 4.

The numerical experiments are divided in two parts. In the first section we will compare the intrinsic methods with the extrinsic ones.

In the second section we will illustrate the performance of the methods based on retractions when applied to the computation of the Lyapunov exponents of a ring of oscillators.

6.1. Intrinsic versus extrinsic. As a first experiment we consider the following initial value problem:

$$y' = F(y) = A(y)y + \lambda y(I_k - y^T y) \tag{13}$$

for $y(0) \in \mathcal{V}_{n,k}$, $n = 1000$, $k = 4$. Here $A(y)$ is a banded skew-symmetric $n \times n$ matrix whose nonzero entries are

$$A_{i,i+m} = -A_{i+m,i} = y_{i,m}, \quad 1 \le i \le n-m, \quad 1 \le m \le k. \tag{14}$$

The initial value, y_0, is obtained as the first factor of the reduced QR factorization of a random 1000×4 matrix, using the command `[y0,r]=qr(rand(n,k),0)` in `Matlab`.

Note that the term $y(I_k - y^T y)$ in (13) is zero for any $y \in \mathcal{V}_{n,k}$, thus the solution of (13) is independent of λ as long as $y_0 \in \mathcal{V}_{n,k}$. It might be desirable that the numerical approximation inherits this property. Although the extra term above may look artificial, it serves the purpose of illustrating that the way the vector field is extended affects the numerical solution obtained by a projection method.

In some cases it might be necessary or advantageous to rewrite (13) in a "strong skew-symmetric form"; i.e., $F(t,y) = H(t,y)y$ with $H(t,y) \in \mathfrak{so}(n)$ (for the RKMK methods for example). Note that in the numerical experiments we always assume the value of the vector field $F(t,y)$ to be given by a black box program that we are not allowed to modify. Since $A(y)$ is a banded matrix the cost of computing F, as given in (13), is effectively $\mathcal{O}(nk^2)$ flops.

Given $F = F(t,y)$ in the tangent space at y to the Stiefel manifold we then consider the following strong skew-symmetric formulation:

$$F = \big(\alpha(y)y^T - y\alpha^T(y)\big)\, y, \quad \alpha(y) = y\,\mathrm{tril}(y^T F) + (F - yy^T F). \tag{15}$$

Formulation (15) is then computed just assuming the knowledge of F and requires $\mathcal{O}(nk^2)$ flops.. Note that the formulation in the strong skew-symmetric form, as pointed out in [2], can be of crucial importance for extrinsic methods, as it makes $\mathcal{V}_{n,k}$ a strong invariant manifold. We apply the projection methods to this reformulation for problem (13) in the case $\lambda \neq 0$.

In Figure 1, for different values of the stepsize h, we plotted the norm of the difference of the numerical approximations produced by the methods after one time step for problem (13) with $\lambda = 0$ and $\lambda = 10$, respectively. The dependence on λ is quite evident for both the extrinsic methods, although the norm of the difference of the two numerical solutions decreases as the stepsize goes to zero. Note that the lines of the PRK and the SPRK lie on top of each other in the plot.

For the intrinsic methods the global error remains more or less the same for the two different values of λ, and as we can see from the figure the difference of the two approximate solutions is of the order of machine accuracy. Similar experiments gave analogous results for the RKMK methods.

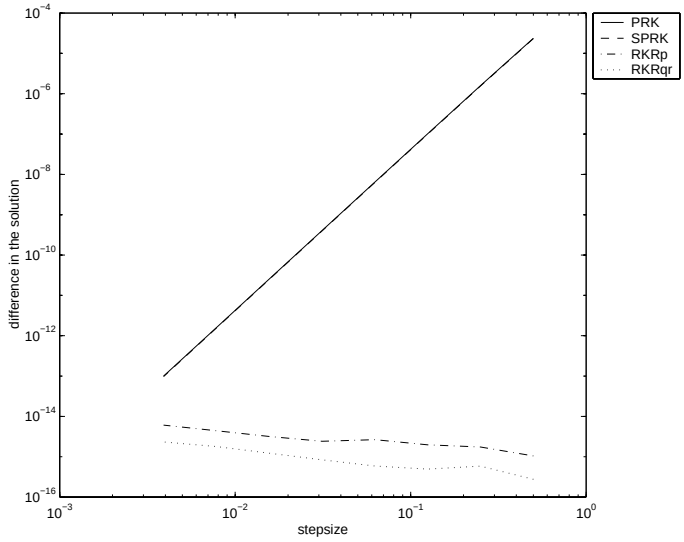


FIG. 1. *Difference in the solution for* $\lambda = 0$ *and* $\lambda = 10$.

In the next two figures we compare the cost of the four methods as applied to problem (13). We plot the global error at $T = 1$ on the y-axis against the number of flops on the x-axis in Figure 2 for $\lambda = 0$, and in Figure 3 for $\lambda = 10$.

As we can notice in the case $\lambda = 0$, the achieved global error per amount of flops for the two extrinsic methods is lower in norm than for the intrinsic methods.

However, for $\lambda = 10$ the RKRqr and the RKRp perform better than the two extrinsic methods.

The use of the RKMK methods in these experiments led to much more time consuming calculations. It seems that a naive implementation of these techniques causes unacceptable computational costs. Recently, new implementation techniques that can reduce the cost of the RKMK methods to a cost of $\mathcal{O}(nk^2)$ flops have been investigated [4].

6.2. Computing Lyapunov exponents. The Lyapunov exponents LEs of a continuous dynamical system $x' = F(x)$, $(x(t) \in \mathbb{R}^n)$ provide a qualitative measure of its complexity and can be defined as follows. Consider the linearization $A(t)$ of $x' = F(x)$ along a trajectory $x(t)$ and the solution U of the matrix problem

$$\dot{U} = A(t)U, \quad U(0) = U_0, \quad n \times n;$$

then the logarithms of the eigenvalues of the matrix

$$\Lambda = \lim_{t\to\infty} \left(U(t)^T U(t)\right)^{\frac{1}{2t}}$$

are the LEs for the given dynamical system. In [8] the authors describe a procedure for computing just k of the n LEs of a dynamical system. The strategy is based on solving a suitable initial value problem on $\mathcal{V}_{n,k}$ and computing a quadrature of the diagonal entries of a $k \times k$ matrix valued function. The initial value problem is defined as follows:

$$\dot{Q} = \left(A - QQ^T A + QSQ^T\right) Q,$$

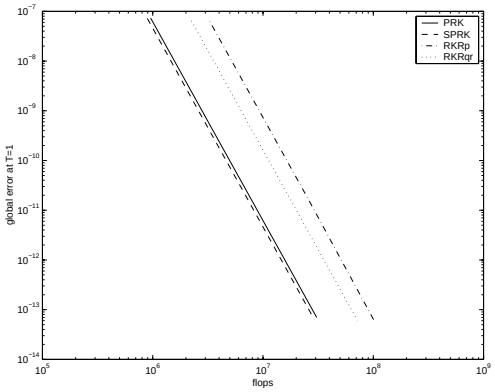


FIG. 2. *Global error versus the number of flops for* $\lambda = 0$.

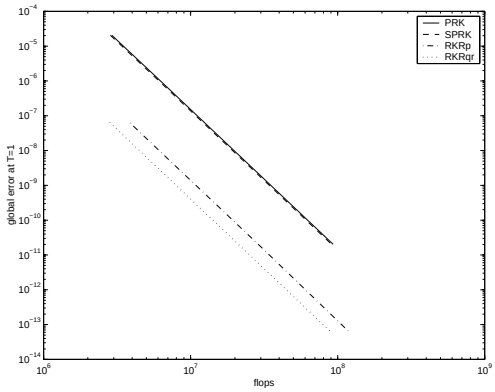


FIG. 3. *Global error versus the number of flops for* $\lambda = 10$.

with random initial value in $\mathcal{V}_{n,k}$ and

$$S_{k,j} = \begin{cases} (Q^T AQ)_{k,j}, & k > j, \\ 0, & k = j, \\ -(Q^T AQ)_{j,k}, & k < j, \end{cases} \qquad k, j = 1, \ldots, p.$$

It can be shown that the ith LE λ_i can be obtained as

$$\lambda_i = \lim_{t \to \infty} \frac{1}{t} \int_0^t B_{i,i}(s) ds, \quad i = 1, \ldots, k, \tag{16}$$

and

$$B = Q^T AQ - S.$$

In the numerical experiments we use the trapezoidal rule to approximate integral (16) and compute λ_i $(i = 1, \ldots, k)$, and we refer to the original paper [8] for further details on the method.

We first consider the following test problem previously proposed in [2]. Given a $n \times n$ diagonal matrix D we define the time dependent ODE

$$\dot{x} = A(t)x, \quad x \in \mathbb{R}^n, \quad A(t) = W(t)DW(t)^{\mathcal{T}} + J,$$

where J is the skew-symmetric part of a randomly generated matrix and $W(t) = \exp(tJ)$. The diagonal entries of D are the LEs of the constructed linear system. We took $D_{i,i} = C(3 - i)/n$ for $i = 1, \ldots, n$ and C a constant. Only the first three LEs are positive.

We considered the methods PRK, SPRK, RKRqr. In the tables we report the 2-norm of the error in the first k LEs and in the Stiefel matrix Q whose columns approximate the first k columns of $W(t)$, and in the last column we report the number of flops per step. The stepsize is the same for all methods and is $h = 0.1$.

The results of the first experiment are reported in Table 1, where we considered $n = 20$, $k = 6$, and $C = 1$, and the methods perform similarly.

In Table 2 are reported the results of the second experiment, where we have taken $n = 25$, $k = 4$, and $C = 27.5$.

TABLE 1
LE, first experiment: $n = 20$, $k = 6$, $C = 1$.

Methods	Error in LEs	Error in Q	Megaflops
PRK	0.0031	4.1361e-05	1.537
SPRK	0.0031	4.7595e-05	1.536
RKRqr	0.0030	2.6738e-05	1.566

TABLE 2
LE, second experiment: $n = 25$, $k = 4$, $C = 27.5$.

Methods	Error in LEs	Error in Q	Megaflops
PRK	0.0023	0.0013	2.992
SPRK	0.0023	0.0013	2.991
RKRqr	0.0014	0.0012	3.008

In the last numerical experiment we apply the RKRqr method to the computation of the LEs of the following ODEs system describing a ring of m Duffing oscillators:

$$\begin{aligned} &\ddot{y} + \alpha\left(y^2 - 1\right)\dot{y} + \omega^2 y = 0, \\ &\ddot{x}_i + d\dot{x}_i + \beta\left[V'(x_i - x_{i-1}) - V'(x_{i+1} - x_i)\right] = \sigma y \delta_{i,1}, \quad i = 1, \ldots, m. \end{aligned} \tag{17}$$

The ring is forced externally by $y(t)$. Here $V(x) = (\frac{x^2}{2}) + (\frac{x^4}{4})$, $\delta_{i,j} = 0$ for $i \neq j$ and $\delta_{i,i} = 1$, and we impose periodic boundary conditions ($x_0 = x_m$ and $x_{m+1} = x_1$). In the experiments $m = 15$ and $\alpha = 1$, $\omega = 1.6$, $\beta = 1$, $\sigma = 2$, and $d = 0.4$. This test problem has been considered in [11, 8, 2].

In this experiment we have considered a time interval $[0, 4000]$ and the stepsize $h = 0.01$. The trajectory is computed numerically using a midpoint rule with stepsize $h/8$.

Considering the error introduced by substituting the integral (16) with the quadrature, the results are in good agreement with those obtained in [11] and give the correct qualitative information about the considered dynamical system. (All LEs are negative.) The results are reported in Figures 4 and 5.

7. Conclusion. We have developed intrinsic integration methods of complexity $\mathcal{O}(n\,k^2)$ for solving ODEs on orthogonal Stiefel manifolds. We would like to emphasize that in order to see this complexity in actual computations it is important that the evaluation of the vector field F (as in $y' = F(t, y)$) does not involve cost of higher arithmetic complexity. It is, for instance, quite common to phrase problems on $\mathcal{V}_{n,k}$ in the form

$$\dot{y} = F(t, y) = A(t, y)y, \tag{18}$$

where $A(t, y) : \mathbb{R} \times \mathcal{V}_{n,k} \to \mathfrak{so}(n)$. If there is no sparsity in $A(t, y)$, it seems impossible to calculate $F(t, y)$ in less than $\mathcal{O}(n^2\,k)$ flops, and the gain in using the presented type of integration methods may not be significant. However, since each tangent space $T_p\mathcal{V}_{n,k}$ has only dimension $d_{n,k} = nk - k(k-1)/2$ it is always possible to find local parametrizations of the vector field F using only $d_{n,k}$ degrees of freedom.

There is no doubt that the methods presented here seem to have something in common with projection methods. Although we have seen no precise definition of a projection method, we still claim that there are important differences between the methods presented here and those normally referred to as projection methods. We have therefore chosen to focus on intrinsicness as the main feature to distinguish

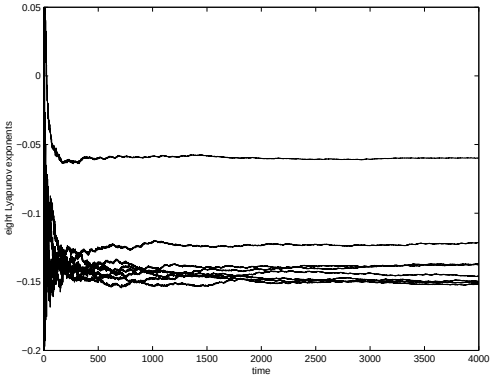


FIG. 4. *LEs for a ring of oscillators: RKRqr on the interval* $[0, 4000]$.

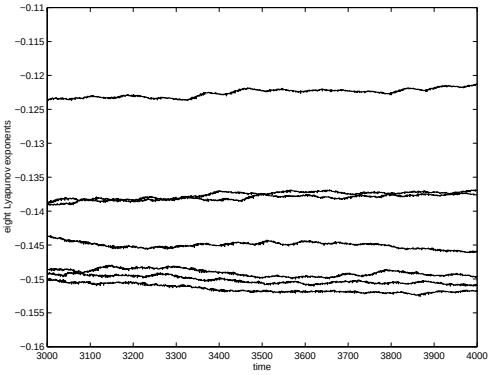


FIG. 5. *LEs for a ring of oscillators: RKRqr detail on the interval* $[3000, 4000]$.

the two types of methods. The term projection often indicates a map from higher to lower dimension, but in our case the retraction map is seen as a map from each tangent space $T_p\mathcal{V}_{n,k}$ onto a neighborhood of $p \in \mathcal{V}_{n,k}$, and in this sense it is a local diffeomorphism. Following the recent threads in the area of geometric integration, we find it useful to make a point out of constructing and analyzing schemes without appealing to the particular embedding of the manifold in a Euclidean space. This means, for instance, that we aim to avoid as much as possible working directly with algebraic constraints since they are closely linked to the realization of the manifold. Only in the explicit construction of specific retraction maps have we found it useful to apply properties of the embedding as a tool. Still, considering the retraction map as a black box, there are no traces of the embedding to be found in the method design. Furthermore, we have been reluctant to allow linear combinations of tangent vectors belonging to different tangent spaces. One may, of course, ask whether there are good reasons for requiring such properties in an integration method for ODEs on manifolds. We believe that there might be situations where intrinsic properties of the manifold itself are of importance for the behavior of the method, and it seems natural in such cases to apply integration methods that are not affected by how the manifold is embedded in a Euclidean space. Nevertheless, these new methods should be thought of as complementing rather than substituting the vast selection of DAE and projection methods available today. We also believe that more studies should be done towards understanding better the relation between established DAE methods and the approach presented here.

The main feature of the proposed methods is that they make no use of an extension of the vector field F to points not lying on the Stiefel manifold. It is certainly true that one can for all practical purposes find such an extension of the vector field F to all of $\mathbb{R}^{n\times k}$; for instance, the imbedding theorem of Nash ensures this for Riemannian manifolds. However, the extension is not unique, and we find it unnatural that the numerical approximation produced should depend on the particular extension which is chosen. Our numerical experiments confirm that different extensions can lead to completely different numerical approximations for projection methods, whereas the ones presented here yield identical results modulo rounding errors.

Acknowledgments. The authors would like to thank Luciano Lopez and Nicoletta Del Buono for inspiring discussion on this topic, and Nick Higham for providing

`Matlab` functions for the polar decomposition.

REFERENCES

[1] U. Ascher and S. Reich, *On some difficulties in integrating highly oscillatory Hamiltonian systems*, in Computational Molecular Dynamics, Lecture Notes in Comput. Sci. Eng. 4, Springer-Verlag, Berlin, 1999, pp. 281–296.

[2] T. J. Bridges and S. Reich, *Computing Lyapunov exponents on a Stiefel manifold*, Phys. D, 156 (2001), pp. 219–238.

[3] E. Celledoni and B. Owren, *A Class of Low Complexity Intrinsic Schemes for Orthogonal Integration*, Technical report Numerics 1/2001, The Norwegian University of Science and Technology, Trondheim, Norway, 2001.

[4] E. Celledoni and B. Owren, *On the Implementation of Lie Group Methods on the Stiefel Manifold*, Technical report Numerics 9/2001, The Norwegian University of Science and Technology, Trondheim, Norway, 2001, Numer. Algorithms, submitted.

[5] I. Chavel, *Riemannian Geometry: A Modern Introduction*, Cambridge Tracts in Math. 108, Cambridge University Press, Cambridge, UK, 1993.

[6] P. E. Crouch and R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci., 3 (1993), pp. 1–33.

[7] L. Dieci, R. D. Russell, and S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal., 34 (1997), pp. 402–423.

[8] L. Dieci and E. S. Van Vleck, *Computation of a few Lyapunov exponents for continuous and discrete dynamical systems*, Appl. Numer. Math., 17 (1995), pp. 275–291.

[9] L. Dieci and E. S. Van Vleck, *Computation of orthonormal factors for fundamental solution matrices*, Numer. Math., 83 (1999), pp. 599–620.

[10] F. Diele, L. Lopez, and R. Peluso, *The Cayley transform in the numerical solution of unitary differential systems*, Adv. Comput. Math., 8 (1998), pp. 317–334.

[11] U. Dressler, *Symmetry property of the Lyapunov spectra of a class of dissipative dynamical systems with viscous damping*, Phys. Rev. A (3), 38 (1988), pp. 2103–2109.

[12] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[13] E. Eich-Soellner and C. Führer, *Numerical Methods in Multibody Dynamics*, European Consort. Math. Indust., B. G. Teubner, Stuttgart, 1998.

[14] K. Geist, U. Parlitz, and W. Lauterborn, *Comparison of different methods for computing Lyapunov exponents*, Progr. Theoret. Phys., 83 (1990), pp. 875–893.

[15] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[16] E. Hairer, *Symmetric projection methods for differential equations on manifolds*, BIT, 40 (2000), pp. 726–734.

[17] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Ser. Comput. Math. 31, Springer-Verlag, Berlin, 2002.

[18] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.

[19] S. J. Hammarling, *Numerical solution of the stable non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.

[20] D. J. Higham, *Time-stepping and preserving orthonormality*, BIT, 37 (1997), pp. 24–36.

[21] H. Munthe-Kaas, *High order Runge–Kutta methods on manifolds*, Appl. Numer. Math., 29 (1999), pp. 115–127.

[22] B. Owren and A. Marthinsen, *Integration methods based on canonical coordinates of the second kind*, Numer. Math., 87 (2001), pp. 763–790.

[23] F. A. Potra and W. C. Rheinboldt, *On the numerical solution of Euler-Lagrange equations*, Mech. Structures Mach., 19 (1991), pp. 1–18.

[24] F. A. Potra and J. Yen, *Implicit numerical integration for Euler-Lagrange equations via tangent space parametrization*, Mech. Structures Mach., 19 (1991), pp. 77–98.

[25] M. Shub, *Some remarks on dynamical systems and numerical analysis*, in Dynamical Systems and Partial Differential Equations, L. Lara-Carrero and J. Lewowicz, eds., Equinoccio, Caracas, Venezuela, 1986, pp. 69–92.

[26] N. Trendafilov, *Projected Gradient Approach to Multivariate Data Analysis, A Review*, http://people.mech.kuleuven.ac.be/∼ntrendaf//pubs.htm.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2085–2104

NUMERICAL METHODS FOR p-HARMONIC FLOWS AND APPLICATIONS TO IMAGE PROCESSING*

LUMINITA A. VESE† AND STANLEY J. OSHER†

Abstract. We propose in this paper an alternative approach for computing p-harmonic maps and flows: instead of solving a constrained minimization problem on S^{N-1}, we solve an unconstrained minimization problem on the entire space of functions. This is possible, using the projection on the sphere of any arbitrary function. Then we show how this formulation can be used in practice, for problems with both isotropic and anisotropic diffusion, with applications to image processing, using a new finite difference scheme.

Key words. energy minimization, p-harmonic maps, p-harmonic flows, directional diffusion, heat flow, total variation minimization, partial differential equations, finite differences, denoising

AMS subject classifications. 35, 49, 65

PII. S0036142901396715

1. Introduction.

This paper is concerned with the minimization of constrained functionals, and in particular with p-harmonic maps. This problem has applications to liquid crystals, as well as to directional diffusion and chromaticity denoising.

Let $\Omega \subset \mathbb{R}^M$ be an open and bounded domain, and let S^{N-1} be the unit sphere in $\mathbb{R}^N$, for $M \geq 1$ and $N \geq 2$.

We first recall the following notations and terminology. The Euclidean norm of a vector y will be denoted by $|\cdot|$. The vector-valued function $U : \Omega \to \mathbb{R}^N$ belongs to S^{N-1} if and only if $|U(x)| = 1$, a.e. (for almost every) $x \in \Omega$.

The component gradient ∇U_i and its Euclidean norm are, respectively, defined by

$$\nabla U_i = \left(\frac{\partial U_i}{\partial x_1}, \frac{\partial U_i}{\partial x_2}, \ldots, \frac{\partial U_i}{\partial x_M}\right), \quad |\nabla U_i| = \sqrt{\left(\frac{\partial U_i}{\partial x_1}\right)^2 + \left(\frac{\partial U_i}{\partial x_2}\right)^2 + \cdots + \left(\frac{\partial U_i}{\partial x_M}\right)^2},$$

and the gradient matrix and its norm of the vector-valued function U are, respectively, defined by

$$\nabla U = \begin{pmatrix} \nabla U_1 \\ \vdots \\ \nabla U_N \end{pmatrix} = \begin{pmatrix} \frac{\partial U_1}{\partial x_1} & \cdots & \frac{\partial U_1}{\partial x_M} \\ \vdots & & \vdots \\ \frac{\partial U_N}{\partial x_1} & \cdots & \frac{\partial U_N}{\partial x_M} \end{pmatrix}, \quad |\nabla U| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{\partial U_i}{\partial x_j}\right)^2}.$$

For $U : \Omega \to S^{N-1}$ and $p \geq 1$, we consider the p-energy

$$E_p(U) = \int_\Omega |\nabla U|^p dx, \tag{1.1}$$

*Received by the editors October 17, 2001; accepted for publication (in revised form) May 13, 2002; published electronically December 13, 2002.
http://www.siam.org/journals/sinum/40-6/39671.html

†Department of Mathematics, University of California at Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1555 (lvese@math.ucla.edu, sjo@math.ucla.edu). The research of the first author was supported by grants NSF ITR-0113439, NSF DMS-9973341, and ONR N00014-02-1-0015. The research of the second author was supported by grants NSF DMS-0074735 and ONR N00014-97-1-0027.

which is finite if U belongs to the Sobolev class

$$W^{1,p}(\Omega, S^{N-1}) = \{U \in W^{1,p}(\Omega, \mathbb{R}^N),\ |U| = 1 \text{ a.e.}\}.$$

Minimizing E_p over $U : \Omega \to S^{N-1}$, with associated boundary conditions on $\partial\Omega$, is a constrained minimization problem. Mappings which are stationary for E_p are called p-harmonic maps.

The associated boundary conditions can be, for example, the following: $U|_{\partial\Omega}$ equals a given map in $S^{N-1}(\partial\Omega)$, or the Neumann boundary conditions $\frac{\partial U}{\partial \vec{n}}|_{\partial\Omega} = 0$, where $\vec{n}$ denotes the exterior unit normal to $\partial\Omega$.

Many authors have studied harmonic maps between manifolds (existence, uniqueness or nonuniqueness, regularity; essentially most of them worked on the case $p = 2$): Bethuel, Brezis, and Coron [5]; Bethuel, Brezis, and Helein [6], [7]; Schoen and Uhlenbeck [29], [30], [31]; Struwe [34], [35], [36]; Courilleau and Demengel [15]; Coron and Gulliver [14]; Brezis, Coron, and Lieb [8]; and others. There are fewer results for the case $p = 1$ (for example by Giaquinta, Modica, and Soucek [19]).

We would also like to mention the following important contributions on harmonic maps for liquid crystals, both in theory and practice: Hardt, Kinderlehrer, and Luskin [20]; Lin and Luskin [24]; Cohen, Lin, and Luskin [12]; Cohen et al. [13]. The work on computational aspects of harmonic maps [13] was proposed before the analysis results on this problem. This paper deals with alternative formulations and numerical methods for computing harmonic maps.

There are difficulties finding numerically the minimizers or the p-harmonic maps, due to nonconvexity (the constraint $|U(x)| = 1$ a.e. is not convex), nonregularity, and nonuniqueness of minimizers.

There are several classical approaches used to solve the minimization problem (1.1).

A first approach is to solve the Euler–Lagrange equations associated with the minimization problem. These consist of a set of coupled PDEs:

$$-\text{div}(|\nabla U|^{p-2}\nabla U) = U|\nabla U|^p. \tag{1.2}$$

The above system of equations holds if and only if $U \in S^{N-1}$. However, in practice, the numerical solution does not necessarily satisfy the constraint $|U| = 1$ everywhere. To correct the numerical error, several authors [18], [38], [39] replace the solution U_*^n obtained at each iteration n by $U^n = \frac{U_*^n}{|U_*^n|}$, but then the question is whether one still decreases the energy. In this framework, we also refer the reader to [13]. It is known [1] that the energy decrease is guaranteed after this renormalization if $|U_*^n| \geq 1$, but the behavior of the energy is not known if $|U_*^n| < 1$. Also, if we would like to extend this numerical procedure involving the projection at each step to other manifolds, then the energy decrease is guaranteed only when the manifold is the boundary of a convex set, and again if, in addition, U_*^n does not belong to the interior of that convex domain.

This problem has been solved in [1] for the S^2 case and in three dimensions, where an interesting convergent algorithm is proposed, but it still involves a renormalization step at each iteration (ensuring now that the energy decreases after the renormalization step). Numerical methods for p-harmonic flows are also proposed in [18] and [13], again based on the renormalization procedure at each step.

The second classical approach is given by the Ginzburg–Landau functionals [6], [7]. Here, the problem is solved by approximation to eliminate the constraint. The minimization of the energy E_p from (1.1) under the constraint $|U(x)| = 1$ a.e. is approxi-

mated by the unconstrained minimization of the following energies, as $\varepsilon \to 0$:

$$E_\varepsilon(U) = \int_\Omega |\nabla U|^p dx + \frac{1}{\varepsilon}\int_\Omega (1-|U|^2)^2 dx. \tag{1.3}$$

In this paper, we introduce a different approach to solving minimization problems on S^{N-1}. We solve an unconstrained minimization problem on the entire space of functions and not only on S^{N-1}. The method uses the projection of an arbitrary function V to the sphere S^{N-1}. We will present our alternative approach for the case of S^{N-1}. Then we discuss how this approach can be extended to more general manifolds, and in particular to manifolds defined implicitly, via a level set function. By proposing numerical schemes in the S^1 and S^2 cases, we also show how our formulations can be used in practice, and in particular for applications to directional diffusion and color image denoising.

In the framework of image processing and directional diffusion, related works are [27], [33], [9], [38], [22], [39], [42], and [32], [23]. We also refer the reader to [17] for manifold constrained variational problems. In the framework of energy minimization with values in S^2, we refer the reader to [16], where the algorithm from [1] is applied in the presence of a data term.

Our main idea is as follows. For $U : \Omega \to S^{N-1}$, with $\Omega \subset \mathbb{R}^M$, consider $V : \Omega \to \mathbb{R}^N \setminus \vec{0}$ such that

$$U = \frac{V}{|V|}.$$

We minimize without constraint the corresponding energy with respect to V:

$$\inf_V \left\{ F(V) = \int_\Omega \left| \nabla\left(\frac{V}{|V|}\right)\right|^p dx \right\}, \tag{1.4}$$

and then we recover U, a minimizer of (1.1), projecting back on S^{N-1}, by $U = \frac{V}{|V|}$, where V is a minimizer of (1.4).

We would like to mention that the idea of solving constrained minimization problems for harmonic maps by associating unconstrained minimization problems has been used as a theoretical tool by Chen and Lin [10] and Struwe [37]. They find a smooth energy-minimizing harmonic map U as a weak limit of minimizers U_L to an unconstrained variational problem for $L \to \infty$.

The outline of the paper is as follows. In section 2, we consider the S^1 case: we derive the Euler–Lagrange equations associated with the unconstrained minimization problem, and in subsection 2.1 we propose a numerical scheme for this case. Similarly, section 3 and subsection 3.1 are devoted to the S^2 case. In section 4 we validate the proposed models and numerical schemes on several experimental results: in subsection 4.1, we consider the case with prescribed boundary conditions, and we make a comparison with the classical formulation (1.2) with the renormalization step at each iteration (we will see that, by the proposed approach, the numerical accuracy is improved in a test case where we know the exact solution); in subsection 4.2, we consider the case of directional diffusion, with Neumann boundary conditions, and applications to chromaticity denoising for color images. Finally, in section 5 we conclude with a discussion for more general manifolds.

2. The S^1 case. To develop our main idea, let us first consider the particular case $N = 2$ of S^1. Then, for $U : \Omega \to S^1$, consider $V = (u, v) : \Omega \to \mathbb{R}^2$ such that $U = \frac{V}{|V|}$.

In order to obtain in an elegant way the Euler–Lagrange equations associated with the minimization problem (1.4), we consider the orientation formulation (which is not always equivalent with the directional formulation). Let $U = (\cos\theta, \sin\theta)$, and let $V = (r\cos\theta, r\sin\theta)$. Then $u^2 + v^2 = r^2$, and we have

$$\left|\nabla\left(\frac{V}{|V|}\right)\right|^2 = |\nabla\theta|^2.$$

For $p = 2$ (the heat flow for harmonic maps), solving

$$\inf_\theta \int_\Omega |\nabla\theta|^2 dx,$$

and parameterizing the descent direction by an artificial time t, we obtain (denoting $u_t = \frac{\partial u}{\partial t}$, $v_t = \frac{\partial v}{\partial t}$)

$$\theta_t = \triangle\theta, \quad r_t = 0.$$

Using

$$\theta = \tan^{-1}\left(\frac{v}{u}\right), \quad \nabla\theta = \frac{u\nabla v - v\nabla u}{u^2 + v^2},$$

we first deduce that

$$\frac{uv_t - vu_t}{u^2 + v^2} = \operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2 + v^2}\right).$$

Now, using $uu_t + vv_t = 0$ (from $r_t = 0$), we obtain the associated Euler–Lagrange equations for $p = 2$:

$$u_t = -v\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2 + v^2}\right), \quad v_t = +u\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2 + v^2}\right). \tag{2.1}$$

For $p = 1$ (the total variation minimization of Rudin, Osher, and Fatemi [28]), on solving

$$\inf_\theta \int_\Omega |\nabla\theta| dx,$$

we obtain

$$\theta_t = \operatorname{div}\left(\frac{\nabla\theta}{|\nabla\theta|}\right), \quad r_t = 0.$$

Then, in a similar way, the associated Euler–Lagrange equations for $p = 1$ are

$$u_t = -v\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{|u\nabla v - v\nabla u|}\right), \quad v_t = +u\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{|u\nabla v - v\nabla u|}\right). \tag{2.2}$$

In the general case, i.e., for any $p \geq 1$, the corresponding linear system in u_t and v_t is

$$\begin{aligned} & uu_t + vv_t = 0, \\ & \frac{uv_t - vu_t}{u^2+v^2} = \operatorname{div}\left[\left(\frac{|u\nabla v - v\nabla u|}{u^2+v^2}\right)^{p-2}\left(\frac{u\nabla v - v\nabla u}{u^2+v^2}\right)\right]. \end{aligned}$$

Solving this linear system in the unknowns u_t and v_t yields similar equations in u and v, like those for the cases $p = 2$ and $p = 1$ from (2.1) and (2.2), respectively.

We will associate with the problems (2.1) and (2.2) initial conditions in the following form: $u(0, x) = u_0(x)$ and $v(0, x) = v_0(x)$ in Ω; at the boundary, we can prescribe either Dirichlet boundary conditions $V(t, x)/|V(t, x)| = F(x)$, with $F : \partial\Omega \to S^1$ given, for $t \geq 0$ and $x \in \partial\Omega$; or Neumann boundary conditions $\frac{\partial u}{\partial \vec{n}} = 0$ and $\frac{\partial v}{\partial \vec{n}} = 0$ on $\partial\Omega$.

We could add data terms in the energy, as in [9] or [16].

Remark. With these formulations, with both $p = 1$ and $p = 2$ (and, in fact, for any $p \geq 1$), we always have, for any fixed $x \in \Omega$, $u(t, x)u_t(t, x) + v(t, x)v_t(t, x) = 0$, or $u^2(t, x) + v^2(t, x) =$ constant in time for fixed x.

Remark. Note that we have used an artificial time, even if we compute a stationary solution of the problem. This is a common technique, and this artificial time represents a parameterization of the descent direction. It can be shown, in general, that the energy is decreasing in time, under such a time-dependent flow, for both Dirichlet or Neumann boundary conditions (as explained in detail in the appendix).

2.1. The numerical algorithm for the S^1 case. To discretize the systems (2.1) and (2.2), we use finite differences. Assume for simplicity that $U : [0, 1]^M \to S^1$, let h be the space step, and let $\triangle t$ be the time step. We denote by u^n and v^n the approximations of $u(n\triangle t, x)$ and of $v(n\triangle t, x)$, respectively, where x is a grid point. (To simplify the notation, we will not explicitly indicate the discrete point $x_{i,j}$ where the approximation is considered; for instance, if $M = 2$, u^n means $u^n_{i,j}$, etc.; similarly, any expression of the form $(E)^n$ denotes an approximation of the quantity E at $(n\triangle t, x)$, at the same discrete point x; this notational convention will allow us to consider any dimension $M \geq 1$.)

We use the following semi-implicit scheme for (2.1) ($p = 2$):

$$\begin{aligned} \frac{u^{n+1} - u^n}{\triangle t} &= -\frac{v^{n+1} + v^n}{2}\left[\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2+v^2}\right)\right]^n, \\ \frac{v^{n+1} - v^n}{\triangle t} &= +\frac{u^{n+1} + u^n}{2}\left[\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2+v^2}\right)\right]^n, \end{aligned}$$

and similarly for (2.2) ($p = 1$).

Denoting by $(Div)^n$ an approximation of the expression $\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2+v^2}\right)$ evaluated at $(n\triangle t, ih, jh, \ldots)$, and solving the previous algebraic system in u^{n+1} and v^{n+1}, we obtain, for both $p = 1$ and $p = 2$,

$$\begin{aligned} u^{n+1} &= \frac{u^n - \left(2v^n + u^n\frac{\triangle t(Div)^n}{2}\right)\frac{\triangle t(Div)^n}{2}}{1 + \left(\frac{\triangle t(Div)^n}{2}\right)^2}, \\ v^{n+1} &= \frac{v^n + \left(2u^n - v^n\frac{\triangle t(Div)^n}{2}\right)\frac{\triangle t(Div)^n}{2}}{1 + \left(\frac{\triangle t(Div)^n}{2}\right)^2}. \end{aligned}$$

To discretize the expression $\operatorname{div}\left(\frac{u\nabla v - v\nabla u}{u^2+v^2}\right)$, we use the finite difference scheme proposed in [28] for $\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)$ and which has also been used in [2] for a more general case.

Remark. As in the continuous case, it is easy to verify that the numerical solution exactly satisfies

$$(u^{n+1})^2 + (v^{n+1})^2 = (u^n)^2 + (v^n)^2$$

at any grid point x. This proves that the scheme produces bounded solutions independent of the relation between $\triangle t$ and h.

Remark. Note that we do not need to apply a renormalization step at every iteration. Only in the end of the algorithm we let $U = \frac{V}{|V|}$, with $V = (u, v)$. Note also that if the initial data $V_0 = (u_0, v_0)$ already satisfies $|V_0| = 1$ everywhere, then, due to the previous remark, this equality will be preserved in time, and therefore, in the end, the numerical solution U will be directly given by V. (In other words, in this case, there is no need to renormalize V at the steady state; we will simply have $U = V$.)

Remark. Although the solutions remain bounded regardless of the magnitude of $\triangle t$, the numerical domain of dependence of u^{n+1}, v^{n+1} is such that convergence for $p = 2$ is possible only if $\triangle t \leq Ch^2$. This follows from the fact that θ satisfies the heat equation. We verified this by numerical experiments, and found that the quantity θ is noisy if $\triangle t$ is too large, although the solution is bounded. (See Figure 7 for a comparison of results obtained for several values of $\triangle t$.) Convergence for $p = 1$ requires a more restrictive constraint on $\triangle t$, typical of that for total variation minimization [28] in θ.

Remark. Note that additional penalty terms obtained by imposing constraints on V or on $\frac{V}{|V|}$ could be added to the energy or to the Euler–Lagrange equations without any difficulty.

3. The S^2 case. We will follow the same idea as in the previous case, in order to derive the Euler–Lagrange equations associated with the unconstrained minimization problem (1.4), for any $M \geq 1$ and $N = 3$.

Using spherical coordinates, we let

$$U = (\cos\theta_1 \cos\theta_2, \cos\theta_1 \sin\theta_2, \sin\theta_1) \in S^2$$

and

$$V = (r\cos\theta_1 \cos\theta_2, r\cos\theta_1 \sin\theta_2, r\sin\theta_1) = (u, v, w).$$

We then have $r^2 = u^2 + v^2 + w^2$,

$$\theta_1 = \tan^{-1}\left(\frac{w}{\sqrt{u^2+v^2}}\right), \quad \theta_2 = \tan^{-1}\left(\frac{v}{u}\right),$$

and it can be shown that

$$|\nabla U|^2 = |\nabla\theta_1|^2 + \cos^2\theta_1 |\nabla\theta_2|^2.$$

Let us consider first the case $p = 2$. From

$$\inf_{\theta_1,\theta_2} \int_\Omega |\nabla\theta_1|^2 + \cos^2\theta_1 |\nabla\theta_2|^2 dx,$$

we obtain (parameterizing the descent directions by an artificial time t)

$$\theta_{1,t} = \triangle\theta_1 + \sin\theta_1 \cos\theta_1 |\nabla\theta_2|^2, \tag{3.1}$$

$$\theta_{2,t} = \operatorname{div}(\cos^2\theta_1 \nabla\theta_2). \tag{3.2}$$

Let us denote by E_1 and E_2, respectively, the expressions on the right-hand sides of (3.1) and (3.2), i.e.,

$$\theta_{1,t} = E_1, \quad \theta_{2,t} = E_2. \tag{3.3}$$

Again, from $r_t = 0$, we deduce that

$$uu_t + vv_t + ww_t = 0. \tag{3.4}$$

Computing and using

$$\nabla\theta_1 = \frac{(u^2+v^2)(\nabla w) - uw(\nabla u) - vw(\nabla v)}{(u^2+v^2+w^2)\sqrt{u^2+v^2}}, \tag{3.5}$$

$$\nabla\theta_2 = \frac{u(\nabla v) - v(\nabla u)}{u^2+v^2}, \tag{3.6}$$

we can then express E_1 and E_2 as functions of (u, v, w) by

$$E_1 = \triangle\theta_1 + \frac{w\sqrt{u^2+v^2}}{u^2+v^2+w^2}|\nabla\theta_2|^2, \quad E_2 = \operatorname{div}\left(\frac{u(\nabla v) - v(\nabla u)}{u^2+v^2+w^2}\right).$$

On the other hand, we have

$$\theta_{1,t} = \frac{(u^2+v^2)w_t - uwu_t - vwv_t}{(u^2+v^2+w^2)\sqrt{u^2+v^2}}, \quad \theta_{2,t} = \frac{uv_t - vu_t}{u^2+v^2}.$$

We now consider the system formed by (3.3), (3.4) in the unknowns u_t, v_t, and w_t:

$$uu_t + vv_t + ww_t = 0, \quad \frac{(u^2+v^2)w_t - uwu_t - vwv_t}{(u^2+v^2+w^2)\sqrt{u^2+v^2}} = E_1, \quad \frac{uv_t - vu_t}{u^2+v^2} = E_2.$$

Solving this linear system in the unknowns u_t, v_t, and w_t, we deduce the associated Euler–Lagrange equations

$$u_t = -\frac{uw}{\sqrt{u^2+v^2}}E_1 - vE_2, \tag{3.7}$$

$$v_t = -\frac{vw}{\sqrt{u^2+v^2}}E_1 + uE_2, \tag{3.8}$$

$$w_t = \sqrt{u^2+v^2}E_1. \tag{3.9}$$

For the case $p = 1$ of the total variation minimization of Rudin, Osher, and Fatemi [28], we consider first the problem in $\theta = (\theta_1, \theta_2) \in [-\frac{\pi}{2}, \frac{\pi}{2}]^2$:

$$\inf_{\theta_1,\theta_2} \int_\Omega \sqrt{|\nabla\theta_1|^2 + \cos^2\theta_1 |\nabla\theta_2|^2}\, dx,$$

which yields the equations

$$\theta_{1,t} = \operatorname{div}\left(\frac{\nabla\theta_1}{\sqrt{|\nabla\theta_1|^2+\cos^2\theta_1|\nabla\theta_2|^2}}\right) + \frac{\sin\theta_1\cos\theta_1|\nabla\theta_2|^2}{\sqrt{|\nabla\theta_1|^2+\cos^2\theta_1|\nabla\theta_2|^2}},$$
$$\theta_{2,t} = \operatorname{div}\left(\cos^2\theta_1\frac{\nabla\theta_2}{\sqrt{|\nabla\theta_1|^2+\cos^2\theta_1|\nabla\theta_2|^2}}\right).$$

Denoting again by E_1 and E_2 the expressions on the right-hand sides of the above equations (corresponding now to the case $p=1$), these can be expressed as functions of (u,v,w) using (3.5) and (3.6). The Euler–Lagrange equations for the case $p=1$, in (u,v,w), are therefore as in (3.7)–(3.9) but with the corresponding differential operators E_1 and E_2 for $p=1$.

3.1. The numerical algorithm for the S^2 case. The expressions E_1 and E_2 are discretized following [28] and [2] for both $p=1$ and $p=2$. (We will still denote their discretizations at a given point by E_1 and E_2.)

Let us denote by u^n, v^n, w^n the discrete solutions at a discrete point in two or three dimensions (but without writing $u^n_{i,j}$ or $u^n_{i,j,k}$, for simplicity). We discretize the system (3.7)–(3.9) using the following implicit scheme:

$$u^{n+1} = u^n - \frac{\triangle t}{\sqrt{(u^n)^2+(v^n)^2}}u^n\left(\frac{w^{n+1}+w^n}{2}\right)E_1 - \left(\frac{v^{n+1}+v^n}{2}\right)E_2\triangle t,$$
$$v^{n+1} = v^n - \frac{\triangle t}{\sqrt{(u^n)^2+(v^n)^2}}v^n\left(\frac{w^{n+1}+w^n}{2}\right)E_1 + \left(\frac{u^{n+1}+u^n}{2}\right)E_2\triangle t,$$
$$w^{n+1} = w^n + \triangle t\sqrt{(u^n)^2+(v^n)^2}E_1.$$

We will use the notations

$$A = \frac{E_1\triangle t}{2\sqrt{(u^n)^2+(v^n)^2}}, \quad B = \frac{E_2\triangle t}{2}, \quad C = \triangle t\sqrt{(u^n)^2+(v^n)^2}E_1.$$

The linear system in $u^{n+1}, v^{n+1}, w^{n+1}$ is nonsingular and has the unique solution

$$u^{n+1} = \frac{R_1 - BR_2}{1+B^2}, \quad v^{n+1} = \frac{R_2+BR_1}{1+B^2}, \quad w^{n+1} = w^n + C,$$

where $R_1 = u^n - Au^n(2w^n+C) - v^nB$ and $R_2 = v^n - Av^n(2w^n+C) + u^nB$.

Remark. The numerical scheme will exactly satisfy the relation

$$(u^{n+1})^2 + (v^{n+1})^2 + (w^{n+1})^2 = (u^n)^2 + (v^n)^2 + (w^n)^2$$

at each grid point, if in the above discretizations the expression $\sqrt{(u^n)^2+(v^n)^2}$ is replaced by $\sqrt{u^n\left(\frac{u^n+u^{n+1}}{2}\right)+v^n\left(\frac{v^n+v^{n+1}}{2}\right)}$, but this yields a nonlinear system in the unknowns u^{n+1}, v^{n+1}, and w^{n+1}, which could be solved by a fixed-point iteration.

4. Numerical experiments. In this section we present numerical experiments in the cases $M=2$, $N=2$, and $M=2,3$ and $N=3$. We will consider the cases with Dirichlet boundary conditions (subsection 4.1) and Neumann boundary conditions (subsection 4.2).

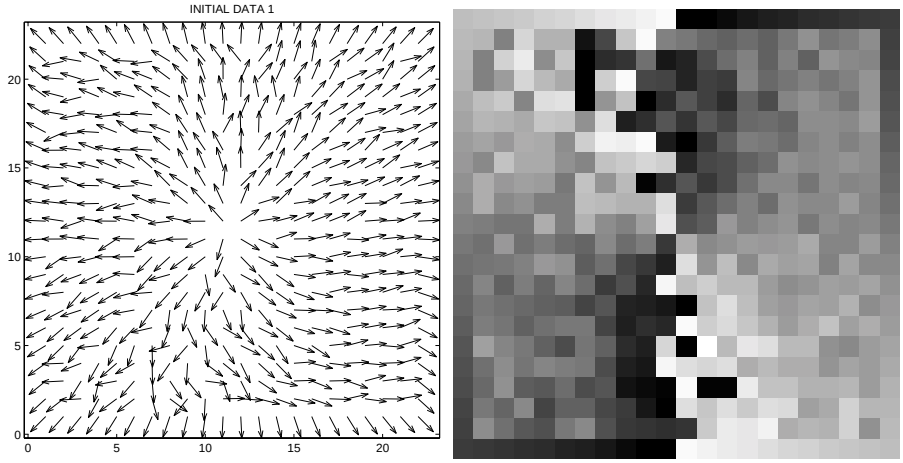


FIG. 1. *Left: initial condition for the Dirichlet problem, as a perturbation of* $\frac{x-x_0}{|x-x_0|}$ *in* $(0,1)^2$*, given by* (4.1)–(4.2)*, and agreeing with* $\frac{x-x_0}{|x-x_0|}$ *at the boundary, where* $x_0 = (0.5, 0.5)$*. Right: corresponding initial angle* $\theta = \tan^{-1}\left(\frac{v^0}{u^0}\right)$.

4.1. Numerical results for prescribed boundary conditions. In the S^1 case, we first consider the Dirichlet problem, with the boundary condition $U(x) = \frac{x-x_0}{|x-x_0|}$ on $\partial\Omega$, with $x_0 = (0.5, 0.5)$, where $\Omega = (0,1)^2$. In this case, it is known that the map $x \mapsto \frac{x-x_0}{|x-x_0|}$ is an exact solution and minimizer in $\overline{\Omega}$. We will show that the numerical solution has the correct behavior, approximating very well the exact solution.

Following [13], an initial condition $V^0 = (u^0, v^0)$ inside Ω can be a perturbation of $\frac{x-x_0}{|x-x_0|}$ (shown in Figure 1, after normalization):

$$u^0(x_1, x_2) = \frac{x_1 - .5}{|x - x_0|} + .6(1 + x_1^2 - x_2^2) - .8\eta, \tag{4.1}$$

$$v^0(x_1, x_2) = \frac{x_2 - .5}{|x - x_0|} + .6(x_1 - 2x_2) + .8\eta, \tag{4.2}$$

for all $(x_1, x_2) \in \Omega$, where η is random noise.

We will also consider another initial condition in this case, defined using the distance function to the boundary as follows: for $(x, y) \in \Omega$, find $(x_b, y_b) \in \partial\Omega$ as the closest point to the boundary $\partial\Omega$ from (x, y). Then let $(u^0(x, y), v^0(x, y)) = U(x_b, y_b)$, where U defines the boundary conditions on $\partial\Omega$. (This second initial condition is shown in Figure 2.)

We now consider the case $p = 2$ for these two initial conditions. For the initial data 1, we also compare the results (the error and the energy decrease) with the classical harmonic map formulation with numerical renormalization at each time step by solving the semidiscrete problem (using central difference approximations for the space derivatives, and with the same prescribed boundary conditions and the same time and space steps):

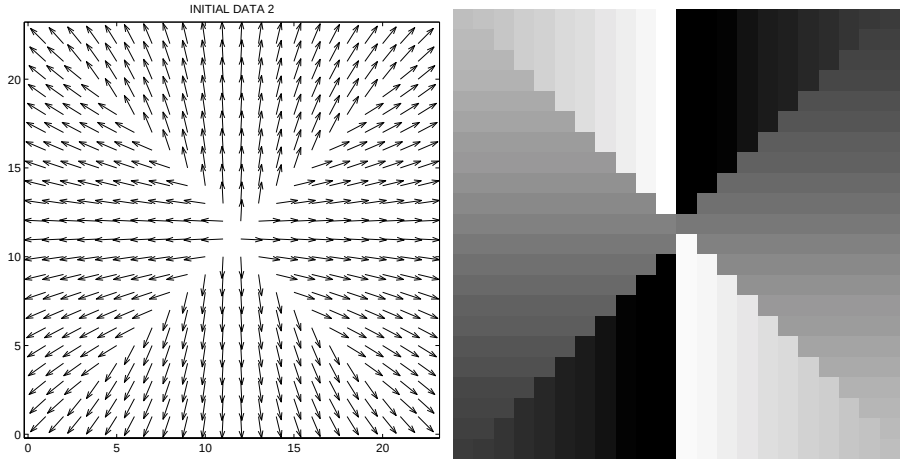


FIG. 2. *Left: another initial condition for the Dirichlet problem, constructed using the closest point to the boundary, and agreeing with* $\frac{x-x_0}{|x-x_0|}$ *at the boundary, where* $x_0 = (0.5, 0.5)$. *Right: corresponding initial angle* $\theta = \tan^{-1}\left(\frac{v^0}{u^0}\right)$.

$$\frac{u_*^{n+1} - u^n}{\triangle t} = \triangle u^n + u^n\Big[(u_x^n)^2 + (u_y^n)^2 + (v_x^n)^2 + (v_y^n)^2\Big],$$
$$\frac{v_*^{n+1} - v^n}{\triangle t} = \triangle v^n + v^n\Big[(u_x^n)^2 + (u_y^n)^2 + (v_x^n)^2 + (v_y^n)^2\Big],$$
$$(u^{n+1}, v^{n+1}) = \frac{(u_*^{n+1}, v_*^{n+1})}{|(u_*^{n+1}, v_*^{n+1})|}.$$

We show the energy decrease and the error versus iterations for the results obtained with the classical harmonic maps applied to the initial data 1, and with the proposed model applied to both initial data 1 and 2 (see Figure 3). Using the proposed model, the error is much smaller. Also, note that the initial data 2 produces a result very fast. For both initial data 1 and 2, by our proposed model, the numerical solution $U(x) = \frac{V(x)}{|V(x)|}$ at the steady state approximates very well the exact solution $\frac{x-x_0}{|x-x_0|}$ in $\Omega = [0,1]^2$, and it is better than using the classical harmonic map scheme with the renormalization at each step.

The results obtained with the proposed model for $p = 2$, for both data, are shown in Figure 4, together with the angle $\theta = \tan^{-1}\left(\frac{v}{u}\right)$.

Corresponding results obtained with the proposed model for $p = 1$ are shown in Figures 5 and 6.

In Figure 7 we show the angle $\theta = \tan^{-1}\left(\frac{v}{u}\right)$, obtained with the initial data 1, for $p = 2$, at the steady state using the proposed model, for different decreasing values of $\triangle t$. This test proves again that if $\triangle t$ is too large, then θ is noisy, but the numerical solution (u^n, v^n) remains bounded. Similar results can be obtained for $p = 1$, with a slightly stronger condition on $\triangle t$, to guarantee the stability of the numerical scheme.

We show next a numerical result for maps with values in S^2 in the three-dimensional case. Following [1], we perform a test, which shows again that, for the Dirichlet boundary conditions, the numerical solution approximates well the exact solution for $p = 2$: in Figure 8, the initial data is to the left, and the result is on the right. We see that the singularity has moved in the center of the domain, this being therefore

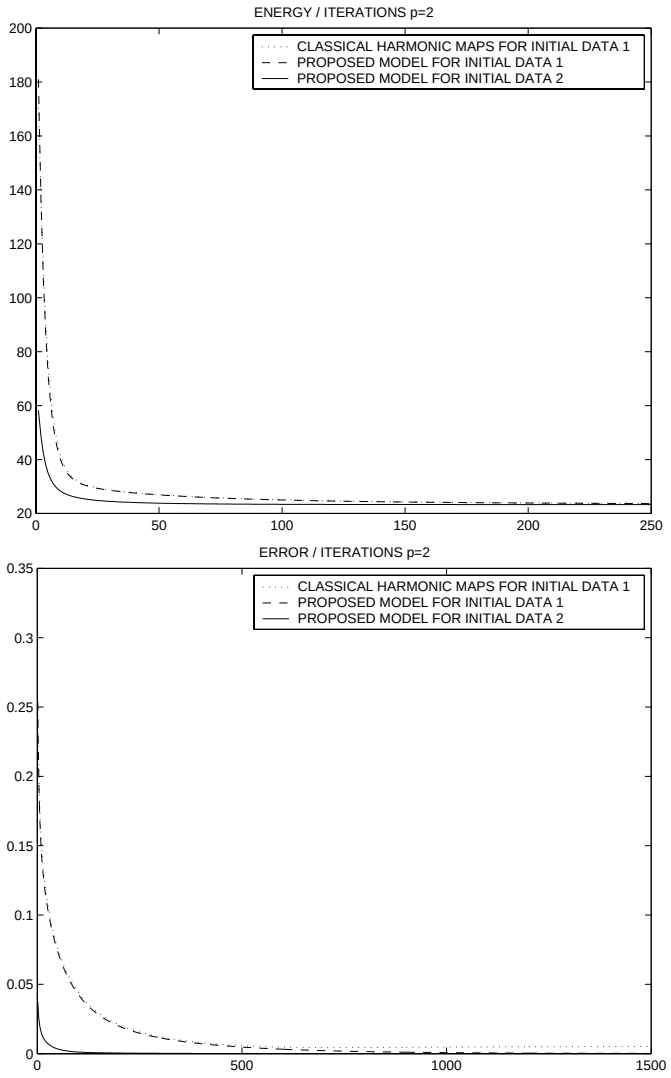


FIG. 3. *Energy and error versus iterations, for the classical harmonic map scheme applied to the initial data* 1*, and for the proposed model applied to both initial data* 1 *and* 2*. Note a much better accuracy obtained with the proposed model, compared with the classical formulation (we use the same* $\triangle t = 0.0001$*,* $h = 1./21$*), for both formulations.*

an approximation of $\frac{x-x_0}{|x-x_0|}$, with $x_0 = (0.5, 0.5, 0.5)$.

4.2. Application to directional denoising and color image denoising. Next, we consider the case with Neumann boundary conditions. For the initial data in Figure 9, the results for $p = 1$ and $p = 2$ are presented in Figure 10. Note that, for $p = 1$ (left), the "edges" are very well preserved, thanks to the total variation minimization [28], while denoising in the homogeneous regions. (We show the results at the steady state and without any fitting term.)

Finally, we show applications more related to denoising of color RGB images. In the first test (Figure 11), we consider a map from $\mathbb{R}^2 \to S^2$, but instead of vectors

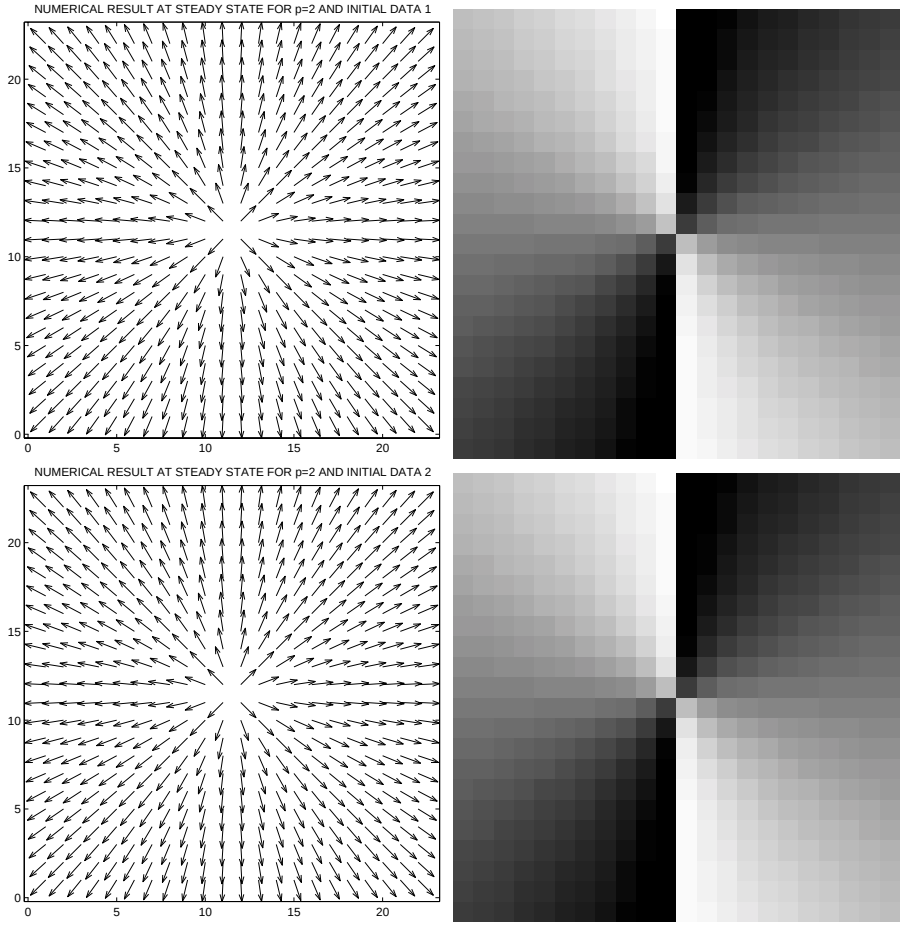


FIG. 4. *Left: numerical result, approximating well the exact solution and minimizer for* $p = 2$*, with Dirichlet boundary conditions* ($\triangle t = 0.0001$, $h = 1/21$)*. Right: corresponding angle* $\theta = \tan^{-1}\left(\frac{v}{u}\right)$.

we plot colors, using the rectangular color space RGB: in Figure 11 (left), we show an initial image of noisy directions. (The components of the unit vector (u, v, w) are visualized as channels in a color RGB picture.) We show in Figure 11 (middle and right) two numerical results in the case of directional diffusion, with $p = 1$ (middle) and $p = 2$ (right), with Neumann boundary conditions. As expected, in the case of the total variation [28], the edges are well preserved, while these are smeared out with the heat flow.

We end the paper with an application to denoising of color RGB images. We consider a color image $I = (I_R, I_G, I_B) \in \mathbb{R}^3$ from which we can extract the intensity or brightness $|I| = \sqrt{I_R^2 + I_G^2 + I_B^2}$ and the chromaticity

$$\frac{I}{|I|} = \left(\frac{I_R}{\sqrt{I_R^2 + I_G^2 + I_B^2}}, \frac{I_G}{\sqrt{I_R^2 + I_G^2 + I_B^2}}, \frac{I_B}{\sqrt{I_R^2 + I_G^2 + I_B^2}} \right) \in S^2.$$

Let us assume that noise has been added to the image but only to the chromaticity $\frac{I}{|I|}$.

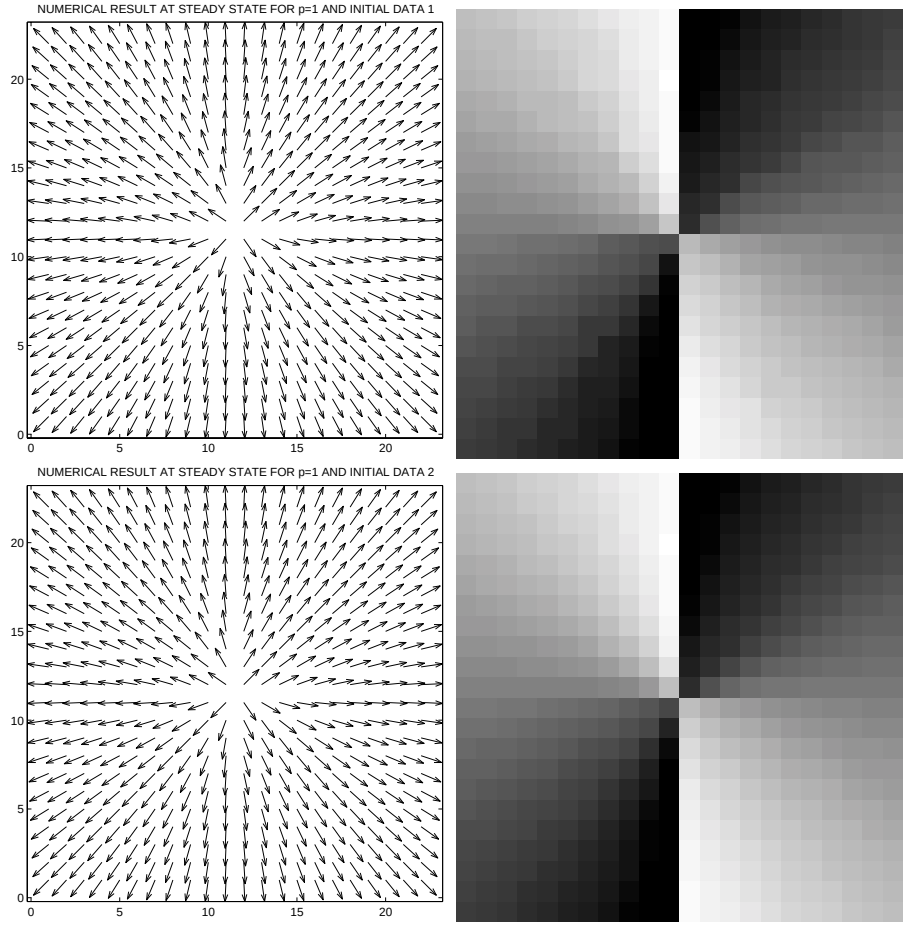


FIG. 5. *Left: numerical result, approximating well the exact solution and minimizer for* $p = 1$*, with Dirichlet boundary conditions* ($\triangle t = 0.00001$*,* $h = 1/21$*). Right: corresponding angle* $\theta = \tan^{-1}\left(\frac{v}{u}\right)$.

Then we can apply the above directional denoising method, with $p = 1$, to the chromaticity. (In this test case, we do not add noise to the brightness $|I|$.) If noise were also added to the brightness, then this could have been denoised, for example, with the corresponding total variation minimization [28] or any other anisotropic diffusion PDE. With the processed result, we obtain a denoised version of the image, using the unchanged brightness. We mention that the idea of decomposing a color RGB image into its brightness and chromaticity, and processing these two quantities separately, has been already used in other works (for example in [21], [40], [41], [38], [39], [9], [32], [23], [33]).

This type of application is illustrated in the last numerical example. In Figure 12, we show an original color RGB image $I = (I_R, I_G, I_B) \in \mathbb{R}^3$ (left), a noisy version (middle), where only the directions $\frac{I}{|I|}$ (the chromaticity) were noisy, keeping the brightness $|I|$ or magnitude of the vectors unchanged, and a denoised version obtained with $p = 1$ (right), where only the chromaticity or directions were denoised, keeping the brightness or magnitude unchanged from the original image, equal to $|I|$.

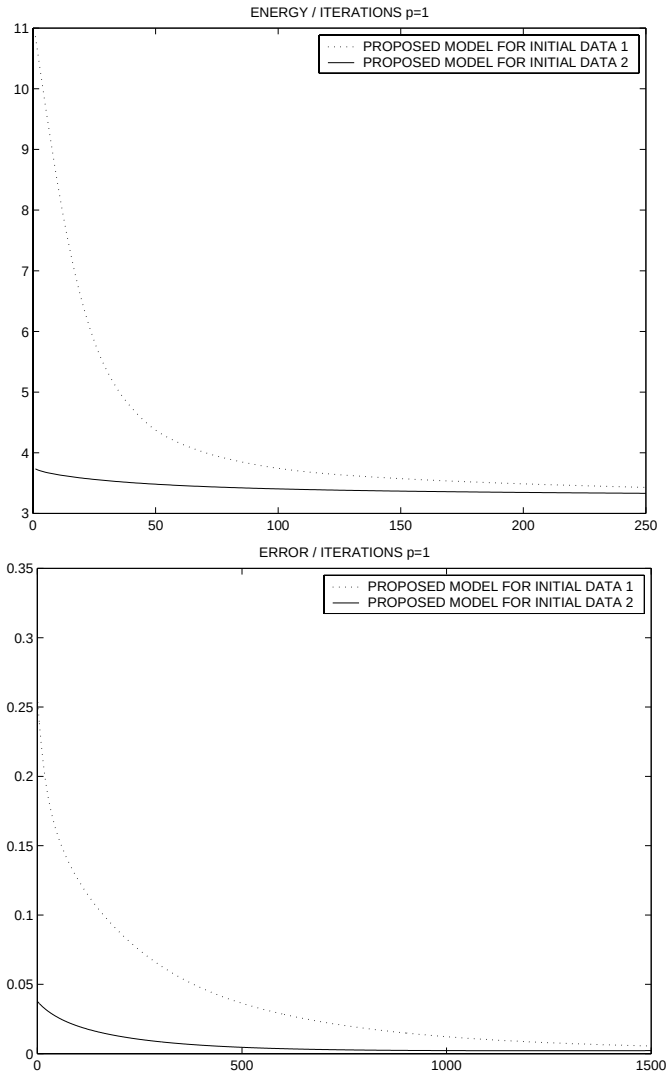


FIG. 6. *Energy and error versus iterations for $p = 1$ with Dirichlet boundary conditions, corresponding to the results in Figure* 5.

5. Concluding remarks and discussions for more general manifolds. In this paper, we have proposed an alternative approach for computing harmonic maps and harmonic flows. We have illustrated the proposed methods by experimental results and comparisons with classical schemes, and applications to directional diffusion and image processing.

It is easy to see that the minimization problems (1.1) and (1.4) have the same infimum, and that solving one problem yields a minimizer for the other one, and vice versa. Of course we cannot expect to have uniqueness of minimizers for (1.4), because λV is a minimizer for any nonzero constant λ if V is a minimizer. Showing the existence of minimizers for (1.4) may be a difficult problem, because the energy is not convex. We have also posed the following question: given Dirichlet boundary

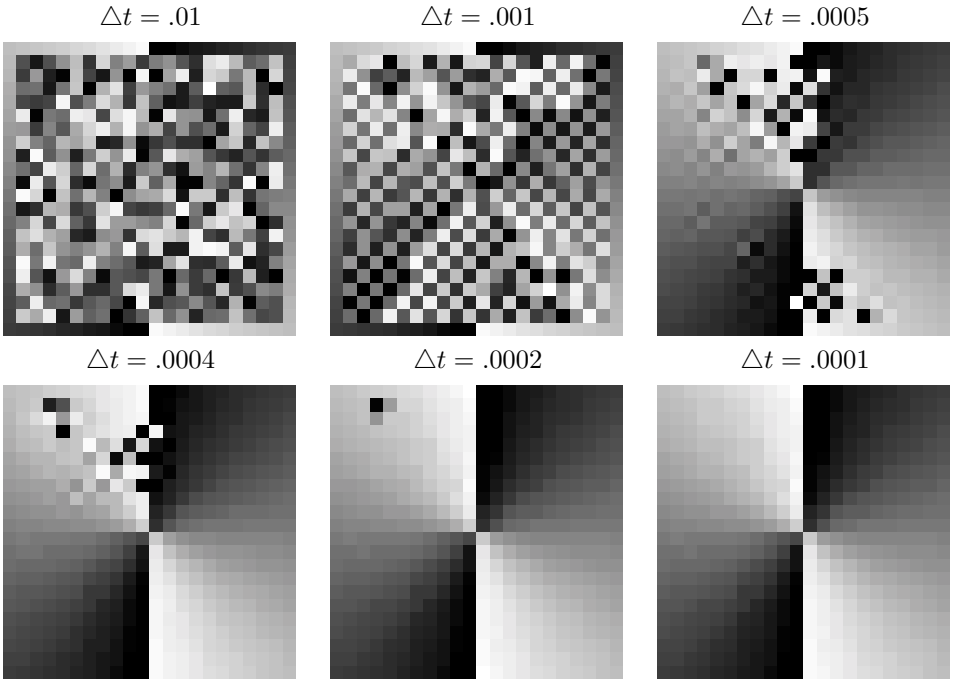


FIG. 7. *The angle* $\theta = \tan^{-1}\left(\frac{v}{u}\right)$ *at the steady state for* $p = 2$*, obtained using the proposed model with Dirichlet boundary conditions, for decreasing values of* $\triangle t$*. (If* $\triangle t$ *is too large,* θ *is noisy, but the numerical solution* (u^n, v^n) *always remains bounded.)*

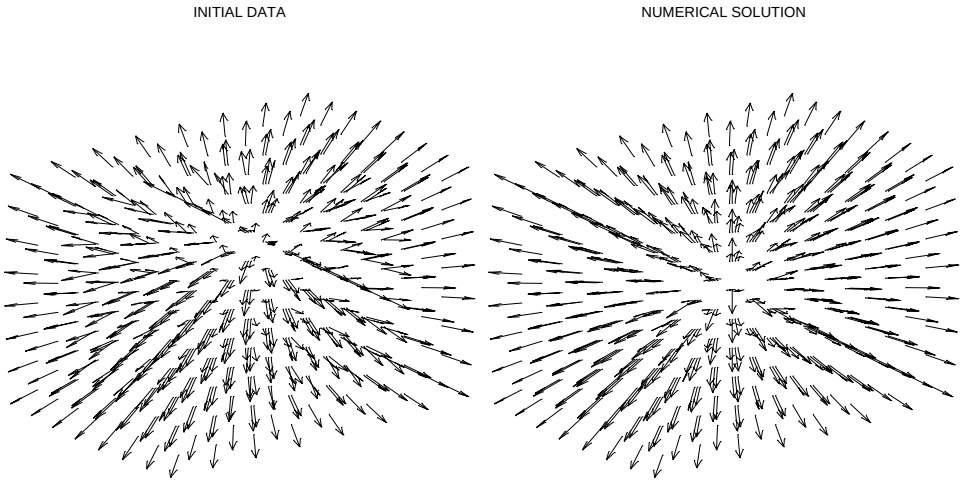


FIG. 8. *Top: initial flow* $\frac{x-x_1}{|x-x_1|}$ *from* $(0,1)^3$ *into* S^2 *with prescribed Dirichlet boundary conditions equal to* $\frac{x-x_0}{|x-x_0|}$*, where* $x_0 = (0.5, 0.5, 0.5)$ *and* $x_1 = (0.64, 0.64, 0.64)$*. Bottom: numerical solution obtained for* $p = 2$*. The singularity has moved to the center of the domain, approximating well the exact solution and minimizer (*$\triangle t = 0.00001$*,* $h = 1/7$*).*

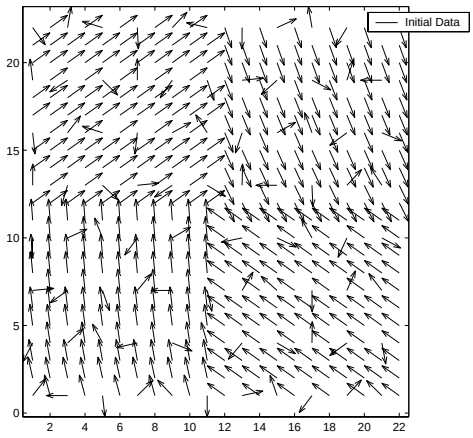


FIG. 9. *Initial noisy data for the case with Neumann boundary conditions.*

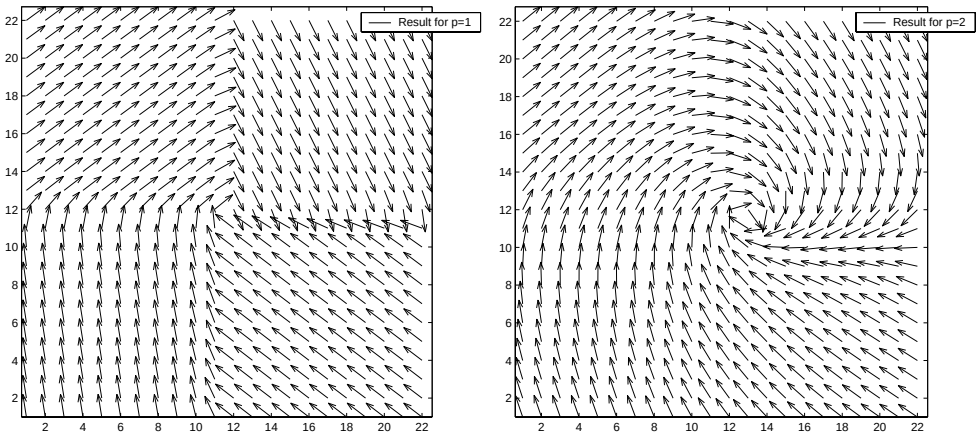


FIG. 10. *Numerical results with the initial noisy data from Figure* 9*, for Neumann boundary conditions and* $p = 1$ *(left), with* $\triangle t = 0.00005$*,* $h = 1$*, steady state, and* $p = 2$ *(right), with* $\triangle t = 0.00005$*,* $h = 1$*, steady state.*

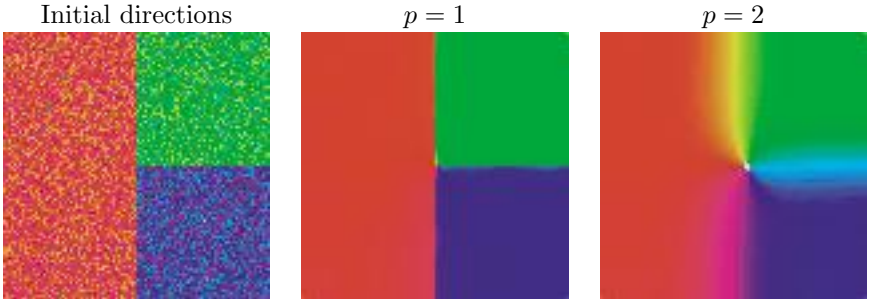


FIG. 11. *Directions denoising with* $p = 1$ *(middle) and* $p = 2$ *(right). The unit vectors are represented as RGB colors (*$\triangle t = 0.01$*,* $h = 1$*).*

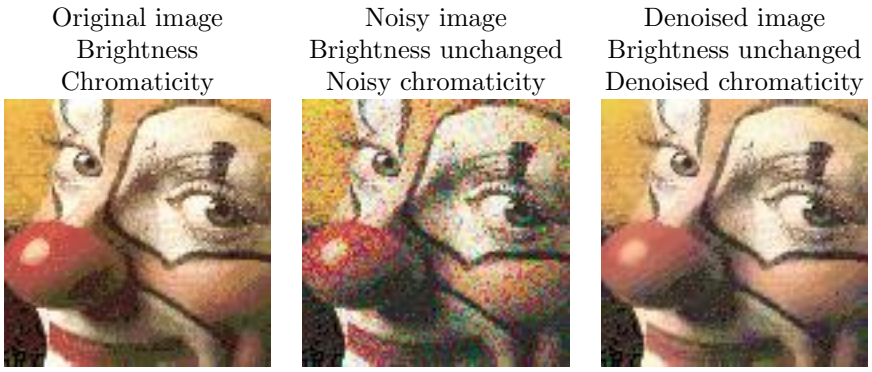


FIG. 12. *Chromaticity denoising with* $p = 1$. *The brightness is kept unchanged from the original image* ($\triangle t = 0.01$, $h = 1$, 50 *iterations*).

conditions on $\partial\Omega$, what would be a good initial condition in Ω to guarantee a fast computation of a minimizer? (To find a particular initial condition, we have used the distance function to the boundary $\partial\Omega$, although perhaps other choices could also be constructed.)

This method can be extended to more general manifolds. For instance, if we consider a manifold $\mathcal{M} \subset \mathbb{R}^N$, then the associated constrained minimization problem can be formulated as follows:

$$\inf_{U:\Omega\to\mathcal{M}} F(U) = \int_\Omega |\nabla U|^p dx.$$

The proposed method for the case when $\mathcal{M} = S^{N-1}$ can be extended to such general cases if we assume, for example, that $\mathcal{M}$ can be represented implicitly, via a level set function, given by the signed distance function to $\mathcal{M}$, from any other point in $\mathbb{R}^N$. (We refer the reader to [26] for definitions and dynamics of closed hypersurfaces defined implicitly, via level set functions and signed distance functions.) Then we can write $\mathcal{M} = \{x \in \mathbb{R}^N : \ d(x) = 0\}$, where d is the signed distance function to $\mathcal{M}$ (in particular a Lipschitz function, taking real values). To any $U : \Omega \to \mathcal{M}$, we associate $V : \Omega \to \mathbb{R}^N$ such that U is the projection of V on the manifold $\mathcal{M}$. This can be done using the closest point or the projection $U = V - d(V)\nabla_V d(V)$, and we have $d(U) = 0$. Then we can associate the unconstrained minimization problem

$$\inf_{V:\Omega\to\mathbb{R}^N} \int_\Omega |\nabla(V - d(V)\nabla_V d(V))|^p dx.$$

This is a generalization of the case $\mathcal{M} = S^{N-1}$, because in this case we have $d(V) = |V| - 1$, and $V - d(V)\nabla_V d(V) = \frac{V}{|V|}$. We plan to consider in the future the solution of this general unconstrained minimization problem.

We would like to mention that the case of more general manifolds, and in particular of manifolds defined implicitly, has been considered in [3], [4] for the manifold of origin and in [25] for the target manifold, but using different formulations.

We would also like to mention that the idea of solving constrained minimization problems for harmonic maps by associating unconstrained minimization problems has been used as a theoretical tool by Chen and Lin [10], Chen and Struwe [11], and Struwe [37]. They find a smooth energy-minimizing harmonic map U as a weak limit

of minimizers U_L to an unconstrained variational problem for $L \to \infty$. They construct in a different way the unconstrained variational problems.

Appendix. We show here that, parameterizing the descent direction by an artificial time, the energy is still decreasing under the associated flow. (See the last remark from subsection 2.1.) We show this property in a general framework. In order to solve the minimization problem

$$\inf_{u_1,\dots,u_N} \int_\Omega f(u_1,\dots,u_N,\nabla u_1,\dots,\nabla u_N)dx,$$

where $\Omega \subset \mathbb{R}^N$, $x = (x_1,\dots,x_M) \in \mathbb{R}^M$, we associate the time-dependent coupled PDEs for $1 \le i \le N$ (the functions u_i take real values, and we use the notations $u = (u_1,\dots,u_N)$, $\nabla u = (\nabla u_1,\dots,\nabla u_N)$, $(u_i)_{x_j} = \frac{\partial u_i}{\partial x_j}$):

$$\frac{\partial u_i}{\partial t} = -\frac{\partial f(u,\nabla u)}{\partial u_i} + \sum_{j=1}^{M} \frac{\partial}{\partial x_j}\left(\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}\right),$$

with the initial conditions $u_i(0,x) = u_{0,i}(x)$ in Ω. On the boundary $\partial\Omega$, we can assume Dirichlet boundary conditions $u_i(t,x) = u_{0,i}(x)$ for $x \in \partial\Omega$ and $t > 0$, or free boundary conditions in the form $\sum_{j=1}^{M} \frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})} n_j = 0$, where $\vec{n} = (n_1,\dots,n_M)$ is the exterior unit normal to $\partial\Omega$.

We formally compute now $\frac{d}{dt}\int_\Omega f(u_1,\dots,u_N,\nabla u_1,\dots,\nabla u_N)dx = \frac{d}{dt}\int_\Omega f(u,\nabla u)dx$, and we show that this quantity is always negative or zero; therefore the energy is decreasing in time:

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega f(u,\nabla u)dx \\
&= \sum_{i=1}^{N}\int_\Omega \left(\frac{\partial f(u,\nabla u)}{\partial u_i}\right)\left(\frac{\partial u_i}{\partial t}\right)dx + \sum_{i=1}^{N}\int_\Omega \left[\sum_{j=1}^{M}\left(\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}\right)\left(\frac{\partial (u_i)_{x_j}}{\partial t}\right)\right]dx \\
&= \sum_{i=1}^{N}\int_\Omega \left(\frac{\partial f(u,\nabla u)}{\partial u_i}\right)\left(\frac{\partial u_i}{\partial t}\right)dx + \sum_{i=1}^{N}\int_\Omega \left[\sum_{j=1}^{M}\left(\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}\right)\left(\frac{\partial (u_i)_t}{\partial x_j}\right)\right]dx \\
&= \sum_{i=1}^{N}\int_\Omega \left(\frac{\partial f(u,\nabla u)}{\partial u_i}\right)\left(\frac{\partial u_i}{\partial t}\right)dx + \sum_{i=1}^{N}\int_\Omega \left\{-\sum_{j=1}^{M}\left[\frac{\partial}{\partial x_j}\left(\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}\right)\right]\left(\frac{\partial u_i}{\partial t}\right)\right\}dx \\
&\quad + \sum_{i=1}^{N}\left\{\int_{\partial\Omega}\left(\frac{\partial u_i}{\partial t}\right)\left(\sum_{j=1}^{M}\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}n_j\right)dS\right\} \\
&= \sum_{i=1}^{N}\int_\Omega \left(\frac{\partial u_i}{\partial t}\right)\left[\frac{\partial f(u,\nabla u)}{\partial u_i} - \sum_{j=1}^{M}\frac{\partial}{\partial x_j}\left(\frac{\partial f(u,\nabla u)}{\partial((u_i)_{x_j})}\right)\right]dx \\
&= -\sum_{i=1}^{N}\int_\Omega \left(\frac{\partial u_i}{\partial t}\right)^2 dx \;\le\; 0.
\end{aligned}$$

Acknowledgments. The authors would like to thank Guillermo Sapiro for his interest in this work and for his very useful remarks and suggestions. The authors started to work on this subject after one of his visits at UCLA. We also thank him

for providing us the "clown" picture. We would also like to thank Ron Kimmel for interesting discussions and for pointing out to us some related references. Finally, we are grateful to the first unknown referee for his numerous typographical and stylistic corrections, helping to improve the presentation of the paper.

REFERENCES

[1] F. Alouges, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.

[2] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[3] M. Bertalmio, G. Sapiro, L.-T. Cheng, and S. Osher, *Variational problems and partial differential equations on implicit surfaces*, in Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, Vancouver, BC, Canada, 2001, pp. 186–193.

[4] M. Bertalmio, L.-T. Cheng, S. Osher, and G. Sapiro, *Variational problems and partial differential equations on implicit surfaces*, J. Comput. Phys., 174 (2001), pp. 759–780.

[5] F. Bethuel, H. Brezis, and J.M. Coron, *Relaxed energies for harmonic maps*, in Variational Problems, H. Beresticky, J. M. Coron, and I. Ekeland, eds., Birkhauser, Paris, Basel, 1988.

[6] F. Bethuel, H. Brezis, and F. Helein, *Asymptotics for the minimization of a Ginzburg-Landau functional*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 123–148.

[7] F. Bethuel, H. Brezis, and F. Helein, *Singular limit for the minimization of Ginzburg-Landau functionals*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 891–895.

[8] H. Brezis, J.M. Coron, and E. H. Lieb, *Harmonic maps with defects*, Comm. Math. Phys., 107 (1986), pp. 649–705.

[9] T. Chan and J. Shen, *Variational restoration of nonflat image features: Models and algorithms*, SIAM J. Appl. Math., 61 (2000), pp. 1338–1361.

[10] Y. Chen and F.-H. Lin, *Remarks on approximate harmonic maps*, Comment. Math. Helv., 70 (1995), pp. 161–169.

[11] Y. Chen and M. Struwe, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.

[12] R. Cohen, S.-Y. Lin, and M. Luskin, *Relaxation and gradient methods for molecular orientation in liquid crystals*, Comput. Phys. Comm., 53 (1989), pp. 455–465.

[13] R. Cohen, R. Hardt, D. Kinderlehrer, S.-Y. Lin, and M. Luskin, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, Springer-Verlag, New York, 1987, pp. 99–122.

[14] J.M. Coron and R. Gulliver, *Minimizing p-harmonic maps into spheres*, J. Reine Angew. Math., 401 (1989), pp. 82–100.

[15] P. Courilleau and F. Demengel, *Heat flow for p-harmonic maps with values in the circle*, Nonlinear Anal., 41 (2000), pp. 689–700.

[16] P. Courilleau, S. Dumont, and R. Hadiji, *Regularity of Minimizing Maps with Values in S^2 and Some Numerical Simulations*, CMLA-ENS Technical report 9905, University of California, Los Angeles, CA, 1999.

[17] B. Dacorogna, I. Fonseca, J. Maly, and K. Trivisa, *Manifold constrained variational problems*, Calc. Var. Partial Differential Equations, 9 (1999), pp. 185–206.

[18] W. E and X.-P. Wang, *Numerical methods for the Landau–Lifshitz equation*, SIAM J. Numer. Anal., 38 (2000), pp. 1647–1665.

[19] M. Giaquinta, G. Modica, and J. Soucek, *Variational problems for maps of bounded variation with values in S^1*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 87–121.

[20] R. Hardt, D. Kinderlehrer, and M. Luskin, *Remarks about the Mathematical Theory of Liquid Crystals*, in Calculus of Variations and Partial Differential Equations, Lecture Notes in Math. 1340, S. Hildebrandt, D. Kinderlehrer, and M. Miranda, eds., Springer-Verlag, New York, 1988, pp. 123–138.

[21] D. Karakos and P.E. Trahanias, *Generalized multi-channel image-filtering structures*, IEEE Trans. Image Process., 6 (1997), pp. 1038–1045.

[22] R. Kimmel, R. Malladi, and N. Sochen, *Images as embedded maps and minimal surfaces: Movies, color, texture, and volumetric medical images*, Int. J. Comput. Vision, 39 (2000), pp. 111–129.

[23] R. Kimmel and N. Sochen, *Orientation diffusion or how to comb a porcupine*, Journal of Visual Communication and Image Representation, 13 (2002), pp. 238–248.

[24] S.-Y. Lin and M. Luskin, *Relaxation methods for liquid crystal problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1310–1324.

[25] F. Memoli, G. Sapiro, and S. Osher, *Solving Variational Problems and Partial Differential Equations Mapping into General Target Manifolds*, UCLA CAM Report 02-04, University of California, Los Angeles, CA, 2002.
[26] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulation*, J. Comput. Phys., 79 (1988), pp. 12–49.
[27] P. Perona, *Orientation diffusion*, IEEE Trans. Image Process., 7 (1998), pp. 457–467.
[28] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[29] R. Schoen and K. Uhlenbeck, *A regularity theory for harmonic maps*, J. Differential Geom., 17 (1982), pp. 307–335.
[30] R. Schoen and K. Uhlenbeck, *Boundary regularity and the Dirichlet problem for harmonic maps*, J. Differential Geom., 18 (1983), pp. 253–268.
[31] R. Schoen and K. Uhlenbeck, *Regularity of minimizing harmonic maps into the sphere*, Invent. Math., 78 (1984), pp. 89–100.
[32] N. A. Sochen and R. Kimmel, *Combing a porcupine via stereographic direction diffusion*, in Scale-Space 2001, Lecture Notes in Comput. Sci. 2106, M. Kerckhove, ed., Springer-Verlag, Berlin, 2001, pp. 308–316.
[33] N. Sochen, R. Kimmel, and R. Malladi, *A general framework for low level vision*, IEEE Trans. Image Process., 7 (1998), pp. 310–318.
[34] M. Struwe, *On the evolution of harmonic mappings of Riemannian surfaces*, Comment. Math. Helv., 60 (1985), pp. 558–581.
[35] M. Struwe, *Variational Methods*, Springer-Verlag, New York, 1990.
[36] M. Struwe, *The Evolution of Harmonic Maps: Existence, Partial Regularity and Singularities*, Progr. Nonlinear Differential Equations Appl. 7, Birkhäuser Boston, Boston, MA, 1992, pp. 485–491.
[37] M. Struwe, *Uniqueness of harmonic maps with small energy*, Manuscripta Math., 96 (1998), pp. 463–486.
[38] B. Tang, G. Sapiro, and V. Caselles, *Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case*, Int. J. Comput. Vision, 36 (2000), pp. 149–161.
[39] B. Tang, G. Sapiro, and V. Caselles, *Color image enhancement via chromaticity diffusion*, IEEE Trans. Image Process., 10 (2001), pp. 701–707.
[40] P. E. Trahanias and A.N. Venetsanopoulos, *Vector directional filters—a new class of multichannel image processing filters*, IEEE Trans. Image Process., 2 (1993), pp. 528–534.
[41] P. E. Trahanias, D. Karakos and A.N. Venetsanopoulos, *Directional processing of color images: Theory and experimental results*, IEEE Trans. Image Process., 5 (1996), pp. 868–880.
[42] D. Tschumperle and R. Deriche, *Regularization of orthonormal vector sets using coupled PDE's*, in Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, Vancouver, BC, Canada, 2001, pp. 3–10.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2105–2132

PIECEWISE SELF-SIMILAR SOLUTIONS AND A NUMERICAL SCHEME FOR SCALAR CONSERVATION LAWS*

YONG-JUNG KIM†

Abstract. The solution of the Riemann problem was a building block for general Cauchy problems in conservation laws. A Cauchy problem is approximated by a series of Riemann problems in many numerical schemes. But, since the structure of the Riemann solution holds locally in time only, and, furthermore, a Riemann solution is not piecewise constant in general, there are several fundamental issues in this approach such as the stability and the complexity of computation.

In this article we introduce a new approach which is based on piecewise self-similar solutions. The scheme proposed in this article solves the problem without the time marching process. The approximation error enters in the step for the initial discretization only, which is given as a similarity summation of base functions. The complexity of the scheme is linear. Convergence to the entropy solution and the error estimate are shown. The mechanism of the scheme is introduced in detail together with several interesting properties of the scheme.

Key words. self-similarity, characteristics, front tracking, gridless scheme

AMS subject classifications. 65M25, 35L65

PII. S0036142901381364

1. Introduction.

Self-similarity of the Cauchy problem for one-dimensional conservation laws,

$$\begin{aligned} &v_t + f(v)_x = 0, \\ &v(x,0) = v_0(x), \end{aligned} \qquad x, v \in \boldsymbol{R},\ t > 0, \tag{1.1}$$

with Riemann initial data

$$v(x,0) = \begin{cases} v_-, & x \le 0, \\ v_+, & x > 0, \end{cases} \tag{1.2}$$

has been the basis of various schemes devised for general initial value problems; see Glimm [10] and Godunov [11], for example. The self-similarity of the Riemann problem is the property that the solution is a function of the self-similarity variable $\xi = x/t$. In other words, the solution is constant along the self-similarity lines

$$\frac{x}{t} = \text{constant}. \tag{1.3}$$

The basic idea of the Godunov scheme for a general initial value problem is to approximate the initial data by a piecewise constant function and then apply the self-similarity structure to the series of Riemann problems.

There are two basic issues we have to consider immediately in this approach. First, since the self-similarity for a piecewise constant solution holds locally in time only, the structure of the Riemann solution can be applied for a small time period. In other words, the scheme is not free from the CFL condition (see [4], [5]), and,

*Received by the editors June 7, 2001; accepted for publication (in revised form) July 8, 2002; published electronically December 13, 2002. This work was supported in part by Korea Science and Engineering Foundation (grant R11-2002-103).

http://www.siam.org/journals/sinum/40-6/38136.html

†School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (yongkim@math.umn.edu).

hence, the scheme can march just a little amount of time every time step and it costs computation time. Furthermore, since rarefaction waves appear immediately, the solution is not piecewise constant anymore. So a numerical scheme contains a process which rearranges the rarefaction wave into a piecewise constant function every time step. The numerical viscosity enters in this process, and tracking down the behavior of the scheme becomes extremely hard.

LeVeque [20] considers a large time step technique based on the Godunov method for the genuinely nonlinear problem. In the scheme the CFL number may go beyond 1, and it is even possible to solve the propagation of a simple wave in a single step, i.e., $\Delta t = T$ for the given final time $T > 0$. However, the scheme handles interactions between waves incorrectly if the CFL number is so large.

One way to avoid the rearranging process is to consider a modified equation,

$$\begin{array}{l} u_t + h(u)_x = 0, \\ u(x,0) = u_0(x), \end{array} \qquad x, u \in \boldsymbol{R},\ t > 0, \tag{1.4}$$

where h and u_0 approximate f and v_0, respectively. Dafermos [7] considers a polygonal approximation $h \sim f$, i.e., h is continuous and piecewise linear. In this case the exact solution of (1.4) is piecewise constant. So the method does not require a rearranging process, and, therefore, the numerical viscosity is not introduced and the error is controlled by refining the polygonal approximation h. In this approach, the exact behavior of the numerical solution can be monitored more closely and we may get a more detailed understanding of the scheme. This idea has been developed in Holden and Holden [12], and it has been extended to multidimensional problems in Holden and Risebro [14] and to systems of conservation laws in Holden, Lie, and Risebro [13]. In particular, we refer to Bressan [2], [3] for systems. This front tracking method has been developed as a computational tool (e.g., [21], [22]).

Lucier [24] approximates the actual flux f by a piecewise parabolic function h and achieves a second order scheme. In this case, the initial data $v_0(x)$ are approximated by a piecewise linear function u_0 and the solution remains piecewise linear. The difference between the solutions of the original problem (1.1) and the modified problem (1.4) is estimated by

$$||v(\cdot,t) - u(\cdot,t)||_1 \le ||v_0 - u_0||_1 + t||f' - h'||_\infty ||v_0||_{BV}. \tag{1.5}$$

Since the linear approximation is of second order, he achieves a second order scheme for a fixed time $t > 0$.

If we want to design a numerical scheme which represents the exact solution, we have to find a way to choose grid points correctly. If they are simply fixed, it is clear that the scheme cannot represent the exact solution and, hence, we need to rearrange the solution to fit the solution to the fixed grid points. So it is natural to consider the moving mesh method; see Miller [25]. In Lucier [24] the moving mesh method is used to find the exact solution of (1.4), where mesh points move along characteristics. Another option is not to use any grid point. In numerical schemes based on the front tracking method mentioned earlier, grid points are used just for the initial discretization. The scheme we develop in this article does not use any grid point either.

This article has two goals. The first one is to introduce the mathematical idea which is behind the piecewise self-similar solutions. The second one is to demonstrate how to implement the idea into a numerical scheme and show properties of the scheme. From the study of the Burgers equation (see [17] or Whitham [26]), it is well known

that the primary structure which dominates the evolution is a saw-tooth profile. In fact, this profile is a series of N-waves and eventually the solution evolves to a single N-wave; see Liu and Pierre [23]. The starting point of our scheme is to use this structure as the unit of the scheme.

If a solution $u(x,t)$ is a function of the self-similarity variable $\xi = x/t$, then we can easily derive from (1.1) that

$$f'(u(x,t)) = x/t.$$

Roughly speaking, a piecewise self-similarity (initial) profile has the structure of

$$f'(u(x,0)) = \frac{x - c_k}{t_k}, \qquad x \in (a_k, b_k),\ c_k, t_k \in \boldsymbol{R}. \tag{1.6}$$

Note that the time index t_k can be a negative number. In this article, we observe that the solution of (1.1) with piecewise self-similarity initial profile has such a structure for all $t > 0$, i.e.,

$$f'(u(x,t)) = \frac{x - c_k}{t + t_k}, \qquad x \in (a_k(t), b_k(t)),\ c_k, t_k \in \boldsymbol{R},\ t \in \boldsymbol{R}^+, \tag{1.7}$$

and we give the explicit formula for this kind of solution in several situations. First we consider a convex flux with positive wave speed,

$$f''(u) \geq 0, \quad f'(u) \geq 0, \tag{H}$$

where f is locally Lipschitz continuous. The convexity of the flux $f''(u) \geq 0$ is used to get the explicit formula $g(x)$ of the self-similarity profile such that $f'(g(x)) = x$, and the self-similarity profile (1.7) can be written as

$$u(x,t) = g\Big(\frac{x - c_k}{t + t_k}\Big), \qquad x \in (a_k(t), b_k(t)),\ c_k, t_k \in \boldsymbol{R},\ t \in \boldsymbol{R}^+. \tag{1.8}$$

Note that the equality is included for the second derivative of the flux in (H), and, hence, the monotonicity of f' is not strict and $g(x)$ is not exactly the inverse function of f', and $g(f'(u)) \neq u$ in general. In this approach, we may include a piecewise linear flux of the front tracking method; see Remark 6.4.

Our approach is as follows. We start our discussion reviewing the self-similarity property in conservation laws in section 2. This discussion leads us to the study of piecewise self-similar solutions, which is the case when the self-similarity lines and characteristics are compatible. In section 3 we consider a piecewise self-similar solution which can be written as a *self-similarity summation* (or simply *S-summation*),

$$\bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x), \quad c_n < \cdots < c_2 < c_1, \tag{1.9}$$

of a finite number of base functions. We give definitions for the S-summation and base functions in the section and show that $u(x,t) = \bigodot_{k=1}^{n} B_{m_k,t+t_k,c_k}(x)$ is the solution of (1.1) with initial data $u_0(x) = \bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x)$; see Theorem 3.6. We consider u as an approximation of the solution v with the original initial data v_0. Then the L^1 contraction theory of conservation laws (see Hörmander [15], Kruzhkov [18], [19]) implies

$$||v(\cdot,t) - u(\cdot,t)||_1 \leq ||v_0 - u_0||_1. \tag{1.10}$$

It is the estimate corresponding to the error estimate (1.5), which does not have the time dependent term anymore. It is natural to expect that the error increases in time if the flux is changed. In our approach, we use the original flux and the error decreases in time. In fact, the left-hand side of (1.10) is of order $O(t^{-1})$; see [16]. The convergence of the scheme is now clear (see Theorem 3.6, Corollary 3.7). Note that the S-summation (1.9) represents only a special kind of piecewise self-similar profile in (1.6), which has positive indexes $t_k > 0$ and is ordered appropriately, i.e., $c_n < \cdots < c_2 < c_1$ if $a_n \le b_n \le \cdots \le a_2 \le b_2 \le \cdots \le a_1 \le b_1$.

The S-summation is successfully coded for a numerical scheme in section 4. This scheme has several unique properties. First, it does not require a time marching procedure. So the complexity of the scheme is of order $O(N)$, not $O(N^2)$. CPU times for several cases are compared in section 4.3. Second, it captures the shock location very well even if a small number of base functions (or mesh points) are used; see Figure 4.5. In the figure it is clearly observed that the solution with finer mesh always passes through bigger artificial shocks. Since it does not introduce numerical viscosity at all, we obtain a very good resolution for an inviscid problem. This scheme also distinguishes physical shocks and artificial ones clearly. Table 4.4 shows the time when the physical shock appears.

In section 5 we generalize the method. For a general convex flux, i.e.,

$$f''(u) \ge 0, \tag{H1}$$

the method is applied through the transformations (5.1) and (5.3). If the flux has inflection points, then the scheme becomes considerably complicated and it is beyond the purpose of this article. But, if the flux has only one inflection point, for example,

$$f''(u) \le 0 \text{ for } u \le A, \quad f''(u) \ge 0 \text{ for } u \ge A, \tag{H2}$$

then we can easily apply the scheme through a similar transformation (5.4). Dafermos [8] considers a flux with a single inflection point through generalized characteristics. The Buckley–Leverett equation satisfies this condition. The flux $f(u) = u^2 - u^3$, which appears in thin film flows (see Bertozzi, Münch, and Shearer [1]), also belongs to this category. Figure 5.2 shows the strength of our scheme over the Godunov scheme in this case.

The scheme is not good enough for a short time behavior $t \ll 1$ since the initial error $||v_0 - u_0||_1$ is not controlled efficiently. To resolve this issue we add an extra structure to base functions in section 6. Using these base functions, we can approximate the initial data with second order accuracy and solve the solution for the modified initial datum. Furthermore, a general piecewise self-similarity profile (1.6) can be written in terms of S-summation of these modified base functions.

2. Self-similarity in conservation laws. Consider one-dimensional scalar conservation laws,

$$\begin{array}{l} u_t + f(u)_x = 0, \\ u(x,0) = u_0(x), \end{array} \quad x, u \in \boldsymbol{R},\ t > 0, \tag{2.1}$$

where the flux f is locally Lipschitz continuous. For a nonlinear flux $f(u)$ the solution may have a singularity, and hence the solution is considered in the weak sense with the entropy admissibility condition:

$$\frac{f(\tilde{u}) - f(u_-)}{\tilde{u} - u_-} \ge \frac{f(u_+) - f(u_-)}{u_+ - u_-}, \tag{2.2}$$

for any number $\tilde{u}$ lying between $u_+ = u(x+,t)$ and $u_- = u(x-,t)$. It is well known that the self-similarity of a conservation law is inherited from the fact that a rescaled function,

$$w(x,t) = u(ax,at), \quad a > 0, \tag{2.3}$$

is also the solution of (2.1) if and only if the initial profile $u_0(x)$ satisfies

$$u_0(x) = u_0(ax), \quad a > 0. \tag{2.4}$$

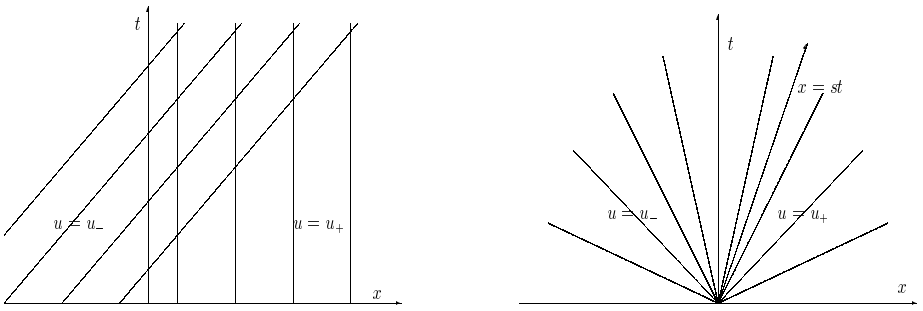
It is clear that, if the Riemann initial value,

$$u(x,0) = \begin{cases} u_-, & x < 0, \\ u_+, & x > 0, \end{cases} \tag{2.5}$$

is given, (2.4) is satisfied and, hence, $u(x,t) = u(ax,at)$ for all $a > 0$. Therefore, $u(x,t)$ is a function of the self-similarity variable,

$$u(x,t) = u(\xi), \qquad \xi = x/t. \tag{2.6}$$

The structure of a Riemann solution is given in Figure 2.1 together with characteristic lines. Note that, even though a self-similarity line $x/t = \xi$, $\xi \in \boldsymbol{R}$, is not a characteristic line, the solution is constant along it. This is a special property of the Riemann problem, and it is not expected in a general Cauchy problem.



(a) *Characteristic lines* (b) *Self-similarity lines*

FIG. 2.1. *Let $f'(u_+) = 0$ and $f'(u_-) = 1$. Then self-similarity lines are different from characteristic lines. However, the solution is constant along self-similarity lines.*

If the total mass of the initial data $u_0(x)$ is finite, i.e.,

$$\int |u_0(x)|dx < \infty, \tag{2.7}$$

then the relation (2.4) cannot be satisfied since the transformation $u_0(x) \to u_0(ax)$ does not preserve the total mass. So the solution cannot be a function of self-similarity variable $\xi = x/t$. In the following, we consider techniques to achieve the Riemann solution like self-similarity for general Cauchy problems.

Suppose that characteristic lines of the solution $u(x,t)$ pass through the origin. Then the relation between the wave speed and characteristics gives

$$f'(u) = \frac{x}{t}. \tag{2.8}$$

Since the right-hand side diverges as $t \to 0$, we consider the initial datum as the profile at a given time $t_0 > 0$. The simplest case of L^1 initial datum of the kind is

$$f'(u(x,0)) = \frac{x}{t_0} \text{ if } 0 < x < s_0, \qquad u(x,0) = 0 \quad \text{otherwise.} \tag{2.9}$$

Characteristic lines of this initial profile are given in Figure 2.2. Nonvertical characteristics pass through the point $(0, -t_0)$, and there is a region in which characteristic lines overlap with each other. The solution is given by finding the shock characteristic $x = s(t)$ correctly. In this case, the shock characteristic $x = s(t)$ is not a straight line and the solution is not a function of $x/(t+t_0)$. However, the solution is a function of $x/(t+t_0)$ in the region $0 < x < s(t)$, i.e.,

$$f'(u(x,t)) = \frac{x}{t+t_0} \text{ if } 0 < x < s(t), \qquad u(x,t) = 0 \quad \text{otherwise.} \tag{2.10}$$

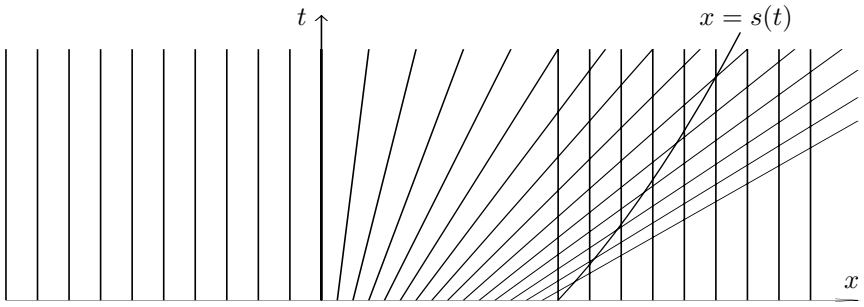


FIG. 2.2. *Characteristic lines of a self-similarity solution are similar to self-similarity lines. The main difference is that the shock characteristic is not a straight line anymore.*

Since the shock speed $s'(t)$ satisfies the Rankine–Hugoniot jump condition, the shock location $s(t)$ can be found by its integral form. On the other hand, if the convexity of the flux f is assumed, i.e.,

$$f''(u) \geq 0,$$

we may consider the self-similarity profile $g(x)$ such that $f'(g(x)) = x$. In this case we obtain

$$u(x,t) = g\big(x/(t+t_0)\big), \qquad 0 < x < s(t), \tag{2.11}$$

and we can easily find the shock location $s(t)$ using the equal area rule,

$$\int_0^{s(t)} g(x/(t+t_0))dx = \int_0^{s_0} g(x/t_0)dx, \quad t > 0. \tag{2.12}$$

Since the conservation law (2.1) does not explicitly depend on the x variable, we may translate the initial data (2.9) in the x-direction. We can also consider initial data which consist of a finite number of structures in (2.9). A simple example is

$$u_0(x) = \sum_{k=1}^{N} g\Big(\frac{x - c_k}{t_k}\Big)\chi_{(c_k, s_k)}, \tag{2.13}$$

where centers c_k and shock locations s_k satisfy

$$-\infty < c_N < s_N < \cdots < c_1 < s_1 < \infty. \tag{2.14}$$

The time indexes $t_k > 0$ in (2.13) decide the slope of the initial profile, and they do not need to be identical. Condition (2.14) implies that all the profiles in (2.13) are separated. If not, the simple summation in (2.13) breaks down the self-similarity structure we want to keep. In section 3 we consider an S-summation which preserves this structure. Figure 2.3 shows characteristic lines for initial data (2.13) with $N = 4$. In this case, tracking down a shock is more complicated and (2.12) is not valid anymore. The following section is devoted to handling the general case.

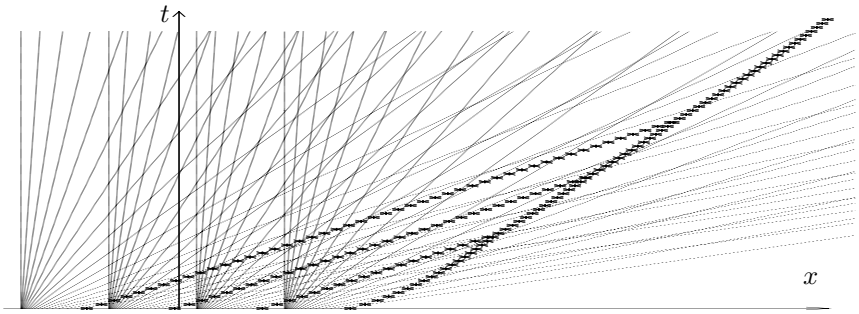


FIG. 2.3. *Shock characteristics (dots) are merged together after contacts among them, and a bigger shock appears which is a physical one.*

3. Piecewise self-similar solutions. In this section we define the S-summation and show that, if the initial value $u_0(x)$ is given as an S-summation, then so is the solution $u(x,t)$ of (2.1) at any given time $t > 0$. We consider the flux under the hypothesis,

$$f''(u) \geq 0, \quad f'(u) \geq 0, \tag{H}$$

and the self-similarity profile $g(x)$ satisfies $f'(g(x)) = x$. We may assume $f'(0) = 0$ without the loss of generality, and it implies that the solution is actually assumed to be positive under (H). The results in this section are generalized in section 5.

3.1. Base functions. As was mentioned earlier, the self-similarity profile

$$u(x,t) = g(x/t), \qquad t > 0, \tag{3.1}$$

represents the asymptotic behavior of the conservation law (1.1). A triple index function $B_{t,c,s}(x)$, defined by

$$B_{t,c,s}(x) = \begin{cases} g\big((x-c)/t\big), & c < x < s, \\ 0, & \text{otherwise}, \end{cases} \tag{3.2}$$

serves as a base function in this article. A base function has the self-similarity profile over the interval between the *center* c and the *shock location* s. The *area* (or *mass*) m enclosed by the x-axis and the base function is given by

$$m = \int_c^s B_{t,s,c}(x)dx = \int_0^{s-c} g(x/t)dx =: m(t,c,s). \tag{3.3}$$

It is convenient to consider the mass m as the fourth index of the base function, say, $B_{m,t,c,s}(x)$, or any three of them as an index set. In any case we always assume that indexes m, t, c, s satisfy the relation (3.3), i.e., if any three of them are given, the fourth one is decided by the relation.

Consider a Cauchy problem,

$$\begin{aligned} u_t + f(u)_x &= 0, \\ u(x,0) &= B_{m_0,t_0,c_0,s_0}(x). \end{aligned} \tag{3.4}$$

It is already observed in (2.10) that the solution $u(\cdot,t)$ has the self-similarity profile with time index $t+t_0$ between the original center c_0 and a new shock location $s(t)$. Since the initial total mass m_0 should be preserved, the solution of (3.4) is

$$u(x,t) = B_{m_0,t+t_0,c_0}(x), \tag{3.5}$$

where the shock location $x = s(t)$ is decided by the relation (3.3).

Remark 3.1. If we take a δ-function as the initial datum, for example, $u_0(x) = m_0\delta(x-c_0)$, then the solution is $u(x,t) = B_{m_0,t,c_0}(x)$. So the slope of the base function represents the time of the evolution starting from the δ-function-like initial data, and that is why we take index t for the base function.

Remark 3.2. For the Burgers case, $f(u) = u^2/2$, the self-similarity profile is the identity function, $g(x) = x$. In this case, (3.3) gives explicit relations,

$$m = (s-c)^2/(2t), \quad t = (s-c)^2/2m, \quad s = c + \sqrt{2mt}, \quad c = s - \sqrt{2mt}\,. \tag{3.6}$$

Remark 3.3. The rescaling (2.3) does not preserve the total mass. So it can not measure the invariance property for L^1 solutions of conservation laws. For the Burgers case, $f(u) = u^2/2$, we may consider

$$v(x,t) = au(ax, a^2t), \quad a > 0, \tag{3.7}$$

where the rescaling preserves the total mass. We can easily check that variables

$$w = \sqrt{t+t_0}\,u, \qquad \zeta = (x-c_0)/\sqrt{t+t_0}, \qquad \tau = \ln(t+t_0), \tag{3.8}$$

are invariant under the rescaling after the translation $x - c_0 \to x$, $t + t_0 \to t$. These variables are called self-similarity variables for L^1 Cauchy problems, and the Cauchy problem (3.4) is transformed to

$$\begin{aligned} w_\tau + \tfrac{1}{2}(w^2 - \zeta w)_\zeta &= 0, \\ w(\zeta, \ln(t_0)) &= B_{m_0,t_0=1,c_0=0}(\zeta). \end{aligned} \tag{3.9}$$

We can easily check that $B_{m_0,t_0=1,c_0=0}(\zeta)$ is an admissible steady state of the equation, and hence $w(\zeta,\tau) = B_{m_0,t_0=1,c_0=0}(\zeta)$ is the solution of (3.9). If we transform the variables back to u, t, x, then we get $u(x,t) = B_{m_0,t+t_0,c_0=0}(x)$. This is another way to show (3.5). In this example we can see that the approach with piecewise self-similar solutions captures the self-similarity of the general Cauchy problems exactly. For a detailed study of the transformed problem (3.9), we refer to [17].

3.2. S-summation. Since the solution of (3.4) is given by (3.5), we can easily guess that

$$u(x,t) = \sum_{k=1}^{n} B_{m_k,t_k+t,c_k}(x) \tag{3.10}$$

is the solution of the conservation law with initial data

$$u_0(x) = \sum_{k=1}^{n} B_{m_k,t_k,c_k}(x), \quad c_n < \cdots < c_2 < c_1, \tag{3.11}$$

if all the supports of the base functions in (3.10) are disjoint. But it is not usually the case since the support of a base function expands in time. The S-summation,

$$B_n(x) = \bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x), \quad c_n < \cdots < c_2 < c_1, \tag{3.12}$$

is to handle the case that supports of base functions overlap with each other. The definition is given inductively in the following.

Let $B_1(x) = B_{m_1,t_1,c_1}(x)$. Suppose that $B_{j-1}(x) = \bigodot_{k=1}^{j-1} B_{m_k,t_k,c_k}(x)$, $j \le n$, is well defined, $\mathrm{supp}(B_{j-1}) \subset [c_{j-1}, \infty)$, and that $\int_{c_{j-1}}^{\infty} B_{j-1}(x)dx = \sum_{k=1}^{j-1} m_k$. Consider a point $\xi_j \in \boldsymbol{R}$ such that $c_j < \xi_j$,

$$g\big((x-c_j)/t_j\big) > B_{j-1}(x), \qquad c_j < x < \xi_j, \tag{3.13}$$

$$\int_{c_j}^{\xi_j} g\big((x-c_j)/t_j\big)dx + \int_{\xi_j}^{\infty} B_{j-1}(x)dx = \sum_{k=1}^{j} m_k. \tag{3.14}$$

Under assumption (3.13), the left-hand side of (3.14) is monotone in ξ_j and, hence, such a point is unique. If there exists such a point $\xi_j > c_j$, we define

$$B_j(x) = \bigodot_{k=1}^{j} B_{m_k,t_k,c_k}(x) = \begin{cases} g\big((x-c_j)/t_j\big), & c_j < x < \xi_j, \\ B_{j-1}(x), & \text{otherwise.} \end{cases} \tag{3.15}$$

Clearly, $\mathrm{supp}(B_j) \subset [c_j, \infty)$ and $\int_{c_j}^{\infty} B_j(x)dx = \sum_{k=1}^{j} m_k$, and we may continue the inductive argument. If not, the S-summation (3.12) is not defined.

Base functions are ordered by centers c_k, and then the S-summation is given from the right-hand side. It is because of the positiveness assumption for the wave speed, $f'(u) \ge 0$, in (H). If the order of the summation is changed, the result is different. So the S-summation is not associative.

Remark 3.4. If the time indexes are identical, $t_k = t_0$, for all k, then we can show the S-summation (3.12) is well defined. Then, since the self-similarity profile $g(x)$ is an increasing function, we have $g\big((x-c_j)/t_0\big) > g\big((x-c_k)/t_0\big)$ for all $k < j$. Since $B_{j-1}(x)$ has values of $g\big((x-c_k)/t_0\big)$, $k < j$, piecewise, the inequality (3.13) is satisfied for all $\xi_j > c_j$. Furthermore the left-hand side of (3.14) has value $\sum_{k=1}^{j-1} m_k$ for $\xi_j = c_j$ and diverges to ∞ as $\xi_j \to \infty$. So there exists a point ξ_j satisfying (3.14), and the S-summation is well defined.

Remark 3.5. We may consider ξ_j as the location of the jth (artificial) shock generated by the base function B_{m_j,t_j,c_j}. Suppose that $\xi_{j-1} < \xi_j$, i.e., the jth shock caught up the $(j-1)$st shock. The definition (3.15) implies that the self-similarity profile $g\big((x-c_{j-1})/t_{j-1}\big)$ disappears. We can easily check that we will get the same S-summation (3.15) if we remove the $(j-1)$st base function and increase m_j by adding m_{j-1}. This property represents the irreversibility of conservation laws and plays the key role in the numerical scheme (see section 4.2, Step 2).

THEOREM 3.6. *Suppose that* $f''(u) \geq 0$ *and* $f'(u) \geq 0$. *If the S-summation* $u_0(x) \equiv \bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x)$ *is well defined, then* $u(x,t) \equiv \bigodot_{k=1}^{n} B_{m_k,t_k+t,c_k}(x)$ *is also well defined for all* $t > 0$ *and it solves* (1.1) *with its initial value* $u_0(x)$. *If* $v(x,t)$ *is the entropy solution of* (1.1) *with its initial value* $v_0 \in L^1$, *then*

$$||v(\cdot,t) - u(\cdot,t)||_1 \leq ||v_0 - u_0||_1. \tag{3.16}$$

Proof. We may assume $f'(0) = 0$ without the loss of generality. The proof is completed through inductive arguments. In section 2, we have shown the theorem for $n = 1$. Now we show the theorem for $n = j > 1$ assuming that it holds for $n = j - 1$. Note that, from the definition, the S-summation $\bigodot_{k=1}^{i} B_{m_k,t_k,c_k}(x)$ is well defined for any $i \leq n$.

Let $u_{j-1}(x,t)$ be the solution of (1.1) with its initial value $\bigodot_{k=1}^{j-1} B_{m_k,t_k,c_k}(x)$. From the assumption, $u_{j-1}(x,t) = \bigodot_{k=1}^{j-1} B_{m_k,t_k+t,c_k}(x)$. Let $u_j(x,t)$ be the solution with $u_j(x,0) = \bigodot_{k=1}^{j} B_{m_k,t_k,c_k}(x)$ and $x = \xi_j(t)$ be the shock characteristic given by the jth base function, i.e., $\xi_j(0)$ is the same as the ξ_j in (3.13)–(3.14). Consider a backward characteristic, associated with $u_j(x,t)$, that emanates from a point (x,t), $x > \xi_j(t)$. From the admissibility of the shock, it does not interact with $x = \xi_j(\tau)$, $\tau < t$, and, hence, it is actually the one associated with $u_{j-1}(x,t)$. So we have $u_j(x,t) = u_{j-1}(x,t)$.

For $x < c_j$, $u_j(x,t) = 0$ since the (vertical) forward characteristic that emanates from a point $(x,0)$, $x < c_j$, does not intersect with shock characteristics which move to the right-hand side under the assumption $f'(u) \geq 0$. The backward characteristic that emanates from a point (x,t), $c_j < x < \xi_j(t)$, is a straight line connecting $(c_j, -t_j)$ since the initial profile over the interval $(c_j, \xi_j(0))$ is self-similar. Hence, $u_j(x,t) = g\big((x-c_j)/(t+t_j)\big)$ for $c_j < x < \xi_j(t)$, and the shock location $x = \xi_j(t)$ should satisfy

$$\int_{c_j}^{\xi_j(t)} g\big((x-c_j)/(t+t_j)\big)dx + \int_{\xi_j(t)}^{\infty} u_{j-1}(x,t)dx = \sum_{k=1}^{j} m_k$$

since the total mass is preserved. So $u_j(x,t) = \bigodot_{k=1}^{j} B_{m_k,t_k+t,c_k}(x)$ from the definition of the S-summation, and the first part of the proof is complete. The second part (3.16) is simply the L^1 contraction theory for conservation laws. ∎

In the proof we employ the theory of characteristics (see [9, Chap. 11]). The error estimate (3.16) implies that the initial error decreases in time. In fact, the error is of order $O(t^{-1})$ as $t \to \infty$ (see [16] for detail). The scheme has ideal properties for the study of asymptotic behavior.

Now we consider $u_0(x) = \bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x)$ as an approximation of L^1 initial value v_0. Let a partition $\mathcal{C} = \{c_n < \cdots < c_1\}$ be the set of centers. Its norm is defined by $||\mathcal{C}|| = \max |c_k - c_{k-1}|$. There can be many ways to discretize the initial value. To guarantee the convergence of the scheme, we need the existence of $\delta, L > 0$ such that

$$||v_0(x) - u_0(x)||_1 \leq \varepsilon \quad \text{if} \quad ||\mathcal{C}|| \leq \delta \text{ and } c_n < -L, L < c_1, \tag{3.17}$$

where a constant $\varepsilon > 0$ is given. An example of such a discretization is given in section 4.2. The convergence of the scheme satisfying (3.17) is clear from (3.16).

COROLLARY 3.7 (convergence). *The scheme of the S-summation* $u(x,t) = \bigodot_{k=1}^{n} B_{m_k,t+t_k,c_k}(x)$ *with initial discretization* $u_0(x) = \bigodot_{k=1}^{n} B_{m_k,t_k,c_k}(x)$ *satisfying* (3.17) *converges to the entropy solution* $v(x,t)$ *with initial data* $v_0 \in L^1(\boldsymbol{R})$ *as* $\delta \to 0$, $L \to \infty$.

Remark 3.8. Now we consider the S-summation between two base functions, $\bigodot_{k=1}^{2} B_{m_k,t_k,c_k}(x)$, $c_2 \le c_1$ (see Figure 3.1). It gives a good example for figuring out the meaning of the S-summation. Furthermore, in the numerical computation, we can possibly compare only two base functions each time and, hence, it is worth considering it in detail. If these two base functions are separated, i.e., $s_2 < c_1$, then the shock place ξ of the definition (3.15) is simply $\xi = s_2$. If $c_1 < s_2$, then ξ satisfies

$$\int_{c_2}^{\xi} g\Big(\frac{x-c_2}{t_2}\Big)dx + \int_{\xi}^{\max(\xi,s_1)} g\Big(\frac{x-c_1}{t_1}\Big)dx = m_1 + m_2. \tag{3.18}$$

If $\xi > s_1$, (3.15) implies that two base functions are merged, i.e.,

$$\bigodot_{k=1}^{2} B_{m_k,t_k,c_k}(x) = B_{m_1+m_2,t_2,c_2}(x) \qquad \text{if} \quad s_1 < \xi. \tag{3.19}$$

For the Burgers case, $f(u) = u^2/2$, (3.18) implies that the trapezoid $BCs_2\xi$ in Figure 3.1 has the same area as the triangle $Ac_1\xi$.

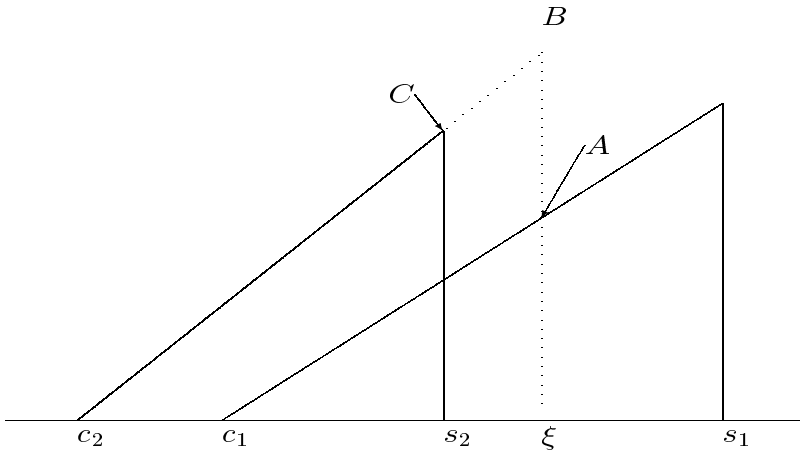


FIG. 3.1. *The equal area rule gives the shock location when two base functions interact together.*

4. S-summation as a numerical scheme. In this section we show how the S-summation can be implemented into a numerical scheme. We assume that the flux is convex $f''(u) \ge 0$ and the solution is positive and compactly supported. More general cases are considered in the following sections. To see what is really happening in each step, we consider a Cauchy problem for the Burgers equation,

$$\begin{array}{c} v_t + vv_x = 0, \\ v(x,0) = v_0(x), \end{array} \qquad v_0(x) = \begin{cases} \sin(\pi x)/\pi, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

This simple example helps us to visualize the mechanism of the scheme. In section 4.3 we consider more complicated examples and compare CPU times of each computation to check the complexity of the method which is of order $O(N)$. Several properties of this scheme are compared with those of the Godunov method.

4.1. Preliminaries. First, we consider basic properties of the self-similarity profile $g(x)$, $x > 0$. If the flux is strictly convex, $f''(u) > 0$, then $f'(u)$ is invertible and the self-similarity profile is simply the inverse function of $f'(u)$. For example, if the flux is given by a power law

$$f(u) = \frac{1}{\gamma}\, u^{\gamma}, \qquad u \geq 0,\ \gamma > 1, \tag{4.2}$$

then the self-similarity profile $g(x)$ is simply

$$g(x) = \sqrt[\gamma-1]{x}, \qquad x > 0. \tag{4.3}$$

This is a case that the self-similarity profile is given explicitly. In general, the value of the self-similarity profile $\bar{u} = g(\bar{x})$ at a given point $\bar{x} \geq 0$ is obtained from the basic relation $f'(g(\bar{x})) = \bar{x}$, i.e., we need to solve

$$f'(\bar{u}) - \bar{x} = 0, \qquad \bar{u} \geq 0. \tag{4.4}$$

The relation between the self-similarity profile and the wave speed, $f'(g(x)) = x$, also makes it easy to handle the integrals of a base function. Let $\bar{u} = g(\bar{x}/t)$. Using the change of variables $u = g(x/t)$, we obtain

$$\int_0^{\bar{u}} t f'(u)du = \int_0^{\bar{x}} \frac{x}{t} g'\Big(\frac{x}{t}\Big)dx = \bar{x} g\Big(\frac{\bar{x}}{t}\Big) - \int_0^{\bar{x}} g\Big(\frac{x}{t}\Big)dx.$$

So the integral of the self-similarity profile is written as a function of $\bar{u}$ or $\bar{x}$ only, i.e.,

$$\int_0^{\bar{x}} g\Big(\frac{x}{t}\Big)dx = \bar{x} g\Big(\frac{\bar{x}}{t}\Big) - t f\Big(g\Big(\frac{\bar{x}}{t}\Big)\Big) = t\bar{u} f'(\bar{u}) - t f(\bar{u}). \tag{4.5}$$

Now we consider a simple lemma which is used in deciding the initial time index $t_0 > 0$. This lemma implies that the graph $y = g((x-c)/t_0)$, $x > c$, crosses over $y = v_0(x)$ just once.

LEMMA 4.1. *Suppose that the (smooth and bounded) initial value $v_0(x)$ satisfies*

$$v_0'(x) < \frac{1}{t_0 f''(v_0(x))}. \tag{4.6}$$

Then the point $\bar{x} \geq c$ satisfying $g((\bar{x}-c)/t_0) = v_0(\bar{x})$ is unique.

Proof. Differentiating both sides of $f'(g(x)) = x$, we obtain

$$g'(x) = \frac{1}{f''(g(x))} = \frac{1}{f''(v)}, \tag{4.7}$$

where $v = g(x)$. Let $g((x_1-c)/t_0) = v_0(x_1)$ for a point $x_1 \geq c$. Since

$$v_0'(x_1) < \frac{1}{t_0 f''(v_0(x_1))} = \frac{1}{t_0} g'\Big(\frac{x_1-c}{t_0}\Big) = \partial_x g\Big(\frac{x_1-c}{t_0}\Big), \tag{4.8}$$

we may choose $\delta > 0$ such that $g((x-c)/t_0) > v_0(x)$ for $x \in (x_1, x_1+\delta)$. Now we show that $g((x-c)/t_0) > v_0(x)$ for all $x > x_1$, which completes the proof. Suppose that $g((x_2-c)/t_0) = v_0(x_2)$ for $x_2 > x_1$. We may take x_2 as the smallest one. Then $g((x-c)/t_0) > v_0(x)$ on (x_1, x_2) and it implies $v_0'(x_2) \geq \partial_x g((x_2-c)/t_0)$. It contradicts the fact that (4.8) holds for $x = x_2$. ∎

4.2. Implementation. Here we introduce a gridless scheme based on the S-summation.

Step 1 (initial discretization). The first step is to design a method to approximate the initial value $v_0(x)$ by an S-summation $u_0(x)$ which satisfies (3.17). Consider n base functions $B[k]$, $k = 1, 2, \dots, n$. Each element $B[k]$ consists of two members $B[k].m, B[k].c$, which represent the mass (or area) and the center of the base function. We use the identical time index $t_k = t_0$ for all k, and, hence, we do not need an extra member for the time index. The first thing to do is to choose the time index $t_0 > 0$ satisfying (4.6). If (4.6) does not hold for any $t_0 > 0$, we need to use a different discretization (see section 6).

Next we decide the two members of the kth base function, $B[k].c$ and $B[k].m$. Let $[L_1, L_2]$ be the support of the initial value $v_0(x)$ and $L_1 = x_n < \cdots < x_1 < x_0 = L_2$ be mesh points. Consider the self-similarity profiles that emanate from points $(x_k, v_0(x_k))$. Then the center c_k satisfies $g((x_k - c_k)/t_0) = v_0(x_k)$. Taking the wave speed f' to both sides we get $c_k = x_k - t_0 f'(v_0(x_k))$. We assign this center to $B[k].c$, i.e.,

$$B[k].c = x_k - t_0 f'(v_0(x_k)). \tag{4.9}$$

Since $g((\bar{x} - c_1)/t_0) = g((\bar{x} - c_2)/t_0)$ at any point $\bar{x} > c_1, c_2$ implies $c_1 = c_2$, we can easily see that two self-similarity profiles with the same time index never cross over to each other. So Lemma 4.1 implies that those centers are ordered by $L_1 = B[n].c < \cdots < B[2].c < B[1].c$. Note that there is no self-similarity profile that emanates from the point $(x_0, v_0(x_0))$.

The value of the second member $B[k].m$ is given as the area enclosed by four (or three) curves, $y = v_0(x)$, $y = 0$, $y = g((x - c_k)/t_0)$, and $y = g((x - c_{k-1})/t_0)$, i.e.,

$$B[k].m = \int_{x_k}^{x_{k-1}} v_0(x)dx + \int_{c_k}^{x_k} g\Big(\frac{x - c_k}{t_0}\Big)dx - \int_{c_{k-1}}^{x_{k-1}} g\Big(\frac{x - c_{k-1}}{t_0}\Big)dx,$$

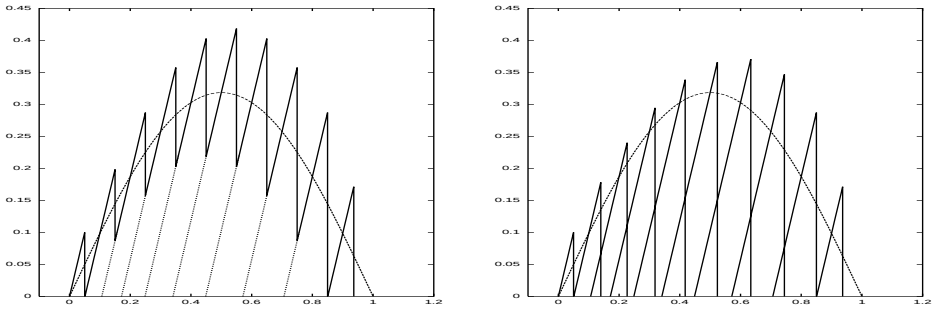
where $c_0 = L_2$. Using relation (4.5), this is written in terms of the initial value and the flux:

$$\begin{aligned} B[k].m = \int_{x_k}^{x_{k-1}} v_0(x)dx + t_0 v_0(x_k) f'(v_0(x_k)) - t_0 f(v_0(x_k)) \\ - t_0 v_0(x_{k-1}) f'(v_0(x_{k-1})) + t_0 f(v_0(x_{k-1})). \end{aligned} \tag{4.10}$$

Consider the Cauchy problem (4.1) as an example. Since $f''(v) = 1$ and $v_x(x, 0) \le 1$, we may take any $t_0 < 1$. In the following examples we use $t_0 = 0.5$. In Figure 4.1(a), 10 self-similarity profiles are shown which emanate from 10 points $(j/10, v_0(j/10))$, $j = 0, 1, \dots, 9$. The centers are their x-intercepts.

In Figure 4.1(b), 10 base functions are displayed with initial value $v_0(x)$. Supports of these base functions are overlapped with each other. Their S-summation $u_0(x) = \bigodot_{k=1}^{n} B[k]$ is considered as the initial discretization, which is the saw-tooth profile (solid lines) in Figure 4.1(a). Let u_0^ε be such an approximation with a uniform mesh size $x_{k-1} - x_k = \varepsilon$. Then the sizes of the triangle-like areas in Figure 4.1(a), added to and subtracted from the area enclosed by $y = v_0(x)$ and the x-axis, are proportional to ε^2, and the total number of them has order $O(1/\varepsilon)$. So we have $||v_0 - u_0^\varepsilon||_1 = O(\varepsilon)$ as $\varepsilon \to 0$, where $u_0^\varepsilon(x) = \bigodot_{k=1}^{n} B[k]$ with $t_k = t_0$ for all k. (Step 1 is complete.)

Theorem 3.6 implies that $u(x, t) = \bigodot_{k=1}^{n} B[k]$ with $t_k = t_0 + t$ is the solution with the modified initial data u_0. So the rest of the scheme is focused on how to



(a) *The equal area rule* (b) *Base functions with overlaps*

FIG. 4.1. *The initial value is approximated by a piecewise self-similar function, which is a saw-tooth profile in* (a). *It turns out to be an S-summation of base functions in* (b).

display the given solution. Even if it is possible to follow the inductive arguments of the definition, we will get serious complexity in the coding if behind shocks capture the front ones, $\xi_{j-1} < \xi_j$, where ξ_j is the shock location generated by the jth base function (see the definition (3.13)–(3.15)). Since the S-summation is not changed even if these two base functions are merged before the summation (see Remark 3.5), we consider the merging process first.

Suppose that $\bigodot_{k=1}^{j-1} B[k]$ is achieved and $\xi_{j-1} < \xi_{j-2} < \cdots < \xi_1$. To obtain $\bigodot_{k=1}^{j} B[k]$ we need to check if $\xi_j < \xi_{j-1}$. Since $\xi_{j-1} \neq s_{j-1}$ in general, an equation corresponding to (3.18) does not provide the information we need. In the following we define an operator using a modified version of (3.18).

DEFINITION 4.2. *We define a binary operator "$*$" between two base functions B_{m_k,t_k,c_k,s_k}, $k = 1, 2$, satisfying $c_2 < c_1$. First, if $s_2 \le c_1$, we define $B_{m_2,t_2,c_2} * B_{m_1,t_1,c_1} \equiv s_2$. If $c_1 < s_2$, $B_{m_2,t_2,c_2} * B_{m_1,t_1,c_1}$ $(\equiv \xi)$ is defined as the solution of*

$$F(\xi) \equiv \int_{c_2}^{\xi} g\Big(\frac{x-c_2}{t_2}\Big)dx - \int_{c_1}^{\xi} g\Big(\frac{x-c_1}{t_1}\Big)dx - m_2 = 0. \tag{4.11}$$

Let $\xi = B[j] * B[j-1]$. From (3.14) we can clearly see that $\xi = \xi_j$ if and only if $\xi \le \xi_{j-1}$. If $\xi_{j-1} < \xi$, we also have $\xi_{j-1} < \xi_j$ and we may merge two base functions, $B[j]$ and $B[j-1]$, before the S-summation. On the other hand, since we have assumed $\xi_{j-1} < \xi_{j-2} < \cdots < \xi_1$, we have $\xi_{j-1} = B[j-1] * B[j-2]$. So we may conclude that

$$\xi_j > \xi_{j-1} \quad \text{if and only if} \quad B[j] * B[j-1] > B[j-1] * B[j-2]. \tag{4.12}$$

So this operator gives the criterion for deciding if two base functions should be merged together or not. Furthermore, after the merging process, it gives the correct (artificial) shock locations $\xi_j(t)$ for the S-summation.

Step 2 (merging). In this step base functions are re-indexed for $k = 1, 2, \ldots, n'$ whenever two base functions are merged together and the total number of base functions is decreased. Suppose that this merging procedure has been completed for all $k < j$ and $\xi_{j-1} < \cdots < \xi_2 < \xi_1 = s_1$ holds. Then $\xi_k = B[k] * B[k-1]$ for $k = 2, \ldots, j-1$. Now we check the next base function $B[j]$.

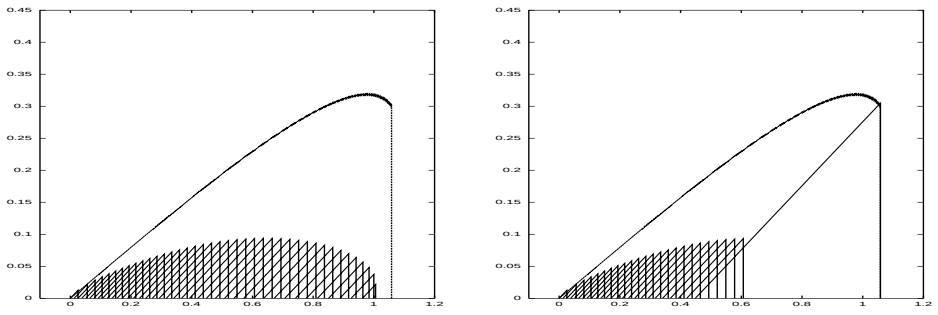
If $B[j] * B[j-1] < B[j-1] * B[j-2]$, then $\xi_j = B[j] * B[j-1]$ and this step is completed for $k \le j$. Suppose that $B[j] * B[j-1] > B[j-1] * B[j-2]$. Then (4.12)

implies that $\xi_j > \xi_{j-1}$, and we may merge $B[j]$ and $B[j-1]$ (see Remark 3.5). Put

$$B[j].m = B[j].m + B[j-1].m, \tag{4.13}$$

remove $B[j-1]$, and then rearrange the array $B[\cdot]$ from $k=1$ to $k=n'-1$, where n' is the total number of base functions left after the previous step. Since the combined base function may take over another one again, we continue this process until we get $\xi_j < \xi_{j-1}$ or $j=1$, decreasing the index j by 1. We continue this procedure from $j=2$ to $j=n'$. Note that there is no base function $B[0]$ and we use a convention $B[1]*B[0] := B[1].s$ in (4.12) for $j=2$, where $B[1].s$ is the shock location of the base function given by the relation (3.3).

In Figure 4.2(a), 40 base functions $B_{m_k,c_k,t+t_0}$, $k=1,\dots,40$, are given at $t=1.5$ together with the exact solution we want to find. During the merging step, Step 2, 16 of them are merged together and a big base function emerges. The location and the size of the discontinuity of the newborn base function are almost identical to those of the physical shock. This big base function can be considered as an accumulation of small artificial shocks, and it represents the physical shock. (Step 2 is complete.)



(a) 40 *base functions before merging* (b) 25 *base functions left after merging*

FIG. 4.2. 40 *base functions have the slope* $1/(t_0+t)$ *at time* $t>0$, *which is* 0.5 *in* (a). *After the merging process, Step* 2, 16 *of them are merged together and a big base function emerges in* (b). *The outside wave is the exact solution we want to approximate.*

Remark 4.3. In the previous algorithm, we solve (4.11) instead of doing the time marching. This relation gives the correct location of artificial shocks if the merging step is completed. Using the relation (4.5), the function $F(\xi)$ in (4.11) is written as

$$F(\xi) \equiv (\xi-c_2)g\Big(\frac{\xi-c_2}{t_2}\Big) - t_2 f\Big(g\Big(\frac{\xi-c_2}{t_2}\Big)\Big) - (\xi-c_1)g\Big(\frac{\xi-c_1}{t_1}\Big) + t_1 f\Big(g\Big(\frac{\xi-c_1}{t_1}\Big)\Big) - m_2. \tag{4.14}$$

So we can simplify the integral equation (4.11). To find the zero of $F(\xi)$ we may use the bisection method. If $B[j]*B[j-1] > B[1].s$, clearly we need to merge $B[j]$ and $B[j-1]$. So we may use $(B[j-1].c, B[1].s)$ as the initial interval.

If we use Newton's method, we need to study the structure of the self-similarity profile first. Let $t_1 = t_2$ $(\equiv t)$. The first two derivatives of $F(\xi)$ are

$$F'(\xi) = g\Big(\frac{\xi-c_2}{t}\Big) - g\Big(\frac{\xi-c_1}{t}\Big), \qquad F''(\xi) = \frac{1}{t}\Big(g'\Big(\frac{\xi-c_2}{t}\Big) - g'\Big(\frac{\xi-c_1}{t}\Big)\Big).$$

Since the self-similarity profile $g(x)$ is an increasing function, we have $F'(\xi) > 0$ and (4.11) has a unique solution. On the other hand, since there is no monotonicity on $g'(x)$ in general, we need to consider the structure $g(x)$ for the initial guess.

Remark 4.4. With the power law $f(u) = u^\gamma/\gamma$, $u > 0$, and the identical time index $t_1 = t_2 \equiv t$, (4.11) is written as

$$F(\xi) = (\xi - c_2)^{\frac{\gamma}{\gamma-1}} - (\xi - c_1)^{\frac{\gamma}{\gamma-1}} - \frac{\gamma}{\gamma-1} m_2 t^{\frac{1}{\gamma-1}} = 0.$$

For the Burgers case, $\gamma = 2$, the operator is explicitly given by

$$B_{m_2,t,c_2} * B_{m_1,t,c_1} = \frac{2m_2 t + c_1^2 - c_2^2}{2(c_1 - c_2)}. \tag{4.15}$$

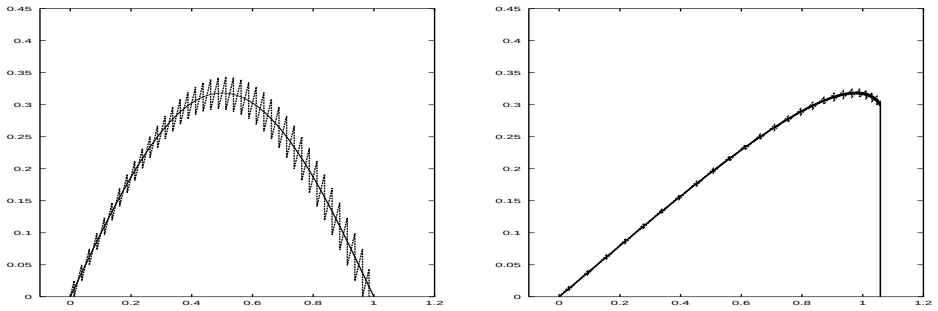
Remark 4.5. If there is no base function merged, there will be $n-1$ comparisons of (4.12). If m base functions are merged, then $n-m$ base functions are left and the maximum number of comparisons (4.12) is $n+m-1 < 2n$, which is of order $O(N)$.

Step 3 (displaying). Now we are ready to display the solution. Suppose that base functions $B[j]$, $j = 1, \ldots, n'$, are left after the merging step. Let $\xi_j = B[j] * B[j-1]$. Then the right- and the left-hand side limits are

$$\begin{aligned} u(\xi_j+, T) &= g\big((\xi_j - B[j-1].c)/(t+t_0)\big), \\ u(\xi_j-, T) &= g\big((\xi_j - B[j].c)/(t+t_0)\big). \end{aligned} \tag{4.16}$$

So to display the solution it is enough to plot the points $(\xi_j, u(\xi_j+, T)), (\xi_j, u(\xi_j-, T))$ for $j = 1, \ldots, n'$. Between these points the solution has the self-similarity profile. So if we connect these points with the self-similarity profile with time index $t + t_0$ and center $B[j].c$, we get the solution.

In Figure 4.3(b), the S-summation of the 25 base functions at time $t = 1.5$ (see Figure 4.2(b)) has been displayed. We may observe that the exact solution passes through the artificial discontinuities of the approximation. Furthermore, we can clearly see that the initial error $||v_0(x) - u_0(x)||_1$ has been decreased a lot. (Step 3 is complete.)



(a) *Initial discretizations* (b) *Solutions at* $t = 1.5$

FIG. 4.3. *The initial value of the problem* (4.1) *has been discretized using* 40 *base functions in* (a). 25 *base functions are left after the merging step with* $t = 1.5$*, Figure* 4.2(b)*, and the S-summation gives the final approximation for the solution. We may observe that the exact solution, which has obtained using finer mesh points, passes through each of the artificial shocks.*

Remark 4.6. One of the main features of the scheme introduced is that it has a complexity of order $O(N)$. Even though we have introduced extra complexities for solving (4.4) and (4.11), this does not increase the order of the complexity. On the other hand, for the convenience of the explanation, we have rearranged the whole array of base functions whenever one of them is merged to another. Since this rearranging

TABLE 4.1

CPU time comparison: Computations for (4.17) with $t_0 = \pi/2$, $t = 4.0$. The number of base functions used initially is N, and L of them are left after the merging step. CPU times for each step of the scheme are compared.

N	L	Discretization	Merging	Displaying	Total	Order α
10	8	0	0	1	1	
100	64	0	5	4	9	0.954
1000	628	0	57	34	91	1.005
10000	6272	2	617	357	976	1.030
100000	62707	22	7763	3956	11741	1.080

process will increase the order of the complexity, we need to use a different strategy in the actual computation. We may link the base functions pointing the adjacent ones so that one of them can be easily removed. These kinds of techniques are classical and we omit the details.

4.3. CPU time comparison. In this section we consider several numerical examples and show the CPU time for each case. In fact, the Burgers equation is the case that the self-similarity profile $g(x)$ and the binary operator "$*$" between base functions are given explicitly, (4.3) and (4.15). So that case does not show the complexity of the scheme well. In the examples in this section, we numerically solve (4.4) and (4.11) using Newton's method.

First, consider a Cauchy problem with the cubic law, $f(v) = v^3/3$,

$$\begin{array}{c} v_t + v^2 v_x = 0, \\ v(x,0) = v_0(x), \end{array} \qquad v_0(x) = \begin{cases} \sin(\pi x)/\pi, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.17}$$

In this case the self-similarity profile $g(x)$ is concave. Since $v_0'(x) = \cos(\pi x)$ and $f''(v) = 2v$, the condition (4.6) is written as

$$\cos(\pi x) < \frac{\pi}{2t_0 \sin(\pi x)}, \qquad 0 < x < 1.$$

We can easily check that it is satisfied for $t_0 = \pi/2$. In Table 4.1 we have compared the CPU time of the computations as we increase the number of base functions (or mesh points). The solution is computed for time $t = 4$.

Suppose that the CPU time $T(N)$ for the computation with N mesh points is $T(N) = cN^\alpha$ for some constants $c, \alpha > 0$. Then we can easily check that

$$\alpha = \frac{\ln(T(N_1)/T(N_2))}{\ln(N_1/N_2)}. \tag{4.18}$$

This number represents the complexity order of the scheme, and it is computed and shown in Table 4.1. We may observe that the order is about $\alpha = 1.08$. These computational results confirm that the complexity of the scheme is almost linear. The extra growth in the CPU time is caused by Newton's method. If we use finer base functions, we need to use smaller tolerances in finding the shock location.

Next we consider a problem with the flux $f(u) = \frac{2}{3}u^{3/2}$, where the self-similarity profile $g(x)$ is convex. In this case we cannot find the time index $t_0 > 0$ that satisfies (4.6) for the initial value given in the previous example. So we consider

$$\begin{array}{c} v_t + \sqrt{v}\, v_x = 0, \\ v(x,0) = v_0(x), \end{array} \qquad v_0(x) = \begin{cases} 5x^2(x-1)^2, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.19}$$

TABLE 4.2

CPU time comparison: Computations for (4.19) *with* $t_0 = 0.1$, $t = 1.0$. *The number of base functions used initially is N, and L of them are left after the merging step. CPU times for each step of the scheme are compared.*

N	L	Discretization	Merging	Displaying	Total	Order α
10	7	0	1	1	2	
100	62	0	13	7	20	1.000
1000	610	0	209	92	301	1.178
10000	6092	1	2785	1113	3899	1.112
100000	60914	12	34068	13105	47185	1.083

TABLE 4.3

CPU time comparison: Computations for (4.20) *with* $t_0 = 0.1$, $t = 0.5$. *The number of base functions used initially is N, and L of them are left after the merging step. CPU times for each step of the scheme are compared.*

N	L	Discretization	Merging	Displaying	Total	Order α
10	7	0	1	0	1	
100	60	0	7	5	12	1.079
1000	589	1	113	53	167	1.144
10000	5878	1	1635	633	2269	1.133
100000	58767	13	19965	7114	27092	1.077

Since $v_0'(x) = 10x(x-1)(2x-1)$ and $f''(v) = 1/2\sqrt{v}$ in this case, the condition (4.6) is written as

$$-10(2x-1) < 2/t_0, \qquad 0 < x < 1.$$

It is satisfied for $t_0 < 0.2$, and we choose $t_0 = 0.1$. The solution has been computed at time $t = 1$, and their CPU times and the complexity of the scheme have been compared in Table 4.2. We observe a similar complexity order, $\alpha = 1.083$, as we do in the previous example.

As the last example, we consider a combination of three power laws,

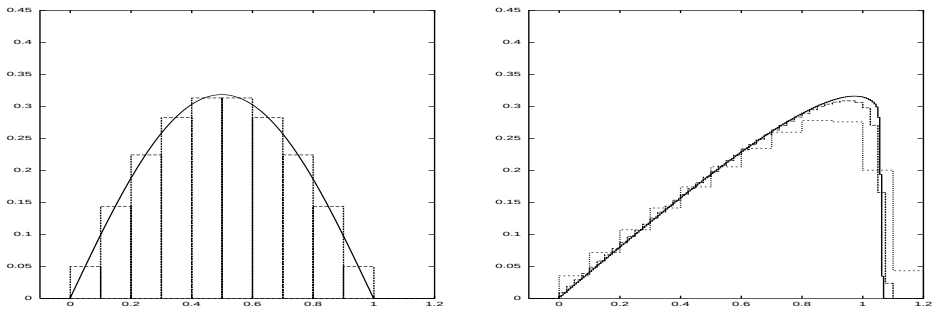
$$(4.20) \qquad \begin{array}{c} v_t + (\sqrt{v} + v + v^2)\, v_x = 0, \\ v(x,0) = v_0(x), \end{array} \qquad v_0(x) = \begin{cases} 5x^2(x-1)^2, & 0 < x < 1. \\ 0, & \text{otherwise.} \end{cases}$$

The solution has been computed at time $t = 0.5$ using an initial time index $t_0 = 0.1$. Their CPU times and the complexity of the scheme have been compared in Table 4.3. We observe a complexity order $\alpha = 1.077$ which is similar to the previous examples.

4.4. Comparison with Godunov. A typical way to discretize the initial data is to take the cell average (see Figure 4.4(a)). The Godunov scheme solves a series of Riemann problems between each cell for a short amount of time Δt and then repeats the process until it reaches a given time $t > 0$. In Figure 4.4(b) we can see that the numerical solution converges to the same limit as the S-summation shown in Figure 4.3(b), as $\Delta x \to 0$.

Remark 4.7 (computation time). Let N be the number of mesh points. Then the number of operations for the S-summation is of order $O(N)$ since the time marching process is not required, Theorem 3.6. The number of operations is almost independent from the final time $t > 0$. On the other hand, the Godunov scheme has operations of order $O(N^2)$ and the situation becomes worse if the final time t is increased.

Remark 4.8 (error estimate). We can clearly observe that the exact solution v of (4.1) (or $||\mathcal{C}|| \to 0$ limit of the S-summation) passes through artificial shocks of



(a) *Data discretization* (b) *Solutions at* $t = 1.5$

FIG. 4.4. *Three approximations by Godunov using* $\Delta x = 1/10, 1/40, 1/160$. *The scheme is convergent to the same limit of the S-summation. We can observe that numerical solutions are separated near the shock, and it is hard to guess where the limit is from a single computation.*

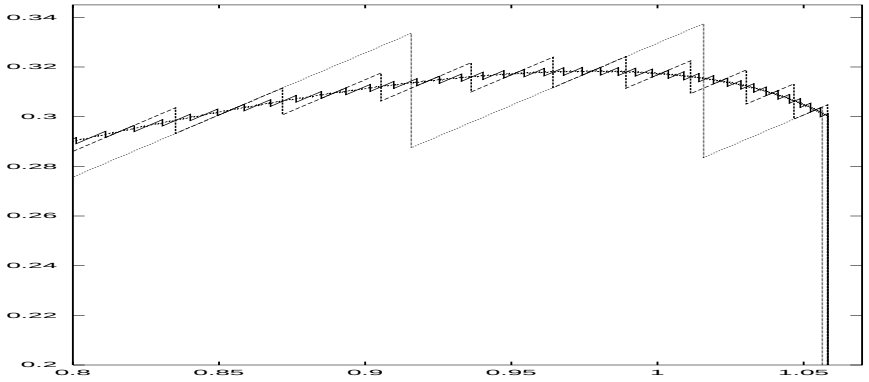


FIG. 4.5. *A magnification of Figure* 4.3(b) *near the physical shock shows that an S-summation with a finer mesh passes through the middle of the artificial shocks.* 10*,* 40*, and* 160 *base functions are used.*

self-similarity solutions (see Figure 4.3(b)). In Figure 4.5 a part of Figure 4.3(b) near the physical shock is magnified together with other similarity summations consisting of 10 and 160 base functions. In this figure we can also observe that S-summations are attached to each other in the middle of self-similar profiles. Noting that the sizes of artificial shocks decrease in time with order of $O(1/(t + t_0))$, these observations show the possibility for a good error estimate.

Remark 4.9 (shock appearance time). In a numerical scheme the solution is approximated by piecewise continuous functions, and it is hard to see if a discontinuity represents the physical shock or not. In our scheme, as we can see from Figure 4.2, the accumulation of base functions represents the physical shock. So, if a base function is merged to its behind one in the sense of (4.12), we may conclude that a physical shock has appeared. The physical shock appears at time $t = 1$ in the example (4.1) since $\min(\partial_x v_0(x)) = -1$. We can easily check whether (4.12) happens around that time. Table 4.4 shows the time when the number of initial base functions decreases by one.

TABLE 4.4

Shock appearance time. The exact solution with initial data (4.1) *blows up at* $t = 1$. *The time of shock appearance can be measured by counting the base functions after the merging step.*

Initial number of base functions	The time when the number is decreased by 1
25	$T = 1.02$
50	$T = 1.005$
100	$T = 1.0015$
200	$T = 1.0008$
400	$T = 1.0002$
800	$T = 1.00005$

5. General cases. The S-summation has been considered under hypothesis (H). In this section we generalize it under hypotheses (H1) and (H2).

5.1. General convex flux. We consider L^1 initial function u_0 which is uniformly bounded, say, $-A \le u_0(x) \le B$. Then the solution of (2.1) is bounded above and below:

$$-A \le u(x,t) \le B, \qquad A, B \in \boldsymbol{R}^+.$$

Consider a general convex flux, i.e.,

$$f''(u) \ge 0. \tag{H1}$$

If the flux satisfies $f''(u) \le 0$, we may change the variable $y = -x$ and get an equation $u_t + \bar{f}(u)_y = 0$ with $\bar{f}(u) = -f(u)$, where $\bar{f}$ satisfies (H1). Note that we include the equality in (H1) and a piecewise linear flux can be considered.

We can easily check that a new flux,

$$h(w) = f(w - A) - f'(-A)w - f(-A), \tag{5.1}$$

satisfies the hypothesis (H) and $h'(0) = 0$. Let $w(x,t)$ be the solution of

$$w_t + h(w)_x = 0, \qquad w(x,0) = u_0(x) + A. \tag{5.2}$$

We can easily check that

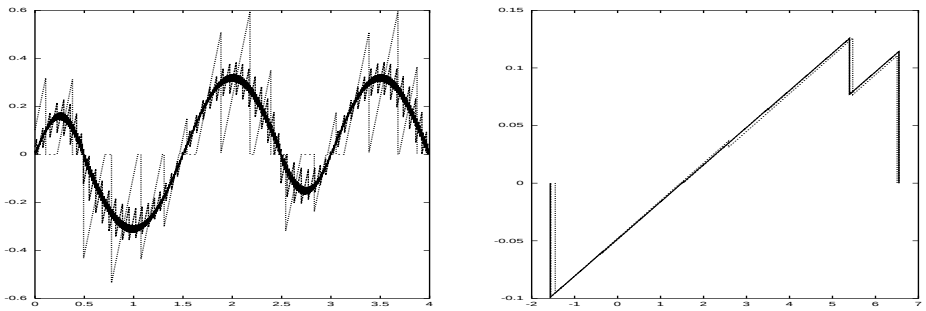
$$u(x,t) = w(x - f'(-A)t, t) - A \tag{5.3}$$

is the solution with the original flux f and initial data u_0. Since $u \ge -A$, the solution $w(x,t)$ is positive. Now we are in the exact same situation as in the previous sections, except with respect to the structure of the initial data. The initial data $w(\cdot, 0)$ is not L^1 anymore. To handle the situation, we consider two special base functions with infinite mass,

$$B_{t,c=-\infty,s}(x) = \begin{cases} A, & x < s + t\,h(A)/A, \\ 0, & x > s + t\,h(A)/A, \end{cases}$$

$$B_{t,c,s=\infty}(x) = \begin{cases} \max\bigl(g(\frac{x-c}{t}), A\bigr), & x > c, \\ 0, & x < c, \end{cases}$$

where $h'(g(x)) = x$, i.e., $g(x)$ is the similarity profile under the flux $h(w)$, not $f(u)$. These base functions handle the transformation $u_0(x) \to u_0(x) + A$. Note that the speed of the shock connecting the state $w = A$ and $w = 0$ is $h(A)/A$ in our case. The S-summation including these two base functions can be defined in a similar way. We omit the details. Figure 5.1 shows how the self-similar solution evolves for the Burgers case. In the figure even the solution with very rough initial discretization with only 16 base functions represents the asymptotic behavior very correctly.



(a) *Data discretization* (b) *Solutions at* $t = 30$

FIG. 5.1. *Three S-summations are displayed using* 16, 64, *and* 256 *base functions. It handles sign changing solutions correctly. This figure shows the time convergence to an inviscid N-wave.*

5.2. Flux without convexity. Consider a flux with a single inflection point:

$$f''(u) \le 0 \text{ for } u \le A, \quad f''(u) \ge 0 \text{ for } u \ge A. \tag{H2}$$

Then, under the change of variables,

$$h(w) = f(w + A) - f'(A)w - f(A), \qquad u(x,t) = w(x - f'(A)t, t) + A, \tag{5.4}$$

the problem (2.1) is transformed to

$$w_t + h(w)_x = 0, \qquad w(x,0) = u_0(x) - A.$$

Then the new flux h satisfies

$$h''(w) \le 0 \text{ for } w \le 0, \quad h''(w) \ge 0 \text{ for } w \ge 0, \quad h'(w) \ge 0 \text{ for all } w, \tag{5.5}$$

and $h(0) = h'(0) = h''(0) = 0$. Since A is not the lower bound of the solution $u(x,t)$ in general, we cannot expect $w \ge 0$. So in this case we have to consider the positive part and the negative part together. It is possible since $h'(w)$ is monotone on $(-\infty, 0)$ and $(0, \infty)$, respectively. All we have to do is to consider negative base functions together with positive ones. Since the wave speed $h'(w)$ is positive, the S-summation is defined from the right-hand side as in the previous cases.

Example 5.1. Consider an inviscid thin film flow in [1],

$$\begin{aligned} u_t + (u^2 - u^3)_x &= 0, \\ u(x,0) &= u_0(x), \end{aligned} \tag{5.6}$$

where the initial datum is compactly supported $\mathrm{supp}(u_0) \subset [L_1, L_2]$. The flux $f(u) = u^2 - u^3$ has a single inflection point $A = 1/3$ and, under the transformation (5.4), we get the flux $h(w) = -w^3$. It satisfies

$$h''(w) \ge 0 \text{ for } w \le 0, \qquad h''(w) \le 0 \text{ for } w \ge 0, \qquad h'(w) \le 0 \text{ for all } w,$$

which is not exactly the same as (5.5) but has the opposite direction in the inequalities. We may do the S-summation from the left-hand side instead of changing the space variable using $y = -x$. Now the original problem (5.6) is transformed to

$$\begin{aligned} w_t - (w^3)_x &= 0, \\ w(x,0) = w_0(x) &:= u_0(x) - A. \end{aligned} \tag{5.7}$$

In this case the self-similarity profile (2.8) is given by

$$g_{\pm}(x) = \pm\sqrt{-x/3}\,, \qquad x < 0, \tag{5.8}$$

and the corresponding base functions are

$$B^{\pm}_{t,s,c}(x) = \begin{cases} g_{\pm}((x-c)/t), & s < x < c, \\ 0, & \text{otherwise.} \end{cases} \tag{5.9}$$

The initial data $v_0(x)$ converges to $-A$ as $x \to \pm\infty$, and we need to consider two base functions with infinite mass that are

$$B_{t,s=L_2,c=\infty}(x) = \begin{cases} -A, & x > L_2 + t\,h(-A)/(-A), \\ 0, & x < L_2 + t\,h(-A)/(-A), \end{cases}$$

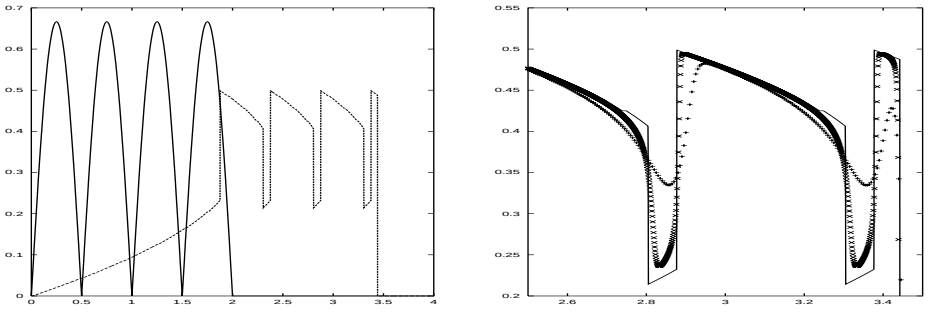
$$B_{t,s=-\infty,c=L_1}(x) = \begin{cases} \max\bigl(g_-((x-L_1)/t), -A\bigr), & x < L_1, \\ 0, & x > L_1. \end{cases}$$

Note again that in our example (5.6) the infinite state is $-A = -1/3$ and the shock speed is $h(-A)/(-A) = -1/9$.

Numerical solutions of (5.6) with initial data,

$$u_0(x) = \begin{cases} \frac{2}{3}|\sin(2\pi x)|, & 0 < x < 2, \\ 0, & \text{otherwise,} \end{cases} \tag{5.10}$$

are shown in Figure 5.2. The first picture shows the initial data and the S-summation of 200 base functions at time $t = 6$. A part of the summation has been magnified with numerical approximations of the Godunov scheme in the second picture. We can clearly see that the solution of the Godunov scheme converges to the S-summation. This example shows that the S-summation gives a very accurate resolution using a small number of mesh points.



(a) *Initial data and S-summation at* $t = 6$ (b) *Comparison with Godunov*

FIG. 5.2. *Flux is* $f(u) = u^2 - u^3$. (a) *shows the initial data and the S-summation at* $t = 6$. (b) *shows that the Godunov scheme converges to the S-summation.* 200 *base functions are used in the S-summation and* 800 *and* 4,000 *meshes are used in the Godunov scheme.*

5.3. Space dependent flux in multidimensional space. The self-similarity of the problem (2.1) relies on the fact that the flux depends only on the solution, i.e., $f = f(u)$. We have no clue how to generalize our scheme to a problem with a general space dependent flux, $f = f(u, x)$. However, if the space dependence is given by

$$u_t + a(x)f'(u)u_x = 0, \tag{5.11}$$

the equation is transformed into

$$u_t + f(u)_y = 0 \tag{5.12}$$

under the change of variable $y(x) = \int_0^x 1/a(s)ds$, and our scheme can be applied.

Since the self-similarity of hyperbolic conservation laws is the one-dimensional property, it should be possible to expand the scheme to multidimensional problems. Consider a two-dimensional problem,

$$u_t + f'(u)(a(x_1, x_2)u_{x_1} + b(x_1, x_2)u_{x_2}) = 0, \tag{5.13}$$

with a velocity vector field satisfying

$$\partial_{x_1} a(x_1, x_2) + \partial_{x_2} b(x_1, x_2) = 0. \tag{5.14}$$

Cvetkovic and Dagans [6] suggest space variables y_1, y_2 satisfying

$$\frac{dy_1}{dx_1} = \frac{1}{a(x_1, \eta)}, \qquad y_2 = x_2 - \eta, \qquad \frac{d\eta}{dx_1} = \frac{b(x_1, \eta)}{a(x_1, \eta)}, \tag{5.15}$$

which transform (5.14) into

$$u_t + f(u)_{y_1} = 0, \qquad u = u(y_1, y_2, t). \tag{5.16}$$

Problem (5.16) can be considered as a set of one-dimensional problems, and, hence, the complexity of the scheme for it is of order $O(N^2)$. Since the transformation (5.15) also has the complexity of $O(N^2)$, we eventually get a scheme of $O(N^2)$ for a two-dimensional problem. In this approach, each channel of the velocity vector field is considered separately and, hence, it seems useful to channel problems.

6. Second order approximation. The scheme introduced in the previous sections exactly solves the problem with modified initial data, and the size of the initial error decreases in time. However, the scheme is not good enough for the short time behavior since the error generated by the initial discretization can be huge. Here we add an extra structure to base functions and make the initial data discretization to be of second order. In this way we can handle general piecewise self-similar solutions in (1.7).

6.1. Modified base functions. The base function considered in the previous sections has three indexes, say, m, t, c. In this section we introduce two extra indexes, h and $\bar{t}$. Note that there are two time indexes t and $\bar{t}$ which play different roles. We assume $0 \le t < \infty$ and $-\infty < \bar{t} \le \infty$. For simplicity we assume (H). It can be easily generalized, as it was in section 5.

To figure out the structure of the new base function $B^{h,\bar{t}}_{m,t,c}(x)$, we introduce

$$x^* = c + tf'(h), \qquad 0 \le t < \infty, \tag{6.1}$$

and

$$\bar{c} = x^* - \bar{t} f'(h), \qquad -\infty < \bar{t} < \infty, \tag{6.2}$$

(see Figures 6.1 and 6.2). Let $g(x)$ be the self-similarity profile. As an intermediate step we define $B_{t,c}^{h,\bar{t}}(x)$ first. For $0 < \bar{t} < \infty$ it is defined by

$$B_{t,c}^{h,\bar{t}}(x) = \begin{cases} g\big((x-c)/t\big), & c < x < x^*, \\ g\big((x-\bar{c})/\bar{t}\,\big), & x^* < x, \\ 0, & \text{otherwise}, \end{cases} \tag{6.3}$$

and, for $-\infty < \bar{t} \le 0$, it is defined by

$$B_{t,c}^{h,\bar{t}}(x) = \begin{cases} g\big((x-c)/t\big), & c < x < x^*, \\ g\big((x-\bar{c})/\bar{t}\,\big), & x^* < x < \bar{c}, \\ 0, & \text{otherwise}. \end{cases} \tag{6.4}$$

The constant $\bar{c}$ is the center of the top self-similarity profile with time index $\bar{t}$, and the constant x^* is the x-coordinate of the intersection point between two self-similarity profiles with indexes t and $\bar{t}$ (see Figures 6.1 and 6.2). We can easily see from (6.2) that $\bar{c} < x^*$ for $\bar{t} > 0$ and $\bar{c} > x^*$ for $\bar{t} < 0$. The function $B_{t,c}^{h,\bar{t}}(x)$ is well defined for $t = 0$, $\bar{t} = 0$ since the corresponding domain is empty. For $\bar{t} = \infty$, we consider

$$B_{t,c}^{h,\infty}(x) = \begin{cases} g\big((x-c)/t\big), & c < x < x^*, \\ h, & x^* < x, \\ 0, & \text{otherwise}. \end{cases} \tag{6.5}$$

Now we introduce the index $m > 0$, which decides the support of the base function. Let $\xi > c$ be the solution of

$$\int_c^{\xi} B_{t,c}^{h,\bar{t}}(x)dx = m. \tag{6.6}$$

For $\bar{t} > 0$ it always has a solution. For $\bar{t} \le 0$ it has a solution only if $m < \int_c^{\bar{c}} B_{t,c}^{h,\bar{t}}(x)dx$. The base function is now defined by

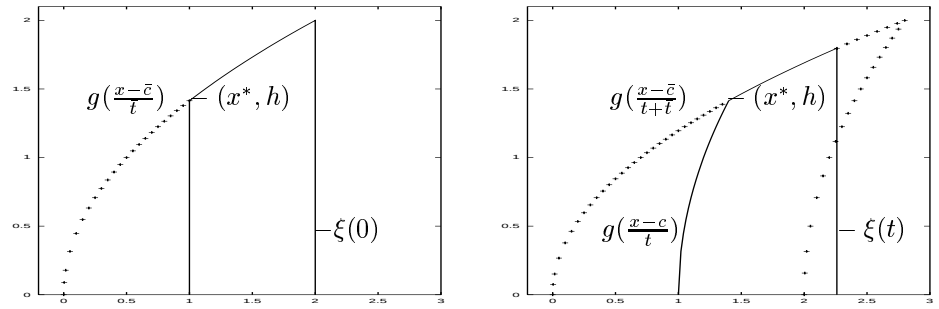
$$B_{m,t,c}^{h,\bar{t}}(x) = \begin{cases} B_{t,c}^{h,\bar{t}}(x), & c < x < \xi, \\ 0, & \text{otherwise}. \end{cases} \tag{6.7}$$

Let $u(x,t)$ be the solution of the conservation law $u_t + f(u)_x = 0$ with its initial value $u(x,0) = B_{m,0,c}^{h,\bar{t}}(x)$. Then, from the well-known technique of equal area construction, we may easily see that the solution is simply $u(x,t) = B_{m,t,c}^{h,\bar{t}+t}(x)$ (see Figures 6.1 and 6.2). For this solution $u(x,t)$, the point $x^* = x^*(t)$ in (6.1) satisfies

$$x^*(t) = c + tf'(h) = c + tf'(u(x^*(t),t)), \qquad x^*(t) < \xi(t),$$

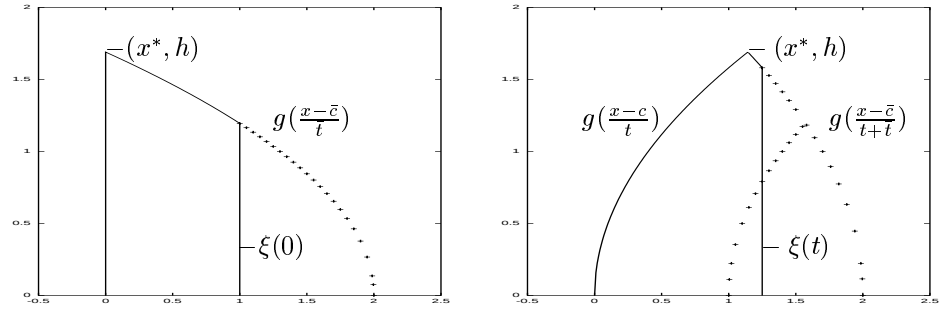
where $\xi = \xi(t)$ is the solution of (6.6). So $x = x^*(t)$ is a characteristic line for $x^*(t) < \xi(t)$. On the other hand, $\bar{c}$ is a constant with respect to $t > 0$:

$$\bar{c} = x^*(t) - (t + \bar{t})f'(h) = c - \bar{t} f'(h).$$



(a) *Base function* $B^{h,\bar{t}}_{m,0,1}(x)$, $\bar{t} > 0$ (b) *Base function* $B^{h,t+\bar{t}}_{m,t,1}(x)$

FIG. 6.1. *If the flux is* $f(v) = v^3/3$, *the self-similarity profile is* $g(x) = \sqrt{x}$. *Base function* $B^{h,\bar{t}}_{m,0,c}(x)$ *with* $c = 1$, $\bar{t} = 0.5$, $h = \sqrt{2}$, $m = \int_1^2 g(x/0.5)dx$ *is given in* (a) *(solid lines). We can easily check that* $\bar{c} = 0$. *If* $u(x,0) = B^{h,\bar{t}}_{m,0,c}(x)$, *the solution of the conservation law is* $u(x,t) = B^{h,t+\bar{t}}_{m,t,c}(x)$ *and it is given in* (b) *(solid lines) with* $t = 0.2$.



(a) *Base function* $B^{h,\bar{t}}_{m,0,0}(x)$, $\bar{t} < 0$ (b) *Base function* $B^{h,t+\bar{t}}_{m,t,0}(x)$

FIG. 6.2. *If the flux is* $f(v) = v^3/3$, *the self-similarity profile is* $g(x) = \sqrt{x}$. *Base function* $B^{h,\bar{t}}_{m,0,c}(x)$ *with* $c = 0$, $\bar{t} = -0.7$, $h = \sqrt{2/0.7}$, $m = \int_0^1 g(2 - x/0.7)dx$ *is given in* (a) *(solid lines). We can easily check that* $\bar{c} = 2$. *If* $u(x,0) = B^{h,\bar{t}}_{m,0,c}(x)$, *the solution of the conservation law is* $u(x,t) = B^{h,t+\bar{t}}_{m,t,c}(x)$ *and it is given in* (b) *(solid lines) with* $t = 0.4$.

In Figures 6.1 and 6.2 base functions are displayed for positive and negative $\bar{t}$ together with self-similarity profiles. In these figures we can clearly observe the different roles of two self-similarity profiles generated by two index sets $\{c, t\}$ and $\{\bar{c}, \bar{t}\,\}$.

The S-summation among these base functions can be similarly defined using the profile $g\big((x-c)/t\big)$ in the domain $c < x < x^*$ and the profile $g\big((x-\bar{c})/\bar{t}\,\big)$ for $x^* < x$. We omit the details. We may consider the base function (3.2) as a special case of (6.7) with $\bar{t} = 0$.

6.2. Initial discretization and the approximation. Suppose the initial function $v_0 \in L^1$ has a compact support $\mathrm{supp}(v_0) \subset [L_1, L_2]$. Let $\mathcal{C} = \{c_n = L_1 < \cdots < c_1 < c_0 = L_2\}$ be a partition of the interval $[L_1, L_2]$. We can approximate v_0 with self-similarity profiles over interval (c_k, c_{k-1}) with time index $\bar{t}_k \in \boldsymbol{R}$, which is second order. For the Burgers case it is simply a piecewise linear approximation. The

approximation u_0 can be written as

$$u_0(x) = \bigodot_{k=1}^{n} B^{h_k,\bar{t}_k}_{m_k,0,c_k}(x) = \sum_{k=1}^{n} B^{h_k,\bar{t}_k}_{m_k,0,c_k}(x), \tag{6.8}$$

where $m_k = \int_{c_k}^{c_{k-1}} u_0(x)dx$ and $h_k = u_0(c_k)$. Initially, the supports of base functions are disjoint, and, hence, the self-similarity summation is the usual summation. The exact solution $v(x,t)$ of the conservation law (1.1) is approximated by

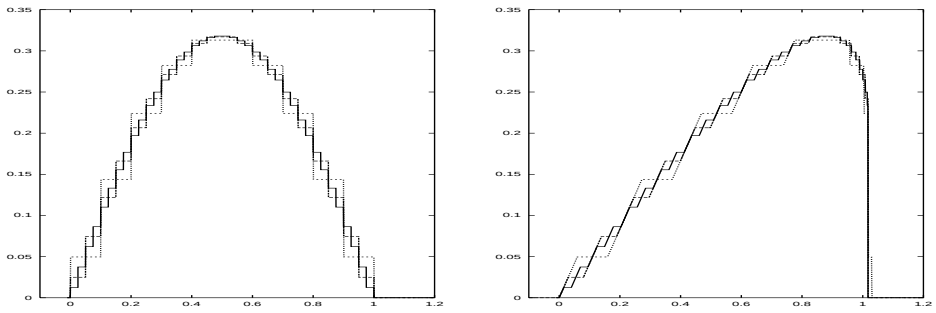
$$u(x,t) = \bigodot_{k=1}^{n} B^{h_k,t+\bar{t}_k}_{m_k,t,c_k}(x), \tag{6.9}$$

and we expect an error estimate similar to (3.16), i.e.,

$$||v(x,t) - u(x,t)||_1 \le ||v_0(x) - u_0(x)||_1 = O(||\mathcal{C}||^2) \quad \text{as} \quad ||\mathcal{C}|| \to 0. \tag{6.10}$$

Remark 6.1. The initial discretization (6.8) is trivial in comparison with Step 1 in section 4.2. It is an additional advantage we obtain when the modified base function is used in a numerical scheme. However, this additional structure may cause extra complexity when it is used as an analytical tool.

Remark 6.2 (piecewise constant data). In many cases initial data are given as piecewise constant functions from the beginning. In this case the initial data can be considered as a summation of base functions with $\bar{t} = \infty$; see (6.5). In Figure 6.3 we consider the Burgers case (4.1) using base functions $B^{h,\infty}_{m,t,c}(x)$. We can clearly see that these approximations represent the shock location very well. Unlike the previous case, the solution with finer mesh always passes though the constant parts of coarse ones.

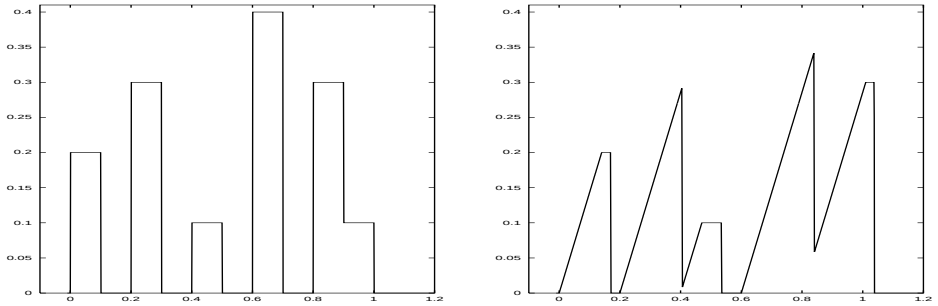


(a) *Data discretization* (b) *Solutions at* $t = 1.2$

FIG. 6.3. *The S-summation for the modified base functions* (6.7) *with* $\bar{t} = \infty$ *gives a piecewise constant, piecewise self-similar solution. In the figure,* 3 *summations are displayed together using* 10, 20, 40 *base functions. We observe that the finer one always passes the constant parts.*

Remark 6.3 (singular initial data). If singular initial data are given, then extra mesh points are usually introduced to capture the effect of the singularity of the data. But since our method handles initial data individually, extra mesh points are not needed. In Figure 6.4(a) the Burgers equation is solved with singular initial data. We use 6 modified base functions with $\bar{t} = \infty$.

Remark 6.4 (front tracking). It is possible to consider the front tracking method in terms of the S-summation. Consider an L^1 solution of the Burgers equation



(a) *Singular initial data* (b) *Solutions at* $t = 0.7$

FIG. 6.4. *The scheme does not require extra meshes to handle singular initial data* (a). *In the S-summation every datum is handled exactly by a base function. Only* 6 *base functions solve this example.*

bounded by $0 \le u(x,t) \le 1$. Let $h(u)$ be the polygonal approximation of the flux $f(u) = u^2/2$ with the partition $\{0, 1/n, \dots, n/n = 1\}$. Then $h'(u)$ is a step function,

$$h'(u) = (2k-1)/2n, \qquad (k-1)/n < u < k/n, \quad k = 1, \dots, n, \tag{6.11}$$

and the self-similarity profile $g(x)$ is given by

$$g(x) = (k-1)/n, \qquad (2k-1)/2n < x < (2k+1)/2n, \quad k = 1, \dots, n. \tag{6.12}$$

So the values of $g(x)$ are the *breaking points* of the flux $h(u)$. We can approximate the given initial data v_0 by taking a cell average, not just breaking points. Then the initial discretization u_0 can be written in the form of (6.8) with $\bar{t}_k = \infty$. This is a simplified version of the front tracking method under (H).

Acknowledgments. The author would like to thank Professor A. E. Tzavaras. He gave the author the motivation and valuable remarks for this work. The author also would like to thank Professor Giovanni Russo and the anonymous referees for their helpful suggestions. He is also grateful to people in the Institute for Mathematics and its Applications (IMA) for helpful discussions and support.

REFERENCES

[1] A. L. BERTOZZI, A. MÜNCH, AND M. SHEARER, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.
[2] A. BRESSAN, *Global solutions of systems of conservation laws by wave-front tracking*, J. Math. Anal. Appl., 170 (1992), pp. 414–432.
[3] A. BRESSAN, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, Oxford, 2000.
[4] R. COURANT, K. FRIEDRICHS, AND H. LEWY, *Uber die partiellen Differenzengleichungen der mathematisches Physik*, Math. Ann., 100 (1928), pp. 32–74.
[5] R. COURANT, K. FRIEDRICHS, AND H. LEWY, *On the partial difference equations of mathematical physics*, IBM J. Res. Develop., 11 (1967), pp. 215–234.
[6] V. CVETKOVIC AND G. DAGAN, *Transport of kinetically sorbing solute by steady random velocity in heterogenous porous formations*, J. Fluid Mech., 265 (1994), pp. 189–215.
[7] C. M. DAFERMOS, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.
[8] C. M. DAFERMOS, *Regularity and large time behaviour of solutions of a conservation law without convexity*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1985), pp. 201–239.

[9] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer–Verlag, New York, 2000.
[10] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.
[11] S. K. Godunov, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb. (N.S.), 47 (1959), pp. 271–306 (in Russian).
[12] H. Holden and L. Holden, *On scalar conservation laws in one dimension*, in Ideas and Methods in Mathematical Analysis, Stochastics, and Applications (Oslo, 1988), Cambridge Univ. Press, Cambridge, UK, 1992, pp. 480–509.
[13] H. Holden, K. A. Lie, and N. H. Risebro, *An unconditionally stable method for the Euler equations*, J. Comput. Phys., 150 (1999), pp. 76–96.
[14] H. Holden and N. H. Risebro, *A method of fractional steps for scalar conservation laws without the CFL condition*, Math. Comp., 60 (1993), pp. 221–232.
[15] L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Math. Appl. 26, Springer-Verlag, Berlin, 1997.
[16] Y.-J. Kim, *Asymptotic behavior of solutions to scalar conservation laws and optimal convergence orders to N-waves*, J. Differential Equations, to appear.
[17] Y. J. Kim and A. E. Tzavaras, *Diffusive N-waves and metastability in Burgers equation*, SIAM J. Math. Anal., 33 (2001), pp. 607–633.
[18] S. N. Kruzhkov, *First order quasilinear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.
[19] S. N. Kruzhkov, *First order quasilinear equations in several independent variables*, Mat. Sb., 123 (1970), pp. 228–255.
[20] R. J. LeVeque, *Large time step shock-capturing techniques for scalar conservation laws*, SIAM J. Numer. Anal., 19 (1982), pp. 1091–1109.
[21] K.-A. Lie, *Front tracking for one-dimensional quasilinear hyperbolic equations with variable coefficients*, Numer. Algorithms, 24 (2000), pp. 275–298.
[22] K.-A. Lie, V. Haugse, and K. H. Karlsen, *Dimensional splitting with front tracking and adaptive grid refinement*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 627–648.
[23] T.-P. Liu and M. Pierre, *Source-solutions and asymptotic behavior in conservation laws*, J. Differential Equations, 51 (1984), pp. 419–441.
[24] B. Lucier, *A moving mesh numerical method for hyperbolic conservation laws*, Math. Comp., 46 (1986), pp. 59–69.
[25] K. Miller, *Moving finite elements.* II, SIAM J. Numer. Anal., 18 (1981), pp. 1033–1057.
[26] G. Whitham, *Linear and Nonlinear Waves*, Pure and Applied Mathematics, Wiley-Interscience, New York, 1974.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2133–2150

A NEW SUPERCONVERGENCE FOR MIXED FINITE ELEMENT APPROXIMATIONS*

RICHARD E. EWING†, MINGJUN LIU‡, AND JUNPING WANG§

Abstract. A new superconvergence result is established for numerical solutions of elliptic problems obtained from the mixed finite element method of Raviart–Thomas over rectangular partitions. The well-known optimal order error estimate in L^2-norm for the flux approximation is of order $\mathcal{O}(h^{k+1})$, where $k \geq 0$ is the order of polynomials employed in the Raviart–Thomas element. The new superconvergence shows an improved accuracy of order $\mathcal{O}(h^{k+3})$ between the mixed finite element approximation and an appropriately defined local projection of the flux variable when $k > 0$. A postprocessing technique using local projection methods is proposed and analyzed in order to provide a new approximate solution with the superconvergent order $\mathcal{O}(h^{k+3})$.

Key words. superconvergence, mixed finite element method, error estimates, elliptic problems

AMS subject classifications. 65N30, 65N15, 65N10

PII. S0036142901391141

1. Introduction.

In this paper we are concerned with error analysis for numerical solutions of elliptic problems by mixed finite element methods over rectangular finite element partitions. In particular, we shall investigate some superconvergence properties of the numerical solution obtained from the Raviart–Thomas rectangular element [15, 7].

The model problem under consideration seeks $p \in H^1(\Omega)$ satisfying

$$-\nabla \cdot (\boldsymbol{a} \nabla p) = f \qquad \text{in } \Omega \tag{1.1}$$

and the boundary condition

$$\boldsymbol{a} \nabla p \cdot \mathbf{n} = g \qquad \text{on } \partial\Omega, \tag{1.2}$$

where Ω is an open bounded domain in $\mathbb{R}^2$, $\boldsymbol{a} = \boldsymbol{a}(x, y)$ is a 2×2 tensor which is symmetric and uniformly positive definite in Ω, $\mathbf{n}$ is the outward unit normal vector on $\partial\Omega$, and $f = f(x, y)$, $g = g(x, y)$ are two given functions defined on Ω and its boundary $\partial\Omega$, respectively. The functions f and g are assumed to satisfy the compatibility condition $\int_\Omega f(x, y) d\Omega + \int_{\partial\Omega} g(x, y) ds = 0$ so that (1.1) and (1.2) has a solution. The mixed finite element method [15, 7, 8] is a method that approximates the scalar $p = p(x, y)$ and the flux variables $\mathbf{u} = -\boldsymbol{a}\nabla p$ simultaneously in two different finite element spaces. The corresponding scheme is outlined in section 2.

Optimal order error estimates have been derived in [8] for all the existing mixed finite elements [4, 5, 6, 10, 15] satisfying the inf-sup condition of Brezzi [3] and

*Received by the editors June 18, 2001; accepted for publication (in revised form) June 11, 2002; published electronically December 13, 2002.

http://www.siam.org/journals/sinum/40-6/39114.html

†Institute for Scientific Computation, Texas A&M University, College Station, TX 77843 (richard-ewing@tamu.edu). The research of this author was supported in part by the NSF grants DMS-9706985 and DMS-9972147.

‡Imagine Software, Inc., New York, NY 10017 (mingjunl@imagine-sw.com).

§Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (jwang@mines.edu). The research of this author was supported in part by the NSF grant DMS-9706985.

Babuška [1]. For the Raviart–Thomas element of order $k \geq 0$ on triangles or rectangles, the optimal order error estimate in L^2-norm for the flux and the scalar variables is of order $\mathcal{O}(h^{k+1})$, where h is the mesh size of the corresponding finite element partition. The optimal order error estimate is generally the best one can get between the exact solution and its numerical approximation when measured globally on the computational domain. But when compared with a certain local projection of the exact solution, the finite element solution often possesses a supercloseness property over the optimal order error estimate. More precisely, if $\mathbf{u}_h$ is the finite element approximation, then for an appropriately defined local projection $\pi_h\mathbf{u}$ [8, 7, 9] one may have

$$\|\mathbf{u}_h - \pi_h\mathbf{u}\|_{L^2(\Omega)} \leq C(\mathbf{u})h^{k+s} \tag{1.3}$$

for some parameters $s > 1$. For example, for rectangular finite element partitions, the mixed finite element solution from using either the Raviart–Thomas or the Brezzi–Douglas–Fortin–Marini (BDFM) [5] elements was proved to satisfy

$$\|\mathbf{u}_h - \pi_h\mathbf{u}\|_{L^2(\Omega)} \leq C(\mathbf{u})h^{k+2}, \tag{1.4}$$

where $C(\mathbf{u})$ is a constant independent of the mesh size h. The estimate (1.4) is certainly much better than the optimal order error estimate and can be used to construct a new numerical solution with the same order of accuracy. We refer to [9, 13, 11, 12, 16] for a detailed discussion of (1.4) on rectangular elements. The estimate (1.4) was also derived for the triangular Raviart–Thomas element of order $k = 0$ when the underlying partition is uniform [2].

In their numerical experiments conducted in 1989 for reservoir simulation, Ewing and Shen [14] observed that the numerical flux $\mathbf{u}_h$ actually approximates the exact solution $\mathbf{u}$ at an order of $\mathcal{O}(h^4)$ on the two diagonal lines of each rectangular element when $k = 1$ for the Raviart–Thomas element. This means that the true rate of superconvergence for the mixed finite element approximation was significantly underestimated by the supercloseness estimate (1.4). Our main objective of this paper is to provide a theoretical justification for the super-superconvergence observed by Ewing and Shen. In particular, for diagonal tensors $\boldsymbol{a} = \boldsymbol{a}(x, y)$ with piecewise constant entries, we show that there is a constant $C = C(\mathbf{u})$ independent of the mesh size h such that

$$\|\mathbf{u}_h - \pi_h\mathbf{u}\|_{L^2(\Omega)} \leq C(\mathbf{u})h^{k+3} \tag{1.5}$$

for any $k \geq 1$.

The analysis for (1.5) is different from that of (1.4) as presented in [9, 13, 11, 12]. The difference lies in the treatment of a linear form

$$\mathcal{F}(\mathbf{v}) = (\boldsymbol{a}^{-1}(\mathbf{u} - \pi_h\mathbf{u}), \mathbf{v}) \tag{1.6}$$

for any finite element function $\mathbf{v}$. Traditionally, this linear form is estimated by expanding the interpolation error $\mathbf{u} - \pi_h\mathbf{u}$ as a Taylor-like series. Leading terms in the expansion are proved to be orthogonal with arbitrary finite element functions $\mathbf{v}$. If so, one would be able to derive a superconvergence estimate. In the analysis to be presented in this paper, the above linear form is studied by expanding $\mathbf{v}$ as a Taylor series involving only finite number of terms. Each of the terms in the Taylor expansion is a polynomial. The orthogonality of $\mathbf{u} - \pi_h\mathbf{u}$ with a certain class of polynomials then plays an important role for the desired superconvergence result.

The paper is organized as follows. In section 2, we provide a short review of the mixed finite element method. In section 3, we establish a general framework of superconvergence. In section 4, we derive some estimates useful for analyzing the linear form (1.6). In section 5, we obtain some superconvergence results for the mixed finite element solutions by combining the results of sections 3 and 4. In section 6, we present a local projection method that can be used to provide a numerical solution with the desired superconvergence.

2. Preliminaries in mixed methods. With the flux variable $\mathbf{u} = -\boldsymbol{a}\nabla p$, our model problem (1.1) can be rewritten as a system of linear equations

$$\begin{aligned} \boldsymbol{a}^{-1}\mathbf{u} + \nabla p &= 0 \qquad \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= f \qquad \text{in } \Omega, \\ \mathbf{u} \cdot \mathbf{n} &= -g \qquad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

The new variable $\mathbf{u}$ is known as Darcy's velocity in the numerical simulation of fluid flow in porous media.

The mixed finite element method for (1.1) is based on a weak form for the mixed formulation (2.1). To this end, let $(\cdot,\cdot)$ denote the standard inner product in $L^2(\Omega)$ or $\left[L^2(\Omega)\right]^2$, as appropriate. The corresponding norm is denoted by $\|\cdot\|_0$. Let

$$\mathbf{V} = \{\mathbf{v} = \mathbf{v}(x) : \quad \mathbf{v} \in L^2(\Omega),\, \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$$

be a Sobolev space equipped with the norm

$$\|\mathbf{v}\|_{\mathbf{V}} = \left(\|\mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2\right)^{1/2}.$$

Let $W = L_0^2(\Omega)$ consisting of functions in $L^2(\Omega)$ with mean value zero, and

$$\mathbf{V}_g = \{\mathbf{v} : \quad \mathbf{v} \in \mathbf{V},\, \mathbf{v} \cdot \mathbf{n} = g \text{ on } \partial\Omega\},$$

where $g = g(x,y) \in H^{-1/2}(\partial\Omega)$ is a given function on the boundary $\partial\Omega$. A weak form for (2.1) seeks $\mathbf{u} \in \mathbf{V}_g$ and $p \in W$ satisfying

$$\begin{aligned} (\boldsymbol{a}^{-1}\mathbf{u}, \mathbf{v}) - (\nabla \cdot \mathbf{v}, p) &= 0, \qquad && \mathbf{v} \in \mathbf{V}_0, \\ (\nabla \cdot \mathbf{u}, w) &= (f, w), && w \in W. \end{aligned} \tag{2.2}$$

The mixed finite element method for (1.1)–(1.2) is based on the weak formulation (2.2) and two finite element subspaces $\mathbf{V}_h \subset \mathbf{V}$ and $W_h \subset W$ associated with a prescribed finite element partition $\mathcal{T}_h$ for the domain Ω. For the purpose of this paper, we consider a special case of the finite element partition $\mathcal{T}_h$ consisting of rectangular elements only. This means that the domain Ω would have to be made of rectangular subdomains with boundaries parallel to either the x or y axis.

Of particular interest in this paper, we consider the Raviart–Thomas finite elements on the rectangular partition $\mathcal{T}_h$. Let $Q_{r,s}$ be the space of polynomials with degree no more than r in the x direction and no more than s in the y direction. The Raviart–Thomas finite element space of order $k \geq 0$ is given as follows:

$$\begin{aligned} \mathbf{V}_h &= \{\mathbf{v} \in \mathbf{V},\, \mathbf{v}|_e \in Q_{k+1,k} \times Q_{k,k+1},\, e \in \mathcal{T}_h\}, \\ W_h &= \{w \in W,\, w|_e \in Q_{k,k},\, e \in \mathcal{T}_h\}. \end{aligned}$$

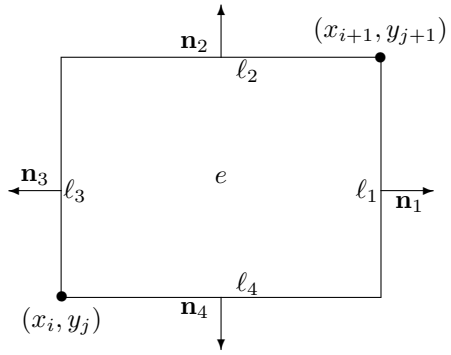


FIG. 1. *A rectangular element e and its four edges.*

Let Λ_h be a finite element space on $\partial\Omega$ obtained as the normal component of functions in $\mathbf{V}_h$:

$$\Lambda_h = \{\mathbf{v}\cdot\mathbf{n} \quad \forall \mathbf{v} \in \mathbf{V}_h\}.$$

It is not hard to see that Λ_h contains piecewise polynomials of degree k in either the x or y directions. Denote by $g_h \in \Lambda_h$ an approximation of the boundary data g such that

$$\langle g_h, \chi\rangle = g(\chi), \qquad \chi \in \Lambda_h,$$

where $g(\chi)$ is the duality pair and $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $L^2(\partial\Omega)$. Let

$$\mathbf{V}_{g,h} = \{\mathbf{v} \in \mathbf{V}_h,\ \mathbf{v}\cdot\mathbf{n} = g_h \text{ on } \partial\Omega\}$$

be a hyper-plane in $\mathbf{V}_h$. The Raviart–Thomas mixed finite element approximation is given by $\mathbf{u}_h \in \mathbf{V}_{g,h}$ and $p_h \in W_h$ satisfying

$$\begin{aligned} (\boldsymbol{a}^{-1}\mathbf{u}_h, \mathbf{v}) - (\nabla\cdot\mathbf{v}, p_h) &= 0, \qquad && \mathbf{v} \in \mathbf{V}_{0,h}, \\ (\nabla\cdot\mathbf{u}_h, w) &= (f, w), && w \in W_h. \end{aligned} \tag{2.3}$$

We emphasize that $\mathbf{V}_{0,h}$ consists of all finite element functions which have a vanishing component on the boundary $\partial\Omega$ in the normal direction.

The discrete problem (2.3) is known to satisfy the inf-sup condition of Brezzi [3] and Babuška [1]. The inf-sup condition can be verified by using a projection operator $\pi_h : \mathbf{V} \to \mathbf{V}_h$ satisfying

$$(\nabla\cdot(\mathbf{v} - \pi_h\mathbf{v}), w) = 0 \tag{2.4}$$

for all $w \in W_h$. The required boundedness of the projection operator π_h can be easily verified.

To construct such a projection operator π_h, let e be any rectangular element depicted in Figure 1. For any sufficiently smooth vector-valued function $\mathbf{v} \in \mathbf{V}$, define its projection $\pi_e\mathbf{v} \in Q_{k+1,k} \times Q_{k,k+1}$ by using the following system of linear equations:

$$\begin{aligned} \int_{\ell_i} (\mathbf{v} - \pi_e\mathbf{v})\cdot\mathbf{n}_i\phi ds &= 0, \qquad && \phi \in P_k(\ell_i),\ i = 1,2,3,4, \\ \int_e (\mathbf{v} - \pi_e\mathbf{v})\cdot\psi dxdy &= 0, && \psi \in Q_{k-1,k}(e) \times Q_{k,k-1}(e), \end{aligned}$$

where $\{\ell_i,\ i=1,2,3,4\}$ stand for the edges of the element e.

Using the local operator π_e, we can define a global projection operator $\pi_h : \mathbf{V} \to \mathbf{V}_h$ by setting

$$(\pi_h \mathbf{v})(x,y) = \pi_e \mathbf{v}(x,y) \qquad \forall (x,y) \in e,\ e \in \mathcal{T}_h. \tag{2.5}$$

It is not hard to see that the projection operator π_h satisfies the desired relation (2.4). In addition, the operator π_h can be split into two components:

$$\pi_h \mathbf{v} = (\pi_1 v_1, \pi_2 v_2),$$

where π_1 and π_2 are defined independently to each other.

Denote by P_h the L^2 projection from $L^2(\Omega)$ onto the pressure finite element space W_h.

3. A general framework in superconvergence. Let $(\mathbf{u}_h; p_h)$ be the mixed finite element approximation of (2.1) arising from the scheme (2.3). Our objective here is to establish a general framework for superconvergence estimation of the errors

$$\mathbf{e}_h = \mathbf{u}_h - \pi_h \mathbf{u}; \qquad \xi_h = p_h - P_h p.$$

THEOREM 3.1. *Assume that* $(\mathbf{u}; p)$ *solves the mixed problem* (2.2). *Let* $(\mathbf{u}_h; p_h)$ *solve* (2.3). *Assume that there is a constant* $M = M(\mathbf{u}, p)$ *and a parameter* $s \geq 0$ *satisfying*

$$\sup_{\mathbf{v} \in \mathbf{V}_{0,h}} \frac{(\boldsymbol{a}^{-1}(\mathbf{u} - \pi_h \mathbf{u}), \mathbf{v})}{\|\mathbf{v}\|_{\mathbf{V}}} \leq M(\mathbf{u}, p) h^{k+s}. \tag{3.1}$$

Then there is a generic constant C *independent of* h *and the solution* p *such that*

$$\|\mathbf{u}_h - \pi_h \mathbf{u}\|_0 + \|p_h - P_h p\|_0 \leq C M(\mathbf{u}, p) h^{k+s}.$$

Proof. By subtracting (2.3) from (2.2) we obtain

$$\begin{aligned} (\boldsymbol{a}^{-1}(\mathbf{u} - \mathbf{u}_h), \mathbf{v}) - (\nabla \cdot \mathbf{v}, p - p_h) &= 0, \qquad \mathbf{v} \in \mathbf{V}_{0,h}, \\ (\nabla \cdot (\mathbf{u} - \mathbf{u}_h), w) &= 0, \qquad w \in W_h. \end{aligned} \tag{3.2}$$

Using the error equation (3.2) and (2.4) we see that

$$\begin{aligned} (\boldsymbol{a}^{-1}\mathbf{e}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, \xi_h) &= (\boldsymbol{a}^{-1}(\mathbf{u} - \pi_h \mathbf{u}), \mathbf{v}), \qquad \mathbf{v} \in \mathbf{V}_{0,h}, \\ (\nabla \cdot \mathbf{e}_h, w) &= 0, \qquad w \in W_h. \end{aligned} \tag{3.3}$$

The second equation in (3.3) implies that

$$\nabla \cdot \mathbf{e}_h = 0. \tag{3.4}$$

By letting $\mathbf{v} = \mathbf{e}_h$ in the first equation and $w = \xi_h$ in the second equation of (3.3), we arrive at

$$(\boldsymbol{a}^{-1}\mathbf{e}_h, \mathbf{e}_h) - (\nabla \cdot \mathbf{e}_h, \xi_h) = (\boldsymbol{a}^{-1}(\mathbf{u} - \pi_h \mathbf{u}), \mathbf{e}_h).$$

Substituting (3.4) into above equation we obtain

$$(\boldsymbol{a}^{-1}\mathbf{e}_h, \mathbf{e}_h) = (\boldsymbol{a}^{-1}(\mathbf{u} - \pi_h \mathbf{u}), \mathbf{e}_h). \tag{3.5}$$

It follows from the condition (3.1) that

$$(\boldsymbol{a}^{-1}\mathbf{e}_h, \mathbf{e}_h) \le Ch^{k+s} M(\mathbf{u}, p)\|\mathbf{e}_h\|_{\mathbf{V}},$$

which, along with (3.4), leads to

$$(\boldsymbol{a}^{-1}\mathbf{e}_h, \mathbf{e}_h) \le Ch^{k+s} M(\mathbf{u}, p)\|\mathbf{e}_h\|_0. \tag{3.6}$$

Thus,

$$\|\mathbf{e}_h\|_0 \le Ch^{k+s} M(\mathbf{u}, p). \tag{3.7}$$

This completes the estimate for the vector component.

Next, we estimate the L^2-norm of ξ_h. From the inf-sup condition and (3.2) we obtain

$$\begin{aligned}
\|\xi_h\|_0 &\le C \sup_{\mathbf{v}\in\mathbf{V}_{0,h}} \frac{(\nabla\cdot\mathbf{v}, \xi_h)}{\|\mathbf{v}\|_{\mathbf{V}}} \\
&= C \sup_{\mathbf{v}\in\mathbf{V}_{0,h}} \frac{(\nabla\cdot\mathbf{v}, p_h - p)}{\|\mathbf{v}\|_{\mathbf{V}}} \\
&= C \sup_{\mathbf{v}\in\mathbf{V}_{0,h}} \frac{(\boldsymbol{a}^{-1}\mathbf{v}, \mathbf{u}_h - \mathbf{u})}{\|\mathbf{v}\|_{\mathbf{V}}} \\
&= C \sup_{\mathbf{v}\in\mathbf{V}_{0,h}} \frac{(\boldsymbol{a}^{-1}\mathbf{v}, \mathbf{u}_h - \pi_h\mathbf{u}) + (\boldsymbol{a}^{-1}\mathbf{v}, \pi_h\mathbf{u} - \mathbf{u})}{\|\mathbf{v}\|_{\mathbf{V}}}, \\
&\le C \sup_{\mathbf{v}\in\mathbf{V}_{0,h}} \frac{(\boldsymbol{a}^{-1}\mathbf{v}, \mathbf{u} - \pi_h\mathbf{u})}{\|\mathbf{v}\|_{\mathbf{V}}} + C\|\mathbf{u}_h - \pi_h\mathbf{u}\|_0,
\end{aligned} \tag{3.8}$$

which, together with the error estimate (3.7) and the condition (3.1), gives

$$\|\xi_h\|_0 \le Ch^{k+s} M(\mathbf{u}, p). \tag{3.9}$$

The proof is then completed by combining (3.7) with (3.9). ∎

4. Some estimates for linear forms. Theorem 3.1 indicates that a superconvergence is guaranteed if one can establish a corresponding estimate for the linear form $\mathcal{F}(\mathbf{v}) = (\boldsymbol{a}^{-1}(\mathbf{u}-\pi_h\mathbf{u}), \mathbf{v})$ in the mixed finite element space $\mathbf{V}_{0,h}$. Our objective here is to study this linear form and derive some useful estimates.

Denote by

$$|\mathbf{u}|_{m,q;\Omega} = \sum_{|\alpha|=m} \|D^\alpha \mathbf{u}\|_{L^q(\Omega)}$$

the seminorm in the Sobolev space $W^{m,q}(\Omega)$ with integer $m \ge 0$ and real number $q \ge 1$. The norm in $W^{m,q}(\Omega)$ is denoted by

$$\|\mathbf{u}\|_{m,q;\Omega} = \sum_{j=0}^{m} |\mathbf{u}|_{j,q;\Omega}.$$

In addition, denote by

$$|w|_{m,q,h} = \sum_{|\alpha|=m} \left(\sum_{e\in\mathcal{T}_h} \|D^\alpha w\|^q_{L^q(e)} \right)^{1/q}$$

a discrete seminorm for any piecewise polynomials $w = w(x, y)$. The following estimate turns out to be critical in our analysis.

LEMMA 4.1. *Let $e \in \mathcal{T}_h$ be a rectangular element as illustrated in Figure 1 with $e = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$. Let ψ be a sufficiently smooth function defined on e, and satisfy*

$$\int_{\ell_i} \psi dy = 0, \qquad i = 1, 3, \tag{4.1}$$

$$\int_e x^j \psi dx dy = 0, \qquad 0 \le j \le \tilde{k} - 1. \tag{4.2}$$

Then for any integer $m \le \tilde{k} + 1$, we have

$$\int_e (x - x_e)^m \psi dx dy = \frac{(-1)^m m!}{(2m+2)!} \int_e E^{m+1}(x) \partial_x^{m+2} \psi dx dy, \tag{4.3}$$

where ℓ_1, ℓ_3 are two vertical edges of the element e, $2h_e = x_{i+1} - x_i$ is the length of ℓ_2, (x_e, y_e) is the center of e, and $E(x) = (x - x_e)^2 - h_e^2$.

Proof. Let us apply the integration by parts to

$$\int_e E^{m+1}(x) \partial_x^{m+2} \psi(x, y) dx dy.$$

Notice that the function $E(x)$ vanishes at edges ℓ_1 and ℓ_3 where $x = x_{j+1}$ and $x = x_i$, respectively. Thus,

$$\int_e E^{m+1}(x) \partial_x^{m+2} \psi dx dy = (-1)^{m+1} \int_e \partial_x^{m+1} E^{m+1}(x) \partial_x \psi dx dy, \tag{4.4}$$

where contributions from the boundary integral are trivial due to the fact that $\partial_x^t E^{m+1}(x) = 0$ for $0 \le t < m + 1$ at $x = x_i$ and $x = x_{i+1}$. Next, using the integration by parts again we obtain

$$\int_e E^{m+1}(x) \partial_x^{m+2} \psi dx dy = (-1)^m \int_e \partial_x^{m+2} E^{m+1}(x) \psi dx dy, \tag{4.5}$$

where the boundary contribution vanishes because of the assumption (4.1).

Now we observe that

$$\partial_x^{m+2} E^{m+1}(x) = \frac{(2m+2)!}{m!} (x - x_e)^m + r(x), \tag{4.6}$$

where $r(x)$ is a polynomial of degree no more than $m - 2 \le \tilde{k} - 1$. By substituting (4.6) into (4.5) and then using the condition (4.2) we obtain

$$\int_e E^{m+1}(x) \partial_x^{m+2} \psi dx dy = \frac{(-1)^m (2m+2)!}{m!} \int_e (x - x_e)^m \psi dx dy,$$

which is (4.3). ∎

Let $\boldsymbol{a}^{-1}$ be a diagonal tensor given by

$$\boldsymbol{a}^{-1} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}. \tag{4.7}$$

By expressing the vector-valued function $\mathbf{v}$ as $\mathbf{v} = (v_1, v_2)$, we arrive at

$$\mathcal{F}(\mathbf{v}) = (\alpha_1(u_1 - \pi_1 u_1), v_1) + (\alpha_2(u_2 - \pi_2 u_2), v_2). \tag{4.8}$$

The following lemma is concerned with the first linear form on the right-hand side of (4.8).

LEMMA 4.2. *Let $k \geq 1$ be an integer and u_1 be a sufficiently smooth function on the rectangle depicted in Figure* 1. *Let v_1 be a polynomial of degree no more than $k+1$ in x and k in y, respectively. Let v_2 be any sufficiently smooth function on e and $\mathbf{v} = (v_1, v_2)$. Then there is a $J_{1,e}$ such that*

$$(u_1 - \pi_1 u_1, v_1)_e = J_{1,e} + \frac{(-1)^k}{(2k+2)!}\left(\int_{\ell_4} - \int_{\ell_2}\right) E^{k+1}(x)\partial_x^{k+2} u_1 \partial_x^{k-1} v_2 dx. \tag{4.9}$$

The term $J_{1,e}$ can be represented as area integrals over the element e with the following estimate:

$$\begin{aligned} |J_{1,e}| \leq \frac{h_e^{2k+2}}{(2k+2)!}\Bigg(&\frac{h_e^2}{2k+3}|u_1|_{k+3,p;e}|v_1|_{k+1,q;e} \\ &+ |u_1|_{k+2,p;e}|\nabla\cdot\mathbf{v}|_{k-1,q;e} + |u_1|_{k+3,p;e}|v_2|_{k-1,q;e}\Bigg), \end{aligned} \tag{4.10}$$

where $2h_e = x_{i+1} - x_i$ is the width of the element and q is the conjugate of $p \geq 1$ satisfying $1/p + 1/q = 1$.

Proof. Let us expand the polynomial v_1 in x as follows:

$$v_1(x,y) = \sum_{i=0}^{k+1} \frac{1}{i!}(x - x_e)^i \partial_x^i v_1(x_e, y).$$

Each of $\partial_x^i v_1(x_e, y)$ is a polynomial of degree no more than k in y. The definition of π_1 implies that $u_1 - \pi_1 u_1$ is orthogonal to the polynomial space $Q_{k-1,k}$ (polynomials of degree no more than $k-1$ in x and k in y). Thus,

$$\begin{aligned} (u_1 - \pi_1 u_1, v_1)_e &= \frac{1}{k!}\int_e (x - x_e)^k (u_1 - \pi_1 u_1)\partial_x^k v_1(x_e, y) dxdy \\ &+ \frac{1}{(k+1)!}\int_e (x - x_e)^{k+1}(u_1 - \pi_1 u_1)\partial_x^{k+1} v_1(x_e, y) dxdy \\ &= I_1 + I_2, \end{aligned} \tag{4.11}$$

where I_j, $j = 1, 2$, are defined accordingly. Notice that $\partial_x^k v_1(x_e, y) \in Q_{0,k}$. By letting $\psi = (u_1 - \pi_1 u_1)\partial_x^k v_1(x_e, y)$ we see that conditions of Lemma 4.1 are satisfied with $\tilde{k} = k - 1$ and $m = k$. Thus, it follows from (4.3) that

$$I_1 = \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2} u_1 \partial_x^k v_1(x_e, y) dxdy. \tag{4.12}$$

Since $\partial_x^k v_1(x, y)$ is linear in x, then

$$\partial_x^k v_1(x_e, y) = \partial_x^k v_1(x, y) + (x_e - x)\partial_x^{k+1} v_1(x, y).$$

Substituting the above into (4.12) gives

$$\begin{aligned} I_1 &= \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^k v_1(x,y)dxdy \\ &\quad - \frac{(-1)^k}{(2k+2)!}\int_e (x-x_e)E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k+1}v_1(x,y)dxdy. \end{aligned} \tag{4.13}$$

To deal with the second term of (4.13), we use

$$(x-x_e)E^{k+1}(x) = \frac{1}{2k+4}\partial_x E^{k+2}(x)$$

to obtain

$$\begin{aligned} &\int_e (x-x_e)E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k+1}v_1(x,y)dxdy \\ &= \frac{1}{2k+4}\int_e \partial_x E^{k+2}(x)\partial_x^{k+2}u_1\partial_x^{k+1}v_1(x,y)dxdy \\ &= \frac{-1}{2k+4}\int_e E^{k+2}(x)\partial_x^{k+3}u_1\partial_x^{k+1}v_1(x,y)dxdy. \end{aligned} \tag{4.14}$$

Substituting (4.14) into (4.13) yields

$$\begin{aligned} I_1 &= \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^k v_1(x,y)dxdy \\ &\quad + \frac{(-1)^k}{(2k+4)(2k+2)!}\int_e E^{k+2}(x)\partial_x^{k+3}u_1\partial_x^{k+1}v_1(x,y)dxdy \\ &= I_{11} + I_{12}. \end{aligned} \tag{4.15}$$

With any given smooth function v_2, we rewrite the first term I_{11} of I_1 as follows:

$$\begin{aligned} I_{11} &= \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}(\partial_x v_1 + \partial_y v_2)dxdy \\ &\quad - \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}\partial_y v_2 dxdy. \end{aligned} \tag{4.16}$$

The second term above can be further simplified by using the integration by parts in y, yielding

$$\begin{aligned} I_{11} &= \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}\nabla\cdot\mathbf{v}\ dxdy \\ &\quad + \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_y\partial_x^{k+2}u_1\partial_x^{k-1}v_2 dxdy \\ &\quad + \frac{(-1)^k}{(2k+2)!}\left(\int_{\ell_4} - \int_{\ell_2}\right)E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}v_2 dx. \end{aligned} \tag{4.17}$$

Combining the above equations we see that

$$(u_1 - \pi_1 u_1, v_1)_e = I_1 + I_2 = I_{11} + I_{12} + I_2,$$

where I_{11} is given in (4.17), I_{12} can be seen in (4.15), and I_2 is the corresponding integral in (4.11). By letting $J_{1,e}$ represent the combined area integrals,

$$\begin{aligned} J_{1,e} = I_{12} + I_2 &+ \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}\nabla\cdot\mathbf{v}dxdy \\ &+ \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_y\partial_x^{k+2}u_1\partial_x^{k-1}v_2dxdy, \end{aligned} \tag{4.18}$$

we arrive at

$$(u_1 - \pi_1 u_1, v_1)_e = J_{1,e} + \frac{(-1)^k}{(2k+2)!}\left(\int_{\ell_4} - \int_{\ell_2}\right) E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}v_2dx.$$

Now we estimate the total area integrals $J_{1,e}$. The term I_{12} is given in (4.15), which is already well expressed. The term I_2 can be estimated by using Lemma 4.1. To this end, observe that the conditions of Lemma 4.1 are satisfied with $\tilde{k} = k$, $m = k+1$, and $\psi = (u_1 - \pi_1 u_1)\partial_x^{k+1}v_1$. Thus, it follows from Lemma 4.1 that

$$\begin{aligned} I_2 &= \frac{1}{(k+1)!}\int_e (x - x_e)^{k+1}(u_1 - \pi_1 u_1)\partial_x^{k+1}v(x,y)dxdy \\ &= \frac{(-1)^{k+1}}{(2k+4)!}\int_e E^{k+2}(x)\partial_x^{k+3}u_1\partial_x^{k+1}v_1dxdy. \end{aligned} \tag{4.19}$$

Substituting the above into (4.18) we obtain

$$\begin{aligned} J_{1,e} = I_{12} &+ \frac{(-1)^{k+1}}{(2k+4)!}\int_e E^{k+2}(x)\partial_x^{k+3}u_1\partial_x^{k+1}v_1dxdy \\ &+ \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_x^{k+2}u_1\partial_x^{k-1}\nabla\cdot\mathbf{v}dxdy \\ &+ \frac{(-1)^k}{(2k+2)!}\int_e E^{k+1}(x)\partial_y\partial_x^{k+2}u_1\partial_x^{k-1}v_2dxdy. \end{aligned} \tag{4.20}$$

Since $|E(x)| \le h_e^2$, using the Hölder inequality, we obtain the following estimate:

$$\begin{aligned} |J_{1,e}| \le \frac{h_e^{2k+2}}{(2k+2)!}\Bigg(&\frac{h_e^2}{2k+3}|u_1|_{k+3,p;e}|v_1|_{k+1,q;e} \\ &+ |u_1|_{k+2,p;e}|\nabla\cdot\mathbf{v}|_{k-1,q;e} + |u_1|_{k+3,p;e}|v_2|_{k-1,q;e}\Bigg), \end{aligned} \tag{4.21}$$

which is the desired inequality in Lemma 4.2. ∎

A similar estimate can be derived for the second linear form on the right-hand side of (4.8). The result is stated as follows without any proof.

LEMMA 4.3. *Let $k \ge 1$ be an integer and u_2 be a sufficiently smooth function on the rectangle depicted in Figure* 1. *Let v_2 be a polynomial of degree no more than $k+1$ in y and k in x, respectively. Let v_1 be any sufficiently smooth function on e*

and $\mathbf{v} = (v_1, v_2)$. *Then there is a term* $J_{2,e}$ *such that*

$$(u_2 - \pi_2 u_2, v_2)_e = J_{2,e} + \frac{(-1)^k}{(2k+2)!}\left(\int_{\ell_3} - \int_{\ell_1}\right) E^{k+1}(y)\partial_y^{k+2} u_2 \partial_y^{k-1} v_1 dy. \tag{4.22}$$

The term $J_{2,e}$ *can be represented as area integrals over the element* e *with the following estimate:*

$$\begin{aligned} |J_{2,e}| \le \frac{\tau_e^{2k+2}}{(2k+2)!}\Bigg(&\frac{\tau_e^2}{2k+3}|u_2|_{k+3,p;e}|v_2|_{k+1,q;e} \\ &+ |u_2|_{k+2,p;e}|\nabla\cdot\mathbf{v}|_{k-1,q;e} + |u_2|_{k+3,p;e}|v_1|_{k-1,q;e}\Bigg), \end{aligned} \tag{4.23}$$

where $2\tau_e = y_{j+1} - y_j$ *is the height of the element and* q *is the conjugate of* $p \ge 1$ *satisfying* $1/p + 1/q = 1$.

5. Superconvergence. Our objective here is to derive some superconvergence for the mixed finite element methods by combining Theorem 3.1 with the estimates established in the previous section. From Theorem 3.1, it is sufficient to investigate the condition (3.1) with a maximum value for the parameter s.

Let us recall that the linear form $\mathcal{F}(\mathbf{v})$ consists of two components:

$$\mathcal{F}_1(v_1) = (\alpha_1(u_1 - \pi_1 u_1), v_1), \qquad \mathcal{F}_2(v_2) = (\alpha_2(u_2 - \pi_2 u_2), v_2).$$

We intend to deal with $\mathcal{F}_1(v_1)$ and $\mathcal{F}_2(v_2)$ by using Lemmas 4.2 and 4.3, respectively.

Assume that α_1 and α_2 are two constants in the computational domain Ω. It is easy to see that

$$\mathcal{F}_1(v_1) = \alpha_1 \sum_{e\in\mathcal{T}_h} (u_1 - \pi_1 u_1, v_1)_e.$$

On each element e, Lemma 4.2 can be employed to give

$$\mathcal{F}_1(v_1) = \alpha_1 \sum_{e\in\mathcal{T}_h} J_{1,e} + \alpha_{1,k} \sum_{e\in\mathcal{T}_h} \left(\int_{\ell_4} - \int_{\ell_2}\right) E^{k+1}(x)\partial_x^{k+2} u_1 \partial_x^{k-1} v_2 dx, \tag{5.1}$$

where $\alpha_{1,k} = \frac{(-1)^k \alpha_1}{(2k+2)!}$ and v_2 is chosen to be the second component of the vector-valued function $\mathbf{v}$. Observe that the line integral over ℓ_4 is given as an integral along the bottom edge and the integral over ℓ_2 is given as one on the top edge of e. If ℓ_4 is not a boundary edge, then there will be another element, say $\tilde{e}$, for which ℓ_4 is seen as the top edge and the corresponding integral on ℓ_4 has an opposite sign from the contribution of e. Thus, all the line integrals over interior edges must cancel each other in the second summation of (5.1). The line integrals on the boundary edges vanish due to the fact that $v_2 = 0$ along all the horizontal boundary edges (which stems from $\mathbf{v} \in \mathbf{V}_{0,h}$). Thus, we obtain

$$\mathcal{F}_1(v_1) = \alpha_1 \sum_{e\in\mathcal{T}_h} J_{1,e}. \tag{5.2}$$

Now we use the estimate (4.10) with $p = q = 2$ to obtain

$$
\begin{aligned}
|\mathcal{F}_1(v_1)| &\le \alpha_1 \sum_{e\in\mathcal{T}_h} |J_{1,e}| \\
&\le \alpha_1 \sum_{e\in\mathcal{T}_h} \frac{h_e^{2k+2}}{(2k+2)!} \left(\frac{h_e^2}{2k+3} |u_1|_{k+3,2;e} |v_1|_{k+1,2;e} \right. \\
&\qquad \left. + |u_1|_{k+2,2;e} |\nabla\cdot\mathbf{v}|_{k-1,2;e} + |u_1|_{k+3,2;e} |v_2|_{k-1,2;e} \right) \qquad (5.3) \\
&\le \alpha_1 \frac{\tilde{h}^{2k+2}}{(2k+2)!} \left(\frac{\tilde{h}^2}{2k+3} |u_1|_{k+3,2,h} |v_1|_{k+1,2,h} \right. \\
&\qquad \left. + |u_1|_{k+2,2,h} |\nabla\cdot\mathbf{v}|_{k-1,2,h} + |u_1|_{k+3,2,h} |v_2|_{k-1,2,h} \right),
\end{aligned}
$$

where $\tilde{h} = \max_{e\in\mathcal{T}_h} h_e$. By applying the standard inverse inequality to various norms of $\mathbf{v}$ in (5.3), we arrive at

$$
\begin{aligned}
|\mathcal{F}_1(v_1)| &\le C\tilde{h}^{k+3} \|u_1\|_{k+3,2;\Omega} \left(\|\mathbf{v}\|_0 + \|\nabla\cdot\mathbf{v}\|_0 \right) \\
&\le C\tilde{h}^{k+3} \|u_1\|_{k+3,2;\Omega} \|\mathbf{v}\|_{\mathbf{V}}.
\end{aligned} \qquad (5.4)
$$

The above argument can be extended to the linear form $\mathcal{F}_2(v_2)$ by using Lemma 4.3. With $\tilde{\tau} = \max_{e\in\mathcal{T}_h} \tau_e$, the corresponding estimate is given by

$$
|\mathcal{F}_2(v_2)| \le C\tilde{\tau}^{k+3} \|u_2\|_{k+3,2;\Omega} \|\mathbf{v}\|_{\mathbf{V}}. \qquad (5.5)
$$

Substituting (5.4) and (5.5) into (4.8), we obtain

$$
|\mathcal{F}(\mathbf{v})| \le C(\tilde{h}^{k+3} + \tilde{\tau}^{k+3}) \|\mathbf{u}\|_{k+3,2;\Omega} \|\mathbf{v}\|_{\mathbf{V}}.
$$

Thus, the condition (3.1) holds true with $s = 3$ and $h = \max(\tilde{h}, \tilde{\tau})$. To summarize, we have proved the following superconvergence result.

THEOREM 5.1. *Assume that* $(\mathbf{u}; p)$ *solves the mixed problem* (2.2) *and* $\mathbf{u} \in \left[H^{k+3}\right]^2$. *Let the tensor* $\boldsymbol{a}^{-1}$ *in* (2.2) *be given by* (4.7) *with constant entries* α_i. *Let* $(\mathbf{u}_h; p_h)$ *solve* (2.3). *Then there is a generic constant* C *independent of* h *and the solution* p *such that*

$$
\|\mathbf{u}_h - \pi_h\mathbf{u}\|_0 + \|p_h - P_h p\|_0 \le Ch^{k+3} \|\mathbf{u}\|_{k+3,2;\Omega}. \qquad (5.6)
$$

The cancellation of line integrals in (5.1) is crucial for the superconvergence estimate (5.6). Let us recall that in (5.1), the line segments ℓ_2 and ℓ_4 correspond to the top and bottom edge of the element e (see Figure 1). The cancellation of the line integral over ℓ_2 occurs if either ℓ_2 is on the boundary of Ω (hence, $v_2 = 0$) or there is another integral over ℓ_2 with opposite sign that was contributed from the element $\tilde{e}$ right above e. In the later case, the function $\alpha_1 = \alpha_1(x, y)$ must have the same value on e and $\tilde{e}$ in order to have a complete cancellation. Thus, all the line integrals in (5.1) cancel each other as long as the function $\alpha_1 = \alpha_1(x, y)$ is a constant on each vertical strip $\{[x_i, x_{i+1}] \times (-\infty, \infty)\} \cap \Omega$. Similarly, the function $\alpha_2 = \alpha_2(x, y)$ must be a constant on each horizontal strip $\{(-\infty, \infty) \times [y_j, y_{j+1}]\} \cap \Omega$ in order to have a full cancellation of the line integrals in the linear form $\mathcal{F}_2(v_2)$. The remaining area

integrals can be handled in the same manner by using the estimates (4.10) and (4.23). The result is summarized as follows.

THEOREM 5.2. *Assume that* $(\mathbf{u};p)$ *solves the mixed problem* (2.2) *and* $\mathbf{u} \in \left[H^{k+3}\right]^2$. *Let the tensor* $\boldsymbol{a}^{-1}$ *be given in* (4.7), *where* $\alpha_1 = \alpha_1(x)$ *and* $\alpha_2 = \alpha_2(y)$ *are piecewise constant functions. Let* $(\mathbf{u}_h;p_h)$ *solve* (2.3). *Then there is a generic constant* C *independent of* h *and the solution* p *such that*

$$\|\mathbf{u}_h - \pi_h\mathbf{u}\|_0 + \|p_h - P_hp\|_0 \le Ch^{k+3}\|\mathbf{u}\|_{k+3,2;\Omega}. \tag{5.7}$$

In the rest of this section, we derive a superconvergence for (2.3) when the tensor $\boldsymbol{a}^{-1}$ has piecewise constant entries in both x and y direction. To this end, let $\bar{\Omega} = \bigcup_{s=1}^{m} \bar{\Omega}_s$ be a nonoverlapping coarse partition of the domain Ω. Assume that the entries $\alpha_i = \alpha_i(x,y)$, $i = 1,2$, have constant values on each subdomain Ω_s. Assume that the finite element partition $\mathcal{T}_h$ aligns with the above coarse partition in the sense that each element $e = [x_i, x_{i+1}] \times [y_j, y_{j+1}] \in \mathcal{T}_h$ intersects with one and only one of the subdomains Ω_s, $s = 1,\ldots,m$.

How do the line integrals cancel each other in the representation (5.1) for the linear form $\mathcal{F}(\mathbf{v})$? It is clear that all the line integrals which are interior to any subdomain Ω_s cancel each other in the way that was explained earlier. But the line integrals along the boundary of Ω_s will not go away from (5.1) because there is no counterpart from their neighbors for a complete cancellation. Let Γ_s be the part of the boundary of Ω_s where the line integrals do not vanish. Using the Hölder inequality, such line integrals can be estimated by

$$B = Ch^{2k+2}|u_1|_{k+2,\infty;\Omega_s}\int_{\Gamma_s} |\partial_\gamma^{k-1}v_2|d\Gamma_s, \tag{5.8}$$

where ∂_γ is the tangential derivative operator along Γ_s. It is routine to show that

$$\int_{\Gamma_s} |\partial_\gamma^{k-1}v_2|d\Gamma_s \le C\left(h^{-1/2}|v_2|_{k-1,2;\Omega_s} + h^{1/2}|v_2|_{k,2;\Omega_s}\right). \tag{5.9}$$

By first substituting the above into (5.8) and then using the inverse inequality on $|v_2|_{k-1,2;\Omega_s}$ and $|v_2|_{k,2;\Omega_s}$, we obtain

$$B \le Ch^{k+2.5}|u_1|_{k+2,\infty;\Omega_s}\|v_2\|_{0,2;\Omega_s}.$$

The corresponding superconvergence is summarized as follows.

THEOREM 5.3. *Assume that* $(\mathbf{u};p)$ *solves the mixed problem* (2.2) *and* $\mathbf{u} \in \left[H^{k+3}\right]^2$. *Let the tensor* $\boldsymbol{a}^{-1}$ *be given in* (4.7), *where* $\alpha_1 = \alpha_1(x,y)$ *and* $\alpha_2 = \alpha_2(x,y)$ *are piecewise constant functions on* Ω. *Let* $(\mathbf{u}_h;p_h)$ *solve* (2.3). *Then there is a generic constant* C *independent of* h *and the solution* p *such that*

$$\|\mathbf{u}_h - \pi_h\mathbf{u}\|_0 + \|p_h - P_hp\|_0 \le Ch^{k+2.5}\left(\|\mathbf{u}\|_{k+2,\infty;\Omega} + \|\mathbf{u}\|_{k+3,2;\Omega}\right). \tag{5.10}$$

6. Postprocessing by patch recovery. The results developed in Theorems 5.1–5.3 indicate a supercloseness between the mixed finite element approximation $\mathbf{u}_h$ and the locally defined projection $\pi_h\mathbf{u}$ of the exact flux variable $\mathbf{u}$. This supercloseness estimate does not mean any superconvergence between $\mathbf{u}$ and its finite element approximation $\mathbf{u}_h$ because no superconvergence is known between $\pi_h\mathbf{u}$ and the exact solution $\mathbf{u}$. In general, $\pi_h\mathbf{u}$ does not have any superconvergence to $\mathbf{u}$ globally in the L^2-norm.

Our objective in this section is to provide a new approximation based on a patch recovery approach that postprocesses $\mathbf{u}_h$ locally on each element. A general framework can be described as follows. Let $\mathbf{u}_h$ be a certain finite element approximation of the exact solution $\mathbf{u}$ with the following error estimate:

$$\|\mathbf{u}_h - \pi_h \mathbf{u}\|_0 \leq Ch^{k+3}\|\mathbf{u}\|_{k+3}, \tag{6.1}$$

where $\pi_h\mathbf{u}$ is a projection of $\mathbf{u}$ into the same finite element space. The interpolation error between $\mathbf{u}$ and $\pi_h\mathbf{u}$ is assumed to be worse than the order $\mathcal{O}(h^{k+3})$. However, thanks to the estimate (6.1) and the locality of $\pi_h\mathbf{u}$, it is possible to construct a new approximate solution based on $\mathbf{u}_h$ which approximates $\mathbf{u}$ with the superconvergence order of $\mathcal{O}(h^{k+3})$. This new approximate solution is often realized through an operator Q_h from the finite element space to a new finite element space consisting of high order (e.g., of order $k+s$ on each element) of polynomials with the following property:

$$Q_h\pi_h\mathbf{u} = Q_h\mathbf{u}. \tag{6.2}$$

If so, the accuracy of the new approximate solution $Q_h\mathbf{u}_h$ can be justified as follows:

$$\begin{aligned} \|\mathbf{u} - Q_h\mathbf{u}_h\|_0 &\leq \|\mathbf{u} - Q_h\mathbf{u}\|_0 + \|Q_h\mathbf{u} - Q_h\mathbf{u}_h\|_0 \\ &\leq \|\mathbf{u} - Q_h\mathbf{u}\|_0 + \|Q_h(\pi_h\mathbf{u} - \mathbf{u}_h)\|_0 \\ &\leq Ch^{k+s+1}\|\mathbf{u}\|_{k+s+1} + Ch^{k+3}\|\mathbf{u}\|_{k+3}. \end{aligned} \tag{6.3}$$

Here we have assumed the L^2-boundedness of the operator Q_h. In general, the boundedness of Q_h depends on the property of the corresponding finite element partition and the choice of interpolating spaces.

The rest of this section will give a constructive approach to a postprocessing operator Q_h for the mixed finite element approximations.

6.1. Some technical tools. It is well known that polynomials can be uniquely determined by their values at a set of points. In the mixed finite element method, the local projection operator π_h is defined by using both line and area moments which must be examined carefully in the construction of our postprocessing operator Q_h. Let us provide some new characterization for polynomials by using line moments. To this end, let

$$M_0(x) = 1 - x, \quad M_1(x) = 1 + x, \quad M_j(x) = \frac{d^{j-2}\left((1-x^2)^{j-1}\right)}{dx^{j-2}}, \qquad j \geq 2,$$

be a set of polynomials. The set of derivatives $\{L_j(x) \equiv M_{j+1}^{'}(x),\ j = 0, 1, \dots\}$ is the set of well-known Legendre polynomials.

LEMMA 6.1. *Let $I = [-1, 1]$ and x_0 be a fixed, but arbitrary, point in $(-1, 1)$. Then any polynomial $w = w(x) \in P_{k+2}(I)$ is uniquely determined by the following set of degrees of freedom:*

$$w(\pm 1), \quad \int_{-1}^{1} w(s)\phi(x)dx, \quad \int_{-1}^{x_0} w(s)(x - x_0)^{k-1}dx \tag{6.4}$$

for all $\phi \in P_{k-1}(I)$.

Proof. The total number of degrees of freedom in (6.4) is given by $2+k+1 = k+3$, which is the same as the dimension of the polynomial space $P_{k+2}(I)$. It is therefore

sufficient to verify the uniqueness. To this end, let $w \in P_{k+2}(I)$ satisfy

$$w(\pm 1) = 0, \tag{6.5}$$

$$\int_{-1}^{1} w(s)\phi(x)dx = 0 \qquad \forall \phi \in P_{k-1}(I), \tag{6.6}$$

$$\int_{-1}^{x_0} w(s)(x - x_0)^{k-1}dx = 0. \tag{6.7}$$

Let us show that $w \equiv 0$. We express $w = w(x)$ as follows:

$$w(x) = \sum_{i=0}^{k+2} \alpha_i M_i(x), \tag{6.8}$$

where α_i are real numbers. Notice that $M_j(\pm 1) = 0$ for any $j \geq 2$. Thus, the condition $w(\pm 1) = 0$ implies that $\alpha_0 = \alpha_1 = 0$.

Next, let us test (6.8) against any $M_j^{''}(x)$ for $2 \leq j \leq k+1$. It follows from the condition (6.6) that

$$\int_{-1}^{1} w(x)M_j^{''}(x)dx = 0.$$

On the other hand, using the integration by parts, we obtain

$$\int_{-1}^{1} w(x)M_j^{''}(x)dx = -\int_{-1}^{1} w'(x)M_j^{'}(x)dx = \alpha_j c_j$$

for some constant $c_j \neq 0$. Hence, $\alpha_j = 0$ for any $2 \leq j \leq k+1$.

It remains to show that $\alpha_{k+2} = 0$. To this end, we use the condition (6.7) to obtain

$$\begin{aligned}
0 &= \int_{-1}^{x_0} w(x)(x - x_0)^{k-1}dx \\
&= \alpha_{k+2} \int_{-1}^{x_0} M_{k+2}(x)(x - x_0)^{k-1}dx \\
&= \alpha_{k+2} \int_{-1}^{x_0} \left((1 - x^2)^{k+1}\right)^{(k)} (x - x_0)^{k-1}dx \\
&= \alpha_{k+2}(-1)^{k-1}(k-1)! \int_{-1}^{x_0} ((1 - x^2)^{k+1})^{'}dx \\
&= \alpha_{k+2}(-1)^{k-1}(k-1)!(1 - x_0^2)^{k+1},
\end{aligned}$$

which implies that $\alpha_{k+2} = 0$. This completes the proof. □

Our next result will determine a polynomial of degree $k+2$ by using a different set of degrees of freedom.

LEMMA 6.2. *Let $w = w(x) \in P_{k+2}(I)$ be a polynomial of degree $k+2$ satisfying*

$$\int_{-1}^{1} w(s)\phi(x)dx = 0 \qquad \forall \phi \in P_k(I), \tag{6.9}$$

$$\int_{-1}^{x_0} (x - x_0)^{k-i} w(x)dx = 0, \qquad i = 0, 1, \tag{6.10}$$

where $x_0 \in (-1, 1)$ is a fixed, but arbitrary, point. Then one has $w \equiv 0$.

Proof. The polynomial $w = w(x)$ can be represented by using the Legendre polynomials as follows:

$$w(x) = \sum_{i=0}^{k+2} \alpha_i L_i(x),$$

where $L_i(x) = M_{i+1}^{'}(x)$ is the Legendre polynomial of order i. From (6.9) we have

$$\alpha_i = 0, \qquad i = 0, 1, \ldots, k.$$

It remains to show that $\alpha_{k+1} = \alpha_{k+2} = 0$. To this end, we use the second momentum condition (6.10) to obtain

$$\begin{aligned}
0 &= \int_{-1}^{x_0} (x - x_0)^k w(x)dx \\
&= \sum_{i=k+1}^{k+2} \alpha_i \int_{-1}^{x_0} (x - x_0)^k L_i(x) \\
&= \sum_{i=k+1}^{k+2} \alpha_i \int_{-1}^{x_0} (x - x_0)^k M_{i+1}^{'}(x) \\
&= (-1)^k k! \alpha_{k+1} (1 - x_0^2)^{k+1} + \alpha_{k+2} (-1)^k k! \int_{-1}^{x_0} \frac{d^2 \left((1 - x^2)^{k+2}\right)}{dx^2} dx \\
&= (-1)^k k! (1 - x_0^2)^{k+1} \left(\alpha_{k+1} - 2x_0 \alpha_{k+2}(k+2)\right)
\end{aligned}$$

and

$$\begin{aligned}
0 &= \int_{-1}^{x_0} (x - x_0)^{k-1} w(x)dx \\
&= \sum_{i=k+1}^{k+2} \alpha_i \int_{-1}^{x_0} (x - x_0)^{k-1} M_{i+1}^{'}(x) \\
&= (-1)^{k-1} (k-1)! \left(\alpha_{k+1} \frac{d\left((1 - x^2)^{k+1}\right)}{dx} + c_{k+2} \frac{d^2 (1 - x^2)^{k+2}}{dx^2} \right) \Bigg|_{x = x_0} \\
&= 2(-1)^k (k-1)! (1 - x_0^2)^k \left(\alpha_{k+1}(k+1)x_0 + \alpha_{k+2}(k+2)(1 - 2x_0^2(k+1)) \right).
\end{aligned}$$

The above two equations imply that

$$\begin{aligned}
\alpha_{k+1} - 2x_0(k+2)\alpha_{k+2} &= 0, \\
x_0(k+1)\alpha_{k+1} + (k+2)(1 - 2x_0^2(k+1))\alpha_{k+2} &= 0,
\end{aligned}$$

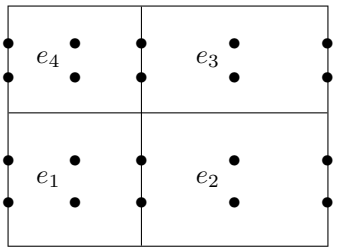


FIG. 2. *A patch K of four rectangular elements.*

which has only a trivial solution since the determinant of the coefficient matrix is

$$(k+2)(1-2x_0^2(k+1))+2x_0^2(k+1)(k+2)=k+2\neq 0.$$

This completes the proof of the lemma. □

6.2. Patch recovery for mixed finite element approximations. Recall that the general theory requires a locally defined operator Q_h satisfying the property (6.2). Here we construct such an operator by using the technical results developed in the previous subsection.

Let $e_1 \in \mathcal{T}_h$ be any element in the finite element rectangular partition. First, we find three neighboring elements, e_i, $i = 2, 3, 4$, such that $K = \bigcup_{i=1}^4 e_i$ forms a larger rectangle (it is a patch of rectangular elements). Second, we construct a polynomial of degree $k + 2$ in both x and y direction on the patch K by using certain degrees of freedom. Does $\pi_h \mathbf{u}$ contain enough information on K to guarantee the existence of an operator Q_h satisfying (6.2)? To answer this question, let us take for example $k = 1$ and consider only the first component of $\pi_h \mathbf{u}$. As a piecewise polynomial, the first component $\pi_h u_1$ has 20 degrees of freedom on K (dotted points in Figure 2). On the other hand, the new interpolation space $Q_{3,3}(K)$ has dimension 16. This simple dimension counting indicates that $\pi_h \mathbf{u}$ *may* contain enough information for the construction of a projection operator Q_h satisfying (6.2). The rest of this section gives a rigorous approach.

For the sake of discussion, let the patch K be given by

$$K = [-1, 1] \times [-1, 1],$$

which consists of four elements:

$$\begin{aligned} e_1 &:= [-1, x_0] \times [-1, y_0], \qquad & e_2 &= [x_0, 1] \times [-1, y_0], \\ e_3 &:= [x_0, 1] \times [y_0, 1], \qquad & e_4 &:= [-1, x_0] \times [y_0, 1]. \end{aligned}$$

Notice that the patch K is normally nonuniform. Our projection operator Q_h can be given as the product of two operators S_1^x and S_2^y defined in the x and the y direction, respectively. More precisely, the operator S_1^x takes a function to a polynomial $\phi = \phi(x)$ of degree $k + 2$ by using the degrees of freedom specified in Lemma 6.1 and the operator S_2^y makes use of the degrees of freedom given in Lemma 6.2. On the target element e_1, the first component of $Q_h \mathbf{u}$ is defined as the restriction of $S_2^y S_1^x u_1$ to e_1. Similarly, the second component of $Q_h \mathbf{u}$ is defined as the restriction of $S_2^x S_1^y u_2$ to e_1. This operator Q_h can be easily verified to satisfy the crucial condition (6.2) on the target element e_1. It is also not hard to see that Q_h is bounded in the L^2-norm

for regular partitions. Consequently, we have proved the following superconvergence result.

THEOREM 6.1. *Let Q_h be the locally defined operator outline as above and $\mathbf{u}_h$ be the mixed finite element approximation obtained from* (2.3)*. Under the same assumption of Theorem* 5.1*, there exists a constant C independent of the exact solution $(\mathbf{u}; p)$ and the mesh size h such that*

$$\|\mathbf{u} - Q_h\mathbf{u}_h\|_0 \leq Ch^{k+3}\|\mathbf{u}\|_{k+3}.$$

Similar superconvergence can be established for the mixed finite element method by using Theorems 5.2 and 5.3. The details are omitted.

REFERENCES

[1] I. BABUŠKA, *The finite element method with Lagrangian multiplier*, Numer. Math., 20 (1973), pp. 179–192.

[2] J. H. BRANDTS, *Superconvergence and a posteriori error estimation for triangular mixed finite elements*, Numer. Math., 68 (1994), pp. 311–324.

[3] F. BREZZI, *On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 2 (1974), pp. 129–151.

[4] F. BREZZI, J. DOUGLAS, R. DURÁN, AND L. MARINI, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 52 (1987), pp. 237–250.

[5] F. BREZZI, J. DOUGLAS, M. FORTIN, AND L. MARINI, *Efficient rectangular mixed finite elements in two and three spaces variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[6] F. BREZZI, J. DOUGLAS, AND L. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer–Verlag, New York, 1991.

[8] J. DOUGLAS AND J. ROBERT, *Global estimate for mixed finite elements methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[9] J. DOUGLAS AND J. WANG, *Superconvergence of mixed finite element spaces on rectangular domains*, Calcolo, 26 (1989), pp. 121–134.

[10] J. DOUGLAS AND J. WANG, *A new family of spaces in mixed finite element methods for rectangular elements*, Comput. Appl. Math., 12 (1993), pp. 183–197.

[11] R. DURÁN, *Superconvergence for rectangular mixed finite elements*, Numer. Math., 58 (1990), pp. 287–298.

[12] R. E. EWING, M. M. LIU, AND J. WANG, *Superconvergence of mixed finite element approximations over quadrilaterals*, SIAM J. Numer. Anal., 36 (1999), pp. 772–787.

[13] R. E. EWING, R. D. LAZAROV, AND J. WANG, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[14] R. E. EWING AND J. SHEN, *private communication*, 1989.

[15] P. RAVIART AND J. THOMAS, *A mixed finite element method for 2nd order elliptic problems*, in Mathematics Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer–Verlag, New York, 1977, pp. 292–315.

[16] J. WANG, *Superconvergence and extrapolation for mixed finite element methods on rectangular domains*, Math. Comp., 56 (1991), pp. 477–503.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2151–2170

THE $\mathcal{P}^{K+1} - \mathcal{S}^K$ LOCAL DISCONTINUOUS GALERKIN METHOD FOR ELLIPTIC EQUATIONS*

CLINT DAWSON†

Abstract. A local discontinuous Galerkin (LDG) method using approximating spaces of order $k+1$ for the gradient and order k for the state variable, with $k \geq 0$, is presented for an elliptic boundary value problem. A priori error estimates are derived and numerical results presented, which demonstrate that this method is order $k+1$ for both the gradient and state variable in the L^2 norm. This improves upon earlier LDG methods which use equal order approximating spaces for both variables, but lose an order of accuracy in the gradient approximation. This approach is also the first convergent LDG method which allows for piecewise constant approximations for the state variable.

Key words. discontinuous Galerkin method, elliptic flow problem, a priori error estimate

AMS subject classifications. 65M12, 65M30

PII. S0036142901397599

1. Introduction.

In this paper, we consider the solution $(\mathbf{u}, p)$ of the second order elliptic problem,

$$cp + \nabla \cdot \mathbf{u} = f \qquad \text{in } \Omega, \tag{1}$$

$$\mathbf{u} = -K\nabla p \qquad \text{in } \Omega, \tag{2}$$

satisfying the boundary conditions

$$p = g_D \qquad \text{on } \partial\Omega, \tag{3}$$

by a local discontinuous Galerkin (LDG) approach. Because of the similarity to porous media flow, we will often refer to the variable p as a fluid "pressure" and $\mathbf{u}$ as fluid "velocity." In this context, c is a compressibility factor and K a permeability tensor. We assume K is uniformly symmetric and positive definite, with

$$K^* \geq K^{-1} \geq K_* > 0, \tag{4}$$

and throughout most of the paper, we assume

$$c^* \geq c(x) \geq c_* > 0. \tag{5}$$

We will consider the case $c = 0$ in section 6.

The LDG method is a type of classical mixed method, whereby the state variable (p) and its flux ($\mathbf{u}$) are simultaneously approximated. The LDG method was introduced by Cockburn and Shu in [13], and since then has been rigorously studied by a number of researchers for convection-diffusion problems [9, 1, 8, 15], purely elliptic problems [7, 11, 6], and Stokes problems [12]. The analysis and implementation of the LDG method to date has primarily used equal order approximating spaces for

*Received by the editors November 6, 2001; accepted for publication (in revised form) May 7, 2002; published electronically December 13, 2002. This work was funded in part by the National Science Foundation, projects DMS-9805491, DMS-9873326, and DMS-0107247.
http://www.siam.org/journals/sinum/40-6/39759.html
†Texas Institute for Computational and Applied Mathematics, The University of Texas at Austin, Austin, TX 78712 (clint@ticam.utexas.edu).

the state variable and the gradient. We refer to this approach as the $\mathcal{P}^k - \mathcal{P}^k$ LDG method, where $\mathcal{P}^k$ denotes the set of complete polynomials of degree k. This method has been shown to give optimal order (h^{k+1}) convergence for p in the L^2 norm, but loses an order of accuracy for $\mathbf{u}$. It also requires that $k \geq 1$. In [12], an LDG method with the gradient approximated in $\mathcal{P}^{k-1}$ and the state variable in $\mathcal{P}^k$ was investigated for the Stokes system, but no advantage in efficiency was seen over the $\mathcal{P}^k - \mathcal{P}^k$ method.

The initial motivation for the work presented here was to develop an LDG method that allows for the pressure to be approximated in $\mathcal{P}^0$, as the ability to use piecewise constant approximations is often desirable in practice. This led to the development of the so-called $\mathcal{P}^1 - \mathcal{P}^0$ LDG method [14], which uses a piecewise constant approximation of p and a piecewise linear approximation of $\mathbf{u}$. This approach relies on the construction of a sufficiently accurate and stable numerical flux on the interface between elements. The construction of this flux differs from that proposed in any discontinuous Galerkin (DG) method to date. In fact, the method does not fall within the unified approach for DG methods given in [2], and thus cannot be analyzed using the techniques described therein. Still, first order convergence in L^2 for p and $\mathbf{u}$ was demonstrated in [14].

In this paper, we outline an extension of the $\mathcal{P}^1 - \mathcal{P}^0$ method to any order $k \geq 0$. We call the resulting approach the $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG method. In this approach, $\mathbf{u}$ is approximated by polynomials of degree $k+1$, and p is approximated in a space $\mathcal{S}^k$ predominantly consisting of polynomials of degree k, although some terms of order $k+1$ may be needed in the approximation of p when $k > 0$, as we will see below. Thus, in general $\mathcal{S}^k$ is somewhere "between" $\mathcal{P}^k$ and $\mathcal{P}^{k+1}$, but in the case $k = 0$, $\mathcal{S}^0 = \mathcal{P}^0$. A primary motivation for this approach is that, unlike the standard LDG methods mentioned above, pressure and velocity are approximated to the same order of accuracy. In many physical applications, porous media flow being a prime example, it is the velocity $\mathbf{u}$ that is the quantity of interest, therefore it is important to approximate this variable as accurately as p is approximated. The $\mathcal{P}^{k+1} - \mathcal{S}^k$ method gives h^{k+1} convergence for both variables, while still requiring only the solution of a linear system for the approximation to p. As in the $\mathcal{P}^1 - \mathcal{P}^0$ method, a necessary ingredient in our approach is the construction of a stable, accurate numerical flux. We will also improve upon our earlier results in [14] for the $\mathcal{P}^1 - \mathcal{P}^0$ method by extending our analysis to the case $c = 0$ in (1).

Since the LDG method is a type of classical mixed method, the $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG scheme is very similar in spirit to the mixed finite element method as described in, for example, [20, 5]. Mixed finite element approximating spaces, for example the Raviart–Thomas spaces, typically employ polynomials of degree $k+1$ for the velocity and k for pressure, with convergence rates of order h^{k+1} for both variables. In the standard mixed method, the velocities are constrained to be in $H(\Omega; \mathrm{div})$, which requires that the normal component of velocity must be continuous across element faces. In the LDG method, we drop this continuity requirement on the normal flux, the result being that the velocity unknowns can be resolved locally (element by element) in terms of the pressure unknowns.

In the next section, we will outline the $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG method. We then give an error analysis of the method in standard L^2 norms. In section 4, we discuss in some detail the construction of a numerical flux, $\widehat{w}$, which is crucial to the method. In section 5, we present some numerical results which confirm our theory. Finally, in section 6, we consider the case $k = 0$ again with the assumption that the coefficient

$c = 0$.

2. The $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG method. In this section, we define the $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG method. First, we introduce some notation.

Let $\mathcal{T}_h$ denote a triangulation of Ω into elements Ω_e. We will make the following fairly standard assumptions on $\mathcal{T}_h$; see [4]. We assume each element Ω_e is affinely equivalent to one of several reference elements. We also assume $\mathcal{T}_h$ is regular; that is, the ratio of the diameter h_e of Ω_e and the diameter ρ_e of the largest ball contained in Ω_e (the "chunkiness parameter" [4]) is bounded by a positive constant:

$$\frac{h_e}{\rho_e} \leq \sigma_1 \qquad \forall \Omega_e. \tag{6}$$

The interior element boundaries need not align; however, we assume that Ω_e has a fixed, maximum number of neighbors. A neighbor of Ω_e is defined to be any element $\Omega_{e'}$ such that

$$\text{interior of } \partial\Omega_e \cap \partial\Omega_{e'} \neq \emptyset. \tag{7}$$

For such neighboring elements, we will assume a local quasi uniformity; that is, we assume a positive constant $\sigma_2 < 1$ exists, independent of Ω_e and $\Omega_{e'}$, such that

$$\sigma_2 \leq \frac{h_e}{h_{e'}} \leq \sigma_2^{-1}. \tag{8}$$

Finally, let

$$h = \max_e h_e.$$

We denote by $(\cdot,\cdot)_R$ the usual L^2 inner product over a d dimensional domain R; to emphasize that the integration is over $(d-1)$ dimensional surfaces, we write $\langle \cdot,\cdot \rangle_{\partial R}$. We denote by $||\cdot||_R$ the L^2 norm on R. Norms in other Sobolev spaces $H(R)$ will be denoted by $||\cdot||_{H(R)}$.

Suppose Ω_e^- and Ω_e^+ are adjacent elements with unit outward normals $\mathbf{n}^-$ and $\mathbf{n}^+$, and $(\mathbf{v}, w)$ are smooth functions defined on these elements, with $\mathbf{v}$ vector-valued and w a scalar. Let $(\mathbf{v}^\pm, w^\pm)$ denote the traces of $(\mathbf{v}, w)$ on the face γ between Ω_e^+ and Ω_e^- from the interiors of the elements. We define the average $\{\cdot\}$ and jump $[\![\cdot]\!]$ for $\mathbf{x} \in \gamma$ as follows:

$$\{w\} = (w^- + w^+)/2, \qquad \{\mathbf{v}\} = (\mathbf{v}^- + \mathbf{v}^+)/2, \tag{9}$$

$$[\![w]\!] = w^+\mathbf{n}^+ + w^-\mathbf{n}^-, \qquad [\![\mathbf{v}]\!] = \mathbf{v}^+ \cdot \mathbf{n}^+ + \mathbf{v}^- \cdot \mathbf{n}^-. \tag{10}$$

Let W_h, $\mathbf{V}_h$ denote discontinuous, piecewise polynomial approximating spaces defined on the triangulation $\mathcal{T}_h$. The variable p will be approximated in the space W_h. We will assume W_h on Ω_e contains $\mathcal{P}^k(\Omega_e)$ and may also contain some but not necessarily all polynomial terms of degree $k+1$. For example, an xy term may be included with terms of degree one. We call the resulting space $\mathcal{S}^k$. Therefore,

$$W_h = \{w \in L^2(\Omega) : w|_{\Omega_e} \in \mathcal{S}^k(\Omega_e) \ \forall\ \Omega_e \in \Omega\},$$

where $\mathcal{P}^k(\Omega_e) \subset \mathcal{S}^k(\Omega_e) \subset \mathcal{P}^{k+1}(\Omega_e)$. The variable $\mathbf{u}$ will be approximated in the space $\mathbf{V}_h$, where

$$\mathbf{V}_h = \{\mathbf{v} \in L^2(\Omega)^d : \mathbf{v}|_{\Omega_e} \in (\mathcal{P}^{k+1}(\Omega_e))^d \ \forall\ \Omega_e \in \Omega\}.$$

We will utilize the fact that

$$\nabla \cdot \mathbf{V}_h|_{\Omega_e} \subset W_h|_{\Omega_e} \tag{11}$$

in our analysis.

The $\mathcal{P}^{k+1} - \mathcal{S}^k$ LDG solution $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times W_h$ is determined by requiring that

$$\begin{aligned} \mathcal{A}_{LDG}(\mathbf{u}_h, p_h; w_h) &\equiv \sum_e [(cp_h, w_h)_{\Omega_e} + (\nabla \cdot \mathbf{u}_h, w_h)_{\Omega_e}] - \langle [\![\mathbf{u}_h]\!], \widehat{w}_h \rangle_{\mathcal{E}_i} \\ &= \sum_e (f, w_h)_{\Omega_e}, \end{aligned} \tag{12}$$

$$\begin{aligned} \mathcal{B}_{LDG}(\mathbf{u}_h, p_h; \mathbf{v}_h) &\equiv \sum_e [(K^{-1}\mathbf{u}_h, \mathbf{v}_h)_{\Omega_e} - (p_h, \nabla \cdot \mathbf{v}_h)_{\Omega_e}] + \langle \widehat{p}_h, [\![\mathbf{v}_h]\!] \rangle_{\mathcal{E}_i} \\ &= -\langle g_D, \mathbf{v}_h \cdot \mathbf{n} \rangle_{\partial\Omega} \end{aligned} \tag{13}$$

for all $(\mathbf{v}_h, w_h) \in \mathbf{V}_h \times W_h$, where $\mathbf{n}$ is the outward unit normal to $\partial\Omega$, and $\mathcal{E}_i$ denotes the set of all interior faces in the triangulation $\mathcal{T}_h$. It remains to define the numerical flux $\widehat{w}_h$ on each interior face for $w_h \in W_h$.

The numerical flux $\widehat{w}_h$ should be easy to compute. We also need it to be sufficiently accurate, as we will see below. We will give specific examples of $\widehat{w}_h$ in section 4. For now, we outline two requirements on $\widehat{w}_h$ needed for the error analysis to go through.

The numerical flux $\widehat{w}_h$ should be constructed on each interior face γ from the degrees of freedom of w_h in an $\mathcal{O}(h)$ neighborhood of γ. Let $\{\Omega_{e_\gamma}\}$ denote the collection of elements adjacent to γ, that is, the interior of $\partial\Omega_{e_\gamma} \cap \gamma \neq \emptyset$, along with any other elements used to compute $\widehat{w}_h$. Let $h_\gamma = \max_{e_\gamma} h_{e_\gamma}$. The requirements on $\widehat{w}_h$ are outlined as follows.

(A1) We require $\widehat{w}_h$ to satisfy the trace (or stability) inequality,

$$||\widehat{w}_h||_\gamma^2 \leq C_1 h_\gamma^{-1} \sum_{e_\gamma} ||w_h||_{\Omega_{e_\gamma}}^2, \qquad w_h \in W_h, \tag{14}$$

where C_1 is independent of h. In the examples in section 4, we will see that C_1 can depend on σ_1 and σ_2.

(A2) For ϕ a smooth function, let $\pi\phi$ be the L^2 projection of ϕ into W_h:

$$(\phi - \pi\phi, w_h)_{\Omega_e} = 0, \qquad w_h \in W_h. \tag{15}$$

We require the following accuracy condition, namely,

$$||\phi - \widehat{\pi\phi}||_\gamma \leq C_{2,\gamma} h_\gamma^{k+3/2}, \tag{16}$$

where

$$\sum_\gamma C_{2,\gamma}^2 \equiv C_2^2, \tag{17}$$

and C_2 is independent of h. In general, C_2 will depend on the smoothness of ϕ (e.g., $\phi \in H^{k+2}(\Omega)$ as we will see in section 4) and the triangulation $\mathcal{T}_h$ (i.e., σ_1 and σ_2).

Returning to (12)–(13), note that in (13) the variable $\mathbf{u}_h$ can be easily eliminated from the equations since it can be expressed *locally* on each element Ω_e in terms of p_h

by choosing $\mathbf{v}_h$ such that it is zero everywhere except on Ω_e. Substituting into (12), the system can be reduced to a system in the variable p_h alone.

If we define the bilinear form

$$\mathcal{A}(\mathbf{q}, r; \mathbf{v}, w) = \mathcal{A}_{LDG}(\mathbf{q}, r; w) + \mathcal{B}_{LDG}(\mathbf{q}, r; \mathbf{v}) \tag{18}$$

and the linear form

$$\mathcal{C}(\mathbf{v}, w) = (f, w)_\Omega - \langle g_D, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial\Omega}, \tag{19}$$

then we can characterize the approximate solution $(\mathbf{u}_h, p_h)$ of the method as the element of $\mathbf{V}_h \times W_h$ such that

$$\mathcal{A}(\mathbf{u}_h, p_h; \mathbf{v}, w) = \mathcal{C}(\mathbf{v}, w) \qquad \forall\, (\mathbf{v}, w) \in \mathbf{V}_h \times W_h. \tag{20}$$

Existence and uniqueness of the approximate solution can be easily shown. Setting the data to zero, that is, $f = 0$ and $g_D = 0$, implies that

$$\mathcal{A}(\mathbf{u}_h, p_h; \mathbf{v}, w) = 0 \qquad \forall\, (\mathbf{v}, w) \in \mathbf{V}_h \times W_h.$$

Setting $\mathbf{v} = \mathbf{u}_h$ and $w = p_h$ and using the definitions of $\mathcal{A}_{LDG}$ and $\mathcal{B}_{LDG}$, we see that

$$0 = \mathcal{A}(\mathbf{u}_h, p_h; \mathbf{u}_h, p_h) = \sum_e [(K^{-1}\mathbf{u}_h, \mathbf{u}_h)_{\Omega_e} + (cp_h, p_h)_{\Omega_e}]; \tag{21}$$

thus, by (4) and (5), $\mathbf{u}_h = p_h = 0$.

3. Error estimates. In this section, we present our main results: error bounds for $||\mathbf{u} - \mathbf{u}_h||_\Omega$ and $||p - p_h||_\Omega$.

THEOREM 3.1. *Let* $(\mathbf{u}, p)$ *be the solution of problem* (1), (2), (3) *and let* $(\mathbf{u}_h, p_h)$ *be the approximate solution given by the* $\mathcal{P}^{k+1} - \mathcal{S}^k$ *LDG method* (12)–(13) *for* $k \geq 0$. *Assume the numerical flux* $\widehat{w}_h$ *satisfies* (14) *and* (16), *and the triangulation* $\mathcal{T}_h$ *satisfies* (6) *and* (8). *Assume* K *and* c *satisfy the bounds* (4) *and* (5). *Then for* $(\mathbf{u}, p)$ *sufficiently smooth,*

$$||\mathbf{u} - \mathbf{u}_h||_\Omega + ||p - p_h||_\Omega \leq C_3 h^{k+1},$$

where C_3 *is a constant independent of* h, *but depends on* C_1, C_2, c^*, c_*, K^*, K_*, σ_1, σ_2, $||p||_{H^{k+1}(\Omega)}$, *and* $||\mathbf{u}||_{H^{k+2}(\Omega)}$.

3.1. Proof of Theorem 3.1. In the arguments below, C will denote a generic, positive constant, independent of h, and ϵ will denote a generic, small positive constant. The dependence of C on quantity a will be denoted by $C(a)$.

We begin by noticing that the error $(e_\mathbf{u}, e_p) = (\mathbf{u} - \mathbf{u}_h, p - p_h)$ satisfies

$$\mathcal{A}(e_\mathbf{u}, e_p; \mathbf{v}, w) = \langle \widehat{p} - p, [\![\mathbf{v}]\!] \rangle_{\mathcal{E}_i} \qquad \forall\, (\mathbf{v}, w) \in \mathbf{V}_h \times W_h. \tag{22}$$

If we write $(e_\mathbf{u}, e_p) = (\psi_\mathbf{u} - \theta_\mathbf{u}, \psi_p - \theta_p)$, where $(\theta_\mathbf{u}, \theta_p) = (\mathbf{u}_h - \pi\mathbf{u}, p_h - \pi p)$ belongs to the finite element space $\mathbf{V}_h \times W_h$, and $(\psi_\mathbf{u}, \psi_p) = (\mathbf{u} - \pi\mathbf{u}, p - \pi p)$ is the interpolation error, we find, setting $(\mathbf{v}, w) = (\theta_\mathbf{u}, \theta_p)$,

$$\begin{aligned}\sum_e \left[||K^{-1/2}\theta_\mathbf{u}||^2_{\Omega_e} + ||c^{1/2}\theta_p||^2_{\Omega_e} \right] &= \mathcal{A}(\theta_\mathbf{u}, \theta_p; \theta_\mathbf{u}, \theta_p) \\ &= \mathcal{A}(\psi_\mathbf{u}, \psi_p; \theta_\mathbf{u}, \theta_p) - \langle \widehat{p} - p, [\![\theta_\mathbf{u}]\!] \rangle_{\mathcal{E}_i}.\end{aligned}$$

From this equality, and using the assumed lower bounds on K^{-1} and c, we obtain an estimate of $\sum_e[||\theta_{\mathbf{u}}||^2_{\Omega_e} + ||\theta_p||^2_{\Omega_e}]$. The theorem then follows from the triangle inequality and an estimate on $\sum_e[||\psi_{\mathbf{u}}||^2_{\Omega_e} + ||\psi_p||^2_{\Omega_e}]$, given below.

Thus

$$\begin{aligned}\sum_e[||K^{-1/2}\theta_{\mathbf{u}}||^2_{\Omega_e} + ||c^{1/2}\theta_p||^2_{\Omega_e}] &= \mathcal{A}(\psi_{\mathbf{u}}, \psi_p; \theta_{\mathbf{u}}, \theta_p) - \langle \widehat{p} - p, [\![\theta_{\mathbf{u}}]\!]\rangle_{\mathcal{E}_i} \\ &= \Theta_1 + \Theta_2, \end{aligned} \tag{23}$$

where

$$\Theta_1 = \sum_e \left[(K^{-1}\psi_{\mathbf{u}}, \theta_{\mathbf{u}})_{\Omega_e} - (\psi_p, \nabla\cdot\theta_{\mathbf{u}})_{\Omega_e}\right] + \langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!]\rangle_{\mathcal{E}_i}, \tag{24}$$

and

$$\Theta_2 = \sum_e \left[(c\psi_p, \theta_p)_{\Omega_e} + (\nabla\cdot\psi_{\mathbf{u}}, \theta_p)_{\Omega_e}\right] - \langle [\![\psi_{\mathbf{u}}]\!], \widehat{\theta}_p\rangle_{\mathcal{E}_i}. \tag{25}$$

We take πp and $\pi\mathbf{u}$ to be the standard L^2 projections of p and $\mathbf{u}$ into W_h and $\mathbf{V}_h$, respectively. It is well known that

$$||p - \pi p||_{\Omega_e} \le C(\sigma_1)||p||_{H^{k+1}(\Omega_e)} h_e^{k+1}, \tag{26}$$

$$||\mathbf{u} - \pi\mathbf{u}||_{\Omega_e} + h_e||\mathbf{u} - \pi\mathbf{u}||_{H^1(\Omega_e)} \le C(\sigma_1)||\mathbf{u}||_{H^{k+2}(\Omega_e)} h_e^{k+2}. \tag{27}$$

With these definitions of πp and $\pi\mathbf{u}$, we immediately see by (11) that

$$\Theta_1 = \sum_e (K^{-1}\psi_{\mathbf{u}}, \theta_{\mathbf{u}})_{\Omega_e} + \langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!]\rangle_{\mathcal{E}_i}. \tag{28}$$

To bound Θ_1 and Θ_2, we will utilize the following well-known results [4]. For a function w in any of the finite dimensional spaces described above, we have that

$$||w||_{H^1(\Omega_e)} \le C(\sigma_1) h_e^{-1} ||w||_{L^2(\Omega_e)}. \tag{29}$$

Also, for any sufficiently smooth function ϕ,

$$||\phi||^2_{L^2(\partial\Omega_e)} \le C(\sigma_1)||\phi||_{L^2(\Omega_e)}||\phi||_{H^1(\Omega_e)}. \tag{30}$$

Thus if ϕ is in one of the finite dimensional spaces above,

$$||\phi||^2_{L^2(\partial\Omega_e)} \le C(\sigma_1) h_e^{-1} ||\phi||^2_{L^2(\Omega_e)}. \tag{31}$$

We will also use the Cauchy–Schwarz inequality, and the standard inequality,

$$ab \le \frac{1}{2\delta}a^2 + \frac{\delta}{2}b^2 \tag{32}$$

for real numbers a, b, and $\delta > 0$.

It is easily seen that Θ_2 is bounded by

$$\sum_e [||c^{1/2}\psi_p||_{\Omega_e}||c^{1/2}\theta_p||_{\Omega_e} + ||c^{-1/2}\nabla\cdot\psi_{\mathbf{u}}||_{\Omega_e}||c^{1/2}\theta_p||_{\Omega_e}] - \langle [\![\psi_{\mathbf{u}}]\!], \widehat{\theta}_p\rangle_{\mathcal{E}_i}. \tag{33}$$

For the last term, let γ be an interior face, and $\{\Omega_{e_\gamma}\}$ and h_γ defined as above. Then

$$\begin{aligned}\langle [\![\psi_{\mathbf{u}}]\!], \widehat{\theta}_p\rangle_\gamma &\le ||[\![\psi_{\mathbf{u}}]\!]||_\gamma ||\widehat{\theta}_p||_\gamma \\ &\le C(\epsilon^{-1}) h_\gamma^{-1} ||[\![\psi_{\mathbf{u}}]\!]||_\gamma^2 + \epsilon h_\gamma ||\widehat{\theta}_p||_\gamma^2 \\ &\le C(\sigma_1, \sigma_2, \epsilon^{-1}) \sum_{e_\gamma} h_{e_\gamma}^{-1} ||\psi_{\mathbf{u}}||_{\Omega_{e_\gamma}} ||\psi_{\mathbf{u}}||_{H^1(\Omega_{e_\gamma})} + C_1 \epsilon \sum_{e_\gamma} ||\theta_p||_{\Omega_{e_\gamma}}^2, \end{aligned} \tag{34}$$

where we have used (30), (8), and (14). Summing over γ, choosing ϵ appropriately, and applying (27), we find

$$\langle [\![\psi_{\mathbf{u}}]\!], \widehat{\theta}_p\rangle_{\mathcal{E}_i} \le C(C_1, \sigma_1, \sigma_2, c_*^{-1}) \sum_e h_e^{2(k+1)} ||\mathbf{u}||_{H^{k+2}(\Omega_e)}^2 + \frac{c_*}{4} \sum_e ||\theta_p||_{\Omega_e}^2. \tag{35}$$

Thus, by (33), (35), (26), and (27),

$$\begin{aligned}\Theta_2 \le{}& \frac{1}{4} \sum_e ||c^{1/2}\theta_p||_{\Omega_e}^2 + \frac{c_*}{4} \sum_e ||\theta_p||_{\Omega_e}^2 \\ &+ C(c_*^{-1}, c^*, C_1, \sigma_1, \sigma_2) \sum_e \left[||\mathbf{u}||_{H^{k+2}(\Omega_e)}^2 + ||p||_{H^{k+1}(\Omega_e)}^2 \right] h_e^{2k+2}. \end{aligned} \tag{36}$$

Now we turn to Θ_1. By (27), we easily see that

$$\sum_e (K^{-1}\psi_{\mathbf{u}}, \theta_{\mathbf{u}})_{\Omega_e} \le C(\sigma_1, K^*) \sum_e ||\mathbf{u}||_{H^{k+2}(\Omega_e)}^2 h_e^{2k+4} + \frac{1}{4} \sum_e ||K^{-1/2}\theta_{\mathbf{u}}||_{\Omega_e}^2. \tag{37}$$

Next, consider $\langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!]\rangle_{\mathcal{E}_i}$. By (30), (16), and (8),

$$\begin{aligned}\langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!]\rangle_\gamma &\le C(\epsilon^{-1}) h_\gamma^{-1} ||p - \widehat{\pi p}||_\gamma^2 + \epsilon h_\gamma ||[\![\theta_{\mathbf{u}}]\!]||_\gamma^2 \\ &\le C(\epsilon^{-1}) C_{2,\gamma}^2 h_\gamma^{2k+2} + C(\sigma_1, \sigma_2) \epsilon \sum_{e_\gamma} ||\theta_{\mathbf{u}}||_{\Omega_{e_\gamma}}^2. \end{aligned} \tag{38}$$

Summing over γ and choosing ϵ appropriately, we find

$$\langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!]\rangle_{\mathcal{E}_i} \le C(K_*^{-1}, \sigma_1, \sigma_2) C_2^2 h^{2k+2} + \frac{K_*}{4} \sum_e ||\theta_{\mathbf{u}}||_{\Omega_e}^2. \tag{39}$$

Substituting the bounds (36), (37), and (39) into (23), we obtain

$$\begin{aligned}&K_* ||\theta_{\mathbf{u}}||^2 + c_* ||\theta_p||^2 \\ &\le C(C_1, c_*^{-1}, c^*, K^*, K_*^{-1}, \sigma_1, \sigma_2) \\ &\quad \times \left\{ \sum_e \left[||\mathbf{u}||_{H^{k+2}(\Omega_e)}^2 + ||p||_{H^{k+1}(\Omega_e)}^2 \right] h_e^{2k+2} + C_2^2 h^{2k+2} \right\}, \end{aligned}$$

from which the result follows.

4. Examples of $\widehat{w}$. Suppose $w \in W_h$. In this section, we discuss the construction of $\widehat{w}$ in more detail and show that, in fact, (14) and (16) can be satisfied for certain standard choices of meshes and spaces W_h. Recall the space $\mathcal{S}^k$ used to define W_h contains $\mathcal{P}^k$ and possibly some functions in $\mathcal{P}^{k+1}$.

On most standard elements Ω_e, we have or can construct an orthogonal set of polynomials $\mathcal{B} = \{l_{\mathbf{i}}(\mathbf{x})\}_{\{\mathbf{i}\in I, |\mathbf{i}|\le k+1\}}$ which serve as basis functions for W_h. In one

dimension, for example, we can use standard Legendre polynomials. In higher dimensions on rectangular or hexahedral elements, tensor products of Legendre polynomials can be used to construct an orthogonal basis. On triangles, such a basis is given in [16]. In general, Gram–Schmidt orthogonalization can be used to construct such a basis on each element.

If ϕ is a smooth function, and $\pi\phi$ its L^2 projection into W_h, then we have an error expansion

$$\phi(\mathbf{x}) - \pi\phi(\mathbf{x}) = \sum_{\{\mathbf{i}:|\mathbf{i}|=k+1,\mathbf{i}\notin I\}} \alpha_{\mathbf{i}} l_{\mathbf{i}}(\mathbf{x}) + \mathcal{O}(h^{k+2}), \qquad \mathbf{x} \in \Omega_e. \tag{40}$$

Thus, at the common roots $\bar{\mathbf{x}}$ of the functions $l_{\mathbf{i}}(\mathbf{x})$, where $|\mathbf{i}| = k+1$ and $\mathbf{i} \notin I$ (if such roots exist), we have superconvergence of $\pi\phi$:

$$\phi(\bar{\mathbf{x}}) - \pi\phi(\bar{\mathbf{x}}) = \mathcal{O}(h^{k+2}). \tag{41}$$

If we define $\pi_{k+1}\phi$ to be the L^2 projection of ϕ into $\mathcal{P}^{k+1}$, then

$$\pi_{k+1}\phi(\mathbf{x}) = \pi\phi(\mathbf{x}) + \sum_{\{\mathbf{i}:|\mathbf{i}|=k+1,\mathbf{i}\notin I\}} \alpha_{\mathbf{i}} l_{\mathbf{i}}(\mathbf{x}), \tag{42}$$

and the superconvergence at $\bar{x}$ can be easily seen since

$$\pi\phi(\bar{\mathbf{x}}) = \pi_{k+1}\phi(\bar{\mathbf{x}}). \tag{43}$$

At such superconvergence points, one can interpolate $\pi\phi$ by a higher order polynomial, say a complete polynomial of degree $k+1$, and this interpolant is a more accurate approximation to ϕ. Such an interpolation procedure will be used in defining the numerical flux $\widehat{w}$. Furthermore, by (43),

$$\widehat{\pi\phi} = \widehat{\pi_{k+1}\phi}. \tag{44}$$

We consider some examples.

***k* = 0.** Suppose W_h is the space of piecewise constants; then as we stated in the introduction, $\mathcal{S}^0 = \mathcal{P}^0$. It is well known that the L^2 projection $\pi\phi$ of ϕ into piecewise constants is superconvergent at the barycenter $\mathbf{x}_e$ of each element Ω_e. The barycentric values can be interpolated to construct a linear approximation to ϕ which is second order accurate. This linear approximation can then be used to define the numerical flux $\widehat{w}$ on interior faces.

In one space dimension, given any interior face point γ, we can linearly interpolate the piecewise constant values of w on either side of γ and evaluate this linear interpolant at γ to determine $\widehat{w}$. In particular, if $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$ are elements on either side of γ with midpoints x_1, x_2, and diameters h_1 and h_2 satisfying (8), let α_i be the linear satisfying

$$\alpha_i(x_j) = \delta_{ij}, \qquad i,j = 1,2.$$

At γ, define

$$\widehat{w}(\gamma) = \sum_{i=1}^{2} w(x_i)\alpha_i(\gamma).$$

Let $\Omega_\gamma = \Omega_{1,\gamma} \cup \Omega_{2,\gamma}$, $h_\gamma = \max(h_1, h_2)$.

By the trace inequality (30),

$$||\widehat{w}||_\gamma^2 = \left| \sum_{i=1}^{2} w(x_i)\alpha_i(\gamma) \right|^2 \leq C \sum_{i=1}^{2} |w(x_i)|^2 ||\alpha_i||_{L^2(\Omega_\gamma)} ||\alpha_i||_{H^1(\Omega_\gamma)}.$$

It is easily seen that

$$||\alpha_i||_{L^2(\Omega_\gamma)} \leq C h_\gamma^{1/2},$$
$$||\alpha_i||_{H^1(\Omega_\gamma)} \leq C(\sigma_2) h_\gamma^{-1/2}.$$

Thus,

$$||\widehat{w}||_\gamma^2 \leq C(\sigma_2) \sum_{i=1}^{2} |w(x_i)|^2 = C(\sigma_2) h_\gamma^{-1} \sum_{i=1}^{2} ||w||_{\Omega_{i,\gamma}}^2,$$

and (14) is satisfied.

If ϕ is a smooth function, then $\widehat{\phi}$ is the linear interpolant of ϕ given by

$$\widehat{\phi} = \sum_{i=1}^{2} \phi(x_i)\alpha_i. \tag{45}$$

By interpolation theory [4],

$$|(\phi - \widehat{\phi})(\gamma)|^2 \leq C h_\gamma^{2m+1} ||\phi||_{H^{m+1}(\Omega_\gamma)}^2, \tag{46}$$

where $m = 1$ in this case, thus $2m + 1 = 3$. Recalling (44), consider

$$|(\widehat{\phi} - \widehat{\pi_1 \phi})(\gamma)|.$$

The error in the L^2 projection into the space $\mathcal{P}^{k+1}$ satisfies

$$|\phi(x_i) - \pi_{k+1}\phi(x_i)| \leq C h_i^{k+2-n/2} ||\phi||_{H^{k+2}(\Omega_{i,\gamma})}, \tag{47}$$

where $n = 1$ (space dimension) and $k = 0$ in this case; hence $k + 2 - n/2 = 3/2$. Thus

$$\begin{aligned} |(\widehat{\phi} - \widehat{\pi_1 \phi})(\gamma)|^2 &= |(\phi(x_1) - \pi_1\phi(x_1))\alpha_1(\gamma) + (\phi(x_2) - \pi_1\phi(x_2))\alpha_2(\gamma)|^2 \\ &\leq C h_\gamma^3 ||\phi||_{H^2(\Omega_\gamma)}^2. \end{aligned}$$

Writing

$$\phi - \widehat{\pi\phi} = \phi - \widehat{\phi} + \widehat{\phi} - \widehat{\pi_1 \phi},$$

(16) is satisfied, with $C_{2,\gamma}^2$ dependent on $||\phi||_{H^2(\Omega_\gamma)}^2$. Summing on γ we find

$$\begin{aligned} \sum_\gamma C_{2,\gamma}^2 &= C \sum_\gamma ||\phi||_{H^2(\Omega_\gamma)}^2 \\ &\leq C ||\phi||_{H^2(\Omega)}^2 \\ &\equiv C_2. \end{aligned}$$

For a mesh consisting of rectangles and/or triangles in two space dimensions, the construction of $\widehat{w}$ is similar to the piecewise linear reconstruction procedures commonly used in finite volume schemes for conservation laws; see, for example, [3, 10, 17]. In order to simplify the discussion, suppose $\mathcal{T}_h$ is a regular, conforming triangulation. If γ is an interior face, suppose $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$ are the two elements sharing face γ; see Figure 1. Given a piecewise constant w associated with the barycenter of each triangle, to construct a linear interpolant from these values we obviously need three triangles. If we look to the neighbors of the two elements to obtain this third value, we have many possibilities from which to choose. One such procedure for constructing $\widehat{w}$ then would be the following.

First, suppose $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$ are interior elements, each with two other neighboring elements. Label these four elements $\Omega_{3,\gamma}, \ldots, \Omega_{6,\gamma}$. Consider the four triangles obtained by connecting the midpoints of $\Omega_{1,\gamma}$, $\Omega_{2,\gamma}$ and each of the four neighboring elements. Let T_γ be the triangle with smallest chunkiness parameter $h_{T_\gamma}/\rho_{T_\gamma}$; see Figure 1. Construct a linear interpolant $L_\gamma(w)$ using the values of w at the three vertices of T_γ. Take

$$\widehat{w}|_\gamma = L_\gamma(w)|_\gamma.$$

The chunkiness parameters for T_γ should satisfy

$$\frac{h_{T_\gamma}}{\rho_{T_\gamma}} \leq C(\sigma_1, \sigma_2). \tag{48}$$

If this is not the case, then one can widen the search for T_γ, for example, by examining the triangle obtained by connecting the midpoint of $\Omega_{1,\gamma}$ with those of the other two neighbors of $\Omega_{2,\gamma}$.

If $\Omega_{1,\gamma}$ (and/or $\Omega_{2,\gamma}$) is near the boundary, the number of possible triangles to construct a T_γ satisfying (48) may be limited. In the most extreme case, if $\Omega_{1,\gamma}$ is a boundary element, it may have no neighbor other than $\Omega_{2,\gamma}$, and $\Omega_{2,\gamma}$ could have only one other neighbor. In this case, it is possible (but highly unlikely) that the three midpoints of these triangles could be colinear, and it would be impossible to construct a linear interpolant using these midpoints. In any case where (48) is not satisfied to within some reasonable tolerance, we may have to widen our search to include neighbors of neighboring elements. Essentially, any three triangles near the interface γ could be used to construct T_γ.

Consider the construction of L_γ and assume (48) holds. For ease of notation, denote by Ω_1, Ω_2, and Ω_3 the triangles whose midpoints are used to construct T_γ. Let (x_i, y_i) denote the barycenter of Ω_i (which is also a vertex of T_γ), let h_i denote the element diameter, and let w_i denote the value of $\widehat{w}$ at (x_i, y_i), $i = 1, 2, 3$. Define linear basis functions α_i by

$$\alpha_i(x_j, y_j) = \delta_{ij}, \qquad i, j, = 1, 2, 3.$$

Since we are working in two space dimensions, (48) implies

$$||\alpha_i||_{\Omega_{j,\gamma}} \leq C(\sigma_1, \sigma_2) h_{j,\gamma}, \tag{49}$$

$$||\alpha_i||_{H^1(\Omega_{j,\gamma})} \leq C(\sigma_1, \sigma_2), \tag{50}$$

where $h_{j,\gamma}$ is the diameter of $\Omega_{j,\gamma}$, $j = 1, 2$. To show that (14) is satisfied, by (30) and the above bounds on α_i,

$$||\widehat{w}||_\gamma^2 = ||L_\gamma(w)||_\gamma^2 = \left\| \sum_{i=1}^{3} w_i \alpha_i \right\|_\gamma^2$$

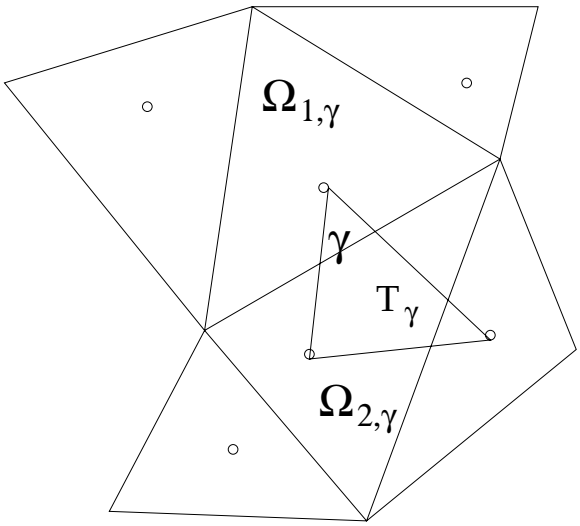


FIG. 1. *An interior face γ with neighbors $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$.*

$$\begin{aligned} &\leq C \sum_{i=1}^{3} |w_i|^2 ||\alpha_i||_{\Omega_{j,\gamma}} ||\alpha_i||_{H^1(\Omega_{j,\gamma})} \\ &\leq C h_{j,\gamma} \sum_{i=1}^{3} |w_i|^2 \\ &\leq C h_\gamma^{-1} \sum_{i=1}^{3} ||w||_{\Omega_i}^2, \end{aligned} \tag{51}$$

where C depends on σ_1 and σ_2.

To derive (16), let Ω_γ denote $\cup_{i=1}^3 \Omega_i \cup \Omega_{1,\gamma} \cup \Omega_{2,\gamma}$. Applying (30) and error estimates similar to (46) and (47) with $m = 1$, $k = 0$, and $n = 2$, and using (49) and (50), we find for smooth ϕ,

$$\begin{aligned} ||\phi - L_\gamma(\pi_1\phi)||_\gamma^2 &\leq C||\phi - L_\gamma(\pi_1\phi)||_{\Omega_{j,\gamma}} ||\phi - L_\gamma(\pi_1\phi)||_{H^1(\Omega_{j,\gamma})} \\ &\leq C h_\gamma^3 ||\phi||_{H^2(\Omega_\gamma)}^2, \end{aligned}$$

where C depends on σ_1 and σ_2. Thus (16) holds, with

$$C_{2,\gamma} = C(\sigma_1, \sigma_2)||\phi||_{H^2(\Omega_\gamma)}.$$

Summing on all faces γ we find

$$\sum_\gamma C_{2,\gamma}^2 \leq C(\sigma_1, \sigma_2)||\phi||_{H^2(\Omega)}^2 \equiv C_2,$$

and C_2 is independent of h.

We remark that the above procedure can be applied to nonconforming triangulations as long as a triangle T_γ in a neighborhood of γ satisfying (48) can be found.

***k* = 1.** Next, suppose W_h contains the space of piecewise linears. In this case, we may need to add second degree terms into the definition of $\mathcal{S}^1$. The procedure is

then to find the common roots of the quadratic orthogonal polynomials omitted from the definition of $\mathcal{S}^1$, use these roots in interpolating w by a complete quadratic, and evaluate this interpolant on interior faces to determine $\widehat{w}$.

In one dimension, we can take $\mathcal{S}^1 = \mathcal{P}^1$. The L^2 projection $\pi\phi$ of ϕ into the space of piecewise linears is superconvergent at the roots of the quadratic Legendre polynomial. On any interval $[a, b]$, these roots are located at

$$\frac{a+b}{2} - \frac{b-a}{2\sqrt{3}}, \quad \frac{a+b}{2} + \frac{b-a}{2\sqrt{3}}. \tag{52}$$

At an interface γ between two elements $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$, one can interpolate w at any collection of three such roots within these two elements by a quadratic (or if desired, at all four roots by a cubic) and evaluate this interpolant at γ to determine $\widehat{w}$. For example, labeling the four roots x_j, $j = 1, \ldots, 4$, we construct four cubic basis functions $\alpha_i(x)$, satisfying $\alpha_i(x_j) = \delta_{ij}$. Then

$$\widehat{w}(\gamma) = \sum_{i=1}^{4} w(x_i)\alpha_i(\gamma).$$

Following the above arguments with $k = 1$, $m = 2$, and $n = 1$, one can show this interpolant satisfies (14) and (16), with C_2 dependent on $||\phi||_{H^3(\Omega)}$.

On rectangles, using tensor product linears to define $\mathcal{S}^1$ and hence W_h, the L^2 projection is superconvergent at the tensor products of the roots in (52). Thus there are four roots per element. Given rectangles $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$ sharing a face γ, we can interpolate w at any six of these roots in the two elements by a quadratic in x and y (or at all eight roots by a tensor product quadratic without the x^2y^2 term). By arguments similar to those above with $k = 1$, $m = 2$, and $n = 2$, this quadratic satisfies (14) and (16).

On triangles, a set of orthogonal polynomials through second degree on the reference element $\widehat{\Omega}$ bounded by $\xi = 0$, $\eta = 0$, $\xi + \eta = 1$ is given by

$$\begin{aligned}
l_{0,0}(\xi, \eta) &= 1, \\
l_{1,0}(\xi, \eta) &= 1 - 3\xi, \\
l_{0,1}(\xi, \eta) &= 1 - \xi - 2\eta, \\
l_{0,2}(\xi, \eta) &= (5\xi - 4)\xi + (-15\eta + 12)\eta - 1, \\
l_{1,1}(\xi, \eta) &= (3\xi + 8\eta - 4)\xi + (3\eta - 4)\eta + 1, \\
l_{2,0}(\xi, \eta) &= (10\xi - 8)\xi + 1.
\end{aligned}$$

We choose $\mathcal{S}^1(\widehat{\Omega})$ to be the span of the linear terms plus one of the quadratic terms, in this case $l_{0,2}$. A basis for $\mathcal{S}^1(\Omega_e)$ and hence W_h is obtained by mapping these functions through an affine map from $\widehat{\Omega}$ to Ω_e. Thus the L^2 projection $\pi\phi$ into W_h is superconvergent to ϕ at the common roots of $l_{2,0}$ and $l_{1,1}$. There are three such roots in the master element, located at

$$(.155051, .706707), \quad (.155051, .213157), \quad (.644949, .191431).$$

For neighboring triangles $\Omega_{1,\gamma}$ and $\Omega_{2,\gamma}$ sharing face γ, we interpolate w by a complete quadratic at the six roots contained in these two elements and evaluate the quadratic

TABLE 1
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{S}^0$ *LDG method, uniform mesh.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0916	–	.3701	–
32	.0459	1.0	.1879	1.0
64	.0230	1.0	.0943	1.0
128	.0115	1.0	.0472	1.0

TABLE 2
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{S}^0$ *LDG method, nonuniform mesh.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0696	–	.4266	–
32	.0347	1.0	.2169	1.0
64	.0174	1.0	.1088	1.0
128	.0087	1.0	.0545	1.0

at γ to determine $\widehat{w}$. That is, labeling the six roots (x_j, y_j), $j = 1, \ldots, 6$, we construct six quadratic basis functions α_i satisfying $\alpha_i(x_j, y_j) = \delta_{ij}$; then

$$\widehat{w}|_\gamma = \sum_{i=1}^{6} w(x_i, y_i)\alpha_i|_\gamma.$$

5. Numerical results. In this section, we give one and two dimensional examples which confirm our theoretical results, and in one dimension, we compare the results of our method to a more standard implementation of the LDG method with equal order approximating spaces for p and $\mathbf{u}$ (the $\mathcal{P}^k - \mathcal{P}^k$ method).

5.1. One dimensional results. First, we consider the method (12)–(13) in one space dimension with $k = 0$. Recall that in this case, $\mathcal{S}^0 = \mathcal{P}^0$. The numerical flux $\widehat{w}$ is constructed using the linear interpolation procedure outlined in section 4. We consider the problem

$$cp - p_{xx} = (c + \pi^2)\sin(\pi x), \qquad 0 < x < 1, \tag{53}$$

$$p(0) = p(1) = 0, \tag{54}$$

which has the solution $p(x) = \sin(\pi x)$. We will take $c = 1$. Suppose the domain $\Omega = [0, 1]$ is partitioned into subintervals $\Omega_i = [x_{i-1/2}, x_{i+1/2}]$ of length h_i, $i = 1, \ldots, N$, with midpoint x_i.

First, consider a sequence of uniform meshes $h_i = 1/N$. The L^2 errors for p and $\mathbf{u}$ for the $\mathcal{P}^1 - \mathcal{S}^0$ method are given in Table 1. First order convergence for both variables is seen, as expected from the theory. Next, we consider a sequence of nonuniform meshes, given by

$$h_i = \begin{cases} \frac{1}{2N}, & i \text{ odd}, \\ \frac{3}{2N}, & i \text{ even}. \end{cases} \tag{55}$$

The L^2 errors for p and $\mathbf{u}$ are given in Table 2. Again, first order convergence for both variables is observed, with errors comparable to the uniform mesh case, though somewhat smaller for p and somewhat larger for $\mathbf{u}$.

Next, we consider (12)–(13) with $k = 1$. For comparison purposes, we also consider the standard $\mathcal{P}^1 - \mathcal{P}^1$ LDG method. To avoid confusion, we denote the numerical

TABLE 3
Case 1-1D*:* $\mathcal{P}^2 - \mathcal{S}^1$ *LDG method, uniform mesh.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0080	–	.0117	–
32	.0020	2.0	.0029	2.0
64	.0005	2.0	.0007	2.0
128	.000124	2.0	.00019	2.0

TABLE 4
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{P}^1$ *LDG method,* $C_{11} = 0$*, uniform mesh,* $\tilde{w} = \{w\}$*.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0061	–	.1349	–
32	.0017	1.84	.0627	1.1
64	.00053	1.64	.0335	.90
128	.00021	1.33	.0167	1.0

TABLE 5
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{P}^1$ *LDG method,* $C_{11} = 1$*, uniform mesh,* $\tilde{w} = \{w\}$*.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0059	–	.1297	–
32	.00147	2.0	.063	1.0
64	.00037	2.0	.032	1.0
128	.00009	2.0	.016	1.0

flux for this method by $\tilde{w}$ instead of $\widehat{w}$. We will consider two choices for the numerical flux: $\tilde{w} = \{w\} = (w^- + w^+)/2$ and $\tilde{w} = w^-$. In [8], it was shown that in one space dimension, with the latter choice of $\tilde{w}$, the $\mathcal{P}^k - \mathcal{P}^k$ method gives optimal convergence for both p and $\mathbf{u}$. We also add a stabilization term involving $[\![p_h]\!]$ to (12). Thus, we define the standard LDG method by

$$\begin{aligned}\mathcal{A}^s_{LDG}(\mathbf{u}_h, p_h; w_h) &\equiv \sum_e [(cp_h, w_h)_{\Omega_e} + (\nabla \cdot \mathbf{u}_h, w_h)_{\Omega_e}] - \langle [\![\mathbf{u}_h]\!], \tilde{w}_h \rangle_{\mathcal{E}_i} \\ &\quad + \langle C_{11} [\![p_h]\!], [\![w_h]\!] \rangle_{\mathcal{E}_i} \\ &= \sum_e (f, w_h)_{\Omega_e}, \end{aligned} \tag{56}$$

$$\begin{aligned}\mathcal{B}^s_{LDG}(\mathbf{u}_h, p_h; \mathbf{v}_h) &\equiv \sum_e [(K^{-1}\mathbf{u}_h, \mathbf{v}_h)_{\Omega_e} - (p_h, \nabla \cdot \mathbf{v}_h)_{\Omega_e}] + \langle \tilde{p}_h, [\![\mathbf{v}_h]\!] \rangle_{\mathcal{E}_i} \\ &= -\langle g_D, \mathbf{v}_h \cdot \mathbf{n} \rangle_{\partial\Omega}, \end{aligned} \tag{57}$$

where $C_{11} \geq 0$.

For the $\mathcal{P}^2 - \mathcal{S}^1$ method, following section 4, the numerical flux $\widehat{w}$ is constructed by interpolating w at the roots of the second degree Legendre polynomial on each interval Ω_i, given by (52). Specifically, in neighboring intervals $\Omega_{1,\gamma}$, $\Omega_{2,\gamma}$ between interface point γ, we compute the cubic Lagrange polynomial which interpolates w at each of the two roots in each element and evaluate this interpolant at the interface point γ to obtain $\widehat{w}$.

First, consider (53)–(54) with the domain discretized with a sequence of uniform meshes. The L^2 errors for p and $\mathbf{u}$ for the $\mathcal{P}^2 - \mathcal{S}^1$ method are given in Table 3. Second order accuracy is observed for both p_h and $\mathbf{u}_h$, as expected from the theory.

We also computed the $\mathcal{P}^1 - \mathcal{P}^1$ LDG solution given by (56)–(57), with $C_{11} = 0$ (Table 4) and $C_{11} = 1$ (Table 5), and $\tilde{w} = \{w\}$. These tables indicate that to obtain

TABLE 6
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{P}^1$ *LDG method,* $C_{11} = 0$*, uniform mesh,* $\tilde{w} = w^-$.

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.246	–	.0074	–
32	.087	1.5	.0019	1.96
64	.030	1.54	.00048	1.98
128	.010	1.58	.00011	2.12

TABLE 7
Case 1-1D*:* $\mathcal{P}^1 - \mathcal{P}^1$ *LDG method,* $C_{11} = 1$*, uniform mesh,* $\tilde{w} = w^-$.

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0048	–	.0074	–
32	.0011	2.12	.0018	2.0
64	.00026	2.08	.00047	1.98
128	.00007	1.90	.00011	2.12

TABLE 8
Case 2-1D*:* $\mathcal{P}^2 - \mathcal{S}^1$ *LDG method, nonuniform mesh.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.014	–	.0228	–
32	.0035	2.0	.0057	2.0
64	.0009	2.0	.0014	2.0
128	.00022	2.0	.00036	2.0

TABLE 9
Case 2-1D*:* $\mathcal{P}^1 - \mathcal{P}^1$ *LDG method,* $C_{11} = 1$*, nonuniform mesh,* $\tilde{w} = w^-$.

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0105	–	.0140	–
32	.0023	2.19	.0036	1.96
64	.00053	2.11	.0009	2.0
128	.00013	2.03	.00023	1.97

full second order accuracy for p in this version of the LDG method, $C_{11} > 0$ is necessary. First order accuracy in $\mathbf{u}_h$ was observed; furthermore, note that the errors in $\mathbf{u}_h$ are an order of magnitude larger when compared to the $\mathcal{P}^2 - \mathcal{S}^1$ solution. We repeated this experiment with $\tilde{w} = w^-$. Again, the accuracy of the pressure solution depends on having a positive penalty parameter C_{11}. In this case, as seen in Tables 6 and 7, full second order accuracy for $\mathbf{u}$ is observed for both choices of C_{11}, with errors actually smaller than the $\mathcal{P}^2 - \mathcal{S}^1$ method. This optimal convergence for $\mathbf{u}$ agrees with the one dimensional theory in [8].

Next, we consider the nonuniform mesh given by (55). The errors for p_h and $\mathbf{u}_h$ for the $\mathcal{P}^2 - \mathcal{S}^1$ LDG method are given in Table 8. We compare these results to the $\mathcal{P}^1 - \mathcal{P}^1$ LDG method results with $\tilde{w} = w^-$ and $C_{11} = 1$ given in Table 9.

5.2. Two dimensional results. In this section, we present two dimensional results for the $\mathcal{P}^1 - \mathcal{S}^0$ and $\mathcal{P}^2 - \mathcal{S}^1$ methods on triangular grids. Convergence tests for the $\mathcal{P}^k - \mathcal{P}^k$ LDG method ($k \geq 1$) in two dimensions can be found in [7], which demonstrate order $k+1$ convergence for p and k for $\mathbf{u}$ in L^2. We consider

$$cp - \Delta p = f, \quad \Omega, \tag{58}$$

$$p = g, \quad \partial\Omega. \tag{59}$$

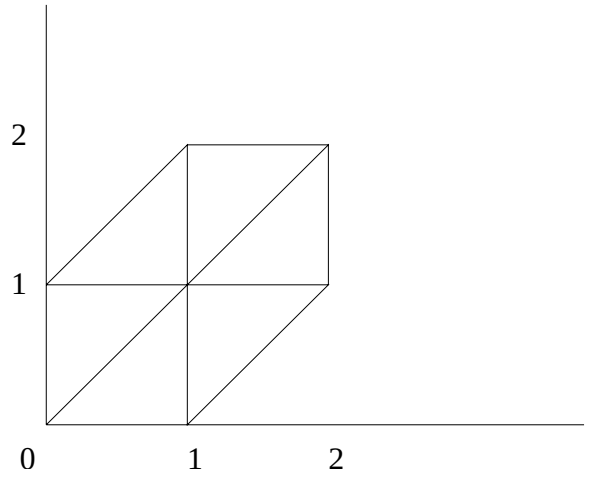


FIG. 2. *Domain Ω for two dimensional test problem* (58)*, discretized by six triangles.*

TABLE 10
Case 3-2D*: $\mathcal{P}^1 - \mathcal{S}^0$ LDG method for two dimensional case with $c = 1$, $p = x^3 + y^3$.*

mesh	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
1	3.3309	–	4.6800	–
2	1.7554	.92	3.5178	.41
3	.8691	1.0	2.1460	.71
4	.4299	1.0	1.0352	1.05
5	.2146	1.0	.5863	.82
6	.1071	1.0	.2674	1.13

TABLE 11
Case 3-2D*: $\mathcal{P}^2 - \mathcal{S}^1$ LDG method for two dimensional case with $c = 1$, $p = x^3 + y^3$.*

mesh	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
1	.5714	–	.5006	–
2	.1343	2.1	.1691	1.56
3	.0398	1.75	.0463	1.86
4	.0081	2.3	.0121	1.94

The domain Ω is given in Figure 2, with a coarse mesh of six elements discretizing Ω. All subsequent meshes are obtained from uniform refinements of this initial mesh.

For this test case, we have chosen $c = 1$ and f and g so that $p = x^3 + y^3$. The construction of the numerical flux $\widehat{w}$ follows the procedures described in section 4. The errors in L^2 for p and $\mathbf{u}$ for the two methods are given in Tables 10 ($k = 0$) and 11 ($k = 1$). Here mesh 1 refers to the initial coarse mesh, meshes 2–6 are uniform refinements of this mesh. As seen in the tables, first order convergence is observed for p when $k = 0$ and second order convergence when $k = 1$. The errors in $\mathbf{u}$ are also approaching first ($k = 0$) and second order ($k = 1$) as the mesh is refined.

6. The case $c = 0$ and $k = 0$. We again consider the important case where W_h is the space of piecewise constant functions and $\mathbf{V}_h$ is in the space of piecewise linears. We will consider the purely elliptic case with $c = 0$ and assume that Ω is discretized by a triangulation $\mathcal{T}_h$ as defined above, with the exception that $\mathcal{T}_h$ is conforming. Let $(W_{h,RT}, \mathbf{V}_{h,RT})$ denote the lowest order Raviart–Thomas mixed finite element

space on $\mathcal{T}_h$. The lowest order space consists of piecewise constant approximations for p ($W_{h,RT} = W_h$) and piecewise linear approximations for $\mathbf{u}$. The properties of $(W_h, \mathbf{V}_{h,RT})$ we require are

$$\mathbf{V}_{h,RT} \subset H(\mathrm{div}; \Omega), \tag{60}$$

$$\mathbf{V}_{h,RT} \subset \mathbf{V}_h, \tag{61}$$

and

$$\nabla \cdot \mathbf{V}_{h,RT} = W_h. \tag{62}$$

(60) implies that if $\mathbf{v} \in \mathbf{V}_{h,RT}$, then the normal component of $\mathbf{v}$ is continuous across interior faces. See [20, 19] for descriptions of such spaces on standard types of elements (rectangular parallelepipeds, triangles, tetrahedra, and prisms).

The method is given by (12)–(13) with $c = 0$. The numerical flux $\widehat{p}_h$ is assumed to still satisfy (16). Assuming zero data, existence and uniqueness of $\mathbf{u}_h$ is proved by the same arguments which lead to (21). Uniqueness of p_h is proved by a duality argument. Let $\sigma_h \in \mathbf{V}_{h,RT}$ be such that $\nabla \cdot \sigma_h = p_h$. This is possible by (62). Then by (13),

$$\begin{aligned} ||p_h||^2 &= \sum_e (p_h, \nabla \cdot \sigma_h)_{\Omega_e} \\ &= \sum_e (K^{-1}\mathbf{u}_h, \sigma_h)_{\Omega_e} + \langle \widehat{p}_h, [\![\sigma_h]\!] \rangle_{\mathcal{E}_i} \\ &= 0, \end{aligned}$$

since $\mathbf{u}_h = 0$ and $[\![\sigma_h]\!] = 0$ (by (60)) on interior faces. Thus existence and uniqueness of p_h are established.

The error estimates follow slightly different arguments. In particular, instead of comparing $\mathbf{u}_h$ to the L^2 projection of $\mathbf{u}$, we compare it to the "divergence projection" $\pi\mathbf{u} \in \mathbf{V}_{h,RT}$ satisfying

$$\sum_e (\nabla \cdot (\mathbf{u} - \pi\mathbf{u}), w)_{\Omega_e} = 0, \qquad w \in W_h. \tag{63}$$

This projection is a standard projection used in mixed finite element analysis. In [20], it is shown that

$$||\mathbf{u} - \pi\mathbf{u}|| \leq C(\sigma_1, \sigma_2)||\mathbf{u}||_{H^1(\Omega)} h. \tag{64}$$

Furthermore $[\![\pi\mathbf{u}]\!] = 0$ across interior faces. Again using the L^2 projection πp of p, and defining $\theta_{\mathbf{u}}$, θ_p, $\psi_{\mathbf{u}}$, and ψ_p as in section 3, we find that

$$\begin{aligned} ||K^{-1/2}\theta_{\mathbf{u}}||^2 &= \mathcal{A}(\psi_{\mathbf{u}}, \psi_p; \theta_{\mathbf{u}}, \theta_p) - \langle \widehat{p} - p, [\![\theta_{\mathbf{u}}]\!] \rangle_{\mathcal{E}_i} \\ &= \Theta_1 + \Theta_2, \end{aligned} \tag{65}$$

where

$$\Theta_1 = \sum_e (K^{-1}\psi_{\mathbf{u}}, \theta_{\mathbf{u}})_{\Omega_e} + \langle p - \widehat{\pi p}, [\![\theta_{\mathbf{u}}]\!] \rangle_{\mathcal{E}_i}, \tag{66}$$

and

$$\Theta_2 = 0. \tag{67}$$

Following the same arguments given above to bound Θ_1, we find that

$$||K^{-1/2}(\mathbf{u}-\mathbf{u}_h)||_\Omega \le C(\sigma_1,\sigma_2,K_*^{-1},K^*)\left[||\mathbf{u}||_{H^1(\Omega)}+C_2\right]h. \tag{68}$$

To estimate the error in p_h, we again use a duality argument and define ϕ and σ to satisfy

$$\nabla\cdot\sigma \equiv -\nabla\cdot(\nabla\phi) = e_p, \quad \Omega, \tag{69}$$

$$\phi = 0, \quad \partial\Omega. \tag{70}$$

By elliptic regularity [18], we have that

$$||\sigma||_{H^1(\Omega)} \le C||e_p||_\Omega. \tag{71}$$

Therefore, letting σ_h be the divergence projection of σ,

$$\begin{aligned} ||e_p||_\Omega^2 &= \sum_e (e_p, \nabla\cdot\sigma)_{\Omega_e} \\ &= \sum_e [(e_p, \nabla\cdot(\sigma-\sigma_h))_{\Omega_e} + (e_p, \nabla\cdot\sigma_h)_{\Omega_e}] \\ &= \sum_e [(e_p, \nabla\cdot(\sigma-\sigma_h))_{\Omega_e} + (K^{-1}e_{\mathbf{u}}, \sigma_h)_{\Omega_e}], \end{aligned} \tag{72}$$

where in the last step we have used (13) and the fact that $[\![\sigma_h]\!] = 0$. By the definition of σ_h,

$$\begin{aligned} \sum_e (e_p, \nabla\cdot(\sigma-\sigma_h))_{\Omega_e} &= \sum_e (p, \nabla\cdot(\sigma-\sigma_h))_{\Omega_e} \\ &= \sum_e (p-\pi p, \nabla\cdot(\sigma-\sigma_h))_{\Omega_e} \\ &\le \sum_e ||p-\pi p||_{\Omega_e}\ ||\nabla\cdot(\sigma-\sigma_h)||_{\Omega_e} \\ &\le C(\sigma_1)h||p||_{H^1(\Omega)}||\sigma||_{H^1(\Omega)} \\ &\le C(\sigma_1)h||p||_{H^1(\Omega)}||e_p||_\Omega. \end{aligned}$$

Furthermore, by (68) and (71),

$$\begin{aligned} \sum_e (K^{-1}e_{\mathbf{u}}, \sigma_h)_{\Omega_e} &\le C(K^*)||e_{\mathbf{u}}||_\Omega\ ||\sigma_h||_\Omega \\ &\le Ch||e_p||_\Omega. \end{aligned}$$

Substituting these bounds into (72), we obtain the following result.

THEOREM 6.1. *Let $(\mathbf{u},p)$ be the solution of problem* (1), (2), (3) *with $c=0$. Assume K satisfies* (4). *Let $(\mathbf{u}_h,p_h)$ be the approximate solution given by the $\mathcal{P}^1-\mathcal{S}^0$ LDG method* (12)–(13). *Assume a conforming triangulation $\mathcal{T}_h$, with the lowest order Raviart–Thomas mixed finite element space $W_h\times\mathbf{V}_{h,RT}$ satisfying* (60)–(62) *defined on $\mathcal{T}_h$. Assume the numerical flux $\widehat{p}$ satisfies* (16) *with $k=0$. Then for $(\mathbf{u},p)$ sufficiently smooth,*

$$||\mathbf{u}-\mathbf{u}_h||_\Omega + ||p-p_h||_\Omega \le C_4 h,$$

TABLE 12
Case 2-1D*: $\mathcal{P}^1 - \mathcal{S}^0$ LDG method, nonuniform mesh, $c = 0$.*

N	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
16	.0917	–	.3695	–
32	.0459	1.0	.1878	1.0
64	.0230	1.0	.0943	1.0
128	.0115	1.0	.0472	1.0

TABLE 13
Case 3-2D*: $\mathcal{P}^1 - \mathcal{S}^0$ LDG method for two dimensional case with $c = 0$, $p = x^3 + y^3$.*

mesh	$\|\|p - p_h\|\|$	rate	$\|\|\mathbf{u} - \mathbf{u}_h\|\|$	rate
1	3.3403	–	4.6781	–
2	1.7665	1.0	3.5167	.41
3	.8717	1.0	2.1455	.71
4	.4302	1.0	1.0351	1.05
5	.2146	1.0	.5863	.82
6	.1072	1.0	.2674	1.13

where C_4 is a constant independent of h but depends on C_2, K_*^{-1}, K^*, σ_1, σ_2, $||\mathbf{u}||_{H^1(\Omega)}$, and $||p||_{H^1(\Omega)}$.

Numerical validation of this result in one and two dimensions can be seen as follows. First, consider (53)–(54), now with $c = 0$. We consider the nonuniform mesh given by (55). In Table 12, we measure the errors $||p - p_h||_\Omega$ and $||\mathbf{u} - \mathbf{u}_h||_\Omega$ for the $\mathcal{P}^1 - \mathcal{S}^0$ approximation. In this case, we obtain the expected first order convergence rates for p and $\mathbf{u}_h$. Second, we consider (58) with $c = 0$ and $p = x^3 + y^3$. The L^2 errors for p and $\mathbf{u}$ are given in Table 13; again, first order convergence is observed.

Theorem 6.1 can be extended to the case $k > 0$ if spaces W_h and $\mathbf{V}_{h,RT}$ satisfying (16) and (60)–(62) can be found. Since W_h must be a rich enough space so that (16) is satisfied, it may not in general be possible to satisfy both this property and (62).

7. Concluding remarks. In this paper, we have formulated and analyzed a variant of the LDG method which uses higher degree polynomials for approximating the flux $\mathbf{u} = -K\nabla p$ over those used for approximating p. The advantage of this approach over the standard LDG method, which uses equal order spaces for $\mathbf{u}$ and p is that it gives a more accurate approximation of the flux. Furthermore, this approach allows for the possibility of piecewise constant approximations to p, which is useful for a number of fluid applications.

REFERENCES

[1] V. AIZINGER, C. DAWSON, B. COCKBURN, AND P. CASTILLO, *Local discontinuous Galerkin method for contaminant transport*, Adv. in Water Res., 24 (2000), pp. 73–87.

[2] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] J. B. BELL, C. N. DAWSON, AND G. R. SHUBIN, *An unsplit, higher order Godunov method for scalar conservation laws in two space dimensions*, J. Comput. Phys., 74 (1988), pp. 1–24.

[4] S. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer–Verlag, New York, 1994.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer–Verlag, New York, 1991.

[6] P. CASTILLO, *Performance of discontinuous Galerkin methods for elliptic PDEs*, SIAM J. Sci. Comput., 24 (2002), pp. 524–547.

[7] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[8] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab, *An optimal a priori error estimate for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems*, Math. Comp., 71 (2001), pp. 455–478.

[9] B. Cockburn and C. Dawson, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions*, in The Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X, J. Whiteman, ed., Elsevier, Oxford, UK, 2000, pp. 225–238.

[10] B. Cockburn, S. Hou, and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[11] B. Cockburn, G. Kanschat, I. Perugia, and D. Schötzau, *Superconvergence of the local discontinuous Galerkin method for elliptic problems on Cartesian grids*, SIAM J. Numer. Anal., 39 (2001), pp. 264–285.

[12] B. Cockburn, G. Kanschat, D. Schötzau, and C. Schwab, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.

[13] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[14] C. Dawson, *A Local Discontinuous Galerkin Method with Lowest Order Approximating Spaces for Flow Problems*, TICAM technical report 01-26, The University of Texas at Austin, Austin, TX, 2001.

[15] C. Dawson and J. Proft, *A priori error estimates for interior penalty version of the local discontinuous Galerkin method applied to transport equations*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 545–564.

[16] M. Dubiner, *Spectral methods on triangles and other domains*, J. Sci. Comput., 6 (1991), pp. 345–390.

[17] L. J. Durlofsky, B. Engquist, and S. Osher, *Triangle based adaptive stencils for the solution of hyperbolic conservation laws*, J. Comput. Phys., 98 (1992), pp. 64–73.

[18] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[19] J. C. Nedelec, *Mixed finite element methods in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[20] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer–Verlag, Berlin, 1977, pp. 292–315.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2171–2194

MIXED *hp*-DGFEM FOR INCOMPRESSIBLE FLOWS*

DOMINIK SCHÖTZAU†, CHRISTOPH SCHWAB‡, AND ANDREA TOSELLI‡

Abstract. We consider several mixed discontinuous Galerkin approximations of the Stokes problem and propose an abstract framework for their analysis. Using this framework, we derive a priori error estimates for *hp*-approximations on tensor product meshes. We also prove a new stability estimate for the discrete divergence bilinear form.

Key words. *hp*-FEM, discontinuous Galerkin methods, Stokes problem

AMS subject classifications. 65N30, 65N35, 65N12, 65N15

PII. S0036142901399124

1. Introduction.

Discontinuous Galerkin (DG) methods for incompressible flow problems allow one to use discrete velocity spaces consisting of piecewise polynomial functions with no interelement continuity. Well-posedness of the discrete formulations is then achieved by numerical fluxes, i.e., by introducing suitable bilinear forms defined on the interfaces between the elements of the mesh. This choice presents considerable advantages for certain types of problems, especially those modeling phenomena where transport is dominant; see the state-of-the-art surveys in [18], the monograph [15], the recent review [20], and the references therein. In addition, DG approximations allow for nonconforming meshes.

Even if transport may be the dominant effect of a problem, diffusive terms still need to be accounted for and correctly discretized in a DG framework. For the Oseen or the incompressible Navier–Stokes equations, for instance, if advective terms are properly treated by, e.g., suitable upwinding techniques, stability and convergence only depend on the diffusive part of the operator and can then be studied for the simpler Stokes problem; see, e.g., [39, 25, 10, 33, 9]. In particular, suitable velocity-pressure space pairs are required to ensure stability and convergence. This separation of advective and diffusive effects was employed in [5] for the first definition of DG methods for convection-diffusion problems, in [19, 16, 7] for the so-called local discontinuous Galerkin and Baumann–Oden methods, respectively, and also in [27] for the *hp*-DG approximation of scalar advection-diffusion problems.

The recent works in [32] and [2] have unified the formulation and analysis of DG approximations for purely diffusive problems, where virtually all the available DG methods can be analyzed in a unified framework. In particular, several assumptions on the discrete spaces and bilinear forms have been given and analyzed that can be used to ensure a priori error estimates for the methods.

While extensive work has been done for diffusion or advection-diffusion problems, there are considerably fewer works for DG discretizations of saddle-point problems describing, e.g., nearly incompressible solids or incompressible fluid flows. We mention [4, 28], where an interior penalty approximation with discontinuous, piecewise

*Received by the editors December 4, 2001; accepted for publication (in revised form) July 19, 2002; published electronically January 7, 2003.
http://www.siam.org/journals/sinum/40-6/39912.html
†Department of Mathematics, University of Basel, Rheinsprung 21, CH-4051 Basel, Switzerland (schotzau@math.unibas.ch).
‡Seminar für Angewandte Mathematik (SAM), ETH Zürich, CH-8092 Zürich, Switzerland (schwab@sam.math.ethz.ch, toselli@sam.math.ethz.ch). The last two authors were partially supported by the Swiss National Science Foundation under project 20-63397.00.

divergence-free velocities and continuous pressures is employed for the Stokes and incompressible Navier–Stokes equations, respectively. In [17], a local discontinuous Galerkin approximation for the Stokes problem is proposed. There, the introduction of certain pressure stabilization terms allows one to choose velocity and pressure spaces of the same polynomial order k. Optimal error estimates for h-approximations are proved. In [26], an h-approximation for incompressible and nearly incompressible elasticity based on an interior penalty DG method is introduced and studied. Triangular and tetrahedral meshes are employed, together with polynomial spaces of total degree k and $k-1$ for the velocity and pressure, respectively. Optimal error estimates in h are derived, which remain valid in the incompressible limit. A similar approach was considered in [40] for hp-approximations of the Stokes problem on tensor product meshes in two and three dimensions. Stability estimates for the discrete divergence bilinear form that are explicit in h and k are obtained. Numerical results point out that these estimates are not sharp in the order k, at least for conforming two-dimensional meshes. In the present work we indeed prove sharper estimates for the same DG approximation.

The present work has two purposes. In the first part, we develop an abstract framework for mixed DG approximations of the Stokes problem. In particular, we give a set of assumptions on the approximation spaces and on the velocity and divergence bilinear forms which allows us to obtain a priori error estimates. All available mixed DG methods for the Stokes problem can be analyzed in the presented framework by introducing lifting operators similar to the ones used in [2] for the Laplace equation. However, unlike in the analysis of [2], our error estimates are derived by using a variant of Strang's lemma, combined with the techniques developed in [40] that give abstract estimates for the errors in the velocity and the pressure. With respect to the use of Strang's lemma, our approach is closely related to the setting proposed in [30, 31] for the analysis of local discontinuous Galerkin methods for purely elliptic problems.

Our second result is a new proof of the inf-sup condition of the discrete DG divergence bilinear form for tensor product meshes and $\mathbb{Q}_k - \mathbb{Q}_{k-1}$ elements. In particular, we prove a bound sharper than that given in [40]. Our analysis is valid for shape-regular two- and three-dimensional tensor product meshes, possibly with hanging nodes. Even though our estimate does not appear to be sharp, at least in two dimensions (see the numerical results in [40]), we are able to ensure the same convergence rate for the velocity and the pressure as that of conforming $\mathbb{Q}_k - \mathbb{Q}_{k-2}$ elements in three dimensions, but with a gap in the polynomial degree of the velocity-pressure pair of just one.

Our framework and analysis can be adapted to the case of nearly incompressible elasticity in a straightforward way. We note that equal-order conforming discretizations are possible both in nearly incompressible elasticity and incompressible flows, but that the bilinear forms need to be suitably modified. These stabilization techniques typically rely on local terms that are added to the bilinear forms and are constructed with the residual of the differential equations on each element; see [22, 21, 24]. The calculation of these terms is not often a simple matter for higher-order hp-approximations. On the other hand, DG approximations allow us to narrow or eliminate the polynomial degree gap between the velocity and pressure spaces by employing a discontinuous velocity space and suitable bilinear forms on the interfaces. This results in an increase of the velocity degrees of freedom that, in the case of p- and hp-approximations, is not, however, of the same order of magnitude as the number of degrees of freedom of the corresponding conforming discretization, as is the case for lower-order approximations.

The rest of this paper is organized as follows: We start by reviewing the Stokes problem in section 2 and then present our abstract framework in section 3. Suitable assumptions on the bilinear forms allow us to derive a priori error estimates in section 4. In section 5 we discuss some particular choices for the bilinear forms and the approximation spaces. Section 6 contains the proofs of the inf-sup condition of the discrete divergence bilinear form. In sections 7 and 8 we establish the remaining assumptions for our DG approximations. Finally, we derive hp-error estimates in section 9.

2. The Stokes problem. Let Ω be a bounded polygonal or polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$, respectively, with $\mathbf{n}$ denoting the outward normal unit vector to its boundary $\partial\Omega$. Given a source term $\mathbf{f} \in L^2(\Omega)^d$ and a Dirichlet datum $\mathbf{g} \in H^{1/2}(\partial\Omega)^d$ satisfying the usual compatibility condition $\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n}\, ds = 0$, the Stokes problem in incompressible fluid flow is to find a velocity field $\mathbf{u}$ and a pressure p such that

$$\begin{aligned} -\nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= \mathbf{g} && \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

If we define

$$\mathbf{V} := H^1(\Omega)^d, \qquad Q := L^2_0(\Omega) = \left\{ q \in L^2(\Omega) : \int_\Omega q\, d\mathbf{x} = 0 \right\},$$

and

$$A(\mathbf{u}, \mathbf{v}) = \int_\Omega \nu \nabla\mathbf{u} : \nabla\mathbf{v}\, d\mathbf{x}, \qquad B(\mathbf{u}, p) = -\int_\Omega p\, \nabla \cdot \mathbf{u}\, d\mathbf{x},$$

then the corresponding variational problem consists of finding $(\mathbf{u}, p) \in \mathbf{V} \times Q$, with $\mathbf{u} = \mathbf{g}$ on $\partial\Omega$, such that

$$\begin{cases} A(\mathbf{u}, \mathbf{v}) + B(\mathbf{v}, p) = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\mathbf{x}, \\ B(\mathbf{u}, q) = 0 \end{cases} \tag{2.2}$$

for all $\mathbf{v} \in H^1_0(\Omega)^d$ and $q \in Q$.

The well-posedness of (2.2) is ensured by the continuity of $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$, the coercivity of $A(\cdot,\cdot)$, and the following inf-sup condition

$$\inf_{0 \neq q \in L^2_0(\Omega)} \sup_{\mathbf{0} \neq \mathbf{v} \in H^1_0(\Omega)^d} \frac{-\int_\Omega q\, \nabla \cdot \mathbf{v}\, d\mathbf{x}}{|\mathbf{v}|_1 \|q\|_0} \geq \gamma > 0, \tag{2.3}$$

with an inf-sup constant γ only depending on Ω; see, e.g., [10, 25]. Here, we denote by $\|\cdot\|_{s,\mathcal{D}}$ and $|\cdot|_{s,\mathcal{D}}$ the norm and seminorm of $H^s(\mathcal{D})$ and $H^s(\mathcal{D})^d$, $s \geq 0$. In case $\mathcal{D} = \Omega$, we drop the subscript.

3. Mixed discretizations with nonconforming velocity spaces. Let $\mathbf{V}_h$ be a nonconforming finite element space approximating the velocities. We introduce the space

$$\mathbf{V}(h) := \mathbf{V} + \mathbf{V}_h$$

and endow it with a suitable norm $\|\cdot\|_h$. Furthermore, let $Q_h \subset Q$ be a conforming finite element space for the pressure, equipped with the L^2-norm $\|\cdot\|_0$.

Given linear forms $A_h : \mathbf{V}(h) \times \mathbf{V}(h) \to \mathbb{R}$, $B_h : \mathbf{V}(h) \times Q \to \mathbb{R}$ and continuous linear functionals $F_h : \mathbf{V}_h \to \mathbb{R}$, $G_h : Q_h \to \mathbb{R}$, chosen to discretize the Laplacian and the divergence constraint, we consider mixed methods of the following form: find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that

$$\left\{ \begin{array}{rcrcl} A_h(\mathbf{u}_h, \mathbf{v}) & + & B_h(\mathbf{v}, p_h) & = & F_h(\mathbf{v}), \\ B_h(\mathbf{u}_h, q) & & & = & G_h(q) \end{array} \right. \tag{3.1}$$

for all $(\mathbf{v}, q) \in \mathbf{V}_h \times Q_h$.

Let us make precise our assumptions on the forms A_h and B_h. First, they are assumed to satisfy the following continuity properties:

$$|A_h(\mathbf{v}, \mathbf{w})| \le \alpha_1 \|\mathbf{v}\|_h \|\mathbf{w}\|_h, \qquad \mathbf{v}, \mathbf{w} \in \mathbf{V}(h), \tag{3.2}$$

$$|B_h(\mathbf{v}, q)| \le \alpha_2 \|\mathbf{v}\|_h \|q\|_0, \qquad (\mathbf{v}, q) \in \mathbf{V}(h) \times Q, \tag{3.3}$$

with constants $\alpha_1 > 0$ and $\alpha_2 > 0$. Further, let us define $\mathbf{Z}(G_h) \subset \mathbf{V}_h$ by

$$\mathbf{Z}(G_h) = \{\, \mathbf{v} \in \mathbf{V}_h : B_h(\mathbf{v}, q) = G_h(q) \ \forall q \in Q_h \,\}. \tag{3.4}$$

We require the form A_h to be coercive on the kernel of B_h, i.e.,

$$A_h(\mathbf{v}, \mathbf{v}) \ge \beta \|\mathbf{v}\|_h^2, \qquad \mathbf{v} \in \mathbf{Z}(0), \tag{3.5}$$

for a coercivity constant $\beta > 0$. The form B_h is assumed to satisfy the discrete inf-sup condition

$$\inf_{0 \ne q \in Q_h} \sup_{\mathbf{0} \ne \mathbf{v} \in \mathbf{V}_h} \frac{B_h(\mathbf{v}, q)}{\|\mathbf{v}\|_h \|q\|_0} \ge \gamma_h, \tag{3.6}$$

with a stability constant $\gamma_h > 0$. Finally, we assume the exact solution $\mathbf{u} \in \mathbf{V}$ to fulfill the consistency condition

$$B_h(\mathbf{u}, q) = G_h(q) \qquad \forall q \in Q_h. \tag{3.7}$$

We do not impose any consistency requirements on the form A_h; instead we will work with the residual

$$R_h(\mathbf{u}, p; \mathbf{v}) := A_h(\mathbf{u}, \mathbf{v}) + B_h(\mathbf{v}, p) - F_h(\mathbf{v}), \qquad \mathbf{v} \in \mathbf{V}_h, \tag{3.8}$$

where $(\mathbf{u}, p) \in \mathbf{V} \times Q$ again is the exact solution. Our abstract error estimates will then be expressed in terms of $\mathcal{R}_h(\mathbf{u}, p)$ given by

$$\mathcal{R}_h(\mathbf{u}, p) := \sup_{\mathbf{0} \ne \mathbf{v} \in \mathbf{V}_h} \frac{|R_h(\mathbf{u}, p; \mathbf{v})|}{\|\mathbf{v}\|_h}. \tag{3.9}$$

For all the DG methods we introduce in section 5, the quantity $\mathcal{R}_h(\mathbf{u}, p)$ is optimally convergent.

The mixed problem (3.1) has a unique solution $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ and $\mathbf{Z}(G_h)$ is nonempty.

REMARK 3.1. *If $\mathbf{V}_h \subset \mathbf{V}$ is chosen to be a conforming finite element space, the setting of this section coincides with the standard mixed finite element setting; see* [10].

REMARK 3.2. *For the DG forms in section* 5 *the constants α_1 and β depend on the viscosity ν whereas α_2 and γ_h are independent of ν. More precisely, we have that $\alpha_1 = \nu\bar{\alpha}_1$ and $\beta = \nu\bar{\beta}$ with $\bar{\alpha}_1$ and $\bar{\beta}$ independent of ν.*

4. Abstract error estimates. Abstract error bounds for the mixed method in (3.1) can be obtained by proceeding as in [40, sect. 8]. We give the details of the proofs for the sake of completeness.

4.1. Error in the velocity. First, we prove an error estimate for the velocities following [40, Lemma 8.1].

PROPOSITION 4.1. *Let $(\mathbf{u},p)\in\mathbf{V}\times Q$ be the exact solution of the Stokes problem and $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times Q_h$ the mixed finite element approximation. Under the assumptions of section* 3*, we have*

$$\|\mathbf{u}-\mathbf{u}_h\|_h\le\left(1+\frac{\alpha_1}{\beta}\right)\left(1+\frac{\alpha_2}{\gamma_h}\right)\inf_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{u}-\mathbf{v}\|_h+\frac{\alpha_2}{\beta}\inf_{q\in Q_h}\|p-q\|_0+\beta^{-1}\mathcal{R}_h(\mathbf{u},p).$$

Proof. First, we fix $\mathbf{w}\in\mathbf{Z}(G_h)$ and $q\in Q_h$. Since $\mathbf{w}-\mathbf{u}_h\in\mathbf{Z}(0)$, (3.5) and the definition of the residual yield

$$\begin{aligned}\beta\|\mathbf{w}-\mathbf{u}_h\|_h^2&\le A_h(\mathbf{w}-\mathbf{u}_h,\mathbf{w}-\mathbf{u}_h)\\&=A_h(\mathbf{w}-\mathbf{u},\mathbf{w}-\mathbf{u}_h)-B_h(\mathbf{w}-\mathbf{u}_h,p-p_h)+R_h(\mathbf{u},p;\mathbf{w}-\mathbf{u}_h).\end{aligned}$$

Since $\mathbf{w}-\mathbf{u}_h\in\mathbf{Z}(0)$, we can replace p_h by q in the form B_h. Using the continuity properties in (3.2), (3.3), and the triangle inequality, we obtain

$$\|\mathbf{u}-\mathbf{u}_h\|_h\le\left(1+\frac{\alpha_1}{\beta}\right)\|\mathbf{u}-\mathbf{w}\|_h+\frac{\alpha_2}{\beta}\|p-q\|_0+\beta^{-1}\mathcal{R}_h(\mathbf{u},p)\tag{4.1}$$

for any $\mathbf{w}\in\mathbf{Z}(G_h)$ and $q\in Q_h$.

Second, we fix $\mathbf{v}\in\mathbf{V}_h$ and consider the problem of finding $\mathbf{z}(\mathbf{v})\in\mathbf{V}_h$ such that

$$B_h(\mathbf{z}(\mathbf{v}),q)=B_h(\mathbf{u}-\mathbf{v},q)\qquad\forall q\in Q_h.$$

Thanks to the discrete inf-sup condition in (3.6), the continuity of B_h in (3.3) and [10, Proposition 1.2, p. 39], a solution $\mathbf{z}(\mathbf{v})$ is defined. Furthermore,

$$\gamma_h\|\mathbf{z}(\mathbf{v})\|_h\le\sup_{0\neq q\in Q_h}\frac{B_h(\mathbf{z}(\mathbf{v}),q)}{\|q\|_0}=\sup_{0\neq q\in Q_h}\frac{B_h(\mathbf{u}-\mathbf{v},q)}{\|q\|_0}\le\alpha_2\|\mathbf{u}-\mathbf{v}\|_h,\tag{4.2}$$

where we have used the continuity of B_h. By construction and assumption (3.7), we have $\mathbf{z}(\mathbf{v})+\mathbf{v}\in\mathbf{Z}(G_h)$. Inserting $\mathbf{z}(\mathbf{v})+\mathbf{v}$ in (4.1) yields

$$\|\mathbf{u}-\mathbf{u}_h\|_h\le\left(1+\frac{\alpha_1}{\beta}\right)\|\mathbf{u}-\mathbf{v}\|_h+\left(1+\frac{\alpha_1}{\beta}\right)\|\mathbf{z}(\mathbf{v})\|_h+\frac{\alpha_2}{\beta}\|p-q\|_0+\beta^{-1}\mathcal{R}_h(\mathbf{u},p).$$

This, together with (4.2), proves the assertion. ∎

REMARK 4.2. *Assuming that α_1, α_2, and β are independent of the discretization parameter h, the bound in Proposition* 4.1 *can be expressed in a simpler fashion as*

$$\|\mathbf{u}-\mathbf{u}_h\|_h\le C\big[\gamma_h^{-1}\inf_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{u}-\mathbf{v}\|_h+\inf_{q\in Q_h}\|p-q\|_0+\mathcal{R}_h(\mathbf{u},p)\big].$$

4.2. Error in the pressure. Next, we prove an error estimate for the pressure following the arguments in [40, Lemma 8.2].

PROPOSITION 4.3. *Let $(\mathbf{u},p)\in\mathbf{V}\times Q$ be the exact solution of the Stokes problem and $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times Q_h$ the mixed finite element approximation. Under the assumptions of section* 3*, we have*

$$\|p-p_h\|_0\le\left(1+\frac{\alpha_2}{\gamma_h}\right)\inf_{q\in Q_h}\|p-q\|_0+\frac{\alpha_1}{\gamma_h}\|\mathbf{u}-\mathbf{u}_h\|_h+\gamma_h^{-1}\mathcal{R}_h(\mathbf{u},p).$$

Proof. Fix $q \in Q_h$. From the inf-sup condition in (3.6) we have

$$\gamma_h \|q - p_h\|_0 \le \sup_{\mathbf{0} \neq \mathbf{v} \in \mathbf{V}_h} \frac{B_h(\mathbf{v}, q - p_h)}{\|\mathbf{v}\|_h}.$$

Since $B_h(\mathbf{v}, q - p_h) = B_h(\mathbf{v}, q - p) - A_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) + R_h(\mathbf{u}, p; \mathbf{v})$ for any $\mathbf{v} \in \mathbf{V}_h$, we obtain from the continuity properties in (3.2) and (3.3) and the definition of $\mathcal{R}_h$ in (3.9)

$$\gamma_h \|q - p_h\|_0 \le \alpha_2 \|p - q\|_0 + \alpha_1 \|\mathbf{u} - \mathbf{u}_h\|_h + \mathcal{R}_h(\mathbf{u}, p).$$

The assertion then follows from the triangle inequality. ∎

REMARK 4.4. *Taking into account the estimate for* $\|\mathbf{u} - \mathbf{u}_h\|_h$ *in Proposition* 4.1 *and assuming again that* α_1, α_2, *and* β *are independent of the discretization parameter* h, *the bound in Proposition* 4.3 *reduces to*

$$\|p - p_h\|_0 \le C\big[\gamma_h^{-1} \inf_{q \in Q_h} \|p - q\|_0 + \gamma_h^{-2} \inf_{\mathbf{v} \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}\|_h + \gamma_h^{-1} \mathcal{R}_h(\mathbf{u}, p)\big].$$

5. Discontinuous Galerkin discretizations. In this section, we give several examples of mixed discontinuous Galerkin methods that can be cast into the setting of section 3 by using lifting operators similar to the ones introduced in [2] for the Laplacian.

5.1. Triangulations and finite element spaces. Let $\mathcal{T}_h$ be a shape-regular affine quadrilateral or hexahedral mesh on Ω. We denote by h_K the diameter of the element $K \in \mathcal{T}_h$. Further, we assign to each element $K \in \mathcal{T}_h$ an approximation order $k_K \ge 1$. The local quantities h_K and k_K are stored in the vectors $\underline{h} = \{h_K\}_{K \in \mathcal{T}_h}$ and $\underline{k} = \{k_K\}_{K \in \mathcal{T}_h}$, respectively. We set $h = \max_{K \in \mathcal{T}_h} h_K$ and $|\underline{k}| = \max_{K \in \mathcal{T}_h} k_K$. Finally, $\mathbf{n}_K$ denotes the outward normal unit vector to the boundary ∂K.

An interior face of $\mathcal{T}_h$ is the (nonempty) interior of $\partial K^+ \cap \partial K^-$, where K^+ and K^- are two adjacent elements of $\mathcal{T}_h$. Similarly, a boundary face of $\mathcal{T}_h$ is the (nonempty) interior of $\partial K \cap \partial\Omega$ which consists of entire faces of ∂K. We denote by $\mathcal{E}_\mathcal{I}$ the union of all interior faces of $\mathcal{T}_h$, by $\mathcal{E}_\mathcal{D}$ the union of all boundary faces, and set $\mathcal{E} = \mathcal{E}_\mathcal{I} \cup \mathcal{E}_\mathcal{D}$. Here and in the following, we refer generically to a "face" even in the two-dimensional case.

We allow for irregular meshes, i.e., meshes with hanging nodes (see [37, sect. 4.4.1]), in general, but suppose that the intersection between neighboring elements is either a common vertex, or a common edge, or a common face, or an entire face of one of the two elements. We also assume the local mesh-sizes and approximation degrees to be of bounded variation, that is, there is a constant $\kappa > 0$ such that

$$\kappa h_K \le h_{K'} \le \kappa^{-1} h_K, \qquad \kappa k_K \le k_{K'} \le \kappa^{-1} k_K, \tag{5.1}$$

whenever K and K' share a common face.

We wish to approximate the velocities and pressures in the discontinuous finite element spaces $\mathbf{V}_h$ and Q_h given by

$$\begin{aligned} \mathbf{V}_h &= \{\, \mathbf{v} \in L^2(\Omega)^d : \mathbf{v}|_K \in \mathbb{Q}_{k_K}(K)^d,\ K \in \mathcal{T}_h \,\}, \\ Q_h &= \{\, q \in L^2_0(\Omega) : q|_K \in \mathbb{Q}_{k_K - 1}(K),\ K \in \mathcal{T}_h \,\}, \end{aligned} \tag{5.2}$$

respectively, where $\mathbb{Q}_k(K)$ is the space of polynomials of maximum degree k in each variable on K.

For the derivation and analysis of the methods, we will make use of the auxiliary space $\underline{\Sigma}_h$ defined by

$$\underline{\Sigma}_h := \{\, \underline{\tau} \in L^2(\Omega)^{d\times d} : \underline{\tau} \in \mathbb{Q}_{k_K}(K)^{d\times d},\ K \in \mathcal{T}_h \,\}.$$

Note that $\nabla_h \mathbf{V}_h \subset \underline{\Sigma}_h$, where ∇_h is the discrete gradient, taken elementwise, and given by $[\nabla \mathbf{v}]_{ij} = \partial_j v_i = \frac{\partial v_i}{\partial x_j}$ on $K \in \mathcal{T}_h$.

5.2. Trace operators. In this section, we define the trace operators needed in our discontinuous Galerkin discretizations. To this end, let $e \subset \mathcal{E}_\mathcal{I}$ be an interior face shared by K^+ and K^-. Let $(\mathbf{v}, q, \underline{\tau})$ be a function smooth inside each element $K^\pm$ and let us denote by $(\mathbf{v}^\pm, q^\pm, \underline{\tau}^\pm)$ the traces of $(\mathbf{v}, q, \underline{\tau})$ on e from the interior of $K^\pm$. Then, we define the mean values $\{\!\{\cdot\}\!\}$ and normal jumps $[\![\cdot]\!]$ at $\mathbf{x} \in e$ as

$$\begin{aligned}
\{\!\{\mathbf{v}\}\!\} &:= (\mathbf{v}^+ + \mathbf{v}^-)/2, & [\![\mathbf{v}]\!] &:= \mathbf{v}^+ \cdot \mathbf{n}_{K^+} + \mathbf{v}^- \cdot \mathbf{n}_{K^-},\\
\{\!\{q\}\!\} &:= (q^+ + q^-)/2, & [\![q]\!] &:= q^+ \,\mathbf{n}_{K^+} + q^- \,\mathbf{n}_{K^-},\\
\{\!\{\underline{\tau}\}\!\} &:= (\underline{\tau}^+ + \underline{\tau}^-)/2, & [\![\underline{\tau}]\!] &:= \underline{\tau}^+ \,\mathbf{n}_{K^+} + \underline{\tau}^- \,\mathbf{n}_{K^-}.
\end{aligned}$$

Note that the jumps $[\![q]\!]$ and $[\![\underline{\tau}]\!]$ are both vectors whereas the jump $[\![\mathbf{v}]\!]$ is a scalar. We also need to define a jump of the velocity $\mathbf{v}$ which is a matrix, namely,

$$\underline{[\![\mathbf{v}]\!]} := \mathbf{v}^+ \otimes \mathbf{n}_{K^+} + \mathbf{v}^- \otimes \mathbf{n}_{K^-},$$

where, for two vectors $\mathbf{a}$ and $\mathbf{b}$, we set $[\mathbf{a} \otimes \mathbf{b}]_{ij} = a_i b_j$.

On a boundary face $e \subset \mathcal{E}_\mathcal{D}$ given by $e = \partial K \cap \partial\Omega$, we accordingly set

$$\{\!\{\mathbf{v}\}\!\} := \mathbf{v}, \qquad \{\!\{q\}\!\} := q, \qquad \{\!\{\underline{\tau}\}\!\} := \underline{\tau},$$

as well as

$$[\![\mathbf{v}]\!] := \mathbf{v} \cdot \mathbf{n}, \qquad \underline{[\![\mathbf{v}]\!]} := \mathbf{v} \otimes \mathbf{n}, \qquad [\![q]\!] := q\mathbf{n}, \qquad [\![\underline{\tau}]\!] := \underline{\tau}\,\mathbf{n}.$$

We remark that, for the exact solution $(\mathbf{u}, p) \in \mathbf{V} \times Q$, there holds $\underline{[\![\mathbf{u}]\!]} = \underline{0}$ and $[\![\nu\nabla\mathbf{u} - p\underline{I}]\!] = \mathbf{0}$ on $\mathcal{E}_\mathcal{I}$. The last property follows from the fact that $\nu\nabla\mathbf{u} - p\underline{I}$ belongs to $H(\mathrm{div}; \Omega)$; see [40].

5.3. Lifting operators. We introduce the following lifting operators. First, for a face $e \subset \mathcal{E}$ we define $\underline{\mathcal{L}}_e : \mathbf{V}(h) \to \underline{\Sigma}_h$ by

$$\int_\Omega \underline{\mathcal{L}}_e(\mathbf{v}) : \underline{\tau}\, d\mathbf{x} = \int_e \underline{[\![\mathbf{v}]\!]} : \{\!\{\underline{\tau}\}\!\}\, ds \qquad \forall \underline{\tau} \in \underline{\Sigma}_h.$$

Note that the support of $\underline{\mathcal{L}}_e(\mathbf{v})$ is contained in the elements that share the face e. For a boundary face $e \subset \mathcal{E}_\mathcal{D}$, we introduce the lifting $\underline{\mathcal{G}}_e \in \underline{\Sigma}_h$ of the Dirichlet datum $\mathbf{g}$ given by

$$\int_\Omega \underline{\mathcal{G}}_e : \underline{\tau}\, d\mathbf{x} = \int_e (\mathbf{g} \otimes \mathbf{n}) : \underline{\tau}\, ds \qquad \forall \underline{\tau} \in \underline{\Sigma}_h.$$

For the exact solution $\mathbf{u} \in \mathbf{V}$, we have

$$\underline{\mathcal{L}}_e(\mathbf{u}) = \underline{0} \quad \forall e \subset \mathcal{E}_\mathcal{I}, \qquad \underline{\mathcal{L}}_e(\mathbf{u}) = \underline{\mathcal{G}}_e \quad \forall e \subset \mathcal{E}_\mathcal{D}. \tag{5.3}$$

Globally, we define $\underline{\mathcal{L}} : \mathbf{V}(h) \to \underline{\Sigma}_h$ and $\underline{\mathcal{G}} \in \underline{\Sigma}_h$ by

$$\underline{\mathcal{L}} := \sum_{e \subset \mathcal{E}} \underline{\mathcal{L}}_e, \qquad \underline{\mathcal{G}} := \sum_{e \subset \mathcal{E}_\mathcal{D}} \underline{\mathcal{G}}_e.$$

These operators can be characterized by

$$\int_\Omega \underline{\mathcal{L}}(\mathbf{v}) : \underline{\tau}\, d\mathbf{x} = \int_\mathcal{E} \underline{[\![\mathbf{v}]\!]} : \{\!\{\underline{\tau}\}\!\}\, ds \qquad \forall \underline{\tau} \in \underline{\Sigma}_h,$$
$$\int_\Omega \underline{\mathcal{G}} : \underline{\tau}\, d\mathbf{x} = \int_{\mathcal{E}_\mathcal{D}} (\mathbf{g} \otimes \mathbf{n}) : \underline{\tau}\, ds \qquad \forall \underline{\tau} \in \underline{\Sigma}_h.$$

Finally, we need the lifting operator $\mathcal{M} : \mathbf{V}(h) \to Q_h$ defined by

$$\int_\Omega \mathcal{M}(\mathbf{v}) \varphi\, d\mathbf{x} = \int_\mathcal{E} [\![\mathbf{v}]\!]\, \{\!\{\varphi\}\!\}\, ds \qquad \forall \varphi \in Q_h.$$

For the exact solution $\mathbf{u} \in \mathbf{V}$, there holds

$$\int_\Omega \mathcal{M}(\mathbf{u})\, \varphi\, d\mathbf{x} = \int_{\mathcal{E}_\mathcal{D}} \varphi\, \mathbf{g} \cdot \mathbf{n}\, ds \qquad \forall \varphi \in Q_h. \tag{5.4}$$

5.4. Mixed discontinuous Galerkin problems. We introduce mixed discontinuous Galerkin methods of the form (3.1) for the mixed-order spaces in (5.2): find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that

$$\left\{ \begin{array}{lcl} A_h(\mathbf{u}_h, \mathbf{v}) + B_h(\mathbf{v}, p_h) &=& F_h(\mathbf{v}), \\ B_h(\mathbf{u}_h, q) &=& G_h(q) \end{array} \right. \tag{5.5}$$

for all $(\mathbf{v}, q) \in \mathbf{V}_h \times Q_h$.

The form $B_h : \mathbf{V}(h) \times Q \to \mathbb{R}$ and the functional $G_h : Q_h \to \mathbb{R}$ will always be chosen as

$$B_h(\mathbf{v}, q) = -\int_\Omega q\, [\nabla_h \cdot \mathbf{v} - \mathcal{M}(\mathbf{v})]\, d\mathbf{x}, \qquad \mathbf{v} \in \mathbf{V}(h),\ q \in Q,$$
$$G_h(q) = \int_{\mathcal{E}_\mathcal{D}} q\, \mathbf{g} \cdot \mathbf{n}\, ds, \qquad q \in Q_h,$$

Restricted to discrete functions $(\mathbf{v}, q) \in \mathbf{V}_h \times Q_h$, we have

$$B_h(\mathbf{v}, q) = -\int_\Omega q\, \nabla_h \cdot \mathbf{v}\, d\mathbf{x} + \int_\mathcal{E} \{\!\{q\}\!\} [\![\mathbf{v}]\!]\, ds. \tag{5.6}$$

Thus, we exactly obtain the form B_h and the functional G_h considered in the mixed DG approaches in [17, 26, 40]. We remark that (3.7) is satisfied thanks to (5.4).

For discrete functions, we equivalently have

$$B_h(\mathbf{v}, q) = \int_\Omega \nabla_h q \cdot \mathbf{v}\, d\mathbf{x} - \int_{\mathcal{E}_\mathcal{I}} [\![q]\!] \cdot \{\!\{\mathbf{v}\}\!\}\, ds, \qquad (\mathbf{v}, q) \in \mathbf{V}_h \times Q_h. \tag{5.7}$$

This follows from integration by parts and elementary manipulations; see equation (4.7) in [40].

The space $\mathbf{V}(h) = \mathbf{V} + \mathbf{V}_h$ is endowed with the broken norm

$$\|\mathbf{v}\|_h^2 = \sum_{K\in\mathcal{T}_h} |\mathbf{v}|_{1,K}^2 + \int_{\mathcal{E}} \sigma |\underline{[\![\mathbf{v}]\!]}|^2 \, ds, \qquad \mathbf{v} \in \mathbf{V}(h), \tag{5.8}$$

where $\sigma \in L^\infty(\mathcal{E})$ is the so-called discontinuity stabilization function that we choose in terms of the local mesh-sizes and the polynomial degrees as follows. Define the functions $\mathtt{h} \in L^\infty(\mathcal{E})$ and $\mathtt{k} \in L^\infty(\mathcal{E})$ by

$$\mathtt{h}(\mathbf{x}) := \begin{cases} \min\{h_K, h_{K'}\}, & \mathbf{x} \text{ in the interior of } \partial K \cap \partial K', \\ h_K, & \mathbf{x} \text{ in the interior of } \partial K \cap \partial\Omega, \end{cases}$$

$$\mathtt{k}(\mathbf{x}) := \begin{cases} \max\{k_K, k_{K'}\}, & \mathbf{x} \text{ in the interior of } \partial K \cap \partial K', \\ k_K, & \mathbf{x} \text{ in the interior of } \partial K \cap \partial\Omega. \end{cases}$$

Then we set

$$\sigma = \sigma_0 \mathtt{h}^{-1}\mathtt{k}^2, \tag{5.9}$$

with a parameter $\sigma_0 > 0$ that is independent of $\mathtt{h}$ and $\mathtt{k}$.

For the form A_h related to the Laplacian, several choices are possible. Let us discuss the stable and consistent forms in the sense of [2].

The interior penalty forms A_h. The symmetric interior penalty (IP) form has been used in the mixed DG method introduced in [26]. It is obtained by first defining the stabilization form I_h^σ as

$$I_h^\sigma(\mathbf{u},\mathbf{v}) := \nu \int_{\mathcal{E}} \sigma \underline{[\![\mathbf{u}]\!]} : \underline{[\![\mathbf{v}]\!]} \, ds, \qquad \mathbf{u},\mathbf{v} \in \mathbf{V}(h), \tag{5.10}$$

where σ is the discontinuity stabilization function in (5.9), and then by taking, for $\mathbf{u},\mathbf{v} \in \mathbf{V}(h)$,

$$\begin{aligned} A_h(\mathbf{u},\mathbf{v}) &= \int_\Omega \nu \big[\nabla_h \mathbf{u} : \nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{u}) : \nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v}) : \nabla_h \mathbf{u}\big] \, d\mathbf{x} + I_h^\sigma(\mathbf{u},\mathbf{v}), \\ F_h(\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v} \, d\mathbf{x} - \nu \int_\Omega \underline{\mathcal{G}} : \nabla_h \mathbf{v} \, d\mathbf{x} + \nu \int_{\mathcal{E}_\mathcal{D}} \sigma \mathbf{g}\cdot\mathbf{v} \, ds. \end{aligned} \tag{5.11}$$

Restricted to discrete functions $\mathbf{u},\mathbf{v} \in \mathbf{V}_h$, we have

$$A_h(\mathbf{u},\mathbf{v}) = \int_\Omega \nu \nabla_h \mathbf{u} : \nabla_h \mathbf{v} \, d\mathbf{x} - \int_{\mathcal{E}} \big(\{\!\{\nu \nabla_h \mathbf{v}\}\!\} : \underline{[\![\mathbf{u}]\!]} + \{\!\{\nu \nabla_h \mathbf{u}\}\!\} : \underline{[\![\mathbf{v}]\!]} \big) \, ds + I_h^\sigma(\mathbf{u},\mathbf{v}).$$

The nonsymmetric variant of the IP form has been studied in the mixed DG approach in [40] (see also [35, 27] for scalar convection-diffusion problems). It is obtained by choosing, for $\mathbf{u},\mathbf{v} \in \mathbf{V}(h)$,

$$\begin{aligned} A_h(\mathbf{u},\mathbf{v}) &= \int_\Omega \nu \big[\nabla_h \mathbf{u} : \nabla_h \mathbf{v} + \underline{\mathcal{L}}(\mathbf{u}) : \nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v}) : \nabla_h \mathbf{u}\big] \, d\mathbf{x} + I_h^\sigma(\mathbf{u},\mathbf{v}), \\ F_h(\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v} \, d\mathbf{x} + \nu \int_\Omega \underline{\mathcal{G}} : \nabla_h \mathbf{v} \, d\mathbf{x} + \nu \int_{\mathcal{E}_\mathcal{D}} \sigma \mathbf{g}\cdot\mathbf{v} \, ds. \end{aligned} \tag{5.12}$$

REMARK 5.1. *For $\sigma \equiv 0$ the form A_h in (5.12) coincides with the form given by the so-called Baumann–Oden method* [7, 29]. *Further, realizations of the methods of Baker, Jureidini, and Karakashian* [4] *and Karakashian and Jureidini* [28] *are obtained with the IP form A_h in* (5.11), *if we choose the spaces $\widetilde{\mathbf{V}}_h = \{\mathbf{v} \in \mathbf{V}_h : \mathbf{v}|_K$ is divergence free on each $K \in \mathcal{T}_h\}$ and $\widetilde{Q}_h = Q_h \cap C^0(\overline{\Omega})$, respectively.*

The LDG form A_h. The local discontinuous Galerkin (LDG) form is closely related to the IP forms since it is also expressed in terms of the stabilization form I_h^σ in (5.10). In the context of the Stokes problem, it has been studied in [17] (see also [19, 13, 31]). In the primal variables, the LDG form is given by taking, for $\mathbf{u}, \mathbf{v} \in \mathbf{V}(h)$,

$$\begin{aligned} A_h(\mathbf{u},\mathbf{v}) &= \int_\Omega \nu \left[\nabla_h \mathbf{u} - \underline{\mathcal{L}}(\mathbf{u})\right] : \left[\nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v})\right] d\mathbf{x} + I_h^\sigma(\mathbf{u},\mathbf{v}), \\ F_h(\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} - \nu \int_\Omega \underline{\mathcal{G}} : \left(\nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v})\right) d\mathbf{x} + \nu \int_{\mathcal{E}_\mathcal{D}} \sigma \mathbf{g}\cdot\mathbf{v}\, ds. \end{aligned} \tag{5.13}$$

The Bassi–Rebay forms A_h. These forms were inspired by the original Bassi–Rebay method in [5], which, in fact, is unstable. They are defined by introducing a different stabilization form I_h^η given by

$$I_h^\eta(\mathbf{u},\mathbf{v}) = \nu \sum_{e\subset\mathcal{E}} \int_\Omega \eta\, \underline{\mathcal{L}}_e(\mathbf{u}) : \underline{\mathcal{L}}_e(\mathbf{v})\, d\mathbf{x}, \qquad \mathbf{u},\mathbf{v} \in \mathbf{V}(h), \tag{5.14}$$

for a parameter $\eta > 0$. The first form we present here was introduced in [6] and is obtained by choosing, for $\mathbf{u}, \mathbf{v} \in \mathbf{V}(h)$,

$$\begin{aligned} A_h(\mathbf{u},\mathbf{v}) &= \int_\Omega \nu \left[\nabla_h \mathbf{u} : \nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{u}) : \nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v}) : \nabla_h \mathbf{u}\right] d\mathbf{x} + I_h^\eta(\mathbf{u},\mathbf{v}), \\ F_h(\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} - \nu \int_\Omega \underline{\mathcal{G}} : \nabla_h \mathbf{v}\, d\mathbf{x} + \nu \sum_{e\subset\mathcal{E}_\mathcal{D}} \int_\Omega \eta\, \underline{\mathcal{G}}_e : \underline{\mathcal{L}}_e(\mathbf{v})\, d\mathbf{x}. \end{aligned} \tag{5.15}$$

In [11], the following variant of the Bassi–Rebay form has been proposed:

$$\begin{aligned} A_h(\mathbf{u},\mathbf{v}) &= \int_\Omega \nu \left[\nabla_h \mathbf{u} - \underline{\mathcal{L}}(\mathbf{u})\right] : \left[\nabla_h \mathbf{v} - \underline{\mathcal{L}}(\mathbf{v})\right] d\mathbf{x} + I_h^\eta(\mathbf{u},\mathbf{v}), \\ F_h(\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} - \nu \int_\Omega \underline{\mathcal{G}} : \nabla_h \mathbf{v}\, d\mathbf{x} + \nu \sum_{e\subset\mathcal{E}_\mathcal{D}} \int_\Omega \eta\, \underline{\mathcal{G}}_e : \underline{\mathcal{L}}_e(\mathbf{v})\, d\mathbf{x}. \end{aligned} \tag{5.16}$$

6. Divergence stability. In this section, we establish an inf-sup condition for the form $B_h(\cdot,\cdot)$ with respect to the norm $\|\cdot\|_h$ in (5.8)–(5.9) and for the $\mathbb{Q}_k - \mathbb{Q}_{k-1}$ spaces in (5.2). We recall that the divergence bilinear form is the same for all the methods that we consider.

6.1. The discrete inf-sup condition. Let us begin by stating our main stability result.

PROPOSITION 6.1. *Let $k_K \geq 2$ for all $K \in \mathcal{T}_h$. Then there are constants $c_1 > 0$ and $c_2 > 0$, independent of $\underline{h}$ and $\underline{k}$, such that for each $q \in Q_h$ there exists a discrete velocity field $\mathbf{v} \in \mathbf{V}_h$ such that*

$$B_h(\mathbf{v},q) \geq c_1 \|q\|_0^2, \qquad \|\mathbf{v}\|_h \leq c_2 |\underline{k}|\, \|q\|_0.$$

From the above result, we immediately find the following stability result.

THEOREM 6.2. *There exists a constant $c > 0$, independent of $\underline{h}$ and $\underline{k}$, such that, for $k_K \geq 2$,*

$$\inf_{0\neq q\in Q_h} \sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{V}_h} \frac{B_h(\mathbf{v},q)}{\|\mathbf{v}\|_h \|q\|_0} \geq \gamma_h \geq c\, |\underline{k}|^{-1}. \tag{6.1}$$

REMARK 6.3. *Theorem* 6.2 *establishes an hp-version divergence stability result for the* $\mathbb{Q}_k - \mathbb{Q}_{k-1}$ *element family where the difference in the approximation orders for the velocity and the pressure is exactly one. It is well known that these elements are unstable in a conforming setting, although they are optimal in terms of the approximation properties of the finite element spaces. The use of discontinuous velocities overcomes these usual stability problems in a natural way. In addition, the bound* (6.1) *holds in two and three dimensions, and it is identical to the bound established in* [38] *for conforming mixed hp-FEM in three dimensions, although there* $\mathbb{Q}_k - \mathbb{Q}_{k-2}$ *spaces have been used.*

REMARK 6.4. *The technique we use to prove this result is a combination of the h-version approach in* [26] *that makes use of* $H(\mathrm{div})$*-conforming projectors and of the work* [40] *that allows us to deal with hanging nodes. Indeed, we also decompose the pressure into piecewise constants and polynomials whose mean values vanish elementwise as in* [40] *(see also the analysis for conforming hp-methods in* [38]*) and use the low-order stability results in two and three dimensions of* [36, 41] *for* $\mathbb{Q}_2 - \mathbb{Q}_0$ *elements on irregular meshes. That is the reason why we assume* $k_K \geq 2$ *in Proposition* 6.1. *We remark that for* $\mathbb{Q}_1 - \mathbb{Q}_0$ *elements and conforming meshes, divergence stability can be obtained by establishing directly a Fortin property. We report on this case in more detail in section* 6.5.

REMARK 6.5. *The numerical tests in* [40] *show that in two dimensions a stability constant independent of* h and k *is expected, indicating that the dependence on* k *in* (6.1) *is not likely to be sharp.*

REMARK 6.6. *As can be inferred from its proof, the result of Theorem* 6.2 *holds, in fact, for the strictly smaller velocity space* $\widetilde{\mathbf{V}}_h \subset \mathbf{V}_h$ *given by*

$$\widetilde{\mathbf{V}}_h = \{\mathbf{v} \in \mathbf{V}_h \cap H_0(\mathrm{div};\Omega) : \mathbf{v}|_K \in RT_{k_K-1}(K),\ K \in \mathcal{T}_h\},$$

for the Raviart–Thomas space $RT_{k_K-1}(K)$ *of degree* $k_K - 1$ *introduced in the next section. Here,* $H_0(\mathrm{div};\Omega) = \{\mathbf{v} \in L^2(\Omega)^d : \nabla \cdot \mathbf{v} \in L^2(\Omega),\ \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega\}$. *Whether or not the dependence of the inf-sup constant* γ_h *on the polynomial degree in* (6.1) *is sharp for this space is an open issue.*

The remaining part of this section is devoted to the proof of Proposition 6.1. We will carry out the proof for the three-dimensional case and note that the result in two dimensions is obtained completely analogously. We start in section 6.2 by defining Raviart–Thomas interpolation operators that we shall use as Fortin operators. In section 6.3, we establish new stability results for these operators. The proof of Proposition 6.1 is then given in section 6.4. In section 6.5, we report on some extensions of our stability result to uniform approximation orders and conforming meshes, also including $\mathbb{Q}_1 - \mathbb{Q}_0$ elements.

6.2. Raviart–Thomas spaces and interpolants. Given the reference cube $\hat{K} = (-1,1)^3$ and an integer $k \geq 0$, we consider the space

$$RT_k(\hat{K}) = \mathbb{Q}_{k+1,k,k}(\hat{K}) \times \mathbb{Q}_{k,k+1,k}(\hat{K}) \times \mathbb{Q}_{k,k,k+1}(\hat{K}),$$

where $\mathbb{Q}_{k_1,k_2,k_3}(\hat{K})$ is the space of polynomials of degree at most k_i in the ith variable. For an affinely mapped element $K \in \mathcal{T}_h$ the space $RT_k(K)$ is defined by suitably mapping functions in $RT_k(\hat{K})$ using a Piola transformation; see [10, sect. 3.3] or [1, sect. 3.3] for further details.

We denote the faces of $\hat{K}$ by γ_m, $m = 1, \dots, 6$. In particular, we set

$$\begin{aligned} \gamma_1 &= \{x = -1\}, \qquad \gamma_2 = \{x = 1\}, \\ \gamma_3 &= \{y = -1\}, \qquad \gamma_4 = \{y = 1\}, \\ \gamma_5 &= \{z = -1\}, \qquad \gamma_6 = \{z = 1\}. \end{aligned}$$

We use the same notation for an affinely mapped element K, where the faces are obtained by mapping the corresponding ones of $\hat{K}$. Moreover, we denote by $\mathbb{Q}_{k,k}(\gamma_m)$ the space of polynomials of degree at most k in each variable on the face γ_m.

On the reference cube, there is a unique interpolation operator $\Pi_{\hat{K}} : H^1(\hat{K})^3 \to RT_k(\hat{K})$ such that

$$\begin{aligned} &\int_{\hat{K}} \big(\Pi_{\hat{K}}\mathbf{w} - \mathbf{w}\big) \cdot \mathbf{r}\, d\mathbf{x} = 0, \quad \mathbf{r} \in \mathbb{Q}_{k-1,k,k}(\hat{K}) \times \mathbb{Q}_{k,k-1,k}(\hat{K}) \times \mathbb{Q}_{k,k,k-1}(\hat{K}), \\ &\int_{\gamma_m} \big(\Pi_{\hat{K}}\mathbf{w} - \mathbf{w}\big) \cdot \mathbf{n}\, \varphi\, ds = 0, \quad \varphi \in \mathbb{Q}_{k,k}(\gamma_m), \quad m = 1, \dots, 6; \end{aligned} \tag{6.2}$$

see [10] or [1]. For $k = 0$, the first condition in (6.2) is void. For an element $K \in \mathcal{T}_h$, the interpolant $\Pi_K : H^1(K)^3 \to RT_k(K)$ can be defined by using a Piola transform in such a way that the orthogonality conditions in (6.2) also hold for Π_K; see, e.g., [1, sect. 3.5].

6.3. Stability of the Raviart–Thomas interpolant. In order to prove our stability results for the operator Π_K, we need to introduce a representation formula, originally proposed in [1] for the two-dimensional case. We start by defining some additional operators for the reference cube $\hat{K}$. Given integers k_1, k_2, and k_3, we define

$$\hat{Q}_{k_1,k_2,k_3} = \pi^z_{k_3} \otimes \pi^y_{k_2} \otimes \pi^x_{k_1} : L^2(\hat{K}) \to \mathbb{Q}_{k_1,k_2,k_3}(\hat{K})$$

as the L^2-orthogonal projection onto $\mathbb{Q}_{k_1,k_2,k_3}(\hat{K})$. We note that $\hat{Q}_{k_1,k_2,k_3}$ is the tensor product of one-dimensional L^2-projections π_{k_i} on the reference interval $I = (-1, 1)$.

We next introduce extension operators from the faces γ_m. To that end, we denote by L_k, $k \geq 0$, the Legendre polynomial of degree k in I; see [9, sect. 3]. For the face γ_1, we define $\mathcal{E}^{\gamma_1}_k : \mathbb{Q}_{k,k}(\gamma_1) \to \mathbb{Q}_{k+1,k,k}(\hat{K})$ as

$$(\mathcal{E}^{\gamma_1}_k \varphi)(x, y, z) := M^{\gamma_1}_k(x)\varphi(y, z), \qquad M^{\gamma_1}_k(x) := \frac{(-1)^{k+1}}{2}(L_{k+1}(x) - L_k(x)).$$

We note that

$$(\mathcal{E}^{\gamma_1}_k \varphi)|_{\gamma_1} = \varphi, \qquad (\mathcal{E}^{\gamma_1}_k \varphi)|_{\gamma_2} = 0,$$

and that $(\mathcal{E}^{\gamma_1}_k \varphi)|_{\gamma_m}$, $m = 3, \dots, 6$, does not vanish in general. Analogous definitions hold for the other faces γ_m, $m = 2, \dots, 6$.

Similar to [1, Lemma 3], for $\mathbf{w} = (w_x, w_y, w_z) \in H^1(\hat{K})^3$, the interpolant $\mathbf{v} = \Pi_{\hat{K}}\mathbf{w}$ can be written as $\mathbf{v} = (v_x, v_y, v_z)$ with

$$\begin{aligned} v_x &= \hat{Q}_{k-1,k,k} w_x + \sum_{m=1}^{2} \mathcal{E}^{\gamma_m}_k (\pi^y_k \circ \pi^z_k) \left(w_x - \hat{Q}_{k-1,k,k} w_x \right), \\ v_y &= \hat{Q}_{k,k-1,k} w_y + \sum_{m=3}^{4} \mathcal{E}^{\gamma_m}_k (\pi^x_k \circ \pi^z_k) \left(w_y - \hat{Q}_{k,k-1,k} w_y \right), \\ v_z &= \hat{Q}_{k,k,k-1} w_z + \sum_{m=5}^{6} \mathcal{E}^{\gamma_m}_k (\pi^x_k \circ \pi^y_k) \left(w_z - \hat{Q}_{k,k,k-1} w_z \right), \end{aligned} \tag{6.3}$$

where, e.g., $(\pi_k^y \circ \pi_k^z)(w_x - \hat{Q}_{k-1,k,k} w_x)$ is understood as $\pi_k^y \circ \pi_k^z$ applied to the restriction of $(w_x - \hat{Q}_{k-1,k,k} w_x)$ to γ_m, $m = 1, 2$.

Before proving our stability results, we need some technical lemmas. The results in the following lemma can be found using similar techniques as in Theorem 2.2 in [12], Lemma 3.9 in [27], and Theorems 3.91 and 3.92 in [37].

LEMMA 6.7. *We have the following estimates.*

1. *Let $w \in H^1(\hat{K})$. Then there exists a constant $C > 0$ independent of k such that*

$$|\hat{Q}_{k-1,k,k} w|^2_{1,\hat{K}} \le C\, k\, |w|^2_{1,\hat{K}}, \tag{6.4}$$

$$\|w - \hat{Q}_{k-1,k,k} w\|^2_{0,\gamma_m} \le C\, k^{-1}\, |w|^2_{1,\hat{K}}, \qquad m = 1, \dots, 6. \tag{6.5}$$

2. *Let $I = (-1, 1)$ and $w \in \mathbb{Q}_k(I)$. Then there exists a constant $C > 0$ such that*

$$|w|_{1,I} \le Ck^2 \|w\|_{0,I}, \tag{6.6}$$

$$\|w\|_{\infty,I} \le Ck \|w\|_{0,I}. \tag{6.7}$$

The following lemma can be proved by using the properties of the Legendre polynomials given, e.g., in Theorem 3.2 and Remark 3.2 in [9], and Theorem 3.96 in [37].

LEMMA 6.8. *Let*

$$M_k^{\gamma_1}(x) = \frac{(-1)^{k+1}}{2}(L_{k+1}(x) - L_k(x)).$$

Then

$$\|M_k^{\gamma_1}\|^2_{0,I} \le Ck^{-1}, \qquad |M_k^{\gamma_1}|^2_{1,I} \le Ck^3.$$

Similar estimates hold for the other faces γ_m.

We have the following stability result.

LEMMA 6.9. *There exists a constant $C > 0$, independent of h_K and k, such that, for $\mathbf{w} \in H^1(K)^3$,*

$$|\Pi_K \mathbf{w}|^2_{1,K} \le Ck^2 |\mathbf{w}|^2_{1,K}.$$

Proof. Let $\mathbf{w} = (w_x, w_y, w_z)$. We set $\mathbf{v} = \Pi_K \mathbf{w}$ and $\mathbf{v} = (v_x, v_y, v_z)$. We only find a bound for the first component v_x. Bounds for v_y and v_z can be similarly obtained. In addition, we only consider the reference cube $\hat{K} = (-1, 1)^3$ since a bound for an affinely mapped K can be easily deduced using a scaling argument. We consider the two terms of v_x in (6.3). Thanks to (6.4), we have

$$|\hat{Q}_{k-1,k,k} w_x|_{1,\hat{K}} \le Ck^{\frac{1}{2}} |w_x|_{1,\hat{K}}. \tag{6.8}$$

We now consider the face γ_1. Using Lemma 6.8 and the stability of the L^2-projection, we can write

$$\begin{aligned}\|\partial_x(\mathcal{E}_k^{\gamma_1}(\pi_k^y \circ \pi_k^z)(w_x - \hat{Q}_{k-1,k} w_x))\|^2_{0,\hat{K}} &= |M_k^{\gamma_1}|^2_{1,I}\, \|(\pi_k^y \circ \pi_k^z)(w_x - \hat{Q}_{k-1,k,k} w_x)\|^2_{0,\gamma_1} \\ &\le Ck^3 \|w_x - \hat{Q}_{k-1,k,k} w_x\|^2_{0,\gamma_1},\end{aligned}$$

and, thanks to the estimate (6.5),

$$\|\partial_x(\mathcal{E}_k^{\gamma_1}(\pi_k^y \circ \pi_k^z)(w_x - \hat{Q}_{k-1,k,k} w_x))\|^2_{0,\hat{K}} \le Ck^2 |w_x|^2_{1,\hat{K}}. \tag{6.9}$$

Using Lemma 6.8 and the inverse estimate (6.6), we find

$$\begin{aligned}\|\partial_y(\mathcal{E}_k^{\gamma_1}(\pi_k^y\circ\pi_k^z)(w_x-\hat{Q}_{k-1,k,k}w_x))\|_{0,\hat{K}}^2 &= \|M_k^{\gamma_1}\|_{0,I}^2\,\|\partial_y((\pi_k^y\circ\pi_k^z)(w_x-\hat{Q}_{k-1,k}w_x))\|_{0,\gamma_1}^2\\ &\le Ck^{-1}k^4\|w_x-\hat{Q}_{k-1,k,k}w_x\|_{0,\gamma_1}^2,\end{aligned}$$

and, due to the estimate (6.5),

$$\|\partial_y(\mathcal{E}_k^{\gamma_1}(\pi_k^y\circ\pi_k^z)(w_x-\hat{Q}_{k-1,k,k}w_x))\|_{0,\hat{K}}^2 \le Ck^2|w_x|_{1,\hat{K}}^2. \tag{6.10}$$

Analogously,

$$\|\partial_z(\mathcal{E}_k^{\gamma_1}(\pi_k^y\circ\pi_k^z)(w_x-\hat{Q}_{k-1,k,k}w_x))\|_{0,\hat{K}}^2 \le Ck^2|w_x|_{1,\hat{K}}^2. \tag{6.11}$$

Similar estimates can be found for the face γ_2. The proof is completed by combining (6.3), (6.8), (6.9), (6.10), and (6.11) with a scaling argument. ∎

On the boundary ∂K of an element K, we have the following bound.

LEMMA 6.10. *There exists a constant $C>0$, independent of h_K and k, such that, for $\mathbf{w}\in H^1(K)^3$,*

$$\|\mathbf{w}-\Pi_K\mathbf{w}\|_{0,\partial K}^2 \le Ch_K|\mathbf{w}|_{1,K}^2.$$

Proof. Let $\mathbf{v}=\Pi_K\mathbf{w}$. First, we find a bound for the first component v_x of $\mathbf{v}$ on the reference cube $\hat{K}=(-1,1)^3$.

On the face γ_1, we have

$$w_x-v_x = w_x-\hat{Q}_{k-1,k,k}w_x-(\pi_k^y\circ\pi_k^z)(w_x-\hat{Q}_{k-1,k,k}w_x).$$

Hence, by the triangle inequality, (6.5), and by the stability of the L^2-projection, we obtain

$$\|w_x-v_x\|_{0,\gamma_1}^2 \le Ck^{-1}|w_x|_{1,\hat{K}}^2.$$

An analogous estimate holds on γ_2.

Consider now the face γ_3. We have

$$\begin{aligned}\|w_x-v_x\|_{0,\gamma_3}^2 \le{}& C\|w_x-\hat{Q}_{k-1,k,k}w_x\|_{0,\gamma_3}^2\\ &+C\sum_{m=1}^{2}\left\|\mathcal{E}_k^{\gamma_m}(\pi_k^y\circ\pi_k^z)\left(w_x-\hat{Q}_{k-1,k,k}w_x\right)\right\|_{0,\gamma_3}^2.\end{aligned}$$

The first term above can be bounded again by using (6.5). Further, using Lemma 6.8, the inverse estimate (6.7), and the estimate (6.5), we find, for $m=1,2$,

$$\begin{aligned}&\int_{\gamma_3}\left(\mathcal{E}_k^{\gamma_m}(\pi_k^y\circ\pi_k^z)\left(w_x-\hat{Q}_{k-1,k,k}w_x\right)\right)^2 dx\,dz\\ &\quad= \|M_k^{\gamma_m}\|_{0,I}^2\int_{-1}^{1}\left((\pi_k^y\circ\pi_k^z)\left(w_x-\hat{Q}_{k-1,k,k}w_x\right)\right)^2_{|_{y=-1}} dz\\ &\quad\le C\,k^{-1}\,k^2\int_{\gamma_m}\left((\pi_k^y\circ\pi_k^z)\left(w_x-\hat{Q}_{k-1,k,k}w_x\right)\right)^2 dy\,dz\\ &\quad\le C\,|w_x|_{1,\hat{K}}^2.\end{aligned}$$

Hence, we obtain

$$\|w_x - v_x\|^2_{0,\gamma_3} \le C\,|w_x|^2_{1,\hat{K}}.$$

The analogous bounds are obtained on γ_4, γ_5, and γ_6. This gives the desired result for the first component of $\mathbf{w} - \Pi_K \mathbf{w}$.

The proof is completed by observing that the same techniques give analogous bounds for the other components of $\mathbf{w} - \Pi_K \mathbf{w}$ and by a scaling argument. ∎

6.4. Proof of Proposition 6.1. Fix $q \in Q_h$. We first proceed as in [40, Lemma 6.3] (see also [38] for conforming mixed *hp*-FEM) and decompose q into

$$q = q_0 + \bar{q}, \tag{6.12}$$

where q_0 is the L^2-projection of q into the subspace of $L^2_0(\Omega)$ consisting of piecewise constant pressures.

Owing to the results in [25, 38] for conforming meshes and the results in [36, 41] (valid for two- and three-dimensional domains) for meshes with hanging nodes (see also [40]), there exists a piecewise quadratic velocity field $\mathbf{v}_0 \in \mathbf{V}_h \cap H^1_0(\Omega)^3$ such that

$$B_h(\mathbf{v}_0, q_0) = -\int_\Omega q_0 \nabla \cdot \mathbf{v}_0 \, d\mathbf{x} \ge \|q_0\|^2_0, \qquad \|\mathbf{v}_0\|_h = |\mathbf{v}_0|_1 \le C_0 \|q_0\|_0. \tag{6.13}$$

Further, for $K \in \mathcal{T}_h$, we set $\bar{q}_K = \bar{q}|_K$ and have, by construction, $\int_K \bar{q}_K \, d\mathbf{x} = 0$. Due to the continuous inf-sup condition [10, 25], there is a velocity field $\bar{\mathbf{w}}_K \in H^1_0(K)^3$ such that

$$-\int_K \bar{q}_K \nabla \cdot \bar{\mathbf{w}}_K \, d\mathbf{x} \ge \|\bar{q}_K\|^2_{0,K}, \qquad |\bar{\mathbf{w}}_K|_{1,K} \le C \|\bar{q}_K\|_{0,K}, \tag{6.14}$$

with a constant $C > 0$ solely depending on the shape-regularity of the mesh. Define $\bar{\mathbf{w}} \in H^1_0(\Omega)^3$ by $\bar{\mathbf{w}}|_K = \bar{\mathbf{w}}_K$ for all $K \in \mathcal{T}_h$, and let $\bar{\mathbf{v}} \in \mathbf{V}_h$ be given by

$$\bar{\mathbf{v}}|_K = \bar{\mathbf{v}}_K := \Pi_K \bar{\mathbf{w}}_K \in RT_{k_K - 1}(K), \qquad K \in \mathcal{T}_h,$$

for the Raviart–Thomas projector Π_K of degree $k_K - 1$ on K. Since $\bar{\mathbf{w}}_K \in H^1_0(K)^3$, we have

$$\bar{\mathbf{v}}_K \cdot \mathbf{n}_K = 0 \qquad \text{on } \partial K \tag{6.15}$$

due to the second conditions in (6.2) (valid for an affinely mapped element), and hence $[\![\bar{\mathbf{v}}]\!] = 0$ on $\mathcal{E}$. From the definition of B_h in (5.6), we thus have

$$B_h(\bar{\mathbf{v}}, \bar{q}) = -\int_\Omega \bar{q}\, \nabla_h \cdot \bar{\mathbf{v}} \, d\mathbf{x} = \sum_{K \in \mathcal{T}_h} \int_K \nabla \bar{q}_K \cdot \bar{\mathbf{v}}_K \, d\mathbf{x}.$$

Mapping $\bar{q}_K$ to $\hat{\bar{q}}_{\hat{K}}$ via the usual pullback operator and $\bar{\mathbf{v}}_K$ to $\hat{\bar{\mathbf{v}}}_{\hat{K}}$ via the Piola transformation, we obtain from [10, sect. 3.1]

$$\int_K \nabla \bar{q}_K \cdot \bar{\mathbf{v}}_K \, d\mathbf{x} = \int_{\hat{K}} \hat{\nabla} \hat{\bar{q}}_{\hat{K}} \cdot \hat{\bar{\mathbf{v}}}_{\hat{K}} \, d\hat{\mathbf{x}}.$$

We then note that, since $\hat{\bar{q}}_{\hat{K}} \in \mathbb{Q}_{k_K - 1}(\hat{K})$, we have

$$\hat{\nabla} \hat{\bar{q}}_{\hat{K}} \in \mathbb{Q}_{k_K - 2, k_K - 1, k_K - 1}(\hat{K}) \times \mathbb{Q}_{k_K - 1, k_K - 2, k_K - 1}(\hat{K}) \times \mathbb{Q}_{k_K - 1, k_K - 1, k_K - 2}(\hat{K}).$$

Using the orthogonality conditions in (6.2), we obtain

$$\int_{\hat{K}} \hat{\nabla}\hat{\bar{q}}_{\hat{K}} \cdot \hat{\bar{\mathbf{v}}}_{\hat{K}}\, d\hat{\mathbf{x}} = \int_{\hat{K}} \hat{\nabla}\hat{\bar{q}}_{\hat{K}} \cdot \hat{\bar{\mathbf{u}}}_{\hat{K}}\, d\hat{\mathbf{x}} = \int_K \nabla \bar{q}_K \cdot \bar{\mathbf{w}}_K\, d\mathbf{x}$$

and, therefore, from (6.14),

$$B_h(\bar{\mathbf{v}}, \bar{q}) = \sum_{K\in\mathcal{T}_h} \int_K \nabla \bar{q}_K \cdot \bar{\mathbf{w}}_K\, d\mathbf{x} = -\sum_{K\in\mathcal{T}_h} \int_K \bar{q}_K \nabla\cdot \bar{\mathbf{w}}_K\, d\mathbf{x} \geq \|\bar{q}\|_0^2. \tag{6.16}$$

Further, from the stability result in Lemma 6.9 and (6.14), we obtain

$$\sum_{K\in\mathcal{T}_h} |\bar{\mathbf{v}}_K|_{1,K}^2 \leq C \sum_{K\in\mathcal{T}_h} k_K^2 |\bar{\mathbf{w}}_K|_{1,K}^2 \leq C|\underline{k}|^2 \|\bar{q}\|_0^2. \tag{6.17}$$

Then, since $\underline{[\![\bar{\mathbf{w}}]\!]} = \underline{0}$ on $\mathcal{E}$, we have with Lemma 6.10 and (5.1)

$$\begin{aligned}\int_{\mathcal{E}} \sigma \underline{[\![\bar{\mathbf{v}}]\!]}^2 ds &= \int_{\mathcal{E}} \sigma \underline{[\![\bar{\mathbf{w}} - \bar{\mathbf{v}}]\!]}^2 ds \\ &\leq C \sum_{K\in\mathcal{T}_h} \frac{k_K^2}{h_K} \|\bar{\mathbf{w}}_K - \bar{\mathbf{v}}_K\|_{0,\partial K}^2 \leq C|\underline{k}|^2 \sum_{K\in\mathcal{T}_h} |\bar{\mathbf{w}}_K|_{1,K}^2 \leq C|\underline{k}|^2 \|\bar{q}\|_0^2.\end{aligned}$$

Combining this estimate with (6.16) and (6.17) yields

$$B_h(\bar{\mathbf{v}}, \bar{q}) \geq \|\bar{q}\|_0, \qquad \|\bar{\mathbf{v}}\|_h^2 \leq \bar{C}|\underline{k}|^2 \|\bar{q}\|_0. \tag{6.18}$$

Next, we define

$$\mathbf{v} = \mathbf{v}_0 + \delta \bar{\mathbf{v}}$$

for a parameter $\delta > 0$ still at our disposal. First, we note that from (5.6) and (6.15),

$$B_h(\bar{\mathbf{v}}, q_0) = -\sum_{K\in\mathcal{T}_h} q_0|_K \int_K \nabla\cdot \bar{\mathbf{v}}_K\, d\mathbf{x} = -\sum_{K\in\mathcal{T}_h} q_0|_K \int_{\partial K} \bar{\mathbf{v}}_K \cdot \mathbf{n}_K\, ds = 0$$

since q_0 is piecewise constant. Further, $\mathbf{v}_0 \in \mathbf{V}_h \cap H_0^1(\Omega)^3$ and, therefore, we obtain from (6.13) and the arithmetic-geometric mean inequality

$$|B_h(\mathbf{v}_0, \bar{q})| = \left| \int_\Omega \bar{q}\, \nabla\cdot \mathbf{v}_0\, d\mathbf{x} \right| \leq C\|q_0\|_0 \|\bar{q}\|_0 \leq \frac{C_1}{\varepsilon}\|q_0\|_0^2 + \varepsilon C_2 \|\bar{q}\|_0^2,$$

with another parameter $\varepsilon > 0$ to be properly chosen. Combining the above results with (6.13) and (6.18) gives

$$\begin{aligned}B_h(\mathbf{v}, q) &= B_h(\mathbf{v}_0, q_0) + B_h(\mathbf{v}_0, \bar{q}) + \delta B_h(\bar{\mathbf{v}}, \bar{q}) \\ &\geq \left(1 - \frac{C_1}{\varepsilon}\right) \|q_0\|_0^2 + (\delta - \varepsilon C_2)\|\bar{q}\|_0^2.\end{aligned}$$

It is then clear that we can choose δ and ε in such a way that

$$B_h(\mathbf{v}, q) \geq c_1 \|q\|_0^2, \tag{6.19}$$

with a constant c_1 independent of $\underline{h}$ and $\underline{k}$. Furthermore, from (6.13) and (6.18),

$$\|\mathbf{v}\|_h \leq |\mathbf{v}_0|_1 + \delta \|\bar{\mathbf{v}}\|_h \leq c_2 |\underline{k}| \|q\|_0, \tag{6.20}$$

with c_2 independent of $\underline{h}$ and $\underline{k}$. The assertion of Proposition 6.1 follows from (6.19) and (6.20).

6.5. Uniform approximation degrees and conforming meshes. For uniform approximation degrees $k_K = k$, $K \in \mathcal{T}_h$, and conforming meshes, the decomposition (6.12) is not necessary and we can establish the inf-sup condition directly via a Fortin property. In particular, this allows us to cover the case of $\mathbb{Q}_1 - \mathbb{Q}_0$ elements as well.

To do this, define the global interpolation operator Π by

$$\Pi \mathbf{w}_{|_K} = \Pi_K \mathbf{w}, \quad K \in \mathcal{T}_h,$$

where Π_K is the Raviart–Thomas projector of degree $k-1$ on K. We note that $\Pi\mathbf{w}$ belongs to $\mathbf{V}_h$ and, in case $\mathbf{w} \in H_0^1(\Omega)^3$, the normal component of $\Pi\mathbf{w}$ is continuous across the interelement boundaries and vanishes on $\partial\Omega$, i.e., $[\![\Pi\mathbf{w}]\!] = 0$ on $\mathcal{E}$. This last property is no longer true if the mesh has hanging nodes.

We have the following Fortin property.

LEMMA 6.11. *Assume that $\mathcal{T}_h$ is conforming and $k_K = k$, $K \in \mathcal{T}_h$. We have, for $\mathbf{w} \in H_0^1(\Omega)^3$ and $k \geq 1$,*

$$B_h(\Pi\mathbf{w}, q) = -\int_\Omega q \nabla \cdot \mathbf{w}\, d\mathbf{x}, \qquad q \in Q_h, \tag{6.21}$$

$$\|\Pi\mathbf{w}\|_h \leq Ck|\mathbf{w}|_1, \tag{6.22}$$

where $C > 0$ is independent of h and k.

Proof. We first note that, from (5.7), we have

$$\begin{aligned} B_h(\Pi\mathbf{w}, q) &= \sum_{K\in\mathcal{T}_h} \int_K \Pi\mathbf{w} \cdot \nabla q d\mathbf{x} - \int_{\mathcal{E}_\mathcal{I}} (q^+ - q^-)\, \frac{(\Pi\mathbf{w})^+ \cdot \mathbf{n}_{K^+} + (\Pi\mathbf{w})^- \cdot \mathbf{n}_{K^-}}{2}\, ds \\ &= \sum_{K\in\mathcal{T}_h} \int_K \Pi\mathbf{w} \cdot \nabla q d\mathbf{x} - \int_{\mathcal{E}_\mathcal{I}} (q^+ - q^-)\, \Pi\mathbf{w} \cdot \mathbf{n}_{K^+}\, ds, \end{aligned}$$

where we have used obvious notation to express the jumps and mean values. Again using the orthogonality conditions in (6.2), valid for an affinely mapped element K, we find

$$\begin{aligned} B_h(\Pi\mathbf{w}, q) &= \sum_{K\in\mathcal{T}_h} \int_K \mathbf{w} \cdot \nabla q d\mathbf{x} - \int_{\mathcal{E}_\mathcal{I}} (q^+ - q^-)\, \mathbf{w} \cdot \mathbf{n}_{K^+}\, ds \\ &= -\int_\Omega q \nabla \cdot \mathbf{w}\, d\mathbf{x}. \end{aligned}$$

The stability estimate in (6.22) follows from Lemma 6.9 and Lemma 6.10 as in the proof of Proposition 6.1. ∎

We note that the previous lemma is not true for irregular meshes. Combining Lemma 6.11 and the inf-sup condition (2.3) of the continuous problem, we find the following stability result.

THEOREM 6.12. *Assume that $\mathcal{T}_h$ is conforming and $k_K = k$, $K \in \mathcal{T}_h$. There exists a constant $c > 0$, independent of h and k, such that for $k \geq 1$*

$$\inf_{0\neq q\in Q_h} \sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{V}_h} \frac{B_h(\mathbf{v}, q)}{\|\mathbf{v}\|_h \|q\|_0} \geq \gamma_h \geq ck^{-1}. \tag{6.23}$$

We emphasize that, in particular, this result holds true for $k = 1$, thus covering $\mathbb{Q}_1 - \mathbb{Q}_0$ elements. We also remark that a similar nonconforming Stokes element,

the so-called $\widetilde{\mathbb{Q}}_1 - \mathbb{Q}_0$ element, has been proposed and studied in [34, 8]. However, this element can be viewed as a natural quadrilateral analogue of the well known Crouzeix–Raviart element whereas the $\mathbb{Q}_1 - \mathbb{Q}_0$ element here is based on completely discontinuous finite element spaces.

7. Continuity and coercivity. In this section, we establish the continuity and coercivity of the forms $A_h(\cdot,\cdot)$ and $B_h(\cdot,\cdot)$ with respect to the norm $\|\cdot\|_h$ in (5.8)–(5.9).

7.1. Stability of the lifting operators. We start by investigating the stability properties of the lifting operators. To this end, we need the following lemma concerning traces of polynomials, where we denote by $\mathbb{Q}_k(\gamma_m)$ the polynomials of degree at most k in each variable on the face γ_m.

LEMMA 7.1. *Let $K \in \mathcal{T}_h$ and γ_m be a face of ∂K. Then we have*

$$\|\varphi\|_{0,\gamma_m} \le C h_K^{-\frac{1}{2}} k \|\varphi\|_{0,K} \qquad \forall \varphi \in \mathbb{Q}_k(K), \tag{7.1}$$

with a constant $C > 0$ just depending on the shape-regularity of the mesh.

Conversely, for $\varphi \in \mathbb{Q}_k(\gamma_m)$ there is a polynomial extension $E(\varphi) \in \mathbb{Q}_k(K)$ with $E(\varphi)|_{\gamma_m} = \varphi$ and

$$\|E(\varphi)\|_{0,K} \le C h_K^{\frac{1}{2}} k^{-1} \|\varphi\|_{0,\gamma_m}, \tag{7.2}$$

with a constant $C > 0$ just depending on the shape-regularity of the mesh.

Proof. The first assertion follows from standard inverse inequalities; see, e.g., [37, Theorem 4.76].

We prove the second assertion only in three dimensions (the two-dimensional case is completely analogous). To this end, we consider first the reference cube $\hat{K} = (-1,1)^3$ and may assume that the face γ_m is given by $x = 1$. Fix $\varphi \in \mathbb{Q}_k(\gamma_m)$. Moreover, we consider the case where k is even and set

$$E(\varphi)(x,y,z) = \left(\frac{2}{k} \sum_{j=\frac{k}{2}+1}^{k} L_j(x) \right) \varphi(y,z),$$

where L_j denotes the Legendre polynomial of degree j on $(-1,1)$. Since $L_j(1) = 1$, we have

$$E(\varphi)|_{\gamma_m} = E(\varphi)(1,y,z) = \frac{2}{k}\,\frac{k}{2}\,\varphi(y,z) = \varphi(y,z).$$

Further,

$$\|E(\varphi)\|_{0,\hat{K}}^2 = \|\varphi\|_{0,\gamma_m}^2 \,\frac{4}{k^2} \sum_{j=\frac{k}{2}+1}^{k} \frac{2}{2j+1}.$$

We have

$$\begin{aligned} \sum_{j=\frac{k}{2}+1}^{k} \frac{2}{2j+1} &= \sum_{j=\frac{k}{2}+1}^{k} \frac{1}{(j+1)-\frac{1}{2}} \le \int_{\frac{k}{2}+1}^{k+1} \frac{1}{t-\frac{1}{2}}\,dt \\ &= \log\left(k+\frac{1}{2}\right) - \log\left(\frac{k}{2}+\frac{1}{2}\right) = \log\left(\frac{2k+1}{k+1}\right). \end{aligned}$$

The bound $\log(\frac{2k+1}{k+1}) \le C$, independent of k, proves the assertion for k even. If k is odd, the extension $E(\cdot)$ can be constructed similarly. This proves the assertion on the reference cube; the general case follows from a standard scaling argument. ∎

We are now ready to prove the following stability result for the lifting $\underline{\mathcal{L}}_e$.

LEMMA 7.2. *For a face $e \subset \mathcal{E}$, we have*

$$\|\underline{\mathcal{L}}_e(\mathbf{v})\|_0^2 \ge C_1 \int_e \mathtt{k}^2\mathtt{h}^{-1}|\underline{[\![\mathbf{v}]\!]}|^2\,ds \qquad \forall \mathbf{v} \in \mathbf{V}_h,$$
$$\|\underline{\mathcal{L}}_e(\mathbf{v})\|_0^2 \le C_2 \int_e \mathtt{k}^2\mathtt{h}^{-1}|\underline{[\![\mathbf{v}]\!]}|^2\,ds \qquad \forall \mathbf{v} \in \mathbf{V}(h),$$

with constants $C_1 > 0$ and $C_2 > 0$ depending on the shape-regularity of the mesh. If e contains a hanging node, C_1 also depends on κ in (5.1).

Proof. To prove the first estimate, fix $\mathbf{v} \in \mathbf{V}_h$ and let K be the element such that e is an entire face of ∂K. By Lemma 7.1, we can find a polynomial $\underline{\tau} \in \mathbb{Q}_{k_K}(K)^{d\times d}$ such that $\underline{\tau}|_e = \underline{[\![\mathbf{v}]\!]}$ and such that

$$\|\underline{\tau}\|_{0,K} \le C h_K^{\frac{1}{2}} k_K^{-1} \|\underline{[\![\mathbf{v}]\!]}\|_{0,e}.$$

Extending $\underline{\tau}$ by zero, we obtain a function also denoted by $\underline{\tau}$ in the finite element space $\underline{\Sigma}_h$. By definition of $\underline{\mathcal{L}}_e$ and construction of $\underline{\tau}$, we have

$$\frac{1}{2}\|\underline{[\![\mathbf{v}]\!]}\|_{0,e}^2 = \int_e \underline{[\![\mathbf{v}]\!]} : \{\!\{\underline{\tau}\}\!\}\,ds \le \int_K |\underline{\mathcal{L}}_e(\mathbf{v}) : \underline{\tau}|\,d\mathbf{x} \le C h_K^{\frac{1}{2}} k_K^{-1} \|\underline{\mathcal{L}}_e(\mathbf{v})\|_0 \|\underline{[\![\mathbf{v}]\!]}\|_{0,e}.$$

If e is also an entire face of a possible neighboring element K', we combine the above bound with the one for K' and obtain the desired result. If e is not an entire face of a neighboring element, we invoke (5.1) and obtain the bound.

Conversely, for $\mathbf{v} \in \mathbf{V}(h)$, we have

$$\begin{aligned}
\|\underline{\mathcal{L}}_e(\mathbf{v})\|_0 &= \sup_{\underline{\tau}\in\underline{\Sigma}_h} \frac{\int_\Omega \underline{\mathcal{L}}_e(\mathbf{v}) : \underline{\tau}\,d\mathbf{x}}{\|\underline{\tau}\|_0} = \sup_{\underline{\tau}\in\underline{\Sigma}_h} \frac{\int_e \underline{[\![\mathbf{v}]\!]} : \{\!\{\underline{\tau}\}\!\}\,ds}{\|\underline{\tau}\|_0} \\
&\le \sup_{\underline{\tau}\in\underline{\Sigma}_h} \frac{\big(\int_e \mathtt{k}^2\mathtt{h}^{-1}|\underline{[\![\mathbf{v}]\!]}|^2\,ds\big)^{\frac{1}{2}} \big(C\sum_{K\in\mathcal{T}_h} k_K^{-2} h_K \|\underline{\tau}\|_{0,\partial K}^2\big)^{\frac{1}{2}}}{\|\underline{\tau}\|_0} \\
&\le \sup_{\underline{\tau}\in\underline{\Sigma}_h} \frac{\big(\int_e \mathtt{k}^2\mathtt{h}^{-1}|\underline{[\![\mathbf{v}]\!]}|^2\,ds\big)^{\frac{1}{2}} \big(C\sum_{K\in\mathcal{T}_h} \|\underline{\tau}\|_{0,K}^2\big)^{\frac{1}{2}}}{\|\underline{\tau}\|_0} \\
&\le C\left(\int_e \mathtt{k}^2\mathtt{h}^{-1}|\underline{[\![\mathbf{v}]\!]}|^2\,ds\right)^{\frac{1}{2}},
\end{aligned}$$

where we used the definition of $\underline{\mathcal{L}}_e$, the Cauchy–Schwarz inequality, and the trace estimate (7.1) from Lemma 7.1. ∎

REMARK 7.3. *Due to* (5.3)*, we also have*

$$\|\underline{\mathcal{G}}_e\|_0^2 \le C \int_e \mathtt{k}^2\mathtt{h}^{-1}|\mathbf{g}|^2\,ds$$

for any boundary face $e \subset \mathcal{E}_\mathcal{D}$.

In the same manner, we obtain the following stability estimates for $\underline{\mathcal{L}}$, $\mathcal{M}$, and $\underline{\mathcal{G}}$.

LEMMA 7.4. *We have the stability estimates*

$$\|\mathcal{M}(\mathbf{v})\|_0^2 \le C \int_{\mathcal{E}} \mathtt{k}^2\mathtt{h}^{-1} |\underline{[\![\mathbf{v}]\!]}|^2 \, ds, \qquad \mathbf{v} \in \mathbf{V}(h),$$

$$\|\underline{\mathcal{L}}(\mathbf{v})\|_0^2 \le C \int_{\mathcal{E}} \mathtt{k}^2\mathtt{h}^{-1} |\underline{[\![\mathbf{v}]\!]}|^2 \, ds, \qquad \mathbf{v} \in \mathbf{V}(h),$$

as well as

$$\|\underline{\mathcal{G}}\|_0^2 \le C \int_{\mathcal{E}_\mathcal{D}} \mathtt{k}^2\mathtt{h}^{-1} |\mathbf{g}|^2 \, ds,$$

with constants $C > 0$ solely depending on the shape-regularity of the mesh.

7.2. Continuity. The continuity conditions of $A_h(\cdot,\cdot)$ and $B_h(\cdot,\cdot)$ with respect to the the discrete norm $\|\cdot\|_h$ in (5.8) are established in the following lemma.

LEMMA 7.5. *Let σ be given as in* (5.9) *with $\sigma_0 > 0$. Then, we have the following.*

1. *All the forms A_h considered in section* 5.4 *are continuous,*

$$|A_h(\mathbf{v},\mathbf{w})| \le \nu\bar{\alpha}_1 \|\mathbf{v}\|_h \|\mathbf{w}\|_h, \qquad \mathbf{v},\mathbf{w} \in \mathbf{V}(h),$$

   with a constant $\bar{\alpha}_1 > 0$ independent of $\underline{h}$ and $\underline{k}$. Hence, condition (3.2) *is satisfied with $\alpha_1 = \nu\bar{\alpha}_1$.*
2. *The form B_h is continuous,*

$$|B_h(\mathbf{v},q)| \le \alpha_2 \|\mathbf{v}\|_h \|q\|_0, \qquad (\mathbf{v},q) \in \mathbf{V}(h) \times Q,$$

   with a constant $\alpha_2 > 0$ independent of $\underline{h}$ and $\underline{k}$.

Proof. This follows immediately from Lemma 7.2, Lemma 7.4, and Cauchy–Schwarz inequalities. ∎

7.3. Coercivity of A_h. The coercivity condition in (3.5) of the different forms A_h is established in the following lemma.

LEMMA 7.6. *Let σ be given as in* (5.9) *with $\sigma_0 > 0$. Then, we have the following.*

1. *There is a constant $\sigma_{min} > 0$ (independent of $\underline{h}$ and $\underline{k}$) such that for $\sigma_0 \ge \sigma_{min}$ the symmetric interior penalty form A_h in* (5.11) *is coercive,*

$$A_h(\mathbf{v},\mathbf{v}) \ge \nu\bar{\beta} \|\mathbf{v}\|_h^2, \qquad \mathbf{v} \in \mathbf{V}_h,$$

   with a constant $\bar{\beta} > 0$ independent of $\underline{h}$ and $\underline{k}$. Hence, condition (3.5) *is satisfied with $\beta = \nu\bar{\beta}$.*
2. *The nonsymmetric interior penalty form A_h in* (5.12) *is coercive on $\mathbf{V}(h)$ for any $\sigma_0 > 0$, with coercivity constant $\beta = \nu$.*
3. *The LDG form A_h in* (5.13) *is coercive on $\mathbf{V}_h$ for any $\sigma_0 > 0$, with a coercivity constant $\beta = \nu\bar{\beta}$, where $\bar{\beta} > 0$ is independent of $\underline{h}$ and $\underline{k}$.*
4. *There is a constant $\eta_{min} > 0$ (independent of $\underline{h}$ and $\underline{k}$) such that for $\eta \ge \eta_{min}$ the Bassi–Rebay form A_h in* (5.15) *is coercive on $\mathbf{V}_h$, with a coercivity constant $\beta = \nu\bar{\beta}$, where $\bar{\beta} > 0$ is independent of $\underline{h}$ and $\underline{k}$.*
5. *The Bassi–Rebay form A_h in* (5.16) *is coercive on $\mathbf{V}_h$ for any $\eta > 0$, with a coercivity constant $\beta = \nu\bar{\beta}$, where $\bar{\beta} > 0$ is independent of $\underline{h}$ and $\underline{k}$.*

Proof. These coercivity properties are obtained from Lemma 7.2, Lemma 7.4, and the arithmetic-geometric mean inequality $2ab \le \varepsilon a^2 + \varepsilon^{-1} b^2$ for all $\varepsilon > 0$; see [2]. ∎

REMARK 7.7. *We chose to express the continuity and coercivity properties of the Bassi–Rebay methods in terms of the discrete norm* $\|\cdot\|_h$ *in* (5.8)–(5.9) *since this norm is explicit in the mesh-sizes and the approximation degrees. Instead, it is also possible to work with*

$$\|\mathbf{v}\|_h^2 = \sum_{K\in\mathcal{T}_h} |\mathbf{v}|_{1,K}^2 + \sum_{e\subset\mathcal{E}} \int_\Omega \eta |\underline{\mathcal{L}}_e(\mathbf{v})|^2 \, d\mathbf{x}.$$

8. The residual. In this section, we study the residual $R_h(\mathbf{u}, p; \mathbf{v})$ in (3.8) for our DG methods and show that it is optimally convergent.

PROPOSITION 8.1. *Let the exact solution* $(\mathbf{u}, p)$ *of the Stokes system* (2.1) *be in* $H^{s_K+1}(K)^d \times H^{s_K}(K)$ *for all* $K \in \mathcal{T}_h$ *and* $s_K \geq 1$. *Let* $\underline{Q}$ *and* Q *be the* L^2*-projections onto* $\underline{\Sigma}_h$ *and* Q_h, *respectively. Then the residual in* $R_h(\mathbf{u}, p; \mathbf{v})$ *in* (3.8) *is given by*

$$R_h(\mathbf{u}, p; \mathbf{v}) = \nu \int_{\mathcal{E}} \{\!\{\nabla \mathbf{u} - \underline{Q}(\nabla \mathbf{u})\}\!\} : \underline{[\![\mathbf{v}]\!]} \, ds - \int_{\mathcal{E}} \{\!\{p - Qp\}\!\} [\![\mathbf{v}]\!] \, ds \qquad \forall \mathbf{v} \in \mathbf{V}_h,$$

for all forms discussed in section 5.4.

Furthermore, we have that $\mathcal{R}_h(\mathbf{u}, p)$ *in* (3.9) *can be estimated by*

$$\mathcal{R}_h(\mathbf{u}, p)^2 \leq C \sum_{K\in\mathcal{T}_h} \frac{h^{2\min(s_K, k_K)}}{k_K^{2s_K+1}} \big[\nu \|\mathbf{u}\|_{s_K+1,K}^2 + \nu^{-1} \|p\|_{s_K,K}^2\big],$$

with a constant $C > 0$ *independent of* $\underline{h}$, $\underline{k}$, *and* ν.

Proof. By (5.3), we have $\underline{\mathcal{L}}(\mathbf{u}) = \underline{\mathcal{G}}$ and obtain for all forms

$$R_h(\mathbf{u}, p; \mathbf{v}) = \nu \int_\Omega [\nabla \mathbf{u} : \nabla_h \mathbf{v} - \nabla \mathbf{u} : \underline{\mathcal{L}}(\mathbf{v})] \, d\mathbf{x} - \int_\Omega p[\nabla \cdot \mathbf{v} - \mathcal{M}(\mathbf{v})] \, d\mathbf{x} - \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}.$$

Note that

$$\int_\Omega \nabla \mathbf{u} : \underline{\mathcal{L}}(\mathbf{v}) \, d\mathbf{x} = \int_\Omega \underline{Q}(\nabla \mathbf{u}) : \underline{\mathcal{L}}(\mathbf{v}) \, d\mathbf{x} = \int_{\mathcal{E}} \{\!\{\underline{Q}(\nabla \mathbf{u})\}\!\} : \underline{[\![\mathbf{v}]\!]} \, ds$$

and

$$\int_\Omega p\mathcal{M}(\mathbf{v}) \, d\mathbf{x} = \int_\Omega Qp\mathcal{M}(\mathbf{v}) \, d\mathbf{x} = \int_\Omega \{\!\{Qp\}\!\} [\![\mathbf{v}]\!] \, ds.$$

If now the exact solution belongs to $H^2(K)^d \times H^1(K)$ for all $K \in \mathcal{T}_h$, we obtain by integration by parts and elementary manipulations

$$\begin{aligned} R_h(\mathbf{u}, p; \mathbf{v}) = &\int_\Omega [-\nu \Delta \mathbf{u} + \nabla p - \mathbf{f}] \cdot \mathbf{v} \, d\mathbf{x} \\ &+ \nu \int_{\mathcal{E}} \{\!\{\nabla \mathbf{u} - \underline{Q}(\nabla \mathbf{u})\}\!\} : \underline{[\![\mathbf{v}]\!]} \, ds - \int_{\mathcal{E}} \{\!\{p - Qp\}\!\} [\![\mathbf{v}]\!] \, ds. \end{aligned}$$

Here, we also used that $[\![\nu \nabla \mathbf{u} - p\underline{I}]\!] = \mathbf{0}$ on $\mathcal{E}_\mathcal{I}$. From the Stokes equations in (2.1) we obtain the first assertion.

From (5.1) and since $|[\![\mathbf{v}]\!]|^2 \leq C |\underline{[\![\mathbf{v}]\!]}|^2$, the Cauchy–Schwarz equation yields

$$\begin{aligned} &R_h(\mathbf{u}, p; \mathbf{v}) \\ &\leq C \|\mathbf{v}\|_h \left(\nu \sum_{K\in\mathcal{T}_h} \frac{h_K}{k_K^2} \|\nabla \mathbf{u} - \underline{Q}(\nabla \mathbf{u})\|_{0,\partial K}^2 + \nu^{-1} \sum_{K\in\mathcal{T}_h} \frac{h_K}{k_K^2} \|p - Q(p)\|_{0,\partial K}^2 \right)^{\frac{1}{2}}, \end{aligned}$$

from where the error estimate follows with the *hp*-approximation properties of the L^2-projection in [27]. ∎

9. Error estimates. In this section, we make the abstract error estimates in section 4 explicit for our DG methods.

9.1. The main result. First, we consider general meshes with hanging nodes. We have the following result.

THEOREM 9.1. *Let the exact solution* $(\mathbf{u}, p)$ *of the Stokes system* (2.1) *be in* $H^{s_K+1}(K)^d \times H^{s_K}(K)$ *for all* $K \in \mathcal{T}_h$ *and* $s_K \geq 1$. *Then we have*

$$\|\mathbf{u}-\mathbf{u}_h\|_h^2 \leq C \sum_{K\in\mathcal{T}_h} \left[\gamma_h^{-2} \frac{h_K^{2\min(s_K,k_K)}}{k_K^{2s_K-1}} \|\mathbf{u}\|_{s_K+1,K}^2 + \frac{h_K^{2\min(s_K,k_K)}}{k_K^{2s_K}} \|p\|_{s_K,K}^2 \right],$$

$$\|p-p_h\|_0^2 \leq C \sum_{K\in\mathcal{T}_h} \left[\gamma_h^{-4} \frac{h_K^{2\min(s_K,k_K)}}{k_K^{2s_K-1}} \|\mathbf{u}\|_{s_K+1,K}^2 + \gamma_h^{-2} \frac{h_K^{2\min(s_K,k_K)}}{k_K^{2s_K}} \|p\|_{s_K,K}^2 \right],$$

with $C > 0$ *independent of* $\underline{h}$ *and* $\underline{k}$.

Proof. This follows from the choice of the stabilization parameter σ in (5.9), Proposition 4.1, Proposition 4.3, Proposition 8.1, and standard approximation properties of the finite element spaces; see, e.g., [3, Lemma 4.5] or [37]. In particular, we choose $\mathbf{v}$ in Proposition 4.1 and q in Proposition 4.3 as the locally constructed interpolants of $\mathbf{u}$ and p, respectively, given in [3, Lemma 4.5]. ∎

REMARK 9.2. *The above hp-version estimates are optimal in the mesh-size* $\underline{h}$, *and slightly suboptimal in* $\underline{k}$ *(half a power is lost), up to the inf-sup constant* γ_h *(which depends on the polynomial degree* $\underline{k}$*). In the mesh-size* $\underline{h}$, *the same optimal bounds have been obtained in* [26] *for the IP method on simplicial and conforming meshes and for* $\mathcal{P}_k - \mathcal{P}_{k-1}$ *elements, with* $\mathcal{P}_k$ *denoting polynomials of total degree at most* k. *We further note that, in the hp-version context, the same result was recently obtained in* [40] *for the nonsymmetric interior penalty method, with different techniques.*

REMARK 9.3. *The loss of half a power of* k *is typical of DG methods for second-order problems. Indeed, in the case of elliptic diffusion problems in two- or three-dimensional domains, no better p-bounds can be found in the DG literature on general unstructured grids (see, e.g., the hp-version analyses in* [27, 32, 35, 31]*). Improved p-bounds have been obtained in* [14] *for one-dimensional convection-diffusion problems, and recently in* [23] *for two-dimensional reaction-diffusion problems on affine quadrilateral grids containing hanging nodes and for solutions that belong to augmented Sobolev spaces. The latter results can be immediately carried over to the Stokes setting considered here.*

REMARK 9.4. *Combining the above bound with the inf-sup constant* γ_h *in Theorem* 6.2 *results in a loss of* $k^{3/2}$ *in the approximation of the velocity and in a loss of* $k^{5/2}$ *for the approximation of the pressure.*

9.2. Uniform approximation degrees and conforming meshes. In this section, we specialize the result of Theorem 9.1 to the case of uniform approximation orders, $k_K = k$, and conforming meshes with no hanging nodes. We also assume that the Dirichlet boundary datum $\mathbf{g}$ is piecewise polynomial; more precisely, we assume that there is a finite element function $\mathbf{G}_h \in \mathbf{V}_h$ such that $\mathbf{G}_h|_{\partial\Omega} = \mathbf{g}$.

In this particular situation, as in the analysis of [31] for the LDG method for pure diffusion problems, we can choose $\mathbf{v}$ in Proposition 4.1 as an optimal *hp*-approximant for the velocity which is continuous in the whole domain Ω according to [3, Theorem 4.6]. The discrete pressure q in Proposition 4.3 can be chosen as before. Since

the residual $\mathcal{R}_h(\mathbf{u}, p)$ is optimally convergent, we obtain the following result.

THEOREM 9.5. *Let the exact solution* $(\mathbf{u}, p)$ *of the Stokes system* (2.1) *be in* $H^{s+1}(\Omega)^d \times H^s(\Omega)$ *for* $s \geq 1$. *Then we have*

$$\|\mathbf{u} - \mathbf{u}_h\|_h \leq C \frac{h^{\min(s,k)}}{k^s} \Big[\gamma_h^{-1} \|\mathbf{u}\|_{s+1} + \|p\|_s\Big],$$
$$\|p - p_h\|_0 \leq C \frac{h^{\min(s,k)}}{k^s} \Big[\gamma_h^{-2} \|\mathbf{u}\|_{s+1} + \gamma_h^{-1} \|p\|_s\Big],$$

with $C > 0$ *independent of* $\underline{h}$ *and* $\underline{p}$.

This estimate is *optimal* in h and k, up to the inf-sup constant (which is independent of $\underline{h}$). With Theorem 6.2, we obtain exactly the same result as Stenberg and Suri in [38] for conforming mixed *hp*-FEM in three dimensions, but with an optimal gap of one order in the finite element spaces for the velocity and the pressure.

REMARK 9.6. *The estimate in Theorem* 9.5 *also holds on meshes with certain kinds of hanging nodes provided that a conforming and optimal hp-approximant can be constructed. In two dimensions, results in this direction can be found in, e.g.,* [37].

REFERENCES

[1] M. AINSWORTH AND K. PINCHEDEZ, *hp-approximation theory for BDFM and RT finite elements on quadrilaterals*, SIAM J. Numer. Anal., 40 (2002), pp. 2047–2068.

[2] D.N. ARNOLD, F. BREZZI, B. COCKBURN, AND L.D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] I. BABUŠKA AND M. SURI, *The hp-version of the finite element method with quasiuniform meshes*, RAIRO Anal. Numér., 21 (1987), pp. 199–238.

[4] G.A. BAKER, W.N. JUREIDINI, AND O.A. KARAKASHIAN, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.

[5] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[6] F. BASSI, S. REBAY, G. MARIOTTI, S. PEDINOTTI, AND M. SAVINI, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the Second European Conference on Turbomachinery–Fluid Dynamics and Thermodynamics (Antwerpen, Belgium), R. Decuypere and G. Dibelius, eds., 1997, pp. 99–108.

[7] C.E. BAUMANN AND J.T. ODEN, *A discontinuous hp-finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[8] R. BECKER AND R. RANNACHER, *Finite element solution of the incompressible Navier-Stokes equations on anisotropically refined meshes*, in Proceedings of the 10th GAMM Seminar, Notes Numer. Fluid Dynamics, Vieweg, Braunschweig, Germany, 1995, pp. 52–62.

[9] C. BERNARDI AND Y. MADAY, *Spectral methods*, in Handbook of Numerical Analysis 5, North–Holland, Amsterdam, 1997, pp. 209–485.

[10] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer–Verlag, New York, 1991.

[11] F. BREZZI, M. MANZINI, D. MARINI, P. PIETRA, AND A. RUSSO, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 365–378.

[12] C. CANUTO AND A. QUARTERONI, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., 38 (1982), pp. 67–86.

[13] P. CASTILLO, B. COCKBURN, I. PERUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[14] P. CASTILLO, B. COCKBURN, D. SCHÖTZAU, AND C. SCHWAB, *Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection–diffusion problems*, Math. Comp., 71 (2002), pp. 455–478.

[15] B. COCKBURN, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, T. Barth and H. Deconink, eds., Springer–

Verlag, Berlin, 1999, pp. 69–224.
[16] B. Cockburn and C. Dawson, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions*, in The Proceedings of the 10th Conference on the Mathematics of Finite Elements and Applications: MAFELAP X, J. Whiteman, ed., Elsevier, New York, 2000, pp. 225–238.
[17] B. Cockburn, G. Kanschat, D. Schötzau, and C. Schwab, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.
[18] B. Cockburn, G.E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods. Theory, Computation and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer–Verlag, Berlin, 2000.
[19] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time–dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
[20] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection–dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
[21] L. Franca and S. L. Frey, *Stabilized finite element methods*, II. The incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.
[22] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.
[23] E.H. Georgoulis and E. Süli, *hp-DGFEM on Shape–Irregular Meshes: Reaction–Diffusion*, Technical report NA 01–09, Oxford University Computing Laboratory, Oxford University, Oxford, 2001.
[24] P. Gervasio and F. Saleri, *Stabilized spectral element approximation for the Navier-Stokes equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 115–141.
[25] V. Girault and P.A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer–Verlag, New York, 1986.
[26] P. Hansbo and M.G. Larson, *Discontinuous finite element methods for incompressible and nearly incompressible elasticity by use of Nitsche's method*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1895–1908.
[27] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
[28] O.A. Karakashian and W.N. Jureidini, *A nonconforming finite element method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 93–120.
[29] J.T. Oden, I. Babuška, and C.E. Baumann, *A discontinuous hp-finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.
[30] I. Perugia and D. Schötzau, *The hp-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell's equations*, Math. Comp., to appear.
[31] I. Perugia and D. Schötzau, *An hp-analysis of the local discontinuous Galerkin method for diffusion problems*, J. Sci. Comput., 17 (2002), pp. 561–571.
[32] S. Prudhomme, F. Pascal, J.T. Oden, and A. Romkes, *Review of A Priori Error Estimation for Discontinuous Galerkin Methods*, Technical report 2000-27, TICAM, University of Texas at Austin, Austin, Texas, 2000.
[33] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer–Verlag, Berlin, 1994.
[34] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.
[35] B. Rivière, M.F. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems, Part* I, Computational Geosciences, 3 (1999), pp. 337–360.
[36] D. Schötzau, C. Schwab, and R. Stenberg, *Mixed hp-FEM on anisotropic meshes,* II. *Hanging nodes and tensor products of boundary layer meshes*, Numer. Math., 83 (1999), pp. 667–697.
[37] C. Schwab, *p- and hp-FEM – Theory and Application to Solid and Fluid Mechanics*, Oxford University Press, Oxford, 1998.
[38] R. Stenberg and M. Suri, *Mixed hp-finite element methods for problems in elasticity and Stokes flow*, Numer. Math., 72 (1996), pp. 367–389.
[39] R. Témam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North–Holland, Amsterdam, 1979.
[40] A. Toselli, *hp-discontinuous Galerkin approximations for the Stokes problem*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1565–1616.
[41] A. Toselli and C. Schwab, *Mixed hp-Finite Element Approximations on Geometric Edge and Boundary Layer Meshes in Three Dimensions*, Technical report 2001-02, Seminar for Applied Mathematics, ETHZ, Zürich, Switzerland, 2001; Numer. Math., to appear.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2195–2218

COUPLING FLUID FLOW WITH POROUS MEDIA FLOW*

WILLIAM J. LAYTON†, FRIEDHELM SCHIEWECK‡, AND IVAN YOTOV†

Abstract. The transport of substances back and forth between surface water and groundwater is a very serious problem. We study herein the mathematical model of this setting consisting of the Stokes equations in the fluid region coupled with the Darcy equations in the porous medium, coupled across the interface by the Beavers–Joseph–Saffman conditions. We prove existence of weak solutions and give a complete analysis of a finite element scheme which allows a simulation of the coupled problem to be uncoupled into steps involving porous media and fluid flow subproblems. This is important because there are many "legacy" codes available which have been optimized for uncoupled porous media and fluid flow.

Key words. coupled porous media and fluid flow, Stokes and Darcy equations, Beavers–Joseph–Saffman condition, weak solutions, finite element scheme, error estimates

AMS subject classifications. 35Q35, 65N30, 65N15, 76D07, 76S05

PII. S0036142901392766

1. Introduction and the model.

There are many serious problems currently facing the world in which the coupling between groundwater and surface water is important. These include questions such as predicting how pollution discharged into streams, lakes, and rivers makes its way into the water supply. This coupling is also important in technological applications involving filtration.

The aim of our research is to begin the study of the following problem: an incompressible fluid in a region Ω_1 can flow both ways across an interface Γ_I into a domain Ω_2 which is a porous medium saturated with the same fluid. The mathematical theory and numerical analysis of each subproblem is well developed, and reliable codes are available. Nevertheless, the mathematical theory of the coupled problem seems to be not completely understood. The model of this situation which is most accessible to large scale computations consists of the Navier–Stokes equations (or Stokes equations) in the fluid region coupled across an interface with the Darcy equations for the filtration velocity in the porous medium. This leads to mathematical difficulties arising from the coupled system of equations of different orders in different regions. See Jäger and Mikelić [16], Payne and Straughan [22] for the beginning of analytical studies of this problem. (For the Brinkman model of porous media flow this difficulty does not occur; see Jäger and Mikelić [17], Angot [1].) The second issue concerns the correct transmission conditions on the interface. The Beavers–Joseph–Saffman interface conditions [3, 25] are now well established. The third difficulty is technical: where the interface meets the other boundaries, there are incompatibilities between the imposed boundary conditions.

*Received by the editors July 19, 2001; accepted for publication (in revised form) June 26, 2002; published electronically January 7, 2003.
http://www.siam.org/journals/sinum/40-6/39276.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (wjl@pitt.edu, http://www.math.pitt.edu/∼wjl; yotov@math.pitt.edu). The research of the first author was partially supported by NSF grants DMS9972622, INT9814115, and INT9805563. The research of the third author was partially supported by the DOE grant DE-FG03-99ER25371, the NSF grants DMS 9873326 and DMS 0107389, the University of Pittsburgh CRDF grant, and the Sloan Foundation.

‡Institute for Analysis and Numerics, University of Magdeburg, Postfach 4120, D-39016, Magdeburg, Germany (friedhelm.schieweck@mathematik.uni-magdeburg.de, http://www-ian.math.uni-magdeburg.de/home/schieweck).

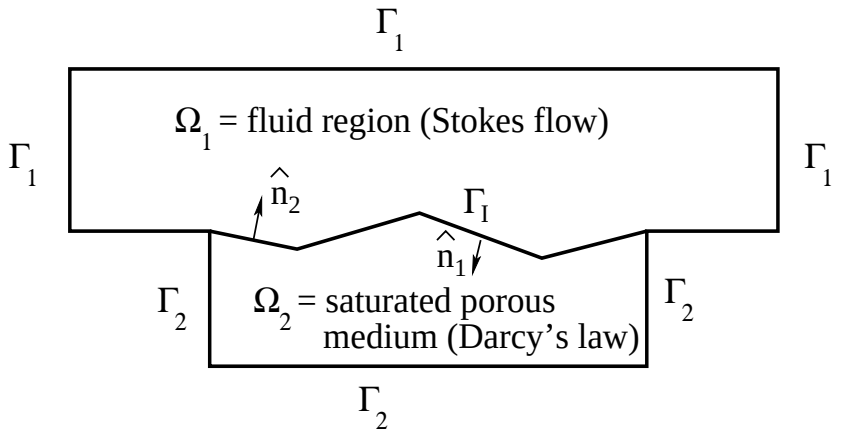


FIG. 1. *The model problem.*

One goal of this report is to find a variational formulation (section 2) for which weak solutions can be guaranteed to exist (section 3) and which can be used as a basis for a domain decomposition strategy for its approximate solution. The main goal is then to develop a finite element procedure with mathematical support (section 4). The method we study imposes the interface conditions using Lagrange multipliers. Thus, it can be used in a heterogeneous domain decomposition procedure in which each subproblem is alternately or simultaneously solved with codes (possibly "legacy" codes) developed and optimized for the physics of fluid motion and of porous media flow. In section 4 we give a complete analysis of this convergent finite element procedure. Because of the importance of the coupled problem, there are many computations of coupled surface water-groundwater flows in the applied literature, using various ad hoc interface decoupling strategies. See, for example, Salinger, Aris, and Derby [26], Gartling, Hickox, and Givler [14], and Prasad [23] for recent and interesting computational studies of the coupled problem.

The coupling strategy via Lagrange multipliers we consider herein has been proven in other applications and we are working towards practical tests of our ideas.

1.1. The model. The model we consider consists of Stokes flow in the fluid region Ω_1 and Darcy's law in the porous medium domain Ω_2. These are separated by an interface Γ_I. Here $\Omega_j \subset \mathbb{R}^d$ ($d = 2$ or 3) are bounded domains with outward unit normal vectors $\hat{n}_j$, $j = 1, 2$. Let $\Gamma_j := \partial\Omega_j \setminus \Gamma_I$. Each interface and boundary is assumed to be polygonal ($d = 2$) or polyhedral ($d = 3$). Figure 1 gives a schematic representation of the geometry.

The fluid velocities and pressures in Ω_1 and Ω_2 are denoted by

$$u_j : \Omega_j \to \mathbb{R}^d, \text{ fluid velocity in } \Omega_j,$$
$$p_j : \Omega_j \to \mathbb{R}\ , \text{ fluid pressure in } \Omega_j.$$

It is important to keep in mind that the velocities and pressures play different mathematical (and physical) roles in the fluid region and in the porous medium.

Recall that the deformation rate tensor $\mathbf{D}$ and stress tensor $\mathbf{T}$ associated with (u_1, p_1) are defined by

$$\mathbf{D}(u_1) := \frac{1}{2}\left(\frac{\partial u_{1i}}{\partial x_j} + \frac{\partial u_{1j}}{\partial x_i}\right), \qquad \mathbf{T}(u_1, p_1) := -p_1\mathbf{I} + 2\mu\mathbf{D}(u_1),$$

where μ is the viscosity. Assuming Stokes flow, (u_1, p_1) satisfies on Ω_1

$$\begin{cases} -\nabla \cdot \mathbf{T}(u_1, p_1) = f_1 \quad \text{in } \Omega_1 \quad \text{(conservation of momentum)}, \\ \nabla \cdot u_1 = 0 \quad \text{in } \Omega_1 \quad \text{(conservation of mass)}, \\ u_1 = 0 \quad \text{on } \Gamma_1 \quad \text{(no slip)}. \end{cases} \tag{1.1}$$

Assuming Darcy's law and no flow through Γ_2, (u_2, p_2) satisfies on Ω_2

$$\begin{cases} u_2 = -k\nabla p_2 \quad \text{in } \Omega_2 \quad \text{(Darcy's law)}, \\ \nabla \cdot u_2 = f_2 \quad \text{in } \Omega_2 \quad \text{(conservation of mass)}, \\ u_2 \cdot \hat{n}_2 = 0 \quad \text{on } \Gamma_2 \quad \text{(no flow)}, \end{cases} \tag{1.2}$$

where k is a symmetric and uniformly positive definite tensor representing the rock permeability divided by the fluid viscosity. The source f_2 is assumed to satisfy the solvability condition

$$\int_{\Omega_2} f_2 \, dx = 0, \tag{1.3}$$

which makes physical sense due to the no-flow boundary condition on $\partial\Omega$ and to (1.4) below. The mixed formulation (1.2) is the most natural one for computations in the porous medium region since it leads to direct approximation of the velocity.

1.2. Interface conditions. The problems (1.1)–(1.2) must be coupled across Γ_I by the correct interface conditions. *Mass conservation* across Γ_I is expressed by

$$u_1 \cdot \hat{n}_1 + u_2 \cdot \hat{n}_2 = 0 \quad \text{on} \quad \Gamma_I. \tag{1.4}$$

The second interface condition is *balance of normal forces* across Γ_I. Recall from, e.g., Serrin [28], that the Cauchy stress vector or traction vector $\vec{t}$ is the force on $\partial\Omega_1$ acting on the fluid volume inside Ω_1 and that

$$\vec{t}(u_1, p_1) = \hat{n}_1 \cdot \mathbf{T}(u_1, p_1)$$

(see Figure 2). Thus, the force on Γ_I exerted by the fluid volume is $-\vec{t}$. The only force in Ω_2 acting on Γ_I is the Darcy pressure p_2. Continuity of forces gives

$$-\vec{t}(u_1, p_1) \cdot \hat{n}_1 = p_2 \quad \text{on} \quad \Gamma_I.$$

This gives the interface condition

$$p_1 - 2\mu \hat{n}_1 \cdot \mathbf{D}(u_1) \cdot \hat{n}_1 = p_2 \quad \text{on} \quad \Gamma_I. \tag{1.5}$$

Finally, since the fluid model is viscous, a condition on the tangential fluid velocity on Γ_I must be given. Let $\hat{\tau}_j, \ j = 1, d-1$, denote an orthonormal system of tangent vectors on Γ_I. The simplest assumption is no-slippage along Γ_I, i.e., $u_1 \cdot \hat{\tau}_j = 0, \ j = 1, d-1$. This is not in good accord with experiment. The boundary condition in best agreement with experimental evidence evolved from the work of Beavers and Joseph [3] and states that

(slip velocity along Γ_I) is proportional to (shear stress along Γ_I).

Mathematically, this can be represented by

$$(u_1 - u_2) \cdot \hat{\tau}_j = \left(\frac{\sqrt{\tilde{k}_j}}{\mu \alpha_1} \right) \ (-\vec{t}(u_1, p_1)) \cdot \hat{\tau}_j, \qquad j = 1, d-1, \quad \text{on} \quad \Gamma_I,$$

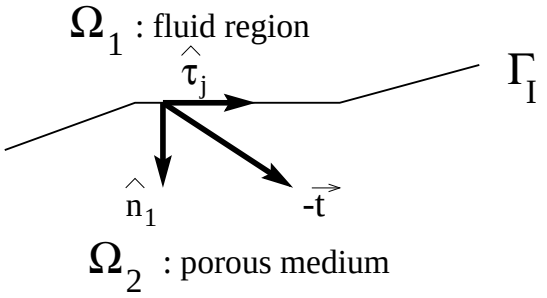


FIG. 2. *The traction vector on* Γ_I.

where $\tilde{k}_j = \hat{\tau}_j \cdot \mu k \cdot \hat{\tau}_j$. However, it is still unclear if this leads to a well-posed problem and it has been observed that the term on the left-hand side "$u_2 \cdot \hat{\tau}_j$" is much smaller than the other terms. Thus, its inclusion in this linear approximation is unclear. The most accepted interface condition was derived by Saffman [25] using a statistical approach and the Brinkman approximation and also by Jones [18] (also see Jäger and Mikelić [17]). This condition, which drops this term, is now known as the Beavers–Joseph–Saffman law and is thus given by

$$u_1 \cdot \hat{\tau}_j = -\frac{\sqrt{\tilde{k}_j}}{\alpha_1}\, 2\hat{n}_1 \cdot \mathbf{D}(u_1) \cdot \hat{\tau}_j, \qquad j = 1, d-1, \text{ on } \Gamma_I. \tag{1.6}$$

Here the form $\sqrt{\tilde{k}_j}/\alpha_1$ for the friction constant arises from dimensional analysis and experimental evidence. The parameter α_1 must be experimentally determined; it seems to depend on many particular features of Γ_I, including its geometry. See, e.g., Beavers and Joseph [3], Payne and Straughan [22], Saffman [25], and Jäger and Mikelić [16, 17] (among roughly 500 papers studying or using this interface condition) for more information.

2. Weak formulation of the coupled problem. This section is devoted to developing suitable weak formulations of the problem (1.1)–(1.6). The weak formulations have two important purposes. One formulation is used to show well-posedness of (1.1)–(1.6). This is already nontrivial because of the incompatibility of the boundary and interface conditions where Γ_I, Γ_1, and Γ_2 meet. Thus, the conditions at these points must be interpreted correctly. A second closely related weak form is developed which is suitable for efficiently splitting the coupled problem into two subproblems. In this formulation the coupling conditions (1.4)–(1.5) are viewed as constraints and imposed via Lagrange multipliers.

Notation. For a subdomain $G \subset \mathbb{R}^d$, the $L^2(G)$ inner product (or duality pairing) and norm are denoted $(\cdot,\cdot)_G$ and $\|\cdot\|_G$, respectively, for scalar, vector, and tensor valued functions. For example, for tensor valued functions $A, B : G \to \mathbb{R}^{d\times d}$,

$$(A,B)_G := \sum_{i,j=1}^{d} \int_G A_{ij}(x) B_{ij}(x) dx = \int_G A : B \; dx.$$

For a connected open subset of the boundary $\Gamma \subset \partial\Omega_1 \cup \partial\Omega_2$, we write $\langle\cdot,\cdot\rangle_\Gamma$ and $\|\cdot\|_\Gamma$ for the $L^2(\Gamma)$ inner product (or duality pairing) and norm, respectively, for

scalar valued functions λ, μ and vector valued functions u, v:

$$\langle \lambda, \mu \rangle_\Gamma := \int_\Gamma \lambda\mu \, ds, \quad \langle u, v \rangle_\Gamma := \int_\Gamma \sum_{i=1}^{d} u_i v_i \, ds.$$

The Sobolev spaces $H^k(\Omega) = W^{k,2}(\Omega)$ are defined in the usual ways for $\Omega = \Omega_1$ or Ω_2 with the usual norm and seminorm $\|\cdot\|_{k,\Omega}$ and $|\cdot|_{k,\Omega}$, respectively. Let

$$X_1 := \{v_1 \in (H^1(\Omega_1))^d : v_1 = 0 \text{ on } \Gamma_1\}, \quad M_1 := L^2(\Omega_1)$$

denote the usual velocity-pressure spaces on Ω_1. The norm on X_1 is given by

$$\|v_1\|_{X_1} := |v_1|_{1,\Omega_1} := \|\nabla v_1\|_{\Omega_1}.$$

The velocity space X_2 on Ω_2 [24, 15, 7] is the subspace of

$$H(\mathrm{div}; \Omega_2) = \{v_2 \in (L^2(\Omega_2))^d : \nabla \cdot v_2 \in L^2(\Omega_2)\}$$

consisting of functions with zero normal trace on Γ_2 and equipped with the norm

$$\|v_2\|_{H(\mathrm{div};\Omega_2)} := (\|v_2\|^2_{\Omega_2} + \|\nabla \cdot v_2\|^2_{\Omega_2})^{1/2}.$$

It is well known [24, 15, 7] that for all $v_2 \in H(\mathrm{div}; \Omega_2)$, $v_2 \cdot \hat{n}_2 \in H^{-1/2}(\partial\Omega_2)$ and there exists a positive constant C such that

$$\|v_2 \cdot \hat{n}_2\|_{-1/2,\partial\Omega_2} \le C \|v_2\|_{H(\mathrm{div};\Omega_2)}. \tag{2.1}$$

The restriction of $v_2 \cdot \hat{n}_2$ to Γ_2, however, may not lie in $H^{-1/2}(\Gamma_2)$. We define the velocity-pressure spaces on Ω_2 as follows [30], [7, sect. III.1]:

$$X_2 := \{v_2 \in H(\mathrm{div}; \Omega_2) : \langle v_2 \cdot \hat{n}_2, w \rangle_{\partial\Omega_2} = 0 \text{ for all } w \in H^1_{0,\Gamma_I}(\Omega_2)\}, \quad M_2 := L^2(\Omega_2),$$

where

$$H^1_{0,\Gamma_I}(\Omega_2) = \{w \in H^1(\Omega_2) : w = 0 \text{ on } \Gamma_I\}.$$

Defining $X := X_1 \times X_2$, a typical $v \in X$ takes the form (v_1, v_2) with $v_i \in X_i$. The norm on X is, as usual,

$$\|v\|_X := (\|v_1\|^2_{X_1} + \|v_2\|^2_{X_2})^{1/2} \qquad \text{for all } v \in X.$$

If $V \subset X$ is any closed subspace, then $\|\cdot\|_X$ is also the induced norm on V. Similarly, let

$$M := \left\{ q = (q_1, q_2) : q_i \in M_i \text{ and } \sum_{i=1}^{2} (q_i, 1)_{\Omega_i} = 0 \right\},$$

with norm

$$\|q\|_M := (\|q_1\|^2_{M_1} + \|q_2\|^2_{M_2})^{1/2}.$$

The coupling across Γ_I between the subproblems in Ω_1 and Ω_2 occurs in the interface conditions (1.4)–(1.5). The procedure for uncoupling the two subproblems is to pick one (we pick the second) and introduce the Lagrange multiplier λ:

$$p_1 - 2\mu\hat{n}_1 \cdot \mathbf{D}(u_1) \cdot \hat{n}_1 = \lambda = p_2 \qquad \text{on } \Gamma_I. \tag{2.2}$$

Considering λ to be known data for each subproblem, the weak formulation is then derived in the usual manner as follows. Beginning with a classical solution of (1.1), multiplying by a sufficiently smooth $v_1 \in X_1$, and integrating by parts gives

$$\begin{aligned}(f_1, v_1)_{\Omega_1} &= (-2\mu\nabla\cdot\mathbf{D}(u_1) + \nabla p_1, v_1)_{\Omega_1} \\ &= 2\mu(\mathbf{D}(u_1), \mathbf{D}(v_1))_{\Omega_1} - (p_1, \nabla\cdot v_1)_{\Omega_1} \\ &+ \langle\{p_1 - 2\mu\hat{n}_1\mathbf{D}(u_1)\hat{n}_1\}, v_1\cdot\hat{n}_1\rangle_{\Gamma_I} \\ &+ \sum_{j=1}^{d}\langle\{-2\mu\hat{n}_1\mathbf{D}(u_1)\hat{\tau}_j\}, v_1\cdot\hat{\tau}_j\rangle_{\Gamma_I}.\end{aligned}$$

The first term in the braces $\{\cdot\}$ is replaced by λ using (2.2) and the second by $(\mu\alpha_1/\sqrt{\tilde{k}_j})\; u_1\cdot\hat{\tau}_j$ using (1.6). Therefore, introducing the bilinear forms

$$a_1(u_1, v_1) := 2\mu(\mathbf{D}(u_1), \mathbf{D}(v_1))_{\Omega_1} + \sum_{j=1}^{d-1}\frac{\mu\alpha_1}{\sqrt{\tilde{k}_j}}\,\langle u_1\cdot\hat{\tau}_j, v_1\cdot\hat{\tau}_j\rangle_{\Gamma_I} \text{ for all } u_1, v_1 \in X_1,$$

and

$$b_1(v_1, q_1) := -(q_1, \nabla\cdot v_1)_{\Omega_1} \qquad \text{for all } v_1 \in X_1, q_1 \in M_1,$$

we obtain for all $v_1 \in X_1$ and $q_1 \in M_1$

$$\begin{aligned}a_1(u_1, v_1) + b_1(v_1, p_1) + \langle\lambda, v_1\cdot\hat{n}_1\rangle_{\Gamma_I} &= (f_1, v_1)_{\Omega_1}, \\ b_1(u_1, q_1) &= 0.\end{aligned}$$

In the porous medium region, multiplication of the first equation in (1.2) by $v_2 \in X_2$, integration over Ω_2, and integration by parts gives

$$0 = (k^{-1}u_2 + \nabla p_2, v_2)_{\Omega_2} = (k^{-1}u_2, v_2)_{\Omega_2} - (p_2, \nabla\cdot v_2)_{\Omega_2} + \langle\lambda, v_2\cdot\hat{n}_2\rangle_{\Gamma_I},$$

where, by (2.2), p_2 is replaced by λ in the last term. Introducing

$$a_2(u_2, v_2) := (k^{-1}u_2, v_2)_{\Omega_2}, \quad b_2(v_2, p_2) := -(p_2, \nabla\cdot v_2)_{\Omega_2},$$

we have

$$\begin{aligned}a_2(u_2, v_2) + b_2(v_2, p_2) + \langle\lambda, v_2\cdot\hat{n}_2\rangle_{\Gamma_I} &= 0 \qquad \text{for all } v_2 \in X_2, \\ b_2(u_2, q_2) &= -(f_2, q_2) \qquad \text{for all } q_2 \in M_2.\end{aligned}$$

The linking across Γ_I occurs through the condition $u_1\cdot\hat{n}_1 + u_2\cdot\hat{n}_2 = 0$ on Γ_I and the definition (2.2) of λ. This linkage is the key to the well-posedness of the coupled problem and it hinges on the choice of the space Λ for the Lagrange multipliers. Define

$$b_I(v, \lambda) := \langle v_1\cdot\hat{n}_1 + v_2\cdot\hat{n}_2, \lambda\rangle_{\Gamma_I} : X\times\Lambda \to \mathbb{R},$$

where Λ is not yet specified. The flux continuity condition (1.4) on Γ_I is then

$$b_I(v, \lambda) = 0 \qquad \text{for all } \lambda \in \Lambda.$$

Since $v_2 \in H(\mathrm{div}, \Omega_2)$, it holds that $v_2 \cdot \hat{n}_2 \in H^{-1/2}(\partial\Omega_2)$. We wish to pick $\Lambda \subset L^2(\Gamma_I)$ to be the largest space for which the pairing $\langle v_2 \cdot \hat{n}_2, \lambda \rangle_{\Gamma_I}$ is well defined. We show in Lemma 2.1 below (see also [20]) that

$$v_2 \cdot \hat{n}_2|_{\Gamma_I} \in (H_{00}^{1/2}(\Gamma_I))^*,$$

where $H_{00}^{1/2}(\Gamma_I)$ is the completion of the smooth functions with compact support in Γ_I with respect to the norm

$$\|\mu\|_{1/2,\partial\Omega_2} := \left(\|\mu\|_{\partial\Omega_2}^2 + \int_{\partial\Omega_2} \int_{\partial\Omega_2} \frac{|\mu(t_1) - \mu(t_2)|^2}{|t_1 - t_2|^d} ds_{t_1} ds_{t_2} \right)^{1/2}.$$

It is well known that $H_{00}^{1/2}(\Gamma_I)$ is the interpolation space

$$H_{00}^{1/2}(\Gamma_I) = [L^2(\Gamma_I), H_0^1(\Gamma_I)]_{1/2}.$$

Any function $\mu \in H_{00}^{1/2}(\Gamma_I)$ has the property that its extension by zero to $\partial\Omega_j$ gives a function $\tilde{\mu}_j \in H^{1/2}(\partial\Omega_j)$ with

$$\|\tilde{\mu}_j\|_{1/2,\partial\Omega_j} \le C \|\mu\|_{H_{00}^{1/2}(\Gamma_I)}, \quad j = 1, 2. \tag{2.3}$$

See Lions and Magenes [19] for background information on $H_{00}^{1/2}(\Gamma_I)$.

Accordingly, choose

$$\Lambda := H_{00}^{1/2}(\Gamma_I) \ (\subset L^2(\Gamma_I)).$$

LEMMA 2.1. *The bilinear form $b_I(\cdot, \cdot)$ is continuous on $X \times \Lambda$.*

Proof. First note that $v_j \cdot \hat{n}_j \in H^{-1/2}(\partial\Omega_j), j = 1, 2$. Let $\mu \in H_{00}^{1/2}(\Gamma_I)$ and let $\tilde{\mu}_j$ be its extension by zero to $\partial\Omega_j$. We have, for $j = 1, 2$,

$$\begin{aligned} \int_{\Gamma_I} v_j \cdot \hat{n}_j \mu \, ds &= \int_{\partial\Omega_j} v_j \cdot \hat{n}_j \tilde{\mu}_j \, ds \le \|v_j \cdot \hat{n}_j\|_{-1/2,\partial\Omega_j} \|\tilde{\mu}_j\|_{1/2,\partial\Omega_j} \\ &\le C \|v\|_X \|\mu\|_\Lambda, \end{aligned}$$

using (2.1) and (2.3) in the last inequality. ∎

Further, define

$$\begin{aligned} a(u, v) &:= \sum_{i=1}^{2} a_i(u_i, v_i) : X \times X \to \mathbb{R}, \\ b(v, p) &:= \sum_{i=1}^{2} b_i(v_i, p_i) : X \times M \to \mathbb{R}, \\ \ell(v) &:= (f_1, v_1)_{\Omega_1}, \quad g(q) := -(f_2, q_2)_{\Omega_2}. \end{aligned}$$

Then, (1.1)–(1.6) has the following weak formulation: find $(u, p, \lambda) \in X \times M \times \Lambda$ satisfying

$$\begin{cases} a(u, v) + b(v, p) + b_I(v, \lambda) = \ell(v) & \text{for all } v \in X, \\ b(u, q) = g(q) & \text{for all } q \in M, \\ b_I(u, \mu) = 0 & \text{for all } \mu \in \Lambda. \end{cases} \tag{2.4}$$

We next derive another weak formulation using the space V of functions in X with trace-continuous normal velocities:

$$V := \{v \in X : b_I(v, \mu) = 0 \text{ for all } \mu \in \Lambda\}.$$

The connection between the two formulations (2.4) and (2.5) is considered in Remark 3.1 in section 3. Note that, due to Lemma 2.1, V is a closed subspace of X, e.g., Brezzi and Fortin [7]. The next lemma indicates that a trace-continuous normal velocity has a well-defined divergence on the whole domain. Let

$$\Omega := \text{ interior}(\overline{\Omega}_1 \cup \overline{\Omega}_2).$$

For a given $v = (v_1, v_2) \in X$, define $\tilde{v} \in (L^2(\Omega))^d$ by $\tilde{v}|_{\Omega_j} := v_j$, $j = 1, 2$. To simplify notation we will omit the tilde in this construction since the meaning whether it is v or $\tilde{v}$ is clear from the context.

LEMMA 2.2. *If $v \in V$, then $v \in H(\text{div}; \Omega)$.*

Proof. Define

$$g(x) = \nabla \cdot v_j(x) \quad \text{for } x \in \Omega_j, \ j = 1, 2.$$

We will show that $g = \nabla \cdot v$. Since $v_j \in H(\text{div}; \Omega_j)$, $j = 1, 2$, we can apply the divergence theorem in each Ω_j. This gives, for all $\phi \in C_0^\infty(\Omega)$,

$$\begin{aligned}\int_\Omega v \nabla \phi \, dx &= \int_{\Omega_1} v_1 \nabla \phi \, dx + \int_{\Omega_2} v_2 \nabla \phi \, dx \\ &= -\int_{\Omega_1} (\nabla \cdot v_1)\phi \, dx - \int_{\Omega_2} (\nabla \cdot v_2)\phi \, dx \\ &\quad + \int_{\Gamma_I} (v_1 \cdot \hat{n}_1 + v_2 \cdot \hat{n}_2)\phi \, dx.\end{aligned}$$

The last term vanishes since $\phi \in C_0^\infty(\Omega)$ implies $\phi_{|\Gamma_I} \in H_{00}^{1/2}(\Gamma_I)$. Thus,

$$\int_\Omega v \nabla \phi \, dx = -\int_\Omega g \phi \, dx.$$

Since $\nabla \cdot v_j \in L^2(\Omega_j)$, $g \in L^2(\Omega)$, and hence g is the weak L^2 divergence of $v \in V$. ∎

We next define the subspace Z,

$$Z := \{v \in V : b(v, q) = 0 \qquad \text{for all } q \in M\}.$$

LEMMA 2.3. *The space Z is a closed subspace of V and X. Moreover, if $v \in Z$, then $\nabla \cdot v = 0$, a.e. $x \in \Omega$.*

Proof. Let $v \in Z$. Since $Z \subset V$, we know by Lemma 2.2 that $v \in H(\text{div}; \Omega)$. Thus, for any $q \in M$

$$b(v, q) = -\int_\Omega \nabla \cdot v \ q \, dx.$$

We claim that $\nabla \cdot v \in M$. Indeed, $\nabla \cdot v \in L^2(\Omega)$ and $\nabla \cdot v$ has zero mean value over Ω:

$$\int_\Omega \nabla \cdot v \, dx = \int_{\partial\Omega} v \cdot \hat{n} \, ds = 0$$

using the divergence theorem. Thus, $\nabla \cdot v \in M$. The second part of the lemma follows by setting $q = \nabla \cdot v$.

The space Z is a closed subspace of V since

$$\begin{aligned} b(v,q) &= -\int_\Omega \nabla \cdot v \; q \; dx \le \|\nabla \cdot v\|_\Omega \|q\|_\Omega \\ &\le \|v\|_X \|q\|_M, \end{aligned}$$

i.e., $b(\cdot,\cdot)$ is continuous on $V \times M$. ∎

Since V is a closed subspace of X, we can write the following variational formulation: find $(u,p) \in V \times M$ satisfying

$$\begin{cases} a(u,v) + b(v,p) = \ell(v) & \text{for all } v \in V, \\ b(u,q) = g(q) & \text{for all } q \in M. \end{cases} \tag{2.5}$$

We end this section noting that, under the solvability condition (1.3), any solution of (2.5) satisfies the mass conservation equations in (1.1) and (1.2). Indeed, define $f \in L^2(\Omega)$ such that $f = 0$ on Ω_1 and $f = f_2$ on Ω_2. If (u,p) is a solution to (2.5), then $\nabla \cdot u \in L^2(\Omega)$ due to Lemma 2.2. The second equation in (2.5) implies that $\nabla \cdot u - f = c$, where c is a constant. The divergence theorem gives

$$c|\Omega| = \int_\Omega (\nabla \cdot u - f)\, dx = \int_{\partial\Omega} u \cdot \hat{n}\, ds - \int_\Omega f\, dx = -\int_{\Omega_2} f_2\, dx = 0$$

using (1.3). Therefore $\nabla \cdot u = 0$ on Ω_1 and $\nabla \cdot u = f_2$ on Ω_2.

3. Analysis of the weak formulation. This section is devoted to a proof of existence of weak solutions to (1.1)–(1.6) based on the weak formulations (2.4) and (2.5). Existence depends on our choice of the Lagrange multiplier space $\Lambda = H_{00}^{1/2}(\Gamma_I)$ so that the problem is neither over nor underconstrained.

We begin with a few simple but useful estimates. Let

$$W_2 := \{v_2 \in X_2 : \nabla \cdot v_2 = 0, \text{ a.e. } x \in \Omega_2\} \subset X_2$$

denote the (closed) subspace of div-free functions in X_2.

LEMMA 3.1. *For $v_i \in H^1(\Omega_i)^d \cap X_i$ $(i = 1,2)$ we have*

$$C_1 \|v_i\|_{\Omega_i} \le \|v_i\|_{X_i} \le C_2 \|v_i\|_{1,\Omega_i}. \tag{3.1}$$

Furthermore, for $i = 1,2$, there holds

$$|a_i(u_i,v_i)| \le C_3 \|u_i\|_{X_i} \|v_i\|_{X_i} \text{ for all } u_i, v_i \in X_i, \tag{3.2}$$

$$a_1(v_1,v_1) \ge C_4 \|v_1\|_{X_1}^2 \text{ for all } v_1 \in X_1, \tag{3.3}$$

$$a_2(v_2,v_2) \ge C_5 \|v_2\|_{X_2}^2 \text{ for all } v_2 \in W_2, \tag{3.4}$$

$$|b_i(v_i,p_i)| \le C_6 \|v_i\|_{X_i}, \|p_i\|_{M_i} \text{ for all } v_i \in X_i, p_i \in M_i, \tag{3.5}$$

$$|a(u,v)| \le C_3 \|u\|_X \|v\|_X \text{ for all } u, v \in X, \tag{3.6}$$

$$|b(v,p)| \le C_6 \|v\|_X \|p\|_M \text{ for all } v \in X,\ p \in M, \tag{3.7}$$

$$a(v,v) \ge \min\{C_4, C_5\} \|v\|_X^2 \text{ for all } v \in X_1 \times W_2. \tag{3.8}$$

Proof. Inequalities (3.1) and (3.2) follow from the Poincaré–Friedrich inequality and the trace theorem. The Korn inequality implies (3.3) while (3.4) and (3.5) are immediate. Inequalities (3.6), (3.7), and (3.8) follow by combining earlier ones. ∎

The next lemma establishes the Ladyzhenskaya–Babuška–Brezzi condition required for the formulation (2.5) in $V \times M$.

LEMMA 3.2. *There is a constant $\beta > 0$ such that*

$$\inf_{q \in M \setminus \{0\}} \sup_{v \in V \setminus \{0\}} \frac{b(v,q)}{\|v\|_X \ \|q\|_M} \geq \beta. \tag{3.9}$$

Proof. Let $q \in M \setminus \{0\}$ be fixed but arbitrary. We construct a $v \in V$ satisfying

$$b(v,q) \geq \beta \|v\|_X \ \|q\|_M.$$

Given $q = (q_1, q_2) \in M$, the function $\tilde{q}(x)$ defined by $\tilde{q}|_{\Omega_i} = q_i$ has mean value zero over Ω; thus $\tilde{q} \in L^2_0(\Omega)$. Thus, (see, e.g., [15, 13]) there exists $\tilde{v} \in (H^1_0(\Omega))^d$ satisfying

$$\nabla \cdot \tilde{v} = \tilde{q}, \text{ in } \Omega, \ \tilde{v} = 0, \text{ on } \partial\Omega, \ \|\tilde{v}\|_{1,\Omega} \leq C_7 \|\tilde{q}\|_\Omega.$$

Given this $\tilde{v}$, define $v = (v_1, v_2) \in X$ by $v_i = \tilde{v}|_{\Omega_i}, (i = 1, 2)$. Since

$$\tilde{v} \in H^1_0(\Omega)^d, \text{ it follows that } v_1|_{\Gamma_1} = 0 \text{ and } v_2 \cdot \hat{n}_2|_{\Gamma_2} = 0.$$

Further, $v_1|_{\Gamma_I} = v_2|_{\Gamma_I} = \tilde{v}|_{\Gamma_I} \in (H^{1/2}_{00}(\Gamma_I))^d$ so that $v_i \cdot \hat{n}_i \in L^2(\Gamma_I)$ $(i = 1, 2)$ and

$$b_I(v, \mu) = \langle v_1 \cdot \hat{n}_1 + v_2 \cdot \hat{n}_2, \ \mu \rangle_{\Gamma_I} = 0$$

for all $\mu \in L^2(\Gamma_I)$. Thus, $v \in V$. Using (3.1) we find

$$\|v\|_X \leq C_2 \|\tilde{v}\|_{1,\Omega} \leq C_2 C_7 \|\tilde{q}\|_{0,\Omega} = C_2 C_7 \|q\|_M.$$

Finally, for this v

$$b(v,q) = \sum_{i=1}^{2} (-\nabla \cdot v_i, q_i) = -(\nabla \cdot \tilde{v}, \tilde{q})_\Omega \tag{3.10}$$

$$= \|\tilde{q}\|^2_{0,\Omega} \geq (C_2 C_7)^{-1} \|v\|_X \|q\|_M, \tag{3.11}$$

completing the proof with $\beta = (C_2 C_7)^{-1}$. ∎

To apply the abstract theory of mixed problems in, e.g., Girault and Raviart [15], Brezzi and Fortin [7], we must show $a(\cdot, \cdot)$ is coercive on the constraint set Z. This is accomplished in the next lemma.

LEMMA 3.3. *$a(\cdot, \cdot)$ is coercive on Z: there is an $\alpha > 0$ such that*

$$a(v,v) \geq \alpha \|v\|^2_X \qquad \textit{for all } v \in Z.$$

Proof. Note that by Lemma 2.3 if $v = (v_1, v_2) \in \ker(B)$, $\nabla \cdot v_2 = 0$, a.e. $x \in \Omega$, i.e., $v_2 \in W_2$. Coercivity now follows from (3.8) of Lemma 3.1. ∎

Lemmas 2.1, 3.2, and 3.3, together with the abstract theory of mixed problems [15, 7], immediately imply existence of a weak solution $(u, p) \in V \times M$ satisfying (2.5).

THEOREM 3.1. *There exists a unique solution $(u, p) \in V \times M$ to the problem (2.5).* ∎

To verify that the solution to (2.5) is also the solution to the formulation (2.4) in $X \times M \times \Lambda$ using the general saddle point problem theory [15, 7], we must verify the inf-sup condition

$$\inf_{0 \neq \lambda \in \Lambda} \sup_{0 \neq v \in X} \frac{b_I(v, \lambda)}{\|v\|_X \|\lambda\|_\Lambda} \geq \beta > 0. \tag{3.12}$$

Due to technical difficulties related to the restriction of $H^{-1/2}(\partial\Omega_2)$ functions to Γ_I, we are only able to show a modified inf-sup condition:

$$\inf_{0\neq\lambda\in\Lambda}\sup_{0\neq v\in X}\frac{b_I(v,\lambda)}{\|v\|_X\|\lambda\|_{1/2,\Gamma_I}}\geq\beta>0. \tag{3.13}$$

LEMMA 3.4. *The* inf-sup *condition* (3.13) *holds.*

Proof. Fix $\lambda\in H^{1/2}_{00}(\Gamma_I)$ and let $\tilde{\lambda}\in H^{1/2}(\partial\Omega_2)$ be its extension by zero to $\partial\Omega_2$. Since $H^{1/2}_{00}(\Gamma_I)\subset H^{1/2}(\Gamma_I)$, there exists $\hat{\lambda}_I\in H^{-1/2}(\Gamma_I)$ such that

$$\frac{\langle\hat{\lambda}_I,\lambda\rangle_{\Gamma_I}}{\|\hat{\lambda}_I\|_{-1/2,\Gamma_I}}\geq\frac{1}{2}\|\lambda\|_{1/2,\Gamma_I}. \tag{3.14}$$

We next define $\hat{\lambda}\in H^{-1/2}(\partial\Omega_2)$ by

$$\langle\hat{\lambda},w\rangle_{\partial\Omega_2}:=\langle\hat{\lambda}_I,w\rangle_{\Gamma_I}\quad\text{for all } w\in H^{1/2}(\partial\Omega_2).$$

We then have

$$\|\hat{\lambda}\|_{-1/2,\partial\Omega_2}=\sup_{0\neq w\in H^{1/2}(\partial\Omega_2)}\frac{\langle\hat{\lambda}_I,w\rangle_{\Gamma_I}}{\|w\|_{1/2,\partial\Omega_2}}\leq\|\hat{\lambda}_I\|_{-1/2,\Gamma_I}. \tag{3.15}$$

Since the normal trace operator maps $H(\mathrm{div},\Omega_2)$ onto $H^{-1/2}(\partial\Omega_2)$ (see [15, Corollary 2.8]) and it is continuous (see (2.1)), by the open mapping theorem there exists $v_2\in H(\mathrm{div},\Omega_2)$ such that $v_2\cdot\hat{n}_2=\hat{\lambda}$ on $\partial\Omega_2$ and

$$\|v_2\|_{X_2}\leq C\|\hat{\lambda}\|_{-1/2,\partial\Omega_2}\leq C\|\hat{\lambda}_I\|_{-1/2,\Gamma_I}, \tag{3.16}$$

using (3.15) for the second inequality. We note that $v_2\in X_2$ since, for all $w\in H^1_{0,\Gamma_I}(\Omega_2)$,

$$\langle v_2\cdot\hat{n}_2,w\rangle_{\partial\Omega_2}=\langle\hat{\lambda},w\rangle_{\partial\Omega_2}=\langle\hat{\lambda}_I,w\rangle_{\Gamma_I}=0.$$

Choosing $v=(0,v_2)\in X$ and using (3.14) and (3.16) we get

$$\begin{aligned}\frac{b_I(v,\lambda)}{\|v\|_X}&=\frac{\langle v_2\cdot\hat{n}_2,\tilde{\lambda}\rangle_{\partial\Omega_2}}{\|v_2\|_{X_2}}=\frac{\langle\hat{\lambda},\tilde{\lambda}\rangle_{\partial\Omega_2}}{\|v_2\|_{X_2}}\\&=\frac{\langle\hat{\lambda}_I,\lambda\rangle_{\Gamma_I}}{\|v_2\|_{X_2}}\geq\frac{1}{C}\frac{\langle\hat{\lambda}_I,\lambda\rangle_{\Gamma_I}}{\|\hat{\lambda}_I\|_{-1/2,\Gamma_I}}\geq\beta\|\lambda\|_{1/2,\Gamma_I}.\end{aligned}$$

∎

REMARK 3.1. *If the porous medium is entirely enclosed within the fluid region, then* $\Gamma_I=\partial\Omega_2$. *In this case there are no incompatible points and it is easy to extend slightly the proof of Lemma* 3.4 *to show that the stronger* inf-sup *condition* (3.12) *holds. In this case, the unique weak solution to* (2.5) *is also the unique weak solution to* (2.4) *and the two formulations are equivalent.*

4. Finite element discretization. This section considers the finite element discretization of the coupled problem. The interface conditions on Γ_I separate into tangential and normal conditions. This splitting on Γ_I introduces interesting features into the finite element procedure and its analysis.

Introduce upon Ω_j a mesh $\mathcal{T}_j^h$ $(j = 1, 2)$ with $\overline{\Omega}_j = \cup_{K\in\mathcal{T}_j^h}\overline{K}$. To simplify the notation we shall assume that the cells $K \in \mathcal{T}_j^h$ are affine equivalent, the grids $\mathcal{T}_1^h$ and $\mathcal{T}_2^h$ match at Γ_I, that Γ_I is polyhedral, and that no point of the interface boundary $\partial\Gamma_I$ belongs to the interior of an element face. We use the notation

$$\mathcal{E}_h(K) := \text{the set of all faces of the element } K,$$
$$\mathcal{E}_h(\Gamma_I) := \text{the set of all element faces } E \text{ with } E \subset \Gamma_I.$$

For the discretization of the fluid's variables we choose finite element spaces X_1^h and M_1^h which are assumed to be div-stable (also called LBB-stable),

$$\begin{cases} X_1^h \subset X_1,\ M_1^h \subset M_1, \text{ and} \\ \inf\limits_{0\neq q_1\in M_1^h}\ \sup\limits_{0\neq v_1\in X_1^h} \frac{b_1(v_1,q_1)}{\|v_1\|_{X_1}\|q_1\|_{M_1}} \ \geq\ \beta_1 > 0, \end{cases} \tag{4.1}$$

and to satisfy a discrete Korn inequality

$$(\mathbf{D}(v_1), \mathbf{D}(v_1))_{\Omega_1} \geq \alpha_1 |v_1|^2_{1,\Omega_1} \quad \text{for all } v_1 \in X_1^h. \tag{4.2}$$

We assume that X_1^h and M_1^h include at least polynomials of degree r_1 and $r_1 - 1$, respectively, $(r_1 \geq 1)$. Specifically, we assume that there exist (quasi) interpolation operators

$$I^h_{X_1} : X_1 \cap (H^s(\Omega_1))^d \to X_1^h \quad \text{and} \quad I^h_{M_1} : M_1 \cap H^s(\Omega_1) \ \to M_1^h$$

such that for all $K \in \mathcal{T}_1^h$

$$\begin{cases} |v_1 - I^h_{X_1} v_1|_{m,K} \leq C h_K^{s-m} |v_1|_{s,\delta(K)}, \quad m = 0, 1,\ 1 \leq s \leq r_1 + 1, \\ \|q_1 - I^h_{M_1} q_1\|_{0,K} \leq C h_K^s |q_1|_{s,\delta(K)}, \quad 0 \leq s \leq r_1. \end{cases} \tag{4.3}$$

Here $\delta(K)$ is equal to K in most cases of usual interpolation operators. However, in cases of quasi interpolation operators suited for H^1 functions like the Clement-operator [9] or the Scott–Zhang-operator [27], $\delta(K)$ denotes the vicinity of K consisting of all elements $\widetilde{K} \in \mathcal{T}_1^h$ that touch element K. We assume the grids $\mathcal{T}_1^h$ and $\mathcal{T}_2^h$ to be shape-regular in the usual sense such that cases with local grid refinement are allowed. For shape-regular grids, changes of the mesh size within the vicinity $\delta(K)$ of an element K are uniformly bounded by a constant C, i.e., in particular for $\mathcal{T}_1^h$,

$$C^{-1}\, h_K \leq h_{\widetilde{K}} \leq C\, h_K \qquad \text{for all } \widetilde{K} \subset \delta(K),\ \ \widetilde{K}, K \in \mathcal{T}_1^h. \tag{4.4}$$

This estimate is used to get rid of the $\delta(K)$-terms in final error estimates.

Examples of spaces satisfying (4.1)–(4.3) include the MINI elements [2], the Taylor–Hood elements [29], and the conforming Crouzeix–Raviart elements [10]. See, e.g., [15, 7], for a more complete list of such spaces.

REMARK 4.1. *The discrete Korn inequality* (4.2) *is inherited from the continuous inequality for all conforming elements. However, nonconforming spaces, in general, do not satisfy* (4.2)*; see* [12].

REMARK 4.2. *The* inf-sup *condition* (4.1) *differs from the usual one verified in the literature* [15, 7] *for various spaces because the pressure space M_1^h is not restricted to have zero mean over Ω_1, i.e., $M_1^h \subset L^2(\Omega_1)$, not $L_0^2(\Omega_1)$. However, the usual discrete* inf-sup *condition is almost enough to prove* (4.1). *The main extra ingredient*

needed is the existence of a (typically locally constructed, see [7, section VI.4]*) operator* $P_1^h : X_1 \to X_1^h$ *(not necessarily the same as* $I_{X_1}^h$*) satisfying, for all* $K \in \mathcal{T}_1^h$ *and all* $v_1 \in X_1$*,*

$$\int_K \nabla \cdot (P_1^h v_1 - v_1)\, dx = 0 \quad \text{and} \quad \|P_1^h v_1\|_{1,\Omega_1} \le C_8 \|v_1\|_{1,\Omega_1}, \tag{4.5}$$

where C_8 *is a constant independent of* v_1 *and* h*. In, e.g.,* [7]*, such an operator is locally constructed for all the aforementioned spaces.*

The following lemma gives sufficient conditions for the discrete LBB-stability (4.1) of the spaces X_1^h and M_1^h.

LEMMA 4.1. *Suppose that an operator* $P_1^h : X_1 \to X_1^h$ *satisfying the condition* (4.5) *exists. Suppose also the spaces* $X_1^h \cap (H_0^1(\Omega_1))^d$ *and* $M_1^h \cap L_0^2(\Omega_1)$ *satisfy the usual discrete* inf-sup *condition. Then, the spaces* X_1^h *and* M_1^h *satisfy* (4.1)*.*

Proof. Let $q_1^h \equiv q_0 \in \mathbb{R}$ be an arbitrary constant function of M_1^h. We first show that there exists a $v_1^h \in X_1^h$ such that

$$b_1(v_1^h, q_1^h) \ge \beta_0 \, \|v_1^h\|_{X_1} \|q_1^h\|_{M_1}$$

with a constant $\beta_0 > 0$ independent of v_1^h and h. To this end, let $\tilde{v}_1$ be a solution of the following problem: find $\tilde{v}_1 \in X_1$ satisfying

$$\nabla \cdot \tilde{v}_1 = q_1^h \ \text{ in } \Omega_1, \quad \tilde{v}_1 = g_1 \ \text{ on } \partial\Omega_1,$$

where g_1 is chosen suitably such that the compatibility condition $\langle g_1 \cdot \hat{n}_1, 1\rangle_{\partial\Omega_1} = (q_1^h, 1)_{\Omega_1} = q_0|\Omega_1|$ is fulfilled and $g_1 \in (H^{1/2}(\partial\Omega_1))^d$. By, e.g., [13, sect. III.3, Exercise 3.4], such a $\tilde{v}_1$ exists and satisfies the estimate

$$\|\tilde{v}_1\|_{1,\Omega_1} \le C_9 \{ \|q_1^h\|_{\Omega_1} + \|g_1\|_{1/2,\partial\Omega_1} \}.$$

For the construction of g_1, let $\varphi_0 \in C(\partial\Omega_1)$ be such that $\varphi_0 \equiv 0$ on Γ_1, φ_0 is quadratic on Γ_I, and $\langle \varphi_0, 1\rangle_{\Gamma_I} = 1$. Then, we choose g_1 as $g_1 := |\Omega_1| q_0 \varphi_0 \hat{n}_1$. One can easily verify that g_1 belongs to $(H^{1/2}(\partial\Omega_1))^d$ and satisfies the compatibility condition as well as the estimate $\|g_1\|_{1/2,\partial\Omega_1} \le c(\Omega_1, \varphi_0)\|q_1^h\|_{\Omega_1}$. This implies

$$\|\tilde{v}_1\|_{1,\Omega_1} \le C_9 \{1 + c(\Omega_1, \varphi_0)\} \, \|q_1^h\|_{\Omega_1}.$$

Defining $v_1^h := -P_1^h \tilde{v}_1$, we have

$$\frac{b_1(v_1^h, q_1^h)}{\|v_1^h\|_{X_1}\|q_1^h\|_{M_1}} = \frac{(\nabla \cdot \tilde{v}_1, q_1^h)_{\Omega_1}}{\|P_1^h \tilde{v}_1\|_{X_1}\|q_1^h\|_{M_1}} \ge \frac{\|q_1^h\|_{M_1}^2}{C_8 \|\tilde{v}_1\|_{X_1}\|q_1^h\|_{M_1}} \ge \beta_0 \tag{4.6}$$

with $\beta_0 := (C_8 C_9 \{1 + c(\Omega_1, \varphi_0)\})^{-1}$. Now, using this result and the assumed discrete inf-sup condition for the spaces $X_1^h \cap (H_0^1(\Omega_1))^d$ and $M_1^h \cap L_0^2(\Omega_1)$, we can show in the same way as in the proof of Theorem 1.12, section II.1.4 in [15] that the spaces X_1^h and M_1^h satisfy the inf-sup condition (4.1). ∎

For the discretization of the porous medium problem in Ω_2, we choose $X_2^h \times M_2^h \subset X_2 \times M_2$ to be any of the well-known mixed finite element spaces (see [7, section III.3]), the RT spaces [24, 21], the BDM spaces [6], the BDFM spaces [5], the BDDF spaces [4], or the CD spaces [8]. We assume that X_2^h and M_2^h contain at least polynomials of degree r_2 and l_2, respectively. It is known for these choices that

$$\nabla \cdot X_2^h = M_2^h$$

and that there exists an interpolation operator $I^h_{X_2} : (H^1(\Omega_2))^d \to X^h_2$ such that for all $v_2 \in (H^1(\Omega_2))^d$

$$(\nabla \cdot I^h_{X_2} v_2, w)_{\Omega_2} = (\nabla \cdot v_2, w)_{\Omega_2}, \quad w \in M^h_2. \tag{4.7}$$

Let $I^h_{M_2} : M_2 \to M^h_2$ be the L^2 orthogonal projection such that for all $q_2 \in M_2$

$$(I^h_{M_2} q_2, w)_{\Omega_2} = (q_2, w)_{\Omega_2}, \quad w \in M^h_2. \tag{4.8}$$

Our next lemma will collect some known useful results for these spaces. Their proof can be found in [7, section III.3].

LEMMA 4.2. *There holds, for all $v_2 \in (H^1(\Omega_2))^d$,*

$$\langle I^h_{X_2} v_2 \cdot \hat{n}_2, \mu \rangle_E = \langle v_2 \cdot \hat{n}_2, \mu \rangle_E \tag{4.9}$$
$$\text{for all } \mu \in R_{r_2}(E) \text{ and for all } E \in \mathcal{E}_h(\Gamma_I),$$

where

$$R_{r_2}(E) := \begin{cases} \mathcal{P}_{r_2}(E) \text{ if } d = 2 \text{ or } E \text{ is a triangle,} \\ \mathcal{Q}_{r_2}(E) \text{ if } d = 3 \text{ and } E \text{ is a quadrilateral,} \end{cases} \tag{4.10}$$

where $P_{r_2}(E)$ and $Q_{r_2}(E)$ are the usual polynomial spaces (see, e.g., [7].) For the restrictions to the element faces,

$$v^h_2 \cdot \hat{n}_2|_E \in R_{r_2}(E) \quad \text{for all } v^h_2 \in X^h_2,\ E \in \mathcal{E}(K),\ K \in \mathcal{T}^h_2. \tag{4.11}$$

Further, the operators $I^h_{X_2}$ and $I^h_{M_2}$ satisfy, for all $K \in \mathcal{T}^h_2$,

$$\|q_2 - I^h_{M_2} q_2\|_{0,K} \le C h^s_K |q_2|_{s,K},\ 0 \le s \le l_2 + 1, \tag{4.12}$$

$$|v_2 - I^h_{X_2} v_2|_{m,K} \le C h^{s-m}_K |v_2|_{s,K},\ m \in \{0, 1\},\ 1 \le s \le r_2 + 1, \tag{4.13}$$

$$\|\nabla \cdot (v_2 - I^h_{X_2} v_2)\|_{0,K} \le C h^s_K |\nabla \cdot\ v_2|_{s,K},\ 0 \le s \le l_2 + 1. \tag{4.14}$$

∎

4.1. The space V^h. Define the finite element spaces over Ω:

$$X^h := X^h_1 \times X^h_2,\ M^h := \left\{ (q_1, q_2) \in M^h_1 \times M^h_2 : \int_{\Omega_1} q_1 dx + \int_{\Omega_2} q_2 dx = 0 \right\}$$

and

$$\Lambda^h := \{\mu^h \in L^2(\Gamma_I) : \mu^h|_E \in \mathcal{R}_{r_2}(E) \text{ for all } E \in \mathcal{E}_h(\Gamma_I)\}.$$

Note that, since function $\mu^h \in \Lambda^h$ does not in general vanish on $\partial\Gamma_I$,

$$\Lambda^h \not\subset \Lambda.$$

With this Λ^h define

$$V^h := \{v = (v_1, v_2) \in X^h : b_I(v, \mu) = 0 \text{ for all } \mu \in \Lambda^h\}.$$

These choices result in an approximation which is nonconforming (since $\Lambda^h \not\subset \Lambda$) and exterior (since $V^h \not\subset V$).

REMARK 4.3. *The space Λ^h is the normal trace of X^h_2 on Γ_I.*

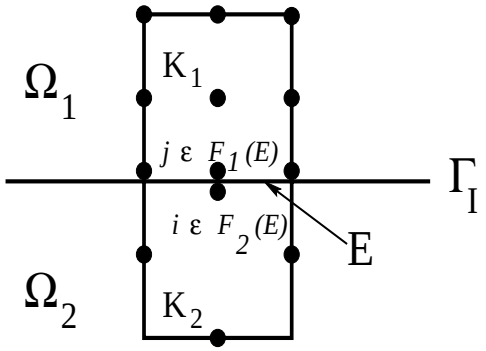


FIG. 3. *Degrees of freedom on* Γ_I.

We consider the following discrete problem: find $(u^h, p^h) \in V^h \times M^h$ satisfying

$$\begin{cases} a(u^h, v^h) + b(v^h, p^h) = \ell(v^h) & \text{for all } v^h \in V^h, \\ b(u^h, q^h) = g(q^h) & \text{for all } q^h \in M^h. \end{cases} \tag{4.15}$$

This is the natural discretization of (2.5). Since $V^h \not\subset V$, conservation of mass across Γ_I holds only in an approximate sense.

It is important to understand in exactly what sense mass conservation across Γ_I holds. To this end, a local characterization of the functions $v = (v_1, v_2) \in V^h$ is needed.

Characterization of $v = (v_1, v_2) \in V^h$. If a function $v = (v_1, v_2) \in X^h$ belongs to V^h, then the nodal values of $v_2 \cdot \hat{n}_2 \in X_2^h$ on Γ_I are linked to those of $v_1 \cdot \hat{n}_1$ on Γ_I. To be specific, let $\mathcal{F}_i$ denote the set of nodes of $X_i^h, i = 1, 2$, and $\mathcal{F}_i(E)$ the set of nodes $j \in \mathcal{F}_i$ belonging to an element face E, and let $\phi_j^{(i)},\ j \in \mathcal{F}_i$ $(i = 1, 2)$, be the associated basis functions of X_i^h. Let $E \in \mathcal{E}_h(\Gamma_I)$ be an element face on Γ_I associated with elements $K_1 \subset \Omega_1$ and $K_2 \subset \Omega_2$,

$$E \in \mathcal{E}(K_1) \cap \mathcal{E}(K_2),\ K_i \in \Omega_i,$$

as depicted in Figure 3.

From the construction of the basis functions, we have for $v = (v_1, v_2) \in X^h$

$$v_i \cdot \hat{n}_i|_E = \sum_{j \in \mathcal{F}_i(E)} (v_j^{(i)} \phi_j^{(i)}) \cdot \hat{n}_i, \qquad i = 1, 2, \tag{4.16}$$

where $v_j^{(i)} \in \mathbb{R}$ are the nodal values of v_i. By (4.10)

$$\dim(R_{r_2}(E)) = \text{ cardinality}(\mathcal{F}_2(E))$$

so that there is a one-to-one *correspondence between nodes* $i \in \mathcal{F}_2(E)$ *and basis functions* $\lambda_{E,i} \in R_{r_2}(E)$ *such that*

$$R_{r_2}(E) = \text{ span } \{\lambda_{E,i} :\ i \in \mathcal{F}_2(E)\}. \tag{4.17}$$

Consider a degree of freedom associated with a node $i \in \mathcal{F}_2(E)$ that is precisely the nodal functional

$$N_i^{(2)}(v_2) := |E|^{-1} \langle v_2 \cdot \hat{n}_2,\ \lambda_{E,i} \rangle_E, \qquad |E| = \text{ measure } (E). \tag{4.18}$$

The basis functions are, by construction, dual with respect to these functionals:

$$N_i^{(2)}(\phi_j^{(2)}) = \delta_{ij} \qquad \text{for all } i,j \in \mathcal{F}_2. \tag{4.19}$$

From (4.18), (4.19), and the formula (4.16) for $v_i \cdot \hat{n}_i|_E$, we get

$$v_i^{(2)} = |E|^{-1}\langle v_2 \cdot \hat{n}_2,\ \lambda_{E,i}\rangle_E \quad \text{for all } i \in \mathcal{F}_2(E), E \in \mathcal{E}_h(\Gamma_I),\ v_2 \in X_2^h. \tag{4.20}$$

Consider the condition defining $V^h, b_I(v,\mu) = 0$ for all $\mu \in \Lambda^h$. Restricting μ to a generic basis function $\lambda_{E,i}$ for Λ^h gives

$$\langle v_2 \cdot \hat{n}_2, \lambda_{E,i}\rangle_E = -\langle v_1 \cdot \hat{n}_1, \lambda_{E,i}\rangle_E \qquad \text{for all } i \in \mathcal{F}_2(E),\ E \in \mathcal{E}_h(\Gamma_I). \tag{4.21}$$

Combining this with (4.20) gives

$$v_i^{(2)} = -|E|^{-1}\langle v_1 \cdot \hat{n}_1, \lambda_{E,i}\rangle_E \qquad \text{for all } i \in \mathcal{F}_2(E), E \in \mathcal{E}_h(\Gamma_I). \tag{4.22}$$

Inserting the expression of v_1 in terms of its nodal values (4.16) into (4.22) gives the following pointwise characterization of the space $v \in V^h$.

PROPOSITION 4.1. *Let $v = (v_1, v_2) \in X^h$ be given. Then $v \in V^h$ is equivalent to the following relation between the nodal values $v_i^{(1)}$ and $v_i^{(2)}$ of v_1 and v_2 on E being satisfied:*

$$v_i^{(2)} = -|E|^{-1} \sum_{j \in \mathcal{F}_1(E)} v_j^{(1)} \langle \phi_j^{(1)} \cdot \hat{n}_1,\ \lambda_{E,i}\rangle_E \quad \text{for all } i \in \mathcal{F}_2(E),\ \ E \in \mathcal{E}_h(\Gamma_I). \tag{4.23}$$

∎

REMARK 4.4. *The relation* (4.23) *can be interpreted to mean that the nodes*

$$i \in \bigcup_{E \in \mathcal{E}_h(\Gamma_I)} \mathcal{F}_2(E)$$

are "hanging nodes" in that values of the function $v \in V^h$ are determined by the corresponding values at the nodes $j \in \cup_{E \in \mathcal{E}_h(\Gamma_I)}\ \mathcal{F}_1(E)$.

4.2. Inf-sup conditions for the coupled problem. The discrete formulation (4.15) leads to the question of an inf-sup condition in $V^h \times M^h$. We show next that the usual fluid's velocity-pressure discrete inf-sup condition (4.1) in fact implies the needed $V^h \times M^h$ inf-sup condition.

LEMMA 4.3. *Suppose that (X_1^h, M_1^h) satisfies the discrete* inf-sup *condition* (4.1). *Then, (V^h, M^h) is LBB-stable as well. Specifically,*

$$\inf_{q^h \in M^h} \sup_{V^h \in V^h} \frac{b(v^h, q^h)}{\|v^h\|_X\ \|q^h\|_M} \geq \beta > 0. \tag{4.24}$$

Proof. Let $q^h = (q_1^h, q_2^h) \in M^h \subset M$ be given and let $\tilde{q} \in L_0^2(\Omega)$ denote the function with $\tilde{q}|_{\Omega_i} = q_i^h$. Then it is known, e.g., [13, 15, 7], that there exists $\tilde{v} \in H^1(\Omega)^d$ with

$$\nabla \cdot \tilde{v} = -\tilde{q} \text{ in } \Omega,\ \tilde{v} = 0 \text{ on } \partial\Omega,$$

satisfying

$$\|\tilde{v}\|_{1,\Omega} \le C\|\tilde{q}\|_{0,\Omega}.$$

Define $v = (v_1, v_2) \in X$ by $v_i = \tilde{v}|_{\Omega_i},\ i = 1, 2$, so that

$$b(v, q^h) = -(\nabla \cdot \tilde{v}, \tilde{q})_\Omega = \|\tilde{q}\|^2_{0,\Omega} = \|q^h\|^2_M.$$

The above a priori bound on $\tilde{v}$ implies

$$b(v, q^h) \ge \frac{1}{C}\ \|\tilde{v}\|_{1,\Omega}\ \|q^h\|_M,$$

which implies an inf-sup condition, similar to (4.24), only over (V, M^h) rather than (V^h, M^h).

To prove the condition (4.24) over (V^h, M^h), we now construct (following Fortin's idea) an operator $\Pi^h : X_1 \times (X_2 \cap (H^1(\Omega_2))^d) \to V^h$ with

$$b(\Pi^h v - v, q^h) = 0 \text{ for all } q^h \in M^h \text{ and } \|\Pi^h v\|_X \le C\|\tilde{v}\|_{1,\Omega}.$$

Indeed, if such an operator exists, then we have

$$\frac{1}{C}\|q^h\|_M \le \frac{b(v, q^h)}{\|\tilde{v}\|_{1,\Omega}} = \frac{b(\Pi^h v, q^h)}{\|\tilde{v}\|_{1,\Omega}} \le \frac{b(\Pi^h v, q^h)}{\frac{1}{C}\|\Pi^h v\|_X} \qquad \text{for all } q^h \in M^h,$$

which would prove (4.24).

Let $\Pi^h v = (\Pi^h_1 v, \Pi^h_2 v) \in X^h_1 \times X^h_2$. To define Π^h_1, note that since (X^h_1, M^h_1) is LBB-stable, by Lemma 1.1 in Chapter II section 1.1 of [15], there exists an operator $i^h_1 : X_1 \to X^h_1$ satisfying, for all $v_1 \in X_1$,

$$b_1(i^h_1 v_1 - v_1, q^h_1) = 0 \qquad \text{for all } q^h_1 \in M^h_1$$

and

$$\|i^h_1 v_1\|_{X_1} \le C\|v_1\|_{X_1}.$$

Thus, define

$$\Pi^h_1 v := i^h_1 v_1 \in X^h_1.$$

Next, construct a $w_2 \in (H^1(\Omega_2))^d$ with

$$\begin{cases} \nabla \cdot w_2 = \nabla \cdot v_2 \text{ in } \Omega_2, \\ w_2 = 0 \text{ on } \Gamma_2 \text{ and } w_2 = \Pi^h_1 v \text{ on } \Gamma_I. \end{cases} \tag{4.25}$$

Indeed, let $g \in L^2(\partial\Omega_2)$ be given by

$$g = \begin{cases} 0 \text{ on } \Gamma_2, \\ \Pi^h_1 v \text{ on } \Gamma_I. \end{cases}$$

Since $\Pi^h_1 v = 0$ on $\partial\Gamma_I, \Pi^h_1 v \in H^{1/2}_{00}(\Gamma_I)^d$. Thus, $g \in H^{1/2}(\partial\Omega_2)^d$ and

$$\begin{aligned} \|g\|_{1/2,\partial\Omega_2} &\le C\|\Pi^h_1 v\|_{1/2,\Gamma_I} \le C\|\Pi^h_1 v\|_{1/2,\partial\Omega_1} \\ &\le C\|\Pi^h_1 v\|_{1,\Omega_1} \le C\|i^h_1 v_1\|_{X_1} \le C\|v_1\|_{1,\Omega_1}. \end{aligned}$$

Thus, there exists an extension $z \in H^1(\Omega_2)^d$ with

$$z = g \text{ on } \partial\Omega, \|z\|_{1,\Omega_2} \leq C\|g\|_{1/2,\partial\Omega_2} \leq C\|v_1\|_{1,\Omega_1}.$$

Next, write $w_2 = z + w_0$, where w_0 satisfies

$$\nabla \cdot w_0 = \nabla \cdot (v_2 - z) \text{ in } \Omega_2, w_0 = 0 \text{ on } \partial\Omega_2.$$

The solution to this problem $w_0 \in H^1(\Omega)^d$ exists [15] and satisfies

$$\begin{aligned}\|w_0\|_{1,\Omega_2} &\leq C\|\nabla \cdot (v_2 - z)\|_{0,\Omega_2} \leq C(\|v_2\|_{1,\Omega_2} + \|z\|_{1,\Omega_2}) \\ &\leq C\{\|v_2\|_{1,\Omega_2} + \|v_1\|_{1,\Omega_1}\} \leq C\|\tilde{v}\|_{1,\Omega}.\end{aligned}$$

The function w_2, so constructed, satisfies (4.25) and

$$\|w_2\|_{1,\Omega_2} \leq C\|\tilde{v}\|_{1,\Omega}. \tag{4.26}$$

Finally, define $\Pi_2^h v$ as the finite element (quasi) interpolant of $w_2 \in X_2$,

$$\Pi_2^h v := I_{X_2}^h w_2 \in X_2^h.$$

From the assumed properties of $I_{X_2}^h$, (4.14) with $s = m = 1$, we get

$$\|I_{X_2}^h w_2\|_{1,K} \leq C\|w_2\|_{1,K},$$

so that (squaring and summing over $K \in \mathcal{T}_2^h$)

$$\begin{aligned}\|I_{X_2}^h w_2\|_{X_2}^2 &= \sum_{K \in \mathcal{T}_2^h} \{\|I_{X_2}^h w_2\|_{0,K}^2 + \|\nabla \cdot I_{X_2}^h w_2\|_{0,K}^2\} \\ &\leq C\|w_2\|_{1,\Omega_2}^2.\end{aligned}$$

This with (4.26) gives

$$\|\Pi_2^h v\|_{X_2} \leq C\|\tilde{v}\|_{1,\Omega},$$

which is one of the two required conditions on Π^h. Next, we show

$$b(\Pi^h v - v, q^h) = 0 \qquad \text{for all } q^h \in M^h.$$

Let $q^h = (q_1^h, q_2^h) \in M^h$. Then, for all $K \in \mathcal{T}_2^h$, $q_2^h|_K \in \mathcal{P}_{r_2}(K)$. We thus get from (4.7) and (4.25) that

$$(\nabla \cdot \Pi_2^h v, q_2^h) = (\nabla \cdot I_{X_2}^h w_2, q_2^h)_K = (\nabla \cdot w_2, q_2^h)_K = (\nabla \cdot v_2, q_2^h)_K.$$

Thus, by summing over K, we get

$$b_2(\Pi_2^h v, q_2^h) = b_2(v_2, q_2^h) \qquad \text{for all } q_2^h \in M_2^h. \tag{4.27}$$

Now, let $E \in \mathcal{E}_h(\Gamma_I)$ be an element face on the interface and let $\mu \in R_{r_2}(E)$. Then, (4.9) in Lemma 4.2 implies (noting that $\Pi_2^h v = I_{X_2}^h w_2$)

$$\langle \Pi_2^h v \cdot \hat{n}_2, \mu\rangle_E = \langle I_{X_2}^h w_2 \cdot \hat{n}_2, \mu\rangle_E = \langle w_2 \cdot \hat{n}_2, \mu\rangle_E = \langle \Pi_1^h v \cdot \hat{n}_2, \mu\rangle_E,$$

where the fact that $w_2 = \Pi_1^h v$ on Γ_I (see (4.25)) was used. Thus

$$\langle \Pi_1^h v \cdot \hat{n}_1 + \Pi_2^h v \cdot \hat{n}_2, \mu \rangle_E = 0 \qquad \text{for all } \mu \in R_{r_2}(E).$$

The definition of Λ^h and summing over $E \subset \Gamma_I$ now implies that

$$\langle \Pi_1^h v \cdot \hat{n}_1 + \Pi_2^h v \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} = 0 \qquad \text{for all } \mu^h \in \Lambda^h. \tag{4.28}$$

In other words, $\Pi^h v = (\Pi_1^h v_1, \Pi_2^h v_2) \in V^h$. Since we have shown

$$b_j(\Pi_j^h v, q_j^h) = b_j(v_j, q_j^h), \ j = 1, 2,$$

it follows that

$$b(\Pi^h v, q^h) = b(v, q^h),$$

completing the proof. □

4.3. Approximation of the coupled problem in V^h. The finite element spaces X_1^h and X_2^h are well understood so the approximation properties of $X^h = X_1^h \times X_2^h$ are known and asymptotically optimal. On the other hand, the finite element space arising in the error analysis is V^h rather than X^h. If $X^h \times \Lambda^h$ satisfied a discrete inf-sup condition similar to (3.13), then the abstract theory of mixed methods [15, 7] would imply that the error in approximation in V^h would be comparable to that in $X^h \times \Lambda^h$. However, $\Lambda^h \not\subset \Lambda$ since functions in Λ^h do not vanish at $\partial\Gamma_I$ (a key condition in the continuous case). Therefore, we do not, in general, expect this discrete inf-sup condition to hold.

Thus, the approximation properties of

$$V^h = \{v^h \in X^h : \langle v_1^h \cdot \hat{n}_1 + v_2^h \cdot \hat{n}_2, \mu \rangle_{\Gamma_I} = 0 \text{ for all } \mu \in \Lambda^h\}$$

must be delineated by a direct construction. Herein, we shall construct an interpolation operator

$$I^h := W \to V^h,$$

where W is a subspace of V of sufficiently smooth functions. To that end, we choose s_i sufficiently large and define W as follows:

$$\begin{aligned} W := \{v = (v_1, v_2) \in X : v_i \in W_i := X_i \cap (H^{s_i}(\Omega_i))^d, \ i = 1, 2, \\ \text{and } v_1 \cdot \hat{n}_2|_{\Gamma_I} = v_2 \cdot \hat{n}_2|_{\Gamma_I} \text{ in } L^2(\Gamma_I)\}. \end{aligned} \tag{4.29}$$

The construction of I^h will be based on the finite element interpolation operators: $I_{X_i}^h : W_i \to X_i^h$ $(i = 1, 2)$. Define $I^h = (I_1^h v, I_2^h v) \in V^h$ via

$$I_1^h v = I_{X_1}^h v_1 \in X_1^h, \qquad I_2^h v = I_{X_2}^h v_2 - \delta_2^h \in X_2^h,$$

where the (small) correction $\delta_2^h \in X_2^h$ is chosen to enforce in a discrete sense continuity of the normal velocities across Γ_I in (4.29).

Construction of the correction δ_2^h enforcing $I^h v \in V^h$. By the choice of $I_{X_2}^h$ and Λ^h we get the following relation for all $\mu^h \in \Lambda^h$:

$$\begin{aligned} &\langle I_1^h v \cdot \hat{n}_1 + I_2^h v \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} \\ &\quad = -\langle I_{X_1}^h v_1 \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} + \langle v_2 \cdot \hat{n}_2, \mu_h \rangle_{\Gamma_I} - \langle \delta_2^h \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} \\ &\quad = \langle (v_1 - I_{X_1}^h v_1) \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} - \langle \delta_2^h \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I}. \end{aligned} \tag{4.30}$$

To construct δ_2^h we shall first construct $\delta_2 \in X_2 \cap (H^1(\Omega_2))^d$ such that

$$\delta_2 = v_1 - I_{X_1}^h v_1 \text{ on } \Gamma_I, \text{ and } \|\delta_2\|_{1,\Omega_2} \le C|v_1 - I_{X_1}^h v_1|_{1,\Omega_1}. \tag{4.31}$$

To this end, let g_2 extend $v_1 - I_{X_1}^h v_1$ by zero to $\partial\Omega_2$:

$$g_2 := \begin{cases} v_1 - I_{X_1}^h v_1 \text{ on } \Gamma_I, \\ 0 \text{ on } \Gamma_2 = \partial\Omega_2 \setminus \Gamma_I. \end{cases}$$

Since $(v_1 - I_{X_1}^h v_1) = 0$ on $\partial\Gamma_I, (v_1 - I_{X_1}^h v_1) \in H_{00}^{1/2}(\Gamma_I)$ so $g_2 \in H^{1/2}(\partial\Omega_2)^d$. Further, we have the bound

$$\begin{aligned} \|g_2\|_{1/2,\partial\Omega_2} &\le C\|v_1 - I_{X_1}^h v_1\|_{1/2,\Gamma_I} \le C\|v_1 - I_{X_1}^h v_1\|_{1/2,\partial\Omega_1} \\ &\le C\|v_1 - I_{X_1}^h v_1\|_{1,\Omega_1} \le C|v_1 - I_{X_1}^h v_1|_{1,\Omega_1}. \end{aligned}$$

Since $H^{1/2}(\partial\Omega_2)^d$ is the range of the trace operator on $H^1(\Omega_2)^d$, we can find a $\delta_2 \in H^1(\Omega_2)^d$ extending g_2 onto Ω_2 and satisfying

$$\|\delta_2\|_{1,\Omega_2} \le C\|g_2\|_{1/2,\partial\Omega_2} \le C|v_1 - I_{X_1}^h v_1|_{1,\Omega_1}.$$

Define δ_2^h as the interpolant of this extension:

$$\delta_2^h := I_{X_2}^h \delta_2. \tag{4.32}$$

The property (4.9) of $I_{X_2}^h(\cdot)$ implies that for $\mu^h \in \Lambda^h$

$$\langle \delta_2^h \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} = \langle \delta_2 \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} = \langle (v_1 - I_{X_1}^h v_1) \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I}.$$

Combining this with (4.30) gives

$$\langle I_1^h v \cdot \hat{n}_1 + I_2^h v \cdot \hat{n}_2, \mu^h \rangle_{\Gamma_I} = 0 \qquad \text{for all } \mu^h \in \Lambda^h, \tag{4.33}$$

implying that $(I_1^h v, I_2^h v) \in V^h$. Thus, this completes the construction of $I^h : W \to V^h$. We shall need an estimate of the correction term $\|\delta_2^h\|_{X_2}$ developed as follows.

From the interpolation error estimates we get, for every $K \in \mathcal{T}_2^h$,

$$\|\delta_2^h\|_{1,K} \le \|\delta_2\|_{1,K} + \|\delta_2 - I_{X_2}^h \delta_2\|_{1,K} \le C\|\delta_2\|_{1,K}.$$

Thus, (summing over $K \subset \Omega_2$)

$$\|\delta_2^h\|_{X_2} = \left\{ \|\delta_2^h\|_{0,\Omega_2}^2 + \|\nabla \cdot \delta_2^h\|_{0,\Omega_2}^2 \right\}^{1/2} \le \left\{ \sum_{K \in \mathcal{T}_2^h} \|\delta_2^h\|_{1,K}^2 \right\}^{1/2},$$

which implies

$$\|\delta_2^h\|_{X_2} \le C\|\delta_2\|_{1,\Omega_2} \le C|v_1 - I_{X_1}^h v_1|_{1,\Omega_1}. \tag{4.34}$$

Bound (4.34) now gives interpolation error estimates for $I_1^h v = I_{X_1}^h v_1$ and $I_2^h v = I_{X_2}^h v_2 - \delta_2^h$:

$$\begin{aligned} \|v - I^h v\|_X &\le |v_1 - I_1^h v|_{1,\Omega_1} + \|v_2 - I_2^h v\|_{X_2} \\ &\le C|v_1 - I_{X_1}^h v_1|_{1,\Omega_1} + \|v_2 - I_{X_2}^h v_2\|_{X_2}. \end{aligned} \tag{4.35}$$

Combining these with the estimates for $I^h_{X_j}$ (see (4.3)),

$$|v_1 - I^h_{X_1} v_1|_{1,\Omega_1} \le C \left\{ \sum_{K \in \mathcal{T}_1^h} \left(h_K^{r_1} |v_1|_{r_1+1,\delta(K)}\right)^2 \right\}^{1/2},$$

$$\|v_2 - I^h_{X_2} v_2\|_{X_2} \le C \left\{ \sum_{K \in \mathcal{T}_2^h} \left(h_K^{r_2+1} (|v_2|_{r_2+1,K} + |\nabla \cdot v_2|_{r_2+1,K})\right)^2 \right\}^{1/2},$$

and using (4.4) and the fact that an element $\widetilde{K}$ can belong at most to a finite number $n(\widetilde{K}) \le C$ of local patches $\delta(K)$ leads to the following result.

PROPOSITION 4.2. *For all $v \in W \subset V$ (given by (4.29)), the interpolation operator $I^h : W \to V^h$ satisfies*

$$\|v - I^h v\|_X \le C \left\{ \sum_{K \in \mathcal{T}_1^h} (h_K^{r_1} |v_1|_{r_1+1,K})^2 + \sum_{K \in \mathcal{T}_2^h} \left(h_K^{r_2+1} \left(|v_2|_{r_2+1,K} + |\nabla \cdot v_2|_{r_2+1,K}\right)\right)^2 \right\}^{1/2}. \qquad \square$$

4.4. Discretization error estimates. Since, as noted above, $\Lambda^h \not\subset \Lambda$ and $V^h \not\subset V$, the associated discretizations of *either* saddle point formulations contain an extra *consistency error* which must be estimated using the earlier constructions. Indeed, the abstract error estimates from Brezzi and Fortin [7, Chap. II, sect. 2.6, Proposition 2.16] give the following.

LEMMA 4.4. *Let $(u, p) \in V \times M$ be the solution of the weak formulation (2.5) of the coupled problem. Let $(u^h, p^h) \in V^h \times M^h$ be the solution of the discrete problem (4.15). Let the finite element spaces be chosen as in subsection 4.1, satisfying LBB-stability (subsection 4.2) and approximability (subsection 4.3). Then,*

$$\|u - u^h\|_X + \|p - p^h\|_M \le C \left\{ \inf_{v^h \in V^h} \|u - v^h\|_X + \inf_{q^h \in M^h} \|p - q^h\|_M \right\} + \mathcal{H}^h,$$

where

$$\mathcal{H}^h := \sup_{v^h \in V^h} \frac{|a(u, v^h) + b(v^h, p) - \ell(v^h)|}{\|v^h\|_X}$$

is the consistency error. $\square$

The error analysis thus depends on obtaining a bound on the consistency error term $\mathcal{H}^h$. To this end, suppose the weak solution (u, p) to the coupled problem is smooth enough (to be made precise soon) and that $\lambda \in H^s(\Gamma_I)$ (for some s depending on the smoothness of (u, p)), where λ is defined in (2.2).

The variational formulation (2.4) of (u, p, λ) in (X, M, Λ) implies that

$$a(u, v^h) + b(v^h, p) + \langle \lambda, v_1^h \cdot \hat{n}_1 + v_2^h \cdot \hat{n}_2 \rangle_{\Gamma_I} = \ell(v^h) \qquad \text{for all } v^h \in X^h.$$

Thus, if we define the consistency error functional

$$\theta(v^h) := a(u, v^h) + b(v^h, p) - \ell(v^h), \qquad v^h \in X^h,$$

it follows that

$$\theta(v^h) = -\langle v_1^h \cdot \hat{n}_1 + v_2^h \cdot \hat{n}_2, \lambda \rangle_{\Gamma_I} \qquad \text{for all } v^h \in V^h \subset X^h.$$

LEMMA 4.5 (consistency error estimate). *For all $v^h \in V^h$, there holds*

$$|\theta(v^h)| \le C \left\{ \sum_{E \in \mathcal{E}_h(\Gamma_i)} (h_E^s |\lambda|_{s,E})^2 \right\}^{1/2} \|v^h\|_X, \tag{4.36}$$

for $0 \le s \le r_2 + 1$.

Proof. Let $\mu^h \in \Lambda^h$ denote the $L^2(\Gamma_I)$ projection of λ into Λ^h. Since Λ^h consists of *discontinuous* piecewise polynomials, the orthogonality relation for μ^h holds edge by edge:

$$\langle \lambda - \mu^h, w \rangle_E = 0 \qquad \text{for all } w \in R_{r_2}(E), \text{ for all } E \in \mathcal{E}_h(\Gamma_I). \tag{4.37}$$

From the definition of V^h it follows that, for all $v^h \in V^h$,

$$\begin{aligned} \theta(v^h) &= \langle v_1^h \cdot \hat{n}_1 + v_2^h \cdot \hat{n}_2,\ \mu^h - \lambda \rangle_{\Gamma_I} \\ &= \langle v_1^h \cdot \hat{n}_1, \mu^h - \lambda \rangle_{\Gamma_I} + \sum_{E \in \mathcal{E}_h(\Gamma_I)} \langle \mu^h - \lambda, v_2^h \cdot \hat{n}_2 \rangle_E. \end{aligned}$$

By Lemma 4.2 we have that

$$w = v_2^h \cdot \hat{n}_2|_E \in R_{r_2}(E) \qquad \text{for all } E \in \mathcal{E}(K),\ K \in \mathcal{T}_2^h,$$

which implies

$$\langle \mu^h - \lambda, v_2^h \cdot \hat{n}_2 \rangle_E = 0 \qquad \text{for all } E \in \mathcal{E}_h(\Gamma_I).$$

Thus, $\theta(v^h) = \langle v_1^h \cdot \hat{n}_1, \mu^h - \lambda \rangle_{\Gamma_I}$, for all $v^h \in V^h$, and it follows that

$$\begin{aligned} |\theta(v^h)| &\le \sum_{E \in \mathcal{E}_h(\Gamma_I)} \|v_1^h\|_{0,E} \|\lambda - \mu^h\|_{0,E} \\ &\le \left(\sum_{E \in \mathcal{E}_h(\Gamma_I)} \|\lambda - \mu^h\|_{0,E}^2 \right)^{1/2} \|v_1^h\|_{0,\Gamma_I}. \end{aligned} \tag{4.38}$$

By the trace theorem and the Poincaré–Friedrichs inequality,

$$\|v_1^h\|_{0,\Gamma_I} \le C \|v^h\|_X.$$

Since μ^h is the $L^2(E)$ projection of λ into $R_{r_2}(E)$ by (4.37), it follows that

$$\|\lambda - \mu^h\|_{0,E} \le C h_E^s |\lambda|_{s,E}, \qquad \text{for } 0 \le s \le r_2 + 1,\ E \in \mathcal{E}_h(\Gamma_I).$$

Using the last two bounds in (4.38) completes the proof. ∎

Lemma 4.4 immediately yields a bound on the consistency error term $\mathcal{H}^h$.

COROLLARY 4.1. *There holds*

$$\mathcal{H}^h \le C \left\{ \sum_{E \in \mathcal{E}_h(\Gamma_I)} (h_E^s |\lambda|_{s,E})^2 \right\}^{1/2}, \quad \textit{for } 0 \le s \le r_2 + 1. \qquad \blacksquare$$

This bound can now be used in the abstract error estimate in Lemma 4.3 to yield a convergence theorem.

THEOREM 4.1. *Let the weak solution (u,p) to (2.5) be sufficiently smooth (that the norms in (4.39) are finite). Let $(u^h,p^h)\in V^h\times M^h$ be the finite element approximation to (u,p). Then,*

$$\begin{aligned}\|u-u^h\|_X+\|p-p^h\|_M \le C\Bigg\{&\Bigg\{\sum_{K\in\mathcal{T}_1^h}\left(h_K^{s_1}(|u_1|_{s_1+1,K}+|p_1|_{s_{1,K}})\right)^2\Bigg\}^{1/2}\\ &+\Bigg\{\sum_{K\in\mathcal{T}_2^h}\left(h_K^{\tilde{s}_2}|u_2|_{\tilde{s}_2,K}+h_K^{s_2}(|p_2|_{s_2,K}+|\nabla\cdot u_2|_{s_2,K})\right)^2\Bigg\}^{1/2}\\ &+\Bigg\{\sum_{E\in\mathcal{E}_h(\Gamma_I)}(h_E^{s_2}|\lambda|_{s_2,E})^2\Bigg\}^{1/2}\Bigg\},\\ &1\le s_1\le r_1,\ 1\le\tilde{s}_2\le r_2+1,\ 0\le s_2\le l_2+1.\end{aligned}\tag{4.39}$$

∎

REMARK 4.5. *Theorem 4.1 implies optimal error bounds in both the fluid region and in the porous medium region.*

REMARK 4.6. *We have just learned of the concurrent work of Discacciati, Miglio, and Quarteroni [11] on a closely related problem. They consider Stokes–Darcy coupling with a free slip condition on Γ_I (i.e., $\alpha_1=0$ in (1.6)) and the formulation of the Darcy model as a Poisson problem rather than as a mixed method, and they obtain interesting results.*

REFERENCES

[1] P. ANGOT, *Analysis of singular perturbations on the Brinkman problem for fictious domain models of viscous flows*, Math. Methods Appl. Sci., 22 (1999), pp. 1395–1412.

[2] D. N. ARNOLD, F. BREZZI, AND M. FORTIN, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[3] G. BEAVERS AND D. JOSEPH, *Boundary conditions at a naturally impermeable wall*, J. Fluid. Mech, 30 (1967), pp. 197–207.

[4] F. BREZZI, J. DOUGLAS, JR., R. DURÀN, AND M. FORTIN, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[5] F. BREZZI, J. DOUGLAS, JR., M. FORTIN, AND L. D. MARINI, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[6] F. BREZZI, J. DOUGLAS, JR., AND L. D. MARINI, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.

[7] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer–Verlag, New York, 1991.

[8] Z. CHEN AND J. DOUGLAS, JR., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[9] P. CLÉMENT, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[10] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[11] M. DISCACCIATI, E. MIGLIO, AND A. QUARTERONI, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Mathematics, 43 (2002), pp. 57–74.

[12] R. S. FALK, *Nonconforming finite element methods for the equations of linear elasticity*, Math. Comput., 57 (1991), pp. 529–550.

[13] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. I, Springer–Verlag, New York, 1994.
[14] D. Gartling, C. Hickox, and R. Givler, *Simulation of coupled viscous and porous flow problems*, Comp. Fluid Dyn., 7 (1996), pp. 23–48.
[15] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer–Verlag, Berlin, 1986.
[16] W. Jäger and A. Mikelić, *On the boundary condition at the interface between a porous medium and a free fluid*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 403–465.
[17] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.
[18] I. P. Jones, *Low Reynolds number flow past a porous spherical shell*, Proc. Camb. Phil. Soc., 73 (1973), pp. 231–238.
[19] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer–Verlag, New York-Heidelberg, 1972.
[20] T. P. Mathew, *Domain Decomposition and Iterative Refinement Methods for Mixed Finite Element Discretizations of Elliptic Problems*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, New York, 1989.
[21] J. C. Nedelec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[22] L. E. Payne and B. Straughan, *Analysis of the boundary condition at the interface between a viscous fluid and a porous medium and related modelling questions*, J. Math. Pures Appl. (9), 77 (1998), pp. 317–354.
[23] V. Prasad, *Convective flow interaction and heat transfer between fluid and porous layers*, in Convective Heat and Mass Transfer in Porous Media, S. Kakac et al., eds., Kluwer, Amsterdam, 1991, pp. 563–615.
[24] R. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer–Verlag, New York, 1977, pp. 292–315.
[25] P. Saffman, *On the boundary condition at the surface of a porous media*, Stud. Appl. Math., 50 (1971), pp. 93–101.
[26] A. Salinger, R. Aris, and I. Derby, *Finite element formulations for large-scale, coupled flows in adjacent porous and open fluid domains*, Internat. J. Numer. Methods Fluids, 18 (1994), pp. 1185–1209.
[27] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[28] J. Serrin, *Principles of Classical Fluid Mechanics*, in Handbuch der Physik, B. 8/1, Springer–Verlag, Berlin, 1959, pp. 125–263.
[29] C. Taylor and P. Hood, *A numerical solution of the Navier-Stokes equations using the finite element technique*, Internat. J. Comput. & Fluids, 1 (1973), pp. 73–100.
[30] J. M. Thomas, *Sur lánalyse numérique des méthodes d'éléments finis hybrides et mixtes, Thèse de Doctorat d'etat*, Universite Pierre et Marie Curie (Paris 6), Paris, 1977.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2219–2239

FLEXIBLE INNER-OUTER KRYLOV SUBSPACE METHODS*

VALERIA SIMONCINI† AND DANIEL B. SZYLD‡

Abstract. Flexible Krylov methods refers to a class of methods which accept preconditioning that can change from one step to the next. Given a Krylov subspace method, such as CG, GMRES, QMR, etc. for the solution of a linear system $Ax = b$, instead of having a fixed preconditioner M and the (right) preconditioned equation $AM^{-1}y = b$ ($Mx = y$), one may have a different matrix, say M_k, at each step. In this paper, the case where the preconditioner itself is a Krylov subspace method is studied. There are several papers in the literature where such a situation is presented and numerical examples given. A general theory is provided encompassing many of these cases, including truncated methods. The overall space where the solution is approximated is no longer a Krylov subspace but a subspace of a larger Krylov space. We show how this subspace keeps growing as the outer iteration progresses, thus providing a convergence theory for these inner-outer methods. Numerical tests illustrate some important implementation aspects that make the discussed inner-outer methods very appealing in practical circumstances.

Key words. flexible or inner-outer Krylov methods, variable preconditioning, nonsymmetric linear system, iterative solver

AMS subject classifications. 65F10, 15A06

PII. S0036142902401074

1. Introduction.

Consider the iterative solution of large sparse (symmetric or) nonsymmetric linear systems of equations of the form

$$Ax = b. \tag{1.1}$$

In recent years, several authors studied Krylov subspace methods with variable (or flexible) preconditioning, i.e., preconditioning with a different (possibly nonlinear) operator at each iteration of a Krylov subspace method. These include [1], [17], [27], [29], [36], [37], and [39]. The usual (right) preconditioning consists of replacing (1.1) by

$$AM^{-1}y = b, \quad \text{with} \quad Mx = y, \tag{1.2}$$

for a suitable preconditioner M. One of the motivations for methods with variable preconditioners is the need to solve each preconditioning equation

$$Mz = v \tag{1.3}$$

only inexactly, as is done, e.g., in [12], using multigrid or, in [40], using a two-stage preconditioner, one of which is inexact; see also [4]. This implies that we have (implicitly) a different M at the kth step of the Krylov method. One can also consider preconditioners which might improve using information from previous iterations; cf. [2], [13], [21].

*Received by the editors January 16, 2002; accepted for publication (in revised form) July 16, 2002; published electronically January 7, 2003.

http://www.siam.org/journals/sinum/40-6/40107.html

†Dipartimento di Matematica, Università di Bologna, I-40126 Bologna, Italy and Istituto di Matematica Applicata e Tecnologie Informati che del CNR, I-27100 Pavia, Italy (val@dragon.imati.cnr.it).

‡Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA 19122-6094 (szyld@math.temple.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-0207525. Part of this paper was prepared while this author visited the Università di Bologna, within the 2001 I.N.d.A.M. project "Problemi di perturbazione singolare: tecniche adattive e di decomposizione dei dominii."

Experiments have been reported in the literature, where the preconditioner in (1.3) is itself a Krylov subspace method. For example, some versions of GMRESR [37] fit this description [6]. We refer to [39] for a more detailed analysis specific to GMRESR versus FGMRES. In [5], [29], one has GMRES for the preconditioner, or inner iterations, and FGMRES as the (outer) flexible method. In [36], QMR is the preconditioner and FQMR is the outer method. In all these cases, we can say that at the kth outer step (or cycle) the inner iterative method is stopped after a certain number m_k of (inner) iterations. This number is either fixed a priori, such as in [5], [29], [37], or is the consequence of some stopping criteria often involving the inner residual [1], [17], [36], [38].

In this paper we analyze the convergence theory of these inner-outer methods, i.e., flexible Krylov subspace methods preconditioned by (a possibly different) Krylov subspace method. The resulting class of methods can be characterized as those choosing at the kth cycle the approximate solution to (1.1) from a particular k-dimensional subspace of a larger Krylov subspace; see sections 2–5. We show that, as the method progresses, the dimensions of this subspace keeps growing, and thus the method converges. This finite termination property is not available for restarted methods, such as GMRES(m). In other words, by restricting ourselves to preconditioners which are Krylov methods, we maintain the global iteration within a larger Krylov subspace.

An alternative view of these inner-outer methods is to consider a Krylov method with polynomial preconditioning (see, e.g., [2], [11], [15], [20], [21], [28], [30]), where a new polynomial is implicitly chosen at each cycle by a Krylov method; cf. [14], [37], [38].

Our approach is very general, and thus our analysis applies to a variety of inner-outer methods. In particular, it applies to truncated Krylov methods, such as DQGMRES [32], when preconditioned with a Krylov subspace method. Several authors have suggested strategies on how to choose which vectors to keep in the truncated basis; see, e.g., [7], [38]. Our results in section 5 provide a theoretical foundation for some of these empirical recommendations.

In section 6 we discuss the possibility of breakdown and stagnation of the inner-outer methods. Since the residual of the inner method is deflated with respect to the previous vectors in the outer basis, stagnation is much less prevalent than in restarted methods.

In section 7 we show in some particular cases why these methods behave better than the corresponding restarted counterparts and illustrate this with additional numerical experiments. In our (albeit limited) computational experience, even in cases where at intermediate steps the restarted methods have a lower residual norm, the inner-outer methods outperform the restarted methods. These experiments demonstrate how competitive these inner-outer methods are. They are a very good alternative to restarted methods such as GMRES(m). We hope that the theory presented here, assuring convergence, together with the numerical evidence, will encourage more practitioners to try these inner-outer methods.

In this paper we concentrate on nonsymmetric systems, but most of our observations are valid for symmetric systems as well, where flexible CG-type methods can be employed [17], [27]. We also note that in our descriptions the coefficient matrix is assumed to be A, as in (1.1), although everything applies to a preconditioned matrix AP^{-1} as in (1.2), or $P^{-1}A$, for a suitable fixed preconditioner P. In fact, some of our experiments in section 8 are of this type, where P corresponds to an incomplete factorization of A.

Finally, we note that exact arithmetic is assumed throughout the paper, and that numerical experiments were carried out using Matlab 5.3 with machine epsilon $\epsilon = 2.2 \cdot 10^{-16}$ [23].

2. General setup. Given an initial guess x_0 to the solution of (1.1), and the corresponding residual $r_0 = b - Ax_0$, a Krylov subspace method generates a basis $\{v_1, v_2, \ldots, v_m\}$ of the Krylov subspace $\mathcal{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\}$. Let $V_m = [v_1, v_2, \ldots, v_m]$. An approximation to the solution of (1.1) is sought in $x_0 + \mathcal{K}_m(A, r_0)$, i.e., of the form $x_m = x_0 + V_m y_m$ for some $y_m \in \mathbb{R}^m$. The different methods arise by different choices of the basis defined by V_m and by different choices of y_m; see, e.g., [19], [30], for detailed description of these methods.

For example, in GMRES [31], the vectors in V_m are orthonormal, produced by the Arnoldi method with $v_1 = r_0/\beta$, $\beta = \|r_0\|$, the norm of the initial residual. Thus, the following relation holds:

$$AV_m = V_{m+1}H_m, \tag{2.1}$$

where H_m is upper Hessenberg of size $(m+1) \times m$. The vector y_m is chosen to minimize the norm of the residual $r_m = b - Ax_m$, i.e., find y_m which is the minimizer of

$$\min_{y \in \mathbb{R}^m} \|r_0 - AV_m y\| = \min_{y \in \mathbb{R}^m} \|V_{m+1}(\beta e_1 - H_m y)\|, \tag{2.2}$$

where e_1 is the first Euclidean vector. Our general analysis also includes truncated methods such as DQGMRES [32] as a possible outer subspace method, where only certain vectors of V_m are kept in storage. For example, one can keep the last ℓ columns of V_m (denote the $n \times \ell$ matrix containing these by $V_{m[\ell]}$). In this case, the minimization (2.2) is replaced by the minimization of

$$\min_{y \in \mathbb{R}^m} \|\beta e_1 - H_m y\|, \tag{2.3}$$

where H_m is banded with upper semibandwidth ℓ. In QMR [16], a relation like (2.1) holds, but V_m is not an orthogonal matrix (the basis is bi-orthogonal to a basis of another Krylov subspace; these bases are obtained via a two-sided Lanczos process), the minimization of (2.3) is performed, and H_m is tridiagonal.

When a standard right preconditioner is used, as in (1.2), the expression (2.1) becomes

$$AM^{-1}V_m = V_{m+1}H_m.$$

When flexible preconditioning is used, this relation is replaced with

$$AZ_m = V_{m+1}H_m, \tag{2.4}$$

cf. (1.3), and storage has to be allocated for both the matrices Z_m and V_m, i.e., one needs approximately twice the storage of the standard preconditioned case; see, e.g., [29], [36].

In order to analyze the inner-outer method, when the outer method is a flexible Krylov subspace method and the inner is a Krylov subspace method, we consider the approximations of the solution of (1.1) at the kth (outer) cycle to be taken from the affine space $x_0 + \text{span}\{z_1, \ldots, z_k\}$, where $Z_k = [z_1, \ldots, z_k]$, i.e., of the form

$$x_k = x_0 + Z_k u_k \quad \text{for some } u_k \in \mathbb{R}^k. \tag{2.5}$$

We change the notation a bit, to make a distinction with the inner method, and rewrite (2.4) as

$$AZ_k = W_{k+1}T_k, \tag{2.6}$$

i.e., $W_k = [w_1, \ldots, w_k]$ is the matrix containing a basis of the outer subspace, and the $(k+1) \times k$ matrix T_k contains the coefficients used in the orthogonalization (or deflation). The matrix T_k is either upper Hessenberg (if orthogonalization is used via the Arnoldi method), banded (if a truncated method is used), or tridiagonal (if the two-sided Lanczos method is used).

In other words, given x_0 and the corresponding residual

$$r_0 = b - Ax_0, \qquad w_1 = r_0/\beta, \qquad \beta = \|r_0\|, \tag{2.7}$$

for each cycle k, first a new vector z_k is computed, which approximates the solution of

$$Az = w_k \tag{2.8}$$

(using an inner Krylov method). Then, the vector Az_k is computed, orthogonalized with respect to the previous vectors w_i, $i \le k$ (or with respect to the ℓ vectors kept in the truncated version), and normalized (or bi-orthogonalized with respect to some other vectors) to obtain the new vector w_{k+1}; cf. (2.6). Thus, the residual at the kth cycle is

$$r_k = b - Ax_k = r_0 - AZ_k u_k = r_0 - W_{k+1}T_k u_k = W_{k+1}(\beta e_1 - T_k u_k). \tag{2.9}$$

We point out that in the case of using an Arnoldi method in the outer scheme, orthogonalizing Az_k with respect to the vectors in W_k is equivalent to deflating the inner residual $w_k - Az_k$ with respect to the vectors in W_k, i.e., with respect to all previous inner Krylov subspace starting vectors. In fact, we have

$$\begin{aligned} t_{k+1,k}w_{k+1} &= (I - W_k W_k^T)Az_k \qquad (2.10)\\ &= w_k - (w_k - Az_k) - w_k + W_k W_k^T(w_k - Az_k)\\ &= -(I - W_k W_k^T)(w_k - Az_k), \qquad (2.11) \end{aligned}$$

where W_k^T stands for the transpose of W_k.

As we said, at the kth cycle a new vector z_k is computed, which approximates the solution of (2.8), using an inner Krylov method. The corresponding (inner) Krylov subspace is $\mathcal{K}_m(A, w_k)$, with $m = m_k$, and its basis is $\{v_1^{(k)}, \ldots, v_m^{(k)}\}$, with

$$v_1^{(k)} = w_k \tag{2.12}$$

(of unit norm). The matrix $V_m^{(k)} = [v_1^{(k)}, \ldots, v_m^{(k)}]$ satisfies (2.1), i.e.,

$$AV_m^{(k)} = V_{m+1}^{(k)} H_m^{(k)}, \tag{2.13}$$

and we have

$$z_k = V_m^{(k)} y_k \quad \text{for some } \; y_k \in \mathbb{R}^m, \; m = m_k. \tag{2.14}$$

It is important to realize that when the inner system is preconditioned from the right, this can be viewed as a global preconditioning strategy. More precisely, consider

preconditioning the inner system (2.8) from the right with a fixed matrix P, so that the inner system transforms into $AP^{-1}\hat{z} = w_k$, with $z = P^{-1}\hat{z}$. An approximation to $\hat{z}$ is determined in $\mathcal{K}_m(AP^{-1}, w_k)$, from which an approximation to z can be readily recovered. Note that we can write $\hat{z}_k = V_m^{(k)} y_k$, where $V_m^{(k)}$ is now a basis for $\mathcal{K}_m(AP^{-1}, w_k)$. Let $Z_k = P^{-1}\hat{Z}_k = [P^{-1}\hat{z}_1, \ldots, P^{-1}\hat{z}_k]$. Relation (2.6) transforms into

$$AP^{-1}\hat{Z}_k = W_{k+1}T_k,$$

which immediately shows that inner preconditioning corresponds to applying the flexible method to the system $AP^{-1}\hat{x} = b$, with $\hat{x} = Px$. In light of these considerations, from now on we shall work with the coefficient matrix A, where A could actually stand for any preconditioned matrix AP^{-1}.

The description of flexible inner-outer methods by (2.5), (2.6), (2.13), and (2.14) is pretty general, and many of our results, including the following proposition, apply to this general setting.

PROPOSITION 2.1. *Each new vector of the outer basis w_{k+1} is a linear combination of the orthogonal projections of the columns $V_{m+1}^{(k)}$ (i.e., of the basis of the inner space $\mathcal{K}_{m+1}(A, w_k)$), onto the (bi-)orthogonal complement of $\mathcal{R}(W_k)$ (or $\mathcal{R}(W_{k[\ell]})$).*

Proof. From (2.14) and (2.13), we have $Az_k = AV_m^{(k)} y_k = V_{m+1}^{(k)}(H_m^{(k)} y_k)$, and from (2.10) the proposition follows. ∎

Note that while for each of the cycles we have the columns of $V_m^{(k)}$ being a basis of a Krylov subspace, neither $\mathcal{R}(W_k)$ nor $\mathcal{R}(Z_k)$, the range of W_k or Z_k, respectively, is a Krylov subspace. This is in contrast to the standard (nonflexible) preconditioned case. As pointed out, e.g., in [9], minimal residual or orthogonal residual methods with these bases would converge, as long as $\dim \mathcal{R}(W_k) = k$ and $\dim \mathcal{R}(Z_k) = k$, i.e., as long as the new vectors z_k and w_k are linearly independent of the previous ones so that the subspaces keep growing. As we shall see, this is the case for the inner-outer methods studied here.

We can say more; the columns of Z_k and those of W_k are bases of (different) subspaces (of dimension k) of a larger Krylov subspace generated by the initial residual r_0.

LEMMA 2.2. *Assume that Z_k and W_{k+1} are full column rank. Then, $\mathcal{R}(Z_k) \subset \mathcal{K}_{q-1}(A, r_0)$, and $\mathcal{R}(W_{k+1}) \subset \mathcal{K}_q(A, r_0)$, where $q = q_k = p_k + k$ and p_k is given by $p = p_k = \sum_{j=1}^{k} m_j$.*

Proof. We use induction on k. From (2.7), $\operatorname{span}\{w_1\} = \operatorname{span}\{r_0\}$. Therefore $z_1 \in \mathcal{K}_{m_1}(A, w_1) = \mathcal{K}_{m_1}(A, r_0)$ and $w_2 \in \operatorname{span}\{Az_1, w_1\} \subset \mathcal{K}_{m_1+1}(A, r_0)$, so that $\mathcal{R}(W_2) \subset \mathcal{K}_q(A, r_0)$ with $q = p_1 + 1 = m_1 + 1$. Assume that the assertions hold for $k-1$, and thus $w_k \in \mathcal{R}(W_k) \subset \mathcal{K}_q(A, r_0)$, with $q = q_{k-1}$. From (2.12), it follows that z_k belongs to the inner Krylov subspace $\mathcal{K}_m(A, w_k)$ $(m = m_k)$, which is a subspace of $\mathcal{K}_q(A, r_0)$ with $q = q_{k-1} + m_k = p_{k-1} + k - 1 + m_k = q_k - 1$. We then have that $w_{k+1} \in \operatorname{span}\{Az_k, W_k\} \subset \mathcal{K}_q(A, r_0)$, with $q = q_k$. ∎

This lemma also applies to truncated Krylov methods, since the vectors kept in storage are chosen from W_{k+1}, i.e., we have

$$\mathcal{R}(W_{k+1[\ell]}) \subset \mathcal{R}(W_{k+1}) \subset \mathcal{K}_q(A, r_0). \tag{2.15}$$

In the special case that all inner subspaces have the same dimension $m_j = m$, we have $p = km$ and $q = km + k = k(m+1)$; cf. [37, Lemma 4.2], where a result similar to Lemma 2.2 is shown for this special case.

We also remark that in [18] and [34] the situation is analyzed when (2.1) is replaced with $AU_{k-1} = U_k B_k$ for some matrix B_k and $\mathcal{R}(U_k)$ is some general subspace; cf. (2.6). The question there is to find an appropriate matrix E for which the columns of U_k are the basis of a Krylov subspace of $A + E$, i.e., $\mathcal{R}(U_k) = \mathcal{K}_k(A + E, r_0)$. Here we instead have subspaces of a larger Krylov subspace of A.

3. The subspaces of the inner-outer methods. In this section, we further characterize the k-dimensional subspaces $\mathcal{R}(Z_k)$ and $\mathcal{R}(W_k)$. Our characterization is of interest but does not necessarily reveal the intrinsic form of these subspaces. It does help us, though, in providing part of the setup used in section 5, where we show how these subspaces grow with each cycle.

The first simple observation is that the approximation x_k can be expressed as a particular linear combination of all the bases of the k inner Krylov subspaces. Indeed, from (2.5) and (2.14), it follows that

$$x_k = x_0 + \sum_{j=1}^{k} (u_k)_j \; z_j = x_0 + \sum_{j=1}^{k} (u_k)_j \; V_{m_j}^{(j)} \; y_j = x_0 + \sum_{j=1}^{k} \sum_{i=1}^{m_j} (u_k)_j \; (y_j)_i \; v_i^{(j)} \; .$$

Equivalently, if we define

$$\mathcal{B}_k = [V_{m_1}^{(1)}, V_{m_2}^{(2)}, \ldots, V_{m_k}^{(k)}] \in \mathbb{R}^{n \times p} \; , \quad Y_k = \begin{bmatrix} y_1 & O & O & \cdots \\ O & y_2 & O & \cdots \\ \vdots & & \ddots & \\ O & \cdots & O & y_k \end{bmatrix} \in \mathbb{R}^{p \times k} \; , \tag{3.1}$$

where O stands for a submatrix with zero entries (in this case $m \times 1$ submatrices, for different values of m), we have, from (2.14),

$$Z_k = \mathcal{B}_k Y_k, \tag{3.2}$$

and thus $x_k = x_0 + Z_k u_k = x_0 + \mathcal{B}_k(Y_k u_k)$. If we write the $n \times q$ matrix $\mathcal{B}'_k = [V_{m_1+1}^{(1)}, V_{m_2+1}^{(2)}, \ldots, V_{m_k+1}^{(k)}]$ and the $q \times p$ matrix $\mathcal{H}_k$ as the "block diagonal collection" of all $H_m^{(i)}$'s, we obtain the relation

$$A\mathcal{B}_k = \mathcal{B}'_k \mathcal{H}_k, \tag{3.3}$$

reminiscent of (2.13) or (2.1). Therefore, using (3.2) and (3.3),

$$r_k = r_0 - AZ_k u_k = r_0 - A\mathcal{B}_k Y_k u_k = r_0 - \mathcal{B}'_k(\mathcal{H}_k Y_k u_k), \tag{3.4}$$

providing an explicit representation of the residual in terms of the complete set of inner bases collected in $\mathcal{B}'_k$.

Remark 3.1. In light of Proposition 2.1 and (2.12), the columns of $\mathcal{B}'_k$ are not linearly independent. We then consider the $n \times (p+1)$ matrix

$$\mathcal{V}_k = [w_1, V_{2:m_1+1}^{(1)}, V_{2:m_2+1}^{(2)}, \ldots, V_{2:m_k+1}^{(k)}], \tag{3.5}$$

where $V_{2:m+1}^{(j)} = [v_2^{(j)}, \ldots, v_m^{(j)}]$, $m = m_j$. If the matrix $\mathcal{V}_k$ is of full rank, we have

$$\mathcal{R}(\mathcal{V}_k) = \mathcal{K}_{p+1}(A, r_0).$$

In other words, the columns of the matrix $\mathcal{V}_k$ (if full rank) provides us with a basis for the Krylov subspace $\mathcal{K}_{p+1}(A, r_0)$ from where the global iterates are computed; see [33], where a similar construct is used. In general, we can conclude only that

$$\mathcal{R}(\mathcal{V}_k) \subseteq \mathcal{K}_{p+1}(A, r_0). \tag{3.6}$$

In light of Lemma 2.2, the relation (3.2) is natural, and by Remark 3.1 we can improve upon the dimension of the Krylov subspace from Lemma 2.2 to obtain

$$\mathcal{R}(Z_k) \subset \mathcal{K}_p(A, r_0), \qquad p = p_k. \tag{3.7}$$

We now want to find a similar explicit dependence of W_{k+1} on $\mathcal{B}'_k$ or $\mathcal{V}_k$. We consider two cases: when the columns of W_{k+1} are produced by an Arnoldi method or by a two-sided Lanczos algorithm.

LEMMA 3.2. *If the Arnoldi method is used to generate Z_k, W_k, the bases of the outer spaces, related by* (2.6)*, and these have full column rank, then,*

$$W_{k+1} = \mathcal{V}_k R_k, \tag{3.8}$$

with $R_k = G_k S_k^{-1}$, where the $(p+1) \times (k+1)$ matrix G_k and the $(k+1) \times (k+1)$ matrix S_k are given by

$$G_k = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ O & -\hat{g}_1 & O & O & \cdots \\ O & O & -\hat{g}_2 & O & \cdots \\ \vdots & & & \ddots & \\ O & O & O & \cdots & -\hat{g}_k \end{bmatrix}, \qquad S_k = \begin{bmatrix} 1 & 0 & t_{1,2} & t_{1,3} & \\ 0 & t_{2,1} & 0 & t_{2,3} & \ddots \\ 0 & 0 & t_{3,2} & 0 & \ddots \\ 0 & 0 & 0 & t_{4,3} & \ddots \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & 0 & t_{k+1,k} \end{bmatrix},$$

$\hat{g}_k \in \mathbb{R}^m$ is defined through $g_k = e_1 - H_m^{(k)} y_k = [\gamma_k; \hat{g}_k^T]^T$, and the $(k+1) \times k$ upper Hessenberg matrix T_k has entries t_{ij}.

Proof. Using (2.13), (2.14), and the definition of g_k,

$$Az_k = AV_m^{(k)} y_k = V_{m+1}^{(k)} H_m^{(k)} y_k = V_{m+1}^{(k)}(e_1 - g_k).$$

Then, by (2.12) and the fact that we use the Arnoldi method,

$$t_{k,k} = w_k^T A z_k = w_k^T V_{m+1}^{(k)}(e_1 - g_k) = e_1^T(e_1 - g_k) = 1 - \gamma_k. \tag{3.9}$$

We can use these relations to write w_{k+1} as a linear combination of the inner bases.

$$\begin{aligned} w_{k+1} t_{k+1,k} &= Az_k - W_k T_{1:k,k} \\ &= V_{m+1}^{(k)}(e_1 - g_k) - W_{k-1} T_{1:k-1,k} - w_k t_{k,k} \\ &= v_1^{(k)}(1 - \gamma_k) - [v_2^{(k)}, \ldots, v_{m+1}^{(k)}]\hat{g}_k - W_{k-1} T_{1:k-1,k} - w_k t_{k,k} \\ &= -[v_2^{(k)}, \ldots, v_{m+1}^{(k)}]\hat{g}_k - [w_1, w_2, \ldots, w_{k-1}] T_{1:k-1,k}, \end{aligned}$$

that is,

$$[w_1, w_2, \ldots, w_{k+1}] \begin{bmatrix} T_{1:k-1,k} \\ 0 \\ t_{k+1,k} \end{bmatrix} = -V_{2:m+1}^{(k)} \hat{g}_k.$$

Collecting all terms, we obtain

$$W_{k+1}S_k = [w_1, V^{(1)}_{2:m+1}, V^{(2)}_{2:m+1}, \ldots, V^{(k)}_{2:m+1}] \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ O & -\hat{g}_1 & O & O & \cdots \\ O & O & -\hat{g}_2 & O & \cdots \\ \vdots & & & \ddots & \\ O & O & O & \cdots & -\hat{g}_k \end{bmatrix}.$$

The lemma follows from the nonsingularity of the upper triangular matrix S_k. ∎

The same result (3.8) holds in the case of two-sided Lanczos. The difference is that the matrix T_k is tridiagonal, so that the entries above the third diagonal in S_k are zero. The proof proceeds in the same manner as that of Lemma 3.2, except that in (3.9) w_k is replaced by the corresponding vector of the bi-orthogonal basis to $\mathcal{R}(W_k)$.

It follows from (3.8) and Remark 3.1 that

$$\mathcal{R}(W_{k+1[\ell]}) \subset \mathcal{R}(W_{k+1}) \subset \mathcal{K}_{p+1}(A, r_0), \qquad p = p_k;$$

cf. (2.15), Lemma 2.2, and (3.7).

To complete the picture, we want to obtain relations of the form (3.3) and (3.4), using the matrix $\mathcal{V}_k$. If full rank, its columns are a basis of the global subspace $\mathcal{K}_{p+1}(A, r_0)$.

Let P be the $q \times q$ permutation matrix that moves all vectors $v_1^{(2)}, v_1^{(3)}, \ldots, v_1^{(k)}$ in $\mathcal{B}'_k$ to the end (that is, as the last columns of the whole matrix). Then we can write $\mathcal{B}'_k P = [\mathcal{V}_k, v_1^{(2)}, v_1^{(3)}, \ldots, v_1^{(k)}] = [\mathcal{V}_k, W_{2:k}]$, where $W_{2:k} = [w_2, w_3, \ldots, w_k] = W_{k+1}[O, I_{k-1}, O]^T$, I_{k-1} is the identity of order $k-1$. Thus, using (3.8), we write

$$\mathcal{B}'_k = \mathcal{V}_k \, [I, R_k[O, I_{k-1}, O]^T]P^T, \tag{3.10}$$

where I is the identity of order $p+1$, obtaining from (3.3) the desired relation

$$A\mathcal{B}_k = \mathcal{V}_k \, [I, R_k[O, I_{k-1}, O]^T]P^T\mathcal{H}_k =: \mathcal{V}_k N_k. \tag{3.11}$$

Similarly, using the definition of N_k in (3.11), replacing (3.10) in (3.4), and using (2.7), we obtain,

$$r_k = r_0 - \mathcal{V}_k N_k Y_k u_k = \mathcal{V}_k(\beta e_1 - N_k Y_k u_k). \tag{3.12}$$

4. Minimal residuals and other methods. The discussion so far applies to any Krylov subspace method including CG, MINRES, GMRES, QMR, FOM, BiCG, etc. In principle, the outer (flexible) method can be any one of them, while the inner method can be any other. In practice, FGMRES-GMRES(m), GCR-GMRES(m), and FQMR-QMR were considered [5], [6], [29], [36], [37]. GMRES is a minimal residual method [31], while QMR [16] can be seen as one if the appropriate inner product and associated norm are considered; see [9]. In fact, the proof in [9, section 4.3] applies equally well to truncated GMRES methods such as DQGMRES; we more generally call these flexible truncated GMRES (FTGMRES). Therefore, when we consider that the inner-outer method is FGMRES-GMRES(m), our analysis will equally apply to the cases of FQMR-QMR and FTGMRES-GMRES(m) with the proviso that the norm at each cycle is a different one.

In FGMRES-GMRES(m), z_k in (2.14) is chosen as to minimize the inner residual in the kth cycle, i.e.,

$$\|w_k - Az_k\| = \|w_k - AV_m^{(k)}y_k\| = \|V_m^{(k)}(e_1 - H_m^{(k)}y_k)\| = \min_{y\in\mathbb{R}^m} \|e_1 - H_m^{(k)}y\|,$$

while u_k in (2.5) is chosen as to minimize the outer residual (2.9), i.e.,

$$\|r_k\| = \min_{u\in\mathbb{R}^k} \|W_{k+1}(\beta e_1 - T_k u)\| = \min_{u\in\mathbb{R}^k} \|\beta e_1 - T_k u\|. \tag{4.1}$$

Remark 4.1. We emphasize that the inner-outer methods we discuss do not consist of just the concatenation of the inner and the outer minimizations—there is also the orthogonalization with respect of the outer basis (see, e.g., Proposition 2.1), i.e., before the outer minimization takes place, there is also a deflation step, cf. [7], [10], where some Krylov spaces with deflation are studied; see further section 7.

In light of the discussion in section 3, the minimization (4.1) can be interpreted as being performed on a subspace of dimension k of the global Krylov space $\mathcal{K}_{p+1}(A, r_0)$. Indeed, from (3.12), now we can write

$$\|r_k\| = \min_{u\in\mathbb{R}^k} \|\mathcal{V}_k(\beta e_1 - N_k Y_k u)\|.$$

We note that the columns of $\mathcal{V}_k$ are not necessarily orthogonal.

One can explicitly say that the subspace of $\mathcal{K}_{p+1}(A, r_0)$ where the minimization takes place is spanned by the (k linearly independent) columns of $\mathcal{V}_k N_k Y_k$. Another characterization of this subspace can be obtained from (2.6) and (3.8) giving us the k columns of $AZ_k = \mathcal{V}_k R_k T_k$.

5. The growing subspaces. The main result of this section is that the subspace from where the approximations are chosen keeps growing. This provides convergence of the inner-outer methods. In exact arithmetic, these methods would then terminate in at most n steps, where n is the order of the matrix A.

We have noted that, in general, we have the inclusion (3.6). It is also well known that the matrix $[r_0, Ar_0, \ldots, A^p r_0]$ may have vectors which are almost linearly dependent. We therefore want to study what we can say about the rank of $\mathcal{B}_k$ defined in (3.1) or, equivalently, that of $\mathcal{V}_k$ defined in (3.5). We show here that as k grows, i.e., as a new cycle is computed, the rank of $\mathcal{B}_k$ is guaranteed to grow as well.

For simplicity of the exposition, we assume in this section that $m_k = m$ for all k, and thus $p = mk$. We further assume that the minimal polynomial of A has degree larger than p. (More precisely, this is assumed with respect to $v_1^{(k-1)}$; see (5.2) below. In other words, we assume that the grade of $v_1^{(k-1)}$ is less than p.) We comment at the end of the section the implications of these assumptions.

LEMMA 5.1. *Let* $\mathcal{R}(V_m^{(k-1)}) = \mathcal{K}_m(A, v_1^{(k-1)})$ *and* $\mathcal{R}(V_m^{(k)}) = \mathcal{K}_m(A, v_1^{(k)})$ *be both of dimension* m *(i.e.,* $V_m^{(k-1)}$ *and* $V_m^{(k)}$ *are of full rank), with*

$$v_1^{(k)} \in \mathcal{R}(AV_m^{(k-1)}) \tag{5.1}$$

and $v_1^{(k)}$ *and* $v_1^{(k-1)}$ *being linearly independent. Then* $\dim \mathcal{R}([V_m^{(k-1)}, V_m^{(k)}]) > m$.

Proof. Let $v = v_1^{(k-1)}$; then any element $w \in \mathcal{R}(V_m^{(k-1)}) = \mathcal{K}_m(A, v)$ can be written as $w = p(A)v$, with p a polynomial of degree $m-1$ at most. Thus, from (5.1) we have that $v_1^{(k)} = Ap(A)v = q(A)v$, with $q(z)$ a nonzero polynomial of degree at most m, such that $q(0) = 0$. In other words, $1 \le \deg q(z) \le m$. In fact, since $0 \neq v_1^{(k)} \neq \alpha v$, for any $\alpha \in \mathbb{R}$, we have $2 \le \deg q(z) \le m$. Now consider

$$A^{m-1} v_1^{(k)} = A^{m-1} q(A) v = \hat{q}(A) v, \tag{5.2}$$

where $\deg \hat{q}(z) = (m-1) + \deg q(z) \geq m+1$; then $A^{m-1}v_1^{(k)} \notin \mathcal{K}_m(A, v_1^{(k-1)})$. Therefore,

$$\dim \mathcal{R}\left([V_m^{(k-1)}, A^{m-1}v_1^{(k)}]\right) = m+1,$$

from which the result follows. ∎

We note that in Lemma 5.1 (as well as Theorem 5.2 below) the hypothesis on the bases $V_m^{(j)}$ of $\mathcal{K}_m(A, v_1^{(j)})$ is that they be of full column rank. There is no requirement that they be orthogonal. Similarly, the vectors $v_1^{(j)}$ and $v_1^{(j-1)}$ need only be linearly independent. This implies that Lemma 5.1 and Theorem 5.2 apply to any method that produces these pairwise linearly independent vectors. In particular, they apply to the cases when either Arnoldi or two-sided Lanczos are used for the construction of the matrix W_k.

In [7] and [38, section 3] it is recommended that when using truncated Krylov methods the last vector of the previous basis, say $V_m^{(k-1)}$, be kept in the new basis. This recommendation was based on empirical evidence on symmetric matrices. Lemma 5.1 provides a theoretical basis for such a recommendation. In fact, the last column of $V_m^{(k-1)}$ has a nonzero component in the direction of (5.2) (or of (5.6) below), which is the key quantity for the subspace to grow.

To evaluate how the subspaces $V_m^{(k-1)}$ and $V_m^{(k)}$ generated during the inner step behave vis-á-vis Lemma 5.1, we shall compute for some specific examples their canonical angles [35]; see later section 7. Consider the two subspaces $\mathcal{R}(V_m^{(k-1)})$ and $\mathcal{R}(V_m^{(k)})$, and let m_* be the smallest integer such that $\dim \mathcal{R}(V_{m_*}^{(j)}) = \dim \mathcal{R}(V_{m_*+1}^{(j)})$, $j = k-1,\ k$. We define $\bar{m} = m$ if $2m \leq m_*$ or $\bar{m} = m_* - m$ if $2m > m_*$. The canonical angles between these two subspaces can be computed as the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\bar{m}}$ of $P_m^{(k-1)}(I - P_m^{(k)})$, where $P_m^{(j)} = V_m^{(j)}(V_m^{(j)})^T$, $j = k-1, k$, are the orthogonal projectors onto $\mathcal{R}(V_m^{(j)})$, $j = k-1,\ k$, respectively.

We present now the main result of this section.

THEOREM 5.2. *Let $\mathcal{R}(V_m^{(j)}) = \mathcal{K}_m(A, v_1^{(j)})$ be of dimension m, for $j = 1, \ldots, k$. Assume that each $v_1^{(j)}$ is such that*

$$v_1^{(j)} = AV_m^{(j-1)}y_{j-1} + \mathcal{B}_{j-2}d_j, \tag{5.3}$$

with $0 \neq y_{j-1} \in \mathbb{R}^m$, $j = 2, \ldots, k$, and that

$$v_1^{(j)} \quad \text{and} \quad v_1^{(j-1)} \quad \text{are linearly independent.} \tag{5.4}$$

Then

$$\dim \mathcal{R}(\mathcal{B}_k) \geq m + k - 1, \qquad k \geq 2. \tag{5.5}$$

Proof. We use induction on k. For $k = 2$, (5.5) holds by Lemma 5.1. Assume that (5.5) holds for $k-1 > 1$. We prove it for k. That is, we show that the next basis $V_m^{(k)}$ of $\mathcal{K}_m(A, v_1^{(k)})$ contains at least a vector that is not a linear combination of elements in $\mathcal{B}_{k-1}$. Consider the starting vector $v_1^{(k)}$ for the new basis $V_m^{(k)}$. From (5.3) we have

$$v_1^{(k)} = AV_m^{(k-1)}y_{k-1} + \mathcal{B}_{k-2}d_k = A\sum_{j=0}^{m-1} A^j v_1^{(k-1)}\eta_j + \mathcal{B}_{k-2}d_k$$

for some values η_j, not all of which are zero. Therefore, for $A^{m-1}v_1^{(k)} \in \mathcal{K}_m(A, v_1^{(k)})$ we have

$$A^{m-1}v_1^{(k)} = A^m \sum_{j=0}^{m-1} A^j v_1^{(k-1)}\eta_j + A^{m-1}\mathcal{B}_{k-2}d_k. \tag{5.6}$$

Since not all η_j's are zero, using the same argument as in the proof of Lemma 5.1,

$$A^m \sum_{j=0}^{m-1} A^j v_1^{(k-1)}\eta_j \notin \mathcal{K}_m(A, v_1^{(k-1)}), \tag{5.7}$$

while the second term of (5.6) lies in $\mathcal{B}_{k-1}$; thus, $A^{m-1}v_1^{(k)} \notin \mathcal{B}_{k-1}$, and the result follows. ∎

Remark 5.3. The hypothesis for Lemma 5.1 and Theorem 5.2 that the degree of the minimal polynomial of A is larger than m and mk, respectively, though seldom violated in practice, is crucial. Otherwise the Krylov subspace might be invariant and one may have, e.g., $\dim \mathcal{R}([V_m^{(k-1)}, V_m^{(k)}]) = m$. We will assume that this hypothesis holds, else one may experience some early breakdown; see section 6 for the definition of breakdown and also [6] for a similar situation.

Several comments on Theorem 5.2 are in order. First, we comment on the hypothesis (5.4). In the case of (flexible) Arnoldi, this is automatically satisfied, since the columns of W_k are orthogonal. In the case of (flexible) two-sided Lanczos, one has a vector $\hat{w} = \hat{w}_{j-1}$ in the basis bi-orthogonal to W_k such that $(v_1^{(j-1)}, \hat{w}) \neq 0$, while $(v_1^{(j)}, \hat{w}) = 0$, implying (5.4); for details about two-sided Lanczos see, e.g., [19], [30]. We mention also that (5.4) implies that these vectors are not the zero vectors, i.e., $w_k \neq 0$, $k = j-1,\ j$. As described in section 6 this is equivalent to assuming that there is no breakdown of the inner-outer method.

Second, this theorem applies to the inner-outer methods described so far, including the truncated methods. The hypothesis (5.3) expresses what we have said in Proposition 2.1 and Lemma 3.2.

Third, we want to emphasize that while the proof of Theorem 5.2 implies that $\mathcal{R}(\mathcal{B}_k)$ grows in each cycle, (5.5) really provides a lower bound on its dimension, and we expect the dimension of $\mathcal{R}(\mathcal{B}_k)$ to be higher. We comment, though, that (5.5) can be rephrased as

$$\dim \mathcal{R}(\mathcal{B}_k) \geq \dim \mathcal{R}([V_m^{(1)}, W_k]) = m + k - 1,$$

where the last equality holds when $V_m^{(1)}$ and W_k are full rank.

Finally, a careful review of the proofs of Lemma 5.1 and Theorem 5.2 indicates that, for the case where m_k changes from one cycle to the next, (5.5) holds for $m = m_1$, and that one needs $m_j \geq m_{j-1}$ to guarantee (5.7).

6. Stagnation and breakdown. In this section we discuss the possibility of breakdown in the construction of the outer basis W_{k+1}, as well as the possibility of stagnation of the inner-outer methods. Some of our observations apply to the general setting described in sections 2, 3, and 5. In some other cases the discussion relates specifically to a minimal residual method, such as FGMRES-GMRES(m).

As we shall see, there are two elements that relate to the possibility of stagnation and breakdown: one is $\bar{T}_k$, the $k \times k$ principal matrix of T_k, and the other is the

next (outer) vector w_{k+1}. The situation on breakdown can be summarized as follows: If $w_{k+1} \neq 0$, there is no breakdown, even if $\bar{T}_k$ is singular. If $w_{k+1} = 0$, then the singularity of $\bar{T}_k$ matters. When $\bar{T}_k$ is singular we have (real) breakdown, while if $\bar{T}_k$ is nonsingular, we have what is usually called a "lucky" breakdown, meaning that this implies that x_k is the exact solution of (1.1); see [29, Proposition 2.2] where breakdown for FGMRES (using any inner solver) is considered. Stagnation of the inner-outer method is discussed at the end of the section.

In the case of FGMRES-GMRES(m), one example of breakdown ($w_{k+1} = 0$) occurs when there is stagnation in the inner iteration, i.e., in GMRES(m). In other words, given w_k, the approximation $z_k \in \mathcal{K}_m(A, w_k)$ is such that $w_k - Az_k = w_k$, cf. (2.11), implying $z_k = 0$. Unless the exact solution was found at the previous outer step, inner stagnation yields incurable breakdown, as the outer iteration no longer proceeds. Prevention of stagnation of the inner iteration has been addressed by other authors. In [37], if inner stagnation occurs, the inner step is replaced by a minimization step. At the end of section 2.2 of [29] there is a comment indicating that all that is needed to avoid stagnation is a direction of steepest descent. This is exactly what is guaranteed by the choice of the stopping criteria of the inner iterations in [17]. We also remark that in [1] a condition on the matrix A is introduced to avoid inner stagnation in the inner-outer generalized conjugate gradient method. Let $Q[w]$ denote the inner outcome. It is shown in [1] that if $AQ[\cdot]$ is coercive, that is, there exists $\delta > 0$ such that $(w, AQ[w]) \geq \delta(w, w)$ for all $w \neq 0$, then the inner iteration does not stagnate. Here (x, y) denotes the usual inner product. Note that in the symmetric positive definite case, coercivity implies that a steepest descent direction is obtained in the inner method; see, e.g., [22]. In the context of FGMRES-GMRES(m), coercivity also implies that the inner GMRES does not stagnate. Indeed, for each outer basis vector w_k and using $z_k = Q[w_k]$, we have

$$(w_k, AQ[w_k]) = (w_k, Az_k) = 1 - \|w_k - Az_k\|^2,$$

where $w_k - Az_k$ is the inner residual. Therefore, imposing $(w_k, AQ[w_k]) > 0$ implies $\|w_k - Az_k\|^2 < 1 = \|w_k\|$, which is equivalent to lack of stagnation. We should mention however that coercivity in our context is a very strong assumption. For GMRES as inner, it holds that $Q[w_k] = p_{m-1}(A)w_k$ for some polynomial p_{m-1} of degree not greater than $m-1$, where the polynomial changes at each outer cycle. Hence, $(w_k, AQ[w_k]) > 0$ is satisfied if the operator $Ap(A)$ is coercive for any polynomial p of degree at most $m-1$.

There are many examples where $\bar{T}_k$ is singular, even if T_k is of full rank. In these cases, having $t_{k+1,k} = \|w_{k+1}\| = 0$ implies real breakdown. In Lemma 5.1 and Theorem 5.2 it is assumed that there is no breakdown, see (5.4) and (2.12), and it is concluded that the space from where the solution is drawn keeps growing.

In the absence of breakdown ($w_{k+1} \neq 0$), $\bar{T}_k$ singular indicates stagnation of the outer process and thus of the overall inner-outer method. The following result is an adaptation of [3, Theorem 3.1] for FGMRES. The proof given in [3] can be used verbatim here, so we do not reproduce it.

THEOREM 6.1. *Suppose that k steps of the (outer) Arnoldi method have been taken, $w_{k+1} \neq 0$, and assume that $\bar{T}_k$ is singular. Then*

$$\min_{u \in \mathbb{R}^k} \|\beta e_1 - T_k u\| = \min_{u \in \mathbb{R}^{k-1}} \|\beta e_1 - T_{k-1} u\|. \tag{6.1}$$

If we denote by u_j the minimizer in (4.1) with j replacing k, for $j = k$ or $k-1$, then $u_k = ((u_{k-1})^T, 0)^T$, and it follows that $x_k = x_{k-1}$. Conversely, suppose that k

steps of the (outer) Arnoldi method have been taken and that (6.1) *holds; then* $\bar{T}_k$ *is singular.*

Note, however that since we assume that there is no breakdown, this stagnation can only be "temporary," since by Theorem 5.2 the subspace keeps growing, and therefore the inner-outer method converges [9].

7. Comparison with restarted methods. Let us consider restarted GMRES, i.e., GMRES(m). This method generates an approximate solution as sum of approximations obtained at each restart, that is,

$$x_k = x_0 + x^{(1)} + x^{(2)} + \cdots + x^{(k)}. \tag{7.1}$$

The single approximations are obtained in the following subspaces:

$$\begin{aligned} &K_m(A, r_0), && r_0 = b - Ax_0, \\ &K_m(A, r_0^{(1)}), && r_0^{(1)} = r_0 - Ax^{(1)} \equiv b - Ax_1, \\ &K_m(A, r_0^{(2)}), && r_0^{(2)} = r_0^{(1)} - Ax^{(2)} \equiv b - Ax_2, \\ &\quad \vdots \end{aligned}$$

where each starting vector $r_0^{(j)}$ is the residual in the previous Krylov subspace.

Intuitively, one can think of improving upon GMRES(m) by considering a linear combination (or weighted average) of the single approximations, say

$$\tilde{x}_k(\alpha) = x_0 + \alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \cdots + \alpha_k x^{(k)}, \tag{7.2}$$

instead of (7.1); cf. [41, section 3]. One could require the parameters $\alpha \in \mathbb{R}^k$ to be constructed, for example, so as to minimize the norm of the residual, in which case we have

$$\|\tilde{r}_k\| = \min_{\alpha \in \mathbb{R}^k} \|b - A\tilde{x}_k(\alpha)\| \leq \|b - Ax_k\|,$$

where the last inequality follows from considering $\alpha = (1, \ldots, 1)^T$. In other words, such a method cannot be worse than restarted GMRES in terms of residual norms.

The inner-outer methods we study, such as FGMRES-GMRES(m), do more than just implicitly choose the k parameters in (7.2). It follows from Proposition 2.1 that one obtains an iteration of the form (7.2), but where the residual obtained for each single (inner) Krylov method is in turn deflated with respect to the previous (outer) vectors; cf. [13], [26]. We thus expect that the overall inner-outer method will perform at least as well as the restarted counterpart.

We prove this explicitly in some particular cases below ($k = 1$ or $m = 1$). For other cases, we show examples where at particular points in the computations, the inner-outer method may have a residual norm which is larger than the corresponding restarted one. Nevertheless, for these, as well as for all other numerical examples we ran, FGMRES-GMRES(m) always converges using fewer matrix-vector multiplications than GMRES(m).

In the following proposition we show that at the very first outer cycle the residual computed by FGMRES-GMRES(m) is the same as that of GMRES(m). Therefore, possible improvements versus the restarted approach are expected starting with the second outer cycle of the flexible method.

PROPOSITION 7.1. *Let r_k^F, r_k^G be the FGMRES-GMRES(m) and GMRES(m) residuals after k cycles, respectively. For $k = 1$ we have $r_1^F = r_1^G$.*

Proof. After m iterations of the first cycle of GMRES(m), we solve the problem (2.2) where V_m spans $\mathcal{K}_m(A, r_0)$, and we obtain $y^G = (H_m^T H_m)^{-1} H_m^T e_1 \beta$, so that $r_1^G = b - AV_m(H_m^T H_m)^{-1} H_m^T e_1 \beta$.

On the other hand, let $w_1 = r_0 \beta^{-1}$ be the first outer basis vector in FGMRES-GMRES(m). After m iterations of the first inner step we solve the problem

$$\min_{y \in \mathbb{R}^m} \|w_1 - AV_m y\| = \min_{y \in \mathbb{R}^m} \|e_1 - H_m y\|,$$

where V_m spans $\mathcal{K}_m(A, r_0) = \mathcal{K}_m(A, w_1)$. We obtain $y^F = (H_m^T H_m)^{-1} H_m^T e_1$ so that $z_1 = V_m y^F$. At the end of the first outer cycle we thus solve the problem

$$\min_{u \in \mathbb{R}} \|b - Az_1 u\| = \beta \min_{\hat{u} \in \mathbb{R}} \|w_1 - Az_1 \hat{u}\|, \qquad \hat{u} = \frac{1}{\beta} u.$$

We obtain $\hat{u} = ((Az_1)^T (Az_1))^{-1} (Az_1)^T w_1$. Explicitly writing $Az_1 = V_{m+1} H_m y^F$ and substituting into the previous expression, we derive $\hat{u} = ((y^F)^T H_m^T H_m y^F)^{-1} (y^F)^T H_m^T e_1 = 1$, from which it follows that $u = \beta \hat{u} = \beta$. Therefore, $r_1^F = b - Az_1 u = b - AV_m(H_m^T H_m)^{-1} H_m^T e_1 \beta = r_1^G$. ∎

Later in this section we provide an example that shows that Proposition 7.1 cannot be generalized to $k > 1$ for general m. Nonetheless, we can prove for $m = 1$ and for any $k > 0$ that the FGMRES-GMRES(m) iterates coincide with those of full GMRES.

PROPOSITION 7.2. *Let x_k^F be the approximate solution obtained after k outer cycles of FGMRES-GMRES(1), and assume that GMRES(1) does not stagnate. Let x_k^G be the approximate solution after k iterations of full GMRES. Then $x_k^F = x_k^G$.*

Proof. At each outer cycle k of FGMRES, the inner solver GMRES approximately solves $Az = w_k$. If GMRES(1) is used, then $z_k = \delta_k w_k$ for some scalar $\delta_k \neq 0$. Setting $D_k = \text{diag}(\delta_1, \ldots, \delta_k)$, we can write $Z_k = W_k D_k$ so that (2.6) becomes

$$AW_k = W_{k+1} T_k D_k^{-1},$$

and the latter is an Arnoldi relation so that $\mathcal{R}(W_k) = \mathcal{K}_k(A, w_1)$. Let $AV_k = V_{k+1} H_k$ be the Arnoldi relation associated with full GMRES. Since $\mathcal{R}(W_k) = \mathcal{R}(V_k)$, there exists an orthogonal $k \times k$ matrix R_k such that $W_k = V_k R_k$ and $H_k = R_{k+1} T_k D_k^{-1} R_k^{-1}$. Since $w_1 = v_1$, $(R_k)_{1,1} = 1$. Using $(H_k^T H_k)^{-1} H_k^T = R_k D_k (T_k^T T_k)^{-1} T_k^T R_{k+1}^T$, the solution $x_k^G = V_k (H_k^T H_k)^{-1} H_k^T e_1 \beta$ becomes

$$x_k^G = V_k R_k D_k (T_k^T T_k)^{-1} T_k^T R_{k+1}^T e_1 \beta = Z_k (T_k^T T_k)^{-1} T_k^T e_1 \beta = x_k^F. \qquad ∎$$

Unfortunately, the result of Proposition 7.2 and the implication that $\|r_k^F\| \leq \|r_k^G\|$ do not carry over to larger values of m, as experimentally shown in the next example.

Example 7.3. We consider the linear system $Ax = b$ of size $n = 100$, where $A = \text{bidiag}(d, 1)$, with $d = [0.01, 0.02, 0.03, 0.04, 10, 11, \ldots, 105] \in \mathbb{R}^n$. In Table 7.1 we report the residual norms for both FGMRES-GMRES(m) (F/G(m) for short) and GMRES(m) (GM(m) for short), with $m = 10$, when the right-hand side is $b_1 = (1, 2, 1, 2, \ldots)^T$ (left) and $b_2 = (1, -2, 1, -2, \ldots)^T$ (right). Both vectors were normalized so that $\|b_1\| = \|b_2\| = 1$. The reported results clearly show that $\|r_k^F\|$ is larger than $\|r_k^G\|$ at an early stage of the iterative process, highlighted in italics. Nonetheless, convergence is achieved after a few more iterations in the flexible method, whereas the restarted approach stagnates.

TABLE 7.1
FGMRES/GMRES(10) vs. GMRES(10) on bidiagonal matrix, with right-hand sides b_1 and b_2.

b_1			b_2		
k	F/G(10)	GM(10)	k	F/G(10)	GM(10)
1	0.197966	0.197966	1	0.168170	168170
2	*0.175303*	*0.170102*	2	0.153462	0.153675
3	0.145040	0.150841	3	*0.139839*	*0.138271*
4	0.145022	0.147060	4	*0.139510*	*0.137050*
⋮			⋮		
12	$6.4 \cdot 10^{-5}$	0.144093	13	0.00029268	0.136947

We recall here that each new vector is deflated before it is used as the initial vector for the inner iteration; see (2.11) or Proposition 2.1. As discussed in section 3, this deflation helps in providing inner subspaces $V_m^{(k)}$, which have a larger angle between them, justifying the good overall performance of the flexible method. This is illustrated in Table 7.2 below, where we considered the matrix in Example 7.3 and, as right-hand side b, the left singular vector corresponding to the smallest singular value of A; see [33]. Note that full GMRES reaches a residual norm less than 10^{-16} in 21 iterations; therefore in our tests we considered that an invariant subspace of A was found for $m_* = 20$.

For both FGMRES-GMRES(m) and GMRES(m), in Table 7.2 we report the smallest sine value of the canonical angles between the subspaces spanned by $V_m^{(k-1)}$ and $V_m^{(k)}$ for $k = 2, 3$; see the discussion in section 3. When using GM(12), we collected the value of the $\bar{m}$th singular value, with $\bar{m} = m_* - m = 20 - 12 = 8$.

TABLE 7.2
Smallest nonzero singular value of $V_m^{(k-1)}(V_m^{(k-1)})^T(I - V_m^{(k)}(V_m^{(k)})^T)$, measuring the sines of canonical angles between $\mathcal{R}(V_m^{(k-1)})$ and $\mathcal{R}(V_m^{(k)})$.

k		F/G(5)	F/G(10)	GM(5)	GM(10)	GMRES(12)
2	σ_{min}	$7 \cdot 10^{-5}$	$4 \cdot 10^{-8}$	$2 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$5 \cdot 10^{-6}$
3	σ_{min}	$1 \cdot 10^{-2}$	$9 \cdot 10^{-4}$	$4 \cdot 10^{-17}$	$3 \cdot 10^{-16}$	$3 \cdot 10^{-4}$

Table 7.2 shows that, for restarted GMRES, the distance between $V_m^{(1)}$ and $V_m^{(2)}$ is around machine precision for $m = 5, 10$, implying that the Krylov subspace generated after one restart is very close to the previous one. The same happens at the next outer cycle, for $V_m^{(2)}$ and $V_m^{(3)}$. Only for $m = 12$ do the subspaces generated after the first restart of GMRES provide new information. This is confirmed by the convergence history of the method, shown in the right plot of Figure 1.

Table 7.2 and Figure 1 also report results for the flexible method. In particular, by only deflating the starting vector of the first inner Krylov subspace, F/G(10) seems to be capable of capturing the information missed by restarted GMRES after one cycle, resulting in faster convergence. On the other hand, we notice that F/G(5) achieves its final approximation to the exact solution with a large delay.

8. Computational considerations. Computational efficiency leads us to consider truncated and restarted versions of optimal algorithms. This is the case both for the original GMRES method as well as for FGMRES [31], [32], [37]. In the case of a flexible method, Theorem 5.2 ensures that in exact arithmetic the method converges as long as condition (5.4) is satisfied. Therefore, if, for example,

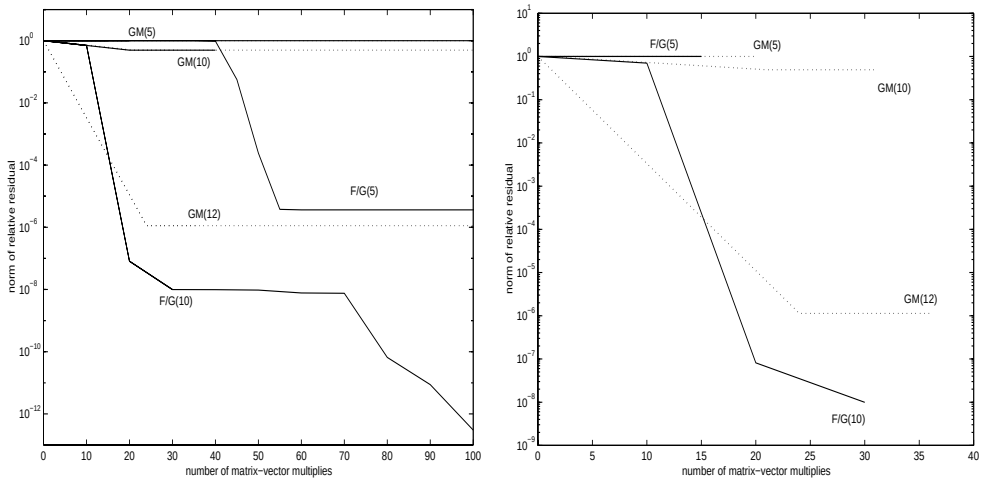


FIG. 1. *Convergence history of flexible and restarted methods for various values of m. Left: Entire picture. Right: Initial phase. See also Table* 7.2.

GMRES is used as the outer method, finite termination is obtained by orthogonalizing only the next outer basis vector w_{k+1} with respect to the previous vector w_k. By orthogonalizing w_{k+1} with respect to more basis vectors, i.e., to more columns of W_k, we expect the convergence to be faster.

In order to make fair comparisons, we recall that k cycles of FGMRES require m vectors for the inner space plus $2k$ vectors for the outer bases. These should be compared with the m vectors required by GMRES(m). A truncated version of FGMRES will instead require $2k_t$ outer vectors, where k_t is fixed beforehand. In our experiments, when using truncated flexible GMRES, we used m_t inner vectors and k_t outer vectors so that $m_t + 2k_t = m$, where m is the number of vectors used in restarted GMRES. We remark that such memory constraint on m_t, k_t forces us to work with a much smaller inner space than in restarted GMRES if outer orthogonality is maintained with respect to several vectors. Different selections of these parameters were analyzed and some of them are reported here. In addition, we mention that the larger k_t, the more expensive the reorthogonalization step. Clearly, the worst case scenario in this context is given by FGMRES, where all outer vectors are kept orthogonal.

We also remark that the matrix-vector multiplication required at each outer step of the flexible method can be avoided by exploiting the available inner residual. Therefore, at least for a low number of outer cycles, restarted and (truncated) flexible methods can be compared in terms of number of matrix-vector products, which in our experiments represents the highest computational cost. It is also customary to further precondition the flexible approach with a matrix P by applying the flexible method to the preconditioned matrix AP^{-1}. Due to the implementation consideration just mentioned, preconditioning the inner-outer method amounts to simply preconditioning the inner solver; see also the discussion in section 2.

Example 8.1. We consider the 900×900 matrix originating from the centered finite difference discretization of the operator

$$L(u) = -\Delta u + \mu x u_x, \qquad \text{for} \quad \mu = 100, 1000, \tag{8.1}$$

on $[0,1] \times [0,1]$ with zero Dirichlet boundary conditions. As a right-hand side we selected $b = Ae$, where e is a vector of all ones.

Figure 2 reports the convergence history of restarted GMRES(20), FGMRES-GMRES(20), and its truncated variant, using FTGMRES-GMRES(m_t, k_t) (FT/G(m-k) for short) as discussed above. In the left plot we displayed the results for $\mu = 100$, in the right plot for $\mu = 1000$. We observe that for $\mu = 100$ truncation is not harmful, especially when the generated inner space is sufficiently large, yielding similar convergence for the flexible method and its truncated variant. For $\mu = 1000$ the picture changes considerably and the convergence of the truncated methods reflects the influence of the two parameters m_t, k_t. Restarted GMRES clearly shows lack of information that is eventually recovered after many more iterations.

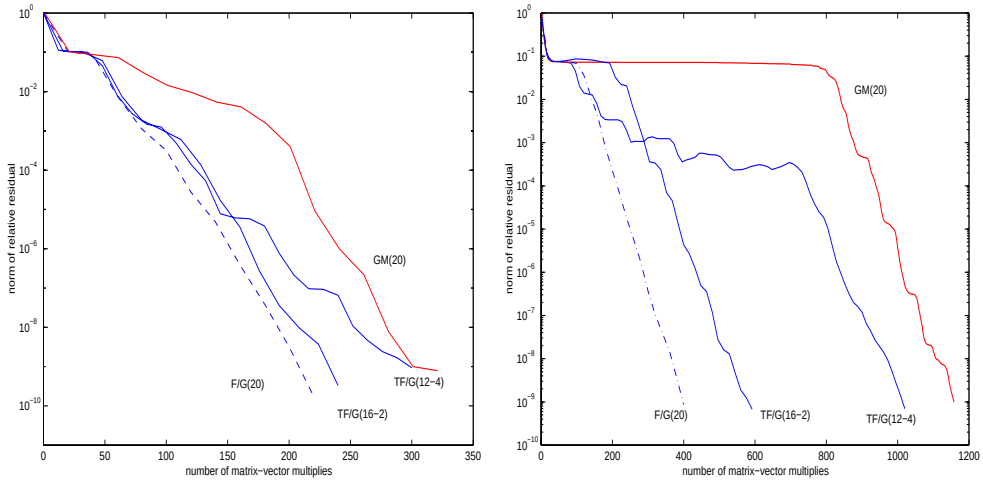


FIG. 2. *Convergence history for operator $L(u)$ in* (8.1). *Left:* $\mu = 100$. *Right:* $\mu = 1000$.

Example 8.2. We consider the operator

$$L(u) = -1000\Delta u + 2e^{4(x^2+y^2)}u_x - 2e^{4(x^2+y^2)}u_y \tag{8.2}$$

on $[0,1] \times [0,1]$, which was studied in [7, Problem 2]. Here we assume zero Dirichlet boundary conditions and a centered finite difference discretization, yielding a coefficient matrix A of size $n = 40000$. The right-hand side is determined as $b = Ae$, where e is the vector of all ones, and then normalized so as to have unit norm. The convergence history of restarted GMRES with $m = 30$ is reported in Figure 3 (left). The figure also shows the curves of the flexible variant and its truncated versions for two different values of the truncation parameter. On this problem, truncation is particularly effective. The results seem to suggest that on this problem, an inner subspace of small dimension suffices for the flexible method to converge rapidly. In the right plot of Figure 3 we show the convergence history of all methods for a larger Krylov subspace dimension, $m = 50$. Restarted GMRES considerably improves its performance; cf. [7]. On the other hand, flexible schemes do not seem to necessitate of a larger inner dimension, implying that information gathering in the outer process is very effective. Our findings corroborate similar experimental evidence in [7]. We also notice that for $m = 50$, the truncated variants converged in less than 20 outer iterations. Therefore, when large values of k_t are selected, truncation takes place only

very late in the convergence stage. As a consequence, the curves in the right plots of Figure 3 associated with the truncated methods FT/G(30-10) and FT/G(20-15) closely resemble the curves one would obtain with FGMRES(m) with $m = 30$ and $m = 20$; cf. the left plot of Figure 3 for $m = 30$.

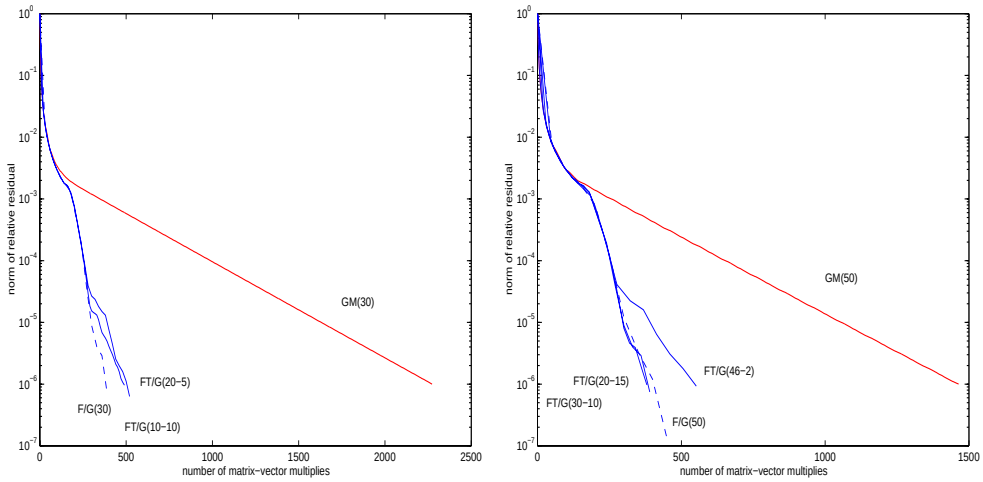


FIG. 3. *Convergence history for operator $L(u)$ in (8.2). Left: $m = 30$. Right: $m = 50$.*

Example 8.3. Our last set of experiments involves three matrices from the "Matrix Market" [8], [24]. For the first two matrices, the right-hand side was chosen to be the vector of all ones. Additional (fixed) incomplete LU preconditioning was applied [25], [30]. In our experiments, we used the Matlab function `luinc` with tolerance *tol* to build the preconditioning matrix P. As mentioned earlier in this section, in the flexible algorithm this amounts to run the inner solver with the preconditioned matrix AP^{-1}.

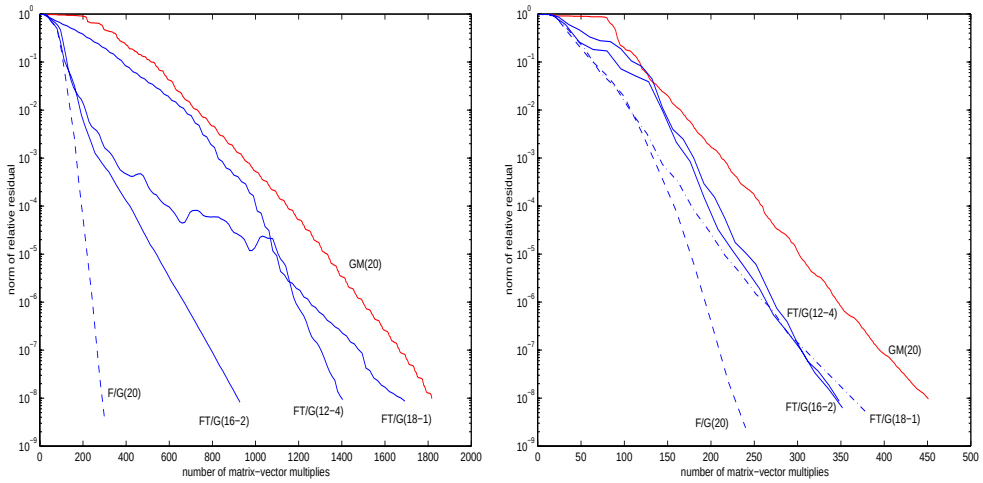


FIG. 4. *Convergence history for flexible and restarted methods on matrix Sherman5. Left: no preconditioning. Right: preconditioning using incomplete LU with tolerance tol* $= 10^{-2}$.

In Figure 4 we report experiments with the matrix SHERMAN5 from the Harwell–

Boeing set, a nonsymmetric 3312×3312 matrix stemming from a fully implicit black oil simulation [8]. The matrix was scaled using the absolute values of its diagonal entries. The fixed ILU preconditioner was built using $tol = 10^{-2}$.

The plots show that the behavior of the methods is more homogeneous after fixed preconditioning, while when no preconditioning is applied the convergence of the truncated flexible method is less predictable.

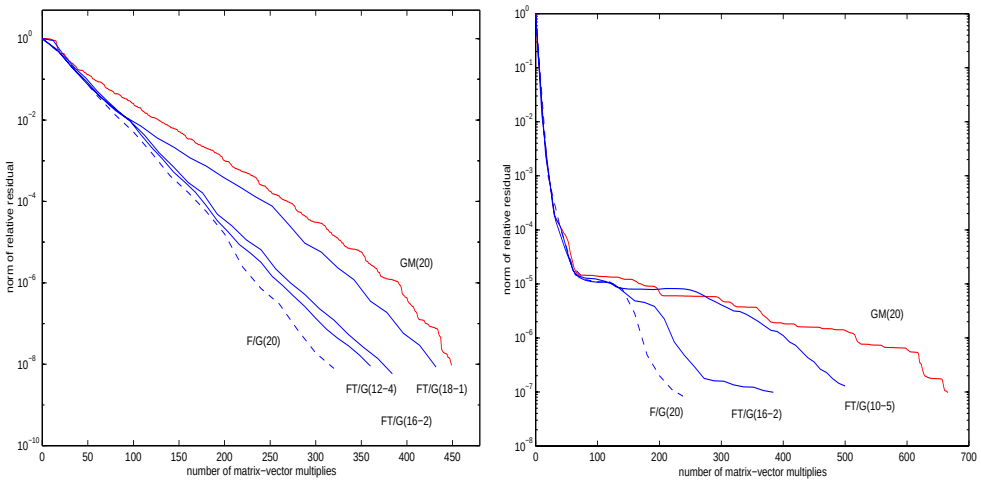


FIG. 5. *Convergence history for flexible and restarted methods on Preconditioned Oilgen (left) and on Scaled Fidap_ex*11 *(right).*

The second matrix in the set is the OILGEN 2205×2205 matrix from the same matrix set, originating from a three-dimensional oil reservoir simulation [8]. The fixed ILU preconditioning was applied with tolerance $tol = 0.5$. Convergence histories are reported in the left plot of Figure 5. The difference in the behavior of the methods is less pronounced.

Finally, we consider a larger matrix from the FIDAP group in the University of Florida collection [8], namely matrix EX11 of size $n = 16614$, which stems from a fully coupled Navier–Stokes problem. We set $b = Ae$, normalized so as to have unit norm, and we use diagonal preconditioning. The results reported in the right plot of Figure 5 show the dramatic improvements of the flexible methods over restarted GMRES with $m = 20$. These results are consistent with the other experiments on smaller matrices reported earlier.

We end this section with some comments on the behavior of the truncated schemes. By comparing the results in Figures 2, 4, and 5, we see that, except for Example 8.2, the curves of the flexible truncated variants quickly abandon the optimal curve of the flexible method, confirming that orthogonalization with respect to the previous inner starting vectors is crucial to obtain fast convergence. Among the choices we have analyzed, however, we see that maintaining orthogonality with respect to the two previous starting vectors ($k_t = 2$) seems to provide the closest to optimal convergence curve. Although in exact arithmetic $k_t = 1$ is sufficient to ensure termination, a larger value of k_t seems to pay off in our tests at the cost of a smaller inner subspace dimension (m_t). Not surprisingly, however, the performance for $k_t = 4$ indicates that a value of m_t that is too small may slow down convergence.

9. Conclusion. We have analyzed a class of flexible inner-outer iterative Krylov subspace methods in an unified manner. These are Krylov methods where the preconditioner is itself a (preconditioned) Krylov method. We have shown convergence of this class of methods by showing that the subspace from where the approximation is chosen keeps growing. This convergence is guaranteed as long as there is no stagnation in the inner iterations.

We have shown experimentally (and in some cases theoretically) that these methods can compete favorably with the standard restarted methods such as GMRES(m).

In the case of truncated methods, our theory indicates one way in which to choose the vectors to keep in order to guarantee convergence. Our experimental evidence confirms the effectiveness of this choice. These truncated methods appear to perform better than the standard restarted versions using the same amount of storage, and in some cases they are almost as good as the untruncated flexible method (which requires more storage).

Further analysis is needed to determine which vectors in the outer basis are the important ones to keep, allowing us to use the storage for more vectors in the inner basis; cf. [7].

Acknowledgments. We are indebted to Eric de Sturler for valuable comments on an earlier version of the manuscript. In particular, they led to Remark 5.3 and to a better proof of Proposition 7.2. We thank the editor, Martin Gutknecht, for his superb job in handling the paper. His questions and observations helped improve our presentation. We also benefited from comments and suggestions from T. Faraj and J. Mas and from the referees.

REFERENCES

[1] O. Axelsson and P. S. Vassilevski, *A black box generalized conjugate gradient solver with inner iterations and variable-step preconditioning*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 625–644.

[2] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1998), pp. 243–269.

[3] P. N. Brown, *A theoretical comparison of the Arnoldi and the GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.

[4] B. Carpentieri, I. S. Duff, and L. Giraud, *Multilevel Preconditioning Techniques for the Solution of Large Dense Linear Systems in Electromagnetism*, Technical report, CERFACS, Toulouse, France, in preparation.

[5] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1998), pp. 43–66.

[6] E. de Sturler, *Nested Krylov methods based on GCR*, J. Comput. Appl. Math., 67 (1996), pp. 15–41.

[7] E. de Sturler, *Truncation strategies for optimal Krylov subspace methods*, SIAM J. Numer. Anal., 36 (1999), pp. 864–889.

[8] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[9] M. Eiermann and O. Ernst, *Geometric aspects in the theory of Krylov subspace methods*, Acta Numer., 2001, pp. 251–312.

[10] M. Eiermann, O. Ernst, and O. Schneider, *Analysis of acceleration strategies for restarted minimal residual methods*, J. Comput. Appl. Math., 123 (2000), pp. 261–292.

[11] S. C. Eisenstat, J. M. Ortega, and C. T. Vaughan, *Efficient polynomial preconditioning for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 859–872.

[12] H. C. Elman, O. G. Ernst, and D. P. O'Leary, *A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations*, SIAM J. Sci. Comput., 23 (2001), pp. 1291–1315.

[13] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES preconditioned by deflation*, J. Comput. Appl. Math., 69 (1996), pp. 303–318.

[14] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.
[15] R. W. Freund, *On conjugate gradient type methods and polynomial preconditioners for a class of complex non-Hermitian matrices*, Numer. Math., 57 (1990), pp. 285–312.
[16] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[17] G. H. Golub and Q. Ye, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999), pp. 1305–1320.
[18] A. Greenbaum, *Comparison of QMR-type methods with GMRES*, presented at Mathematical Journey through Analysis, Matrix Theory and Scientific Computation: A Conference on the Occasion of Richard S. Varga's 70th Birthday, Kent State University, Kent, Ohio, March 24–26, 1999.
[19] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.
[20] O. J. Johnson, C. A. Micchelli, and G. Paul, *Polynomial preconditioners for conjugate gradient calculations*, SIAM J. Numer. Anal., 20 (1983), pp. 362–376.
[21] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for non-symmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
[22] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley, Reading, MA, 1984.
[23] *MATLAB User's Guide*, The MathWorks, Inc., Natick, MA, 1998.
[24] *Matrix Market*, http://math.nist.gov/MatrixMarket (15 July 2002).
[25] J. A. Meijerink and H. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.
[26] R. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[27] Y. Notay, *Flexible conjugate gradient*, SIAM J. Sci. Comput., 22 (2000), pp. 1444–1460.
[28] Y. Saad, *Practical use of polynomial preconditionings for the conjugate gradient method*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 865–881.
[29] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[30] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.
[31] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
[32] Y. Saad and K. Wu, *DQGMRES: A direct quasi-minimal residual algorithm based on incomplete orthogonalization*, Numer. Linear Algebra Appl., 3 (1996), pp. 329–343.
[33] V. Simoncini, *On the convergence of restarted Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 430–452.
[34] G. W. Stewart, *Backward error bounds for approximate Krylov subspaces*, Linear Algebra Appl., 340 (2002), pp. 81–86.
[35] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[36] D. B. Szyld and J. A. Vogel, *FQMR: A flexible quasi-minimal residual method with inexact preconditioning*, SIAM J. Sci. Comput., 23 (2001), pp. 363–380.
[37] H. A. Van der Vorst and C. Vuik, *GMRESR: A family of nested GMRES methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 369–386.
[38] C. Vuik, *Further experiences with GMRESR*, Supercomputer, 55 (1993), pp. 13–27.
[39] C. Vuik, *New insights in GMRES-like methods with variable preconditioners*, J. Comput. Appl. Math., 61 (1995), pp. 189–204.
[40] J. S. Warsa, M. Benzi, T. A. Wareing, and J. E. Morel, *Preconditioning a mixed discontinuous finite element method for radiation diffusion*, Numer. Linear Algebra Appl., to appear.
[41] L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2240–2253

CONVERGENCE OF SEMI-LAGRANGIAN APPROXIMATIONS TO CONVEX HAMILTON–JACOBI EQUATIONS UNDER (VERY) LARGE COURANT NUMBERS*

ROBERTO FERRETTI†

Abstract. We consider a class of semi-Lagrangian high-order approximation schemes for convex Hamilton–Jacobi equations. In this framework, we prove that under certain restrictions on the relationship between Δx and Δt, the sequence of approximate solutions is uniformly Lipschitz continuous and hence, by consistency, that it converges to the exact solution. The argument is suitable for most reconstructions of interest, including high-order polynomials and ENO reconstructions.

Key words. convex Hamilton–Jacobi equations, high-order schemes, semi-Lagrangian schemes, convergence

AMS subject classifications. 65M25, 65M12, 65M06

PII. S0036142901388378

1. Introduction.

Although the numerical solution of first-order Hamilton–Jacobi equations is a widely studied topic in many applications, it is well known that, when high-order schemes are considered (see [OS], [JP], [LT1]), very few convergence results are available. A key tool for proving convergence of numerical approximations would be some stability property for discrete solutions; when working with continuous solutions it is common to require boundedness of the family in $W^{1,\infty}$, so that compactness would be proved by the classical Ascoli–Arzela theorem. Bounds on the Lipschitz norm can generally be proved with reasonable effort in the case of low-order (usually monotone) schemes but can be really troublesome in high-order schemes, in which oscillations may occur. Lack of monotonicity also prevents using the result of Barles and Souganidis (see [BS]) which, roughly speaking, states that consistency, monotonicity, and L^∞ stability imply convergence. In the framework of Hamilton–Jacobi equations, we must also mention that a recent paper (see [LT2]) has proposed a different stability concept, that is, uniform semiconcavity of discrete solutions. A bound on the second incremental ratio is also exploited in [LS] to prove convergence of modified upwind scheme for conservation laws (MUSCL)-type schemes.

This paper is devoted to the study of the Hamilton–Jacobi equation

$$\begin{cases} v_t(x,t) + H(\nabla v(x,t)) = 0 & \text{in } \mathbb{R}^N \times [0,T], \\ v(x,0) = v_0(x). \end{cases} \tag{1.1}$$

The function $H(p)$ will be assumed to be smooth and to satisfy, for some $m_H > 0$, the condition

$$(H_{pp}\xi, \xi) \geq m_H |\xi|^2. \tag{1.2}$$

The purpose of the paper is precisely to obtain a result of convergence via uniform Lipschitz continuity for a semi-Lagrangian, high-order scheme. The scheme is described in several versions in [FF1], [FF3], [FFM], [FG], [CFF] (see also [SC] for a

*Received by the editors April 24, 2001; accepted for publication (in revised form) June 24, 2002; published electronically, January 7, 2003. This work was supported by the MURST project "Metodologie numeriche avanzate per il calcolo scientifico."
http://www.siam.org/journals/sinum/40-6/38837.html
†Dipartimento di Matematica, Universtitá di Roma Tre, largo S. Leonardo Murialdo, 1, 00146 Roma, Italy (ferretti@mat.uniroma3.it).

general review on semi-Lagrangian approach in computational fluid dynamics), and its convergence for the low-order case is discussed in [FG], [FF3]. The high-order, linear case is analyzed in [FF2], whereas for the nonlinear equation (1.2) a weaker preliminary result of $W^{1,\infty}$ boundedness is given in [F].

For the reader's convenience, we briefly sketch the construction of the scheme for the Hamilton–Jacobi equation (1.1). Since the solution (see [L]) may be represented as

$$v(x,t) = \inf_{\alpha\in\mathbb{R}^N}\{tH^*(\alpha) + v_0(x+\alpha t)\} \tag{1.3}$$

(where $H^*(\alpha)$ is the Legendre transform of $H(p)$), using the representation formula (1.3) at each node and over each single time step, we get

$$v(x_j, t_{n-1}+\Delta t) = \inf_{\alpha}\{\Delta t H^*(\alpha) + v(x_j+\alpha\Delta t, t_{n-1})\},$$

and further replacing $v(x_j+\alpha\Delta t, t_{n-1})$ with a numerical reconstruction $I[V^{n-1}](x_j+\alpha\Delta t)$, we obtain at last the general form of the scheme for (1.1), namely,

$$\begin{cases} v_j^n = \min_{\alpha}\{\Delta t H^*(\alpha) + I[V^{n-1}](x_j+\alpha\Delta t)\}, \\ v_j^0 = v_0(x_j). \end{cases} \tag{1.4}$$

In (1.4), we have omitted for simplicity the treatment of boundary conditions, for which an explicit form can be found in [FFM].

It is worth pointing out that in (1.4) there is no need to let $\Delta t \to 0$ since characteristics are straight lines and the best accuracy is achieved with $\Delta t = t$ as in (1.3). However, this is no longer true if $H(p)$ is replaced by $H(x,p)$ (the scheme can also handle this case, see [FFM]), and in order for this simplified convergence analysis to give indications for the more general case, we will let $\Delta t \to 0$ anyway.

The outline of the paper is the following. We will prove in section 2 a bound on the second increment and deduce Lipschitz continuity for numerical solutions under suitable assumptions on the reconstruction operator and on the $\Delta t/\Delta x$ relationship. Section 3 takes into consideration polynomial reconstructions and checks the applicability of the theory. Lastly, section 4 gives the main result of convergence for the scheme.

2. Uniform Lipschitz continuity. Let (1.1) be discretized on an infinite grid with nodes x_j; moreover, let Δx be the space discretization parameter, so that

$$C_-\Delta x \le |x_i - x_j| \le C_+\Delta x$$

for any couple (x_i, x_j) of nodes on the boundary of the same cell. (For shortness of notation, nodes satisfying this condition will be referred to in what follows as *neighboring nodes*.) We will identify the numerical solution at time $n\Delta t$ with the sequence $V^n = \{v_j^n\}_j$, whose l^∞ norm will be defined as usual by $|V^n|_\infty := \sup_j |v_j^n|$. We will denote by $I[V](\cdot)$ the reconstruction (interpolation) operator which extends a sequence V on the whole of $\mathbb{R}^N$, by $U(x)$ the stencil of $I[V](x)$, and by $I_1[V]$ the P_1 (piecewise linear) interpolation on the sequence V.

For simplicity, the main technical difficulty of the paper, Lemma 2.1, will be stated in the one-dimensional case. In this case, we assume that $x_j = j\Delta x$ ($j = 0, \pm1, \pm2, \ldots$) and that given a Lipschitz continuous function $v(x)$ and the sequence $V = \{v_j\}_j = \{v(x_j)\}_j$, the operator $I[V]$ satisfies, for some constant $C < 1$,

$$I[V](x_j) = v(x_j)\ ,\quad |I[V](x) - I_1[V](x)| \le C \max_{x_{j-1},x_j,x_{j+1}\in U(x)} |v_{j+1} - 2v_j + v_{j-1}|, \tag{2.1}$$

where $I_1[V]$ and $U(x)$ have the same meaning as before. In $\mathbb{R}$, we can write $U(x)$ as $U(x) = (x - h_- \Delta x, x + h_+ \Delta x)$; for example, a quadratic Lagrange reconstruction can be performed taking one node on the left and two nodes on the right of the point x (in this case, $h_- = 1$, $h_+ = 2$), or two nodes on the left and one on the right (and in this case, $h_- = 2$, $h_+ = 1$). In a second-order ENO reconstruction (see [S]), both cases are possible depending on the solution (and thus, $h_- = h_+ = 2$). A third-order Lagrange reconstruction is typically performed using two nodes on the left and two on the right, so that $h_- = h_+ = 2$. In the third-order ENO case, $h_- = h_+ = 3$, and so forth.

We recall that condition (1.2) in the one-dimensional case reads as

$$H''(p) \geq m_H,$$

and this implies in terms of the Legendre transform that

$$H^{*''}(\alpha) \leq \frac{1}{m_H}. \tag{2.2}$$

It will also be useful to define the function

$$F_j(\alpha) := \Delta t H^*(\alpha) + I[V^{n-1}](x_j + \alpha \Delta t).$$

In what follows, we will denote by $\bar{\alpha}_j$ the value of α achieving the minimum in (1.4) and in F_j. In order to stress the "locality" in the above definition, the notation explicitly shows the dependence of F on the node index j, although for simplicity the dependence on the time step n has been dropped.

We give now a bound on the second increment of the numerical solution. The bound is globally one-sided, but becomes two-sided at the foot of characteristics (that is, in a neighborhood of fixed radius $h\Delta x$, with $h > \max(h_+, h_-)$, around the point $x_j + \bar{\alpha}_j \Delta t$). More precisely, we have the following technical result.

LEMMA 2.1. *Consider the scheme* (1.4) *applied to equation* (1.1)*, with* $N = 1$*. If* (2.2) *holds, then, for any* $j \in \mathbb{Z}$ *and* $n \geq 1$*,*

$$v_{j+1}^n - 2v_j^n + v_{j-1}^n \leq \frac{\Delta x^2}{m_H \Delta t}. \tag{2.3}$$

Moreover, assuming, in addition, that (2.1) *holds, then, for any* $j \in \mathbb{Z}$ *and* $n \geq 2$*,*

$$\max_{x_{i-1}, x_i, x_{i+1} \in \overline{U}(x_j + \bar{\alpha}_j \Delta t)} |v_{i+1}^{n-1} - 2v_i^{n-1} + v_{i-1}^{n-1}| \leq \bar{C} \frac{\Delta x^2}{\Delta t} \tag{2.4}$$

with $\overline{U}(x) = (x - h\Delta x, x + h\Delta x)$ *(*h *being a fixed positive integer such that* $h > \max(h_+, h_-)$*) and for some positive constant* $\bar{C}$ *depending on* C*,* h*, and* m_H*.*

Proof. We start by proving (2.3). By (1.4) we have, for $n \geq 1$,

$$\begin{aligned}
v_j^n &= \Delta t H^*(\bar{\alpha}_j) + I[V^{n-1}](x_j + \bar{\alpha}_j \Delta t), \\
v_{j-1}^n &= \Delta t H^*(\bar{\alpha}_{j-1}) + I[V^{n-1}](x_{j-1} + \bar{\alpha}_{j-1} \Delta t) \\
&\leq \Delta t H^* \left(\bar{\alpha}_j + \frac{\Delta x}{\Delta t} \right) + I[V^{n-1}](x_j + \bar{\alpha}_j \Delta t), \\
v_{j+1}^n &= \Delta t H^*(\bar{\alpha}_{j+1}) + I[V^{n-1}](x_{j+1} + \bar{\alpha}_{j+1} \Delta t) \\
&\leq \Delta t H^* \left(\bar{\alpha}_j - \frac{\Delta x}{\Delta t} \right) + I[V^{n-1}](x_j + \bar{\alpha}_j \Delta t)
\end{aligned}$$

(where the inequalities come from the use of $\bar\alpha_j \pm \Delta x/\Delta t$ instead of the actual minimizers $\bar\alpha_{j\mp1}$), so that

$$v_{j+1}^n - 2v_j^n + v_{j-1}^n \le \Delta t \left[H^*\left(\bar\alpha_j - \frac{\Delta x}{\Delta t}\right) - 2H^*(\bar\alpha_j) + H^*\left(\bar\alpha_j + \frac{\Delta x}{\Delta t}\right)\right] \le \Delta t \left(\frac{\Delta x}{\Delta t}\right)^2 \sup H^{*\prime\prime}, \tag{2.5}$$

which gives (2.3).

In order to prove (2.4), we take into consideration the values of $F_j(\alpha)$ for $x_j + \alpha\Delta t$ coinciding with a grid node, that is for $\alpha = k\Delta x/\Delta t$. Let k_j denote the index of the node for which $x_j + \bar\alpha_j \Delta t \in [x_{j+k_j}, x_{j+k_j+1}]$. We will denote by δ_j the (signed) maximal second increment of the numerical solution in the neighborhood $\overline{U}(x_j + \bar\alpha_j \Delta t)$ and by $j + k_j + l_j$ the index of the node at which this second increment occurs. If $\delta_j \ge 0$, then (2.4) follows from (2.3) and we have nothing else to prove. We will assume, therefore, that $\delta_j < 0$, so that

$$\begin{aligned}\delta_j &= - \max_{x_{i-1},x_i,x_{i+1}\in\overline{U}(x_j+\bar\alpha_j\Delta t)} |v_{i+1}^{n-1} - 2v_i^{n-1} + v_{i-1}^{n-1}| \\ &= v_{j+k_j+l_j+1}^{n-1} - 2v_{j+k_j+l_j}^{n-1} + v_{j+k_j+l_j-1}^{n-1}.\end{aligned} \tag{2.6}$$

We recall that $|l_j| < h$, and will also assume in what follows that $l_j \ge 0$, the reverse case being similar to prove.

As a first step, we can bound the second increment of the function F_j as follows:

$$\begin{aligned} &F_j\left((k+1)\frac{\Delta x}{\Delta t}\right) - 2F_j\left(k\frac{\Delta x}{\Delta t}\right) + F_j\left((k-1)\frac{\Delta x}{\Delta t}\right) \\ &= \Delta t\left[H^*\left((k+1)\frac{\Delta x}{\Delta t}\right) - 2H^*\left(k\frac{\Delta x}{\Delta t}\right) + H^*\left((k-1)\frac{\Delta x}{\Delta t}\right)\right] \\ &\quad + v_{j+k+1}^{n-1} - 2v_{j+k}^{n-1} + v_{j+k-1}^{n-1} \\ &\le \frac{\Delta x^2}{m_H \Delta t} + v_{j+k+1}^{n-1} - 2v_{j+k}^{n-1} + v_{j+k-1}^{n-1} \le \frac{2\Delta x^2}{m_H \Delta t}, \end{aligned} \tag{2.7}$$

where the second increment of H^* has been estimated as in (2.5), and the last inequality holds for $n \ge 2$.

In the second step, we prove that

$$F_j(\bar\alpha_j) \ge \min\left[F_j\left(k_j\frac{\Delta x}{\Delta t}\right), F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)\right] - \frac{\Delta x^2}{8m_H\Delta t} - C|\delta_j|. \tag{2.8}$$

In fact, set $\bar\alpha_j = (1-\bar\theta)k_j\Delta x/\Delta t + \bar\theta(k_j+1)\Delta x/\Delta t$, with $\bar\theta \in [0,1]$. Taking into account the uniform convexity of H^* implied by (2.2), we have

$$\begin{aligned} H^*(\bar\alpha_j) &\ge (1-\bar\theta)H^*\left(k_j\frac{\Delta x}{\Delta t}\right) + \bar\theta H^*\left((k_j+1)\frac{\Delta x}{\Delta t}\right) \\ &\quad - \frac{1}{2}\bar\theta(1-\bar\theta)\left(\frac{\Delta x}{\Delta t}\right)^2 \sup H^{*\prime\prime} \\ &\ge (1-\bar\theta)H^*\left(k_j\frac{\Delta x}{\Delta t}\right) + \bar\theta H^*\left((k_j+1)\frac{\Delta x}{\Delta t}\right) - \frac{\Delta x^2}{8m_H\Delta t^2} \end{aligned} \tag{2.9}$$

in which the third term has been minimized with respect to $\bar\theta$.

In a similar way, we also get, using (2.1) and the inclusion of $U(x)$ in $\overline{U}(x)$,

$$\begin{aligned}
&I[V^{n-1}](x_j+\bar\alpha_j\Delta t)\\
&\geq I_1[V^{n-1}](x_j+\bar\alpha_j\Delta t) - C\max_{x_{i-1},x_i,x_{i+1}\in U(x_j+\bar\alpha_j\Delta t)}|v_{i+1}-2v_i+v_{i-1}|\\
&\geq I_1[V^{n-1}](x_j+\bar\alpha_j\Delta t) - C\max_{x_{i-1},x_i,x_{i+1}\in \overline{U}(x_j+\bar\alpha_j\Delta t)}|v_{i+1}-2v_i+v_{i-1}|\\
&= (1-\bar\theta)v_{j+k_j}+\bar\theta v_{j+k_j+1} - C|\delta_j|\\
&= (1-\bar\theta)I[V^{n-1}]\left(x_j+k_j\frac{\Delta x}{\Delta t}\Delta t\right)+\bar\theta I[V^{n-1}]\left(x_j+(k_j+1)\frac{\Delta x}{\Delta t}\Delta t\right)-C|\delta_j|,
\end{aligned}$$

which, combined with (2.9), gives

$$F_j(\bar\alpha_j)\geq(1-\bar\theta)F_j\left(k_j\frac{\Delta x}{\Delta t}\right)+\bar\theta F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)-\frac{\Delta x^2}{8m_H\Delta t}-C|\delta_j|,$$

which in turn implies (2.8).

In the third step, we derive from (2.8) an estimate of the increment of F_j between $k_j\Delta x/\Delta t$ and $(k_j+1)\Delta x/\Delta t$.

Assume first that $F_j(k_j\Delta x/\Delta t)<F_j((k_j+1)\Delta x/\Delta t)$. Using (2.8) and the optimality of $\bar\alpha_j$, we obtain

$$F_j\left((k_j-1)\frac{\Delta x}{\Delta t}\right)\geq F_j(\bar\alpha_j)\geq F_j\left(k_j\frac{\Delta x}{\Delta t}\right)-\frac{\Delta x^2}{8m_H\Delta t}-C|\delta_j|,$$

that is, considering the extreme terms

$$F_j\left(k_j\frac{\Delta x}{\Delta t}\right)-F_j\left((k_j-1)\frac{\Delta x}{\Delta t}\right)\leq\frac{\Delta x^2}{8m_H\Delta t}+C|\delta_j|.$$

On the other hand, from (2.7) written with $k=k_j$, we also have

$$F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)-F_j\left(k_j\frac{\Delta x}{\Delta t}\right)\leq F_j\left(k_j\frac{\Delta x}{\Delta t}\right)-F_j\left((k_j-1)\frac{\Delta x}{\Delta t}\right)+\frac{2\Delta x^2}{m_H\Delta t}$$

and therefore, combining the last two inequalities, we get the desired bound on the first increment:

$$F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)-F_j\left(k_j\frac{\Delta x}{\Delta t}\right)\leq\frac{17\Delta x^2}{8m_H\Delta t}+C|\delta_j|. \tag{2.10}$$

Moreover, if $F_j((k_j+1)\Delta x/\Delta t)<F_j(k_j\Delta x/\Delta t)$, then

$$F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)-F_j\left(k_j\frac{\Delta x}{\Delta t}\right)\leq 0$$

and (2.10) is also trivially satisfied.

As a fourth step, we show that in order for $\bar\alpha_j$ to be a minimizer for F_j, δ_j must satisfy (2.4). To this end, we first assume that $F_j(k_j\Delta x/\Delta t)\leq F_j((k_j+1)\Delta x/\Delta t)$ and bound the values $F_j(k\Delta x/\Delta t)$ from above using a function $\tilde F_j(k\Delta x/\Delta t)$ constructed so as to coincide with F_j at $k_j\Delta x/\Delta t$ and to have first and second increments greater or equal to the corresponding increments of F_j. Taking into account the bounds on the first and second increment of F_j obtained in the previous steps of the proof, we

could define $\tilde{F}_j$ so that its first increment between $k_j\Delta x/\Delta t$ and $(k_j+1)\Delta x/\Delta t$ would be given by the right-hand side of (2.10), and the second increment would be obtained in view of (2.6), (2.7) as

$$\tilde{F}_j\left((k+1)\frac{\Delta x}{\Delta t}\right) - 2\tilde{F}_j\left(k\frac{\Delta x}{\Delta t}\right) + \tilde{F}_j\left((k-1)\frac{\Delta x}{\Delta t}\right) = \begin{cases} \dfrac{2\Delta x^2}{m_H\Delta t} & \text{if } k \neq k_j + l_j, \\ \dfrac{\Delta x^2}{m_H\Delta t} + \delta_j & \text{if } k = k_j + l_j. \end{cases} \tag{2.11}$$

We recall that, as it is easy to prove by induction, if a sequence f_i has a constant second increment,

$$f_{i+2} - 2f_{i+1} + f_i \equiv d,$$

then the values of the elements and of the first increments are given by

$$f_{k+l} = f_k + l(f_{k+1} - f_k) + (1 + 2 + \cdots + (l-1))d,$$

$$f_{k+l+1} - f_{k+l} = (f_{k+1} - f_k) + ld.$$

Using the previous equalities, a function $\tilde{F}_j$ suitable for our purpose could be defined more explicitly as

$$\tilde{F}_j\left(k_j\frac{\Delta x}{\Delta t}\right) = F_j\left(k_j\frac{\Delta x}{\Delta t}\right),$$

$$\tilde{F}_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right) = F_j\left(k_j\frac{\Delta x}{\Delta t}\right) + \frac{17\Delta x^2}{8m_H\Delta t} + C|\delta_j|,$$

$$\tilde{F}_j\left((k_j+2)\frac{\Delta x}{\Delta t}\right) = F_j\left(k_j\frac{\Delta x}{\Delta t}\right) + 2\left(\frac{17\Delta x^2}{8m_H\Delta t} + C|\delta_j|\right) + \frac{2\Delta x^2}{m_H\Delta t},$$

$$\vdots$$

$$\tilde{F}_j\left((k_j+l_j)\frac{\Delta x}{\Delta t}\right) = F_j\left(k_j\frac{\Delta x}{\Delta t}\right) + l_j\left(\frac{17\Delta x^2}{8m_H\Delta t} + C|\delta_j|\right) + (1+\cdots+(l_j-1))\frac{2\Delta x^2}{m_H\Delta t},$$

$$\begin{aligned}\tilde{F}_j\left((k_j+l_j+1)\frac{\Delta x}{\Delta t}\right) &= F_j\left(k_j\frac{\Delta x}{\Delta t}\right) + l_j\left(\frac{17\Delta x^2}{8m_H\Delta t} + C|\delta_j|\right) \\ &\quad + (1+\cdots+(l_j-1))\frac{2\Delta x^2}{m_H\Delta t} + \left(\frac{17\Delta x^2}{8m_H\Delta t} + C|\delta_j|\right) \\ &\quad + l_j\frac{2\Delta x^2}{m_H\Delta t} + \delta_j\end{aligned}$$

(note that the computation of $\tilde{F}_j((k_j+l_j+1)\Delta x/\Delta t)$ is "restarted" because of (2.11)) and last, for any integer $m>0$,

$$\begin{aligned}\tilde{F}_j\left((k_j+l_j+m)\frac{\Delta x}{\Delta t}\right) &= F_j\left(k_j\frac{\Delta x}{\Delta t}\right)\\ &\quad+(l_j+m)\left(\frac{17\Delta x^2}{8m_H\Delta t}+C|\delta_j|\right)\\ &\quad+(1+\cdots+(l_j+m-1))\frac{2\Delta x^2}{m_H\Delta t}+m\delta_j\\ &\le F_j\left(k_j\frac{\Delta x}{\Delta t}\right)+(h+m)\left(\frac{17\Delta x^2}{8m_H\Delta t}+C|\delta_j|\right)\\ &\quad+(1+\cdots+(h+m-1))\frac{2\Delta x^2}{m_H\Delta t}+m\delta_j.\end{aligned}\tag{2.12}$$

On the other hand, by (2.8) and the optimality of $\bar{\alpha}_j$, we also have, for any $m>0$,

$$\begin{aligned}\tilde{F}_j\left((k_j+l_j+m)\frac{\Delta x}{\Delta t}\right) &\ge F_j\left((k_j+l_j+m)\frac{\Delta x}{\Delta t}\right)\\ &\ge F_j(\bar{\alpha}_j)\\ &\ge \min\left[F_j\left(k_j\frac{\Delta x}{\Delta t}\right),F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)\right]\\ &\quad-\frac{\Delta x^2}{8m_H\Delta t}-C|\delta_j|.\end{aligned}\tag{2.13}$$

We explicitly note that in (2.13) the first inequality follows from the construction of $\tilde{F}_j$ as an upper bound for F_j, the second one from the optimality of $\bar{\alpha}_j$ in F_j, and the third one from the lower bound (2.8).

Recalling that we have assumed $F_j(k_j\Delta x/\Delta t)\le F_j((k_j+1)\Delta x/\Delta t)$ and $\delta_j<0$, and using (2.12), we obtain from the two extreme terms of (2.13)

$$\begin{aligned}&F_j\left(k_j\frac{\Delta x}{\Delta t}\right)+(h+m)\left(\frac{17\Delta x^2}{8m_H\Delta t}-C\delta_j\right)\\ &\quad+(1+\cdots+(h+m-1))\frac{2\Delta x^2}{m_H\Delta t}+m\delta_j\\ &\ge F_j\left(k_j\frac{\Delta x}{\Delta t}\right)-\frac{\Delta x^2}{8m_H\Delta t}+C\delta_j;\end{aligned}$$

that is,

$$\left[\frac{17}{8}(h+m)+2(1+\cdots+(h+m-1))+1\right]\frac{\Delta x^2}{m_H\Delta t}\ge\left[C(h+m)-m+C\right]\delta_j,$$

which gives, solving for δ_j,

$$-\frac{m(1+o(1))}{1-C}\frac{\Delta x^2}{m_H\Delta t}\le\delta_j<0\tag{2.14}$$

in which it is easy to recognize that the maximum of the left-hand side is $O(\Delta x^2/\Delta t)$, provided $C<1$. Using the reverse estimate (2.3), we get at last (2.4).

If otherwise $F_j((k_j+1)\Delta x/\Delta t) < F_j(k_j\Delta x/\Delta t)$, it is possible to redefine the function $\tilde{F}_j$ so that $\tilde{F}_j((k_j+1)\Delta x/\Delta t) = F_j((k_j+1)\Delta x/\Delta t)$, and the same upper bounds are used on the first and second increments. In this way, we replace (2.12) with

$$\begin{aligned}(2.12') \qquad \tilde{F}_j\left((k_j+l_j+m)\frac{\Delta x}{\Delta t}\right) &= F_j\left((k_j+1)\frac{\Delta x}{\Delta t}\right)\\ &\quad + (l_j+m-1)\left(\frac{17\Delta x^2}{8m_H\Delta t}+C|\delta_j|\right)\\ &\quad + (1+\cdots+(l_j+m-1))\frac{2\Delta x^2}{m_H\Delta t}+m\delta_j,\end{aligned}$$

which again implies (2.14). This construction completely parallels the previous one and is therefore left to the reader. □

Remark 2.1. The technique used in this lemma is not specifically suited for the one-dimensional case, although for higher dimensions (especially on unstructured grids), assumption (2.1) should be suitably redefined. The first part of Lemma 2.1 may be soon extended to higher dimensions, by using the same technique of making the feet of different characteristics coincide. The second part of the lemma would require a more technical construction of the function $\tilde{F}_j$. Also, the treatment of Dirichlet boundary conditions is not a major problem. Simply, the foot of characteristics might coincide with a point on the boundary, requiring again a proper, more technical definition of the functions F_j and $\tilde{F}_j$.

We can now prove Lipschitz continuity. We will assume that, given a Lipschitz continuous function $v(x)$ and the sequence $V = \{v_j\}_j = \{v(x_j)\}_j$, the condition

$$|I[V](x) - I_1[V](x)| \le \tilde{C}\Delta x \tag{2.15}$$

holds for any $x \in \mathbb{R}$, with a positive constant $\tilde{C}$. In the one-dimensional setting, this is a consequence of (2.1) since

$$\begin{aligned}|I[V](x) - I_1[V](x)| &\le C \max_{x_{i-1},x_i,x_{i+1}\in U(x)} |v_{i+1}-2v_i+v_{i-1}|\\ &\le C\left(\sup|v_{i+1}-v_i| + \sup|v_{i-1}-v_i|\right) = 2CL\Delta x.\end{aligned}$$

We will also assume that at the foot of a characteristic the stronger condition

$$|I[V^{n-1}](x_i+\bar{\alpha}_j\Delta t) - I_1[V^{n-1}](x_i+\bar{\alpha}_j\Delta t)| \le \hat{C}\frac{\Delta x^2}{\Delta t} \tag{2.16}$$

holds for any $j \in \mathbb{Z}$, $n \ge 2$ and for any node x_i neighboring x_j, with some positive constant $\hat{C}$. In the case of one-dimensional problems, (2.16) follows from (2.1) and Lemma 2.1. (In fact, $\overline{U}(x_j+\bar{\alpha}_j\Delta t)$ contains all the nodes involved in the reconstructions $I[V^{n-1}](x_{j\pm1}+\bar{\alpha}_j\Delta t)$.)

THEOREM 2.1. *Consider the scheme* (1.4) *applied to* (1.1). *Assume that* (2.15), (2.16) *hold, that* $\Delta x = O(\Delta t^2)$, *and that* v_0 *is Lipschitz continuous with Lipschitz constant* L_0. *Then, the numerical solutions* V^n *satisfy, for any* i *and* j, *the discrete Lipschitz estimate*

$$\frac{|v_i^n - v_j^n|}{|x_i - x_j|} \le L$$

for a constant L independent of Δx and Δt, and for $0 \le n \le T/\Delta t$, as $\Delta t \to 0$.

Proof. It clearly suffices to prove the claim for i, j such that x_i and x_j are neighboring nodes. Assume that at the previous step the discrete solution satisfies, for any i and j,

$$\frac{|v_i^{n-1} - v_j^{n-1}|}{|x_i - x_j|} \le L_{n-1}.$$

Making the argmin explicit and using (2.16), we have

$$\begin{aligned} v_j^n &= \min_\alpha \left\{ \Delta t H^*(\alpha) + I[V^{n-1}](x_j + \alpha \Delta t) \right\} \\ &= \Delta t H^*(\bar\alpha_j) + I[V^{n-1}](x_j + \bar\alpha_j \Delta t) \\ &\ge \Delta t H^*(\bar\alpha_j) + I_1[V^{n-1}](x_j + \bar\alpha_j \Delta t) - \hat C \frac{\Delta x^2}{\Delta t}. \end{aligned} \tag{2.17}$$

In order to estimate the discrete incremental ratio of V^n, we give on v_i^n the bound

$$\begin{aligned} v_i^n &= \Delta t H^*(\bar\alpha_i) + I[V^{n-1}](x_i + \bar\alpha_i \Delta t) \\ &\le \Delta t H^*(\bar\alpha_j) + I[V^{n-1}](x_i + \bar\alpha_j \Delta t) \\ &\le \Delta t H^*(\bar\alpha_j) + I_1[V^{n-1}](x_i + \bar\alpha_j \Delta t) + \hat C \frac{\Delta x^2}{\Delta t}, \end{aligned} \tag{2.18}$$

which results from both the optimality of $\bar\alpha_i$ and (2.16) and holds for any $n \ge 2$. If $n = 1$, applying (2.15) instead of (2.16), we obtain

$$\begin{aligned} v_i^1 &= \Delta t H^*(\bar\alpha_i) + I[V^0](x_i + \bar\alpha_i \Delta t) \\ &\le \Delta t H^*(\bar\alpha_j) + I[V^0](x_i + \bar\alpha_j \Delta t) \\ &\le \Delta t H^*(\bar\alpha_j) + I_1[V^0](x_i + \bar\alpha_j \Delta t) + \tilde C \Delta x. \end{aligned} \tag{2.19}$$

From (2.17) and (2.18) we obtain, for $n \ge 2$, the unilateral estimate

$$\begin{aligned} \frac{v_i^n - v_j^n}{|x_i - x_j|} &\le \frac{1}{|x_i - x_j|} \left[I_1[V^{n-1}](x_i + \bar\alpha_j \Delta t) - I_1[V^{n-1}](x_j + \bar\alpha_j \Delta t) \right] \\ &\quad + 2\hat C \frac{\Delta x^2}{\Delta t} \\ &\le L_{n-1} + \frac{2\hat C}{C_-} \frac{\Delta x}{\Delta t}, \end{aligned} \tag{2.20}$$

in which C_- is the constant defined at the start of the section, and we have used the fact that the first-order reconstruction I_1 at step $n-1$ has also Lipschitz constant L_{n-1}. Interchanging the role of $\bar\alpha_j$ and $\bar\alpha_{j+1}$, we get the reverse estimate

$$\frac{v_j^n - v_i^n}{|x_i - x_j|} \le L_{n-1} + \frac{2\hat C}{C_-} \frac{\Delta x}{\Delta t},$$

and therefore

$$\frac{|v_j^n - v_i^n|}{|x_i - x_j|} \le L_{n-1} + \frac{2\hat C}{C_-} \frac{\Delta x}{\Delta t}. \tag{2.21}$$

A similar computation yields, for $n = 1$,

$$\frac{|v_j^1 - v_i^1|}{|x_i - x_j|} \le L_0 + \frac{2\tilde{C}}{C_-}, \tag{2.22}$$

so that combining (2.21) with (2.22) and iterating back, we have

$$\begin{aligned} L_n &\le L_{n-1} + \frac{2\hat{C}}{C_-}\frac{\Delta x}{\Delta t} \\ &\le \cdots \le L_1 + \frac{T-\Delta t}{\Delta t}\frac{2\hat{C}}{C_-}\frac{\Delta x}{\Delta t} \le L_0 + \frac{T-\Delta t}{\Delta t}\frac{2\hat{C}}{C_-}\frac{\Delta x}{\Delta t} + \frac{2\tilde{C}}{C_-}. \end{aligned} \tag{2.23}$$

Last, it is possible to get a finite limit in (2.23), if and only if $\Delta x = O(\Delta t^2)$. □

Remark 2.2. Once we have proved Lipschitz continuity, we can also obtain L^∞ boundedness using the same arguments as in [F].

Remark 2.3. The condition $\Delta x/\Delta t \to 0$ is not unnatural in this class of schemes. As has been remarked elsewhere, the numerical domain of dependence enlarges (as it is easy to see from (1.4)) as Δt increases, so that very large Courant numbers are allowed without loss of stability. Of course, managing this larger domain of dependence requires some caution to keep the computational complexity as low as possible in the minimization phase. Roughly speaking, practical implementations of the scheme perform the minimization of F_j at any node and any time step by some suitable descent method. (The interested reader can find details in [CFF], [FF3].)

Remark 2.4. The local truncation error of the scheme (see [FF3]) is of order

$$\Delta t^p + \frac{\Delta x^{r+1}}{\Delta t}$$

(with order p of approximation of characteristics, order r of space interpolation) so that at least in the case $r = 0$ (piecewise constant reconstruction), the consistency of the scheme itself requires that the Courant number should diverge.

3. Applications to various reconstruction operators. In this section we prove the applicability of the previous results to one-dimensional high-order reconstruction of polynomial type. We assume the reconstruction to be of rth order in the Newton form

$$\begin{aligned} I[V](x) = V[x_{j_0}] &+ V[x_{j_0}, x_{j_1}](x - x_{j_0}) + \cdots \\ &+ V[x_{j_0}, \ldots, x_{j_r}](x - x_{j_0}) \cdots (x - x_{j_{r-1}}), \end{aligned} \tag{3.1}$$

where $x_{j_0}, \ldots, x_{j_r}$ are $r+1$ adjacent nodes so that $\max(x_{j_0}, \ldots, x_{j_r}) - \min(x_{j_0}, \ldots, x_{j_r}) = r\Delta x$ and, moreover,

$$x \in (\min(x_{j_0}, \ldots, x_{j_r}), \max(x_{j_0}, \ldots, x_{j_r})) \subset U(x). \tag{3.2}$$

This definition includes both Lagrange polynomial interpolations, for which the reconstruction stencil is fixed once x is fixed, and ENO reconstructions, for which it depends on the solution itself (see [S]). The divided differences are defined, as usual, by

$$V[x_{j_0}] = v_{j_0},$$

$$V[x_{j_0}, \ldots, x_{j_k}] = \frac{V[x_{j_1}, \ldots, x_{j_k}] - V[x_{j_0}, \ldots, x_{j_{k-1}}]}{x_{j_k} - x_{j_0}} \quad (k = 1, \ldots, r).$$

Note that, although in principle the nodes $x_{j_0},\dots,x_{j_k}$ need neither to be adjacent nor to satisfy $x\in(\min x_{j_i},\max x_{j_i})$, it is possible to reorder the nodes so that both conditions would be satisfied.

According to its definition, we can bound a generic kth order divided difference as follows:

$$|V[x_{j_0},\dots,x_{j_k}]|\le\frac{2\max|V[x_{j_i},\dots,x_{j_{k+i-1}}]|}{k\Delta x} \tag{3.3}$$

in which the max is performed for $x_{j_i},\dots,x_{j_{k+i-1}}\in U(x)$. Let us now denote more precisely the constant C in (2.1) by C_r (depending on the order r of the reconstruction). To prove (2.1), we start from the second divided difference

$$|V[x_{j_0},x_{j_1},x_{j_2}]|=\frac{|v_{j_0}-2v_{j_1}+v_{j_2}|}{2\Delta x^2},$$

and, hence,

$$\begin{aligned}|V[x_{j_0},x_{j_1},x_{j_2},x_{j_3}]|&\le\frac{2\max|v_{j_i}-2v_{j_{i+1}}+v_{j_{i+2}}|}{3!\Delta x^3}\\&\vdots\\|V[x_{j_0},\dots,x_{j_r}]|&\le\frac{2^{r-2}\max|v_{j_i}-2v_{j_{i+1}}+v_{j_{i+2}}|}{r!\Delta x^r}.\end{aligned}$$

Plugging such bounds into (3.1), we get an estimate in the form (2.1); that is,

$$\begin{aligned}|I[V](x)-I_1[V](x)|&\le|V[x_{j_0},x_{j_1},x_{j_2}](x-x_{j_0})(x-x_{j_1})\\&\quad+\cdots+V[x_{j_0},\dots,x_{j_r}](x-x_{j_0})\cdots(x-x_{j_{r-1}})|\\&\le\frac{|v_{j_0}-2v_{j_1}+v_{j_2}|}{2\Delta x^2}M_2\Delta x^2\\&\quad+\cdots+\frac{2^{r-2}\max|v_{j_i}-2v_{j_{i+1}}+v_{j_{i+2}}|}{r!\Delta x^r}M_r\Delta x^r\\&\le\max|v_{j_i}-2v_{j_{i+1}}+v_{j_{i+2}}|\sum_{k=2}^{r}\frac{M_k2^{k-2}}{k!},\end{aligned} \tag{3.4}$$

where

$$\begin{aligned}M_k&:=\frac{1}{\Delta x^k}\max_{x\in(x_{j_0},x_{j_{k-1}})}|(x-x_{j_0})\cdots(x-x_{j_{k-1}})|\\&=\max_{t\in(0,k-1)}|t(t-1)\cdots(t-k+1)|.\end{aligned}$$

It remains to check that $C_r<1$; that is,

$$\sum_{k=2}^{r}\frac{M_k2^{k-2}}{k!}<1. \tag{3.5}$$

Indeed, the first values M_k may be either computed by algebraic manipulations (up to M_5), or estimated by simply plotting the polynomials $t(t-1)\cdots(t-k+1)$ on the interval $[0,k-1]$. It turns out that $M_2=1/4$, $M_3=2\sqrt{3}/9\approx0.3849$, $M_4=1$,

$M_5 \approx 3.6314$, $M_6 < 17$, and so on. Accordingly, the computation of the left-hand side of (3.5) for various values of r gives $C_2 = 1/8 = 0.125$, $C_3 \approx 0.2533$, $C_4 \approx 0.42$, $C_5 \approx 0.6621$, $C_6 \approx 1.04$. We can conclude that by this technique, polynomial and ENO reconstructions up to the fifth order can be proved to satisfy (2.1).

Remark 3.1. While this paper was under review, we realized (see [CFR]) a property which allows us to extend this theory to Lagrange and weighted essentially nonoscillatory (WENO) interpolations up to the degree $r = 9$, provided the reconstruction stencil is "balanced" in the sense that

$$|h_+ - h_-| \le 1. \tag{3.6}$$

Roughly speaking, the idea is that the interpolation $I[V](x)$ could be obtained as a sum

$$I[V] = w_1(x)p_1(x) + \cdots + w_q(x)p_q(x) \tag{3.7}$$

in which the polynomials w_i and p_i are such that $\deg w_i + \deg p_i = r$ and the p_i are constructed on smaller stencils which include the point x (this form is at the base of WENO interpolations). With $\deg p_i \le 5$, one has $r \le 9$, and it turns out that $w_i(x) \ge 0$ for all i and $\sum_i w_i(x) = 1$, so that by (3.7)

$$\min(p_1(x), \dots, p_q(x)) \le I[V](x) \le \max(p_1(x), \dots, p_q(x))$$

and therefore (2.1), being satisfied by all p_i, is also satisfied by $I[V]$.

4. The main convergence result. Last, we present in this section the main result of the paper, that is, a convergence theorem for the scheme (1.4).

THEOREM 4.1. *Consider the scheme* (1.4) *applied to* (1.1). *Assume that* (2.15), (2.16) *hold, that* $\Delta x = O(\Delta t^2)$, *and that* v_0 *is Lipschitz continuous. Then, the numerical solutions* V^n *satisfy*

$$\|I[V^n] - v(n\Delta t)\|_\infty \to 0$$

for $0 \le n \le T/\Delta t$, *as* $\Delta t \to 0$.

Proof. We rewrite exact and approximate solutions at node x_j and time $n\Delta t$ as

$$v(x_j, n\Delta t) = \Delta t H^*(a_j) + v(x_j + a_j\Delta t, (n-1)\Delta t), \tag{4.1}$$

$$v_j^n = \Delta t H^*(\bar\alpha_j) + I[V^{n-1}](x_j + \bar\alpha_j \Delta t), \tag{4.2}$$

where the argmin in (1.3) and (1.4) has been made explicit. Working as in Theorem 2.1, we give a first unilateral bound as

$$\begin{aligned} v_j^n - v(x_j, n\Delta t) &\le v(x_j + a_j\Delta t, (n-1)\Delta t) - I[V^{n-1}](x_j + a_j\Delta t) \\ &\le |v(x_j + a_j\Delta t, (n-1)\Delta t) - I[W^{n-1}](x_j + a_j\Delta t)| \\ &\quad + |I[W^{n-1}](x_j + a_j\Delta t) - I[V^{n-1}](x_j + a_j\Delta t)|, \end{aligned} \tag{4.3}$$

in which we have used the sequence

$$W^k := \{v(x_i, t_k)\}_i.$$

Under our assumptions, the numerical solution is Lipschitz by Theorem 2.1. Since this is also true for the exact solution (see [L]), all reconstructions in the right-hand

side of (4.3) are performed on Lipschitz sequences and therefore (2.15) applies. (We note that, even in the one-dimensional case, we cannot apply the bound of Lemma 2.1 since a_j is used in the scheme instead of $\bar{\alpha}_j$.) Now, the first term of the right-hand side of (4.3) is the reconstruction error for the Lipschitz continuous function v, and we have

$$\begin{aligned}&|v(x_j + a_j\Delta t, (n-1)\Delta t) - I[W^{n-1}](x_j + a_j\Delta t)| \\ &\le |v(x_j + a_j\Delta t, (n-1)\Delta t) - I_1[W^{n-1}](x_j + a_j\Delta t)| \\ &\quad + |I_1[W^{n-1}](x_j + a_j\Delta t) - I[W^{n-1}](x_j + a_j\Delta t)| \to 0\end{aligned}$$

with first-order convergence with respect to Δx, resulting from the convergence of P_1 interpolation to a Lipschitz function and from (2.15) (written with W^{n-1} instead of V). For the second term in the right-hand side of (4.3) we can write, using (2.15) and the monotonicity of the reconstruction I_1,

(4.4)

$$\begin{aligned}|I[W^{n-1}](x_j + a_j\Delta t) - I[V^{n-1}](x_j + a_j\Delta t)| &\le |I_1[W^{n-1}](x_j + a_j\Delta t) \\ &\quad - I_1[V^{n-1}](x_j + a_j\Delta t)| + 2\tilde{C}\Delta x \\ &\le |W^{n-1} - V^{n-1}|_\infty + 2\tilde{C}\Delta x,\end{aligned}$$

and therefore, using the reverse estimate which can be obtained in the same way,

$$|W^n - V^n|_\infty \le |W^{n-1} - V^{n-1}|_\infty + 2\tilde{C}\Delta x,$$

which implies $|W^n - V^n|_\infty \to 0$ since $|W^0 - V^0|_\infty = 0$ and $\Delta x = O(\Delta t^2)$. Last, by (2.15) and the Lipschitz continuity of v, we easily get

$$\|I[V^n] - v(n\Delta t)\|_\infty \le |W^n - V^n|_\infty + \check{C}\Delta x$$

and this proves the theorem for $\Delta t, \Delta x \to 0$. □

Remark 4.1. In this theorem, the condition $\Delta x = O(\Delta t^2)$ is only required in order to ensure Lipschitz stability. The proof itself requires the weaker condition $\Delta x = o(\Delta t)$, which is in turn related to consistency as already remarked in section 2.

Remark 4.2. Since the scheme experimentally appears to be convergent under any $\Delta t/\Delta x$ relationship, we should infer that this convergence result is not optimal. This might be due to an intrinsic limitation of this technique of proof, as well as to the very weak assumptions on the reconstruction operator. In fact, assumption (3.2) alone does not even allow the scheme to be stable in the sense of Von Neumann (see [FF2] where a sort of "balancing" of the reconstruction stencil turns out to be necessary). Here, the nonlinear min operation plays a crucial role in stabilizing the scheme.

REFERENCES

[BS] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asympt. Anal., 4 (1991), pp. 271–283.

[CFF] E. Carlini, M. Falcone, and R. Ferretti, *An Efficient Algorithm for Hamilton–Jacobi Equations in High Dimension*, submitted.

[CFR] E. Carlini, R. Ferretti, and G. Russo, *A Weighted Essentially Non-Oscillatory, Large Time-Step Scheme for Hamilton–Jacobi Equations*, manuscript.

[FF1] M. Falcone and R. Ferretti, *Discrete–time high–order schemes for viscosity solutions of Hamilton–Jacobi–Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.

[FF2] M. Falcone and R. Ferretti, *Convergence analysis for a class of high- order semi-Lagrangian advection schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 909–940.

[FF3] M. Falcone and R. Ferretti, *Semi–Lagrangian schemes for Hamilton–Jacobi equations, discrete representation formulae and Godunov methods*, J. Comput. Phys., 175 (2002), pp. 559–575.

[FFM] M. Falcone, R. Ferretti, and T. Manfroni, *Optimal discretization steps in semi–Lagrangian approximation of first order PDEs*, in Numerical Methods for Viscosity Solutions and Applications (Heraklion, 1999), Ser. Adv. Math. Appl. Sci. 59, M. Falcone and C. Makridakis, eds., World Scientific, River Edge, NJ, 2001, pp. 95–117.

[FG] M. Falcone and T. Giorgi, *An approximation scheme for evolutive Hamilton–Jacobi equations*, in Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming, W. M. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 289–303.

[F] R. Ferretti, *Equicontinuity of some large time–step approximations to convex Hamilton–Jacobi equations*, in Proceedings of the Workshop "HJB2000," INRIA, Paris, 2000.

[JP] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[L] P. L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, London, 1982.

[LS] P. L. Lions and P. Souganidis, *Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton–Jacobi equations*, Numer. Math., 69 (1995), pp. 441–470.

[LT1] C.-T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.

[LT2] C. T. Lin and E. Tadmor, *L^1 stability and error estimates for Hamilton–Jacobi solutions*, Numer. Math., 87 (2001), pp. 701–735.

[OS] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[S] C. W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations (Cetraro, 1997), Lecture Notes in Math. 1697, Springer–Verlag, Berlin, 1998, pp. 325–432.

[SC] A. Staniforth and J. Coté, *Semi–Lagrangian integration schemes for atmospheric models - a review*, Monthly Weather Review, 119 (1991), pp. 2206–2223.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2254–2269

ON THE CONVERGENCE OF THE FOURIER APPROXIMATION FOR EIGENVALUES AND EIGENFUNCTIONS OF DISCONTINUOUS PROBLEMS*

M. S. MIN† AND D. GOTTLIEB†

Abstract. In this paper, we consider a model eigenvalue problem with discontinuous coefficients in order to study the convergence of the Fourier methods applied to this problem. We prove that the rate of convergence of the Fourier–Galerkin method is third order for the eigenvalues and order 2.5 for the eigenfunctions. For the Fourier collocation method we obtained only second order accuracy.

We also show that the Fourier collocation method can be improved by a preprocessing of the coefficients.

The theory is confirmed by numerical results.

Key words. discontinuous problems, Fourier–Galerkin method, Fourier collocation method, minmax principle

AMS subject classifications. 65N15, 65N25, 65N35

PII. S0036142902403012

1. Introduction.

The paper is motivated by an issue arising in the use of spectral methods in nonlinear optics. The Fourier methods when applied to problems in nonlinear optics are extremely fast, and if the problem is smooth they provide high order accuracy. However, when different media are considered, the coefficients are only piecewise smooth and the accuracy is lost.

In order to understand the phenomenon, and as a first step to improve the accuracy of the Fourier schemes in those circumstances, we consider in this paper a model eigenvalue problem with piecewise constant coefficients and study the convergence of the Fourier–Galerkin and Fourier collocation methods to the eigenvalues and the eigenfunctions of this problem. The surprising fact is that the order of convergence of the eigenvalues obtained by the Fourier–Galerkin method is cubic. When the Fourier collocation method is applied, the results are only second order. Those results are proven and supported by numerical computations.

It turns out that, by preprocessing the discontinuous coefficients, one can improve the accuracy of the collocation method. In fact, if one uses the point values of the finite Fourier series of the coefficients instead of the point values of the coefficients themselves, one recovers third order accuracy for the eigenvalues and order 2.5 for the eigenfunction.

The paper is organized as follows. In section 2, we present the problem and show some of the eigenvalues and eigenfunctions. In section 3, we rewrite the problem in its variational form and quote some relevant facts. In section 4, we discuss the Fourier–Galerkin method and prove the order of accuracy. Section 5 is devoted to the Fourier collocation method and the error estimates of this method. In section 6, we show how to improve the accuracy of the collocation method.

*Received by the editors February 21, 2002; accepted for publication (in revised form) June 24, 2002; published electronically January 7, 2003. This work was supported by AFOSR grant F49620-99-1-0077, DOE grant DE-FG02-98ER25346, and NSF grant DMS-9804985.

http://www.siam.org/journals/sinum/40-6/40301.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (msmin@cfm.brown.edu, dig@cfm.brown.edu).

We regard this paper as the first step toward recovering exponential accuracy for this problem.

2. The discontinuous eigenvalue problem. Consider the following eigenvalue problem with a piecewise constant coefficient:

$$-\frac{d^2u}{dx^2} = \lambda\epsilon(x)u \quad \text{for} \quad x \in (-\pi,\pi), \tag{2.1}$$

where $\epsilon(x) = 1$ in $(-\pi, 0)$ and $\epsilon(x) = \beta^2$ in $[0,\pi)$, $\beta \neq 1$. The $H_p^2[-\pi,\pi]$ eigenfunction $u_l(x)$ (the p stands for periodic) is given by

$$u_l(x) = \begin{cases} C\cos(\sqrt{\lambda}_l x) + \beta D\sin(\sqrt{\lambda}_l x), & -\pi \le x \le 0, \\ C\cos(\beta\sqrt{\lambda}_l x) + D\sin(\beta\sqrt{\lambda}_l x), & 0 \le x \le \pi, \end{cases} \tag{2.2}$$

where the constants C, D and the eigenvalue λ_l are determined by the demand that the system

$$\begin{aligned} C(\cos\sqrt{\lambda}\pi - \cos\beta\sqrt{\lambda}\pi) + D(-\beta\sin\sqrt{\lambda}\pi - \sin\beta\sqrt{\lambda}\pi) = 0, \\ C(\sin\sqrt{\lambda}\pi + \beta\sin\beta\sqrt{\lambda}\pi) + D(\beta\cos\sqrt{\lambda}\pi - \beta\cos\beta\sqrt{\lambda}\pi) = 0 \end{aligned}$$

has a nontrivial solution. Considering the case $\beta = 2$, for $y = \cos\sqrt{\lambda}\pi$, the eigenvalues λ satisfy the equation

$$(y-1)(9y^2 + 9y + 2) = 0,$$

and so there are families of eigenvalues determined by

$$\cos\sqrt{\lambda}\pi = 1, -\frac{1}{3}, \quad \text{or} \quad -\frac{2}{3}. \tag{2.3}$$

The first five analytic eigenvalues (with six digits of precision) and the corresponding eigenvectors are shown in Figure 1. For comparison, we also carry the same procedure for $\beta = 3$, where the analytic eigenvalues are determined by

$$\cos\sqrt{\lambda}\pi = \pm 1 \ \text{ or } \ \pm\frac{1}{4}. \tag{2.4}$$

In this paper, we examine the rate of convergence of the Fourier methods (Galerkin and collocation) as a first step in an effort to improve the rate of convergence and be able to also apply the Fourier methods for this discontinuous problem.

3. The variational formulation. We define two inner products:

$$a(u,v) = \int_{-\pi}^{\pi} u'(x)\overline{v'(x)}dx, \tag{3.1}$$

$$(u,v) = \int_{-\pi}^{\pi} u(x)\overline{v(x)}\epsilon(x)dx. \tag{3.2}$$

Following Strang and Fix [7, p. 220], the eigenvalue problem (2.1) can be presented in the following variational form: finding a scalar λ and a function $u \in H_p^1[-\pi,\pi]$ such that

$$a(u,v) = \lambda(u,v) \tag{3.3}$$

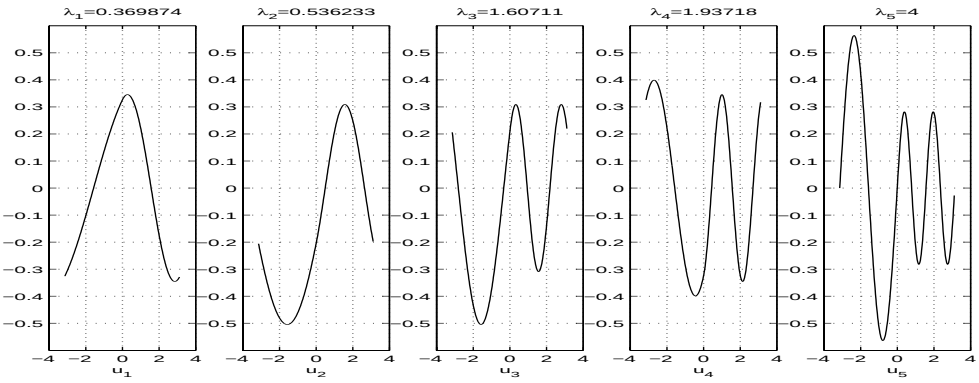


FIG. 1. *The first five analytic eigenvalues and eigenfunctions.*

for all v in the Hilbert space $H_p^1[-\pi,\pi]$. Note that $a(u,v)$ is Hermitian.

Our proofs will use extensively the minmax principle [3].

THEOREM 3.1. *Let λ_l denote the eigenvalues of* (2.1) *and S_l be any l-dimensional subspace of $H_p^1[-\pi,\pi]$. Then, for $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_l \ldots$,*

$$\lambda_l = \min_{S_l \subset H_p^1[-\pi,\pi]} \max_{v \in S_l} \frac{a(v,v)}{(v,v)}. \tag{3.4}$$

In this paper, we will use sharper characterizations of the eigenvalues.

LEMMA 3.2. *Let λ_i be arranged in an ascending order and define*

$$E_{i,j} = span\{u_i, \ldots, u_j\},$$

where u_i is the eigenfunction corresponding to the eigenvalue λ_i. Then

$$\lambda_l = \max_{v \in E_{k,l}} \frac{a(v,v)}{(v,v)}, \quad k \le l, \tag{3.5}$$

$$\lambda_l = \min_{v \in E_{l,m}} \frac{a(v,v)}{(v,v)}, \quad l \le m. \tag{3.6}$$

4. Fourier–Galerkin method. It is natural to consider the Fourier method to approximate the periodic problem. Here, we introduce the Fourier–Galerkin method applied to the variational formulation for approximating the eigenvalues and eigenfunctions.

Let $\mathcal{P}_\mathcal{N}$ be the space of the trigonometric polynomials of degree $N/2$ defined as

$$\mathcal{P}_\mathcal{N} = \text{span}\{e^{ikx} | -N/2 \le k \le N/2\}. \tag{4.1}$$

In this subspace, we look for λ^N and u^N such that

$$a(u^N, v^N) = \lambda^N (u^N, v^N) \text{ for all } v^N \in \mathcal{P}_\mathcal{N}; \tag{4.2}$$

in other words,

$$\int_{-\pi}^{\pi} (u^N(x))' \overline{(v^N(x))'} dx = \lambda^N \int_{-\pi}^{\pi} u^N(x) \overline{v^N(x)} \epsilon(x) dx. \tag{4.3}$$

4.1. Numerical scheme and its results. The approximate eigenfunction u^N in the subspace $\mathcal{P}_\mathcal{N}$ is expanded by

$$u^N = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} (\hat{u}^N)_k e^{ikx}. \tag{4.4}$$

Substituting u^N into the variational formulation (4.2) with $v^N = e^{inx}$, and denoting the vector of the coefficients by $\hat{\mathbf{u}}^N$, we get a generalized eigenvalue problem in a matrix form as

$$K\hat{\mathbf{u}}^N = \lambda^N M\hat{\mathbf{u}}^N, \tag{4.5}$$

where

$$K_{nk} = \int_{-\pi}^{\pi} (k \cdot n) e^{i(k-n)x} dx \quad \text{and} \quad M_{nk} = \int_{-\pi}^{\pi} e^{i(k-n)x} \epsilon(x) dx. \tag{4.6}$$

Solving the matrix eigenvalue problem (4.5) computationally using a proper eigensolver, we obtain the approximate lth eigenvalues, $\lambda_l^N (l \le N)$, and the set of orthogonal vectors $\hat{\mathbf{u}}_l{}^N = [(\hat{u}_l^N)_{-N/2}, \ldots, (\hat{u}_l^N)_{N/2}]^T$ which is used to approximate the lth eigenfunction u_l as a finite Fourier series u_l^N. In Tables 1 and 2, the orders of the relative errors for $\lambda_l^N - \lambda_l$ and the discrete L_2-errors of $u_l - u_l^N$ are provided for the first five eigenvalues in ascending order and for the associated eigenfunctions. We note the surprising fact that the Galerkin approximation to the eigenvalue problem (2.1) converges with *third order accuracy* for the eigenvalues and order 2.5 for the eigenfunctions even though the eigenfunctions are only in H_p^2. In fact, we will show in Lemma 4.5 that the eigenfunctions are in $H_p^{\frac{5}{2}-\epsilon}$ for any $\epsilon > 0$.

4.2. Error estimates for eigenvalues and eigenfunctions. In this section, we provide the error estimates for the approximate eigenvalues and eigenfunctions for the Fourier–Galerkin method.

We first treat the approximate eigenvalues. Let $P_N u$ be the $\frac{N}{2}$th order truncated Fourier series of u. (We will denote also $P = P_N$.) It is clear that it satisfies

$$a(u - P_N u, v^N) = 0 \ \text{ for all } v^N \in \mathcal{P}_\mathcal{N}. \tag{4.7}$$

It is true that the minmax principle is also valid for the Galerkin procedure:

$$\lambda_l^N = \min_{S_l \subset \mathcal{P}_\mathcal{N}} \max_{v \in S_l} \frac{a(v,v)}{(v,v)}. \tag{4.8}$$

LEMMA 4.1. *Let λ_l^N be the approximation to λ_l which is obtained by the Galerkin procedure. Then*

$$\lambda_l \le \lambda_l^N \le \lambda_l \max_{v \in E_{1,l}} \frac{(v,v)}{(Pv,Pv)}.$$

Proof. Due to the minmax principle (3.4) and (4.8), we have

$$\begin{aligned} \lambda_l &= \min_{S_l \subset H_p^1[-\pi,\pi]} \max_{v \in S_l} \frac{a(v,v)}{(v,v)} \\ &\le \min_{S_l \subset \mathcal{P}_\mathcal{N}} \max_{v \in S_l} \frac{a(v,v)}{(v,v)} = \lambda_l^N. \end{aligned}$$

TABLE 1

The relative errors of eigenvalues for the case $\beta = 2$ and the discrete L_2-errors of $u_i - u_i^N$ for the Fourier–Galerkin method.

λ_i	N	λ_i^N	$\frac{(\lambda_i^N - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^N\|_{l_2}$	Order
	16	0.36991	1.0269(-4)		4.1430(-4)	
	32	0.36988	1.5247(-5)	2.7517	8.9399(-5)	2.2124
0.36987	64	0.36988	2.0853(-6)	2.8702	1.7382(-5)	2.3627
	128	0.36988	2.7295(-7)	2.9335	3.2192(-6)	2.4328
	256	0.36987	3.4960(-8)	2.9649	5.8234(-7)	2.4668
	16	0.53628	8.6637(-5)		5.7658(-4)	
	32	0.53624	1.0678(-5)	3.0203	1.0045(-4)	2.5210
0.53623	64	0.53623	1.3287(-6)	3.0066	1.7617(-5)	2.5114
	128	0.53623	1.6580(-7)	3.0025	3.1020(-6)	2.5057
	256	0.53623	2.0643(-8)	3.0057	5.4732(-7)	2.5028
	16	1.60758	2.8694(-4)		1.8606(-3)	
	32	1.60717	3.2907(-5)	3.1243	3.0781(-4)	2.5957
1.60712	64	1.60712	4.0121(-6)	3.0360	5.3130(-5)	2.5344
	128	1.60712	4.9821(-7)	3.0095	9.3136(-6)	2.5121
	256	1.60712	6.2253(-8)	3.0005	1.6412(-6)	2.5046
	16	1.93833	5.9203(-4)		2.4795(-3)	
	32	1.93734	8.1996(-5)	2.8520	4.8917(-4)	2.3416
1.93718	64	1.93720	1.0998(-5)	2.8983	9.2481(-5)	2.4031
	128	1.93718	1.4324(-6)	2.9408	1.6966(-5)	2.4465
	256	1.93718	1.8343(-7)	2.9651	3.0582(-6)	2.4719
	16	4.00094	2.3385(-4)		3.4923(-3)	
	32	4.00002	5.9398(-6)	5.2990	2.6560(-4)	3.7168
4.00000	64	4.00000	1.7675(-7)	5.0706	2.2714(-5)	3.5476
	128	4.00000	5.4572(-9)	5.0174	1.9917(-6)	3.5115
	256	4.00000	1.6993(-10)	5.0051	1.7822(-7)	3.4822

TABLE 2

The relative errors of eigenvalues for the case $\beta = 3$ and the discrete L_2-errors of $u_i - u_i^N$ for the Fourier–Galerkin method.

λ_i	N	λ_i^N	$\frac{(\lambda_i^N - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^N\|_{l_2}$	Order
	16	0.17606	1.5651(-4)		8.6969(-4)	
	32	0.17604	2.3223(-5)	2.7526	1.7589(-4)	2.3059
0.17603	64	0.17603	3.1759(-6)	2.8703	3.2960(-5)	2.4159
	128	0.17603	4.1565(-7)	2.9337	5.9857(-6)	2.4611
	256	0.17603	5.3176(-8)	2.9665	1.0719(-6)	2.4814
	16	0.33704	4.3531(-4)		2.1237(-3)	
	32	0.33691	5.3609(-5)	3.0215	3.7423(-4)	2.5046
0.33690	64	0.33690	6.6757(-6)	3.0055	6.6210(-5)	2.4988
	128	0.33690	8.3365(-7)	3.0014	1.1722(-5)	2.4978
	256	0.33690	1.0418(-7)	3.0004	2.0744(-6)	2.4984
	16	1.00005	4.7257(-5)		2.1107(-3)	
	32	1.00000	1.8636(-6)	4.6643	3.6870(-4)	2.5172
1.00000	64	1.00000	6.6628(-8)	4.8059	6.5495(-5)	2.4930
	128	1.00000	2.2386(-9)	4.8955	1.1610(-5)	2.4959
	256	1.00000	7.2292(-11)	4.9526	2.0544(-6)	2.4986
	16	2.02204	3.4072(-3)		2.6579(-2)	
	32	2.01587	3.4220(-4)	3.3157	3.0104(-3)	3.1423
2.01518	64	2.01526	4.0586(-5)	3.0758	4.5219(-4)	2.7349
	128	2.01519	5.0069(-6)	3.0190	7.4652(-5)	2.5987
	256	2.01518	6.2380(-7)	3.0048	1.2792(-5)	2.5449
	16	2.50541	3.0620(-3)		1.0310(-2)	
	32	2.49866	3.5823(-4)	3.0955	2.1636(-3)	2.2526
2.49776	64	2.49788	4.6049(-5)	2.9596	4.2364(-4)	2.3525
	128	2.49778	5.9303(-6)	2.9570	8.0203(-5)	2.4011
	256	2.49776	7.5555(-7)	2.9725	1.4749(-5)	2.4430

Thus only the upper bound of the approximate eigenvalue is left to be investigated.

Let $PE_{1,l}$ be spanned by $Pu_1, \ldots, Pu_l$. Then it is clear that $PE_{1,l}$ is the l-dimensional subspace of $\mathcal{P}_{\mathcal{N}}$. Using the minmax principle (4.8),

$$\lambda_l^N \le \max_{v \in PE_{1,l}} \frac{a(v,v)}{(v,v)} = \max_{v \in E_{1,l}} \frac{a(Pv,Pv)}{(Pv,Pv)}. \tag{4.9}$$

Note that in [7] we have

$$a(v,v) = a(Pv,Pv) + 2a(v-Pv,Pv) + a(v-Pv,v-Pv). \tag{4.10}$$

From (4.7), we know that $a(v-Pv,Pv)$ always vanishes for all Pv in the space $\mathcal{P}_{\mathcal{N}}$. Then we have $a(Pv,Pv) \le a(v,v)$. Thus,

$$\lambda_l^N \le \max_{v \in E_{1,l}} \frac{a(v,v)}{(Pv,Pv)} = \max_{v \in E1,_l} \frac{a(v,v)}{(v,v)} \cdot \frac{(v,v)}{(Pv,Pv)} \le \lambda_l \cdot \max_{v \in E_{1,l}} \frac{(v,v)}{(Pv,Pv)}.$$

The last inequality is a by-product of (3.5). Thus the lemma is proven. ∎

The issue is how close $(P_N v, P_N v)$ is to (v,v) for $v \in E_{1,l}$. One would expect the second order accuracy in N because of the smoothness of the eigenfunctions u_i. However, we will show that it is really third order. We start by examining the Fourier coefficients of the eigenfunctions.

LEMMA 4.2. *The Fourier coefficients $(\hat{u_l})_k$ of the eigenfunction u_l decay as $O(k^{-3})$; in fact,*

$$(\hat{u_l})_k \le Ck^{-3}\left\{|u_l(0)| + \frac{1}{k}|u_l^{'}(0)| + \frac{\lambda_l}{k}||u_l||\right\}, \tag{4.11}$$

where $||u_l||$ is the L_2-norm of u_l.

Proof. Letting $u_l = u$ for simplicity, and using the fact that $u^{'}$ is continuous,

$$\hat{u}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} ue^{-ikx}dx = \frac{1}{2\pi ik}\int_{-\pi}^{\pi} u'e^{-ikx}dx = -\frac{1}{2\pi k^2}\int_{-\pi}^{\pi} u''e^{-ikx}dx. \tag{4.12}$$

Substituting $u'' = -\lambda\epsilon u$ into (4.12) and, for convenience, using the notation $\mu = \beta^2 - 1$, we have

$$\begin{aligned}
\hat{u}_k &= \frac{1}{2\pi k^2}\int_{-\pi}^{\pi} \lambda\epsilon ue^{-ikx}dx = \frac{\lambda}{2\pi k^2}\left(\int_{-\pi}^{0} ue^{-ikx}dx + \beta^2\int_{0}^{\pi} ue^{-ikx}dx\right) \\
&= \frac{\lambda}{2\pi ik^3}\left(\mu\{(-1)^k u(\pi) - u(0)\} + \int_{-\pi}^{0} u'e^{-ikx}dx + \beta^2\int_{0}^{\pi} u'e^{-ikx}dx\right) \\
&= \frac{\lambda}{2\pi ik^3}\left(\mu[(-1)^k u(\pi) - u(0)] + \frac{\mu}{ik}[(-1)^k u'(\pi) - u'(0)] + \frac{1}{ik}\int_{-\pi}^{\pi} \lambda\epsilon ue^{-ikx}dx\right).
\end{aligned}$$

Therefore, the lemma is proven. ∎

We are ready now for the next lemma.

LEMMA 4.3.

$$\max_{v \in E1,l} \frac{(v,v)}{(Pv,Pv)} \le 1 + ClN^{-3},$$

where the constant C is independent of N and l.

Proof. We first note that

$$\frac{(v,v)}{(Pv,Pv)} = \frac{1}{1-\frac{(v,v)-(Pv,Pv)}{(v,v)}}. \tag{4.13}$$

Since v is in $E_{1,l}$, it can be represented by $v = \sum_{i=1}^{l} \alpha_i u_i$. We also have

$$(Pv,Pv) = (v,v) - (v-Pv,v) - \overline{(v-Pv,v)} + (v-Pv, v-Pv).$$

Thus we get

$$\begin{aligned}\frac{(v,v)-(Pv,Pv)}{(v,v)} &\le \frac{2|(v-Pv,v)|}{(v,v)} \\ &= \frac{2\sum_{i,j=1}^{l}|\alpha_i||\bar{\alpha}_j||(u_i-Pu_i,u_j)|}{(\sum_{i=1}^{l}|\alpha_i|^2)} \\ &\le 2l \max_{i,j=1,\dots,l}|(u_i-Pu_i,u_j)|.\end{aligned}$$

Now we have

$$\begin{aligned}|(u_i-Pu_i,u_j)| &= \left|\left(\sum_{|k|>\frac{N}{2}}(\hat{u_i})_k e^{ikx}, \sum_{n=\infty}^{\infty}(\hat{u_j})_n e^{inx}\right)\right| \\ &= \left|\sum_{|k|>\frac{N}{2}}\sum_{n=\infty}^{\infty}(\hat{u_i})_k\overline{(\hat{u_j})_n}\int_{-\pi}^{\pi}e^{i(k-n)x}\epsilon(x)dx\right| \\ &\le \sum_{|k|>\frac{N}{2}}\sum_{\substack{n=\infty\\ n\neq k}}^{\infty}|(\hat{u_i})_k|\cdot\left|\overline{(\hat{u_j})_n}\right|\cdot\frac{(\beta^2-1)|((-1)^{k-n}-1)|}{|k-n|} \\ &\quad + \sum_{|k|>\frac{N}{2}}\sum_{\substack{n=\infty\\ n=k}}^{\infty}|(\hat{u_i})_k|\cdot\left|\overline{(\hat{u_j})_n}\right|\cdot(\beta^2+1)\pi.\end{aligned}$$

Recalling (4.11), where $|\hat{u}_k|$ decays like $O(k^{-3})$ at least, we get

$$|(u_i-Pu_i,u_j)| \le CN^{-3},$$

where C is a positive constant. Finally, we have

$$\begin{aligned}\frac{(v,v)}{(Pv,Pv)} &= \frac{1}{1-\frac{(v,v)-(Pv,Pv)}{(v,v)}} \\ &\le 1 + 2\frac{(v,v)-(Pv,Pv)}{(v,v)} \\ &\le 1 + ClN^{-3}.\end{aligned}$$

Thus the lemma is proven. ∎

We can now state the following theory.

THEOREM 4.4. *Let λ_l^N be the Fourier–Galerkin approximation to the eigenvalue λ_l. Then*

$$|\lambda_l - \lambda_l^N| \le Cl\lambda_l N^{-3},$$

where the constant C depends only on the values of $u_i(0)$, $u_i'(0)$, and the L_2-norm of u_i for all $i \le l$.

Now we are ready to treat the eigenvectors. Following Strang and Fix [7, p. 234], we can state

$$||u_l - u_l^N|| \le ||u_l - P_N u_l||,$$

where $P_N u_l$ is the finite Fourier series of u_l. For the right-hand side we have the following estimate.

LEMMA 4.5.

$$||u_l - P_N u_l|| \le CN^{-2.5}.$$

Proof. By the Parseval equality, and using Lemma 4.2, which states the Fourier coefficients of u_l decay cubicly, we get

$$\begin{aligned} ||u_l - P_N u_l|| &= \left(\sum_{|k|>\frac{N}{2}} |(\hat{u_l})_k|^2 \right)^{\frac{1}{2}} \\ &\le C \left(\sum_{|k|>\frac{N}{2}} |k|^{-6} \right)^{\frac{1}{2}} \\ &\le CN^{-2.5}. \end{aligned}$$

Thus the lemma is proven. □

We can therefore conclude the following theorem.

THEOREM 4.6. *Let u_l be the lth eigenfunction, and let u_l^N be the solution of the Fourier–Galerkin approximation* (4.2)*; then*

$$||u_l - u_l^N|| \le CN^{-2.5}. \tag{4.14}$$

The numerical results presented in Tables 1 and 2 conform to the theory.

5. Fourier collocation method. Let $\mathcal{I}_{\mathcal{N}}$ be the space of the trigonometric polynomial of degree $N/2$, defined as

$$\mathcal{I}_{\mathcal{N}} = \text{span}\{(\cos(kx)|0 \le k \le N/2) \cup (\sin(kx)|1 \le k \le N/2 - 1)\}. \tag{5.1}$$

For an even integer $N > 0$, we consider the set of points

$$x_j = -\pi + \frac{2\pi j}{N}, \quad j = 0, \dots, N. \tag{5.2}$$

The discrete approximations of the inner products (3.1) and (3.2) are defined by

$$a(u,v)_h = \frac{2\pi}{N} \sum_{j=0}^{N-1} u'(x_j)\overline{v'(x_j)}, \tag{5.3}$$

$$(u,v)_h = \frac{2\pi}{N} \sum_{j=0}^{N-1} u(x_j)\overline{v(x_j)}\epsilon(x_j). \tag{5.4}$$

Alternatively, defining $c_j = 1$, $0 \neq j \neq N$, and $c_N = c_0 = 2$, we can redefine

$$a(u,v)_h = \frac{2\pi}{N}\sum_{j=0}^{N} u'(x_j)\overline{v'(x_j)}\frac{1}{c_j}, \tag{5.5}$$

$$(u,v)_h = \frac{2\pi}{N}\sum_{j=0}^{N} u(x_j)\overline{v(x_j)}\epsilon(x_j)\frac{1}{c_j}. \tag{5.6}$$

Remark 5.1. Note that the bilinear form $a(u,v)_h$ coincides with the inner product $a(u,v)$ for trigonometrical polynomials of the right order:

$$a(u,v)_h = a(u,v) \quad \text{for all} \quad u, v \in \mathcal{I}_{\mathcal{N}}. \tag{5.7}$$

This is a result of the exactness of the quadrature formula if $u'v'$ is up to degree $N-1$ [1]. One can observe that the highest degree N for $u'v'$ is obtained when choosing $\cos(\frac{N}{2}x)$ for both u and v. However, $\{\cos(\frac{N}{2}x)\}' = -\frac{N}{2}\sin(\frac{N}{2}x)$ and $\sin(\frac{N}{2}x)$ vanishes at the grid points x_j so that the quadrature formula still remains valid also for the case of highest degree N. Thus (5.7) is true for any u, v in $\mathcal{I}_{\mathcal{N}}$.

Remark 5.2. Equation (5.6) can be rewritten as

$$\begin{aligned}(v,v)_h = {} & \frac{\pi}{N}(|v(x_0)|^2\epsilon(x_0) + |v(x_{\frac{N}{2}})|^2\epsilon(x_{\frac{N}{2}})) + \frac{2\pi}{N}\sum_{j=1}^{\frac{N}{2}-1}|v(x_j)|^2\epsilon(x_j) \\ & + \frac{2\pi}{N}\sum_{\frac{N}{2}+1}^{N-1}|v(x_j)|^2\epsilon(x_j) + \frac{\pi}{N}(|v(x_N)|^2\epsilon(x_N) + |v(x_{\frac{N}{2}})|^2\epsilon(x_{\frac{N}{2}})).\end{aligned}$$

The first two terms can be identified as the trapezoidal rule [6] for $\int_{-\pi}^{0}|v(\xi)|^2\epsilon(\xi)d\xi$, whereas the other two terms are the same rule for $\int_0^{\pi}|v(\xi)|^2\epsilon(\xi)d\xi$. We can therefore state

$$|(v,v) - (v,v)_h| \leq CN^{-2}\max\left\{\max_{-\pi\leq x<0}(|v|^2\cdot\epsilon)'', \max_{0\leq x\leq\pi}(|v|^2\cdot\epsilon)''\right\}. \tag{5.8}$$

5.1. Numerical scheme and its results. The collocation methods can be defined as finding λ^c and $u^c \in \mathcal{I}_N$ such that

$$a(u^c, v^c)_h = \lambda^c(u^c, v^c)_h \quad \text{for all} \quad v^c \in \mathcal{I}_N. \tag{5.9}$$

There are several ways to realize the abstract definition of the collocation methods, and we will quote one of them: u^c can be presented using the Lagrange trigonometrical polynomials as interpolation polynomials [4] as follows:

$$u^c = \sum_{j=0}^{N-1} u^c(x_j)l_j(x), \tag{5.10}$$

where

$$l_j(x) = \frac{1}{N}\sin\left[N\frac{(x-x_j)}{2}\right]\cot\left[\frac{x-x_j}{2}\right]. \tag{5.11}$$

Taking $v^c(x) = l_n(x)$, we have

$$\begin{aligned} a(u,v)_h &= \frac{2\pi}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} u^c(x_j) l_j'(x_i) l_n'(x_i) \\ &= \frac{2\pi}{N}\sum_{j=0}^{N-1} u^c(x_j)\sum_{i=0}^{N-1} l_j'(x_i) l_n'(x_i) \\ &= \frac{2\pi}{N}\sum_{j=0}^{N-1} u^c(x_j) D_{nj}^2, \end{aligned}$$

where $D^2 = -D \cdot D$ and D is the first order differentiation matrix for even grid points [2], [4], [5]. Also,

$$\begin{aligned} (u,v)_h &= \frac{2\pi}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} u^c(x_j) l_j(x_i) l_n(x_i)\epsilon(x_i) \\ &= \frac{2\pi}{N}\sum_{j=0}^{N-1} u^c(x_j)\sum_{i=0}^{N-1} l_j(x_i) l_n(x_i)\epsilon(x_i) \\ &= \frac{2\pi}{N}\sum_{j=0}^{N-1} u^c(x_j) A_{jn}, \end{aligned}$$

where $A = \mathrm{diag}\{\epsilon(x_0), \ldots, \epsilon(x_{N-1})\}$. Then we solve the matrix equation

$$D^2 \mathbf{u}^c = \lambda^c A \mathbf{u}^c \tag{5.12}$$

to get the approximate eigenvalues λ^c and eigenfunctions $\mathbf{u}^c = [u^c(x_0), \ldots, u^c(x_{N-1})]^T$.

Remark 5.3. In order to make (5.12) compatible with definition (5.6), we should replace $\epsilon(x_0)$ by the average $\frac{\epsilon(x_0)+\epsilon(x_N)}{2}$.

The variational formulation with odd grids can be obtained in a similar way. In Tables 3 and 4, we present the relative error for $\lambda_l^c - \lambda_l$. It is clear that we see *second order accuracy* with even grids as well as with odd grids as N increases. The discrete L_2-error of $u_l - u_l^c$ converges with *second order accuracy* with even and odd grids as N increases.

5.2. Error estimates for eigenvalues and eigenfunctions. Here we provide error estimates for the approximate eigenvalues and eigenfunctions of the Fourier collocation method. We first consider the approximate eigenvalues. Let S_l be any l-dimensional subspace of $\mathcal{I}_\mathcal{N}$.

LEMMA 5.1. *Let λ_l^c be the approximation to λ_l obtained by the collocation procedure, and let S_l be any l-dimensional subspace of $\mathcal{I}_\mathcal{N}$. Then*

$$\lambda_l^c \le \max_{v \in S_l} \frac{a(v,v)_h}{(v,v)_h}. \tag{5.13}$$

Proof. The space S_l is spanned by the eigenfunctions $u_{k_1}^c, \ldots, u_{k_l}^c$ for $k_1 < \cdots < k_l$. It follows that $k_l \ge l$. Now

$$\frac{a(u_{k_l}^c, u_{k_l}^c)_h}{(u_{k_l}^c, u_{k_l}^c)_h} = \lambda_{k_l}^c \ge \lambda_l^c. \tag{5.14}$$

TABLE 3

Fourier collocation with even grids: The relative errors of eigenvalues for the case $\beta = 2$ and the discrete L_2-errors of $u_i - u_i^c$.

λ_i	N	λ_i^c	$\frac{(\lambda_i^c - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^c\|_{l_2}$	Order
	16	0.37224	6.3993(-3)		4.4960(-3)	
	32	0.37049	1.6710(-3)	1.9372	1.1755(-3)	1.9354
0.36987	64	0.37003	4.2818(-4)	1.9644	3.0154(-4)	1.9628
	128	0.36992	1.0846(-4)	1.9811	7.6428(-5)	1.9801
	256	0.36989	2.7298(-5)	1.9903	1.9243(-5)	1.9897
	16	0.53024	-1.1169(-2)		7.3926(-3)	
	32	0.53477	-2.7366(-3)	2.0290	1.7725(-3)	2.0603
0.53623	64	0.53587	-6.7697(-4)	2.0152	4.3472(-4)	2.0276
	128	0.53614	-1.6834(-4)	2.0077	1.0771(-4)	2.0130
	256	0.53621	-4.1974(-5)	2.0038	2.6811(-5)	2.0062
	16	1.63690	1.8531(-2)		1.1564(-2)	
	32	1.61442	4.5477(-3)	2.0267	3.0650(-3)	1.9157
1.60712	64	1.60895	1.1432(-3)	1.9920	7.9696(-4)	1.9433
	128	1.60758	2.8754(-4)	1.9913	2.0372(-4)	1.9679
	256	1.60723	7.2160(-5)	1.9945	5.1533(-5)	1.9830
	16	1.89597	-2.1273(-2)		1.7356(-2)	
	32	1.92826	-4.6037(-3)	2.2081	4.8491(-3)	1.8396
1.93718	64	1.93510	-1.0746(-3)	2.0991	1.2694(-3)	1.9336
	128	1.93668	-2.5988(-4)	2.0478	3.2443(-4)	1.9682
	256	1.93706	-6.3921(-5)	2.0235	8.1999(-5)	1.9842
	16	4.00439	1.0977(-3)		6.8315(-2)	
	32	4.00012	2.9270(-5)	5.2288	1.5700(-2)	2.1214
4.00000	64	4.00000	8.7981(-7)	5.0561	3.8391(-3)	2.0319
	128	4.00000	2.7229(-8)	5.0140	9.5432(-4)	2.0082
	256	4.00000	8.4864(-10)	5.0039	2.3823(-4)	2.0021

TABLE 4

Fourier collocation with odd grids: The relative errors of eigenvalues for the case $\beta = 2$ and the discrete L_2-errors of $u_i - u_i^c$.

λ_i	N	λ_i^c	$\frac{(\lambda_i^c - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^c\|_{l_2}$	Order
	17	0.37037	1.3299(-3)		1.0195(-2)	
	33	0.37002	3.8441(-4)	1.7906	2.6389(-3)	1.9499
0.36987	65	0.36991	1.0289(-4)	1.9016	6.7366(-4)	1.9698
	129	0.36988	2.6595(-5)	1.9519	1.7035(-4)	1.9835
	257	0.36988	6.7596(-6)	1.9761	4.2842(-5)	1.9914
	17	0.53491	-2.4693(-3)		8.0457(-3)	
	33	0.53589	-6.4469(-4)	1.9374	2.0181(-3)	1.9952
0.53623	65	0.53615	-1.6437(-4)	1.9717	5.0695(-4)	1.9931
	129	0.53621	-4.1480(-5)	1.9864	1.2715(-4)	1.9953
	257	0.53623	-1.0418(-5)	1.9933	3.1846(-5)	1.9973
	17	1.61319	3.7829(-3)		1.6275(-2)	
	33	1.60879	1.0434(-3)	1.8581	4.5235(-3)	1.8471
1.60712	65	1.60756	2.7480(-4)	1.9249	1.2086(-3)	1.9041
	129	1.60723	7.0552(-5)	1.9616	3.1368(-4)	1.9460
	257	1.60714	1.7876(-5)	1.9806	7.9996(-5)	1.9713
	17	1.92848	-4.4933(-3)		3.7036(-2)	
	33	1.93509	-1.0779(-3)	2.0595	9.7683(-3)	1.9228
1.93718	65	1.93667	-2.6181(-4)	2.0416	2.5446(-3)	1.9407
	129	1.93706	-6.4278(-5)	2.0261	6.5220(-4)	1.9641
	257	1.93715	-1.5904(-5)	2.0149	1.6529(-4)	1.9803

This concludes the proof. ∎

We are now ready to estimate λ_l^c from above.

LEMMA 5.2.

$$\lambda_l^c \le \lambda_l(1 + CN^{-2}), \tag{5.15}$$

where C is independent of N (but may depend linearly on l).

Proof. Let $J_N(= J)$ be the orthogonal projection (in the usual L^2 sense) of H^2 to $\mathcal{I}_{\mathcal{N}}$. Let $S_l = JE_{1,l} = \text{span}\{Ju_1, \dots, Ju_l\}$ in Lemma 5.1 to get

$$\lambda_l^c \le \max_{v \in JE_{1,l}} \frac{a(v,v)_h}{(v,v)_h} = \max_{v \in E_{1,l}} \frac{a(Jv,Jv)_h}{(Jv,Jv)_h}.$$

Then

$$\lambda_l^c \le \max_{v \in E_{1,l}} \frac{a(v,v)}{(v,v)} \cdot \frac{a(Jv,Jv)_h}{a(v,v)} \cdot \frac{(v,v)}{(Jv,Jv)} \cdot \frac{(Jv,Jv)}{(Jv,Jv)_h}. \tag{5.16}$$

From Lemma 3.2, we have

$$\max_{v \in E_{1,l}} \frac{a(v,v)}{(v,v)} = \lambda_l.$$

Also from the exactness of the trapezoidal rule and the fact that J is an orthogonal projection, it is true that

$$a(Jv,Jv)_h = a(Jv,Jv) \le a(v,v).$$

Due to Lemma 4.3 (with the same proof for J replacing P), we have

$$\max_{v \in E_{1,l}} \frac{(v,v)}{(Jv,Jv)} \le 1 + ClN^{-3}.$$

Also, Remark 5.2 gives

$$\frac{(Jv,Jv)}{(Jv,Jv)_h} \le 1 + CN^{-2},$$

and so the lemma is proven. ∎

We will now try to get a lower bound for λ_l^c. Define $E_{1,l}^c = \text{span}\{u_1^c, \dots, u_l^c\}$. From the minmax theorem, we have

$$\lambda_l \le \max_{v \in E_{1,l}^c} \frac{a(v,v)}{(v,v)}.$$

It is also clear that

$$\lambda_l^c = \max_{v \in E_{1,l}^c} \frac{a(v,v)_h}{(v,v)_h}.$$

We can now state the following lemma.

LEMMA 5.3.

$$\lambda_l \le \lambda_l^c(1 + CN^{-2}), \tag{5.17}$$

where C is independent of N (but may depend linearly on l).

Proof. We start from

$$\begin{aligned}\lambda_l &\le \max_{v\in E^c_{1,l}} \frac{a(v,v)}{(v,v)} \\ &= \max_{v\in E^c_{1,l}} \frac{a(v,v)_h}{(v,v)_h}\cdot\frac{a(v,v)}{a(v,v)_h}\cdot\frac{(v,v)_h}{(v,v)}.\end{aligned}$$

Since $v \in \mathcal{I}_{\mathcal{N}}$, we have $a(v,v) = a(v,v)_h$. Also, because of the trapezoidal rule estimate

$$|(v,v)_h - (v,v)| \le CN^{-2},$$

and therefore the lemma is proven. ∎

We can now conclude the following theorem.

THEOREM 5.4. *Let λ_l^c be the approximation to λ_l obtained by the collocation procedure. Then*

$$|\lambda_l^c - \lambda_l| \le C\lambda_l N^{-2}, \tag{5.18}$$

where C is independent of N.

We now turn to the eigenfunctions. The set $u_1^c, u_2^c, \dots, u_N^c$ forms an orthogonal basis for $\mathcal{I}_{\mathcal{N}}$. Then we can express the orthogonal projection Ju_l of u_l into the subspace $\mathcal{I}_{\mathcal{N}}$ as the following:

$$Ju_l = \sum_{j=1}^{N} (Ju_l, u_j^c)_h u_j^c. \tag{5.19}$$

By subtracting the following variational formulations,

$$\begin{aligned}\lambda_l(u_l, u_j^c) &= a(u_l, u_j^c),\\ \lambda_j^c(Ju_l, u_j^c)_h &= a(Ju_l, u_j^c)_h = a(Ju_l, u_j^c),\end{aligned}$$

we have

$$\begin{aligned}(\lambda_j^c - \lambda_l)(Ju_l, u_j^c)_h &= a(Ju_l, u_j^c) - a(u_l, u_j^c) - \lambda_l[(Ju_l, u_j^c)_h - (u_l, u_j^c)]\\ &= -a(u_l - Ju_l, u_j^c) + \lambda_l[(u_l, u_j^c) - (Ju_l, u_j^c)_h].\end{aligned}$$

Since $a(u_l - Ju_l, u_j^c) = 0$, we have

$$\begin{aligned}|(Ju_l, u_j^c)_h| &\le \frac{\lambda_l}{|\lambda_j^c - \lambda_l|}\cdot|(u_l, u_j^c) - (Ju_l, u_j^c)_h|\\ &\le \frac{\lambda_l}{|\lambda_j^c - \lambda_l|}\cdot\{|(u_l, u_j^c) - (u_l, u_j^c)_h| + |(u_l, u_j^c)_h - (Ju_l, u_j^c)_h|\}.\end{aligned}$$

From the Schwarz inequality and $(u_j^c, u_j^c)_h = 1$, we have

$$\begin{aligned}|(u_l - Ju_l, u_j^c)_h| &\le \|u_l - Ju_l\|_h \cdot \|u_j^c\|_h\\ &\le CN^{-2},\end{aligned}$$

where $\|u\|_h = \sqrt{(u,u)_h}$. Using the trapezoidal rule as in (5.8), we have

$$|(u_l, u_j^c) - (u_l, u_j^c)_h| \le CN^{-2}.$$

Then, following [7],

$$\|Ju_l - \beta u_l^c\|_h = \sqrt{\sum_{\substack{j=1 \\ j\neq l}}^{N} |(Ju_l, u_j^c)_h|^2} \le \rho CN^{-2},$$

where $\beta = (Ju_l, u_l^c)_h$ and ρ is a separation constant for the eigenvalues as in [7, pp. 234–235]. From (5.8) and $\|u_l\| = \|u_l^c\|_h = 1$,

$$\|u_l\|_h \le \|u_l\| + CN^{-2} = \|u_l^c\|_h + CN^{-2}.$$

Then, following [7], we have

$$\|u_l - u_l^c\|_h \le \|u_l - Ju_l\|_h + \|Ju_l - \beta u_l^c\|_h + \|\beta u_l^c - u_l^c\|_h \le CN^{-2}.$$

We can therefore conclude the following theorem.

THEOREM 5.5. *Let u_l be the lth eigenfunction, and let u_l^c be the solution of the Fourier collocation approximation* (5.9)*; then*

$$||u_l - u_l^c||_h \le CN^{-2}.$$

Due to the equivalence of the norms in finite space, the discrete L_2-error of $u_l - u_l^c$ also converges with $O(N^{-2})$.

6. Accuracy enhancement for the collocation method. A simple trick can be used in order to enhance the accuracy of the Fourier collocation method. We expand the discontinuous coefficient function $\epsilon(x)$ in the finite Fourier series represented by

$$\epsilon^N(x) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} (\hat{\epsilon}^N)_k e^{ikx},$$

where the Fourier coefficients are defined as

$$(\hat{\epsilon}^N)_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \epsilon(x) e^{-ikx} dx.$$

Now, instead of (5.4), defining

$$(u,v)_h = \frac{2\pi}{N} \sum_{j=0}^{N-1} u(x_j)\overline{v(x_j)}\epsilon^N(x_j) \tag{6.1}$$

in the variational formulation (5.9), we have the scheme as follows:

$$D^2\mathbf{u}^c = \lambda^c A\mathbf{u}^c, \tag{6.2}$$

where $A = \mathrm{diag}\{\epsilon^N(x_0), \ldots, \epsilon^N(x_{N-1})\}$ and D^2 is the same as defined in (5.12).

The numerical results are presented in the Tables 5 and 6 for even and odd grids, respectively. The accuracy is now the same accuracy as for the Galerkin method! An analysis for this will appear in a future paper.

TABLE 5

The accuracy enhancement for collocation with even grids: The relative errors of eigenvalues for the case $\beta = 2$ and the discrete L_2-errors of $u_i - u_i^c$ using $\epsilon^N(x)$.

λ_i	N	λ_i^c	$\frac{(\lambda_i^c - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^c\|_{l_2}$	Order
	16	0.36995	2.0453(-4)		5.1956(-4)	
	32	0.36988	2.1664(-5)	3.2389	1.0071(-4)	2.3671
0.36987	64	0.36988	2.4904(-6)	3.1208	1.9120(-5)	2.3970
	128	0.36988	2.9842(-7)	3.0609	3.5138(-6)	2.4440
	256	0.36987	3.6546(-8)	3.0296	6.3363(-7)	2.4713
	16	0.53624	7.7676(-6)		9.7460(-4)	
	32	0.53624	5.7076(-6)	0.4445	1.4544(-4)	2.7444
0.53623	64	0.53623	1.0179(-6)	2.4873	2.2753(-5)	2.6764
	128	0.53623	1.4643(-7)	2.7973	3.7110(-6)	2.6161
	256	0.53623	1.9435(-8)	2.9135	6.2557(-7)	2.5686
	16	1.60674	-2.3345(-4)		4.9371(-3)	
	32	1.60714	1.5379(-5)	3.9241	4.7815(-4)	3.3681
1.60712	64	1.60712	3.2157(-6)	2.2577	6.7468(-5)	2.8250
	128	1.60712	4.5610(-7)	2.8177	1.0873(-5)	2.6334
	256	1.60712	5.9832(-8)	2.9304	1.8450(-6)	2.5591
	16	1.94119	2.0698(-3)		6.8043(-3)	
	32	1.93744	1.3393(-4)	3.9500	6.3418(-4)	3.4235
1.93718	64	1.93721	1.3685(-5)	3.2908	1.0359(-4)	2.6140
	128	1.93718	1.5887(-6)	3.1067	1.8498(-5)	2.4855
	256	1.93718	1.9292(-7)	3.0418	3.3183(-6)	2.4789
	16	3.92531	-1.8673(-2)		3.9832(-2)	
	32	3.99878	-3.0425(-4)	5.9395	1.8962(-3)	4.3927
4.00000	64	3.99995	-1.3559(-5)	4.4880	2.0989(-4)	3.1754
	128	4.00000	-7.3809(-7)	4.1993	2.5516(-5)	3.0402
	256	4.00000	-4.3296(-8)	4.0915	3.1659(-6)	3.0107

TABLE 6

The accuracy enhancement for collocation with odd grids: The relative errors of eigenvalues for the case $\beta = 2$ and the discrete L_2-errors of $u_i - u_i^c$ using $\epsilon^N(x)$.

λ_i	N	λ_i^c	$\frac{(\lambda_i^c - \lambda_i)}{\lambda_i}$	Order	$\|u_i - u_i^c\|_{l_2}$	Order
	17	0.36994	1.7576(-4)		5.8880(-4)	
	33	0.36988	1.9991(-5)	3.1362	1.0629(-4)	2.4698
0.36987	65	0.36988	2.3892(-6)	3.0647	1.9258(-5)	2.4644
	129	0.36988	2.9220(-7)	3.0315	3.4570(-6)	2.4778
	257	0.36987	3.6169(-8)	3.0141	6.1624(-7)	2.4880
	17	0.53626	4.4494(-5)		8.1441(-4)	
	33	0.53624	8.1111(-6)	2.4556	1.2636(-4)	2.6882
0.53623	65	0.53623	1.1708(-6)	2.7924	2.0471(-5)	2.6259
	129	0.53623	1.5604(-7)	2.9075	3.4310(-6)	2.5768
	257	0.53623	2.0042(-8)	2.9608	5.8871(-7)	2.5430
	17	1.60723	7.1745(-5)		3.5514(-3)	
	33	1.60716	2.4805(-5)	1.5323	3.9880(-4)	3.1547
1.60712	65	1.60712	3.6264(-6)	2.7740	6.0783(-5)	2.7139
	129	1.60712	4.7742(-7)	2.9252	1.0143(-5)	2.5832
	257	1.60712	6.1050(-8)	2.9672	1.7484(-6)	2.5364
	17	1.93983	1.3694(-3)		5.1304(-3)	
	33	1.93740	1.1435(-4)	3.5820	6.1267(-4)	3.0659
1.93718	65	1.93721	1.2826(-5)	3.1563	1.0292(-4)	2.5736
	129	1.93718	1.5433(-6)	3.0550	1.8186(-5)	2.5006
	257	1.93718	1.9029(-7)	3.0197	3.2305(-6)	2.4930

Acknowledgments. The authors would like to thank Bertil Gustafsson, Seymour Parter, and Wai-Sun Don for several useful suggestions and discussions regarding this paper.

REFERENCES

[1] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer Ser. Comput. Phys., Springer-Verlag, New York, 1988.

[2] D. Funaro, *Polynomial Approximation of Differential Equations,* Springer-Verlag, New York, 1991.

[3] G. H. Golub and C. Van Loan, *Matrix Computations,* The Johns Hopkins University Press, Baltimore, MD, 1996.

[4] D. Gottlieb, M. Y. Hussaini, and S. A. Orszag, *Theory and Applications of Spectral Methods*, in Spectral Methods for Partial Differential Equations, R. Voigt, D. Gottlieb, and M.Y. Hussaini, eds., SIAM, Philadelphia, 1984, pp. 1–54.

[5] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Application*, CMBS-NSF Regional Conf. Ser. Appl. Math. 26, SIAM, Philadelphia, 1977.

[6] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods,* John Wiley & Sons, New York, 1966.

[7] G. Strang and G. Fix, *An Analysis of The Finite Element Method,* Prentice–Hall, Englewood Cliffs, NJ, 1973.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2270–2291

SYMMETRIC ERROR ESTIMATES FOR MOVING MESH MIXED METHODS FOR ADVECTION-DIFFUSION EQUATIONS*

YINGJIE LIU†, RANDOLPH E. BANK‡, TODD F. DUPONT§, SONIA GARCIA¶, AND RAFAEL F. SANTOS‖

Abstract. A mixed method allowing a general class of mesh movements is proposed for an advection-diffusion equation in either conservative or nonconservative form. Several symmetric error estimates are derived for the method under certain conditions. In one space dimension, optimal order L^2 convergence and superconvergence are proved as corollaries of the symmetric estimates.

Key words. mixed methods, parabolic equations, finite elements, moving mesh

AMS subject classifications. 65M60, 65M12

PII. S003614290038073X

1. Introduction.

Moving mesh finite element methods have been widely studied; in [10, 9] methods based on Galerkin formulations were given. In [5, 2] error analysis was provided for related classes of moving mesh finite element methods which allow piecewise time continuous mesh movements. In this work, we examine moving mesh methods for mixed methods that incorporate some of the ideas in [4], where a procedure for including characteristics within finite element methods for advection-diffusion equations was proposed.

A symmetric error estimate is, to within a constant, a best approximation result. That is, if the error *can be* made small in the given norm, then it *is* small in that norm. Somewhat more precisely, there is a norm $\|\!| \cdot |\!\|$ and a constant C such that

$$\|\!|\text{error}|\!\| \le C \|\!|\text{best approximation error}|\!\|.$$

Dupont [5], Bank and Santos [2], Dupont and Liu [6], and sections 5 and 7 of this work establish bounds of this type. In [6] and this paper, the constant C does not increase as the advective term increases in size, provided that the mesh movement approximates the advective term sufficiently well. These results thus make it clear that the mesh movement is actually modeling the advection. Also, the norms in section 5 involve the convective derivative instead of the partial with respect to time, and as

*Received by the editors November 9, 2000; accepted for publication (in revised form) July 19, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/38073.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (yingjie@math.gatech.edu).
‡Department of Mathematics, University of California at San Diego, La Jolla, CA 92093 (rbank@ucsd.edu). The work of this author was supported by the National Science Foundation under contract DMS-9706090.
§Department of Computer Science, University of Chicago, Chicago, IL 60637 (dupont@cs.uchicago.edu). The work of this author was supported by the ASCI Flash Center at the University of Chicago under DOE contract B341495, and by the MRSEC Program of the National Science Foundation under award DMR-9808595.
¶Department of Mathematics, United States Naval Academy, Annapolis, MD 21402 (smg@nadn.navy.mil). The work of this author was supported by the Office of Research of the U.S. Naval Academy, NARC Program.
‖Department of Mathematics, University of Algarve, 8000 Faro, Portugal and Centro de Matemática e Aplicações Fundamentais, Av. Prof. Gama Pinto, 2, 1699 Lisboa Codex, Portugal (rsantos@ualg.pt). The work of this author was supported by Fundação para a Ciência e Tecnologia and PRAXIS XXI grant PRAXIS/2/2.1/MAT/ 125/94.

Douglas and Russell pointed out in [4], for advection dominated problems the convective derivative will typically be much smoother, and therefore easier to approximate well. While symmetric error estimates for parabolic equations have a certain attractiveness in the simplicity of the statement that they make, it is sometimes hard to see the precise meaning of the result because the norms involved are made up of several parts. We exploit the idea of [6] to weaken some of these parts to "concentrate" the norm on certain terms.

Although the motivation for this research was an improved understanding of moving mesh methods, it is worth remarking that the symmetric error estimates provided here are valid even if the mesh does not move. While such estimates for parabolic equations have a thirty year history in the context of Galerkin methods, these are the first symmetric error estimates for mixed methods for parabolic problems even in the fixed mesh case.

Characteristics-type mixed methods have been studied in several papers (see, e.g., Yang [12] and Arbogast and Wheeler [1]), but the analytical understanding of mixed methods in combination with moving meshes is far from complete. Unlike Galerkin methods using conforming finite element spaces, moving mesh methods using mixed formulations and discontinuous approximation spaces can develop singularities in the time derivative at the edges between elements. Therefore it is critical to use directional time derivatives along the mesh movement direction throughout the analysis. The mesh movement that is considered here is more general than just a fixed mesh or a mesh that follows characteristics, for several reasons; two of the most significant are the following. First, the best mesh may not be fixed or follow the characteristics. Diffusion spreads things out and a mesh can follow such patterns; in fact, in [8] it is shown that in some situations mesh movement alone can model diffusion. Thus when diffusion and advection are both present, one may want to use a mesh that reflects the action of the two together. Second, the choice of mesh moving strategy will usually involve in a strong way considerations of the complexity of the program used to implement the mesh movement. One technique that we have used is to guess an analytic form for the mesh transformation based on a coarse grid calculation. Since the estimates here say that if you *can* approximate the solution, you *will*, this very simple-to-code approach is seen as a legitimate way to proceed. This technique is illustrated in an example in section 6.

In this paper, we first introduce our method and prove our symmetric error estimates. Next, an optimal order L^2 error estimate and a superconvergence result are proved for one space dimension as a corollary of the symmetric error estimate. The error bound gives considerable insight into the effectiveness of a given mesh movement. Aligning the mesh movement with the characteristics is not necessary as long as the difference between the advection velocity and the velocity of mesh movement remains bounded. The fact that the constants in the error bounds don't depend directly on the advection coefficient reflects the fact that mesh movement does indeed model advection. Furthermore, the analysis also shows that if the mesh is moved in such a way that it has a finer mesh where the solution has hard-to-approximate regions, then the bound on the error is decreased. These two observations give insight into what are good choices of mesh movement.

The remainder of this paper is organized as follows. In section 2, we discuss the advection-diffusion equation in conservative form, introduce several notations, and formulate the mixed method for general mesh movements. In section 3, we introduce a pseudoinverse operator "A" of "div," and in section 4 we develop the basic

properties of the directional derivative "D/Dt"; these concepts are used in section 5 to get symmetric error estimates. Optimal order error bounds are proved in section 6, and an example is presented that illustrates some of the issues associated with these techniques. In section 7, we consider a mixed method for an advection-diffusion equation in nonconservative form, allowing general mesh movements. Symmetric error analysis and one-dimensional applications are derived in a manner that parallels the earlier analysis.

2. Model problem and mixed method. Consider the following advection-diffusion model problem on $Q = \Omega \times (0,T)$:

$$\begin{cases} \partial_t u - \nabla \cdot (a\nabla u + bu) = f & \text{on } Q, \\ u = 0 & \text{on } \partial\Omega \times (0,T), \\ u = u_0 & \text{for } t = 0, \end{cases} \tag{2.1}$$

where $a(x)$, $b(x)$, and $f(x,t)$ are smooth and bounded and $a_1 \geq a(x) \geq a_0 > 0$ for some constants a_0, a_1. Here Ω is a bounded domain in R^n. For simplicity, we assume that Ω is a fixed polyhedron.

We use $\|\cdot\|_s$ to denote the $H^s(\Omega)$ norm. When $s = 0$, we usually use $\|\cdot\|$. If we use domains other than Ω, we will use $\|\cdot\|_{H^s(\Omega_i)}$ or $\|\cdot\|_{L^2(\Omega_i)}$. The norm for the dual space of $H_0^1(\Omega)$ is denoted $\|\cdot\|_{-1}$, and $\|\xi\|_{L^p(0,T;X)}$ denotes the $L^p(0,T)$ norm of $\|\xi(\cdot,t)\|_X$. We will use $(\cdot,\cdot)$ as the inner product on $L^2(\Omega)$ and on $(L^2(\Omega))^n$, and will rely on context to indicate which is intended.

We will study methods that approximate the solution u of (2.1) on a moving mesh, which is given as a time-dependent image of a fixed reference mesh. Suppose that $\bar{D} = \cup D_i$ is a fixed polyhedron, where D_i's are closed sets with nonvoid disjoint interiors. We need few assumptions on the D_i's for much of the argument, but to keep the discussion simple, we suppose that each D_i is a simplex and that they form a tessellation of $\bar{D}$. Further, we suppose that there is a continuous mapping $\mathcal{G}$ from $\bar{D} \times [0,T]$ onto $\bar{\Omega}$ such that

1. for each t, $\mathcal{G}(\cdot,t)$ is a one-to-one piecewise linear mapping (with respect to $\{D_j\}$) of $\bar{D}$ onto $\bar{\Omega}$;
2. $\mathcal{G}$ is continuously differentiable on each $D_i \times [0,T]$; and
3. $\partial\Omega = \mathcal{G}(\partial D, t)$.

Let $\Omega_i(t) = \mathcal{G}(D_i,t)$, $h_i(t)$ be the diameter of $\Omega_i(t)$, and $h(t) = \max_i\{h_i(t)\}$. Then $\Omega_i(t)$ is also a simplex and $\{\Omega_i(t)\}$ becomes the moving partition of Ω. It is sometimes convenient to think of this moving mesh as being generated by a mapping of Ω onto itself. Let $\mathcal{G}^{-1} = \mathcal{G}^{-1}(\cdot,t)$ denote the inverse of $\mathcal{G}$ as a map of D onto Ω; thus this function can be viewed as being defined on $\bar{Q}$. The partial derivative with respect to t of $\mathcal{G}$ is denoted $\mathcal{G}_t$. The finite element mesh is advected with a flow that is given by

$$\dot{x}(t) = \mathcal{G}_t(\mathcal{G}^{-1}(x,t),t).$$

Given the assumptions on $\mathcal{G}$, the function $\dot{x}$ is a continuous piecewise linear function over the partition $\{\Omega_i\}$ of Ω. Let $\tilde{V}_h$ be a finite-dimensional subspace of $L^2(D)$. Then the corresponding finite element space on Ω is defined by

$$V_h(t) = \{\phi(x,t) : \phi(\mathcal{G}(\cdot,t),t) \in \tilde{V}_h\}.$$

We will take $H_h(t)$ to be a finite-dimensional subspace of $H(div,\Omega)$ so that $div\, H_h = V_h$ for any t. In particular, we will take V_h to be the space of discontinuous polynomials

of total degree at most m, and H_h to be the Raviart–Thomas flux space. Let P_h denote the L^2 projection onto V_h. Let Π_h be the linear operator $H(div,\Omega) \to H_h$ satisfying $(div\,(W - \Pi_h W), r) = 0\ \forall r \in V_h$ and $div\,\Pi_h = P_h div$ as defined by Raviart and Thomas in [11].

Let $\underline{h}(x,t)$ denote the function that has the value $h_i(t)$ on each $\Omega_i(t)$. For a function φ such that its restriction to Ω_i is in $H^s(\Omega_i)$, let

$$\|\varphi\|^2_{\underline{H}^s} = \sum_i \|\varphi\|^2_{H^s(\Omega_i)}.$$

We denote a particular directional derivative DF/Dt as follows:

$$\frac{DF(x,t)}{Dt} = \frac{\partial F(x,t)}{\partial t} + \dot{x}\cdot\nabla_x F(x,t).$$

Note that if $F(\cdot,t) \in V_h(t)$ is differentiable on each Ω_i, then DF/Dt is also in V_h. Even though it might seem that both $\partial F/\partial t$ and $\nabla_x F$ are singular on the boundaries $\partial\Omega_i$, the directions involved in DF/Dt never cross the boundary of any Ω_i.

The first mixed method we consider uses a mesh movement induced flux across subdomain boundaries. Let $\sigma = -(a\nabla u + bu + \dot{x}u)$ and $\alpha = 1/a$, $\beta = b/a$. The exact solution u satisfies

$$\frac{Du}{Dt} + div\,\sigma + (\nabla\cdot\dot{x})u = f.$$

This leads to the following mixed formulation:

$$\text{(2.2)}\qquad \begin{cases} (\alpha\sigma + (\beta + \alpha\dot{x})u, \mathcal{X}) - (u, div\,\mathcal{X}) = 0 & \forall \mathcal{X} \in H(div,\Omega), \\ \left(\dfrac{Du}{Dt} + div\,\sigma + (\nabla\cdot\dot{x})u, r\right) = (f,r) & \forall r \in L^2(\Omega). \end{cases}$$

We define the mixed approximation to be functions $u_h : [0,T] \to V_h$ and $\sigma_h : [0,T] \to H_h$ such that $u_h(0) = P_h u(0)$ and

$$\text{(2.3)}\qquad \begin{cases} (\alpha\sigma_h + (\beta + \alpha\dot{x})u_h, \mathcal{X}) - (u_h, div\,\mathcal{X}) = 0 & \forall \mathcal{X} \in H_h, \\ \left(\dfrac{Du_h}{Dt} + div\,\sigma_h + (\nabla\cdot\dot{x})u_h, r\right) = (f,r) & \forall r \in V_h. \end{cases}$$

Note that this method is *locally conservative*, because the rate of change of the integral of u over each subdomain is given by the integral around the boundary of the normal component of σ, and the normal component of σ is continuous across subdomain boundaries. (If this is less than clear, please see the proof of Lemma 7.)

In proving the symmetric error estimates, we don't need specific approximation properties, but we will need such properties in order to obtain a priori error bounds based on the mesh size and the smoothness of the solution u. We summarize these additional conditions here.

CONDITION 1 (approximation). *There exists a constant C_1 such that for any $w \in H^{s_1}(\Omega)$, $s_1 \geq 0$, and any $t \in [0,T]$,*

$$\|w - P_h w\| \leq C_1 \|\underline{h}^{\min\{m+1,s_1\}} w\|_{\underline{H}^{s_1}},$$

and for any $W \in (H^{s_2}(\Omega))^n$, $s_2 \geq 1$, and any $t \in [0,T]$,

$$\|W - \Pi_h W\| \leq C_1 \|\underline{h}^{\min\{m_1+1,s_2\}} W\|_{\underline{H}^{s_2}},$$

where $m_1 = m+1$ in one dimension and $m_1 = m$ in higher space dimensions.

This condition holds for the Raviart–Thomas spaces, where C_1 depends on m and on a bound for $h_i/\tilde{h}_i$, where $\tilde{h}_i$ is the diameter of the largest ball in R^n contained in Ω_i.

CONDITION 2 (stability of Π_h). *There exists a constant C_2 such that for any $W \in (H^1(\Omega))^n$ and any $t \in [0,T]$*

$$\|\Pi_h W\| \le C_2 \|W\|_1.$$

If Condition 1 holds, then C_2 can be taken to be $1 + C_1 h$. But this condition is strictly weaker than Condition 1; it allows controlled degeneracy in the elements as the mesh size decreases.

CONDITION 3 (H^2 regularity). *The domain Ω is regular enough that there exists a C_3 such that, for any $\xi \in L^2(\Omega)$, the boundary value problem*

$$\begin{cases} \Delta g = \xi & \text{in } \Omega, \\ g = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.4}$$

has a unique solution and $\|g\|_2 \le C_3 \|\xi\|$.

3. A pseudoinverse of *div*. In this section we define and explore the properties of a smoothing mapping that appears naturally in the symmetric error estimates. Let $A : L^2(\Omega) \to H_h$ be the pseudoinverse of *div* in the sense that

$$\begin{gathered} \varphi - div\,(A\varphi) \perp V_h, \\ \|A\varphi\| \text{ is minimal.} \end{gathered}$$

Note that $A(\varphi) = A(P_h\varphi)$; thus we can factor A as $A_{V_h} P_h$, where A_{V_h} is A restricted to V_h. Note that this factorization gives that A^* maps H_h into V_h. Let $H_h = \mathcal{O} \oplus \mathcal{O}^\perp$, where $\mathcal{O} = \{\mathcal{X} \in H_h : div\,\mathcal{X} = 0\}$ and $\mathcal{O}^\perp$ is its orthogonal complement with respect to the $(L^2(\Omega))^n$ inner product. Then *div* is a one-to-one mapping from $\mathcal{O}^\perp$ onto V_h, and A_{V_h} is its inverse. In the case of one dimension with $m = 0$, the operator A can be explicitly described: $A\varphi$ is the piecewise linear interpolant of a constant plus the integral of φ. The following result shows that in more general situations A behaves as a smoothing operator.

THEOREM 4. *If Conditions 1 and 3 hold, then there is a $C = C(C_1, C_3)$ such that for any $\xi \in L^2(\Omega)$*

$$\begin{aligned} \|A\xi\| &\le C\{h\|\xi\| + \|\xi\|_{-1}\}, \\ \|A\xi\| &\le C\{h\|P_h\xi\| + \|P_h\xi\|_{-1}\}. \end{aligned}$$

Proof. Let g be the solution of (2.4) and set $W = \nabla g$. Take $\rho \in H_h$ and $\nu \in V_h$ to be the mixed method approximation of W and g:

$$\begin{cases} (\rho, \mathcal{X}) + (\nu, div\,\mathcal{X}) = 0 & \forall \mathcal{X} \in H_h, \\ (div\,\rho, r) = (\xi, r) & \forall r \in V_h. \end{cases} \tag{3.1}$$

We want to show that $\rho = A\xi$. In fact, the second equation of (3.1) implies $div\,\rho = P_h\xi$, and the first one implies $(\rho, \mathcal{X}) = 0\ \forall \mathcal{X} \in \mathcal{O}$, which in turn implies that $\|\rho\|$ is minimal among all elements in H_h whose divergence is $P_h\xi$.

Next we need an approximation result for mixed methods (see, e.g., [7]) to see that

$$\begin{aligned} \|A\xi\|^2 &= (\rho,\rho) \\ &= (\rho,\rho - W) + (\rho, W) \\ &\le \|\rho\|\{Ch\|g\|_2 + \|W\|\} \\ &\le \|A\xi\|\{Ch\|\xi\| + \|W\|\}. \end{aligned} \tag{3.2}$$

It follows from (2.4) that

$$\|W\| = \|\nabla g\| \le C\|\xi\|_{-1}.$$

From this and (3.2) the first result of this theorem follows. The second follows since $A\xi = AP_h\xi$. ∎

Note that even if Ω fails to have the assumed H^2 regularity, the result may still be proved in some cases. Suppose that Ω can be expanded to $\tilde{\Omega}$, which has H^2 regularity, and the function spaces can be extended to $\tilde{\Omega}$ with the approximation properties still holding. Then extending ξ to be zero on $\tilde{\Omega} - \Omega$ and a slight modification of the above proof gives the conclusions of the theorem. For example, if Ω were an L-shaped region in two space dimensions, H^2 regularity would fail, but the extension to a square might be possible.

On H_h, the operator $A\ div$ does not increase the L^2 norm. Suppose that $\rho \in H_h$ and let $\psi = A\ div\,\rho$. Then $div\,\rho - div\,\psi \perp V_h$. Hence $\psi = \rho + z$, where $z \in \mathcal{O}$. Because $\|\psi\|$ is taken to be minimal and $z \equiv 0$ is possible, we see that

$$\|A\ div\,\rho\| = \|\psi\| \le \|\rho\|. \tag{3.3}$$

In one dimension the choice of discontinuous piecewise polynomial spaces allows a more local version of Theorem 4. In fact, let $\Omega = (x_0, x_N)$ and $\Omega_i = (x_{i-1}, x_i)$; then $A\xi = \int_{x_0}^{x} P_h\xi(s)ds + C$.

THEOREM 5. *If Condition* 2 *holds, then there is a* C *such that for any* $\xi \in L^2(\Omega)$

$$\|A\xi\| \le C\|\xi\|.$$

Proof. Take $\rho \in H_h$ and $\nu \in V_h$ to be defined by (3.1); thus we know that $A\xi = \rho$. From (3.1) with $\chi = A\xi$ and $r = \nu$ we see that

$$\|A\xi\|^2 = -(\xi,\nu) \le \|\xi\|\,\|\nu\|.$$

Let B be a cube that contains Ω, and take φ be the extension of ν to B by zero outside Ω. Take $g \in H_0^1(B)$ such that, on B, $\Delta g = \varphi$. Then, because the cube has H^2 regularity for the Laplacian, we see that ∇g is bounded in $(H^1(B))^2$ by $C\|\varphi\|_{L^2(B)} = C\|\nu\|$. Note that

$$\|\nu\|^2 = (\nu, div\nabla g) = (\nu, div\Pi_n\nabla g) = (-A\xi, \Pi_h\nabla g).$$

The operator Π_h is bounded as a map of H^1 into L^2 by Condition 2. Thus it follows that

$$\|\nu\|^2 \le \|A\xi\|\,\|\Pi_h\nabla g\| \le C\|A\xi\|\,\|g\|_2 \le C\|A\xi\|\,\|\nu\|.$$

The two displayed inequalities then give the desired result. ∎

4. Properties of D/Dt. From the definition of directional derivative we have the following basic relations, which we use later in energy-type arguments.

LEMMA 6.

$$\nabla_x \cdot \dot{x} = \frac{\partial |\det(\nabla_s \mathcal{G}(s,t))|/\partial t}{|\det(\nabla_s \mathcal{G})|}.$$

Proof. Take $D_0 \subset D$ to be an arbitrary small ball and let $\Omega_0(t) = \mathcal{G}(D_0, t)$. Then, with n as the outward normal to Ω_0,

$$\frac{\partial}{\partial t}\int_{\Omega_0(t)} dx = \int_{\partial\Omega_0(t)} \dot{x}\cdot n d\sigma = \int_{\Omega_0(t)} \nabla_x \cdot \dot{x} dx.$$

On the other hand,

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\Omega_0(t)} dx &= \frac{\partial}{\partial t}\int_{D_0} |\det(\nabla_s \mathcal{G}(s,t))| ds = \int_{D_0} \frac{\partial}{\partial t}|\det(\nabla_s \mathcal{G}(s,t))| ds \\ &= \int_{\Omega_0(t)} \frac{\partial |\det(\nabla_s \mathcal{G}(s,t))|/\partial t}{|\det(\nabla_s \mathcal{G})|} dx.\end{aligned}$$

The result follows from the arbitrary choice of D_0. ∎

We will say that a function ξ on Q is piecewise C^1 if, when it is pulled back by $\mathcal{G}$ to $D_i^o \times (0,T)$, it can be extended to be C^1 on $D_i \times [0,T]$. A function that is the limit in $H^1(D_i \times [0,T])$ of piecewise C^1 functions will be called piecewise smooth on Q. We will usually operate formally on piecewise smooth functions without going through the step of approximating them by smooth functions and taking limits, since this is routine.

LEMMA 7. *Suppose that ξ is piecewise smooth on Q; then, with $\mathcal{R} = \Omega$ or Ω_i,*

$$\frac{d}{dt}\int_{\mathcal{R}} \xi dx = \int_{\mathcal{R}} \frac{D\xi}{Dt} dx + \int_{\mathcal{R}} \xi(\nabla_x \cdot \dot{x}) dx.$$

Proof. It suffices to show the result for Ω_i. Note that

$$\begin{aligned}\frac{d}{dt}\int_{\Omega_i} \xi dx &= \frac{d}{dt}\int_{D_i} \xi |\det(\nabla \mathcal{G})| ds \\ &= \int_{D_i} \frac{\partial \xi}{\partial t}|\det(\nabla \mathcal{G})| ds + \int_{D_i} \xi \frac{\partial}{\partial t}|\det(\nabla \mathcal{G})| ds \\ &= \int_{\Omega_i} \frac{D\xi}{Dt} dx + \int_{\Omega_i} \xi \left(\frac{\partial |\det(\nabla_s \mathcal{G}(s,t))|/\partial t}{|\det(\nabla_s \mathcal{G})|}\right) dx.\end{aligned}$$

Using Lemma 6, the proof is complete. ∎

D/Dt also has the following properties for any piecewise smooth functions ξ, η:

$$\begin{aligned}\frac{D}{Dt}(\xi\eta) &= \eta\frac{D\xi}{Dt} + \xi\frac{D\eta}{Dt}, \\ \frac{D}{Dt}\nabla_x \xi &= \nabla_x \frac{D\xi}{Dt} - (\nabla_x \dot{x})^T \nabla_x \xi,\end{aligned}$$

where $\nabla_x \xi$ is a column vector and $\nabla_x \dot{x}$ is the Jacobian of $\dot{x}$ with respect to x.

It easily follows from this and Lemma 7 that

$$\left(\frac{D\xi}{Dt}, \xi\right) = \frac{1}{2}\frac{d}{dt}\|\xi\|^2 - \frac{1}{2}(\xi, \xi(\nabla_x \cdot \dot{x})). \tag{4.1}$$

We denote the pseudoderivative of ξ by

$$D_t\xi = \frac{D\xi}{Dt} + (\nabla \cdot \dot{x})\xi,$$

and now show that D_t commutes with P_h.

LEMMA 8. *For function ξ that is piecewise smooth on Q, $P_h D_t \xi = D_t P_h \xi$.*

Proof. Let $\psi = P_h\xi$; then $(\xi - \psi, r) = 0$ for any $r \in V_h$. Given $t_0 \in [0,T]$, let $\phi(x)$ be any function in $V_h(t_0)$. Let $r(x,t) = \phi(\mathcal{G}(\mathcal{G}^{-1}(x,t),t_0))$. Then $r(x,t_0) = \phi(x)$, $r(\cdot,t) \in V_h(t)$, and $Dr/Dt = 0$ for any $t \in [0,T]$. Thus at t_0,

$$\begin{aligned} 0 &= \frac{d}{dt}(\xi - \psi, r) \\ &= \left(\frac{D}{Dt}(\xi - \psi), \phi\right) + \left(\xi - \psi, \frac{Dr}{Dt}\right) + (\xi - \psi, (\nabla_x \cdot \dot{x})\phi). \end{aligned}$$

That is,

$$0 = (D_t(\xi - \psi), \phi) = (P_h D_t \xi - D_t P_h \xi, \phi).$$

The proof is completed by observing $D_t P_h \xi \in V_h$. ∎

5. Symmetric error estimates. In this section, we prove four symmetric error estimates.

Let F_h be a linear operator $V_h(t) \to H_h(t)$ such that for any $v_h \in V_h(t)$

$$(\alpha F_h(v_h) + (\beta + \alpha\dot{x})v_h, \mathcal{X}) - (v_h, div\, \mathcal{X}) = 0 \quad \forall \mathcal{X} \in H_h.$$

Thus F_h is the flux operator associated with the space V_h. Using F_h and the norms $|||(\cdot,\cdot)|||$ and $|||(\cdot,\cdot)|||_*$ defined by

$$|||(\eta,\psi)|||^2 = \|\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \left\|A\frac{D\eta}{Dt}\right\|^2_{L^2(0,T:L^2(\Omega))} + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))},$$

$$|||(\eta,\psi)|||^2_* = \|P_h\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \left\|A\frac{D\eta}{Dt}\right\|^2_{L^2(0,T:L^2(\Omega))} + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))},$$

we have the following pair of symmetric error estimates.

THEOREM 9. *Suppose that Condition 2 holds and there exist constants c_1, c_2 such that for all $(x,t) \in Q$*

$$|\nabla_x \cdot \dot{x}| \le c_1 \quad \text{and} \quad |\beta + \alpha\dot{x}| \le c_2.$$

Then there exists a constant $C > 0$, depending only on C_2, c_1, c_2, T, the bounds of coefficient a, and Ω, such that for any piecewise smooth function v_h with $v_h(\cdot,t) \in V_h(t)$,

$$\begin{aligned} |||(u - u_h, \sigma - \sigma_h)||| &\le C|||(u - v_h, \sigma - F_h(v_h))|||, \\ |||(u - u_h, \sigma - \sigma_h)|||_* &\le C|||(u - v_h, \sigma - F_h(v_h))|||_*. \end{aligned}$$

Proof. Take v_h to be a piecewise C^1 function such that $v_h(\cdot,t) \in V_h(t)$. With $\mathcal{S}_h = F_h(v_h)$, adopt the notation

$$\begin{aligned} \nu &= u_h - v_h, \qquad \rho = \sigma_h - \mathcal{S}_h, \\ \eta &= u - v_h, \qquad \psi = \sigma - \mathcal{S}_h. \end{aligned}$$

Subtracting (2.2) from (2.3), we obtain the following orthogonalities:

$$(\alpha\rho + (\beta + \alpha\dot{x})\nu, \mathcal{X}) - (\nu, div\, \mathcal{X}) = 0 \qquad \forall \mathcal{X} \in H_h,$$

$$\left(\frac{D\nu}{Dt} + div\, \rho + (\nabla \cdot \dot{x})\nu, r\right) = \left(\frac{D\eta}{Dt} + div\, \psi + (\nabla \cdot \dot{x})\eta, r\right) \qquad \forall r \in V_h. \tag{5.1}$$

With $\mathcal{X} = \rho$ and $r = \nu$, these and (4.1) give

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nu\|^2 + (\alpha\rho + (\beta + \alpha\dot{x})\nu, \rho) \\
&\qquad = \left(\frac{D\eta}{Dt} + div\, \psi, \nu\right) + ((\nabla \cdot \dot{x})\eta, \nu) - \frac{1}{2}\int_\Omega \nu^2(\nabla \cdot \dot{x})dx \\
&\qquad = \left(div\, A\left(\frac{D\eta}{Dt} + div\, \psi\right), \nu\right) + ((\nabla \cdot \dot{x})\eta, \nu) - \frac{1}{2}\int_\Omega \nu^2(\nabla \cdot \dot{x})dx \qquad (5.2) \\
&\qquad = \left(\alpha\rho + (\beta + \alpha\dot{x})\nu, A\left(\frac{D\eta}{Dt} + div\, \psi\right)\right) + ((\nabla \cdot \dot{x})\eta, \nu) \\
&\qquad\quad - \frac{1}{2}\int_\Omega \nu^2(\nabla \cdot \dot{x})dx.
\end{aligned}$$

Therefore

$$\frac{d}{dt}\|\nu\|^2 + \alpha_1\|\rho\|^2 \le C\left\{\|\nu\|^2 + \left\|A\left(\frac{D\eta}{Dt} + div\, \psi\right)\right\|^2 + \|\eta\|^2\right\}, \tag{5.3}$$

where $\alpha_1 = 1/a_1$. It follows from Gronwall's inequality that

$$\|\nu\|^2_{L^\infty(0,T:L^2(\Omega))} + \|\rho\|^2_{L^2(0,T:L^2(\Omega))} \le C\{\|\nu(0)\|^2 + |||(\eta, \psi)|||^2\}.$$

The choice of $u_h(0) = P_h u(0)$ shows $\|\nu(0)\| \le \|\eta(0)\|$, and so the $\|\nu(0)\|$-term is bounded by $|||(\eta, \psi)|||$. Combining these results with (3.3), we see that

$$\|\nu\|^2_{L^\infty(0,T:L^2(\Omega))} + \|A\, div\, \rho\|^2_{L^2(0,T:L^2(\Omega))} \le C|||(\eta, \psi)|||^2.$$

Note that $\nu = P_h\nu$ and $((\nabla \cdot \dot{x})\eta, \nu) = ((\nabla \cdot \dot{x})P_h\eta, \nu)$, since $\nabla \cdot \dot{x}$ is constant on each Ω_i and V_h has no continuity between subdomains. Therefore we can replace $\|\nu\|$ by $\|P_h\nu\|$, $\|\eta\|$ by $\|P_h\eta\|$ in (5.3) to obtain

$$\|P_h\nu\|^2_{L^\infty(0,T:L^2(\Omega))} + \|A\, div\, \rho\|^2_{L^2(0,T:L^2(\Omega))} \le C|||(\eta, \psi)|||^2_*.$$

It remains to estimate $\|A(\frac{D\nu}{Dt})\|^2$. Using (5.1) and Theorem 5,

$$\begin{aligned}
\left(A\frac{D\nu}{Dt}, A\frac{D\nu}{Dt}\right) &= \left(\frac{D\nu}{Dt}, A^*A\frac{D\nu}{Dt}\right) \\
&= -\left(div\, \rho + (\nabla \cdot \dot{x})\nu, A^*A\frac{D\nu}{Dt}\right) \\
&\qquad + \left(\frac{D\eta}{Dt} + div\, \psi + (\nabla \cdot \dot{x})\eta, A^*A\frac{D\nu}{Dt}\right) \qquad (5.4) \\
&= -\left(A\, div\, \rho + A(\nabla \cdot \dot{x})\nu, A\frac{D\nu}{Dt}\right) \\
&\qquad + \left(A\frac{D\eta}{Dt} + A\, div\, \psi + A(\nabla \cdot \dot{x})\eta, A\frac{D\nu}{Dt}\right) \\
&\le C\left\|A\frac{D\nu}{Dt}\right\|\left\{\|A\, div\, \rho\| + \|\nu\| + \left\|A\frac{D\eta}{Dt}\right\| + \|A\, div\, \psi\| + \|\eta\|\right\}.
\end{aligned}$$

Therefore we have

$$\left\| A\frac{D\nu}{Dt} \right\|^2_{L^2(0,T:L^2(\Omega))} \le C|||(\eta,\psi)|||^2.$$

Since

$$\left(A(\nabla\cdot\dot{x})\eta, A\frac{D\nu}{Dt}\right) = \left(AP_h(\nabla\cdot\dot{x})\eta, A\frac{D\nu}{Dt}\right) = \left(A(\nabla\cdot\dot{x})P_h\eta, A\frac{D\nu}{Dt}\right),$$

we also have

$$\left\| A\frac{D\nu}{Dt} \right\|^2_{L^2(0,T:L^2(\Omega))} \le C|||(\eta,\psi)|||^2_*.$$

Hence,

$$|||(\nu,\rho)||| \le C|||(u-v_h, \sigma-\mathcal{S}_h)|||,$$
$$|||(\nu,\rho)|||_* \le C|||(u-v_h, \sigma-\mathcal{S}_h)|||_*.$$

Applying the triangle inequality completes the proof. □

Next we define two additional norms $|||(\cdot,\cdot)|||_{D_t}$, $|||(\cdot,\cdot)|||_{D_t^*}$ by

$$|||(\eta,\psi)|||^2_{D_t} = \|\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \|AD_t\eta\|^2_{L^2(0,T:L^2(\Omega))} + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))},$$
$$|||(\eta,\psi)|||^2_{D_t^*} = \|P_h\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \|AD_t\eta\|^2_{L^2(0,T:L^2(\Omega))} + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))}$$

and use them to get the following pair of symmetric error estimates.

THEOREM 10. *Suppose there exist constants $c_1, c_2 > 0$ such that*

$$-\nabla_x\cdot\dot{x} \le c_1 \quad \text{and} \quad |\beta+\alpha\dot{x}| \le c_2$$

$\forall(x,t)\in Q$. *Then there exists a constant $C>0$, depending only on c_1, c_2, T, the bounds of coefficient a, and Ω, such that, for any piecewise smooth function v_h with $v_h(\cdot,t)\in V_h(t)$,*

$$|||(u-u_h, \sigma-\sigma_h)|||_{D_t} \le C|||(u-v_h, \sigma-F_h(v_h))|||_{D_t},$$
$$|||(u-u_h, \sigma-\sigma_h)|||_{D_t^*} \le C|||(u-v_h, \sigma-F_h(v_h))|||_{D_t^*}.$$

Proof. We slightly modify the proof of Theorem 9. The inequality (5.2) becomes

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nu\|^2 + (\alpha\rho + (\beta+\alpha\dot{x})\nu, \rho) \\
&\qquad = (D_t\eta + div\,\psi, \nu) - \frac{1}{2}\int_\Omega \nu^2(\nabla\cdot\dot{x})dx \\
&\qquad = (div\,A(D_t\eta + div\,\psi), \nu) - \frac{1}{2}\int_\Omega \nu^2(\nabla\cdot\dot{x})dx \qquad (5.5)\\
&\qquad = (\alpha\rho + (\beta+\alpha\dot{x})\nu, A(D_t\eta + div\,\psi)) - \frac{1}{2}\int_\Omega \nu^2(\nabla\cdot\dot{x})dx.
\end{aligned}$$

Therefore

$$\frac{d}{dt}\|\nu\|^2 + \alpha_1\|\rho\|^2 \le C\{\|\nu\|^2 + \|A(D_t\eta + div\,\psi)\|^2\}. \qquad (5.6)$$

It then follows from Gronwall's inequality and (3.3) that

$$\|\nu\|^2_{L^\infty(0,T:L^2(\Omega))} + \|A\ div\ \rho\|^2_{L^2(0,T:L^2(\Omega))} \le C|\!|\!|(\eta,\psi)|\!|\!|^2_{D_t},$$

and, since $P_h\nu = \nu$,

$$\|P_h\nu\|^2_{L^\infty(0,T:L^2(\Omega))} + \|A\ div\ \rho\|^2_{L^2(0,T:L^2(\Omega))} \le C|\!|\!|(\eta,\psi)|\!|\!|^2_{D_t^*}.$$

It remains to estimate $\|AD_t\nu\|^2$.

$$\begin{aligned}(AD_t\nu, AD_t\nu) &= (D_t\nu, A^*AD_t\nu)\\ &= -(div\ \rho, A^*AD_t\nu) - (D_t\eta + div\ \psi, A^*AD_t\nu)\\ &= -(A\ div\ \rho, AD_t\nu) - (AD_t\eta + A\ div\ \psi, AD_t\nu)\\ &\le C\|AD_t\nu\|\{\|A\ div\ \rho\| + \|AD_t\eta\| + \|A\ div\ \psi\|\}.\end{aligned}$$

Therefore

$$\|AD_t\nu\|^2_{L^2(0,T:L^2(\Omega))} \le C(\|\nu(0)\|^2 + |\!|\!|(\eta,\psi)|\!|\!|^2_{D_t})$$

and

$$\|AD_t\nu\|^2_{L^2(0,T:L^2(\Omega))} \le C(\|P_h\nu(0)\|^2 + |\!|\!|(\eta,\psi)|\!|\!|^2_{D_t^*}).$$

As before, the triangle inequality completes the proof. ∎

Note that Theorem 10 uses A but does not rely on Theorem 5; hence it does not require Condition 2 to hold.

6. Optimal order and superconvergent $L^2(\Omega)$ bounds in one space dimension. In one dimension, Ω is an interval. Let c_4 be a constant satisfying $c_4 \ge \frac{1}{2}(a_0 + \frac{\tilde{c}_2}{a_0})$, where $\tilde{c}_2 = \|b + \dot{x}\|_{L^\infty([0,T],L^\infty(\Omega))}$. Assume that a, b are sufficiently regular such that for any $g \in L^2(\Omega)$, the elliptic equation

$$\begin{cases} -\partial_x(a\partial_x w) + (b+\dot{x})\partial_x w + c_4 w = g & \text{in } \Omega,\\ w|_{\partial\Omega} = 0 \end{cases} \tag{6.1}$$

has a unique solution w satisfying $\|w\|_2 \le C\|g\|$.

We have the following optimal order $L^2(\Omega)$ error estimate.

THEOREM 11. *Suppose that Condition* 1 *holds and there exist constants c_1, c_2, c_3 such that, for any $t \in [0,T]$, $\|\partial_x\dot{x}\|_\infty, \|\partial_x b\|_\infty \le c_1$; $\|\beta + \alpha\dot{x}\|_\infty, \|\frac{D}{Dt}(\beta+\alpha\dot{x})\|_\infty \le c_2$; $\|\partial_x a\|_\infty, \|\frac{D\alpha}{Dt}\|_\infty \le c_3$. Then there exists a constant C, depending on C_1, c_1, c_2, c_3, Ω, T, and the bounds of coefficient a, such that, for h sufficiently small,*

$$\begin{aligned}\|u - u_h\| \le C\Bigg\{ &\|\underline{h}^{\min\{m+1,s\}}u\|_{L^\infty[0,T,\underline{H}^s]} + \left\|h\underline{h}^{\min\{m+1,s-1\}}\frac{Du}{Dt}\right\|_{L^2[0,T;\underline{H}^{s-1}]}\\ &+ \|\underline{h}^{\min\{m+2,s\}}\sigma\|_{L^2[0,T;\underline{H}^s]} + \|h^2\underline{h}^{\min\{m+1,s-2\}}\sigma\|_{L^2[0,T;\underline{H}^{s-1}]}\\ &+ \left\|h^2\underline{h}^{\min\{m+1,s-2\}}\frac{D\sigma}{Dt}\right\|_{L^2[0,T;\underline{H}^{s-1}]}\Bigg\}.\end{aligned} \tag{6.2}$$

Proof. This is an application of Theorem 10 using $|\!|\!|\cdot|\!|\!|_{D_t}$. Since $|\!|\!|(u-u_h, \sigma-\sigma_h)|\!|\!|_{D_t}$ dominates the term we want to bound, it suffices to show that $|\!|\!|(u-v_h, \sigma - F_h(v_h))|\!|\!|_{D_t}$ can be bounded by terms on the right-hand side of (6.2) for a suitable choice of v_h.

At each time we take the elliptic projection (v_h, S_h) of (u, σ) into $V_h \times H_h$ to satisfy

$$\begin{cases} (\alpha(S_h - \sigma) + (\beta + \alpha\dot{x})(v_h - u), \mathcal{X}) - (v_h - u, \partial_x \mathcal{X}) = 0 & \forall \mathcal{X} \in H_h, \\ (\partial_x(S_h - \sigma) + c_4(v_h - u), r) = 0 & \forall r \in V_h. \end{cases} \tag{6.3}$$

Notice that $S_h = F_h(v_h)$.

Differentiating (6.3) with respect to time, using Lemma 7 and properties of $\frac{D}{Dt}$, we have

(6.4)

$$\begin{cases} \left(\alpha \dfrac{D}{Dt}(S_h - \sigma) + (\beta + \alpha\dot{x}) \dfrac{D}{Dt}(v_h - u), \mathcal{X}\right) - \left(\dfrac{D}{Dt}(v_h - u), \partial_x \mathcal{X}\right) \\ \qquad = (E_1(S_h - \sigma), \mathcal{X}) + (E_2(v_h - u), \mathcal{X}) & \forall \mathcal{X} \in H_h, \\ \left(\partial_x \dfrac{D}{Dt}(S_h - \sigma) + c_4 \dfrac{D}{Dt}(v_h - u), r\right) = (E_3(v_h - u), r) & \forall r \in V_h, \end{cases}$$

where

$$\begin{aligned} E_1 &= -\left(\frac{D}{Dt}\alpha + \alpha \partial_x \dot{x}\right), \\ E_2 &= -\left(\frac{D}{Dt}(\beta + \alpha\dot{x}) + (\beta + \alpha\dot{x})\partial_x \dot{x}\right), \\ E_3 &= -c_4 \partial_x \dot{x}. \end{aligned}$$

Here we are also using the fact that, for any given $t_0 \in [0, T]$, $\mathcal{X}(x) \in H_h(t_0)$, and $r(x) \in V_h(t_0)$, we can define $\tilde{\mathcal{X}}(x, t) = \mathcal{X}(\mathcal{G}(\mathcal{G}^{-1}(x, t), t_0)) \in H_h(t)$ and $\tilde{r}(x, t) = r(\mathcal{G}(\mathcal{G}^{-1}(x, t), t_0)) \in V_h(t)$ for any $t \in [0, T]$, so that $\tilde{\mathcal{X}}(x, t_0) = \mathcal{X}(x)$, $\tilde{r}(x, t_0) = r(x)$, and $\frac{D}{Dt}\tilde{\mathcal{X}} = \frac{D}{Dt}\tilde{r} = 0$.

Because of (6.1), using the duality lemma in [3], for any h sufficiently small we have

$$\|v_h - P_h u\| \le C\{h\|S_h - \sigma\| + h\|P_h u - u\| + h^2 \|\partial_x(S_h - \sigma)\|\}. \tag{6.5}$$

From the second equation of (6.3) we have

$$\|P_h \partial_x(S_h - \sigma)\| \le C\|v_h - P_h u\|. \tag{6.6}$$

Therefore, using the triangle inequality,

$$\|\partial_x(S_h - \sigma)\| \le C\|v_h - P_h u\| + \|P_h \partial_x \sigma - \partial_x \sigma\|. \tag{6.7}$$

Also from the first equation of (6.3)

$$\begin{aligned} \|S_h - \Pi_h \sigma\|^2 &\le C(\alpha(S_h - \sigma), S_h - \Pi_h \sigma) + C(\alpha(\sigma - \Pi_h \sigma), S_h - \Pi_h \sigma) \\ &= C(v_h - u, \partial_x(S_h - \Pi_h \sigma)) - C((\beta + \alpha\dot{x})(v_h - u), S_h - \Pi_h \sigma) \\ &\qquad + C(\alpha(\sigma - \Pi_h \sigma), S_h - \Pi_h \sigma). \end{aligned} \tag{6.8}$$

Note that

$$(v_h - u, \partial_x(S_h - \Pi_h \sigma)) = (v_h - P_h u, P_h \partial_x(S_h - \sigma)) \le C\|v_h - P_h u\|^2;$$

therefore

$$\begin{aligned}\|S_h - \Pi_h\sigma\|^2 &\le C\{\|v_h - P_h u\|^2 + \|u - P_h u\|^2 + \|\sigma - \Pi_h\sigma\|^2\},\\ \|S_h - \sigma\|^2 &\le C\{\|v_h - P_h u\|^2 + \|u - P_h u\|^2 + \|\sigma - \Pi_h\sigma\|^2\}.\end{aligned} \tag{6.9}$$

Substituting into (6.5), we have

$$\|v_h - P_h u\| \le C\{h\|u - P_h u\| + h\|\sigma - \Pi_h\sigma\| + h^2\|P_h\partial_x\sigma - \partial_x\sigma\|\}. \tag{6.10}$$

Using the triangle inequality,

$$\|v_h - u\| \le C\{\|u - P_h u\| + h\|\sigma - \Pi_h\sigma\| + h^2\|P_h\partial_x\sigma - \partial_x\sigma\|\}. \tag{6.11}$$

Substituting (6.10) into (6.9),

$$\|S_h - \sigma\| \le C\{\|u - P_h u\| + \|\sigma - \Pi_h\sigma\| + h^2\|P_h\partial_x\sigma - \partial_x\sigma\|\}. \tag{6.12}$$

Similarly applying the duality lemmas in [3] to (6.4), noting that $\|E_1\|_\infty$, $\|E_2\|_\infty$, $\|E_3\|_\infty \le C$, we have for h sufficiently small,

$$\begin{aligned}\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| \le C\Bigg\{&h\left\|\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right\| + h\left\|P_h\frac{D}{Dt}u - \frac{D}{Dt}u\right\|\\ &+ h^2\left\|\partial_x\left(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right)\right\| + \|S_h - \sigma\| + \|v_h - u\|\Bigg\}.\end{aligned} \tag{6.13}$$

From the second equation of (6.4), we have

$$\left\|P_h\partial_x\left(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right)\right\| \le C\left\{\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| + \|v_h - P_h u\|\right\}.$$

Therefore a triangle inequality yields

$$\begin{aligned}&\left\|\partial_x\left(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right)\right\|\\ &\qquad\le C\left\{\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| + \|v_h - P_h u\| + \left\|P_h\partial_x\frac{D}{Dt}\sigma - \partial_x\frac{D}{Dt}\sigma\right\|\right\}.\end{aligned}$$

Also, from the first equation of (6.4)

$$\begin{aligned}\left\|\frac{D}{Dt}S_h - \Pi_h\frac{D}{Dt}\sigma\right\|^2 \le C\Bigg\{&\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\|^2 + \|v_h - P_h u\|^2\\ &+ \left\|P_h\frac{D}{Dt}u - \frac{D}{Dt}u\right\|^2 + \left\|\frac{D}{Dt}\sigma - \Pi_h\frac{D}{Dt}\sigma\right\|^2 + \|S_h - \sigma\|^2 + \|v_h - u\|^2\Bigg\},\end{aligned}$$

and the triangle inequality gives the same bound for $\|\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\|^2$. Substituting these into (6.13),

$$\begin{aligned}\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| \le C\Bigg\{&h\|v_h - P_h u\| + h\left\|P_h\frac{D}{Dt}u - \frac{D}{Dt}u\right\|\\ &+ h\left\|\frac{D}{Dt}\sigma - \Pi_h\frac{D}{Dt}\sigma\right\| + \|S_h - \sigma\| + \|v_h - u\| + h^2\left\|P_h\partial_x\frac{D}{Dt}\sigma - \partial_x\frac{D}{Dt}\sigma\right\|\Bigg\}.\end{aligned} \tag{6.14}$$

Choosing v_h in Theorem 10 to be the solution of (6.3) and noticing that S_h of (6.3) is equal to $F_h(v_h)$, $\partial_x \dot{x}$ is piecewise constant and therefore commutes with P_h, we have $\|P_h(u - v_h)\| = \|P_h u - v_h\|$ and

$$\begin{aligned}
\|AD_t(u - v_h)\| &= \|AP_h D_t(u - v_h)\| \\
&= \left\| A\left(P_h \frac{D}{Dt} u - \frac{D}{Dt} v_h\right) + A(\partial_x \dot{x}) P_h(u - v_h) \right\| \\
&\le C \left\| P_h \frac{D}{Dt} u - \frac{D}{Dt} v_h \right\| + C\|P_h u - v_h\| \\
&\le C \left\{ h \left\| P_h \frac{D}{Dt} u - \frac{D}{Dt} u \right\| + h \left\| \frac{D}{Dt}\sigma - \Pi_h \frac{D}{Dt}\sigma \right\| \right. \\
&\qquad + h^2 \left\| P_h \partial_x \frac{D}{Dt}\sigma - \partial_x \frac{D}{Dt}\sigma \right\| + \|u - P_h u\| \\
&\qquad \left. + \|\sigma - \Pi_h \sigma\| + h^2 \|P_h \partial_x \sigma - \partial_x \sigma\| \right\}
\end{aligned}$$

and

$$\begin{aligned}
\|A\partial_x(\sigma - F_h(v_h))\| &\le C\|P_h \partial_x \sigma - \partial_x S_h)\| \\
&\le C\{h\|u - P_h u\| + h\|\sigma - \Pi_h \sigma\| + h^2 \|P_h \partial_x \sigma - \partial_x \sigma\|\}.
\end{aligned}$$

Using approximation properties of P_h and Π_h, the proof of Theorem 11 is then complete. □

With more restrictions on the coefficients and the mesh movement, we can obtain the following superconvergence result.

THEOREM 12. *Suppose that the conditions of Theorem* 11 *hold and that there exist constants* $c_5, c_6, c_7 > 0$ *such that* $\|\partial_x(\frac{D}{Dt}(\beta + \alpha\dot{x}))\|_\infty \le c_5$, $\|\partial_x \frac{D\alpha}{Dt}\|_\infty \le c_6$, *and* $|\partial_x \dot{x}(x_i-) - \partial_x \dot{x}(x_i+)| \le c_7 \min\{h_i, h_{i+1}\}$ $\forall i$. *Then there exists a constant* C, *depending on* $C_1, c_1, c_2, c_3, c_5, c_6, c_7$, Ω, T, *and the bounds of coefficient* a, *such that for any* h *sufficiently small*

$$\begin{aligned}
\|P_h u - u_h\| \le C \Bigg\{ & \|h\underline{h}^{\min\{m+1,s\}} u\|_{L^\infty[0,T;\underline{H}^s]} + \left\| h\underline{h}^{\min\{m+1,s-1\}} \frac{Du}{Dt} \right\|_{L^2[0,T;\underline{H}^{s-1}]} \\
& + \|h\underline{h}^{\min\{m+2,s-1\}} \sigma_{-L^2[0,T;\underline{H}^{s-1}]} + \|h^2 \underline{h}_i^{\min\{m+1,s-2\}} \sigma\|_{L^2[0,T;\underline{H}^{s-1}]} \\
& + \left\| h^2 \underline{h}^{\min\{m+1,s-2\}} \frac{D\sigma}{Dt} \right\|_{L^2[0,T;\underline{H}^{s-1}]} \Bigg\}.
\end{aligned}$$

Proof. We slightly modify the proof of Theorem 11. First we apply the duality argument in [3] to (6.3) to get

$$\|S_h - \sigma\|_{-1} \le C\{h^2 \|\partial_x(S_h - \sigma)\| + \|v_h - P_h u\| + h\|u - P_h u\|\}. \tag{6.15}$$

Let ω be a piecewise linear continuous function on $\{\Omega_i\}$ such that $\omega(x_i) = \{\partial_x \dot{x}(x_i-) + \partial_x \dot{x}(x_i+)\}/2$ for any i. Then it is easy to see that $\|\omega - \partial_x \dot{x}\|_\infty \le Ch$ and $\|\omega\|_1 \le C$.

From the right-hand side of (6.4) we have

$$\begin{aligned}
(E_3(v_h-u),r) &= (E_3(v_h-P_hu),r),\\
(E_2(v_h-u),\mathcal{X}) &= -\left(v_h-u,\frac{D}{Dt}(\beta+\alpha\dot{x})\,\mathcal{X}\right)\\
&\quad -((\partial_x\dot{x})(v_h-u),(\beta+\alpha\dot{x})\cdot\mathcal{X}-P_h((\beta+\alpha\dot{x})\cdot\mathcal{X}))\\
&\quad -((\partial_x\dot{x})(v_h-P_hu),P_h((\beta+\alpha\dot{x})\cdot\mathcal{X}))\\
&\le C\{\|v_h-u\|_{-1}+h\|v_h-u\|+\|v_h-P_hu\|\}\|\mathcal{X}\|_1\\
&\le C\{h\|u-P_hu\|+\|v_h-P_hu\|\}\|\mathcal{X}\|_1,\\
(E_1(S_h-\sigma),\mathcal{X}) &= -\left(S_h-\sigma,\frac{D\alpha}{Dt}\mathcal{X}\right)-(\alpha(\partial_x\dot{x}-\omega)(S_h-\sigma),\mathcal{X})\\
&\quad -(S_h-\sigma,\alpha\omega\mathcal{X})\\
&\le C\{\|S_h-\sigma\|_{-1}+h\|S_h-\sigma\|\}\|\mathcal{X}\|_1.
\end{aligned}$$

Following the duality lemmas in [3] again and also using (6.15), we have

$$\begin{aligned}
\left\|\frac{D}{Dt}v_h-P_h\frac{D}{Dt}u\right\| \le C\Bigg\{&\|v_h-P_hu\|+h\left\|P_h\frac{D}{Dt}u-\frac{D}{Dt}u\right\|\\
&+h\left\|\frac{D}{Dt}\sigma-\Pi_h\frac{D}{Dt}\sigma\right\|+h\|S_h-\sigma\|\\
&+h^2\left\|P_h\partial_x\frac{D}{Dt}\sigma-\partial_x\frac{D}{Dt}\sigma\right\|+h\|u-P_hu\|+h^2\|\partial_x(S_h-\sigma)\|\Bigg\}.
\end{aligned}\tag{6.16}$$

Note that $\|AD_t(u-v_h)\| \le C\|P_h\frac{D}{Dt}u-\frac{D}{Dt}v_h\|+C\|P_hu-v_h\|$, and $\|u_h-P_hu\|$ is dominated by $\|\!|(u-u_h,\sigma-\sigma_h)|\!\|_{D_t^*}$; the rest of the proof is similar to that of Theorem 11. □

We conduct a convergence test using the equation $u_t-(u_x-b_1u)_x=0$, for $(x,t)\in(0,10)\times(0,1)$, $u(0,t)=u(10,t)=0, t\in[0,1]$, $u(x,0)=u_0(x), x\in[0,10]$. Here $u_0(x)$ is a smooth nonnegative function with support in $[3,5]$; see Figure 1. $b_1(x)$ is a C^2 nonnegative function such that $b_1=3.5$ on $[2,7]$, $b_1=0$ on $[0,1]\cup[8,10]$, and b_1 is a 5th order polynomial in $(1,2)$ and $(7,8)$. Three cases are examined. The case referred to as "moving mesh" is based on a specified mesh technique discussed in the next paragraph. There is a characteristic moving mesh case in which the mesh points are moved along characteristics, starting from the same mesh as the first case. There is also a case that uses a fixed uniform mesh. In all cases, we have taken the time step sufficiently small that the time truncation can be ignored, i.e., we are looking at the continuous-time case.

We illustrate a simple, but powerful, moving mesh strategy in which the mesh is specified by giving the mesh at the initial and the final times, and the meshes are then connected. A specified mesh calculation is very easy to program if one has a code that allows for variable mesh spacing; all that is required is a change in the convective term to account for $\alpha\dot{x}$. The selection of the mesh is easier if one can look at a coarse grid calculation. (One can specify the mesh at more than two levels, and various techniques can be used to connect the mesh points.) The initial mesh is taken so that the density of mesh points in $(0,6)$ is about one third higher than the average density across the entire interval. In the specified movement case the mesh at the final time $T=1$ is such that the local mesh density is proportional to $\epsilon+|\frac{\partial^2u}{\partial x^2}|$, where u is approximated by a coarse uniform grid numerical solution, and ϵ is taken to be

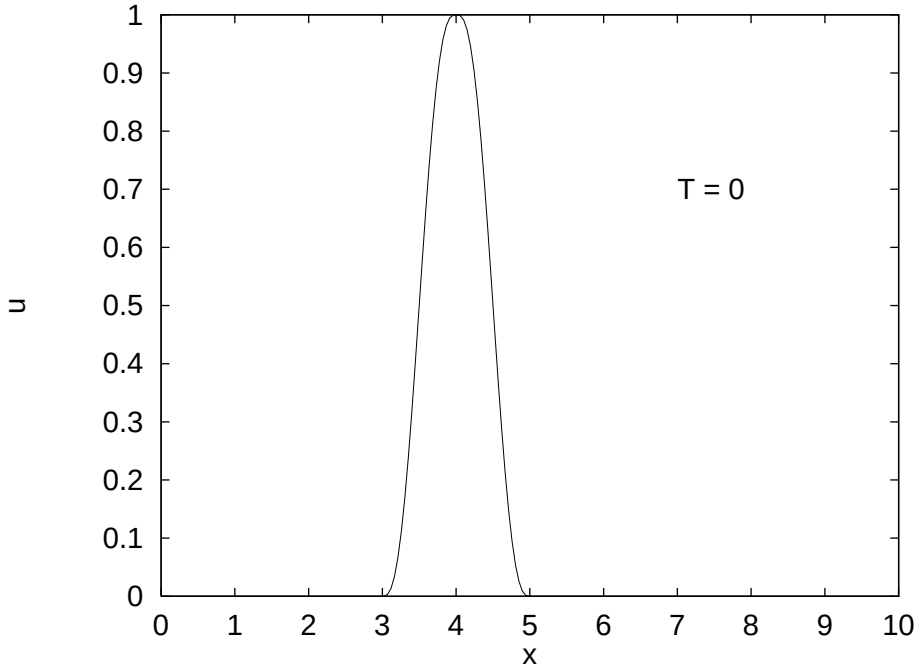


FIG. 1. *Initial value.*

0.2. (The value of ϵ is between the average and the maximum absolute value of the second derivative.) Figure 2 shows the mesh movement in the space time plane with mesh cell number $n = 40$.

In Figure 3, the final solution at $T = 1$ with $n = 20$ for the moving mesh mixed method is compared with the solution from the mixed method with an evenly distributed fixed mesh. Each of these solutions is used to produce a reconstructed continuous piecewise linear approximation $\tilde{u}_h$, through connecting the points $(x_i, u_h(x_i))$, where the x_i's are the cell centers. The "exact" solution is the result of a very fine grid calculation. As expected, higher resolution is achieved for the moving mesh near $(7, 8)$.

In Table 1 the comparison between the moving and fixed meshes is given in quantitative terms. The table clearly shows the first order convergence of the error and the second order convergence of the approximation built on the supercloseness of the midcell values.

TABLE 1
Comparative L^2 and L^∞ errors.

	Moving mesh			*Fixed mesh*		
n	$\|u - u_h\|$	$\|P_h u - u_h\|$	$\|u - \tilde{u}_h\|_\infty$	$\|u - u_h\|$	$\|P_h u - u_h\|$	$\|u - \tilde{u}_h\|_\infty$
20	0.052	0.0040	0.013	0.083	0.013	0.075
40	0.026	0.0010	0.0035	0.040	0.0027	0.025
80	0.013	0.00027	0.0011	0.020	0.00070	0.0066
160	0.0066	0.000072	0.00026	0.010	0.00020	0.0018

For this problem, using the same initial mesh as in the specified movement case, following the characteristics produces an overconcentration of mesh points in $(7, 8)$

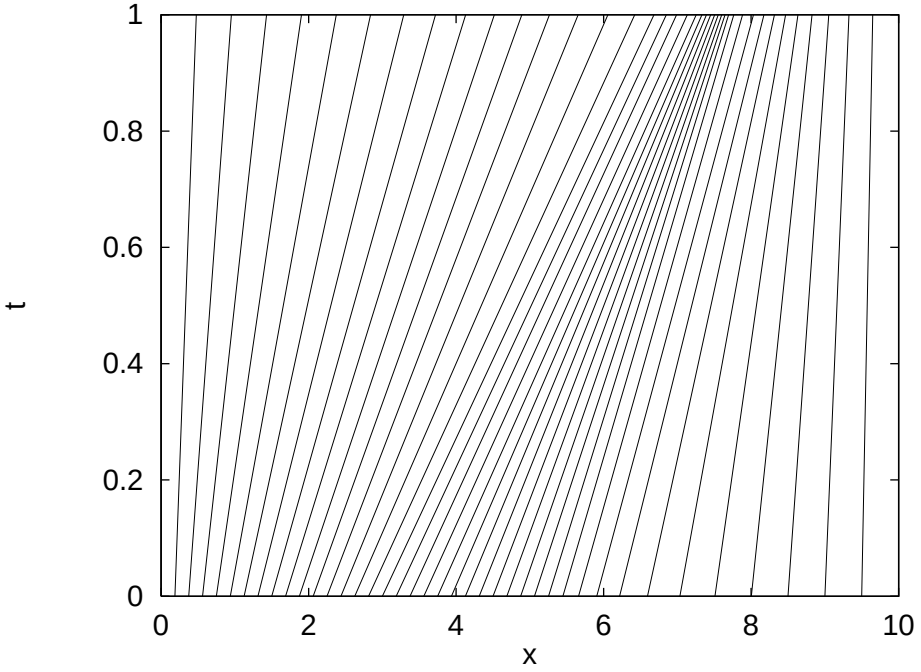


FIG. 2. *Moving mesh in the space time plane.*

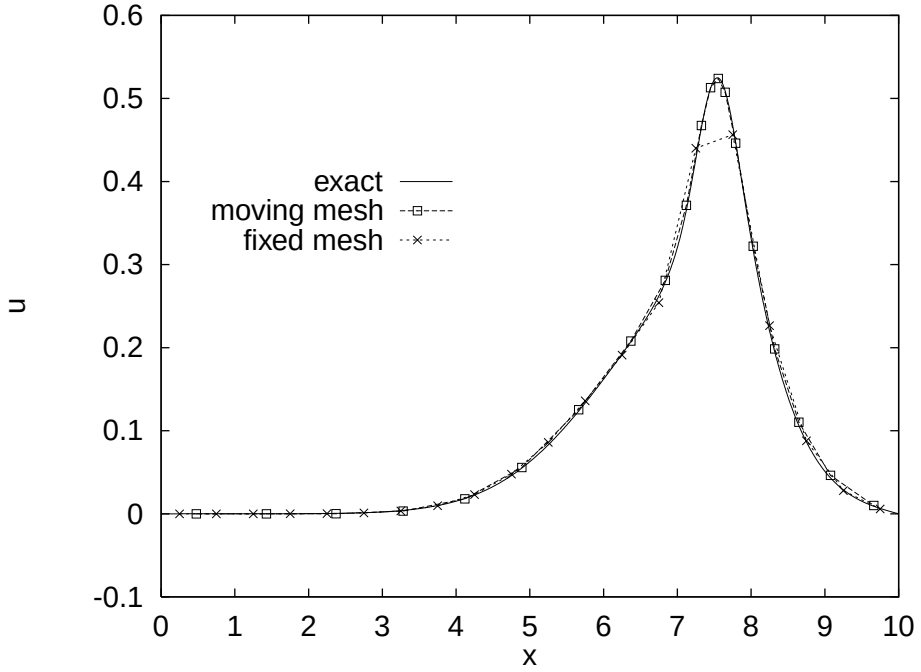


FIG. 3. *Specified and fixed mesh approximations at* $T = 1$.

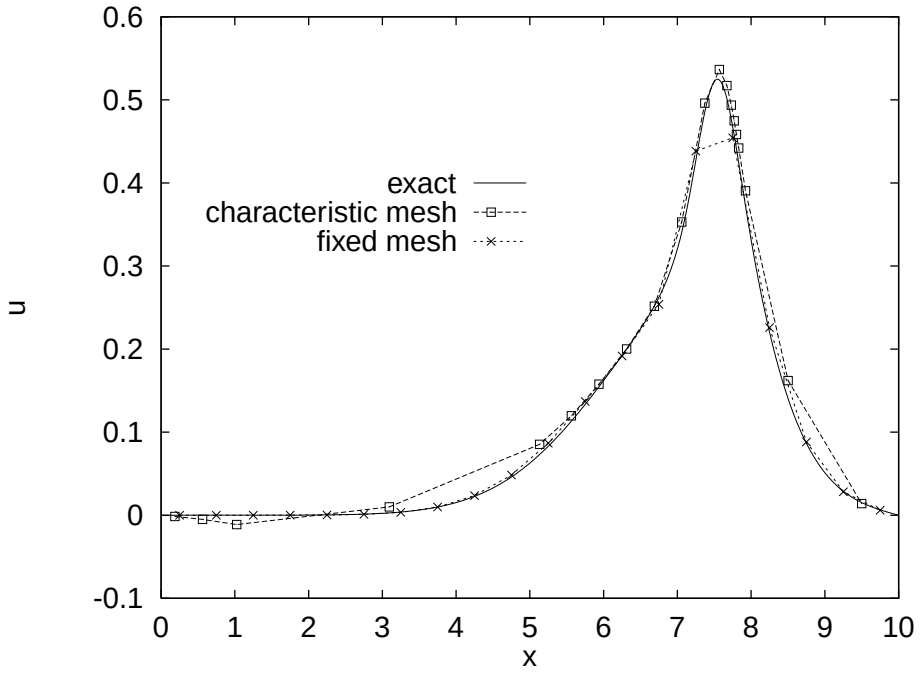


FIG. 4. *Characteristic and fixed mesh approximations at* $T = 1$.

TABLE 2
L^2 *and* L^∞ *errors with mesh moving along characteristics.*

n	$\|u - u_h\|$	$\|P_h u - u_h\|$	$\|u - \tilde{u}_h\|_\infty$
20	0.095	0.019	0.040
40	0.050	0.0063	0.018
80	0.026	0.0014	0.0059
160	0.013	0.00039	0.0016

but too few mesh points in $(1, 5.5)$. In Figure 4 and Table 2 computational results similar to those in Figure 3 and Table 1 are given. In this case the mesh is moving along characteristics.

7. Another mixed method. Consider the nonconservative form of (2.1):

$$\begin{cases} \partial_t u - \nabla \cdot (a \nabla u) + b \cdot \nabla u + cu = f & \text{on } Q, \\ u = 0 & \text{on } \partial\Omega \times (0, T), \\ u = u_0 & \text{for } t = 0. \end{cases} \tag{7.1}$$

Let $\sigma = a\nabla u$ and $\alpha = 1/a$, $\beta = b/a$. A natural mixed form is

$$\begin{cases} (\alpha\sigma, \mathcal{X}) + (u, div\, \mathcal{X}) = 0 & \forall \mathcal{X} \in H(div, \Omega), \\ \left(\dfrac{Du}{Dt} + div\, \sigma, r \right) + ((\beta - \dot{x}\alpha) \cdot \sigma, r) + (cu, r) = (f, r) & \forall r \in L^2(\Omega). \end{cases} \tag{7.2}$$

Note that, with a little abuse of the notations, $a, b, c, u, \alpha, \beta, \sigma$ have been redefined.

We will keep on using the relevant notations and results from previous sections unless otherwise specified.

The mixed method is to find $u_h : [0,T] \to V_h$ and $\sigma_h : [0,T] \to H_h$ such that

$$\begin{cases} (\alpha\sigma_h, \mathcal{X}) + (u_h, div\, \mathcal{X}) = 0 & \forall \mathcal{X} \in H_h, \\ \left(\dfrac{Du_h}{Dt} + div\, \sigma_h, r\right) + ((\beta - \dot{x}\alpha)\cdot\sigma_h, r) + (cu_h, r) = (f, r) & \forall r \in V_h. \end{cases} \tag{7.3}$$

The above formulas are introduced in [1]. But here we deal with general mesh movement, and therefore $\beta - \dot{x}\alpha$ is not necessarily zero.

We define the norm $|\!|\!|(\cdot,\cdot)|\!|\!|_c$ by

$$\begin{aligned} |\!|\!|(\eta,\psi)|\!|\!|_c^2 = \|\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \left\|A\frac{D\eta}{Dt}\right\|^2_{L^2(0,T:L^2(\Omega))} \\ + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))} + \|\psi\|^2_{L^2(0,T:L^2(\Omega))}. \end{aligned} \tag{7.4}$$

Let L_h be a linear operator $V_h(t) \to H_h(t)$ such that for any $v_h \in V_h(t)$

$$(\alpha L_h(v_h), \mathcal{X}) + (v_h, div\, \mathcal{X}) = 0 \quad \forall \mathcal{X} \in H_h.$$

We have the following theorem, whose proof is similar to that of Theorem 9.

THEOREM 13. *Suppose that Condition* 2 *holds and there exist constants* c_1, c_2 *such that*

$$\nabla_x \cdot \dot{x} \le c_1 \quad \text{and} \quad |\beta - \alpha\dot{x}| \le c_2$$

$\forall (x,t) \in Q$. *Then there exists a constant* $C > 0$, *depending only on* C_2, c_1, c_2, T, *the bounds of coefficients* a *and* c, *and* Ω, *such that, for any piecewise smooth function* v_h *with* $v_h(\cdot,t) \in V_h(t)$,

$$|\!|\!|(u - u_h, \sigma - \sigma_h)|\!|\!|_c \le C |\!|\!|(u - v_h, \sigma - L_h(v_h))|\!|\!|_c.$$

Introduce another norm $|\!|\!|(\cdot,\cdot)|\!|\!|_{c^*}$ by

$$\begin{aligned} |\!|\!|(\eta,\psi)|\!|\!|^2_{c^*} = \|P_h\eta\|^2_{L^\infty(0,T:L^2(\Omega))} + \left\|A\frac{D\eta}{Dt}\right\|^2_{L^2(0,T:L^2(\Omega))} + \|A(div\,\psi)\|^2_{L^2(0,T:L^2(\Omega))} \\ + \|A((\beta - \alpha\dot{x})\cdot\psi)\|^2_{L^2(0,T:L^2(\Omega))} + \|A(c\eta)\|^2_{L^2(0,T:L^2(\Omega))}. \end{aligned}$$

We have another theorem whose proof is similar to that of Theorem 10, also using Theorem 5.

THEOREM 14. *Suppose that Condition* 2 *holds and there exist constants* c_1, c_2 *such that*

$$\nabla_x \cdot \dot{x} \le c_1 \quad \text{and} \quad |\beta + \alpha\dot{x}| \le c_2$$

$\forall (x,t) \in Q$. *Then there exists a constant* $C > 0$, *depending only on* C_2, c_1, c_2, T, *the bounds of coefficients* a *and* c, *and* Ω, *such that, for any piecewise smooth function* v_h *with* $v_h(\cdot,t) \in V_h(t)$,

$$|\!|\!|(u - u_h, \sigma - \sigma_h)|\!|\!|_{c^*} \le C |\!|\!|(u - v_h, \sigma - L_h(v_h))|\!|\!|_{c^*}.$$

Parallel to what was done in section 6, we derive an optimal convergence result for one dimension in the next theorem. In particular, the L^2 norm of $u_h - P_h u$ is superconvergent.

Assume that a is sufficiently regular so that for any $\xi \in L^2(\Omega)$ the equation

$$\begin{cases} -\partial_x(a\partial_x w) = g & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega \end{cases} \tag{7.5}$$

has a unique solution satisfying $\|w\|_2 \le C\|g\|$. We have the following theorem.

THEOREM 15. *Suppose that Condition* 1 *holds and there exist constants* c_1, c_2, c_3 *such that* $|\partial_x \dot{x}|, |\frac{D\alpha}{Dt}| \le c_1$; $|\beta + \alpha\dot{x}| \le c_2$; $|\partial_x c| \le c_3$ $\forall (x,t) \in Q$. *Then there exists a constant* $C > 0$, *depending only on* C_1, c_1, c_2, c_3, T, *the bounds on coefficients* a *and* c, *and* Ω, *such that for any* h *sufficiently small*

$$\begin{aligned} \|u_h - P_h u\| \le C\Bigg\{ & \|\underline{h}^{\min\{m+2,s+1\}}\sigma\|_{L^\infty[0,T;\underline{H}^{s+1}]} + \|h\underline{h}^{\min\{m+1,s\}}\sigma\|_{L^2[0,T;\underline{H}^{s+1}]} \\ & + \left\|h\underline{h}^{\min\{m+2,s\}}\frac{D\sigma}{Dt}\right\|_{L^2[0,T;\underline{H}^s]} + \left\|h^2\underline{h}^{\min\{m+1,s-1\}}\frac{D\sigma}{Dt}\right\|_{L^2[0,T;\underline{H}^s]} \\ & + \|h\underline{h}^{\min\{m+1,s\}}u\|_{L^2[0,T;\underline{H}^s]}\Bigg\} \end{aligned}$$

and

$$\begin{aligned} \|u - u_h\| \le C\Bigg\{ & \|\underline{h}^{\min\{m+2,s\}}\sigma\|_{L^\infty[0,T;\underline{H}^s]} + \|h\underline{h}^{\min\{m+1,s-1\}}\sigma\|_{L^2[0,T;\underline{H}^s]} \\ & + \left\|h\underline{h}^{\min\{m+2,s-1\}}\frac{D\sigma}{Dt}\right\|_{L^2[0,T;\underline{H}^{s-1}]} + \left\|h^2\underline{h}_i^{\min\{m+1,s-2\}}\frac{D\sigma}{Dt}\right\|_{L^2[0,T;\underline{H}^{s-1}]} \\ & + \|h\underline{h}^{\min\{m+1,s\}}u\|_{L^2[0,T;\underline{H}^s]}\Bigg\}. \end{aligned}$$

Proof. The proof of the first estimate is an application of Theorem 14. Since $\|\!|(u - u_h, \sigma - \sigma_h)|\!\|_{c^*}$ dominates the term we want to bound, it suffices to show that $\|\!|(u - v_h, \sigma - L_h(v_h))|\!\|_{c^*}$ can be bounded by terms on the right-hand side of the first estimate. The second estimate follows from a triangle inequality.

Consider the following elliptic projection:

$$\begin{cases} (\alpha(S_h - \sigma), \mathcal{X}) + (v_h - u, \partial_x \mathcal{X}) = 0 & \forall \mathcal{X} \in H_h, \\ (\partial_x(S_h - \sigma), r) = 0 & \forall r \in V_h. \end{cases} \tag{7.6}$$

Notice that $S_h = L_h(v_h)$.

Differentiating (7.6) with respect to time and using Lemma 7 and properties of $\frac{D}{Dt}$, we have

$$\begin{cases} \left(\alpha\dfrac{D}{Dt}(S_h - \sigma), \mathcal{X}\right) + \left(\dfrac{D}{Dt}(v_h - u), \partial_x \mathcal{X}\right) \\ \qquad = (E_4(S_h - \sigma), \mathcal{X}), & \forall \mathcal{X} \in H_h, \\ \left(\partial_x \dfrac{D}{Dt}(S_h - \sigma), r\right) = 0 & \forall r \in V_h, \end{cases} \tag{7.7}$$

where $E_4 = -(\frac{D}{Dt}\alpha + \alpha\partial_x\dot{x})$. Using the duality lemma in [3], we have

$$\|v_h - P_h u\| \le C\{h\|S_h - \sigma\| + h^2\|\partial_x(S_h - \sigma)\|\}. \tag{7.8}$$

Also from the second equation of (7.6),

$$\|\partial_x(S_h - \Pi_h\sigma)\| = 0 \quad \text{and} \quad \|P_h\partial_x(S_h - \sigma)\| = 0,$$

so $\|\partial_x(S_h - \sigma)\| = \|P_h\partial_x\sigma - \partial_x\sigma\|$. From the first equation of (7.6),

$$\|S_h - \Pi_h\sigma\| \le C\|\sigma - \Pi_h\sigma\|, \quad \text{and so} \quad \|S_h - \sigma\| \le C\|\sigma - \Pi_h\sigma\|.$$

Therefore

$$\|v_h - P_h u\| \le C\{h\|\sigma - \Pi_h\sigma\| + h^2\|P_h\partial_x\sigma - \partial_x\sigma\|\}.$$

Similarly for equation (7.7),

$$\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| \le C\left\{h\left\|\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right\| + h^2\left\|\partial_x\left(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right)\right\| + \|S_h - \sigma)\|\right\} \tag{7.9}$$

and $\|P_h\partial_x(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma)\| = 0$, $\|\partial_x(\frac{D}{Dt}S_h - \Pi_h\frac{D}{Dt}\sigma)\| = 0$. So

$$\left\|\partial_x\left(\frac{D}{Dt}S_h - \frac{D}{Dt}\sigma\right)\right\| = \left\|P_h\partial_x\frac{D}{Dt}\sigma - \partial_x\frac{D}{Dt}\sigma\right\|.$$

Also from the first equation of (7.7)

$$\left\|\frac{D}{Dt}S_h - \Pi_h\frac{D}{Dt}\sigma\right\| \le C\left\{\|S_h - \sigma\| + \left\|\frac{D}{Dt}\sigma - \Pi_h\frac{D}{Dt}\sigma\right\|\right\}.$$

Therefore

$$\begin{aligned}\left\|\frac{D}{Dt}v_h - P_h\frac{D}{Dt}u\right\| \le C\Bigg\{&\|S_h - \sigma\| + h\left\|\frac{D}{Dt}\sigma - \Pi_h\frac{D}{Dt}\sigma\right\| \\ &+ h^2\left\|\partial_x\frac{D}{Dt}\sigma - P_h\partial_x\frac{D}{Dt}\sigma\right\|\Bigg\}.\end{aligned} \tag{7.10}$$

In Theorem 15, choose v_h to be the solution of (7.6). Note that

$$\begin{aligned}\left\|A\frac{D}{Dt}(u - v_h)\right\| &\le C\left\|P_h\frac{D}{Dt}u - \frac{D}{Dt}v_h\right\|, \\ \|A((\beta - \alpha\dot{x})\cdot(\sigma - S_h))\| &\le C\|\sigma - S_h\| \\ &\le C\|\sigma - \Pi_h\sigma\|, \\ \|A(c(u - v_h))\| &\le \|A((c - \bar{c})(u - v_h))\| + \|A(\bar{c}(u - v_h))\| \\ &\le C\|(c - \bar{c})(u - v_h)\| + \|A(\bar{c}P_h(u - v_h))\| \\ &\le Ch(\|u - P_h u\| + \|P_h u - v_h\|) + C\|P_h u - v_h\|,\end{aligned}$$

where $\bar{c}|_{\Omega_i} \equiv (1/|\Omega_i|)\int_{\Omega_i} c dx\ \forall i$ is a piecewise constant function which commutes with P_h. The proof is completed using the approximation properties of the projections P_h and Π_h. ∎

REFERENCES

[1] T. Arbogast and M. F. Wheeler, *A characteristics-mixed finite element method for advection-dominated transport problems*, SIAM J. Numer. Anal., 32 (1995), pp. 404–424.

[2] R. E. Bank and R. F. Santos, *Analysis of some moving space-time finite element methods*, SIAM J. Numer. Anal., 30 (1993), pp. 1–18.

[3] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[4] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[5] T. F. Dupont, *Mesh modification for evolution equations*, Math. Comp., 39 (1982), pp. 85–107.

[6] T. F. Dupont and Y. Liu, *Symmetric error estimates for moving mesh Galerkin methods for advection-diffusion equations*, SIAM J. Numer. Anal., 40 (2002), pp. 914–927.

[7] C. Johnson and V. Thomée, *Error estimates for some mixed finite element methods for parabolic type problems*, RAIRO Modél. Anal. Numér., 15 (1981), pp. 41–78.

[8] Y. Liu, *The Analysis for a Moving Mesh Finite Element Method*, preprint.

[9] K. Miller, *Moving finite elements, part* II, SIAM J. Numer. Anal., 18 (1981), pp. 1033–1057.

[10] R. Miller and K. Miller, *Moving finite elements, part* I, SIAM J. Numer. Anal., 18 (1981), pp. 1019–1032.

[11] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer–Verlag, Berlin, 1977, pp. 292–315.

[12] D. Q. Yang, *A characteristic mixed method with dynamic finite-element space for convection-dominated diffusion problems*, J. Comput. Appl. Math., 43 (1992), pp. 343–353.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2292–2310

A MONOTONIC METHOD FOR THE NUMERICAL SOLUTION OF SOME FREE BOUNDARY VALUE PROBLEMS*

RAPHAÈLE HERBIN†

Abstract. This work presents an efficient monotonic algorithm for the numerical solution of the obstacle problem and the Signorini problems, when they are discretized either by the finite element method or by the finite volume method. The convergence of this algorithm applied to the discrete problem is proven in both cases.

Key words. variational inequalities, iterative algorithm, obstacle problem, Signorini problem, finite element and finite volume methods

AMS subject classifications. 65K10, 49A29

PII. S0036142900380558

1. Introduction.

We are interested here in the numerical solution of some free boundary problems which are discretized by the finite element or the finite volume method. We introduce an efficient monotonic algorithm which applies to both the obstacle problem and the Signorini problem.

The obstacle problem is one of the simplest unilateral problems; it arises when modelling a constrained membrane in the classical linear elasticity theory. Signorini boundary conditions may be encountered in fluid mechanics and heat transfer problems when modelling, for instance, the flow through semipermeable boundaries. They are also encountered in contact problems in elasticity. The Signorini boundary conditions which we deal with here arise from modelling the so-called triple point of an electrochemical reaction (see [23]) and involve a diffusion operator. Both the obstacle and the Signorini problems may be written as variational inequalities.

The obstacle problem appeared in the mathematical literature in the work of Stampacchia [28] (see also [29], [30]), and the first rigorous analysis of a class of Signorini problems was published in 1963 by Fichera [12], [13]. The mathematical analysis including the study of existence, uniqueness, and regularity of the solution for the obstacle problem and Signorini problem may be found in [24], [25], and [7].

The obstacle problem and the Signorini problem are classically discretized by the finite element method formulated in [21], [16]; see also [27], [2], [1], [3] for more recent work (some of them subsequent to the submission of this paper) on elastic contact problems. In the case of diffusion problems, with which we are concerned here, a cell-center finite volume scheme was also recently applied and shown to converge [19].

The approximate problem can be solved by a duality method [16], [15], [22]. In [16], a point overrelaxation method with projection is also studied and found to be cheaper in terms of computational cost than the duality method. Another candidate for the resolution of the approximated Signorini problem is the penalty method (see [21] and references therein); it has the disadvantage of yielding ill-conditioned systems, while our algorithm deals only with submatrices of the whole discretization matrix.

*Received by the editors November 7, 2000; accepted for publication (in revised form) June 3, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/38055.html
†Université de Provence, CMI, 39 rue Joliot Curie, 13453 Marseille cedex 13, France (herbin@cmi.univ-mrs.fr).

We present here a particularly simple iterative monotonic algorithm that is inspired by a procedure used for multiphase flow modelling [8]. We show that it may be applied to the finite linear element approximation of the obstacle problem and the finite volume discretization of both the obstacle problem and the Signorini problem. In each case, we prove the monotonicity of the algorithm and its convergence in a finite number of iterations towards the exact solution to the discrete problem.

2. The obstacle problem. We consider here the so-called obstacle problem, which arises for instance in the modelling of contact problems (see [7]):

$$\text{(1)} \qquad \begin{cases} u \in \mathcal{K} = \{v \in H_0^1(\Omega),\ v \le \psi \text{ on } \Omega\}, \quad \text{satisfying} \\ \displaystyle\int_\Omega \nabla u(x) \cdot \nabla (v-u)(x) dx \ge \int_\Omega f(x)(v-u)(x) dx \qquad \forall v \in \mathcal{K}, \end{cases}$$

where the following holds.

Assumption 2.1.

1. Ω is a bounded open polygonal subset of $\mathbb{R}^d$, with $d = 2$ or 3.
2. $f \in L^2(\Omega)$ and $\psi \in H^1(\Omega) \cap C(\Omega)$ and $\psi \ge 0$ a.e. in the neighborhood of $\partial\Omega$.

Under these assumptions, it is well known that there exists a unique solution to problem (1), thanks to Stampacchia's theorem. Indeed, the set $\mathcal{K}$ is nonempty since $\min(0, \psi)$ belongs to $\mathcal{K}$. Furthermore, it is now classical that the solution to problem (1) belongs to $H^2(\Omega)$ (see [4]). Thanks to this H^2 regularity of the solution, it is easily shown that the variational inequality (1) can be written as a free boundary problem in the following way.

THEOREM 2.1. *Under Assumption* 2.1, *if u is a solution to the free boundary problem*

$$\text{(2)} \qquad \begin{cases} u \in H^2(\Omega) \cap H_0^1(\Omega), \quad \textit{satisfying} \\ u \le \psi \quad \textit{a.e. on } \Omega, \\ \Delta u + f \ge 0 \quad \textit{a.e. on } \Omega, \\ (\Delta u + f)(\psi - u) = 0 \quad \textit{a.e. on } \Omega, \end{cases}$$

then u is a solution to Problem (1).

Conversely, if $\psi \ge 0$ a.e. on Ω and u is a solution to Problem (1), *then u is a solution to Problem* (2).

We shall study the monotonic algorithm for both the finite element and the finite volume discretization of the above problem. Let us first start with the finite element method.

2.1. Approximation by the finite element method. Let $\mathcal{T}$ denote a "classical" triangulation of Ω (see, e.g., [6]).

DEFINITION 2.2 (triangulation $\mathcal{T}$ of Ω). *Let $\mathcal{T}$ be a finite set of triangles if $d = 2$, or tetrahedra if $d = 3$, such that*

(i) *$T \subset \overline{\Omega}\ \forall T \in \mathcal{T}$, and $\cup_{T\in\mathcal{T}} = \overline{\Omega}$;*

(ii) *for any $(T_1, T_2) \in \mathcal{T}^2$ with $T_1 \ne T_2$, either the $(d-1)$-dimensional Lebesgue measure of $\overline{T}_1 \cap \overline{T}_2$ is 0, or T_1 and T_2 have only a whole common edge (or face if $d = 3$).*

Let Σ be the set of vertices of triangles (tetrahedra) of $\mathcal{T}$ which belong to Ω (i.e., do not lie on the boundary) and $N = \operatorname{card}(\Sigma)$.

The set $H_0^1(\Omega)$ is classically approximated by

$$\text{(3)} \qquad V_h = \{v \in H_0^1(\Omega) \cap C^0(\overline{\Omega}),\ v_{|\partial\Omega} = 0,\ v_{|T} \in P_1\},$$

where $v_{|\partial\Omega}$ is the trace of v on $\partial\Omega$, $v_{|_T}$ denotes the restriction of v to T, and P_1 the space of polynomials in x_1 and x_2 of degree less than or equal to one. Assuming that $\Sigma = \{s_i,\ i \in \{1,\dots,N\}\}$, let $(\varphi_i)_{i\in\{1,\dots,N\}}$ be the N basis functions of V_h such that $\varphi_i(s_i) = 1$ and $\varphi_i(s_j) = 0\ \forall i \neq j$; notice that the functions φ_i are linear on each triangle for which s_i is a vertex.

We then consider the following approximate problem:

$$\text{(4)} \qquad \begin{cases} \tilde{u} \in K_h = \{v \in V_h,\ v(s) \leq \psi(s)\ \forall s \in \Sigma\}, \quad \text{satisfying} \\ \displaystyle\int_\Omega \nabla\tilde{u}(x)\cdot\nabla(v-\tilde{u})(x)dx \geq \int_\Omega f(x)(v-\tilde{u})(x)dx \qquad \forall v \in K_h. \end{cases}$$

By Stampacchia's theorem, problem (4) has a unique solution. Indeed, the set K_h is nonempty since the function $\min(\sum_{i=1,N}\psi_i\varphi_i, 0)$ belongs to K_h. Error estimates for the approximate finite element solution of the elliptic variational inequalities can be found in Falk [10], Mosco and Strang [26], Glowinski, Lions, and Trémolières [16], Ciarlet [6], Brezzi, Hager, and Raviart [5], and Falk and Mercier [11]. Error estimates of order 1 in the discretization step are known for the discretization of the obstacle problem using linear elements [10], [5].

Remark 2.1. In the present paper we shall use linear finite elements, and we shall avoid higher order finite elements for three reasons. First, it is well known that the maximum principle does not hold for higher order finite elements. In our underlying application, where the unknown is a concentration, it is absolutely necessary that it hold, since the electrical current, which we need to compute, depends on the logarithm of the concentration. The discrete maximum principle must therefore hold. Second, the precision obtained with the linear elements is, in general, largely sufficient for diffusion problems such as the one we consider. Third, our proof of convergence of the monotonic algorithm makes heavy use of the discrete maximum principle, and it is therefore not clear how the algorithm would behave in a setting where the maximum principle does not hold (in the case of higher order finite elements, or for the elasticity problem, for instance).

The monotonic algorithm is derived on a "strong formulation" of problem (4), which is easily shown to be equivalent to (4) as follows.

PROPOSITION 2.3. *Let $\tilde{u}$ be the unique solution to problem* (4) *and let $U = (u_1,\dots,u_N) \in \mathbb{R}^N$ be defined by $u_i = \tilde{u}(s_i)\ \forall i \in \{1,\dots,N\}$; then $\tilde{u}$ is a solution to* (4) *if and only if U is a solution to the following complementarity problem:*

$$\text{(5)} \qquad \begin{cases} u_i \leq \psi_i \quad \forall i \in \{1,\dots,N\}, \\ (AU)_i \leq F_i \quad \forall i \in \{1,\dots,N\}, \\ ((AU)_i - F_i)(\psi_i - u_i) = 0 \quad \forall i \in \{1,\dots,N\}, \end{cases}$$

with $\psi_i = \psi(s_i)$, $F_i = \int_\Omega f(x)\varphi_i(x)dx$, and A being the square matrix of order N whose coefficients satisfy $a_{i,j} = \int_\Omega \nabla\varphi_i(x)\cdot\nabla\varphi_j(x)dx$; therefore

$$(AU)_i = \sum_{j=1}^{N} u_j \int_\Omega \nabla\varphi_j(x)\cdot\nabla\varphi_i(x)dx.$$

Problem (5) is nonlinear. We shall solve it by an iterative algorithm that is adapted from a similar one used for multiphase flows in porous media [8]. Let us first remark that for $i \in \{1,\dots,N\}$ the last equation in (5) is equivalent to $(AU)_i = F_i$ or $u_i = \psi_i$. Therefore, there exist two disjoint subsets of $\{1,\dots,N\}$ such that $u_i = \psi_i$

and $(AU)_i \le F_i$ for any i in the first subset, and $(AU)_i = F_i$ and $u_i \le \psi_i$ for i in the second subset.

If we knew two disjoint subsets $\mathcal{J}$ and $\mathcal{I}$ of $\{1,\dots,N\}$ such that

$$\begin{aligned} u_i &\le \psi_i \quad \forall i \in \mathcal{J}, \\ (AU)_i &\le F_i \quad \forall i \in \mathcal{I}, \end{aligned}$$

then problem (5) would be solved by the solution of the following linear system:

$$\left\{ \begin{array}{l} u_i = 0 \quad \forall i \in \{1,\dots,N\} \quad \text{s.t. } s_i \in \partial\Omega \cap K_h, \\ (AU)_i = F_i \quad \forall i \in \mathcal{J}, \\ u_i = \psi_i \quad \forall i \in \mathcal{I}. \end{array} \right. \tag{6}$$

The algorithm which we propose here assumes the sets $\mathcal{J}$ and $\mathcal{I}$ to be known at each iteration, solves problem (6), and corrects the sets $\mathcal{J}$ and $\mathcal{I}$ by looking for the nodes where the corresponding constraints are violated. Let us write this algorithm as follows.

Monotonic algorithm, obstacle problem, finite element discretization.

- Initialization. Let $\mathcal{I}^{(0)}$ and $\mathcal{J}^{(0)}$ be such that

$$\mathcal{I}^{(0)} \subset \{1,\dots,N\} \text{ and } \mathcal{J}^{(0)} = \{1,\dots,N\} \setminus \mathcal{I}^{(0)}. \tag{7}$$

- Step (j), $j \ge 0$. For given sets $\mathcal{I}^{(j)}$ and $\mathcal{J}^{(j)} = \{1,\dots,N\} \setminus \mathcal{I}^{(j)}$, let $U^{(j)} = (u_1^{(j)},\dots,u_N^{(j)}) \in \mathbb{R}^N$ be the solution to the following set of equations:

$$\left\{ \begin{array}{l} (AU^{(j)})_i = F_i \quad \forall i \in \mathcal{J}^{(j)}, \\ u_i^{(j)} = \psi_i \quad \forall i \in \mathcal{I}^{(j)}, \end{array} \right. \tag{8}$$

  where $(AU^{(j)})_i = \sum_{k=1}^N u_k^{(j)} \int_\Omega \nabla\varphi_k(x) \cdot \nabla\varphi_i(x)dx$ and $F_i = \int_\Omega f(x)\varphi_i(x)dx$. Let $\mathcal{I}^{(j+1)}$ and $\mathcal{J}^{(j+1)}$ be defined by

$$\begin{array}{ll} \mathcal{I}^{(j,0)} = \{i \in \mathcal{I}^{(j)}; AU_i^{(j)} \le F_i\}, & \mathcal{I}^{(j,1)} = \mathcal{I}^{(j)} \setminus \mathcal{I}^{(j,0)}, \\ \mathcal{J}^{(j,0)} = \{i \in \mathcal{J}^{(j)}; u_i^{(j)} \le \psi_i\}, & \mathcal{J}^{(j,1)} = \mathcal{J}^{(j)} \setminus \mathcal{J}_j^{(0)}, \\ \mathcal{I}^{(j+1)} = \mathcal{I}^{(j,0)} \cup \mathcal{J}^{(j,1)}, & \mathcal{J}^{(j+1)} = \{1,\dots,N\} \setminus \mathcal{I}^{(j+1)}. \end{array} \tag{9}$$

- The algorithm stops if there exists a step n such that $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$.

Let us first remark that this algorithm is well defined.

Proposition 2.4. *Let $\Sigma = \{s_i, i = 1, N\}$ denote the set of nodes of a given triangulation of Ω, let $\mathcal{I}^{(j)} \subset \{1,\dots,N\}$ and $\mathcal{J}^{(j)} = \{1,\dots,N\} \setminus \mathcal{I}^{(j)}$; then problem (8) has a unique solution.*

Proof. The proof of this result follows immediately from the Lax–Milgram lemma by noting that under Assumptions 2.1 and with the notations of Definition 2.2 and Proposition 2.4, $U^{(j)} = (u_1^{(j)},\dots,u_N^{(j)}) \in \mathbb{R}^N$ is a solution to problem (8) if and only if $\tilde{u}^{(j)}(x) = \sum_{i=1}^N u_i^{(j)} \varphi_i(x)$ is a solution to the following variational problem:

$$\left\{ \begin{array}{l} \tilde{u}^{(j)} \in V_h \quad \text{s.t. } u_i^{(j)} = \psi_i \ \forall i \in \mathcal{I}^{(j)}, \\ \displaystyle\int_\Omega \nabla\tilde{u}^{(j)}(x) \cdot \nabla v(x)dx = \int_\Omega f(x)v(x)dx \quad \forall v \in V_h. \end{array} \right. \tag{10}$$

∎

Let us now show that the algorithm defined by (7)–(9) is monotonic.

LEMMA 2.5. *Under Assumption* 2.1 *and those of Definition* 2.2, *the sequence* $(U^{(j)})_{j\in\mathbb{N}}$ *constructed by the algorithm* (7)–(9), *where* $U^{(j)} = (u_1^{(j)}, \dots, u_N^{(j)})$, *satisfies*

$$u_i^{(j+1)} \leq u_i^{(j)} \quad \forall j \in \mathbb{N}, \forall i \in \{1, \dots, N\}. \tag{11}$$

Equivalently, the sequence of functions $(\tilde{u}^{(j)})_{j\in\mathbb{N}}$ *defined by* $\tilde{u}^{(j)}(x) = \sum_{i=1}^N u_i^{(j)}\varphi_i(x)$ *for all* $x \in \Omega$ *satisfies*

$$\tilde{u}^{(j+1)} \leq \tilde{u}^{(j)} \quad \forall j \in \mathbb{N}. \tag{12}$$

Proof. Let $j \in \mathbb{N}$ and $w_h = \tilde{u}^{(j)} - \tilde{u}^{(j+1)}$. Then

$$\int_\Omega |\nabla w_h^-(x)|^2 dx = -\sum_{i=1}^N w_i^- \int_\Omega \nabla w_h(x)\nabla\varphi_i(x)dx. \tag{13}$$

- If $i \in \mathcal{I}^{(j)} \cap \mathcal{I}^{(j+1)}$, one has $w_i = 0$, and therefore $\int_\Omega \nabla w_h(x)\cdot\nabla(w_i^-\varphi_i(x))dx = 0$.
- If $i \in \mathcal{J}^{(j)} \cap \mathcal{J}^{(j+1)}$, one has $\int_\Omega \nabla\tilde{u}^{(j)}(x) \cdot \nabla\varphi_i(x)dx = \int_\Omega \nabla\tilde{u}^{(j+1)}(x) \cdot \nabla\varphi_i(x)dx$, and therefore
$$\int_\Omega \nabla w_h(x) \cdot \nabla(w_i^-\varphi_i(x))dx = 0.$$
- If $i \in \mathcal{J}^{(j)} \cap \mathcal{I}^{(j+1)}$, one obtains $u_i^{(j)} > \psi_i$ and $u_i^{(j+1)} = \psi_i$, hence $w_i > 0$, and therefore
$$\int_\Omega \nabla w_h(x) \cdot \nabla(w_i^-\varphi_i(x))dx = 0.$$
- Finally if $i \in \mathcal{I}^{(j)} \cap \mathcal{J}^{(j+1)}$, then $(AU^{(j)})_i > F_i$ and $(AU^{(j+1)})_i = F_i$, so that
$$\int_\Omega \nabla w_h(x) \cdot \nabla(w_i^-\varphi_i(x))dx \geq 0.$$

These inequalities and (13) yield that $\int_\Omega |\nabla w_h^-(x)|^2 dx = 0$, and since $w_h^- \in H_0^1(\Omega)$, this implies that $w_h \geq 0$, which concludes the proof of the lemma. ∎

We may now turn to the convergence of the algorithm. We first state that if the sets $\mathcal{I}^{(j)}$ and $\mathcal{J}^{(j)}$ are left unchanged from one iteration to the next, then the algorithm has reached the unique solution to problem (5).

PROPOSITION 2.6. *Assume that the sequence of sets* $(\mathcal{I}^{(j)})_{j\in\mathbb{N}}$ *constructed by the algorithm* (7)–(9) *is such that there exists* $n \in \mathbb{N}$ *such that* $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$; *then the solution* $U^{(n)}$ *to* (8) *is the unique solution to problem* (5).

Proof. Under the assumptions of Proposition 2.6, let $\mathcal{I} = \mathcal{I}^{(n)}$, $\mathcal{J} = \mathcal{J}^{(n)}$; let $U^{(n)} = (u_1, \dots, u_n)$ be the solution to (8) with $j = n$. Since $\mathcal{J}^{(n)} = \mathcal{J}^{(n+1)}$, one has $u_i \leq \psi_i$ for any $i \in \mathcal{J}^{(n)}$. Furthermore, $u_i = \psi_i$ for any $i \in \mathcal{I}^{(n)}$, so that $u_i \leq \psi_i$ for any $i \in \{1, \dots, N\}$. In a similar way, one has that $(AU)_i \leq F_i$ for any $i \in \{1, \dots, N\}$, and from (8) one has that $((AU)_i - F_i)(u_i - \psi_i) = 0$ for any $i \in \{1, \dots, N\}$. ∎

Let us now show that the monotonic algorithm terminates in a finite number of iterations.

THEOREM 2.7. *Under Assumption* (2.1), *there exists* $n \in \mathbb{N}$ *such that the sequence* $(U^{(n)})_{n\in\mathbb{N}}$ *constructed by the algorithm* (7)–(9) *is such that* $U^{(n)}$ *is the exact solution to the discrete problem* (4) *for all* $j \geq n$. *Furthermore the integer* n *satisfies*

$$n \leq N + 1. \tag{14}$$

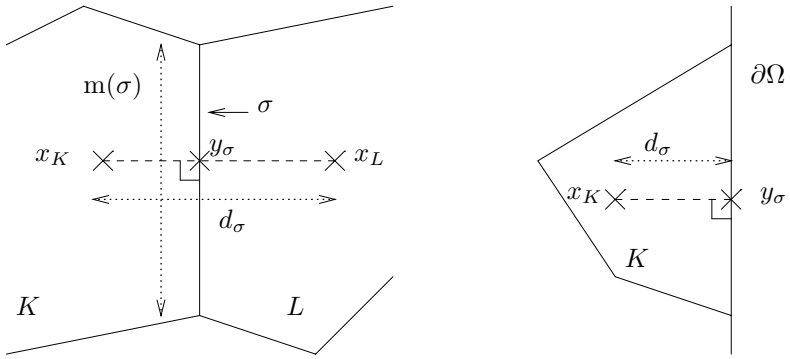


FIG. 1. *Admissible meshes.*

Proof. Let the sets $\mathcal{I}^{(j)}$ and $\mathcal{J}^{(j)}$ be defined by the algorithm (7)–(9) for any step (j); if there exists an integer n such that $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$, then, by Proposition 2.10, $U^{(n)}$ is the exact solution to the discrete problem (4), and the first part of the theorem is proven. It remains to prove that such a step exists and that it satisfies (14).

Let us first remark that for $i \in \{1, \dots, N\}$ if $u_i^{(0)} \leq \psi_i$, then $u_i^{(1)} \leq \psi_i$ by Lemma 2.5, and if $u_i^{(0)} > \psi_i$, then $u_i^{(1)} = \psi_i$ by (9) in the monotonic algorithm. Hence

$$u_i^{(1)} \leq \psi_i \text{ for any } i \in \{1, \dots, N\}. \tag{15}$$

Therefore, by an easy induction, one has that $\mathcal{I}^{(j)} = \mathcal{I}_0^{(j)}$ for any $j > 1$, which yields that $\mathcal{I}^{(j)} \subset \mathcal{I}^{(j+1)}$ for any $j > 1$. Since $\mathcal{I}^{(n)} \subset \{1, \dots, N\}$ is a finite set, this means that there exists an index n such that $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$.

Let us finally show that (14) holds true. Let n be the smallest integer such that $\mathcal{I}^{(n)} = \mathcal{I}^{(n+1)}$. Since $\mathcal{I}^{(j)}$ is strictly included in $\mathcal{I}^{(j+1)}$ for any $j > 1$, one has $N + 1 \geq \text{card}(\mathcal{I}^{(j+1)}) \geq \text{card}(\mathcal{I}^{(j+1)}) + 1$ for any $j < n$, which yields that $n \leq N + 1$. ∎

2.2. Approximation by the finite volume scheme. Let us now define a discretization mesh over Ω, which is assumed (following [9]) to be admissible for finite volumes in the following sense (see Figure 1).

DEFINITION 2.8 (admissible meshes). *Let Ω be an open bounded polygonal domain of $\mathbb{R}^d$. An admissible finite volume mesh of Ω, denoted by $\mathcal{T}$, is given by a family of "control volumes," which are disjoint polygonal convex subsets of Ω, a family of subsets of $\overline{\Omega}$ contained in hyperplanes of $\mathbb{R}^d$, denoted by $\mathcal{E}$ (these are the "sides" of the control volumes), with strictly positive one-dimensional measure, and a family of points of Ω, denoted by $\mathcal{P}$, satisfying the following properties (in fact, we shall denote, somewhat incorrectly, by $\mathcal{T}$ the family of control volumes):*

(i) *The closure of the union of all the control volumes is $\overline{\Omega}$.*
(ii) *For any $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ such that $\partial K = \overline{K} \setminus K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$.*
(iii) *For any $(K, L) \in \mathcal{T}^2$ with $K \neq L$, either the one-dimensional Lebesgue measure of $\overline{K} \cap \overline{L}$ is 0 or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which will then be denoted by $K|L$.*
(iv) *The family $\mathcal{P} = (x_K)_{K \in \mathcal{T}}$ is such that $x_K \in K$ (for all $K \in \mathcal{T}$) and, if $\sigma = K|L$, it is assumed that $x_K \neq x_L$, and the straight line $\mathcal{D}_{K,L}$ going through x_K and x_L is assumed to be orthogonal to $K|L$.*

In what follows, the following notations are used. Let $\mathrm{size}(\mathcal{T}) = \sup\{\mathrm{diam}(K), K \in \mathcal{T}\}$. *For any* $K \in \mathcal{T}$ *and* $\sigma \in \mathcal{E}$, $\mathrm{m}(K)$ *is the two-dimensional Lebesgue measure of* K, *and* $\mathrm{m}(\sigma)$ *the one-dimensional measure of* σ. *The set of interior (resp., boundary) edges is denoted by* $\mathcal{E}_{\mathrm{int}}$ *(resp.,* $\mathcal{E}_{\mathrm{ext}}$*), that is,* $\mathcal{E}_{\mathrm{int}} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ *(resp.,* $\mathcal{E}_{\mathrm{ext}} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$*). The set of neighbors of* K *is denoted by* $\mathcal{N}(K)$, *that is,* $\mathcal{N}(K) = \{L \in \mathcal{T};\ \exists \sigma \in \mathcal{E}_K\ \sigma = K \cap L\}$. *If* $\sigma = K|L$, *we denote by* d_σ *or* $d_{K|L}$ *the Euclidean distance between* x_K *and* x_L *(which is positive). If* $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}$, *let* d_σ *denote the Euclidean distance between* x_K *and* y_σ. *For any* $\sigma \in \mathcal{E}$, *the transmissivity through* σ *is defined by* $\tau_\sigma = \frac{\mathrm{m}(\sigma)}{d_\sigma}$ *if* $d_\sigma \neq 0$.

Remark 2.2. The condition $x_K \neq x_L$ if $\sigma = K|L$ is in fact quite easy to satisfy: two neighboring control volumes K, L, which do not satisfy it, just have to be collapsed into a new control volume M with $x_M = x_K = x_L$, and the edge $K|L$ removed from the set of edges. The new mesh thus obtained is admissible.

We refer to, e.g., [9] or [14] for examples of admissible meshes. These include rectangular meshes, Delaunay triangulations, and Voronoi meshes.

Let us now define a "discrete" functional space and a discrete H_0^1 norm.

DEFINITION 2.9. *Let* Ω *be an open bounded polygonal domain of* $\mathbb{R}^d$, *and* $\mathcal{T}$ *be an admissible mesh in the sense of Definition* 2.8.

Define $Y(\mathcal{T})$ *as the set of the functions defined a.e. from* Ω *to* $\mathbb{R}$ *which are constant over each control volume of the mesh. We shall denote by* u_K *the value taken by* u *on the control volume* K.

For $u \in Y(\mathcal{T})$, *define the discrete* H_0^1 *norm by*

$$\|u\|_{1,\mathcal{T}}^2 = \sum_{\sigma \in \mathcal{E}} \tau_\sigma (D_\sigma u)^2, \tag{16}$$

with

$$|D_\sigma u| = |u_K - u_L| \text{ if } \sigma \in \mathcal{E}_{\mathrm{int}}, \qquad \sigma = K/L, \tag{17}$$

$$D_\sigma u = -u_K \text{ if } \sigma \subset \partial\Omega. \tag{18}$$

Let $\mathcal{T}$ be an admissible finite volume mesh in the sense of Definition 2.8, let $\psi_K = \psi(x_K)$ and $f_K = \frac{1}{m(K)} \int_K f(x)dx$ for any $K \in \mathcal{T}$. A cell-centered finite volume discretization of problem (1) is written with respect to the discrete unknowns $(u_K)_{K \in \mathcal{T}}$ in the following way (see [19] for a description of how this scheme is obtained):

$$-\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} + m(K) f_K \geq 0 \quad \forall K \in \mathcal{T}, \tag{19}$$

$$\left(-\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} + m(K) f_K \right) (\psi_K - u_K) = 0 \quad \forall K \in \mathcal{T}, \tag{20}$$

$$u_K \leq \psi_K \quad \forall K \in \mathcal{T}, \tag{21}$$

$$F_{K,\sigma} = -\tau_\sigma (u_L - u_K) \quad \forall \sigma \in \mathcal{E}_{\mathrm{int}} \text{ if } \sigma = K/L, \tag{22}$$

$$F_{K,\sigma} = \tau_\sigma u_K \quad \forall \sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K. \tag{23}$$

The proof of the existence and uniqueness of the solution to this scheme was given in [19]. It follows for the following remark: let $(u_K)_{K \in \mathcal{T}} \in \mathbb{R}^{\mathrm{card}(\mathcal{T})}$, and let $u_{\mathcal{T}} \in Y(\mathcal{T})$ be defined by $u_{\mathcal{T}}(x) = u_K$ for $x \in K\ \forall K \in \mathcal{T}$. Then one may show

that $(u_K)_{K\in\mathcal{T}}$ is a solution to problem (19)–(23) if and only if $u_\mathcal{T}$ is a solution to the following problem:

$$\left\{\begin{array}{l} u_\mathcal{T} \in \mathcal{K}_\mathcal{T} = \{v \in Y(\mathcal{T}),\ \text{s.t. } v_K \le \psi_K\ \forall K \in \mathcal{T}\}, \\ A(u_\mathcal{T}, v - u_\mathcal{T}) \ge L(v - u_\mathcal{T}) \quad \forall v \in \mathcal{K}_\mathcal{T}, \end{array}\right. \tag{24}$$

where for any $u = (u_K)_{K\in\mathcal{T}}$ and $v = (v_K)_{K\in\mathcal{T}} \in Y(\mathcal{T})$,

$$A(u,v) = \sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} \tau_{K|L}(u_K - u_L)(v_K - v_L) + \sum_{\sigma\in\mathcal{E}_{\rm ext}\cap\mathcal{E}_K} \tau_\sigma u_K v_K, \tag{25}$$

and

$$L(v) = \sum_{K\in\mathcal{T}} m(K) f_K v_K. \tag{26}$$

Our goal here is to construct an algorithm yielding an approximate solution of problem (19)–(23). The iterative process which we described for the finite element discretization is easily adapted to the finite volume framework. Let $K \in \mathcal{T}$; then from (20) one has

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K \quad \text{or} \quad u_K = \psi_K.$$

Therefore, from (19) and (21), there exist two disjoint subsets of $\mathcal{T}$ such that on one subset one has

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K \text{ and } u_K \le \psi_K \quad \text{for } K \text{ in the first subset,}$$

and

$$u_K = \psi_K \text{ and } \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} \le m(K) f_K \quad \text{for } K \text{ in the second subset.}$$

Now assume that we knew two subsets $\mathcal{T}_f$ and $\mathcal{T}_\psi$ of $\mathcal{T}$ such that $\mathcal{T}_f \cup \mathcal{T}_\psi = \mathcal{T}$, $\mathcal{T}_f \cap \mathcal{T}_\psi = \emptyset$, and

$$u_K \le \psi_K \quad \forall K \in \mathcal{T}_f, \tag{27}$$

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} \le m(K) f_K \quad \forall K \in \mathcal{T}_\psi. \tag{28}$$

Then, as in the finite element case, the solution of problem (19)–(23) could be obtained by solving the linear problem

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K \quad \forall K \in \mathcal{T}_f, \tag{29}$$

$$u_K = \psi_K \quad \forall K \in \mathcal{T}_\psi, \tag{30}$$

where the numerical fluxes $F_{K,\sigma}$ are defined by (22)–(23). As in the finite element case, we shall solve (29)–(30) at each iteration and iterate on the sets $\mathcal{T}_f$ and $\mathcal{T}_\psi$ by looking at the constraints which are violated after the solution of (29)–(30).

The algorithm that follows determines $\mathcal{T}_f$ and $\mathcal{T}_\psi$ by an iterative method.

MONOTONIC ALGORITHM, OBSTACLE PROBLEM, FINITE VOLUME DISCRETIZATION.

• Initialization. Let $\mathcal{T}_f^{(0)}$ and $\mathcal{T}_\psi^{(0)}$ be such that

$$\mathcal{T}_f^{(0)} \cap \mathcal{T}_\psi^{(0)} = \emptyset \quad \text{and} \quad \mathcal{T}_f^{(0)} \cup \mathcal{T}_\psi^{(0)} = \mathcal{T} \tag{31}$$

(for example, $\mathcal{T}_f^{(0)} = \mathcal{T}$ and $\mathcal{T}_\psi^{(0)} = \emptyset$).

• Step (j). Assume the sets $\mathcal{T}_f^{(j)}$ and $\mathcal{T}_\psi^{(j)}$ to be known such that $\mathcal{T}_f^{(j)} \cap \mathcal{T}_\psi^{(j)} = \emptyset$ and $\mathcal{T}_f^{(j)} \cup \mathcal{T}_\psi^{(j)} = \mathcal{T}$. Let $(u_K^{(j)})_{K\in\mathcal{T}}$ be the solution to the following set of equations:

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^{(j)} = m(K) f_K \quad \forall K \in \mathcal{T}_f^{(j)}, \tag{32}$$

$$u_K^{(j)} = = \psi_K \quad \forall K \in \mathcal{T}_\psi^{(j)}, \tag{33}$$

$$F_{K,\sigma}^{(j)} = \tau_\sigma (u_K^{(j)} - u_L^{(j)}) \quad \forall \sigma \in \mathcal{E}_{\text{int}} \text{ if } \sigma = K/L, \tag{34}$$

$$F_{K,\sigma}^{(j)} = \tau_\sigma u_K^{(j)} \quad \forall \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K. \tag{35}$$

Let $\mathcal{T}_f^{(j+1)}$ and $\mathcal{T}_\psi^{(j+1)}$ be defined in the following way:

$$\begin{aligned}
&\mathcal{T}_f^{(j,0)} = \{K \in \mathcal{T}_f^{(j)}, ; u_K^{(j)} \le \psi_K\}, && \mathcal{T}_f^{(j,1)} = \mathcal{T}_f^{(j)} \setminus \mathcal{T}_f^{(j,0)},\\
&\mathcal{T}_\psi^{(j,0)} = \left\{ K \in \mathcal{T}_\psi^{(j)};\ \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^{(j)} \le m(K) f_K \right\}, && \mathcal{T}_\psi^{(j,1)} = \mathcal{T}_\psi^{(j)} \setminus \mathcal{T}_\psi^{(j,0)},\\
&\mathcal{T}_f^{(j+1)} = \mathcal{T}_f^{(j,0)} \cup \mathcal{T}_\psi^{(j,1)}, && \mathcal{T}_\psi^{(j+1)} = \mathcal{T} \setminus \mathcal{T}_f^{(j+1)}.
\end{aligned} \tag{36}$$

• The algorithm stops if there exists a step (J) such that $\mathcal{T}_f^{(J)} = \mathcal{T}_f^{(J+1)}$ and $\mathcal{T}_\psi^{(J)} = \mathcal{T}_\psi^{(J+1)}$.

The above algorithm is well defined thanks to the following result.

PROPOSITION 2.10. *Let $\mathcal{T}$ be an admissible finite volume mesh in the sense of Definition* 2.8, *and assume that the sets $\mathcal{T}_f^{(j)}$ and $\mathcal{T}_\psi^{(j)}$ such that $\mathcal{T}_f^{(j)} \cap \mathcal{T}_\psi^{(j)} = \emptyset$ and $\mathcal{T}_f^{(j)} \cup \mathcal{T}_\psi^{(j)} = \mathcal{T}$ are known; then problem* (32)–(35) *admits a unique solution.*

Proof. Under the assumptions of Proposition 2.10, one may find an equivalent "variational" formulation to problem (32)–(35). Let $u_\mathcal{T}^{(j)} \in Y(\mathcal{T})$ be defined by $u_\mathcal{T}^{(j)}(x) = u_K^{(j)}$ for $x \in K$, $\forall K \in \mathcal{T}$; it is easy to prove that $u_\mathcal{T}^{(j)}$ is a solution to problem (32)–(35) if and only if $u_\mathcal{T}^{(j)}$ is a solution to the following problem:

$$\begin{cases}
u_K^{(j)} = \psi_K \quad \forall K \in \mathcal{T}_\psi^{(j)},\\
A(u_\mathcal{T}^{(j)}, v) = L(v) \quad \forall v = (v_K)_{K\in\mathcal{T}} \in Y(\mathcal{T}),\\
\qquad \text{such that } v_K = 0 \quad \forall K \in \mathcal{T}_\psi^{(j)},
\end{cases} \tag{37}$$

with A and L defined by (25) and (26). The existence and uniqueness of the solution to (32)–(35) (and (37)) follow from the Lax–Milgram lemma. □

The algorithm (31)–(36) is therefore well defined; let us now show its monotonicity. This property is much related to the discrete maximum principle, which holds for finite volume discretizations of the Laplace equation; see, e.g., [18].

LEMMA 2.11 (monotonicity of the scheme). *Under Assumption* 2.1, *let* $\mathcal{T}$ *be an admissible finite volume mesh in the sense of Definition* 2.8; *the sequences* $(u_K^{(j)})_{j\in\mathbb{N},K\in\mathcal{T}}$ *which are constructed by the algorithm* (31)–(36) *satisfy*

$$u_K^{(j+1)} \le u_K^{(j)} \quad \text{for } j \in \mathbb{N} \text{ and } K \in \mathcal{T}.$$

Proof. Define $v_{\mathcal{T}} = u_{\mathcal{T}}^{(j)} - u_{\mathcal{T}}^{(j+1)}$, $F_{K,\sigma} = F_{K,\sigma}^{(j)} - F_{K,\sigma}^{(j+1)}$ $\forall K \in \mathcal{T}$, $\forall \sigma \in \mathcal{E}_K$, and $\min(v_{\mathcal{T}}) = \min\{v_K = u_K^{(j)} - u_K^{(j+1)},\ K \in \mathcal{T}\}$; let us show that $\min(v_{\mathcal{T}}) \ge 0$. Let $K_0 \in \mathcal{T}$ such that $\min(v_{\mathcal{T}}) = v_{K_0}$. Then we have the following:

- If $K_0 \in \mathcal{T}_\psi^{(j)} \cap \mathcal{T}_\psi^{(j+1)}$, then $v_{K_0} = 0$ so that $\min(v_{\mathcal{T}}) = 0$.
- Now if $K_0 \in \mathcal{T}_f^{(j)} \cap \mathcal{T}_\psi^{(j+1)}$, one has $u_{K_0}^{(j)} > \psi_{K_0}$ and $u_{K_0}^{(j+1)} > \psi_{K_0}$ so that $\min(v_{\mathcal{T}}) > 0$.
- Assume next that $K_0 \in \mathcal{T}_\psi^{(j)} \cap \mathcal{T}_f^{(j+1)}$; then

$$\sum_{\sigma\in\mathcal{E}_{K_0}} F_{K_0,\sigma}^{(j)} < m(K_0) f_{K_0} < 0 \quad \text{and} \quad \sum_{\sigma\in\mathcal{E}_{K_0}} F_{K_0,\sigma}^{(j+1)} = m(K_0) f_{K_0}.$$

  Therefore $\sum_{\sigma\in\mathcal{E}_{K_0}} F_{K_0,\sigma} > 0$, and, since $v_{K_0} \le v_K$ $\forall K \in \mathcal{T}$, one has $\sum_{\sigma\in\mathcal{E}_{K_0}} F_{K_0,\sigma} \le 0$, which is impossible.
- Let us finally assume that $K_0 \in \mathcal{T}_f^{(j)} \cap \mathcal{T}_f^{(j+1)}$; in this case one has

$$\sum_{\sigma\in\mathcal{E}_{K_0}} F_{K_0,\sigma} = 0. \tag{38}$$

  1. If the control volume K_0 lies near the boundary, that is, $\mathcal{E}_{K_0} \cap \mathcal{E}_{\text{ext}} \ne \emptyset$, then (38) becomes

$$\sum_{\substack{\sigma\in\mathcal{E}_{K_0}\cap\mathcal{E}_{\text{int}} \\ \sigma = K_0|K}} \frac{v_{K_0} - v_{K_\sigma}}{d_\sigma} + v_{K_0}\left(\sum_{\sigma\in\mathcal{E}_{K_0}\cap\mathcal{E}_{\text{ext}}} \frac{1}{d_{K_0,\sigma}}\right) = 0.$$

     Since $\min(v_{\mathcal{T}}) = v_{K_0}$, all the terms in the first sum are nonpositive, and therefore v_{K_0} must be nonnegative, which proves that $\min(v_{\mathcal{T}}) \ge 0$.
  2. Now if the control volume K_0 lies in the interior domain in the sense that $\mathcal{E}_{K_0} \subset \mathcal{E}_{\text{int}}$, then one needs to consider one of the two following subcases:
     (a) There exists a "path" of control volumes, which are all in $\mathcal{T}_f^{(j)} \cap \mathcal{T}_f^{(j+1)}$, leading from K_0 to the boundary; that is, there exists $m \in \mathbb{N}$ and $(K_\ell)_{\ell=0,\dots,m}$ such that $K_\ell \in \mathcal{T}_f^{(j)} \cap \mathcal{T}_f^{(j+1)}$, $\mathcal{E}_{K_\ell} \cap \mathcal{E}_{K_{\ell+1}} \ne \emptyset$ $\forall \ell = 0, \dots, m-1$. In this case, one has $v_{K_0} = v_{K_1} = \cdots = v_{K_m} = \min(v_{\mathcal{T}})$, and since K_m lies near the boundary, $\min(v_{\mathcal{T}}) \ge 0$.
     (b) If there does not exist such a path, then there exists some control volume K which does not belong to $\mathcal{T}_f^{(j)} \cap \mathcal{T}_f^{(j+1)}$ and such that $\min(v_{\mathcal{T}}) = v_K$; this case falls into one of the three cases which were previously analyzed, and for which we proved that $\min(v_{\mathcal{T}}) \ge 0$. ∎

We may now turn to the convergence of the algorithm. As for the finite element discretization, we first state that if the sets $\mathcal{T}_f^{(j)}$ and $\mathcal{T}_\psi^{(j)}$ are left unchanged from

one iteration to the next, then the algorithm has reached the unique solution to problem (5).

PROPOSITION 2.12. *Assume that the sequence of sets* $(\mathcal{T}_f^{(j)})_{j\in\mathbb{N}}$ *and* $(\mathcal{T}_\psi^{(j)})_{j\in\mathbb{N}}$, *which are constructed by the algorithm* (31)–(36), *are such that there exists a step* (J) *such that* $\mathcal{T}_f^{(J)} = \mathcal{T}_f^{(J+1)}$ *and* $\mathcal{T}_\psi^{(J)} = \mathcal{T}_\psi^{(J+1)}$; *then the solution* $(u_K^{(J)})_{K\in\mathcal{T}}$ *to* (32)–(35) *is the unique solution to problem* (19)–(23).

Proof. Let $\mathcal{T}_f = \mathcal{T}_f^{(J)}$, $\mathcal{T}_\psi = \mathcal{T}_\psi^{(J)}$, and $u_\mathcal{T} = u_\mathcal{T}^{(J)}$; hence $u_\mathcal{T}$ satisfies the set of equations (32)–(35). Since $\mathcal{T}_\psi^{(J)} = \mathcal{T}_\psi^{(J+1)}$, one has $-\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + m(K)f_K \geq 0$ $\forall K \in \mathcal{T}_\psi$, and, thanks to (32), one has $-\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + m(K)f_K = 0$ $\forall K \in \mathcal{T}_f$; since $\mathcal{T} = \mathcal{T}_\psi \cup \mathcal{T}_f$, $u_\mathcal{T}$ satisfies (19). Similarly, since $\mathcal{T}_f^{(J)} = \mathcal{T}_f^{(J+1)}$, one has $u_K \leq \psi_K$ $\forall K \in \mathcal{T}_f$, and, thanks to (33), $u_\mathcal{T}$ satisfies (21), and finally, since $\mathcal{T}_f \cup \mathcal{T}_\psi = \mathcal{T}$, $u_\mathcal{T}$ satisfies (20). Hence $u_\mathcal{T}$ is the unique solution to problem (19)–(23). ∎

THEOREM 2.13. *Under Assumption* 2.1, *there exists an integer* $J \in \mathbb{N}$ *such that the sequence* $(u_K^{(j)})_{j\in\mathbb{N}}$, *which is constructed by the algorithm* (31)–(36), *is such that* $(u_K^{(j)}, K \in \mathcal{T})$ *is the exact solution to the discrete problem* (24) *for all* $j \geq J$. *Furthermore the integer* J *satisfies*

$$J \leq \mathrm{card}(\mathcal{T}) + 1, \tag{39}$$

where $\mathrm{card}(\mathcal{T})$ *denotes the number of cells of the mesh.*

Proof. Let the sets $\mathcal{T}_\psi^{(j)}$ and $\mathcal{T}_f^{(j)}$ be defined by the algorithm (31)–(36) for any step (j); if there exists an integer J such that $\mathcal{T}_\psi^{(J)} = \mathcal{T}_\psi^{(J+1)}$, then by Proposition 2.12, $(u_K^{(J)}, K \in \mathcal{T})$ is the exact solution to the discrete problem (24), and the first part of the theorem is proven. There remains to prove that such a step exists and that it satisfies (39).

As in the case of the finite element discretization, let us first remark that for $K \in \mathcal{T}$, if $u_K^{(0)} \leq \psi_K$, then $u_K^{(1)} \leq \psi_i$ by Lemma 2.11, and if $u_K^{(0)} > \psi_K$, then $u_K^{(1)} = \psi_K$ by step (9) of the algorithm. Hence

$$u_K^{(1)} \leq \psi_K \text{ for any } K \in \mathcal{T}; \tag{40}$$

therefore by an easy induction one has that $\mathcal{T}_\psi^{(j)} \subset \mathcal{T}_\psi^{(j+1)}$ for any $j > 1$. Since $\mathcal{T}_\psi^{(j)} \subset \mathcal{T}$, this means that there exists an index J such that $\mathcal{T}_\psi^{(J)} = \mathcal{T}_\psi^{(J+1)}$.

The proof of (39) is identical to the case of the finite element discretization (see the proof of Theorem 2.7). ∎

3. The Signorini problem. Let us now consider the following diffusion problem:

$$-\Delta u(x) = f, \quad x \in \Omega, \tag{41}$$

$$u(x) = 0, \quad x \in \Gamma^1, \tag{42}$$

$$\nabla u(x)\cdot\mathbf{n} = 0, \quad x \in \Gamma^2, \tag{43}$$

with a Signorini condition on a part of the boundary,

$$\left.\begin{array}{rcl} u(x) & \geq & a, \\ \nabla u(x)\cdot\mathbf{n} & \geq & b, \\ (u(x)-a)(\nabla u(x).\mathbf{n}-b) & = & 0, \end{array}\right\} \quad x \in \Gamma_3, \tag{44}$$

where the following holds.

Assumption 3.1.

1. Ω is an open bounded polygonal subset of $\mathbb{R}^d$.
2. The boundary $\partial\Omega$ of Ω is composed of three nonempty, disjoint connected sets Γ^1, Γ^2, and Γ^3 such that $\overline{\Gamma^1} \cup \overline{\Gamma^2} \cup \overline{\Gamma^3} = \overline{\partial\Omega}$.
3. $f \in L^2(\Omega)$, $a \leq 0$, and $b \in \mathbb{R}$.
4. $\mathbf{n}$ is the unit normal vector to $\partial\Omega$ outward to the domain Ω.

Under some regularity assumptions, problem (41)–(44) is equivalent to the following variational problem (see, e.g., [17]):

$$\left\{ \begin{array}{l} u \in \mathcal{K} = \{v \in H^1(\Omega),\ v_{|\partial\Omega} \geq a \text{ a.e.}\}, \quad \text{satisfying} \\ \displaystyle\int_\Omega \nabla u(x) \cdot \nabla(v-u)(x)dx \geq \int_{\partial\Omega} b(\gamma(v)-\gamma(u))(s)ds \quad \forall v \in \mathcal{K}, \end{array} \right. \tag{45}$$

with $v_{\partial\Omega} = \gamma(v)_{\partial\Omega}$, where γ is the trace operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$. By Stampacchia's theorem, problem (45) has a unique solution.

The Signorini problem may be viewed as an obstacle problem in which the obstacle is located on the boundary. However, because the complementarity condition is written on the normal derivative on the boundary, one may not write the monotonic algorithm with piecewise linear finite elements in a straightforward way as in the case of the obstacle problem. Indeed, the normal derivative of the piecewise linear finite element approximate solution is defined on each edge of a triangle neighboring the boundary of the domain, but it is not defined at the nodes of the triangulation lying on the boundary. This problem could be solved by using higher order finite elements, but, as we already mentioned in Remark 2.1, a crucial issue in the underlying electrochemical application is that the maximum principle must hold, and this is not the case with higher order finite element methods. However, there is no such problem when using a finite volume discretization of the Signorini problem; the discrete normal derivative is well defined, and the maximum principle holds (see [19]). Hence the monotonic algorithm may be written quite easily.

We shall use here the same admissible finite volume meshes as for the discretization of the obstacle problem, which were defined in Definition 2.8, with the two following additional assumptions, which are needed because of the Signorini boundary conditions on the boundary:

(v) *For any $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega$, there exists $i \in \{1,2,3\}$ such that $\sigma \subset \Gamma^i$.*

(vi) *For any $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega$, let K be the control volume such that $\sigma \in \mathcal{E}_K$ and $\mathcal{D}_{K,\sigma}$ be the straight line going through x_K and orthogonal to σ; then $y_\sigma = \mathcal{D}_{K,\sigma} \cap \sigma$.*

Let us then define an appropriate "discrete" functional space.

DEFINITION 3.1 (discrete functional space). *Let Ω be an open bounded polygonal domain of $\mathbb{R}^d$, and $\mathcal{T}$ be an admissible mesh in the sense of Definition* 2.8. *Define $X(\mathcal{T})$ as the set of the functions defined a.e. from $\overline{\Omega}$ to $\mathbb{R}$ which are constant over each control volume of the mesh, and which are constant over each edge in $\mathcal{E}_3 = \mathcal{E}_{\text{ext}}$. We shall denote by u_K the value taken by u on the control volume K, and by u_σ the value taken by u on the edge $\sigma \in \mathcal{E}_{\text{ext}}$, $\sigma \subset \Gamma^3$.*

As in the case of the obstacle problem, a classical finite volume formulation is obtained by integrating the diffusion equation (41) over each control volume $\mathcal{T}$, using Green's formula and approximating the normal fluxes by a consistent difference quotient. Let us denote the discrete unknowns by $(u_K)_{K\in\mathcal{T}}$ for any $K \in \mathcal{T}$ and by $(u_\sigma)_{\sigma\subset\Gamma^3}$ for any $\sigma \in \mathcal{E}_{\text{ext}}$, and the "discrete flux" by $F_{K,\sigma}$, which is expected to

approximate the exact flux $-\int_\sigma \nabla u(s).\mathbf{n} ds$; the finite volume scheme can be written

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = 0 \quad \forall K \in \mathcal{T}, \tag{46}$$

with

$$F_{K,\sigma} = -\tau_\sigma(u_L - u_K) \quad \forall \sigma \in \mathcal{E}_{\rm int} \text{ if } \sigma = K/L, \tag{47}$$

$$F_{K,\sigma} = \tau_\sigma u_K \quad \forall \sigma \subset \Gamma^1,\ \sigma \in \mathcal{E}_K, \tag{48}$$

$$F_{K,\sigma} = 0 \quad \forall \sigma \subset \Gamma^2,\ \sigma \in \mathcal{E}_K, \tag{49}$$

$$F_{K,\sigma} = -\tau_\sigma(u_\sigma - u_K) \quad \forall \sigma \in \mathcal{E}_3,\ \sigma \in \mathcal{E}_K, \tag{50}$$

with the Signorini boundary condition

$$u_\sigma \geq a \quad \forall \sigma \in \mathcal{E}_3, \tag{51}$$

$$-F_{K,\sigma} \geq \mathrm{m}(\sigma)\, b \quad \forall \sigma \in \mathcal{E}_3, \tag{52}$$

$$(u_\sigma - a)\left(\frac{F_{K,\sigma}}{\mathrm{m}(\sigma)} + b\right) = 0 \quad \forall \sigma \in \mathcal{E}_3, \tag{53}$$

where $\mathcal{E}_3$ denotes the set of edges of the mesh that are included in Γ^3.

In [19], we prove the following existence result.

PROPOSITION 3.2. *Let $\mathcal{T}$ be an admissible mesh of Ω; problem* (46)–(53) *admits a unique solution $(u_K)_{K\in\mathcal{T}}, (u_\sigma)_{\sigma\in\mathcal{E}_3}$.*

We may therefore define the approximate solution $u_\mathcal{T}$ from a.e. in $\Omega \cup \Gamma^3$ to $\mathbb{R}$ by

$$u_\mathcal{T}(x) = u_K \text{ for } x \in K \text{ and } K \in \mathcal{T}, \qquad u_\mathcal{T}(x) = u_\sigma \text{ for } x \in \sigma \text{ and } \sigma \in \mathcal{E}_3. \tag{54}$$

Remark 3.1. Under regularity assumptions on the exact solution, we give in [19] an estimate of order 1 with respect to the mesh size for the "discrete" H^1 norm and L^2 norm of the error on the solution. If the exact solution is no longer assumed to be regular, the convergence of the discrete solution towards the exact solution may still be proven; see [19].

The monotonic algorithm is again based on the obvious remark that, for a given $\sigma \in \mathcal{E}_3$, (53) is equivalent to $u_\sigma = a$ or $-F_{K,\sigma} = \mathrm{m}(\sigma)\, b$. Hence there exist two disjoint subsets of $\mathcal{E}_3$ such that on one subset one has

$$u_\sigma = a \quad \text{and} \quad -F_{K,\sigma} \geq \mathrm{m}(\sigma)\, b,$$

and on the other one,

$$-F_{K,\sigma} = \mathrm{m}(\sigma)\, b \quad \text{and} \quad u_\sigma \geq a.$$

Now if the subsets $\mathcal{E}_a$ and $\mathcal{E}_b$ of $\mathcal{E}_3$ such that $\mathcal{E}_a \cup \mathcal{E}_b = \mathcal{E}_3$, $\mathcal{E}_a \cap \mathcal{E}_b = \emptyset$ and such that

$$-F_{K,\sigma} \geq \mathrm{m}(\sigma)\, b \quad \forall \sigma \in \mathcal{E}_a, \tag{55}$$

$$u_\sigma \geq a \quad \forall \sigma \in \mathcal{E}_b \tag{56}$$

were known, then the solution to problem (46)–(53) could be obtained by solving the linear problem (46)–(50), together with

$$u_\sigma = a \quad \forall \sigma \in \mathcal{E}_a, \tag{57}$$

$$-F_{K,\sigma} = \mathrm{m}(\sigma)\, b \quad \forall \sigma \in \mathcal{E}_b. \tag{58}$$

The algorithm which follows determines the sets $\mathcal{E}_a$ and $\mathcal{E}_b$ by an iterative method.

MONOTONIC ALGORITHM, SIGNORINI PROBLEM, FINITE VOLUME DISCRETIZATION.

• Initialization. Let $\mathcal{E}_a^{(0)}$ and $\mathcal{E}_b^{(0)} \subset \mathcal{E}_3$ be such that

$$\mathcal{E}_a^{(0)} \cap \mathcal{E}_b^{(0)} = \emptyset \quad \text{and} \quad \mathcal{E}_a^{(0)} \cup \mathcal{E}_b^{(0)} = \mathcal{E}_3. \tag{59}$$

• Step (j). Assume that the sets $\mathcal{E}_a^{(j)}$ and $\mathcal{E}_b^{(j)}$ are known such that $\mathcal{E}_a^{(j)} \cap \mathcal{E}_b^{(j)} = \emptyset$ and $\mathcal{E}_a^{(j)} \cup \mathcal{E}_b^{(j)} = \mathcal{E}_3$.

Let $u_{\mathcal{T}}^{(j)} \in X(\mathcal{T})$ be defined by $u_{\mathcal{T}}^{(j)}(x) = u_K^{(j)}$ for $x \in K$, $\forall K \in \mathcal{T}$, and by $u_{\mathcal{T}}^{(j)}(x) = u_\sigma^{(j)}$ for $x \in \sigma$, $\forall \sigma \in \mathcal{E}_3$, and let $u_{\mathcal{T}}^{(j)}$ be the solution to the set of equations (46)–(50) and

$$u_\sigma^{(j)} = a \quad \forall \sigma \in \mathcal{E}_a^{(j)}, \tag{60}$$

$$F_{K,\sigma}^{(j)} = -\mathrm{m}(\sigma)\, b \quad \forall \sigma \in \mathcal{E}_b^{(j)}. \tag{61}$$

Let $\mathcal{E}_a^{(j+1)}$ and $\mathcal{E}_b^{(j+1)}$ be defined in the following way:

$$\begin{array}{ll} \mathcal{E}_a^{(j,0)} = \{\sigma \in \mathcal{E}_a^{(j)}; -F_{K,\sigma}^{(j)} \geq \mathrm{m}(\sigma)\, b\}, & \mathcal{E}_a^{(j,1)} = \mathcal{E}_a^{(j)} \setminus \mathcal{E}_a^{(j,0)}, \\ \mathcal{E}_b^{(j,0)} = \{\sigma \in \mathcal{E}_b^{(j)}; u_\sigma^{(j)} \geq a\}, & \mathcal{E}_b^{(j,1)} = \mathcal{E}_b^{(j)} \setminus \mathcal{E}_b^{(j,0)}, \\ \mathcal{E}_a^{(j+1)} = \mathcal{E}_a^{(j,0)} \cup \mathcal{E}_b^{(j,1)}, & \mathcal{E}_b^{(j+1)} = \mathcal{E}_3 \setminus \mathcal{E}_b^{(j+1)}. \end{array} \tag{62}$$

• The algorithm stops if there exists a step (J) such that $\mathcal{E}_a^{(J)} = \mathcal{E}_a^{(J+1)}$ and $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$.

The above algorithm is well defined, thanks to the following result.

PROPOSITION 3.3. *Under Assumption* 3.1, *let* $\mathcal{T}$ *be an admissible finite volume mesh in the sense of Definition* 2.8; *then problem* (46)–(50), (60)–(61) *has a unique solution* $u_{\mathcal{T}}^{(j)}$.

Proof. Under Assumption 3.1, let $\mathcal{T}$ be an admissible finite volume mesh in the sense of Definition 2.8, and $u_{\mathcal{T}}^{(j)} \in X(\mathcal{T})$ be defined by $u_{\mathcal{T}}^{(j)}(x) = u_K^{(j)}$ for $x \in K$, $\forall K \in \mathcal{T}$, and by $u_{\mathcal{T}}^{(j)}(x) = u_\sigma^{(j)}$ for $x \in \sigma$, $\forall \sigma \in \mathcal{E}_3$, and let the sets $\mathcal{E}_a^{(j)}$ and $\mathcal{E}_b^{(j)}$ be such that $\mathcal{E}_a^{(j)} \cap \mathcal{E}_b^{(j)} = \emptyset$ and $\mathcal{E}_a^{(j)} \cup \mathcal{E}_b^{(j)} = \mathcal{E}_3$. It is easily seen that $u_{\mathcal{T}}^{(j)}$ is solution to problem (46)–(50), (60)–(61) if and only if $u_{\mathcal{T}}^{(j)}$ is a solution to the following problem:

$$\left\{ \begin{array}{l} u_{\mathcal{T}}^{(j)} \in \mathcal{K}_{\mathcal{T}}^{(j)} = \{v \in X(\mathcal{T}) \text{ s.t. } v_\sigma = a\ \forall \sigma \subset \mathcal{E}_a^{(j)}\} \text{ such that} \\ \mathcal{A}(u_{\mathcal{T}}^{(j)}, v) = \mathcal{L}^{(j)}(v) \quad \forall v \in X(\mathcal{T}) \text{ s.t. } v_\sigma = 0\ \forall \sigma \in \mathcal{E}_a^{(j)}, \end{array} \right. \tag{63}$$

with

$$\begin{array}{rl} \mathcal{A}(u,v) = & \displaystyle\sum_{\sigma = K|L \in \mathcal{E}_{\mathrm{int}}} \tau_{K|L}(u_K - u_L)(v_K - v_L) + \sum_{\sigma \in \mathcal{E}_K,\, \sigma \subset \Gamma^1} \tau_\sigma u_K v_K \\ & + \displaystyle\sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_3} \tau_\sigma (u_\sigma - u_K)(v_\sigma - v_K) \qquad \forall u,\ v \in X(\mathcal{T}), \end{array} \tag{64}$$

$$\mathcal{L}^{(j)}(v) = \sum_{\sigma \in \mathcal{E}_b^{(j)}} b v_\sigma \mathrm{m}(\sigma) \quad \forall v \in X(\mathcal{T}). \tag{65}$$

Then the existence and uniqueness of the solution to (46)–(50) follows by the Lax–Milgram lemma. ∎

Let us now turn to the monotonicity property of the algorithm.

LEMMA 3.4 (monotonicity). *Under Assumption* 3.1, *let* $\mathcal{T}$ *be an admissible finite volume mesh in the sense of Definition* 2.8; *the sequences* $(u_K^{(j)})_{j\in\mathbb{N}}$, $K \in \mathcal{T}$, *and* $(u_\sigma^{(j)})_{j\in\mathbb{N}}$, $\sigma \in \mathcal{E}_3$, *which are constructed by the algorithm* (59)–(62), *satisfy*

$$\begin{aligned} u_K^{(j)} &\le u_K^{(j+1)} \quad \forall j \in \mathbb{N} \text{ and for } K \in \mathcal{T}, \\ u_\sigma^{(j)} &\le u_\sigma^{(j+1)} \quad \forall j \in \mathbb{N} \text{ and for } \sigma \in \mathcal{E}_3. \end{aligned} \tag{66}$$

Proof. Let v be defined by $v = u_{\mathcal{T}}^{(j+1)} - u_{\mathcal{T}}^{(j)}$, and let $\min(v_{\mathcal{T}})$ be defined by

$$\min(v_{\mathcal{T}}) = \min\left\{ \min_{K\in\mathcal{T}} v_K, \min_{\sigma\in\mathcal{E}_3} v_\sigma \right\}.$$

We note that v satisfies the set of equations (46)–(50).

- Assume first that $\min(v_{\mathcal{T}}) = v_{K_0}$, with $K_0 \in \mathcal{T}$ such that $\partial K_0 \cap \Gamma^1 = \emptyset$ or $\partial K_0 \cap \Gamma^1$ is a point. Since $v_{K_0} \le v_K \ \forall K \in \mathcal{T}$ and $v_{K_0} \le v_\sigma \ \forall \sigma \in \mathcal{E}_3$, one has $\min(v_{\mathcal{T}}) = v_K \ \forall K \in \mathcal{T}$, and $\min(v_{\mathcal{T}}) = v_\sigma \ \forall \sigma \in \mathcal{E}_3$. Therefore, the minimum is reached on a control volume neighboring Γ^1, or on an edge included in Γ^3.
- Assume next that $\min(v_{\mathcal{T}}) = v_{K_0}$, with $K_0 \in \mathcal{T}$ such that there exists $\sigma \subset \partial K_0 \cap \Gamma^1$; from (46)–(50), we deduce that $\min(v_{\mathcal{T}}) \ge 0$.
- Now assuming $\sigma \in \mathcal{E}_b^{(j)} \cap \mathcal{E}_b^{(j+1)}$ and $\min(v_{\mathcal{T}}) = v_\sigma$, we obtain $-\tau_\sigma(v_\sigma - v_K) = 0$ with K such that $\partial K \cap \sigma = \sigma$; then $\min(v_{\mathcal{T}}) = v_K \ \forall K \in \mathcal{T}$ and $\min(v_{\mathcal{T}}) = v_\sigma \ \forall \sigma \in \mathcal{E}_3$.
- Next if $\sigma \in \mathcal{E}_b^{(j)} \cap \mathcal{E}_a^{(j+1)}$ and $\min(v_{\mathcal{T}}) = v_\sigma$, one has $u_\sigma^{(j)} < a$ and $u_\sigma^{(j+1)} = a$; hence $\min(v_{\mathcal{T}}) > 0$.
- Finally if $\sigma \in \mathcal{E}_a^{(j)} \cap \mathcal{E}_b^{(j+1)}$ and $\min(v_{\mathcal{T}}) = v_\sigma$, one has $-\tau_\sigma(u_\sigma^{(j)} - u_K^{(j)}) < \mathrm{m}(\sigma)b$ and $-\tau_\sigma(u_\sigma^{(j+1)} - u_K^{(j+1)}) = \mathrm{m}(\sigma)b$; therefore $-\tau_\sigma(v_\sigma - v_K) > 0$, which is in contradiction with $\min(v_{\mathcal{T}}) = v_\sigma$. ∎

We now turn to the convergence of the algorithm.

PROPOSITION 3.5. *Assume that there exists a step* (J) *such that* $\mathcal{E}_a^{(J)} = \mathcal{E}_a^{(J+1)}$ *and* $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$ *and let* $u_{\mathcal{T}}^{(J)}$ *be the solution to* (46)–(50), (60)–(61); *then* $u_{\mathcal{T}}^{(J)}$ *is the unique solution to problem* (46)–(53).

Proof. Let $\mathcal{E}_a = \mathcal{E}_a^{(J)}$, $\mathcal{E}_b = \mathcal{E}_b^{(J)}$, and $u_{\mathcal{T}} = u_{\mathcal{T}}^{(J)}$; hence $u_{\mathcal{T}}$ satisfies the set of equations (46)–(50) and

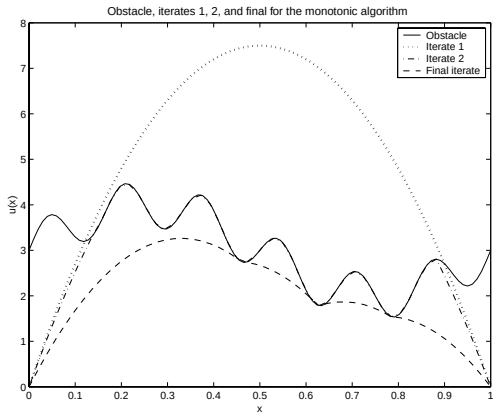
$$\begin{aligned} u_\sigma &= a \quad \forall \sigma \in \mathcal{E}_a, \\ F_{K,\sigma} &= -\mathrm{m}(\sigma)\, b \quad \forall \sigma \in \mathcal{E}_b. \end{aligned}$$

Since $\mathcal{E}_a^{(J)} = \mathcal{E}_a^{(J+1)}$, one has $F_{K,\sigma} \ge -\mathrm{m}(\sigma)\, b \ \forall \sigma \in \mathcal{E}_a$, and since $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$, one has $u_\sigma \ge a \ \forall \sigma \in \mathcal{E}_b$. Therefore, since $\mathcal{E}_a \cup \mathcal{E}_b = \mathcal{E}_3$, $u_{\mathcal{T}}$ satisfies the set of equations (46)–(53). ∎

THEOREM 3.6. *Under Assumption* 3.1, *there exists an integer* $J \in \mathbb{N}$ *such that the sequences* $(u_K^{(j)})_{j\in\mathbb{N}}$ $(u_\sigma^{(j)})_{j\in\mathbb{N}}$, *which are constructed by the algorithm* (31)–(36), *are such that* $(u_K^{(j)}, K \in \mathcal{T})$, $(u_\sigma^{(j)}, K \in \mathcal{E}_3)$ *is the exact solution to the discrete problem* (24) *for all* $j \ge J$. *Furthermore the integer* J *satisfies*

$$J \le \mathrm{card}(\mathcal{E}_3) + 1, \tag{67}$$

where $\mathrm{card}(\mathcal{E}_3)$ *denotes the number of edges of the mesh which are on the Signorini boundary* Γ_3.



N	ITER
20	6
50	7
100	8
500	31
1000	130

FIG. 2. *One-dimensional obstacle problem: the obstacle and some iterates (left), the number of discretization points N and number of iterations to convergence ITER (right).*

Proof. Let the sets $\mathcal{E}_a^{(j)}$ and $\mathcal{E}_b^{(j)}$ be defined by the algorithm (59)–(62) for any step (j); if there exists an integer J such that $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$, then by Proposition 2.12, $(u_K^{(J)}, K \in \mathcal{T})$ is the exact solution to the discrete problem (24), and the first part of the theorem is proven. There remains to prove that such a step exists and that it satisfies (67).

Similarly to the remark on the obstacle problem, let us first note that for $\sigma \in \mathcal{E}_3$ if $u_\sigma^{(0)} \geq a$, then $u_\sigma^{(1)} \geq a$ by Lemma 2.11, and if $u_\sigma^{(0)} < a$, then $u_\sigma^{(1)} = a$ by step (62) of the algorithm. Hence

$$u_\sigma^{(1)} \geq a \text{ for any } \sigma \in \mathcal{E}_3; \tag{68}$$

therefore by an easy induction one has that $\mathcal{E}_a^{(j)} \subset \mathcal{E}_a^{(j+1)}$ for any $j > 1$. Since $\mathcal{E}_a^{(j)} \subset \mathcal{E}_3$, this implies that there exists an index J such that $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$.

Let us now prove that (67) holds. Let J be the smallest integer such that $\mathcal{E}_b^{(J)} = \mathcal{E}_b^{(J+1)}$. Since $\mathcal{E}_b^{(J)} \subset \mathcal{E}_3 \ \forall j \in \mathbb{N}$, one has $\text{card}(\mathcal{E}_3) + 1 \geq \text{card}(\mathcal{E}_b^{(j+1)}) \geq \text{card}(\mathcal{E}_b^{(j)}) + 1$ for any $j > 1$, which yields that $J \leq \text{card}(\mathcal{E}_3) + 1$. □

3.1. Numerical tests. In order to test the efficiency of this new algorithm, some numerical experiments were performed. We first tested the algorithm on a one-dimensional obstacle problem, discretized by either the finite volume or the finite element method (in the one-dimensional case the two schemes differ by only the right-hand side and the boundary conditions). The results proved excellent. We show in Figure 2 a few iterations for $\psi(x) = 3 + \frac{1}{2}\sin(12\varphi_i x) + \sin(2\varphi_i x)$ and for a right-hand side equal to 1. We also give the number of iterations (ITER) required to convergence versus the number N of discretization points. Recall that we have a theoretical bound ITER $\leq$ N: the results show that this bound is far from optimal. Note also that we have taken the solution of the unconstrained problem (i.e., the solution of $-u'' = f$) as an initial guess; of course, one could decrease the number of iterations by taking a better-chosen initial guess.

A less academic study was performed by Herbin and Marchand [20] for a finite volume discretization of an electrochemical problem involving a Signorini boundary

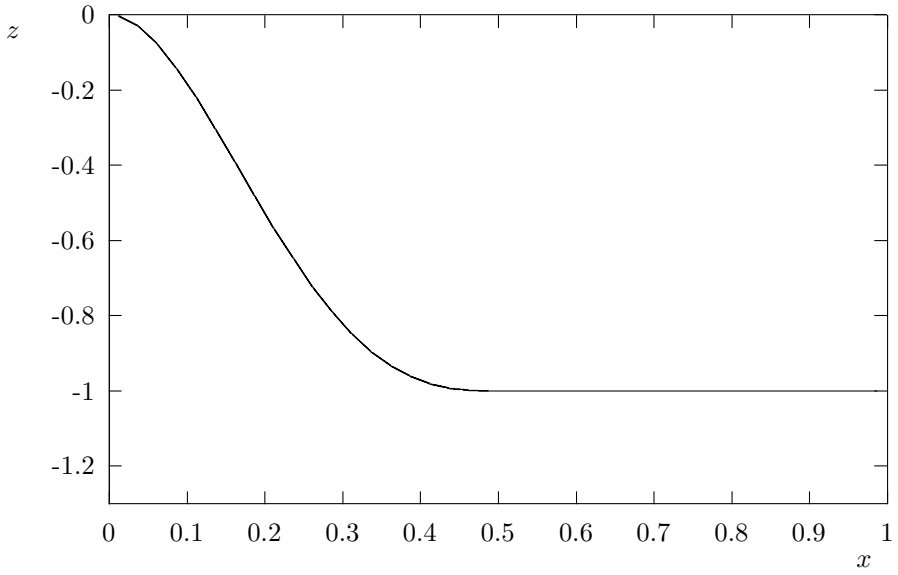


FIG. 3. $z = u(x,0)\ \forall x \in [0, x_m]$ *(on* Γ^3*: Signorini boundary).*

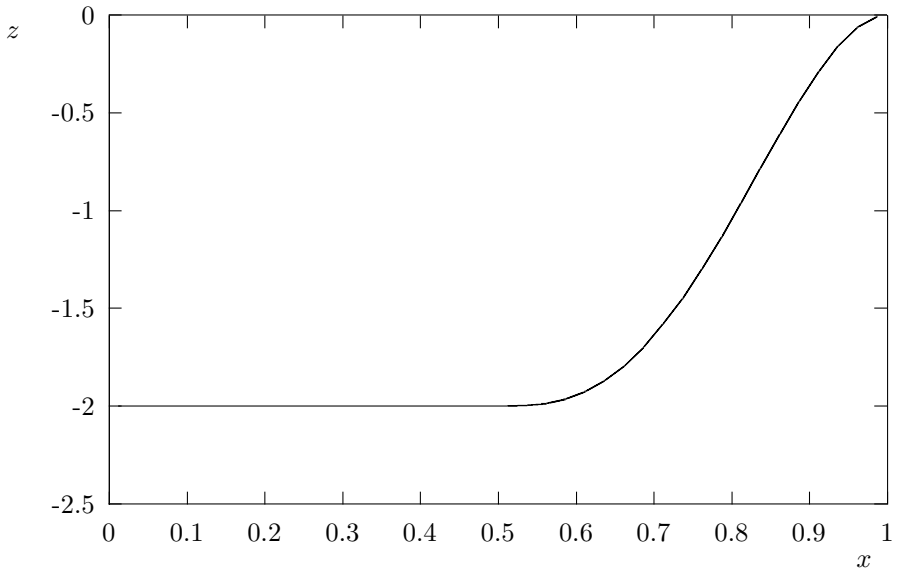


FIG. 4. $z = \nabla u \cdot \mathbf{n}(x,0)\ \forall x \in [0, x_m]$ *(on* Γ^3*: Signorini boundary).*

condition, which was introduced in [23], and which involved a two-dimensional problem. These results illustrate the performance of the monotonic algorithm quite well.

The domain Ω is taken to be the rectangle $(\Omega =]0, x_m[\ \times\ ]0, y_m[\,)$. We set the data such that the exact solution $u \in C^2(\overline{\Omega})$ is known. We show in Figures 3 and 4 the plots of the traces of u and of $\nabla u \cdot \mathbf{n}$ on the Signorini boundary Γ^3.

A rectangular mesh is used on the domain Ω. We vary the discretization step and the initial guess $\mathcal{E}_a^{(0)}$ and give the number of iterates required to converge to the (exact) solution of the discrete problem. Table 1 gives some results when taking the number of cells between 100 and 2500 with a uniform step.

TABLE 1
Number of iterations needed for the monotonicity algorithm.

$\mathcal{E}_3$ / grid size	10×10	20×20	30×30	40×40	50×50
$\mathcal{E}_{\text{ext}}$	4	6	7	7	8
$\emptyset$	4	6	7	7	7
$\{\sigma \subset \{x_m\} \times [0, x_m/2]\}$	4	7	7	7	8
$\{\sigma \subset \{x_m\} \times [x_m/2, x_m]\}$	2	3	2	2	2

These results show that the algorithm is quite efficient. The number of iterations does not vary much with respect to the grid size, and it is considerably less than the number of cells on $\mathcal{E}_3$, which was the upper bound given by Theorem 3.6. Of course, if one has a hint of how the Signorini boundary should be, then a good initial guess lowers the number of iterations, as may be seen from the last line of the table.

Let us also recall that at each iteration there is only one solve of a linear subproblem to be performed. Hence there are no ill-conditioned systems involved, such as those in penalty methods. Finally, let us point out that this algorithm may also be successfully implemented for other free boundary problems: we also tested it on the dam problem and it performs well. It is also used in multiphase problems [8], although in this last case no theoretical convergence result is known (in fact, existence and uniqueness of the solution are an open problem in this last case).

4. Conclusion. The monotonic algorithm which we have introduced for both the obstacle problem and the Signorini problems has been shown to be convergent for the linear finite element and finite volume discretizations in the case of the obstacle problem and the finite volume discretization in the case of the Signorini problem. Furthermore, a bound of the number of iterations is known. An important advantage of this algorithm is that, at each iteration, it necessitates only a linear solve involving a submatrix of the diffusion operator, and therefore no ill-conditioned system must be solved as would be the case with a penalty method. The actual implementation of the algorithm is very easy, and the computational cost is low, since the number of iterations is much lower than the theoretical bound, that is, the number of cells for which the constraint holds, even when the initial guess is not well chosen.

The limitation of the method is linked to the fact that it is proven to converge thanks to the discrete maximum principle, which holds for adequate discretizations of diffusion problems such as the ones we considered here. Note that the discretization is originally chosen such that the maximum principle holds, not because of the monotonic algorithm, but because the discrete maximum principle reflects a physical constraint which the approximate solution needs to satisfy (in the case of a chemical diffusion, the concentration should stay between 0 and 1). Hence it is natural in this type of problem to use the monotonic algorithm.

However, the efficiency (and proof of convergence) of the monotonic algorithm when the maximum principle does not hold (elasticity, higher order finite elements) is still an open question.

REFERENCES

[1] Z. BELHACHMI AND F. BEN BELGACEM, *Eléments finis d'ordre deux pour l'inéquation variationelle de Signorini*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 727–732.

[2] F. BEN BELGACEM, *Numerical simulation of some variational inequalities arisen from unilateral contact problems by the finite element methods*, SIAM J. Numer. Anal., 37 (2000), pp. 1198–1216.

[3] F. BEN BELGACEM, *Méthodes d'éléments finis pour les inéquations variationelles de contact unilatéral*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 811–816.

[4] H. BREZIS AND G. STAMPACCHIA, *Sur la régularité de la solution d'inéquations elliptiques*, Bull.

Soc. Math. France, 96 (1968), pp. 153–180.
[5] F. Brezzi, W. W. Hager, and P. A. Raviart, *Error estimates for the finite element solution of variational inequalities Part* I*—Primal theory*, Numer. Math., 28 (1977), pp. 431–443.
[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1998.
[7] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer–Verlag, Berlin, 1976.
[8] R. Eymard and T. Gallouët, *Traitement de changement de phase dans la modelisation de gisements pétroliers*, Journées Numériques de Besançon, J.M. Crolet and P. Lesaint, eds., Université de Besançon, Besançon, France, 1991.
[9] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handb. Numer. Anal. 7, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 713–1020.
[10] R. Falk, *Error estimates for the approximation of a class of variational inequalities*, Math. Comp., 28 (1974), pp. 963–997.
[11] R. Falk and B. Mercier, *Error estimates for elastoplastic problems*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 11 (1977), pp. 135–144.
[12] G. Fichera, *Problemi elastotatici con vincoli unilateral: Il problema di Signorini con ambigue condizioni al contorno*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Natur. Sez. Ia., 7 1963–1964, pp. 91–140.
[13] G. Fichera, *Boundary value problems of elasticity with unilateral constraints*, in Handbuch der Physik, Vol. VI a/2, Springer–Verlag, Berlin, 1972, pp. 391–424.
[14] T. Gallouët, R. Herbin, and M. H. Vignal, *Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions*, SIAM J. Numer. Anal., 37 (2000), pp. 1935–1972.
[15] R. Glowinski, *Lectures on Numerical Methods for Non-Linear Variational Problems*, notes by M.G. Vijayasundaram and M. Adimurthi, Tata Institute of Fundamental Research Lectures on Mathematics and Physics 65, Tata Institute of Fundamental Research, Bombay, India, Springer-Verlag, Berlin, New York, 1980.
[16] R. Glowinski, J. L. Lions, and R. Trémolières, *Analyse numérique des inequations variationnelles*, Dunod, Paris, 1976.
[17] S. Gerbi, R. Herbin, and E. Marchand, *Existence of a solution to a coupled elliptic system with a Signorini condition*, Adv. Differential Equations, 4 (1999), pp. 225–250.
[18] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.
[19] R. Herbin and E. Marchand, *Finite volume approximation of a class of variational inequalities*, IMA J. Numer. Anal., 21 (2001), pp. 553–585.
[20] R. Herbin and E. Marchand, *Numerical approximation of a nonlinear problem with a Signorini boundary condition*, in Iterative Methods in Scientific Computations, J. Wang, M. B. Allen, B. Chen, and T. Matthew, eds., IMACS Series in Computational and Applied Mathematics 4, IMACS, New Brunswick, NJ, 1998, pp. 283–288.
[21] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity, A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.
[22] N. Kikuchi and Y. J. Song, *Contact problems involving forces and movements for incompressible linearly elastic materials*, Internat. J. Engrg. Sci., 18 (1980), pp. 357–377.
[23] M. Kleitz, L. Dessemond, R. Jimenez, F. Petitbon, R. Herbin, and E. Marchand, *Micromodelling of the cathode and experimental approaches*, in Proceedings of the Second European Solid Oxide Fuel Cell Forum, Oslo, Norway, 1996, B. Thortensen, ed., Dr. Ulf Bossel, Oberrohrdorf, Switzerland, pp. 317–324.
[24] H. Lewy and G. Stampacchia, *On the regularity of the solution of a variational inequality*, Comm. Pure Appl. Math., 22 (1969), pp. 153–188.
[25] J. L. Lions and G. Stampacchia, *Variational inequalities*, Comm. Pure Appl. Math., 20 (1969), pp. 493–519.
[26] U. Mosco and G. Strang, *One sided approximation and variational inequalities*, Bull. Amer. Math. Soc., 80 (1974), pp. 308–312.
[27] L. Slemane, A. Bendali, and P. Laborde, *Mixed formulations for a class of variational inequalities*, C. R. Acad. Sci. Paris Sér. I Math., 334 (2002), pp. 87–92.
[28] G. Stampacchia, *Formes bilinéaire coercivitives sur les ensembles convexes*, C. R. Acad. Sci. Paris, 258 (1964), pp. 4413–4416.
[29] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.
[30] G. Stampacchia, *Equations Elliptiques du Second Ordre à coefficients discontinus*, les Presses de l'université de Montréal, Montréal, 1966.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2311–2317

HALF-RANGE GENERALIZED HERMITE POLYNOMIALS AND THE RELATED GAUSSIAN QUADRATURES*

JAMES S. BALL†

Abstract. A method is developed for calculating the recurrence coefficients for half-range generalized Hermite polynomials. These are orthogonal polynomials with measure $x^\gamma e^{-x^2}$ on the interval $(0,\infty)$. The recurrence coefficients can then be used to generate the weights and nodes of the related Gaussian quadratures. These quadratures allow efficient high accuracy evaluation of many Gaussian integrals encountered in probability functions, statistical mechanics, and quantum mechanics. The number of steps for an accurate numerical calculation of the recurrence coefficients is proportional to N, the number of coefficients obtained. Extended precision arithmetic is not needed in these calculations.

Key words. Gaussian integrals, Gaussian-type quadrature rules, orthogonal polynomials

AMS subject classifications. 41A55, 46N40, 65Q05, 68U01

PII. S0036142900370939

1. Introduction. Gaussian integrals over the range $(0,\infty)$ are encountered quite often in probability calculations when normal distributions are considered, in statistical mechanics where the kinetic energy of a particle is proportional to the velocity squared, and in quantum mechanical calculations involving harmonic oscillator wave functions. The purpose of this paper is to develop a Gaussian-type quadrature for integrals of the following form:

$$\int_0^\infty dx x^\gamma e^{-x^2} f(x) \tag{1.1}$$

for $\gamma > -1$. The special case $\gamma = 0$ was treated by several groups in 1969 [1], [2]. In 1981 Shizgal [3] generalized the treatment to nonzero γ, his interest being solutions to the Boltzmann equation for which the values of $\gamma = 1$ and 2 play a special role. His final methods made use of extended precision arithmetic. Further work by Clarke and Shizgal [4] investigated the instability of the method for determining the recurrence coefficients and proposed the use of asymptotic expansions. Most recently Gautschi [5] used the half-range Hermite weight function in testing his general methods for obtaining recurrence coefficients for orthogonal polynomials with arbitrary weight functions.

2. The calculation of the recurrence coefficients for generalized half-range Hermite polynomials. The orthogonality formula for monic generalized half-range Hermite polynomials $\phi_n^\gamma(x)$ is

$$\int_0^\infty dx x^\gamma e^{-x^2} \phi_n^\gamma(x)\phi_m^\gamma(x) = \delta_{nm} T_n. \tag{2.1}$$

Here the notation follows that of Gautschi [5] and his implementation of the Stieltjes procedure. The three term recurrence formula satisfied by these polynomials is as

*Received by the editors April 19, 2000; accepted for publication (in revised form) July 24, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/37093.html
†Physics Department, University of Utah, Salt Lake City, UT 84112 (ball@physics.utah.edu).

follows:

$$\phi^{\gamma}_{n+1}(x) = (x - \alpha_n)\phi^{\gamma}_n(x) - \beta_n \phi^{\gamma}_{n-1}(x). \tag{2.2}$$

Defining S_n as

$$S_n = \int_0^{\infty} dx x^{\gamma+1} e^{-x^2} \phi^{\gamma}_n(x)\phi^{\gamma}_n(x) \tag{2.3}$$

and applying the orthogonality relation to the recurrence formula, one obtains

$$\beta_n = T_n/T_{n-1} \tag{2.4}$$

and

$$\alpha_n = S_n/T_n. \tag{2.5}$$

Inserting the recurrence formula into the expression for T_n and using the orthogonality, one obtains

$$T_n = \int_0^{\infty} dx x^{\gamma+1} e^{-x^2} \phi^{\gamma}_n(x)\phi^{\gamma}_{n-1}(x). \tag{2.6}$$

Using the recurrence formula to replace $\phi^{\gamma}_n(x)$ results in the following identity:

$$\beta_n + \beta_{n-1} + \alpha^2_{n-1} = \frac{1}{T_{n-1}} \int_0^{\infty} dx x^{\gamma+2} e^{-x^2} (\phi^{\gamma}_{n-1}(x))^2. \tag{2.7}$$

Applying the same operations to S_n, one obtains a second identity:

$$\alpha_n + \alpha_{n-1} = \frac{1}{T_n} \int_0^{\infty} dx x^{\gamma+2} e^{-x^2} \phi^{\gamma}_n(x)\phi^{\gamma}_{n-1}(x). \tag{2.8}$$

Note that these are only algebraic manipulations, and they apply to any orthogonal polynomials.

The simplification that is possible for generalized half-range Hermite polynomials is that the integrals in (2.7), (2.8) can be integrated by parts. This fact was used by Shizgal [3], and although his procedure was somewhat different from the following discussion, it can be shown that the final results are in agreement. The by-parts integrals can be expressed in terms of α's and β's providing algebraic recurrence formulas for these quantities. This goes as follows:

$$\begin{aligned} \int_0^{\infty} dx x^{\gamma+2} e^{-x^2} (\phi^{\gamma}_{n-1}(x))^2 &= -\frac{1}{2} \int_0^{\infty} dx x^{\gamma+1} (\phi^{\gamma}_{n-1}(x))^2 \frac{d}{dx} e^{-x^2} \\ &= \frac{1}{2} \int_0^{\infty} dx e^{-x^2} \frac{d}{dx} [x^{\gamma+1} (\phi^{\gamma}_{n-1}(x))^2] = \frac{2n - 1 + \gamma}{2} T_{n-1}. \end{aligned} \tag{2.9}$$

This yields the following recurrence formula:

$$\beta_n + \beta_{n-1} + \alpha^2_{n-1} = \frac{2n - 1 + \gamma}{2}. \tag{2.10}$$

Integration by parts in (2.8) yields

$$T_n(\alpha_n + \alpha_{n-1}) = \frac{1}{2} \int_0^{\infty} dx x^{\gamma+1} e^{-x^2} \phi'^{\gamma}_n(x)\phi^{\gamma}_{n-1}(x). \tag{2.11}$$

After substituting the recurrence formulas for ϕ_{n-1} and ϕ_n' and using the orthogonality properties, one obtains a second recurrence formula:

$$\beta_n(\alpha_n + \alpha_{n-1}) = \frac{1}{2}\alpha_{n-1} + \beta_{n-1}(\alpha_{n-1} + \alpha_{n-2}). \tag{2.12}$$

Multiplying (2.12) by α_{n-1} and using (2.10) to replace α_{n-1}^2, we obtain

$$\begin{aligned}\alpha_n\alpha_{n-1}\beta_n = &\left[\left(\frac{n+\frac{\gamma}{2}}{2} - \beta_n\right) + \left(\frac{n-1+\frac{\gamma}{2}}{2} - \beta_{n-1}\right)\right] \\ &\cdot \left[\left(\frac{n+\frac{\gamma}{2}}{2} - \beta_n\right) - \left(\frac{n-1+\frac{\gamma}{2}}{2} - \beta_{n-1}\right)\right] + \alpha_{n-1}\alpha_{n-2}\beta_{n-1}.\end{aligned} \tag{2.13}$$

Multiplying out the factors in square brackets and rearranging the terms, we find

$$\alpha_n\alpha_{n-1}\beta_n - \left(\frac{n+\frac{\gamma}{2}}{2} - \beta_n\right)^2 = \alpha_{n-1}\alpha_{n-2}\beta_{n-1} - \left(\frac{n-1+\frac{\gamma}{2}}{2} - \beta_{n-1}\right)^2. \tag{2.14}$$

Evidently the quantities on either side of this equation are constant (i.e., independent of n). The $n = 0$ case determines this constant to be $-\gamma^2/16$. The final expression for the second recurrence formula is

$$\alpha_n\alpha_{n-1}\beta_n = \left(\frac{n+\frac{\gamma}{2}}{2} - \beta_n\right)^2 - \frac{\gamma^2}{16}. \tag{2.15}$$

Note that this equation involves only quantities at n and $n-1$.

The necessary starting values are given below:

$$\begin{aligned}T_0 &= \Gamma\left(\frac{\gamma+1}{2}\right)/2, \\ S_0 &= \Gamma\left(\frac{\gamma}{2}+1\right)/2.\end{aligned} \tag{2.16}$$

3. Evaluation of the recurrence formula. The recurrence procedure now appears straightforward. Given

$$\alpha_0 = \frac{\Gamma(\frac{\gamma}{2}+1)}{\Gamma(\frac{\gamma+1}{2})}, \tag{3.1}$$

one calculates β_1 from (2.10) by using the fact that $\beta_0 = 0$, and then one uses (2.15) to calculate α_1. This process is then repeated to obtain the desired values of α and β.

Unfortunately, as observed in earlier work, while the procedure is simple, the system is rather poorly conditioned, and in practice the error grows more than one order of magnitude per step. How this comes about and what methods can be used to overcome this problem can be seen by writing α_n and β_n in terms of a single set of functions g_n:

$$\alpha_n^2 = \frac{2n+\gamma+1}{3} - g_{n+1} - g_n \tag{3.2}$$

and

$$\beta_n = \frac{(n+\frac{\gamma}{2})}{6} + g_n. \tag{3.3}$$

In terms of these variables, (3.2) is identical to (2.10), and squaring (2.15) produces a nonlinear three term recurrence formula for the g's. The fact that n in all equations appears in the combination $n + \frac{\gamma}{2}$ suggests the following change of variable:

$$Y = Y(n) = 2n + \gamma. \tag{3.4}$$

The final recurrence formula is

$$\left(\frac{Y+1}{3} - g_{n+1} - g_n\right)\left(\frac{Y-1}{3} - g_n - g_{n-1}\right)\left(\frac{Y}{12} + g_n\right)^2 - \left[\left(\frac{Y}{6} - g_n\right)^2 - \frac{\gamma^2}{16}\right]^2 = 0. \tag{3.5}$$

This equation has the following properties: It is linear in g_{n+1} and g_{n-1}; the g_n^4 terms cancel out of this equation, and as a result of the choice relating β to g, the Y^4 terms also cancel out. The large n or Y behavior of (3.5) is determined by keeping only the cubic and quadratic terms in Y:

$$Y^3(14g_n - g_{n+1} - g_{n-1}) = Y^2\left(\frac{1}{3} - \frac{3\gamma^2}{2}\right). \tag{3.6}$$

The solution of interest for large n is that obtained by setting $g_{n+1} \simeq g_n \simeq g_{n-1}$, yielding

$$g_n \simeq \frac{2 - 9\gamma^2}{72Y}. \tag{3.7}$$

On the other hand, it is clear that if one calculates g_{n+1} or g_{n-1} from (3.6), the error in g_n increases by a factor of 14. Another way of saying this is in terms of the solution to the homogeneous version of (3.6) which has the following general solution:

$$g_n = A(7 + \sqrt{48})^n + B(7 + \sqrt{48})^{-n}. \tag{3.8}$$

Thus any error will excite these unwanted solutions, and both forward and backward recurrence will fail.

In previous work this stability problem was overcome with the brute force method of using arbitrary precision arithmetic and simply using as many significant figures as necessary so that the final result has the desired accuracy. For example, 75 significant figures are needed to calculate the first 50 values of α and β to 16 digits.

What is proposed here is a method that turns the factor 14, the source of the instability when one solves (3.6) for g_{n+1} or g_{n-1}, into a virtue. This can be done by solving (3.6) for g_n. What results is a fixed point method in which g_n is calculated in terms of g_{n+1} and g_{n-1}. The iterative solution will now converge rapidly, with the error now falling approximately one order of magnitude per iteration. In order to use this method one must have a set of starting values for $g_n, n = 0, 1, \ldots, N$, that are close enough to the solution to be convergent. Before proposing starting values and the procedure to follow, a couple of points should be made. The first involves the error introduced by the starting value of g_N. In the course of the fixed point iteration this value will stay fixed with whatever error it has. As a result, g_{N-1} will converge to a value with an error of 1/14, the fixed error in g_N. At each step down the ladder the error will decrease by 1/14. This error remains after the procedure has converged and is, therefore, a fixed permanent error. Thus one must adjust the value of N and the accuracy of the trial g_N so that accurate values are determined for the

range of interest. A second point is that if g_{n+1}, g_n, and g_{n-1} are calculated from the recurrence formula, that value of g_n will be unchanged by the iteration procedure and will appear to have converged. As a result one cannot simply apply this method to improve the accuracy of values obtained from the recurrence formula. Furthermore, if the first m trial values are obtained from the recurrence formula, m iterations are required before all of the first m terms have been modified. Since the first term is accurate, this doesn't produce a fixed error but does require a minimum number of iterations for accurate low n results.

Returning now to the actual recurrence formula, we first solve it for the linear term in g_n which will determine the new values from the iterative process. All other quantities are calculated from the old or input values of g_{n-1}, g_n, and g_{n+1}. This iterative process is again rapidly convergent, with behavior very similar to that of the large n case discussed above. This appears to be due to the fact that even the first few values of g are already quite small, making the quadratic and cubic terms which are ignored in (3.6) unimportant.

The remaining ingredient for this calculation is the starting values of the g_n's. Since the values of g_0 and g_1 are known functions which can be evaluated to near machine accuracy, the first 7 or 8 terms can be calculated from the recurrence formula with a loss of about 8 significant figures. For large n or Y, g_n can be expressed in terms of a Taylor series in $1/Y$. Although it has not been proven, it seems likely that this series has a zero radius of convergence and is therefore an asymptotic series. The first 4 terms of this series are

$$g_n(\gamma) = \frac{C_0}{Y} + \frac{C_1}{Y^3} + \frac{C_2}{Y^5} + \frac{C_3}{Y^7}, \tag{3.9}$$

where

$$C_0 = \frac{1}{36} - \frac{\gamma^2}{8}, \tag{3.10}$$

$$C_1 = \frac{23}{432} - \frac{11}{48}\gamma^2 + \frac{3}{32}\gamma^4, \tag{3.11}$$

$$C_2 = \frac{1189}{2592} - \frac{409}{192}\gamma^2 + \frac{75}{64}\gamma^4 + \frac{9}{64}\gamma^6, \tag{3.12}$$

and

$$C_3 = \frac{196057}{20736} - \frac{153559}{3456}\gamma^2 + \frac{7111}{256}\gamma^4 + \frac{639}{128}\gamma^6 + \frac{135}{512}\gamma^8. \tag{3.13}$$

Empirically, using these first 4 terms gives approximately 14 figure accuracy for β_{50}. The accuracy improves for larger n. For $n = 9$ this formula is good to 8 significant figures. Thus for a trial set of g's with minimum error one uses the recurrence formula for small n and switches to the asymptotic series at the point where the accuracy of both methods are about equal. For double precision this transition occurs near $n = 8$. The exact value is not particularly important because the trial input has small enough error to guarantee rapid convergence. For that matter, in the double precision calculation only three terms in the asymptotic expansion need be used if this is compensated for by increasing the number of g's involved in the iteration. Typical results in double precision are as follows: For accurate values of g_n up to $n = 50$, with g_1 and g_{51} fixed, the calculation required 10 iterations on the 49 intermediate g's for convergence, yielding 16 figure accuracy for the β's and α's. For quad precision it was necessary to increase the range to g_{66} to avoid the error caused by truncation.

The transition point could be shifted to a higher value of $n = 16$. The number of iterations required was 30 to produce 33 significant figures for the β's and α's. These results seem to apply to all values of γ, although only values $-1 < \gamma < 1$ were tested. This was considered the most interesting range because integrals with weight functions corresponding to $\gamma + 1$ and $\gamma + 2$ can be evaluated accurately by increasing the number of nodes in the quadrature for γ by one.

While this method is straightforward and has rapid linear convergence, we can in fact do much better by employing a method with quadratic convergence. This is provided by the well-known Newton's method, rather than by solving (3.5) for g_n and treating the resulting equation as a fixed point problem. We consider the problem of finding the N roots, g_n, of the N equations that (3.5) represents. This procedure goes as follows: Define

$$F^i = \left(\frac{Y(i)+1}{3} - g_{i+1} - g_i\right)\left(\frac{Y(i)-1}{3} - g_i - g_{i-1}\right)\left(\frac{Y(i)}{12} + g_i\right)^2 - \left[\left(\frac{Y(i)}{6} - g_i\right)^2 - \frac{\gamma^2}{16}\right]^2 \tag{3.14}$$

and

$$J_{i,j} = \partial_{g_j} F^i. \tag{3.15}$$

Here the g's are the input values. The new values are then given by

$$g_j^{new} = g_j^{old} - \sum_k J_{j,k}^{-1} F^k. \tag{3.16}$$

Note that J is a tridiagonal matrix because the recurrence formula involves only three terms. As a result all of the steps in this iteration are of $O(N)$. The results are rapidly convergent; only one iteration is required in double precision, although in practice a second is required to verify convergence. For quad precision two iterations are necessary, with one more to check convergence. It should be noted that this procedure does not have the problem with the use of the recurrence formula in the calculation of the starting values for small n that were encountered with the fixed point method.

Some of results of the actual calculation are as follows: In double precision the asymptotic expansion introduces no error for values of N greater than 50. For $N = 50$, the calculation on a Sun Ultra Sparc 2 with two processors requires approximately .0005 seconds. The results for larger values of N scale linearly with N. Applying this method to larger N is of no particular interest because (3.9) can be used directly to obtain the α and β for $N > 50$. The quad precision results required $N = 66$ to produce full machine accuracy for the first 50 values. The time required was much longer than that for double precision because quad arithmetic on this machine is implemented in software. The times were .053 seconds for 50 accurate terms, .088 seconds for 100 accurate terms, and .124 seconds for 150 terms. Note that the time grew less rapidly than the number of terms. This is because the number of extra terms necessary to eliminate the truncation error actually decreases with increasing N.

4. Conclusion. The Gaussian quadrature obtained from these coefficients behaves as expected. To check noninteger γ the test integrals used $\gamma = -\frac{1}{4}$. For $f(x) = \exp(-10x)$ the integral can be evaluated to a fractional error of 10^{-14} with a

25 node quadrature. The case $f(x) = \exp(10x)$ requires more nodes. For 40 nodes both integrals have a fractional error of approximately 10^{-14}. In the case of quad precision the case $f(x) = \exp(-10x)$ can be evaluated to 32 significant figures with 40 nodes. As in the double precision case more nodes were required for the accurate evaluation of $f(x) = \exp(10x)$. Obtaining 32 significant figures required 60 nodes.

The results obtained here can be compared to the tables published by Steen, Byrne, and Gelbard [1]. The tables are for the $\gamma = 0$ case only and are tables of the weights and nodes for half-range Hermite quadratures for 2 through 15 nodes. The results are quite surprising. For 8 or fewer nodes the results obtained here agree to 14 significant figures. For 10 nodes one finds 12 figure agreement. As the number of nodes increases, one loses 1 significant figure when the number of nodes is increased by 1. This is exactly the kind of behavior that might be expected from the recurrence calculation based on approximately 20 digit accuracy. For the largest number of nodes, 15, the nodes and weights obtained by Steen, Byrne, and Gelbard have an error in the seventh figure. On the other hand, if one uses these nodes and weights to perform an integral, the results are surprisingly good. Why this is true is not clear, but recall that only the large n polynomials have errors. Perhaps the set of polynomials generated corresponds to a weight in the orthogonality relation that is very close to $\exp(-x^2)$, and the fact that both the nodes and the weights are shifted conspires to produce good results for the integrals. This general kind of behavior was observed by Gautschi [6] in connection with a different Gaussian quadrature which, in spite of having series errors in the nodes, still produced good results for test integrals. See this work for a more detailed discussion of ways to test Gaussian quadratures.

For the special cases $\gamma = 0, 1, 2$ the nodes and weights for the Gaussian quadrature can be compared directly with those published by Shizgal [3]. The results obtained here agree to 15 figures with his $N = 16$ results given in Tables IIa, IIb, and IIc, with one exception, namely the last node in Table IIc, which has 8's as the ninth and tenth figures. This appears to be a misprint, and there should only be one 8. The fact that the location of this node has one more significant figure than the others also points to a misprint.

Finally, the quad precision results for the recurrence coefficients can be compared with the selected values for $\gamma = 0$ published by Gautschi [5] in his Table VIII. The results obtained here agree to all 25 figures published in his article.

It seems likely that the methods used here to calculate the recurrence coefficients can also be applied to other weight functions, such as those considered by Clark and Shizgal [4]. These possibilities are the subject of future work.

REFERENCES

[1] N. M. Steen, G. D. Byrne, and E. M. Gelbard, *Gaussian quadratures for the integrals* $\int_0^\infty dx e^{-x^2} f(x)$ *and* $\int_0^b dx e^{-x^2} f(x)$, Math. Comp., 23 (1969), pp. 661–671.

[2] D. Galant, *Gauss quadrature rule for the evaluation of* $2\pi^{-1/2} \int_0^\infty dx e^{-x^2} f(x)$, Math. Comp., 23 (1969), p. 674.

[3] B. Shizgal, *A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems*, J. Comput. Phys., 41 (1981), pp. 309–328.

[4] A. S. Clarke and B. Shizgal, *On the generation of orthogonal polynomials using asymptotic methods for recurrence coefficients*, J. Comput. Phys., 104 (1993), pp. 140–149.

[5] W. Gautschi, *Algorithm* 726: *ORTHPOL—A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules*, ACM Trans. Math. Software, 20 (1994), pp. 21–62.

[6] W. Gautschi, *How and how not to check Gaussian quadrature formulae*, BIT, 23 (1983), pp. 209–216.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2318–2336

CONVERGENCE THEORY OF RESTRICTED MULTIPLICATIVE SCHWARZ METHODS*

REINHARD NABBEN† AND DANIEL B. SZYLD‡

Abstract. Convergence results for the restricted multiplicative Schwarz (RMS) method, the multiplicative version of the restricted additive Schwarz (RAS) method for the solution of linear systems of the form $Ax = b$, are provided. An algebraic approach is used to prove convergence results for nonsymmetric M-matrices. Several comparison theorems are also established. These theorems compare the asymptotic rate of convergence with respect to the amount of overlap, the exactness of the subdomain solver, and the number of domains. Moreover, comparison theorems are given between the RMS and RAS methods as well as between the RMS and the classical multiplicative Schwarz method.

Key words. restricted Schwarz methods, domain decomposition, multisplittings, parallel algorithms

AMS subject classifications. 65F10, 65F35, 65M55

PII. S003614290138944X

1. Introduction.

We consider restricted Schwarz methods for the solution of linear systems of the form

$$Ax = b, \tag{1}$$

where A is $n \times n$ and nonsingular. These methods were introduced by Tai [26] and by Cai and Sarkis [10] for the parallel solution of (1); see also [8, 9]. In [10], it is shown by numerical examples that the restricted additive Schwarz (RAS) method is an efficient alternative to the classical additive Schwarz preconditioner. RAS preconditioners are widely used and are the default preconditioner in the PETSc software package [1]. In [26] and [10], the multiplicative variant of the RAS method, the restricted multiplicative Schwarz (RMS) method, is also mentioned; see also [9]. Although restricted Schwarz methods work very well in practice, until recently no theoretical results were available. In [16], convergence and comparison results for the RAS method were established when the matrix A in (1) is a (possibly nonsymmetric) M-matrix (or more generally an H-matrix). Those results use a new algebraic formulation of Schwarz methods and a connection with the well-known concept of multisplittings [7, 22]; see [2, 14, 15].

In this paper, we consider the RMS method. Again using the algebraic approach we are able to establish convergence results for the RMS method applied to M-matrices. Thus, this paper is the counterpart to [16] for the multiplicative case, although we prove some new results on RAS iterations as well. Furthermore, we are able to present a comparison result between the RMS and RAS methods. We show that, as measured in a certain norm, the convergence of the RMS method is never

*Received by the editors May 17, 2001; accepted for publication (in revised form) May 7, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/38944.html
†Fakultät für Mathematik, Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany (nabben@mathematik.uni-bielefeld.de).
‡Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA 19122-6094 (szyld@math.temple.edu). The research of this author was supported by National Science Foundation grant DMS-0207525.

worse than that of the RAS method. More precisely, if $M_{RAS}^{-1}A$ and $M_{RMS}^{-1}A$ are the preconditioned matrices using the RAS and RMS methods, respectively, we show that in some norm

$$\|I - M_{RMS}^{-1}A\| \leq \|I - M_{RAS}^{-1}A\|;$$

see section 3. In some cases, this comparison is valid for the spectral radii $\rho_1 := \rho(I - M_{RMS}^{-1}A) \leq \rho_2 := \rho(I - M_{RAS}^{-1}A)$. This implies that the spectrum of the preconditioned matrix with RAS $\sigma(M_{RAS}^{-1}A) \subseteq B(1, \rho_2)$, the ball centered at 1 with radius ρ_2, while the spectrum of the preconditioned matrix with RMS $\sigma(M_{RMS}^{-1}A)$ is contained in the smaller ball $B(1, \rho_1)$. These results remain true if we allow an inexact (or approximate) solution of the subdomain problems; see section 4. We point out that such a theoretical comparison has only recently become available between the classical additive and multiplicative Schwarz methods [20].

In section 3, we prove that the asymptotic rate of convergence of the RMS method is no faster than that of the classical multiplicative Schwarz method. The reason why the restricted Schwarz methods are attractive is that the communication time between processors is reduced, usually converging in less overall computational time [10].

We prove several other comparison theorems. We compare the speed of convergence with respect to the amount of overlap of the domains (section 5), the exactness of the subdomain solver (section 4), and the number of domains (section 6). Some variants of the RMS method are analyzed in section 7. We finish the paper with some comments on coarse grid corrections.

2. The algebraic representation and notations. As in [10, 16] we consider p nonoverlapping subspaces $W_{i,0}$, $i = 1, \ldots, p$, which are spanned by columns of the identity I over $\mathbb{R}^n$ and which are then augmented to produce overlap. For a precise definition, let $S = \{1, \ldots, n\}$, and let

$$S = \bigcup_{i=1}^{p} S_{i,0}$$

be a partition of S into p disjoint, nonempty subsets. For each of these sets $S_{i,0}$ we consider a nested sequence of larger sets $S_{i,\delta}$ with

$$S_{i,0} \subseteq S_{i,1} \subseteq S_{i,2} \subseteq \cdots \subseteq S = \{1, \ldots, n\}, \tag{2}$$

so that we again have $S = \cup_{i=1}^{p} S_{i,\delta}$ for all values of δ, but for $\delta > 0$ the sets $S_{i,\delta}$ are not pairwise disjoint; i.e., there is *overlap*. A common way to obtain the sets $S_{i,\delta}$ is to add those indices to $S_{i,0}$ which correspond to nodes lying at distance δ or less from those nodes corresponding to $S_{i,0}$ in the (undirected) graph of A. This approach is particularly adequate in discretizations of partial differential equations where the indices correspond to the nodes of the discretization mesh; see [6, 8, 9, 10, 13, 25].

Let $n_{i,\delta} = |S_{i,\delta}|$ denote the cardinality of the set $S_{i,\delta}$. For each nested sequence of the form (2) we can find a permutation π_i on $\{1, \ldots, n\}$ with the property that for all $\delta \geq 0$ we have $\pi_i(S_{i,\delta}) = \{1, \ldots, n_{i,\delta}\}$.

We now build matrices $R_{i,\delta} \in \mathbb{R}^{n_{i,\delta} \times n}$ whose rows are precisely those rows j of the identity for which $j \in S_{i,\delta}$. Formally, such a matrix $R_{i,\delta}$ can be expressed as

$$R_{i,\delta} = [I_{i,\delta}|O]\,\pi_i \tag{3}$$

with $I_{i,\delta}$ the identity on $\mathbb{R}^{n_{i,\delta}}$. Finally, we define the weighting (or masking) matrices

$$E_{i,\delta} = R_{i,\delta}^T R_{i,\delta} \ = \pi_i^T \begin{bmatrix} I_{i,\delta} & O \\ O & O \end{bmatrix} \pi_i \quad \in \mathbb{R}^{n\times n} \tag{4}$$

and the subspaces

$$W_{i,\delta} = \text{range}(E_{i,\delta}), \qquad i = 1, \ldots, p.$$

Note the inclusion $W_{i,\delta} \supseteq W_{i,\delta'}$ for $\delta \geq \delta'$ and, in particular, $W_{i,\delta} \supseteq W_{i,0}$ for all $\delta \geq 0$.

We view the matrices $R_{i,\delta}$ as restriction operators and $R_{i,\delta}^T$ as prolongations. We can identify the image of $R_{i,\delta}^T$ with the subspace $W_{i,\delta}$. For each subspace $W_{i,\delta}$ we define a restriction of the operator A on $W_{i,\delta}$ as

$$A_{i,\delta} = R_{i,\delta} A R_{i,\delta}^T.$$

To describe and analyze the classical Schwarz methods, the theory of orthogonal projections plays an important role; see, e.g., [17, Chap. 11], [25], and especially [5]. Therefore let

$$P_{i,\delta} = R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A, \tag{5}$$

provided that $A_{i,\delta}$ is nonsingular. It is not hard to see that this is a projection onto the subspace $W_{i,\delta}$. (In the case of symmetric A, this projection is orthogonal.) The additive Schwarz preconditioner is

$$M_{AS,\delta}^{-1} = \sum_{i=1}^{p} R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} \tag{6}$$

and the preconditioned matrix is

$$M_{AS,\delta}^{-1} A = \sum_{i=1}^{p} P_{i,\delta}.$$

Similarly, the multiplicative Schwarz preconditioner $M_{MS,\delta}^{-1}$ is such that

$$T_{MS,\delta} = I - M_{MS,\delta}^{-1} A = (I - P_{p,\delta})(I - P_{p-1,\delta}) \cdots (I - P_{1,\delta}) = \prod_{i=p}^{1} (I - P_{i,\delta}). \tag{7}$$

Next we describe the *restricted* additive and multiplicative Schwarz preconditioners. We introduce "restricted" operators

$$\tilde{R}_{i,\delta} = R_{i,\delta} E_{i,0} \in \mathbb{R}^{n_{i,\delta}\times n}.$$

The image of $\tilde{R}_{i,\delta}^T = E_{i,0} R_{i,\delta}^T$ can be identified with $W_{i,0}$, so $\tilde{R}_{i,\delta}^T$ "restricts" $R_{i,\delta}^T$ in the sense that the image of the latter, $W_{i,\delta}$, is restricted to its subspace $W_{i,0}$, the space from the nonoverlapping decomposition. In the restricted (additive and multiplicative) Schwarz methods from [8, 10] the prolongation operator $R_{i,\delta}^T$ is replaced by $\tilde{R}_{i,\delta}^T$ and the (oblique) projection

$$Q_{i,\delta} = \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A = E_{i,0} P_{i,\delta}$$

is used; see [16]. Thus, the restricted counterparts to the operators (6) and (7) are

$$M_{RAS,\delta}^{-1} = \sum_{i=1}^{p} \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta}$$

and

$$T_{RMS,\delta} = (I - Q_{p,\delta})(I - Q_{p-1,\delta}) \cdots (I - Q_{1,\delta}) = \prod_{i=p}^{1} (I - Q_{i,\delta}), \tag{8}$$

respectively. The iteration matrix of the RAS method is then

$$T_{RAS,\delta} = I - M_{RAS,\delta}^{-1} A = I - \sum_{i=1}^{p} \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A = I - \sum_{i=1}^{p} Q_{i,\delta}. \tag{9}$$

For practical parallel implementations, replacing $R_{i,\delta}^T$ by $\tilde{R}_{i,\delta}^T$ means that the corresponding part of the computation does not require any communication, since the images of the $\tilde{R}_{i,\delta}^T$ do not overlap. In addition, the numerical results in [10] indicate that the RAS method is faster (in terms of number of iterations and/or CPU time) than the classical one.

For the analysis of preconditioned Krylov subspace methods, the relevant matrices are $M_{AS,\delta}^{-1} A$ and $M_{RAS,\delta}^{-1} A$ for additive Schwarz and $I - T_{MS,\delta}$ and $I - T_{RMS,\delta}$ for multiplicative Schwarz. Alternatively, we can consider and compare the iteration matrices $T_{AS,\delta} = I - M_{AS,\delta}^{-1} A$, $T_{RAS,\delta}$ and $T_{MS,\delta}$, $T_{RMS,\delta}$. These correspond to stationary iterative methods, e.g., of the form

$$x^{k+1} = T_{RAS,\delta} x^k + M_{RAS,\delta}^{-1} b, \qquad k = 0, 1, \ldots,$$

for the RAS case; see, e.g., [18] for another example of such Schwarz iterations.

As in [2, 15, 16], the key to our analysis is a new (algebraic) representation of the restricted Schwarz methods. We construct a set of matrices $M_{i,\delta}$ associated with $R_{i,\delta}$ as follows:

$$M_{i,\delta} = \pi_i^T \begin{bmatrix} A_{i,\delta} & O \\ O & D_{\neg i,\delta} \end{bmatrix} \pi_i \tag{10}$$

and $D_{\neg i,\delta}$ is the diagonal part of the principal submatrix of A "complementary" to $A_{i,\delta}$; i.e.,

$$D_{\neg i,\delta} = \operatorname{diag}\left([O | I_{\neg i,\delta}] \cdot \pi_i \cdot A \cdot \pi_i^T \cdot [O | I_{\neg i,\delta}]^T \right)$$

with $I_{\neg i,\delta}$ the identity on $\mathbb{R}^{n - n_{i,\delta}}$. Here, we assume that $A_{i,\delta}$ and $D_{\neg i,\delta}$ are nonsingular. It can be shown (see [16]) that

$$\tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} = E_{i,0} M_{i,\delta}^{-1}, \qquad i = 1, \ldots, p,$$

and therefore

$$Q_{i,\delta} = \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A = E_{i,0} M_{i,\delta}^{-1} A, \qquad i = 1, \ldots, p.$$

With these fundamental identities the RAS and RMS methods can be described by the iteration matrices

$$T_{RAS,\delta} = I - \sum_{i=1}^{p} E_{i,0} M_{i,\delta}^{-1} A,$$

$$T_{RMS,\delta} = \prod_{i=p}^{1} (I - E_{i,0} M_{i,\delta}^{-1} A). \tag{11}$$

Moreover, we have

$$M_{RAS,\delta}^{-1} = \sum_{i=1}^{p} E_{i,0} M_{i,\delta}^{-1}.$$

In the rest of this section, we list some basic terminology and some well-known results which we use in the rest of the paper.

The natural partial ordering $\leq$ between matrices $A = (a_{ij})$, $B = (b_{ij})$ of the same size is defined componentwise; i.e., $A \leq B$ iff $a_{ij} \leq b_{ij}$ for all i, j. If $A \geq O$, we call A nonnegative. If all entries of A are positive, we say that A is positive and write $A > O$. This notation and terminology carries over to vectors as well.

A nonsingular matrix $A \in \mathbb{R}^{n\times n}$ is called monotone if $A^{-1} \geq O$. A monotone matrix $A \in \mathbb{R}^{n\times n}$ is called a (nonsingular) M-matrix if it has nonpositive off-diagonal elements. The following lemma states some useful properties of M-matrices; see, e.g., [4, 28].

LEMMA 2.1. *Let $A, B \in \mathbb{R}^{n\times n}$ be two nonsingular M-matrices with $A \leq B$. Then we have the following:*

(i) *Every principal submatrix of A or B is again an M-matrix.*
(ii) *Every matrix D such that $A \leq D \leq B$ is an M-matrix. In particular, if $A \leq D \leq \mathrm{diag}(A)$, then D is an M-matrix.*
(iii) *$B^{-1} \leq A^{-1}$.*

Our convergence results are formulated in terms of nonnegative splittings according to the following definition.

DEFINITION 2.2. *Consider the splitting $A = M - N \in \mathbb{R}^{n\times n}$ with M nonsingular. This splitting is said to be*

(i) regular *if $M^{-1} \geq O$ and $N \geq O$,*
(ii) weak nonnegative of the first type *(also called* weak regular*) if $M^{-1} \geq O$ and $M^{-1}N \geq O$,*
(iii) weak nonnegative of the second type *if $M^{-1} \geq O$ and $NM^{-1} \geq O$, and*
(iv) nonnegative *if $M^{-1} \geq O$, $M^{-1}N \geq O$, and $NM^{-1} \geq O$.*

Note that all the above splittings $A = M - N$ are *convergent* splittings for M-matrices A; i.e., the spectral radius $\rho(M^{-1}N)$ of the iteration matrix $M^{-1}N$ is less than one. Given an iteration matrix, there is a unique splitting for it, which is stated by the following result; see [3].

LEMMA 2.3. *Let A and T be square matrices such that A and $I - T$ are nonsingular. Then there exists a unique pair of matrices B, C such that B is nonsingular, $T = B^{-1}C$, and $A = B - C$. The matrices are $B = A(I - T)^{-1}$ and $C = B - A = A((I - T)^{-1} - I)$.*

For a positive vector w, we denote $\|x\|_w$ the weighted max norm in $\mathbb{R}^n$ given by

$$\|x\|_w = \max_{i=1,\ldots,n} |x_i|/w_i.$$

The resulting operator norm in $\mathbb{R}^{n\times n}$ is denoted similarly, and for $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ we have (see, e.g., [24])

$$\|B\|_w = \max_{i=1,\ldots,n} \left(\sum_{j=1}^{n} |b_{ij}| w_j \right) / w_i. \tag{12}$$

The following lemma follows directly from (12).

LEMMA 2.4. *Let $T, \tilde{T}$ be nonnegative matrices. Assume that $Tw \le \tilde{T}w$ for some vector $w > 0$. Then $\|T\|_w \le \|\tilde{T}\|_w$.*

3. Convergence and comparisons of RMS. In this section, we show that for a monotone matrix A the restricted multiplicative Schwarz iteration is convergent. Moreover, we establish that the spectral radius of the RMS iteration matrix is less than or equal to the spectral radius of the RAS iteration matrix, and it is no smaller than the spectral radius of the classical multiplicative Schwarz method (Theorems 3.5 and 3.8).

We begin by stating a lemma proved in [2].

LEMMA 3.1. *Let A be monotone, and let a collection of p triples (E_i, M_i, N_i) be given such that $O \le E_i \le I$, $\sum_{i=1}^{p} E_i \ge I$, and $A = M_i - N_i$ is a weak regular splitting for $i = 1, \ldots, p$. Let*

$$T = (I - E_p M_p^{-1} A)(I - E_{p-1} M_{p-1}^{-1} A) \cdots (I - E_1 M_1^{-1} A).$$

Then T is nonnegative and, for any vector $w = A^{-1}e > 0$ with $e > 0$, $\rho(T) \le \|T\|_w < 1$.

Now we formulate one of the main results of this section. It is the counterpart to Theorem 4.4 [16], where it was shown that the RAS method is convergent, and the iteration matrix (9) induces a weak regular splitting.

THEOREM 3.2. *Let A be a nonsingular M-matrix. Then for each value of $\delta \ge 0$ and for any $w = A^{-1}e > 0$ with $e > 0$, we have $\rho(T_{RMS,\delta}) \le \|T_{RMS,\delta}\|_w < 1$. Furthermore, there exists a unique splitting $A = M_{RMS,\delta} - N_{RMS,\delta}$ such that $T_{RMS,\delta} = M_{RMS,\delta}^{-1} N_{RMS,\delta}$, and this splitting is weak regular (i.e., weak nonnegative of the first type). The matrix $M_{RMS,\delta}$ is given by $M_{RMS,\delta} = A(I - T_{RMS,\delta})^{-1}$.*

Proof. The proof we present is almost the same as the proof of the convergence of the classical multiplicative Schwarz method given in [2]. Let $E_{i,0}$ as in (4) and $M_{i,\delta}$ as in (10). Observe that $O \le E_{i,0} \le I$, $i = 1, \ldots, p$. We have already seen that

$$I - Q_{i,\delta} = I - E_{i,0} M_{i,\delta}^{-1} A, \qquad i = 1, \ldots, p.$$

Moreover, it is not hard to see that the splittings $A = M_{i,\delta} - N_{i,\delta}$ (with $N_{i,\delta} = M_{i,\delta} - A$) are regular. Hence, by Lemma 3.1, $T_{RMS,\delta} \ge O$ and $\rho(T_{RMS,\delta}) \le \|T_{RMS,\delta}\|_w < 1$ for any $w = A^{-1}e > 0$ with $e > 0$. Furthermore, by Lemma 2.3, there exists a unique splitting $A = M_{RMS,\delta} - N_{RMS,\delta}$ such that $T_{RMS,\delta} = M_{RMS,\delta}^{-1} N_{RMS,\delta}$. To prove that the splitting is weak regular it suffices to show that

$$M_{RAS,\delta}^{-1} = (I - T_{RMS,\delta}) A^{-1} \ge O \tag{13}$$

or, equivalently, that $M_{RAS,\delta}^{-1} z \ge 0$ for all $z \ge 0$. Letting $v = A^{-1}z \ge 0$, all we need to show is that $(I - T_{RMS,\delta})v \ge 0$ or $T_{RMS,\delta} v \le v$. This is proved in the same way as Lemma 3.1; see [2]. Hence, the unique splitting $A = M_{RMS,\delta} - N_{RMS,\delta}$ is weak regular. ∎

In Example 3.3, we show that the splittings induced by the RAS method and the RMS method are, in general, not nonnegative, i.e., are not weak nonnegative of the second type. This is in contrast with the classical Schwarz methods; see [2].

Example 3.3. For the RAS method, we have to consider

$$\bar{T} = I - AM_{RAS}^{-1} = I - A\sum_{i=1}^{p} E_{i,0}M_{i,\delta}^{-1},$$

while for the RMS method

$$\tilde{T} = N_{RMS,\delta}M_{RMS,\delta}^{-1} = I - AM_{RMS,\delta}^{-1} = I - A(I - T_{RMS,\delta})A^{-1}.$$

It is not hard to see that $\tilde{T} = \prod_{i=p}^{1}(I - \tilde{Q}_{i,\delta})$ with $\tilde{Q}_{i,\delta} = AE_{i,\delta}M_{i,\delta}^{-1}$. Now let

$$A = \begin{bmatrix} 6 & -1 & -2 \\ -2 & 8 & -3 \\ -1 & -1 & 4 \end{bmatrix}, \quad M_{1,\delta} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad E_{1,0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$M_{2,\delta} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad E_{2,0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$M_{3,\delta} = \begin{bmatrix} 6 & -1 & 0 \\ -2 & 8 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad E_{3,0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We obtain

$$\bar{T} = I - A\sum_{i=1}^{3} E_{i,0}M_{i,\delta}^{-1} = \begin{bmatrix} 0.0435 & -0.0054 & 0.5000 \\ 0.3478 & 0.0435 & 0.7500 \\ 0.1739 & 0.1467 & 0 \end{bmatrix} \not\geq 0,$$

$$\tilde{T} = \prod_{i=3}^{1}(I - AE_{i,0}M_{i,\delta}^{-1}) = \begin{bmatrix} -0.0435 & -0.0054 & -0.0258 \\ 0.3478 & 0.0435 & 0.2065 \\ 0.1739 & 0.1467 & 0.1970 \end{bmatrix} \not\geq 0.$$

In the case of no overlap, the RMS method as well as the classical multiplicative Schwarz method reduce to a block Gauss–Seidel method. Similarly, with no overlap, the RAS and the classical additive Schwarz methods reduce to the block Jacobi method. The classical Stein–Rosenberg theorem (see, e.g., [28]) says that for M-matrices the Gauss–Seidel method converges faster than the Jacobi method. The next theorem extends this statement to the case of overlap. We are able to compare the RMS method with the RAS method. We point out that only recently a similar result was obtained for the classical Schwarz methods [20].

We need the following lemma; see [15, 21].

LEMMA 3.4. *Let $A^{-1} \geq O$. Let $A = \bar{M} - \bar{N} = M - N$ be two weak regular splittings such that*

$$\bar{M}^{-1} \geq M^{-1}.$$

Let $w > 0$ be such that $w = A^{-1}e$ for some $e > 0$. Then $\|\bar{M}^{-1}\bar{N}\|_w \leq \|M^{-1}N\|_w$.

THEOREM 3.5. *Let A be a nonsingular M-matrix, and let $w > 0$ be any positive vector such that $Aw > 0$, e.g., $w = A^{-1}v$ with $v > 0$. Then*

$$\|T_{RMS,\delta}\|_w \le \|T_{RAS,\delta}\|_w.$$

Moreover, if the Perron vector w_δ of $T_{RAS,\delta}$ satisfies $w_\delta > 0$ and $Aw_\delta \ge 0$, then we also have

$$\rho(T_{RMS,\delta}) \le \rho(T_{RAS,\delta}).$$

Proof. We will use Lemma 3.4. The splittings corresponding to the RAS and RMS methods are weak regular splittings, and, in particular, the matrices $M_{RAS,\delta}^{-1}$ and $M_{RMS,\delta}^{-1}$ are nonnegative; see (13). Next we show that

$$M_{RMS,\delta}^{-1} \ge M_{RAS,\delta}^{-1}.$$

To that end, we write explicitly $M_{RMS,\delta}^{-1}$ using (13) and (11) as follows:

$$\begin{aligned} M_{RMS,\delta}^{-1} &= \left(I - \prod_{i=p}^{1} (I - E_{i,0} M_{i,\delta}^{-1} A) \right) A^{-1} \\ &= \left(I - (I - E_{p,0} M_{p,\delta}^{-1} A)(I - E_{p-1,0} M_{p-1,\delta}^{-1} A) \cdots (I - E_{1,0} M_{1,\delta}^{-1} A) \right) A^{-1}. \end{aligned}$$

Thus, by computing the product $M_{RMS,\delta}^{-1}$ can be written as

$$\begin{aligned} M_{RMS,\delta}^{-1} = &\sum_{i=1}^{p} E_{i,0} M_{i,\delta}^{-1} - \sum_{i<j} E_{j,0} M_{j,\delta}^{-1} A E_{i,0} M_{i,\delta}^{-1} \\ &+ \sum_{m=3}^{p} \sum_{\substack{(j_1,\ldots,j_m) \\ j_i \in \{1,\ldots,p\}, \\ j_i > j_k \text{ if } i<k}} (-1)^{m+1} \left(\prod_{k=1}^{m} E_{j_k,0} M_{j_k,\delta}^{-1} A \right) A^{-1}. \end{aligned} \tag{14}$$

Note that in each product above all $E_{j_k,0}$ are different, i.e., $E_{j_k,0} \ne E_{j_i,0}$ for $k \ne i$.

The first sum in (14) is just $M_{RAS,\delta}^{-1}$. Thus, all that remains to be shown is that the remaining part of (14) is a nonnegative matrix. To do so, we first consider matrices of the form

$$E_{j,0} M_{j,\delta}^{-1} A E_{i,0} M_{i,\delta}^{-1}.$$

Since $E_{j,0} M_{j,\delta}^{-1} = E_{j,0} M_{j,\delta}^{-1} E_{j,\delta}$ and $E_{i,0} M_{i,\delta}^{-1} = E_{i,0} M_{i,\delta}^{-1} E_{i,\delta}$ we have that

$$E_{j,0} M_{j,\delta}^{-1} A E_{i,0} M_{i,\delta}^{-1} = E_{j,0} M_{j,\delta}^{-1} \left(E_{j,\delta} A E_{i,0} \right) M_{i,\delta}^{-1} E_{i,\delta}. \tag{15}$$

We consider two cases.

Case (a). $S_{j,\delta} \cap S_{i,0} = \emptyset$. Since A is an M-matrix,

$$\left(E_{j,\delta} A E_{i,0} \right)_{s,t} \begin{cases} \le 0 & \text{if } s \in S_{j,\delta},\ t \in S_{i,0} \\ = 0 & \text{otherwise.} \end{cases}$$

Thus, since $E_{j,0} M_{j,\delta}^{-1}$ and $M_{i,\delta}^{-1}$ are nonnegative, we obtain

$$E_{j,0} M_{j,\delta}^{-1} A E_{i,0} M_{i,\delta}^{-1} \le O.$$

Case (b). $S_{j,\delta} \cap S_{i,0} \neq \emptyset$. With the construction on $M_{j,\delta}^{-1}$ in (10), we obtain here

$$\left(E_{j,0}M_{j,\delta}^{-1}A\right)_{s,t} \begin{cases} = 0 & \text{if } s \notin S_{j,0} \\ = 1 & \text{if } s \in S_{j,0},\ s = t \\ = 0 & \text{if } s \in S_{j,0},\ t \in S_{j,\delta},\ s \neq t \\ \leq 0 & \text{otherwise.} \end{cases} \tag{16}$$

Since $S_{j,0} \cap S_{i,0} = \emptyset$ it follows that

$$E_{j,0}M_{j,\delta}^{-1}AE_{i,0}M_{i,\delta}^{-1} \leq O.$$

Therefore in both cases we obtain

$$-\sum_{i<j} E_{j,0}M_{j,\delta}^{-1}AE_{i,0}M_{i,\delta}^{-1} \geq O.$$

Moreover, in both cases we have with (15) that

$$\left(E_{j,0}M_{j,\delta}^{-1}AE_{i,0}M_{i,\delta}^{-1}\right)_{s,t} \begin{cases} \leq 0 & \text{if } s \in S_{j,0},\ t \in S_{i,\delta} \\ = 0 & \text{otherwise.} \end{cases} \tag{17}$$

Finally, consider the terms of the third sum in (14), i.e., consider

$$\begin{aligned} &(-1)^{m+1}\left(\prod_{k=1}^{m} E_{j_k,0}M_{j_k,\delta}^{-1}A\right)A^{-1} \\ &= (-1)^{m+1}\left(\prod_{k=1}^{m-2} E_{j_k,0}M_{j_k,\delta}^{-1}A\right)E_{j_2,0}M_{j_2,\delta}^{-1}AE_{j_1,0}M_{j_1,\delta}^{-1} \end{aligned}$$

with $m \geq 3$. Each of these contains one factor of the form (17), while the other $m-2$ factors are of the form (16). Since the matrices $E_{j_k,0}$ are different for different values of k, the entries with value 1 in (16) get multiplied by zeros when performing the product. This implies that in this case every entry in (16) is nonpositive.

We proceed now by induction on the number of factors. If we have an even number of factors, we have $(-1)^{m+1} = -1$, but since the factor of the form (17) is nonpositive and each of the other $m-2$ factors of the form (16) is also nonpositive, the product is a nonnegative matrix. Similarly, if m is odd $(-1)^{m+1} = 1$, but we have an odd number of nonpositive factors of the form (16) and the nonpositive factor of the form (17). Thus, the product is a nonnegative matrix. Hence in both cases we have

$$(-1)^{m+1}\left(\prod_{k=1}^{m} E_{j_k,0}M_{j_k,\delta}^{-1}A\right)A^{-1} \geq O.$$

Therefore $M_{RMS,\delta}^{-1} \geq M_{RAS,\delta}^{-1}$. Hence with Lemma 3.4 we obtain

$$\|T_{RMS,\delta}\|_w \leq \|T_{RAS,\delta}\|_w.$$

Now, if the Perron vector w_δ of $T_{RAS,\delta}$ satisfies $w_\delta > 0$ and $Aw_\delta \geq 0$, we also have $\|T_{RAS,\delta}\|_{w_\delta} = \rho(T_{RAS,\delta})$. Thus $\rho(T_{RMS,\delta}) \leq \rho(T_{RAS,\delta})$. ∎

To end this section, we compare the RMS method with its classical version. We need first the following two lemmas. The first one is well known and can be found, e.g., in [4], while the second is from [2].

LEMMA 3.6. *Assume that a square matrix T is nonnegative and that for some $\alpha \geq 0$ and for some nonzero vector $x \geq 0$ we have $Tx \geq \alpha x$. Then $\rho(T) \geq \alpha$. The inequality is strict if $Tx > \alpha x$.*

LEMMA 3.7. *Let $A^{-1} \geq O$. Let $A = M - N$ be a splitting such that $M^{-1} \geq O$ and $NM^{-1} \geq O$. Then $\rho(M^{-1}N) < 1$ and there exists a nonzero vector $x \geq 0$ such that $M^{-1}Nx = \rho(M^{-1}N)x$ and $Ax \geq 0$.*

THEOREM 3.8. *Let A be a nonsingular M-matrix, and let $w > 0$ be any positive vector such that $Aw > 0$. Let $T_{RMS,\delta}$ and $T_{MS,\delta}$ be as in (8) and (7); then, for any $\delta \geq 0$,*

$$\|T_{MS,\delta}\|_w \leq \|T_{RMS,\delta}\|_w < 1.$$

Moreover, $\rho(T_{MS,\delta}) \leq \rho(T_{RMS,\delta})$.

Proof. The proof is similar to that of Theorem 4.7 of [2]. We have already seen that $T_{RMS,\delta}$ and $T_{MS,\delta}$ are nonnegative matrices. By Theorem 3.5 of [2] the iteration matrix $T_{MS,\delta}$ induces a nonnegative splitting of A. Let $x \geq 0$, $x \neq 0$ be an eigenvector of $T_{MS,\delta}$ with eigenvalue $\rho(T_{MS,\delta})$. We will show that

$$T_{RMS,\delta}\, x \;\geq\; T_{MS,\delta}\, x \;=\; \rho(T_{MS,\delta})\, x, \tag{18}$$

so that by Lemma 3.6 we get the desired result $\rho(T_{RMS,\delta}) \geq \rho(T_{MS,\delta})$. Let $x^0 = \bar{x}^0 = x$ and define $x^i := (I - E_{i,\delta}M_{i,\delta}^{-1}A)x^{i-1}$ and $\bar{x}^i := (I - E_{i,0}M_{i,\delta}^{-1}A)\bar{x}^{i-1}$, $i = 1, \ldots, p$. Thus, $x^p = T_{MS,\delta}x$ and $\bar{x}^p = T_{RMS,\delta}x$. To establish (18) we proceed by induction and show that

$$Ax^i \geq 0, \qquad i = 1, \ldots, p-1, \tag{19}$$

and

$$0 \leq x^i \leq \bar{x}^i, \qquad i = 1, \ldots, p. \tag{20}$$

We then have (18) since $x^p = T_{MS,\delta}x$ and $\bar{x}^p = T_{RMS,\delta}x$; see (7) and (8).

For $i = 0$, (20) holds by assumption, while relation (19) is true by Lemma 3.7. Assume now that (19) and (20) are both true for some i. To obtain (19) for $i+1$, observe that $Ax^{i+1} = A(I - E_{i,\delta}M_{i,\delta}^{-1}A)x^i = (I - AE_{i,\delta}M_{i,\delta}^{-1})Ax^i$. We have $I - AE_{i,\delta}M_{i,\delta}^{-1} \geq O$, since

$$\begin{aligned} I - M_{i,\delta}^{-T}E_{i,\delta}^T A^T &= I - E_{i,\delta}M_{i,\delta}^{-T}A^T = I - E_{i,\delta} + E_{i,\delta}(I - M_{i,\delta}^{-T}A^T) \\ &= I - E_{i,\delta} + E_{i,\delta}M_{i,\delta}^{-T}N_{i,\delta}^T \geq O, \end{aligned}$$

with $N_{i,\delta} := M_{i,\delta} - A \geq 0$. Moreover, $Ax^i \geq 0$ by the induction hypothesis, and thus (19) holds for $i + 1$. To prove that (20) holds for $i + 1$, we use (19), the fact $E_{i,0} \leq E_{i,\delta}$, and the induction hypothesis to obtain

$$x^{i+1} = (I - E_{i,\delta}M_{i,\delta}^{-1}A)x^i \leq (I - E_{i,0}M_{i,\delta}^{-1}A)x^i \leq (I - E_{i,0}M_{i,\delta}^{-1}A)\bar{x}^i = \bar{x}^{i+1}.$$

To establish the inequalities for the weighted max norms, one proceeds in precisely the same manner as before (using w instead of x) to show $T_{RMS,\delta}w \geq T_{MS,\delta}w$. Since both matrices are nonnegative, by Lemma 2.4 we get $\|T_{MS,\delta}\|_w \leq \|T_{RMS,\delta}\|_w$. ∎

4. Inexact local solves. In the previous section, the subdomain problems were assumed to be solved exactly, and this is represented by the inverses of the matrices $A_{i,\delta}$. In this section, we consider the case were the subdomain problems are solved approximatively or, in other words, inexactly. We represent this fact by using an approximation $\tilde{A}_{i,\delta}$ of the matrix $A_{i,\delta}$. In practice, one uses, for example, an incomplete factorization of $A_{i,\delta}$; see, e.g., [19, 27].

As in [15], suppose that the inexact solves are such that the splittings

$$A_{i,\delta} = \tilde{A}_{i,\delta} - (\tilde{A}_{i,\delta} - A_{i,\delta}) \qquad \text{are weak regular splittings} \tag{21}$$

for $i = 1, \ldots, p$ or that

$$\tilde{A}_{i,\delta} \ \text{ is an } M\text{-matrix and } \ \tilde{A}_{i,\delta} \geq A_{i,\delta}, \qquad i = 1, \ldots, p. \tag{22}$$

Note that (22) implies (21). The incomplete factorizations satisfy (21) [19].

The restricted multiplicative Schwarz iteration with inexact solves on the subdomains is then given by

$$\tilde{T}_{RMS,\delta} = \prod_{i=p}^{1} (I - \tilde{R}_{i,\delta} \tilde{A}_{i,\delta}^{-1} R_{i,\delta} A).$$

In a way similar to (10), we construct matrices

$$\tilde{M}_{i,\delta} = \pi_i^T \begin{bmatrix} \tilde{A}_{i,\delta} & O \\ O & D_{\neg i,\delta} \end{bmatrix} \pi_i \tag{23}$$

such that

$$\tilde{R}_{i,\delta} \tilde{A}_{i,\delta}^{-1} R_{i,\delta} A = E_{i,0} \tilde{M}_{i,\delta}^{-1} A, \qquad i = 1, \ldots, p,$$

and thus

$$\tilde{T}_{RMS,\delta} = \prod_{i=p}^{1} (I - E_{i,0} \tilde{M}_{i,\delta}^{-1} A). \tag{24}$$

We can now establish our convergence result.

THEOREM 4.1. *Let A be a nonsingular M-matrix. Then the RMS iteration matrix* (24) *with inexact solves satisfying* (21) *satisfies $\rho(\tilde{T}_{RMS,\delta}) \leq \|\tilde{T}_{RMS,\delta}\|_w < 1$ for any $w = A^{-1}e > 0$ with $e > 0$. Furthermore, there exists a unique splitting $A = \tilde{B} - \tilde{C}$ such that $\tilde{T}_{RMS,\delta} = \tilde{B}^{-1}\tilde{C}$, and this splitting is weak regular.*

Proof. The proof proceeds in the same manner as that of Theorem 3.2. All we need to show is that each splitting $A = \tilde{M}_{i,\delta} - \tilde{N}_{i,\delta}$ with $\tilde{M}_{i,\delta}$ as in (23) is weak regular. Since $\tilde{A}_{i,\delta}$ is monotone, it follows from (23) that $\tilde{M}_{i,\delta}^{-1} \geq O$. With

$$\pi_i A \pi_i^T := \begin{bmatrix} A_{i,\delta} & K_i \\ L_i & A_{\neg i,\delta} \end{bmatrix}$$

and $\tilde{N}_{i,\delta} = \tilde{M}_{i,\delta} - A$ we have

$$\pi_i \tilde{M}_{i,\delta}^{-1} \tilde{N}_{i,\delta} \pi_i^T = \begin{bmatrix} \tilde{A}_{i,\delta}^{-1}(\tilde{A}_{i,\delta} - A_{i,\delta}) & -\tilde{A}_{i,\delta}^{-1} K_{i,\delta} \\ -D_{\neg i}^{-1} L_{i,\delta} & D_{\neg i}^{-1}(D_{\neg i} - A_{\neg i,\delta}) \end{bmatrix},$$

which, in view of (21) and the fact that A is an M-matrix, is nonnegative. ∎

Note that if (22) holds, then $\tilde{N}_{i,\delta} = \tilde{M}_{i,\delta} - A \geq O$ and $A = \tilde{M}_{i,\delta} - \tilde{N}_{i,\delta}$ is in fact a regular splitting.

It is shown in [16] that if (21) holds, the inexact RAS method given by

$$\tilde{T}_{RAS,\delta} = I - \sum_{i=1}^{p} \tilde{R}_{i,\delta}^T \tilde{A}_{i,\delta}^{-1} R_{i,\delta} A$$

is also convergent and that this matrix also induces a weak regular splitting. We use these properties to compare the inexact RMS method with the inexact RAS method.

THEOREM 4.2. *Let A be an M-matrix and consider the inexact RAS method and the inexact RMS method where the matrices $\tilde{A}_{i,\delta}$ corresponding to the inexact solves satisfy* (21)*. Then, for any positive vector w such that $Aw > 0$ and any $\delta \geq 0$, we have*

$$\|\tilde{T}_{RMS,\delta}\|_w \leq \|\tilde{T}_{RAS,\delta}\|_w < 1. \tag{25}$$

Moreover, if the Perron vector w_δ of $T_{RAS,\delta}$ satisfies $w_\delta > 0$ and $Aw_\delta \geq 0$, then we also have

$$\rho(\tilde{T}_{RMS,\delta}) \leq \rho(\tilde{T}_{RAS,\delta}). \tag{26}$$

Proof. The proof is similar to the proof of Theorem 3.5. With (21) and (23), all matrices $\tilde{M}_{i,\delta}^{-1}$ are nonnegative. We will show that

$$\tilde{M}_{RMS,\delta}^{-1} \geq \tilde{M}_{RAS,\delta}^{-1}.$$

Following the proof of Theorem 3.5 we have only to modify Case (b). However, with (21) we have $\tilde{A}_{i,\delta}^{-1} A_{i,\delta} \leq I$, and thus

$$\left(E_{j,0}\tilde{M}_{j,\delta}^{-1}A\right)_{s,t} \begin{cases} = 0 & \text{if } s \notin S_{j,0} \\ \leq 1 & \text{if } s \in S_{j,0},\ s = t \\ \leq 0 & \text{otherwise.} \end{cases}$$

We then proceed as in the proof of Theorem 3.5.

If the Perron vector w_δ can be chosen as w, we have $\|T_{RAS,\delta}\|_{w_\delta} = \rho(T_{RAS,\delta})$, so that (25) yields $\|T_{RMS,\delta}\|_{w_\delta} \leq \rho(T_{RAS,\delta})$, and since the spectral radius is never larger than any induced operator norm we have (26). ∎

Next we relate the speed of convergence to the exactness of the subdomain solver.

THEOREM 4.3. *Let A be an M-matrix. Consider two inexact RMS methods where the matrices $\tilde{A}_{i,\delta}$ and $\hat{A}_{i,\delta}$ corresponding to the inexact solves satisfy* (22) *and*

$$O \leq \hat{A}_{i,\delta}^{-1} \leq \tilde{A}_{i,\delta}^{-1} \leq A_{i,\delta}^{-1}, \qquad i = 1, \ldots, p.$$

Let the corresponding iteration matrices be as in (24)*. Then, for any positive vector w such that $Aw > 0$ and any $\delta \geq 0$, we have*

$$\|T_{RMS,\delta}\|_w \leq \|\tilde{T}_{RMS,\delta}\|_w \leq \|\hat{T}_{RMS,\delta}\|_w < 1. \tag{27}$$

Proof. From the hypothesis, (10), and (23) it follows that

$$M_{i,\delta}^{-1} \geq \tilde{M}_{i,\delta}^{-1} \geq \hat{M}_{i,\delta}^{-1}.$$

Following the proof of Theorem 3.5, this establishes (27). ∎

5. The effect of overlap on RMS. We study in this section the effect of varying the overlap. More precisely, we prove comparison results on the spectral radii and/or on weighted max norms for the corresponding iteration matrices

$$T_{RMS,\delta} = I - M_{RMS,\delta}^{-1} A$$

for different values of $\delta \geq 0$.

We start with a result which compares one RMS iterative process, defined through the sets $S_{i,\delta'}$, with another one with more overlap defined through sets $S_{i,\delta}$, where $S_{i,\delta'} \subseteq S_{i,\delta}$, $i = 1, \ldots, p$. We show that the larger the overlap ($\delta \geq \delta'$) the faster the RMS method converges as measured in certain weighted max norms.

THEOREM 5.1. *Let A be a nonsingular M-matrix, and let $w > 0$ be any positive vector such that $Aw > 0$. Then, if $\delta \geq \delta'$,*

$$\|T_{RMS,\delta}\|_w \leq \|T_{RMS,\delta'}\|_w < 1. \tag{28}$$

Moreover, if the Perron vector $w_{\delta'}$ of $T_{RMS,\delta'}$ satisfies $w_{\delta'} > 0$ and $Aw_{\delta'} \geq 0$, then we also have

$$\rho(T_{RMS,\delta}) \leq \rho(T_{RMS,\delta'}). \tag{29}$$

Proof. Since $S_{i,\delta'} \subseteq S_{i,\delta}$, $i = 1, \ldots, p$, we have $A \leq M_{i,\delta} \leq M_{i,\delta'} \leq \operatorname{diag}(A)$. Since A is an M-matrix, this yields

$$M_{i,\delta}^{-1} \geq M_{i,\delta'}^{-1} \quad \text{for} \quad i = 1, \ldots, p.$$

Next we compare the matrices $M_{RMS,\delta}^{-1}$ and $M_{RMS,\delta'}^{-1}$. To do so consider (14) in the proof of Theorem 3.5. Since all the parts in the sum are nonnegative, we get (28).

Now, if the Perron vector $w_{\delta'}$ can be chosen as w, we have $\|T_{RMS,\delta'}\|_{w_{\delta'}} = \rho(T_{RMS,\delta'})$ so that (28) yields (29). ∎

As a special case of Theorem 5.1 above we choose $\delta' = 0$, i.e., a block Gauss–Seidel method. In this case, we do not need any additional assumption for comparing the spectral radii. To that end, we use the following comparison theorem due to Woźnicki [29]; see also [12].

THEOREM 5.2. *Let $A^{-1} \geq O$ and two splittings $A = M - N = \tilde{M} - \tilde{M}$, where one of them is weak nonnegative of the first type (weak regular) and the other is weak nonnegative of the second type. If $M^{-1} \geq \tilde{M}^{-1}$, then*

$$\rho(M^{-1}N) \leq \rho(\tilde{M}^{-1}\tilde{N}).$$

THEOREM 5.3. *Let A be a nonsingular M-matrix. Then, for any value of $\delta \geq 0$, $\rho(T_{RMS,\delta}) \leq \rho(T_{RMS,0})$.*

Proof. The proof follows immediately from the above results and the fact that the block Gauss–Seidel splitting is a regular splitting. ∎

6. Varying the number of domains. In this section, we show how the partitioning of a subdomain into smaller subdomains affects the convergence of the restricted Schwarz method. In the M-matrix case, we show that, for both additive and multiplicative restricted Schwarz methods, the more subdomains the slower the convergence rate.

Formally, consider each block of variables $S_{i,\delta}$ partitioned into k_i subblocks; i.e., we have

$$S_{i_j,\delta} \subset S_{i,\delta}, \qquad j = 1, \ldots, k_i, \tag{30}$$

$\bigcup_{j=1}^{k_i} S_{i_j,\delta} = S_{i,\delta}$, and $S_{i_j,\delta} \cap S_{i_k,\delta} = \emptyset$ if $j \neq k$. Each set $S_{i_j,\delta}$ has associated matrices $R_{i_j,\delta}$ and $E_{i_j,\delta} = R_{i_j,\delta}^T R_{i_j,\delta}$. Since we have a partition,

$$E_{i_j,\delta} \leq E_{i,,\delta}, \qquad j = 1, \ldots, k_i, \quad \text{and} \quad \sum_{j=1}^{k_i} E_{i_j,\delta} = E_{i,\delta}, \qquad i = 1, \ldots, p. \tag{31}$$

We define the matrices $A_{i_j,\delta} = R_{i_j,\delta} A R_{i_j,\delta}^T$, and $M_{i_j,\delta}$ corresponding to the set $S_{i_j,\delta}$ as in (10) so that

$$E_{i_j,\delta} M_{i_j,\delta}^{-1} = R_{i_j,\delta}^T A_{i_j}^{-1} R_{i_j,\delta}, \qquad j = 1, \ldots, k_i, \qquad i = 1, \ldots, p.$$

The iteration matrix of the restricted additive Schwarz method with the refined partition is then

$$\bar{T}_{RAS,\delta} = I - \sum_{i=1}^{p} \sum_{j=1}^{k_i} E_{i_j,0} M_{i_j,\delta}^{-1} A, \tag{32}$$

and the unique induced splitting $A = \bar{M}_{RAS,\delta} - \bar{N}_{RAS,\delta}$ (which is a weak regular splitting) is given by

$$\bar{M}_{RAS,\delta}^{-1} = \sum_{i=1}^{p} \sum_{j=1}^{k_i} E_{i_j,0} M_{i_j,\delta}^{-1}.$$

THEOREM 6.1. *Let A be a nonsingular M-matrix. Consider two sets of subblocks of A defined by* (2) *and* (30)*, respectively, and the two corresponding RAS iterations* (9) *and* (32)*. Then, for every $\delta \geq 0$ and for any vector $w > 0$ for which $Aw > 0$, $\|T_{RAS,\delta}\|_w \leq \|\bar{T}_{RAS,\delta}\|_w$.*

Proof. The inclusion (30) implies that

$$M_{i_j,\delta}^{-1} \leq M_{i,\delta}^{-1}, \qquad j = 1, \ldots, k_i, \qquad i = 1, \ldots, p. \tag{33}$$

Thus, with (31) we have

$$\sum_{j=1}^{k_i} E_{i_j,0} M_{i_j,\delta}^{-1} \leq \sum_{j=1}^{k_i} E_{i_j,0} M_{i,\delta}^{-1} = E_{i,0} M_{i,\delta}^{-1}$$

and therefore $\bar{M}_{RAS,\delta}^{-1} \leq M_{RAS,\delta}^{-1}$, which implies the result, using Lemma 3.4. ∎

Next we consider the RMS method. The iteration matrix for the RMS method corresponding to the finer partition (more subdomains) is given by

$$\tilde{T}_{RMS,\delta} = \prod_{i=p}^{1} \prod_{j=k_i}^{1} (I - Q_{i_j,\delta}), \tag{34}$$

where $Q_{i_j,\delta} = E_{i_j,0} M_{i_j,\delta}^{-1} A = \tilde{R}_{i_j,\delta}^T A_{i_j,\delta}^{-1} R_{i_j,\delta} A$.

THEOREM 6.2. *Let A be a nonsingular M-matrix. Consider two sets of subblocks of A defined by* (2) *and* (30)*, respectively, and the two corresponding RMS iterations* (8) *and* (34)*. Then, for any $\delta \geq 0$ and for any vector $w > 0$ for which $Aw > 0$, $\|T_{RMS,\delta}\|_w \leq \|\tilde{T}_{RMS,\delta}\|_w$.*

Proof. Since each $Q_i = E_{i,0}M_{i,\delta}^{-1}A = R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A$ is a projection [16], we have

$$I - Q_i = (I - Q_i)^2 = \cdots = (I - Q_i)^{k_i}.$$

This allows us to represent $T_{RMS,\delta}$ and $\tilde{T}_{RMS,\delta}$ as a product with the same number of factors. We pair each factor $I - Q_{i_j} = I - E_{i_j,0}M_{i_j}^{-1}A$ of $\tilde{T}_{RMS,\delta}$ in (34) with the corresponding factor $I - Q_i = I - E_{i,0}M_{i,\delta}^{-1}A$ of $T_{RMS,\delta}$ in (8). The corresponding set of indices S_{i_j} and S_i satisfy $S_{i_j} \subseteq S_i$. By (31) and (33) we have that $E_{i_j,0}M_{i_j,\delta}^{-1} \leq E_{i,0}M_{i,\delta}^{-1}$. Therefore we can proceed in exactly the same manner as in the proof of Theorem 3.8 to establish the desired result. ∎

7. RMS variants: MSH, RMSH, WRMS, and WMSH. Cai and Sarkis [10] introduced restricted Schwarz methods with harmonic extension. In these variants, the projections $P_{i,\delta}$ in (5) of the classical Schwarz method are replaced by

$$H_{i,\delta} = R_{i,\delta}^T A_{i,\delta}^{-1} \tilde{R}_{i,\delta} A = R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} E_{i,0} A,$$

in contrast to the restricted methods where

$$Q_{i,\delta} = \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A = E_{i,0} R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} A$$

are used. The additive Schwarz method with harmonic extension (ASH method) can then be described in our notation by the iteration matrix $T_{ASH,\delta} = I - M_{ASH}^{-1}A$, where M_{ASH}^{-1} is given by

$$M_{ASH,\delta}^{-1} = \sum_{i=1}^{p} R_{i,\delta}^T A_{i,\delta}^{-1} \tilde{R}_{i,\delta} = \sum_{i=1}^{p} M_{i,\delta}^{-1} E_{i,0}.$$

Similarly, the multiplicative Schwarz method with harmonic extension (MSH method) is defined by

$$T_{MSH,\delta} = \prod_{i=p}^{1} (I - H_{i,\delta}) = \prod_{i=p}^{1} (I - M_{i,\delta}^{-1} E_{i,0} A).$$

It was observed in [10, Rem. 2.4] that the ASH method and the RAS method used as a preconditioner exhibit a similar convergence behavior. In fact, it was shown in [16] that in the case of a symmetric matrix A the two spectra coincide, i.e., $\sigma(M_{ASH,\delta}^{-1}A) = \sigma(M_{RAS,\delta}^{-1}A)$.

In the following, we establish similar results for the MSH method. We have, for a general nonsingular matrix A,

$$\begin{aligned} T_{MSH,\delta}^T &= \prod_{i=p}^{1} (I - R_{i,\delta}^T A_{i,\delta}^{-1} R_{i,\delta} E_{i,0} A)^T = \prod_{i=p}^{1} (I - A^T E_{i,0} R_{i,\delta}^T A_{i,\delta}^{-T} R_{i,\delta}) \\ &= A^T \left(\prod_{i=p}^{1} (I - E_{i,0} R_{i,\delta}^T A_{i,\delta}^{-T} R_{i,\delta} A^T) \right) A^{-T}. \end{aligned} \tag{35}$$

Hence the spectrum of the MSH method is the same as the spectrum of a RMS method for A^T. So, with the weighted column sum norm $\|\cdot\|_{1,w}$ defined for $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ as

$$\|B\|_{1,w} = \max_{j=1,\ldots,n} \left(\sum_{i=1}^{n} |b_{ij}| w_i \right) / w_j,$$

we immediately obtain the following result.

THEOREM 7.1. *Let A be a nonsingular M-matrix. Then the following hold.*

(i) *For any value of $\delta \geq 0$, the splitting $A = M_{MSH,\delta} - N_{MSH,\delta}$, corresponding to the MSH method, is weak nonnegative of the second type, hence*

$$\rho(T_{MSH,\delta}) < 1.$$

(ii) *If $A = A^T$, then for any value of $\delta \geq 0$*

$$\sigma(T_{MSH,\delta}) = \sigma(T_{RMS,\delta}) \qquad \text{and} \qquad \sigma(M_{MSH,\delta}^{-1}A) = \sigma(M_{RMS,\delta}^{-1}A).$$

(iii) *For any positive vector w such that $w^T A > 0$ and for $\delta \geq \delta'$, we have*

$$\|T_{MSH,\delta}\|_{1,w} \leq \|T_{MSH,\delta'}\|_{1,w}.$$

Moreover, if the Perron vector $w_{\delta'}$ of $T_{MSH,\delta'}^T$ satisfies $w_{\delta'} > 0$ and $A^T w_{\delta'} \geq 0$, then we also have

$$\rho(T_{MSH,\delta}) \leq \rho(T_{MSH,\delta'}).$$

(iv) *For any value of $\delta \geq 0$, $\rho(T_{MSH,\delta}) \leq \rho(T_{MSH,0})$.*

In the same way that we showed that the RMS method is faster than the RAS method (Theorem 3.5), we show that the MSH method is faster than the ASH method.

THEOREM 7.2. *Let A be a nonsingular M-matrix. Then, for any value $\delta \geq 0$, we have*

$$\|T_{MSH,\delta}\|_{1,w} \leq \|T_{ASH,\delta}\|_{1,w}.$$

Proof. By Lemma 2.3, we have that $M_{MSH,\delta} = A(I - T_{MSH,\delta})^{-1}$. Thus, using (35), we write

$$M_{MSH,\delta}^{-T} = A^{-T}(I - T_{MSH,\delta}^T) = \left(I - \prod_{i=p}^{1} (I - E_{i,0} M_{i,\delta}^{-T} A^T) \right) A^{-T}.$$

Since

$$\left(M_{ASH,\delta}^{-1}\right)^T = \sum_{i=1}^{p} \tilde{R}_{i,\delta}^T \left(A_{i,\delta}^{-1}\right)^T R_{i,\delta},$$

every ASH-splitting of A gives rise to a corresponding RAS-splitting of A^T [16]. We can then follow the proof of Theorem 3.5 verbatim considering the M-matrix A^T. ∎

We note that Theorems 7.1 and 7.2 hold if inexact solves on the subdomains are used; see section 4.

Combining the restricted and the harmonic versions we obtain the RASH and RMSH methods of [10] with

$$T_{RASH,\delta} = I - M_{RASH,\delta}^{-1} A = I - \sum_{i=1}^{p} \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} \tilde{R}_{i,\delta} A,$$

$$T_{RMSH,\delta} = I - M_{RMSH,\delta}^{-1} A = \prod_{i=p}^{1} (I - \tilde{R}_{i,\delta}^T A_{i,\delta}^{-1} \tilde{R}_{i,\delta} A).$$

However, the RASH method is, in general, not convergent as observed in [16]. The same holds for the RMSH method, as the following example illustrates.

Example 7.3. Consider the symmetric M-matrix

$$A = \begin{bmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}.$$

Let

$$R_{1,0} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad R_{2,0} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and let $R_{1,1} = R_{2,1} = I$. We then have

$$T_{RMSH,1} = (I - E_{2,0}A^{-1}E_{2,0}A)(I - E_{1,0}A^{-1}E_{1,0}A) = \begin{bmatrix} -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -3 & -3 & 2 & 2 \\ -3 & -3 & 2 & 2 \end{bmatrix}$$

with $\rho(T_{RMSH,1}) = 2$.

Other variants of the classical Schwarz methods are the weighted restricted Schwarz methods introduced by Cai and Sarkis [10]. For these modifications, one introduces weighted restriction operators $R^{\omega}_{i,\delta}$ which result from $R_{i,\delta}$ by replacing the entry 1 in column j by $1/k$, where k is the number of sets $S_{i,\delta}$ the component j belongs to, or more generally by some weights adding up to 1. With this notation, we define

$$\tilde{E}_{i,\delta} = R^T_{i,\delta} R^{\omega}_{i,\delta},$$

and we have

$$\sum \tilde{E}_{i,\delta} = \sum R^T_{i,\delta} R^{\omega}_{i,\delta} = I.$$

Then the weighted restricted additive Schwarz (WRAS) method and the weighted restricted multiplicative Schwarz (WRMS) method can be described in our notation by the iteration matrices

$$T_{WRAS,\delta} = I - M^{-1}_{WRAS}A = I - \sum_{i=1}^{p} (R^{\omega}_{i,\delta})^T A^{-1}_{i,\delta} R_{i,\delta} A = I - \sum_{i=1}^{p} \tilde{E}_{i,\delta} M^{-1}_{i,\delta} A \tag{36}$$

and

$$T_{WRMS,\delta} = \prod_{i=p}^{1} (I - (R^{\omega}_{i,\delta})^T A^{-1}_{i,\delta} R_{i,\delta} A) = \prod_{i=p}^{1} (I - \tilde{E}_{i,\delta} M^{-1}_{i,\delta} A), \tag{37}$$

respectively. Similarly weighted Schwarz methods with harmonic extensions can be defined. Observe that $(R^{\omega}_{i,\delta})^T A^{-1}_{i,\delta} R_{i,\delta} A$ is not a projection.

If we compare the WRMS method with the classical multiplicative Schwarz method we obtain the following result.

THEOREM 7.4. *Let A be nonsingular M-matrix. Then, for any $\delta \geq 0$, we have*

$$\rho(T_{MS,\delta}) \leq \rho(T_{WRMS,\delta}) \leq 1. \tag{38}$$

Proof. Following our analysis in the previous sections, we obtain that $T_{WRMS,\delta}$ induces a weak regular splitting of A. Since $\tilde{E}_{i,\delta} \leq E_{i,\delta}$ we get (38) using the same techniques as in the proof of Theorem 3.8. ∎

A similar result holds for the weighted MSH method.

8. Coarse grid corrections. It has been shown theoretically, and confirmed in practice, that a coarse grid correction improves the performance of the classical Schwarz methods. This coarse grid correction can be applied either additively or multiplicatively; see, e.g., [2, 11, 15, 23, 25]. This corresponds to a two-level scheme, the coarse correction being the second level. In [26], a coarse grid correction was used in connection with RAS iterations.

The analysis done for the RAS case in [16] applies almost without changes to the RMS methods of this paper, so we omit the details. All we will say is that in all cases where we have shown convergence the coarse grid correction can never degrade, and often improves, the convergence rate.

Acknowledgments. We thank Michele Benzi and Andreas Frommer for a careful reading of the manuscript and their suggestions which helped improve our presentation.

REFERENCES

[1] S. Balay, W.D. Gropp, L.C. McInnes, and B.F. Smith, *PETSc* 2.0 *User's Manual*, Technical Report ANL-95/11 - Revision 2.0.22, Argonne National Laboratory, Argonne, IL, 1998; also available online from http://www.mcs.anl.gov/petsc.

[2] M. Benzi, A. Frommer, R. Nabben, and D. B. Szyld, *Algebraic theory of multiplicative Schwarz methods*, Numer. Math., 89 (2001), pp. 605–639.

[3] M. Benzi and D. B. Szyld, *Existence and uniqueness of splittings for stationary iterative methods with applications to alternating methods*, Numer. Math., 76 (1997), pp. 309–321.

[4] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979; reprinted and revised, SIAM, Philadelphia, 1994.

[5] P. E. Bjørstad and J. Mandel, *On the spectra of sums of orthogonal projections with applications to parallel computing*, BIT, 31 (1991), pp. 76–88.

[6] P. Bjørstad and O. B. Widlund, *To overlap or not to overlap: A note on a domain decomposition method for elliptic problems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1053–1061.

[7] R. Bru, V. Migallón, J. Penadés, and D. B. Szyld, *Parallel, synchronous and asynchronous two-stage multisplitting methods*, Electron. Trans. Numer. Anal., 3 (1995), pp. 24–38.

[8] X.-C. Cai, C. Farhat, and M. Sarkis, *A Minimum Overlap Restricted Additive Schwarz Preconditioner and Applications to 3D Flow Simulations*, Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 479–485.

[9] X.-C. Cai and Y. Saad, *Overlapping domain decomposition algorithms for general sparse matrices*, Numer. Linear Algebra Appl., 3 (1996), pp. 221–237.

[10] X.-C. Cai and M. Sarkis, *A restricted additive Schwarz preconditioner for general sparse linear systems*, SIAM J. Sci. Comput., 21 (1999), pp. 792–797.

[11] T. F. Chan and T. P. Mathew, *Domain decomposition methods*, Acta Numerica (1994), pp. 61–143.

[12] J.-J. Climent and C. Perea, *Some comparison theorems for weak nonnegative splittings of bounded operators*, Linear Algebra Appl., 275/276 (1998), pp. 77–106.

[13] M. Dryja and O. B. Widlund, *Towards a unified theory of domain decomposition algorithms for elliptic problems*, in Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T.F. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., SIAM, Philadelphia, 1990, pp. 3–21.

[14] A. Frommer, R. Nabben, and D. B. Szyld, *An algebraic convergence theory for restricted additive and multiplicative Schwarz methods*, in Domain Decomposition Methods in Science and Engineering, N. Debit, M. Garbey, R. Hoppe, J. Périaux, D. Keyes, and Y. Kuznetzov, eds., CIMNE, UPS, Barcelona, 2002, pp. 371–377.

[15] A. Frommer and D. B. Szyld, *Weighted max norms, splittings, and overlapping additive Schwarz iterations*, Numer. Math., 83 (1999), pp. 259–278.

[16] A. Frommer and D. B. Szyld, *An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms*, SIAM J. Numer. Anal., 39 (2001), pp. 463–479.

[17] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1994.

[18] T. P. Mathew, *Uniform convergence of the Schwarz alternating method for solving singularly perturbed advection-diffusion equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1663–1683.
[19] J. A. Meijerink and H. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.
[20] R. Nabben, *Comparisons between additive and multiplicative Schwarz iterations in domain decomposition methods*, Numer. Math., to appear.
[21] M. Neumann and R. J. Plemmons, *Convergence of parallel multisplitting methods for M-matrices*, Linear Algebra Appl., 88/89 (1987), pp. 559–574.
[22] D. P. O'Leary and R. E. White, *Multi-splittings of matrices and parallel solution of linear systems*, SIAM J. Alg. Disc. Meth., 6 (1985), pp. 630–640.
[23] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Clarendon Press, Oxford University Press, New York, 1999.
[24] W. C. Rheinboldt and J. S. Vandergraft, *A simple approach to the Perron-Frobenius theory for positive operators on general partially-ordered finite-dimensional linear spaces*, Math. Comp., 27 (1973), pp. 139–145.
[25] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, New York, Melbourne, 1996.
[26] X.-C. Tai, *A space decomposition method for parabolic problems*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 27–46.
[27] R. S. Varga, *Factorization and normalized iterative methods*, in Boundary Problems in Differential Equations, R. E. Langer, ed., The University of Wisconsin Press, Madison, WI, 1960, pp. 121–142.
[28] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962; 2nd ed., Springer-Verlag, Berlin, 2000.
[29] Z. I. Woźnicki, *Nonnegative splitting theory*, Japan J. Indust. Appl. Math., 11 (1994), pp. 289–342.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2337–2351

CONVERGENCE THEORY OF A NUMERICAL METHOD FOR SOLVING THE CHAPMAN–KOLMOGOROV EQUATION*

YUZHI CAI†

Abstract. A convergence theory has been established for a new numerical method for solving the Chapman–Kolmogorov equation [Y. Cai, *A Numerical Forecasting Procedure for Nonlinear Autoregressive Time Series Model*, manuscript, Department of Mathematics and Statistics, University of Surrey, Surrey, UK, 2001]. The theory has been applied to many types of nonlinear time series models in order to obtain m-step ahead predictive probability density function, predictive cumulative distribution function, predictive mean, and predictive variance.

Key words. convergence theory, numerical procedure, nonlinear autoregressive time series models, forecasting

AMS subject classifications. 65, 62

PII. S0036142901390366

1. Introduction.

The nonlinear autoregressive time series model of order k (NLAR(k)) is given by

$$x_t = \lambda(\mathbf{x}_{t-1}) + \xi_t\,, \tag{1}$$

where $\lambda : R^k \to R$, $\mathbf{x}_s = (x_s, x_{s-1}, \ldots, x_{s-k+1})^\top$ and $\{\xi_t\}$ is a sequence of independently, identically distributed (i.i.d.) random variables with zero mean and constant variance σ^2. Let $g(\cdot)$ be the pdf of ξ_t and $f(x_{t+m} \mid \mathbf{x}_t)$ denote the m-step ahead predictive pdf. For convenience, we do not use a subscript on f but rely on its argument subscripts to indicate the random variables involved. Then the Chapman–Kolmogorov equation can be written in the following form:

$$f(x_{t+m} \mid \mathbf{x}_t) = \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}\,, \tag{2}$$

where

$$f(x_{t+1} \mid \mathbf{x}_t) = g(x_{t+1} - \lambda(\mathbf{x}_t)).$$

The m-step ahead predictive pdf given the history up to time t can be obtained by solving (2) recursively. However, in most cases it is impossible to solve (2) directly. Standard numerical integration methods can be used, but they can be very time consuming if accuracy checking procedures are involved. Otherwise the accuracy can not be guaranteed.

Several authors have tried to use (2) to obtain predictive values. For example, Tong and Moeanaddin [9] studied multistep least squares prediction methods for nonlinear autoregressive models and illustrated these with both real and simulated data. Moeanaddin and Tong [5] also studied the numerical evaluation of distributions in nonlinear autoregressions based on (2). In calculating the m-step ahead predictive

*Received by the editors June 4, 2001; accepted for publication (in revised form) July 10, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/39036.html
†Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom (Y.Cai@surrey.ac.uk).

pdf from (2) recursively, they used the Gauss–Hermite quadrature with weights and abscissae supplied by NAG routine D01BCF directly with no accuracy checking at all. Pemberton [6] presented a numerical integration method for self-exciting threshold autoregressive (SETAR) models. Davies, Pemberton, and Petruccelli [3] developed a package for identification, estimation, and forecasting for SETAR models, where the forecasting method is also based on (2) but the numerical integration is based on Sack and Donovan [7]. Again the accuracy cannot be guaranteed numerically.

Although it was pointed out by Davies, Pemberton, and Petruccelli [3] that the Chapman–Kolmogorov equation could be used for general nonlinear models like (1), they did not carry out any numerical or other investigations. To the author's knowledge, up to now no further work has appeared on this aspect.

However, recently Cai [1] presented a numerical method for solving the Chapman–Kolmogorov equation by introducing an unusual accuracy checking procedure into the algorithm. This numerical method can be used to obtain the predictive pdf, cumulative density function (cdf), mean, and variance directly for a range of nonlinear autoregressive time series models, rather than just SETAR models.

The purpose of this paper is to develop a convergence theory for the algorithm presented in Cai [1] and to apply this theory to a range of nonlinear time series models.

After reviewing the algorithm in section 2, we present the convergence theory in section 3. The implementation issues are discussed in section 4. In section 5, we apply the theory to a range of nonlinear time series models. Conclusion and comments are given in section 6.

2. The algorithm. For simplicity, we assume that ξ_t is normally distributed with mean zero and variance σ^2, denoted by $\xi_t \sim N(0, \sigma^2)$. The algorithm can be easily modified for other distribution functions. The analysis is given in detail for the predictive pdf but can be extended to the predictive cdf, mean, and variance. The numerical algorithm for the predictive pdf can be described as follows.

Initially for $m = 1$, we have

$$f(x_{t+1} \mid \mathbf{x}_t) = g(x_{t+1} - \lambda(\mathbf{x}_t)) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_{t+1}-\lambda(\mathbf{x}_t))^2}{2\sigma^2}},$$

and no numerical integration is required.

For $m = 2$, we note that $f(x_{t+2} \mid \mathbf{x}_t)$ can be approximated by $\int_a^b f(x_{t+2} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}$, where $|a|$ and b are sufficiently large. Let

$$f_I = \sum_{i_1} f(x_{t+2} \mid \mathbf{x}_{t+1} \simeq y_{i_1}^I) w_{i_1}^I ,$$

where $\{w_{i_1}^I\}$ and $\{y_{i_1}^I\}$ are the weights and abscissae obtained by applying a numerical quadrature rule (QR) on 2^I equal subintervals of $[a, b]$ with weight function $f(x_{t+1} \mid \mathbf{x}_t)$, and $\mathbf{x}_{t+1} \simeq y_{i_1}^I$ means the first element of $\mathbf{x}_{t+1}$ is replaced by $y_{i_1}^I$. If

$$|f_I - f_{I-1}| < \epsilon , \tag{3}$$

where ϵ is the required accuracy, then we take f_{I-1} as the approximate value of $f(x_{t+2} \mid \mathbf{x}_t)$ and denote the corresponding weights and abscissae as $\{w_{i_1}\}$ and $\{y_{i_1}\}$. Otherwise it is necessary to continue doubling the number of subintervals until (3) is satisfied. Thus an approximate value of $f(x_{t+2} \mid \mathbf{x}_t)$ is given by

$$f(x_{t+2} \mid \mathbf{x}_t) \approx \sum_{i_1} f(x_{t+2} \mid \mathbf{x}_{t+1} \simeq y_{i_1}) w_{i_1} .$$

Repeating the above procedure for a range of values of x_{t+2}, we obtain a discrete version of $f(x_{t+2} \mid \mathbf{x}_t)$ on $[a, b]$.

Generally, for the case $m \geq 3$ we have

$$\begin{aligned} f(x_{t+m} \mid \mathbf{x}_t) &\approx \textstyle\int_a^b f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} \\ &\approx \textstyle\sum_{i_1} f(x_{t+m} \mid \mathbf{x}_{t+1} \simeq y_{i_1}) w_{i_1} \\ &\approx \cdots \\ &\approx \textstyle\sum_{i_1} \sum_{i_2} \cdots \sum_{i_{m-1}} f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y^I_{i_{m-1} \cdots i_2 i_1}) w^I_{i_{m-1} \cdots i_2 i_1} \cdots w_{i_1}, \end{aligned}$$

where $\{w^I_{i_{m-1} \cdots i_2 i_1}\}$ and $\{y^I_{i_{m-1} \cdots i_2 i_1}\}$ are the weights and abscissae obtained by applying the numerical QR on 2^I subintervals of $[a, b]$ with weight function $f(x_{t+m-1} \mid \mathbf{x}_{t+m-2} \simeq y_{i_{m-2} \cdots i_1})$, and $\mathbf{x}_{t+m-1} \simeq y^I_{i_{m-1} \cdots i_2 i_1}$ means that the first $m-1$ elements of $\mathbf{x}_{t+m-1}$ are replaced by $y^I_{i_{m-1} \cdots i_2 i_1}, \ldots, y_{i_1}$.

Let

$$f_I = \sum_{i_{m-1}} f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y^I_{i_{m-1} \cdots i_2 i_1}) w^I_{i_{m-1} \cdots i_2 i_1}.$$

Then if (3) holds, we find a new set of weights $\{w_{i_{m-1} \cdots i_2 i_1}\}$ and abscissae $\{y_{i_{m-1} \cdots i_2 i_1}\}$. By using both the old and the new sets of weights and abscissae, we obtain

$$f(x_{t+m} \mid \mathbf{x}_t) \approx \sum_{i_1} \cdots \sum_{i_{m-1}} f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1} \cdots i_2 i_1}) w_{i_{m-1} \cdots i_2 i_1} \cdots w_{i_1}.$$

As in the case $m = 2$, we can obtain a discrete version of $f(x_{t+m} \mid \mathbf{x}_t)$ on $[a, b]$.

The above idea can be summarized as follows. Let γ_1 and γ_2 be sufficiently large integers (which could be provided by the user such that the difference between the integration on $(-\infty, \infty)$ and $(-10^{\gamma_1}, 10^{\gamma_2})$ is negligible), $\widetilde{M}$ the largest number of steps ahead we wish to forecast, and N the total number of points on which we want to obtain the values of the predictive pdf $f(x_{t+m} \mid \mathbf{x}_t)$. Furthermore, let $\mathbf{w}$ be the set of weights and $\mathbf{y}$ the set of abscissae. Suppose we use an n-point QR for our numerical method. Here Sack and Donovan's [7] numerical integration method is used to provide weights and abscissae on a finite interval. Then our numerical method can be described in the algorithm `FORECAST` (see Table 1) and in the algorithm `Weight` (see Table 2), which is used to calculate the weights and abscissae for the numerical QR.

It is noted that an unusual accuracy checking procedure is involved in the above algorithm, that is, we check the accuracy only for every two-step ahead predictive pdf. In the next section, we will show that, as long as $|a|$ and b are sufficiently large and the number of subintervals of $[a, b]$ is large enough, then the above accuracy checking procedure is sufficient to guarantee the accuracy of any finite m-step ahead predictive pdf.

3. Convergence theory. In this section we establish a convergence theory for the numerical procedure described in section 2.

First we restate a standard result.

THEOREM 3.1. *Let $w(x) \geq 0$ be a weight function defined on $[-1, 1]$ with corresponding orthonormal polynomials $p^*_n(x)$, and let $k^*_n \neq 0$ be the leading coefficient of*

TABLE 1
FORECAST algorithm. FORECAST is used to obtain the m-step ahead predictive pdf for $m = 1, \ldots, \widetilde{M}$.

```
FORECAST(γ1, γ2, n, M̃, N)
   Set a = −10^γ1 , b = 10^γ2
   For i = 0, ..., N
       For m = 1, ..., M̃
           Set x_{t+m,i} = a + (b−a)i/N
   For i = 0, ..., N
       Calculate f̂(x_{t+1,i} | x_t) = g(x_{t+1,i} − λ(x_t))
       For m = 2, ..., M̃
           I = 1
           Weight(m, w, y, n, a, b, I, x_{t+m,i})
           Calculate
               f̂(x_{t+m,i} | x_t) =
               Σ_{i_1} ··· Σ_{i_{m−1}} f(x_{t+m,i} | x_{t+m−1} ≃ y_{i_{m−1}...i_1}) w_{i_{m−1}...i_1} ··· w_{i_1}
   For m = 1, ..., M̃
       For i = 0, ..., N
           Return f̂(x_{t+m,i} | x_t)
```

TABLE 2
Weight algorithm. Weight is used to obtain the weights and the abscissae of the numerical QR.

```
Weight(m, w, y, n, a, b, I, x_{t+m,i})
   Divide interval (a, b) into 2^I equal subintervals
   Calculate abscissae y^I_{i_{m−1} i_{m−2} ... i_1}
       and weights w^I_{i_{m−1}...i_1} on each subinterval
   If |f_I − f_{I−1}| < ε,
       put weights and abscissae corresponding to 2^{I−1}
           subintervals into w, y
       return w, y
   otherwise,
       I = I + 1
       Weight(m, w, y, n, a, b, I, x_{t+m,i})
```

$p_n^*(x)$. *Suppose that* $f(x) \in C^{2n}_{[-1,1]}$. *Then*

$$\int_{-1}^{1} f(t)w(t)dt - \sum_{k=1}^{n} w_k f(x_k) = \frac{f^{(2n)}(\eta)}{(2n)!k_n^{*\,2}}, \quad -1 < \eta < 1,$$

where $\{w_k\}$ *and* $\{x_k\}$ *are the weights and abscissae, respectively.*

Proof. The proof can be found in most numerical analysis books. For example, see Hildebrand [4, pp. 319–321]. □

THEOREM 3.2. *Assume that the conditions of Theorem* 3.1 *hold. Let the interval* $[a, b]$ *be divided into* M *equal subintervals by* $a = u_0 < u_1 < \cdots < u_M = b$. *If the above*

n-point Gauss QR is applied to each subinterval, then there exists $\eta \in (a,b)$ *such that*

$$|R_M(f)| \le \left(\frac{b-a}{2M}\right)^{2n} \frac{(b-a)}{2} \frac{1}{(2n)!k_n^2} \left|f^{(2n)}(\eta)\right| ,$$

where $k_n \neq 0$ *is the smallest leading coefficient of the orthonormal polynomials in each interval, and*

$$R_M(f) = \int_a^b f(x)w(x)dx - \frac{b-a}{2M} \sum_{i=1}^{M} \sum_{k=1}^{n} f\left(u_{i-1} + \frac{b-a}{2M}(1+y_{ki})\right) w_{ki} ,$$

where y_{ki} (w_{ki}) *is the* k*th abscissa (weight) in the* i*th subinterval. Furthermore,* $R_M(f) \to 0$ *as* $M \to \infty$.

Proof. By applying an n-point Gauss QR on each subinterval and by Theorem 3.1, we have

$$\begin{aligned} R_M(f) &= \sum_{i=1}^{M} \left\{ \frac{b-a}{2M} \int_{-1}^{1} f\left(u_{i-1} + \frac{b-a}{2M}(1+y)\right) w\left(u_{i-1} + \frac{b-a}{2M}(1+y)\right) dy \right. \\ &\qquad \left. - \sum_{k=1}^{n} \frac{b-a}{2M} f\left(u_{i-1} + \frac{b-a}{2M}(1+y_{ki})\right) w_{ki} \right\} \\ &= \sum_{i=1}^{M} \left(\frac{b-a}{2M}\right)^{2n+1} \frac{f^{(2n)}(\eta_i)}{(2n)!k_{ni}^2} , \qquad u_{i-1} < \eta_i < u_i , \end{aligned}$$

where $k_{ni} \neq 0$ is the leading coefficient of the nth orthonormal polynomial on the ith subinterval.

Since $f^{(2n)}(x) \in C_{[a,b]}$, we have $\left|f^{(2n)}(x)\right| \in C_{[a,b]}$ and therefore there exist M_1, M_2 such that

$$M_1 \le \frac{1}{M} \sum_{i=1}^{M} \left|f^{(2n)}(\eta_i)\right| \le M_2 .$$

Therefore, from the intermediate value theorem, there exists $\eta \in (a,b)$ such that

$$\frac{1}{M} \sum_{i=1}^{M} \left|f^{(2n)}(\eta_i)\right| = \left|f^{(2n)}(\eta)\right| .$$

Let $k_n^2 = \min_i\{k_{ni}^2\}$. Then we have

$$|R_M(f)| \le \left(\tfrac{b-a}{2M}\right)^{2n} \tfrac{1}{(2n)!k_n^2} \tfrac{b-a}{2M} \textstyle\sum_{i=1}^{M} \left|f^{(2n)}(\eta_i)\right| = \left(\tfrac{b-a}{2M}\right)^{2n} \tfrac{1}{(2n)!k_n^2} \tfrac{b-a}{2} \left|f^{(2n)}(\eta)\right|$$

and

$$R_M(f) \to 0 \quad \text{as } M \to \infty. \qquad \square$$

Theorem 3.2 tells us that the numerical result of the integration on $[a,b]$ will converge to the theoretical value as the number of subintervals of $[a,b]$ tends to infinity. However, Theorem 3.2 is a rather general result on a finite interval. To apply this general result to our case, we also need to deal with the truncated error.

Let

$$\epsilon_{t+j,m-1} = \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+j}) f(x_{t+j} \mid \mathbf{x}_{t+j-1}) dx_{t+j} - \int_{a}^{b} f(x_{t+m} \mid \mathbf{x}_{t+j}) f(x_{t+j} \mid \mathbf{x}_{t+j-1}) dx_{t+j},$$

where a, b are the two chosen constants and $j = 1, 2, \ldots, m-1$. Then for sufficiently large values of $|a|$, b with $a < 0$, $b > 0$, the value of $\epsilon_{t+j,m-1}$ can be sufficiently small and greater than zero for all possible values of j. In the following, we will denote the truncation error by ϵ, the error due to numerical calculation by R, and the overall error $\epsilon + R$ by $\dot{\epsilon}$, all with appropriate subindices.

LEMMA 3.3.

$$\sum_{i_1} \sum_{i_2} \cdots \sum_{i_{m-1}} w_{i_{m-1} \ldots i_1} \ldots w_{i_2 i_1} w_{i_1} \leq 1 \tag{4}$$

and $0 < w_{i_{m-1} \ldots i_1}, \ldots, w_{i_1} < 1$, *where* $w_{i_{m-1} \ldots i_1}, \ldots, w_{i_1}$ *are obtained from the* `Weight` *algorithm (see Table* 2*).*

Proof. This is easily proved by induction, since any pdf must integrate to one. □

THEOREM 3.4. *Suppose that for* $l = 1, 2, \ldots, m-1$, *where* $m > 1$, $f(x_{t+m} \mid \mathbf{x}_{t+l})$ *has continuous* 2*nth derivatives about* x_{t+l} *on* $(-\infty, \infty)$. *Let* $\underline{m} = \min_k\{m_k\}$, *where* m_k *is the largest number of subintervals among all the numerical integrations for obtaining the* k*-step ahead predictive pdf. Then at any point of* x_{t+m}, *and for any* $\epsilon > 0$, *there exist* $a < 0$, $b > 0$, $\underline{M} > 0$ *such that, whenever* $\underline{m} \geq \underline{M}$,

$$\left| \widetilde{E}(f, x_{t+m}) \right| \leq \epsilon,$$

where

$$\widetilde{E}(f, x_{t+m}) = \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \sum_{i_1} \sum_{i_2} \cdots \sum_{i_{m-1}} f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1} \ldots i_1}) w_{i_{m-1} \ldots i_1} \ldots w_{i_1}.$$

Proof. Since $f^{(2n)}(x_{t+m} \mid \mathbf{x}_{t+l})$ is continuous about x_{t+l} on $(-\infty, \infty)$, where

$$f^{(2n)}(x_{t+m} \mid \mathbf{x}_{t+l}) = \frac{\partial^{2n} f(x_{t+m} \mid \mathbf{x}_{t+l})}{\partial x_{t+l}^{2n}},$$

we have for any $[a_1, b_1] \subset (-\infty, \infty)$ that there exists a constant A such that

$$\mid f^{(2n)}(x_{t+m} \mid \mathbf{x}_{t+l}) \mid \leq A$$

for $x_{t+l} \in [a_1, b_1]$, $l = 1, \ldots, m-1$, $m > 1$. Now for $m = 1$, $f(x_{t+1} \mid \mathbf{x}_t) = g(x_{t+1} - \lambda(\mathbf{x}_t))$, so there is no need to use the numerical QR.

For $m = 2$,

$$\sum_{i_1} f(x_{t+2} \mid \mathbf{x}_{t+1} \simeq y_{i_1}) w_{i_1} = \int_{-\infty}^{\infty} f(x_{t+2} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \dot{\epsilon}_{t+1}^{1},$$

where

$$\dot{\epsilon}^1_{t+1} = R_{t+1,1} + \epsilon_{t+1,1}\,,$$

$$R_{t+1,1} = \int_{a_1}^{b_1} f(x_{t+2} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \sum_{i_1} f(x_{t+2} \mid \mathbf{x}_{t+1} \simeq y_{i_1}) w_{i_1}\,,$$

$$\epsilon_{t+1,1} = \int_{-\infty}^{\infty} f(x_{t+2} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \int_{a_1}^{b_1} f(x_{t+2} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}\,.$$

It follows from Theorem 3.2 that there exist $\eta_1^{(1)} \in (a_1, b_1)$ such that

$$\begin{aligned} |R_{t+1,1}| &\le \left(\frac{b_1 - a_1}{2m_1}\right)^{2n} \frac{b_1 - a_1}{2(2n)! k_n^2} \left| f^{(2n)}(x_{t+2} \mid \mathbf{x}_{t+1} \simeq \eta_1^{(1)}) \right| \\ &\le \left(\frac{b_1 - a_1}{2m_1}\right)^{2n} \frac{b_1 - a_1}{2(2n)! k_n^2} A = \overline{R}_{t+1,1}, \qquad \text{say,} \end{aligned}$$

where

$$f^{(2n)}(x_{t+2} \mid \mathbf{x}_{t+1} \simeq \eta_1^{(1)}) = \frac{\partial^{2n} f(x_{t+2} \mid \mathbf{x}_{t+1})}{\partial x_{t+1}^{2n}} \mid_{x_{t+1} = \eta_1^{(1)}}$$

and m_1 is the number of subintervals used for the integration.

Generally, for $m \ge 3$, we have

$$\begin{aligned} &\textstyle\sum_{i_1} \sum_{i_2} \cdots \sum_{i_{m-1}} f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1} \ldots i_1}) w_{i_{m-1} \ldots i_1} \ldots w_{i_1} \\ &= \textstyle\int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \dot{\epsilon}_{t+1}^{m-1} - \sum_{i_1} w_{i_1} \dot{\epsilon}_{t+2}^{i_1} \\ &\quad - \cdots - \textstyle\sum_{i_1} \cdots \sum_{i_{m-2}} w_{i_{m-2}} \ldots w_{i_1} \dot{\epsilon}_{t+m-1}^{i_{m-2} \ldots i_1}\,, \end{aligned}$$

where

$$\dot{\epsilon}_{t+1}^{m-1} = R_{t+1,m-1} + \epsilon_{t+1,m-1}$$

and for $k = 2, 3, \ldots, m-1$

$$\dot{\epsilon}_{t+k}^{i_{k-1} \ldots i_1} = R_{t+k}^{i_{k-1} \ldots i_1} + \epsilon_{t+k}^{i_{k-1} \ldots i_1}$$

and

$$R_{t+1,m-1} = \int_{a_1}^{b_1} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} - \sum_{i_1} f(x_{t+m} \mid \mathbf{x}_{t+1} \simeq y_{i_1}) w_{i_1}\,,$$

$$\begin{aligned} \epsilon_{t+1,m-1} &= \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} \\ &\quad - \int_{a_1}^{b_1} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}\,, \end{aligned}$$

$$\begin{aligned} R_{t+k}^{i_{k-1} \cdots i_1} &= \int_{a_1}^{b_1} f(x_{t+m} \mid \mathbf{x}_{t+k}) f(x_{t+k} \mid \mathbf{x}_{t+k-1} \simeq y_{i_{k-1} \cdots i_1}) dx_{t+k} \\ &\quad - \sum_{i_k} f(x_{t+m} \mid \mathbf{x}_{t+k} \simeq y_{i_k \cdots i_1}) w_{i_k \cdots i_1}\,, \end{aligned}$$

$$\begin{aligned} \epsilon_{t+k}^{i_{k-1} \ldots i_1} &= \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+k}) f(x_{t+k} \mid \mathbf{x}_{t+k-1} \simeq y_{i_{k-1} \ldots i_1}) dx_{t+k} \\ &\quad - \int_{a_1}^{b_1} f(x_{t+m} \mid \mathbf{x}_{t+k}) f(x_{t+k} \mid \mathbf{x}_{t+k-1} \simeq y_{i_{k-1} \ldots i_1}) dx_{t+k}\,. \end{aligned}$$

Then it follows from Theorem 3.2 that, for $k = 2, 3, \dots, m-1$, there exist $\eta_{i_{k-1}\cdots i_1}$ and $\eta_1^{(m-1)} \in (a_1, b_1)$ such that

$$\begin{aligned}\left|R_{t+k}^{i_{k-1},\cdots,i_1}\right| &\le \left(\tfrac{b_1-a_1}{2m_{i_{k-1}\cdots i_1}}\right)^{2n} \tfrac{b_1-a_1}{2(2n)!k_n^2}\left|f^{(2n)}(x_{t+m} \mid \mathbf{x}_{t+k} \simeq \eta_{i_{k-1}\dots i_1})\right| \\ &\le \left(\tfrac{b_1-a_1}{2m_{i_{k-1}\cdots i_1}}\right)^{2n} \tfrac{b_1-a_1}{2(2n)!k_n^2} A = \overline{R}_{t+k}^{i_{k-1}\cdots i_1}, \quad \text{say,}\end{aligned}$$

and

$$\begin{aligned}|R_{t+1,m-1}| &\le \left(\frac{b_1-a_1}{2m_1}\right)^{2n} \frac{b_1-a_1}{2(2n)!k_n^2}\left|f^{(2n)}(x_{t+m} \mid \mathbf{x}_{t+1} \simeq \eta_1^{(m-1)})\right| \\ &\le \left(\frac{b_1-a_1}{2m_1}\right)^{2n} \frac{b_1-a_1}{2(2n)!k_n^2} A = \overline{R}_{t+1,m-1}, \quad \text{say.}\end{aligned}$$

Therefore

$$\begin{aligned}\left|\widetilde{E}(f, x_{t+m})\right| \le \left|\dot{\epsilon}_{t+1}^{m-1}\right| &+ \sum_{i_1}\left|\dot{\epsilon}_{t+2}^{i_1}\right| w_{i_1} + \cdots \\ &+ \sum_{i_1}\sum_{i_2}\cdots\sum_{i_{m-2}}\left|\dot{\epsilon}_{t+m-1}^{i_{m-2}\dots i_1}\right| w_{i_{m-2}\dots i_1}\cdots w_{i_1}.\end{aligned}$$

For $k = 2, 3, \dots, m-1$, set

$$\hat{\epsilon}_{t+k} = \max_{i_{k-1},\cdots,i_1}\left\{\left|\epsilon_{t+k}^{i_{k-1},\cdots,i_1}\right|\right\}, \qquad \hat{R}_{t+k} = \max_{i_{k-1},\cdots,i_1}\left\{\overline{R}_{t+k}^{i_{k-1},\cdots,i_1}\right\},$$

$$\hat{\epsilon}_{t+1} = |\epsilon_{t+1,m-1}|, \qquad \hat{R}_{t+1} = |R_{t+1,m-1}|.$$

Then it follows from Lemma 3.3 that

$$\begin{aligned}&\left|\widetilde{E}(f, x_{t+m})\right| \le |\epsilon_{t+1,m-1}| + |R_{t+1,m-1}| + \textstyle\sum_{i_1}\left(\left|\epsilon_{t+2}^{i_1}\right| + \left|R_{t+2}^{i_1}\right|\right) w_{i_1} + \cdots \\ &\qquad + \textstyle\sum_{i_1}\cdots\sum_{i_{m-2}}\left(\left|\epsilon_{t+m-1}^{i_{m-2}\dots i_1}\right| + \left|R_{t+m-1}^{i_{m-2}\dots i_1}\right|\right) w_{i_{m-2}\dots i_1}\cdots w_{i_1} \qquad (5) \\ &\qquad \le \textstyle\sum_{k=1}^{m-1}\hat{\epsilon}_{t+k} + \sum_{k=1}^{m-1}\hat{R}_{t+k} = \widetilde{\epsilon} + \widetilde{R}, \ \text{ say.}\end{aligned}$$

From the definition of $\hat{\epsilon}_{t+k,}$ we have

$$\hat{\epsilon}_{t+k} \to 0 \quad \text{as} \quad a_1 \to -\infty, \quad b_1 \to \infty.$$

Therefore, for any $\epsilon > 0$, there exist $a < 0, b > 0$ such that $\widetilde{\epsilon} < \frac{\epsilon}{2}$.

From the definition of $\overline{R}_{t+k}^{i_{k-1},\cdots,i_1}$ we have

$$\overline{R}_{t+k}^{i_{k-1},\cdots,i_1} \to 0 \quad \text{as} \qquad m_{i_{k-1},\cdots,i_1} \to \infty.$$

If we let

$$m_k = \max_{i_{k-1},\cdots,i_1}\left\{m_{i_{k-1},\cdots,i_1}\right\},$$

then from the definition of $\hat{R}_{t+k}$ we have $\hat{R}_{t+k} \to 0$ as $m_k \to \infty$. It follows that if $\underline{m} = \min_k\{m_k\}$, then

$$\widetilde{R} \to 0 \quad \text{as} \quad \underline{m} \to \infty.$$

That is, given $\epsilon > 0$, there exists $\underline{M} > 0$ such that, whenever $\underline{m} > \underline{M}$, $\widetilde{R} < \frac{\epsilon}{2}$.

Therefore, for any $\epsilon > 0$, there exist $a < 0, b > 0$, $\underline{M} > 0$ such that, whenever $\underline{m} > \underline{M}$,

$$\left|\widetilde{E}(f, x_{t+m})\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\,. \qquad \square$$

It follows from (5) that the following corollary holds.

COROLLARY 3.5. *If for $k = 1, 2, \ldots, m-1$, we have $\hat{R}_{t+k} < \epsilon$ and $\hat{\epsilon}_{t+k} < \epsilon$, then* $|\widetilde{E}(f, x_{t+m})| \leq 2(m-1)\epsilon$.

Theorem 3.4 tells us that the numerical value obtained from our algorithm will converge to

$$\int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}$$

as $a \to -\infty$, $b \to \infty$ and the number of subintervals of $[a, b]$ tends to infinity. Corollary 3.5 shows that when m increases, the upper bound of the error will usually also increase. However, for any given finite value of m, we can still control the accuracy by appropriate choice of ϵ. They also show that as long as $|a|, b$ are sufficiently large and the number of subintervals of $[a, b]$ is large enough, such that the numerical result of every two-step ahead predictive pdf converges to the corresponding theoretical result, then the numerical result of any finite m-step ahead predictive pdf converges to the corresponding theoretical result. This justifies the accuracy checking procedure in our numerical method.

Note that the above results hold only at one value of x_{t+m}. The following theorem indicates that, under certain conditions, the accuracy can also be guaranteed at other values of x_{t+m} by using the same set of weights and abscissae.

THEOREM 3.6. *Suppose $f(x_{t+m} \mid \mathbf{x}_{t+m-1})$ is continuous about $x_{t+m} \in (-\infty, \infty)$, and we use Theorem 3.4 to calculate the value of $f(x^*_{t+m} \mid \mathbf{x}_t)$ such that*

$$\left|\widetilde{E}(f, x^*_{t+m})\right| < 2(m-1)\epsilon\,. \tag{6}$$

Then there exists a region $\widetilde{D}(x^*_{t+m})$ centered at x^*_{t+m} such that

$$\left|\widetilde{E}(f, x_{t+m})\right| < 6(m-1)\epsilon$$

holds for any $x_{t+m} \in \widetilde{D}(x^*_{t+m})$.

Proof.

$$\begin{aligned}
&\left|\widetilde{E}(f, x_{t+m})\right| \\
&\leq \textstyle\sum_{i_1} \cdots \sum_{i_{m-1}} \left| f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1} \ldots i_1}) \right. \\
&\quad \left. - f(x^*_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1} \ldots i_1}) \right| w_{i_{m-1} \ldots i_1} \cdots w_{i_1} \\
&\quad + \left|\widetilde{E}(f, x^*_{t+m})\right| + \textstyle\int_{-\infty}^{\infty} \left| f(x^*_{t+m} \mid \mathbf{x}_{t+1}) - f(x_{t+m} \mid \mathbf{x}_{t+1}) \right| f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} \\
&= I_1 + I_2 + I_3, \text{ say.}
\end{aligned}$$

Consider first I_1. Since $f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1}\ldots i_1})$ is continuous at any $x_{t+m} \in (-\infty, \infty)$, then for any $\epsilon > 0$ there exists $\delta > 0$ such that, whenever $\left|x_{t+m} - x^*_{t+m}\right| < \delta$,

$$\left|f(x_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1}\ldots i_1}) - f(x^*_{t+m} \mid \mathbf{x}_{t+m-1} \simeq y_{i_{m-1}\ldots i_1})\right| < 2(m-1)\epsilon\,.$$

It therefore follows from Lemma 3.3 that $I_1 < 2(m-1)\epsilon$.

For I_2, it follows from (6) that

$$I_2 = \left|\widetilde{E}(f, x^*_{t+m})\right| < 2(m-1)\epsilon\,.$$

For I_3, note that for $k = 1, 2, \ldots, m-2$ we have

$$\begin{aligned}&\left|f(x^*_{t+m} \mid \mathbf{x}_{t+k}) - f(x_{t+m} \mid \mathbf{x}_{t+k})\right| \\ &\le \textstyle\int_{-\infty}^{\infty} \left|f(x^*_{t+m} \mid \mathbf{x}_{t+k+1}) - f(x_{t+m} \mid \mathbf{x}_{t+k+1})\right| f(x_{t+k+1} \mid \mathbf{x}_{t+k}) dx_{t+k+1},\end{aligned}$$

and whenever $\left|x^*_{t+m} - x_{t+m}\right| < \delta$ we have

$$\left|f(x^*_{t+m} \mid \mathbf{x}_{t+m-1}) - f(x_{t+m} \mid \mathbf{x}_{t+m-1})\right| < 2(m-1)\epsilon\,.$$

Therefore

$$\begin{aligned}&\left|f(x^*_{t+m} \mid \mathbf{x}_{t+m-2}) - f(x_{t+m} \mid \mathbf{x}_{t+m-2})\right| \\ &\le 2(m-1)\epsilon \textstyle\int_{-\infty}^{\infty} f(x_{t+m-1} \mid \mathbf{x}_{t+m-2}) dx_{t+m-1} = 2(m-1)\epsilon\,.\end{aligned}$$

It is easy to see that for any $k = 1, 2, \ldots, m-1$ we have

$$\left|f(x^*_{t+m} \mid \mathbf{x}_{t+k}) - f(x_{t+m} \mid \mathbf{x}_{t+k})\right| \le 2(m-1)\epsilon\,. \tag{7}$$

Hence, by substituting (7) into I_3 we have

$$\begin{aligned}I_3 &= \textstyle\int_{-\infty}^{\infty} \left|f(x^*_{t+m} \mid \mathbf{x}_{t+1}) - f(x_{t+m} \mid \mathbf{x}_{t+1})\right| f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} \\ &\le 2(m-1)\epsilon \textstyle\int_{-\infty}^{\infty} f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} = 2(m-1)\epsilon\,.\end{aligned}$$

In summary, for any $\epsilon > 0$ there exists $\delta > 0$ such that whenever $|x_{t+m} - x^*_{t+m}| < \delta$ we have

$$\left|\widetilde{E}(f, x_{t+m})\right| \le I_1 + I_2 + I_3 = 6(m-1)\epsilon\,.$$

Therefore we can take

$$\widetilde{D}(x^*_{t+m}) = \{x_{t+m} : \ |x_{t+m} - x^*_{t+m}| < \delta\}$$

as required. □

Theorem 3.6 tells us that, with little loss of accuracy, we can use the same weights and abscissae to calculate the values of the predictive pdf $f(x_{t+m} \mid \mathbf{x}_t)$ at different points, say, x_{t+m} and x^*_{t+m}, as long as x_{t+m} and x^*_{t+m} are "close" enough. That means, in order to guarantee the accuracy on the whole interval $[a, b]$, we need to check the accuracy at a number of points. However, if the number of the points is infinite, we would not be able to do so. The following theorem solves this problem theoretically.

THEOREM 3.7. *Suppose* $f(x_{t+m} \mid \mathbf{x}_{t+m-1})$ *is continuous on* $(-\infty, \infty)$ *about* x_{t+m} *and that, for any* $x^*_{t+m} \in [a,b]$, $|\widetilde{E}(f, x^*_{t+m})| < 2(m-1)\epsilon$. *Then there exists a finite number of regions* $\widetilde{D}(x^{(i)}_{t+m})$ *centered at* $x^{(i)}_{t+m} \in [a,b]$, $i = 1, \ldots, l$, *such that* $[a,b] \subset \bigcup_{i=1}^{l} \widetilde{D}(x^{(i)}_{t+m})$ *and for any* $x_{t+m} \in [a,b]$ *we have*

$$\left|\widetilde{E}(f, x_{t+m})\right| \leq 6(m-1)\epsilon\,.$$

Proof. It follows from Theorem 3.6 that for any $x^*_{t+m} \in [a,b]$ there exists a region $\widetilde{D}(x^*_{t+m})$ centered at x^*_{t+m} such that for any $x_{t+m} \in \widetilde{D}(x^*_{t+m})$

$$\left|\widetilde{E}(f, x_{t+m})\right| < 6(m-1)\epsilon\,. \tag{8}$$

Therefore $[a,b] \subset \bigcup \widetilde{D}(x^*_{t+m})$. Then it follows from the Heine–Borel theorem that there exist a finite number of regions $\widetilde{D}(x^{(i)}_{t+m})$, say, $i = 1, \ldots, l$, such that $[a,b] \subset \bigcup_{i=1}^{l} \widetilde{D}(x^{(i)}_{t+m})$ and (8) holds for any $x_{t+m} \in [a,b]$. □

Theorem 3.7 tells us that we can find a finite number of nonoverlapping sets of weights and abscissae such that the numerical value of $f(x_{t+m} \mid \mathbf{x}_t)$ can be calculated at any $x_{t+m} \in [a, b]$. As we have seen from the above results, the error bound grows linearly with the lead time m. Therefore, in order to achieve the same accuracy we should decrease ϵ as m increases. However, given ϵ all the errors are upper bounded by $6(\widetilde{M}-1)\epsilon$, where $\widetilde{M}$ is the max-value of m. Further discussion about applications of the results will be given in the next section.

We have established a convergence theory for the numerical method for solving the Chapman–Kolmogorov equation in the case of the predictive pdf. It is not difficult to develop the corresponding theory for the predictive cdf, mean, and variance, because we can derive the required formulae as follows.

For the predictive cdf, we have

$$F(x_{t+m} \mid \mathbf{x}_t) = \int_{-\infty}^{\infty} F(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}, \quad m = 2, 3, \ldots,$$

$$F(x_{t+1} \mid \mathbf{x}_t) = \int_{-\infty}^{x_{t+1}} f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} = \int_{-\infty}^{x_{t+1}} g(x_{t+1} - \lambda(\mathbf{x}_t)) dx_{t+1}.$$

For the predictive mean, we have

$$E(x_{t+m} \mid \mathbf{x}_t) = \int_{-\infty}^{\infty} E(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}, \qquad m = 2, 3, \ldots,$$

$$E(x_{t+1} \mid \mathbf{x}_t) = \lambda(\mathbf{x}_t).$$

For the predictive variance, we have

$$\operatorname{var}(x_{t+m} \mid \mathbf{x}_t) = E(x^2_{t+m} \mid \mathbf{x}_t) - E^2(x_{t+m} \mid \mathbf{x}_t),$$

$$E(x^2_{t+m} \mid \mathbf{x}_t) = \int_{-\infty}^{\infty} E(x^2_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}, \quad m = 2, 3, \ldots,$$

$$\operatorname{var}(x_{t+1} \mid \mathbf{x}_t) = \sigma^2(\mathbf{x}_t),$$

$$E(x_{t+1} \mid \mathbf{x}_t) = \lambda(\mathbf{x}_t), \quad E(x^2_{t+1} \mid \mathbf{x}_t) = \sigma^2(\mathbf{x}_t) + \lambda^2(\mathbf{x}_t).$$

It is seen that the predictive cdf, mean, and variance can also be obtained by solving similar equations recursively. We do not give any details here for reasons of space. However, it is worth pointing out that the predictive pdf obtained from the numerical method is only a discrete version of the theoretical predictive pdf. We could use this discrete pdf to estimate the predictive cdf, mean, and variance. Alternatively, we could also use the numerical method to obtain a discrete version of the predictive cdf or find the predictive mean and variance. As long as the functions $F(x_{t+m} \mid \mathbf{x}_{t+l})$, $E(x_{t+m} \mid \mathbf{x}_{t+l})$, and $E(x_{t+m}^2 \mid \mathbf{x}_{t+l})$ satisfy similar conditions to those satisfied by $f(x_{t+m} \mid \mathbf{x}_{t+l})$, for $l = 1, 2, \ldots, m-1$ and $m > 1$, then the accuracy check procedure guarantees their accuracy too.

4. Implementation issues. In the previous section we have presented the convergence theory of our numerical integration method for solving the Chapman–Kolmogorov equation. Theoretically, under certain conditions, the numerical results will converge to the true values. However, several practical issues in applying the theory will be discussed in this section.

The first question is how to choose a and b. Generally, $|a|$ and b should be chosen as large as possible such that the difference between the integrations on $(-\infty, \infty)$ and $[a, b]$ is negligible. In our examples Cai [1] we take $\gamma_1 = \gamma_2 = 7$, which gives $a = -10^7$ and $b = 10^7$. Our experience indicates that, for the models we studied, the truncated errors are negligible for such large values of $|a|$ and b.

Since the metric space $(-\infty, \infty)$ is complete, it follows from Theorem 3.4 and its corollaries that we can check the accuracy according to (3). This is an unusual accuracy checking procedure, but it guarantees the accuracy and increases the speed of the calculations.

We have investigated different ways of obtaining the sets of weights and abscissae. Since our main purpose here is to obtain a discrete version of the m-step ahead predictive pdf, for $m = 1, 2, \ldots, \widetilde{M}$, we could divide the interval $[a, b]$ into N equal subintervals and then calculate $f(x_{t+m} \mid \mathbf{x}_t)$ at each end point of the subintervals with the accuracy checking procedure applied. Hence, as long as N is large enough, we can obtain a good discrete version of $f(x_{t+m} \mid \mathbf{x}_t)$ on $[a, b]$. In our examples in Cai [1] we take $N = 50$, which works very well and the results are very good.

We also carried out the following experiment. Since $f(x_{t+m} \mid \mathbf{x}_{t+m-1})$ is a normal density, given $\epsilon > 0$, we can actually write a program and find the value of a_i such that $a = a_0 < a_1 < \cdots < a_N = b$ and $|f(a_i \mid \mathbf{x}_{t+m-1}) - f(a_{i-1} \mid \mathbf{x}_{t+m-1})| < \epsilon$ for $i = 1, 2, \ldots, N$. Theorem 3.7 guarantees that N must be finite. Of course, in this case the lengths of the subintervals are not the same. Then we check the accuracy at each point a_i. Theorem 3.6 guarantees that we can also calculate $f(x_{t+m} \mid \mathbf{x}_t)$ for any $x_{t+m} \in (a_{i-1}, a_i)$ by using the set of weights and abscissae obtained at a_{i-1} with the desired accuracy. The disadvantage of this method is that for a very small value of ϵ, N can be very large, so the calculation can be time consuming. It is hoped that, with the development of modern technology, this will not be a serious problem in the near future.

In practice, what value of n in the n-point QR should be used? Because the accuracy is checked by doubling the number of subintervals, the length of the subintervals is usually small. Our experience shows that two- or three-point Gauss QR is good enough for satisfactory results. The accuracy of all the results in Cai [1] is so high that it is difficult to distinguish the numerical and theoretical values by graphs. All the results presented in the tables in Cai [1] show that our procedure works very well.

5. Applications to specific NLAR models. In this section, we will show that a range of NLAR time series models satisfy the conditions of Theorem 3.4, so the numerical procedure can be applied to these models. All the models mentioned below can be found in Tong [8].

Consider the following general nonlinear time series model:

(9)
$$x_t = \left(a_0 + \sum_{i=1}^{k} a_i x_{t-i}\right) + \left((b_0 G(u) + d_0 H(v)) + \sum_{i=1}^{k} (b_i G(u) + d_i H(v))\, x_{t-i}\right) + \xi_t\,,$$

where $k \geq 0$ is the order of the model, ξ_t are i.i.d. $N(0, \sigma^2)$ random variables, and u and v are functions of x_{t-d}, where $d \geq 1$ is the delay of the model. Then different models can be derived from (9) as follows.

1. Let $d_i = 0$ for $i = 0, 1, \ldots, k$, $u = (x_{t-d} - r)/c$, $G(u) = \Phi(u)$, where $\Phi(u)$ is the standard normal distribution function. Then we obtain a smooth threshold autoregressive model (STAR).
2. Let $d_i = 0$ for $i = 0, 1, \ldots, k$, $u = r_1(x_{t-d} - c_1)$, $r_1 > 0$, $G(u) = (1 + e^{-u})^{-1}$. Then we obtain a logistic STAR (LSTAR) model.
3. Let $b_i = 0$ for $i = 0, 1, \ldots, k$, $v = r_2(x_{t-d} - c_2)^2$, $r_2 > 0$, $H(v) = 1 - e^{-v}$. Then we obtain an exponential STAR (ESTAR) model.
4. Let $u, G(u)$ and $v, H(v)$ be given by 2 and 3 above. Then we obtain a hybrid STAR (HYSTAR) model.
5. Let $b_0 = 0$ and $d_i = 0$ for $i = 0, 1, \ldots, k$, $u = cx_{t-d}^2$, $G(u) = e^{-u}$. Then we obtain an exponential autoregressive (EXPAR) model.

For the above models, $g(\cdot)$ is a normal density function, g, G, H, u, $v \in C^{\infty}_{[a,b]}$. Furthermore, it follows from the Chapman–Kolmogorov equation that, for any $1 \leq l \leq m-2$,

$$f(x_{t+m} \mid \mathbf{x}_{t+l}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+m-1}) \ldots f(x_{t+l+1} \mid \mathbf{x}_{t+l}) dx_{t+m-1} \ldots dx_{t+l+1}\,.$$

Therefore, the integrand $f(x_{t+m} \mid \mathbf{x}_{t+m-1}) \ldots f(x_{t+l+1} \mid \mathbf{x}_{t+l}) \in C^{\infty}_{[a,b]}$. It can be shown that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^{2n+1} f(x_{t+m} \mid \mathbf{x}_{t+m-1}) \ldots f(x_{t+l+1} \mid \mathbf{x}_{t+l})}{\partial x_{t+l}^{2n+1}} dx_{t+m-1} \ldots dx_{t+l+1} < \infty.$$

Therefore $f(x_{t+m} \mid \mathbf{x}_{t+l}) \in C^{2n}_{[a,\,b]}$ (for $1 \leq l \leq m-2$) about x_{t+l}. On the other hand, since $f(x_{t+m} \mid \mathbf{x}_{t+m-1})$, $F(x_{t+m} \mid \mathbf{x}_{t+m-1})$, $E(x_{t+m} \mid \mathbf{x}_{t+m-1})$, and $E(x_{t+m}^2 \mid \mathbf{x}_{t+m-1})$ are all continuous about $x_{t+m} \in (-\infty, \infty)$, we can use the methods mentioned in the previous sections to obtain forecast distributions for the above models.

Note that in the STAR model there is a parameter d. The predictive pdfs for up to d steps are easily shown to be

$$f(x_{t+m} \mid \mathbf{x}_t) = \frac{1}{\sigma_m \sqrt{2\pi}} e^{-\frac{(x_{t+m} - \mu_m)^2}{2\sigma_m^2}}\,, \qquad m \leq d\,,$$

where $\mu_m = E(x_{t+m} \mid \mathbf{x}_t)$, $\sigma_m^2 = \mathrm{var}(x_{t+m} \mid \mathbf{x}_t)$. Therefore, as long as we can obtain μ_m, σ_m, we can obtain the whole predictive pdf and cdf for $m \leq d$.

The following theorem provides the formulae for calculating the predictive means and variances when $m \leq d$. The detailed proof can be found in Cai [2].

THEOREM 5.1. *For the STAR model let $m \leq d$ and*

$$\widehat{x}_t(l) = \begin{cases} E\left(x_{t+l} \mid \mathbf{x}_t\right), & l > 0, \\ x_{t+l}, & l \leq 0. \end{cases}$$

Then the following hold:

$$\widehat{x}_t(m) = \left(a_0 + \sum_{i=1}^{k} a_i \widehat{x}_t(m-i)\right) + \left(b_0 + \sum_{i=1}^{k} b_i \widehat{x}_t(m-i)\right) \Phi\left(\frac{x_{t+m-d} - r}{c}\right)$$

and

$$\operatorname{var}(x_{t+m} \mid \mathbf{x}_t) = \sigma^2 \sum_{i=1}^{m} \left(\alpha_i^{(m)}\right)^2 ,$$

where

$$\begin{aligned} \alpha_m^{(m)} &= 1\,, \\ \alpha_{m-1}^{(m)} &= a_1 \alpha_{m-1}^{(m-1)} + b_1 \alpha_{m-1}^{(m-1)} \Phi\left(\tfrac{x_{t+m-d}-r}{c}\right)\,, \\ &\vdots \\ \alpha_{m-l}^{(m)} &= \textstyle\sum_{i=1}^{l} a_i \alpha_{m-l}^{(m-i)} + \left(\sum_{i=1}^{l} b_i \alpha_{m-l}^{(m-i)}\right) \Phi\left(\tfrac{x_{t+m-d}-r}{c}\right)\,, \\ &\vdots \\ \alpha_1^{(m)} &= \textstyle\sum_{i=1}^{m-1} a_i \alpha_1^{(m-i)} + \left(\sum_{i=1}^{m-1} b_i \alpha_1^{(m-i)}\right) \Phi\left(\tfrac{x_{t+m-d}-r}{c}\right)\,. \end{aligned}$$

Therefore, when we apply our numerical procedure to the STAR model, we need to use Theorem 5.1 to obtain the m-step ahead predictive pdf for $m \leq d$, while for $m > d$, we apply our numerical procedure.

Similarly, we can prove that our numerical method can be applied to the other models and that a similar result to Theorem 5.1 holds for each of these models.

Finally, a SETAR model is given by

$$x_t = a_0^{(j)} + \sum_{i=1}^{k} a_i^{(j)} x_{t-i} + \xi_t^{(j)} \qquad \text{if } x_{t-d} \in (r_{j-1}, r_j],$$

where $-\infty \leq r_0 < r_1 < \cdots < r_l \leq \infty$ are the threshold values. The order of the model is $k \geq 0$, the delay of the model is $d \geq 1$, and $\xi_t^{(j)}$ are i.i.d. $N(0,\ \sigma_j^2)$, $j = 1, \ldots, l$.

Using the expression

$$\begin{aligned} f(x_{t+m} \mid \mathbf{x}_t) &= \int_{-\infty}^{\infty} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1} \\ &= \sum_{i=1}^{l} \int_{r_{i-1}}^{r_i} f(x_{t+m} \mid \mathbf{x}_{t+1}) f(x_{t+1} \mid \mathbf{x}_t) dx_{t+1}, \end{aligned}$$

it is not difficult to modify the above theory in order to deal with SETAR model. Again we omit the details here. A similar result to Theorem 5.1 for the SETAR model was quoted by Tong [8, p. 356]. The examples in Cai [1] show that our forecasting procedure works very well for SETAR models.

6. Comments and conclusion. We have presented a convergence theory for our numerical method for solving the Chapman–Kolmogorov equation. We have shown that the method can be applied to a range of nonlinear autoregressive time series models, not only for obtaining the predictive pdf, but also for obtaining the predictive cdf, mean, and variance. The examples in Cai [1] indicate that our numerical method works very well for these models.

From our experience we see that the actual storage required by the numerical method depends on the models. Thus different forecasting horizons can be achieved with different models. The relationship between the storage requirement and the model needs to be investigated in the future.

We considered the case when the ξ_t are i.i.d. normal. Theoretically, for other continuous distributions we can also obtain the desired results. The performance of the algorithm for different distributional assumptions on ξ_t also needs to be compared and investigated in the future.

It may be possible to calculate predictive cdf, mean, and variance by using the weights and abscissae obtained for pdf. However, whether the accuracy can be guaranteed by the accuracy check on the pdf needs to be investigated in the future.

Finally, it is possible to extend the theory to autoregressive conditional heteroskedastic (ARCH) models in certain circumstances. The details will be presented elsewhere.

Acknowledgment. I would like to express my sincere thanks to the referees and to Professor Trevor Sweeting for their thoughtful comments which greatly enhanced the presentation of this paper.

REFERENCES

[1] Y. Cai, *A Numerical Forecasting Procedure for Nonlinear Autoregressive Time Series Models*, manuscript, Department of Mathematics and Statistics, University of Surrey, Surrey, UK, 2001

[2] Y. Cai, *Convergence of a Forecasting Procedure for Nonlinear Autoregressive Time Series Models*, manuscript, Department of Mathematics and Statistics, University of Surrey, Surrey, UK, 2001.

[3] N. Davies, J. Pemberton, and J. D. Petruccelli, *An automatic procedure for identification, estimation and forecasting univariate self exciting threshold autoregressive models*, The Statistician, 37 (1988), pp. 199–204.

[4] F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd ed., Mcgraw–Hill, New York, 1974, pp. 319–321.

[5] R. Moeanaddin and H. Tong, *Numerical evaluation of distribution in nonlinear autoregression*, Journal Time Ser. Anal., 11 (1990), pp. 38–48.

[6] J. Pemberton, *Exact least squares multi-step prediction from nonlinear time series models*, Journal Time Ser. Anal., 8 (1987), pp. 443–448.

[7] R. A. Sack and A. F. Donovan, *An algorithm for Gaussian quadrature given modified moments*, Numer. Math., 18 (1972), pp. 465–478.

[8] H. Tong, *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press, New York, 1990.

[9] H. Tong and R. Moeanaddin, *On multi-step nonlinear least squares prediction*, The Statistician, 37 (1988), pp. 101–110.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2352–2367

STRONG SEMISMOOTHNESS OF EIGENVALUES OF SYMMETRIC MATRICES AND ITS APPLICATION TO INVERSE EIGENVALUE PROBLEMS*

DEFENG SUN† AND JIE SUN‡

Abstract. It is well known that the eigenvalues of a real symmetric matrix are not everywhere differentiable. A classical result of Ky Fan states that each eigenvalue of a symmetric matrix is the difference of two convex functions, which implies that the eigenvalues are semismooth functions. Based on a recent result of the authors, it is further proved in this paper that the eigenvalues of a symmetric matrix are strongly semismooth everywhere. As an application, it is demonstrated how this result can be used to analyze the quadratic convergence of Newton's method for solving inverse eigenvalue problems (IEPs) and generalized IEPs with multiple eigenvalues.

Key words. symmetric matrices, eigenvalues, strong semismoothness, Newton's method, inverse eigenvalue problems, quadratic convergence

AMS subject classifications. 65F15, 65F18, 65H10, 65H17

PII. S0036142901393814

1. Introduction.

The theory of semismooth functions developed in the last decade has been successful in analyzing the quadratic convergence of Newton's method for nondifferentiable (nonsmooth) equations; it is well received by the optimization community, but is perhaps not well known by researchers in numerical analysis. In this paper we take the inverse eigenvalue problem (IEP) as an example to show how this theory can be used in analyzing matrix-related equations. For applications of the IEP the interested reader is referred to the paper of Friedland, Nocedal, and Overton [10], the book of Xu [27], and the references therein. For general nonsmooth analysis involving eigenvalues of symmetric matrices and a survey on eigenvalue optimization, see Lewis [12] and Lewis and Overton [13], respectively.

Let $\mathcal{S}$ be the linear space of symmetric matrices of size n. Let $A : \Re^n \to \mathcal{S}$ be continuously differentiable. Given n real numbers $\{\lambda_i^*\}_{i=1}^n$, which are arranged in the decreasing order $\lambda_1^* \geq \cdots \geq \lambda_n^*$, the IEP is to find a vector $c^* \in \Re^n$ such that $\lambda_i(A(c^*)) = \lambda_i^*$ for $i = 1, \ldots, n$. A typical choice for $A(c)$ is

$$A(c) = A_0 + \sum_{j=1}^{n} c_j A_j, \tag{1}$$

where $A_0, A_1, \ldots, A_n \in \mathcal{S}$. In this case, $A(c)$ is an affine function of c.

*Received by the editors August 16, 2001; accepted for publication (in revised form) July 10, 2002; published electronically January 14, 2003.
http://www.siam.org/journals/sinum/40-6/39381.html

†Department of Mathematics, National University of Singapore, Singapore 117543, Republic of Singapore (matsundf@nus.edu.sg). This author's research was partially supported by grant R146-000-035-101 from the National University of Singapore.

‡School of Business and Singapore-MIT Alliance, National University of Singapore, Singapore 119620, Republic of Singapore (jsun@nus.edu.sg). This author's research was partially supported by grant R314-000-028/042-112 from the National University of Singapore and a grant from the Singapore-MIT Alliance.

Define $F : \Re^n \to \Re^n$ by

$$F(c) = \begin{bmatrix} \lambda_1(A(c)) - \lambda_1^* \\ \vdots \\ \lambda_n(A(c)) - \lambda_n^* \end{bmatrix}. \tag{2}$$

Then the IEP is equivalent to finding $c^* \in \Re^n$ to be a solution of the following equation:

$$F(c) = 0. \tag{3}$$

Of course, there are other ways to formulate the IEP as a system of equations. For instance, we may solve $F(c) = 0$, where

$$F(c) = \begin{bmatrix} \det(A(c) - \lambda_1^* I) \\ \vdots \\ \det(A(c) - \lambda_n^* I) \end{bmatrix}. \tag{4}$$

A Newton method was proposed by Biegler-König [2] for model (4), which generalizes an algorithm of Lancaster [11]. However, as analyzed by Friedland, Nocedal, and Overton [10], model (2) seems to be always preferred over model (4) both from theoretical and computational points of view. Thus, we concentrate on model (2) in this paper. The convergence theory we are going to present is based on a property of F called strong semismoothness (defined later). It is well known that for $X \in \mathcal{S}$ the eigenvalues of X, as functions of X, are not everywhere differentiable. However, we shall show that they are strongly semismooth and therefore quadratic convergence of Newton's method is a natural result when applied to equations involving eigenvalues. In doing so, we also give a constructive proof for a difficult result of Chen and Tseng [4] on upper semicontinuity of a set-valued mapping of orthogonal matrices.

The concept of semismoothness of functionals was originally studied by Mifflin [14] while strong semismoothness was introduced by Qi and Sun in [18] for vector valued functions. Recently, both concepts are further extended to matrix valued functions [24]. Generally speaking, *strong semismoothness* of an equation is tied with quadratic convergence of the Newton method applied to the equation and *semismoothness* corresponds to superlinear convergence. It was shown that smooth functions, piecewise smooth functions, and convex and concave functions are semismooth functions. They are not, however, necessarily strongly semismooth functions.

To see the motivation of this paper more clearly, let us consider the following example:

$$X = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix},$$

where x_1, x_2, and x_3 are parameters. In this case, we have

$$\lambda_1(X) = \frac{x_1 + x_3 + \sqrt{(x_1 - x_3)^2 + 4x_2^2}}{2} \quad \text{and} \quad \lambda_2(X) = \frac{x_1 + x_3 - \sqrt{(x_1 - x_3)^2 + 4x_2^2}}{2}. \tag{5}$$

Since $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are not differentiable at X with $x_1 = x_3$ and $x_2 = 0$, a gradient-dependent numerical method (e.g., Newton's method) may get into trouble when

hitting those points. In addition, theoretical analysis gets tricky without differentiability. Further inspection reveals that $\lambda_1(\cdot)$ is a convex function and $\lambda_2(\cdot)$ is a concave function. Hence, both of them are semismooth functions and a nonsmooth version of Newton's method [18] might be applied to equations containing $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$. This should be not a coincidence. Let $f_m(X)$ be the sum of m largest eigenvalues of X. Then, Ky Fan's maximum principle [8, 1] says that for each $i = 1, \ldots, n$, $f_i(\cdot)$ is a convex function. This result implies that

- $\lambda_1(\cdot)$ is a convex function and $\lambda_n(\cdot)$ is a concave function; and,
- for $i = 2, \ldots, n-1$, $\lambda_i(\cdot)$ is the difference of two convex functions.

Since convex and concave functions are semismooth and the difference of two semismooth functions is still a semismooth function [14], Ky Fan's result shows that $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$ are all semismooth functions. It is therefore expected, when applying the nonsmooth Newton method to IEPs, the convergence rate is at least superlinear. A more interesting question is, Are all $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$ *strongly* semismooth functions (therefore implying quadratic convergence)? In this paper, based on a recent result of the authors [24], we will give an affirmative answer to the above question.

The organization of this paper is as follows. Some basic facts on semismoothness are presented in section 2. Some nonsmooth versions of the Newton method, which we call *relative generalized Newton methods,* are introduced in section 3. Section 4 concentrates on showing the strong semismoothness of eigenvalues of a symmetric matrix. The quadratic convergence of the relative generalized Newton methods for IEPs and generalized IEPs is proved in section 5. Section 6 gives a summary and a few possible future research topics.

Some notations to be used are as follows.

- $\mathcal{S}$ is the set of real symmetric matrices; $\mathcal{O}$ is the set of all $n \times n$ orthogonal matrices.
- A superscript "T" represents the transpose of matrices and vectors. For a matrix M, $M_{i\cdot}$, and $M_{\cdot j}$ represent the ith row and jth column of M, respectively.
- Unless otherwise specified, all vector norms are 2-norms and matrix norms are Frobenius norms: $\|M\| := \mathrm{trace}\,(M^T M)^{1/2}$.
- A diagonal matrix is written as $\mathrm{diag}\,(\beta_1, \ldots, \beta_n)$ and a block-diagonal matrix is denoted by $\mathrm{diag}\,(B_1, \ldots, B_s)$, where $B_1, \ldots, B_s$ are matrices.
- The eigenvalues of $X \in \mathcal{S}$ is designated by $\lambda_i(X)$, $i = 1, \ldots, n$, and $\Lambda(X) := \mathrm{diag}\,(\lambda_1(X), \ldots, \lambda_n(X))$.
- We write $X = O(\alpha)$ (respectively, $o(\alpha)$) if $\|X\|/|\alpha|$ is uniformly bounded (respectively, tends to zero) as $\alpha \to 0$.

2. Some basic facts on semismoothness.

2.1. Semismooth functions. Let $G : \Re^n \to \Re^m$ be a locally Lipschitz continuous function. We regard the $r \times r$ symmetric matrix space as a special case of $\Re^s$ with $s = r(r+1)/2$. Hence the discussions of this subsection apply to matrix variable and/or matrix valued functions as well.

According to Rademacher's theorem, G is differentiable almost everywhere. Let D_G be the set of differentiable points of G and let G' be the Jacobian of G whenever it exists. Denote

$$\partial_B G(x) := \{V \in \Re^{m\times n} |\ V = \lim_{x^k \to x} G'(x^k),\ x^k \in D_G\}.$$

Then Clarke's generalized Jacobian [5] is

$$\partial G(x) = \mathrm{conv}\{\partial_B G(x)\}, \tag{6}$$

where "conv" stands for the convex hull in the usual sense of convex analysis [20].

DEFINITION 2.1. *Suppose that $G : \Re^n \to \Re^m$ is a locally Lipschitz continuous function. G is said to be semismooth at $x \in \Re^n$ if G is directionally differentiable at x and for any $V \in \partial G(x + \Delta x)$*

$$G(x + \Delta x) - G(x) - V(\Delta x) = o(\|\Delta x\|).$$

G is said to be p-order ($0 < p < \infty$) semismooth at x if G is semismooth at x and

$$G(x + \Delta x) - G(x) - V(\Delta x) = O(\|\Delta x\|^{1+p}). \tag{7}$$

In particular, G is called strongly semismooth at x if G is 1-order semismooth at x.

A function G is said to be a (strongly) semismooth function if it is (strongly) semismooth everywhere on $\Re^n$. It is shown that the composition of (strongly) semismooth functions is still a (strongly) semismooth function (see [14, 9]).

The next result [24, Theorem 3.7] provides a convenient tool for proving strong semismoothness.

THEOREM 2.2. *Suppose that $G : \Re^n \to \Re^m$ is locally Lipschitzian and directionally differentiable in a neighborhood of x. Then for any $p \in (0, \infty)$ the following two statements are equivalent:*

(a) *for any $V \in \partial G(x + \Delta x)$,*

$$G(x + \Delta x) - G(x) - V(\Delta x) = O(\|\Delta x\|^{1+p});$$

(b) *for any $x + \Delta x \in D_G$,*

$$G(x + \Delta x) - G(x) - G'(x + \Delta x)(\Delta x) = O(\|\Delta x\|^{1+p}). \tag{8}$$

2.2. Generalized Newton methods. Suppose that $G : \Re^n \to \Re^n$ is locally Lipschitz continuous. Based on $\partial G(x)$, Qi and Sun [18] proposed the following Newton method for solving $G(x) = 0$.

Generalized Newton method I. Given $x^0 \in \Re^n$, for $k = 0, 1, \ldots,$

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \tag{9}$$

where $V_k \in \partial G(x^k)$.

The following convergence theorem for the generalized Newton method I is established in [18].

THEOREM 2.3. *Suppose that $G(x^*) = 0$. If all $V \in \partial G(x^*)$ are nonsingular and G is semismooth at x^*, then there exists a neighborhood $N(x^*)$ of x^* such that for any $x^0 \in N(x^*)$ the generalized Newton method* I *is well defined and is Q-superlinearly convergent. Moreover, if G is strongly semismooth at x^*, then* (9) *converges Q-quadratically.*

To relax the nonsingularity assumption on $\partial G(x^*)$, Qi [17] introduced the following method based on the concept of $\partial_B G(x)$.

Generalized Newton method II. Given $x^0 \in \Re^n$, for $k = 0, 1, \ldots,$

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \tag{10}$$

where $V_k \in \partial_B G(x^k)$.

The convergence theorem for the generalized Newton method II is the same as Theorem 2.3 except that ∂G is replaced by $\partial_B G$.

Now, let us consider the following composite nonsmooth equation:

$$G(x) := \Phi(\Psi(x)) = 0, \tag{11}$$

where $\Phi : \Re^n \to \Re^m$ is nonsmooth but of special structure and $\Psi : \Re^m \to \Re^n$ is continuously differentiable. It is noted that neither $\partial G(x)$ nor $\partial_B G(x)$ is easy to compute even if $\partial\Phi(y)$, $\partial_B\Phi(y)$, and $\Psi'(x)$ are available. To circumvent the difficulty in computing $\partial G(x)$ and $\partial G_B(x)$, Potra, Qi, and Sun [16] introduced the following concept of generalized Jacobian:

$$\partial_Q G(x) = \partial_B\Phi(\Psi(x))\Psi'(x),$$

where "Q" stands for "quasi." We shall see in the later discussion that $\partial_Q G(x)$ is more convenient to compute than $\partial G(x)$ and $\partial_B G(x)$ for IEPs.

Generalized Newton method III. Given $x^0 \in \Re^n$, for $k = 0, 1, \ldots,$

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \tag{12}$$

where $V_k \in \partial_Q G(x^k)$.

The following convergence theorem for the generalized Newton method III for solving (11) is proved in [16, Theorem 5.3].

THEOREM 2.4. *Suppose that G is defined by* (11) *and $G(x^*) = 0$. If all $V \in \partial_Q G(x^*)$ are nonsingular and Φ is semismooth at $\Psi(x^*)$, then there exists a neighborhood $N(x^*)$ of x^* such that for any $x^0 \in N(x^*)$ the generalized Newton method* III *is well defined and is Q-superlinearly convergent. Moreover, if Φ is strongly semismooth at $\Psi(x^*)$ and Ψ' is Lipschitz continuous around x^*, then* (12) *converges Q-quadratically.*

3. Relative generalized Newton methods. It should be noted that, apart from the semismoothness, another key assumption for the superlinear convergence of the generalized Newton methods I–III is the nonsingularity of $\partial G(x^*)$, $\partial_B G(x^*)$, or $\partial_Q G(x^*)$. However, this may not be satisfied in general for IEPs with multiple eigenvalues. In order to weaken the nonsingularity assumption on the generalized Jacobians, we shall introduce the concept of *relative* generalized Jacobians and the corresponding generalized Newton methods based on the concept of relative generalized gradient introduced by Clarke [5, p. 231].

Let S be a subset of $\Re^n$. For instance, in the context of matrix functions, S could represent the set of all nonsingular matrices. The S-relative generalized Jacobian $\partial|_S G(x)$ of G at x is defined by

$$\partial|_S G(x) := \{V \mid V \text{ is a limit of } V_i \in \partial G(y_i),\ y_i \in S, y_i \to x\}.$$

The following result can be proved in an analogous way to [5, Proposition 6.2.1]. We omit the details.

LEMMA 3.1. *Let G be Lipschitz continuous near x. Then we have the following:*

(a) *$\partial|_S G(x)$ is a compact subset of $\partial G(x)$.*

(b) *$\partial|_S G(x) = \partial G(x)$ if x lies in the interior part of S; $\partial|_S G(x) = \emptyset$ if $(x+\varepsilon B)\cap S = \emptyset$ for some $\varepsilon > 0$; and $\partial|_S G(x)$ is nonempty if $x \in \mathrm{cl}(S)$, the closure of S.*

(c) *$\partial|_S G(\cdot)$ is upper semicontinuous at x.*

Now, we can introduce our first relative generalized Newton method for solving $G(x) = 0$.

Relative generalized Newton method I. Given $x^0 \in \Re^n$, for $k = 0, 1, \ldots,$ and $x^k \in S$,

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \tag{13}$$

where $V_k \in \partial|_S G(x^k)$.

In the following analysis, we assume that the relative generalized Newton method I does not find a solution of $G(x) = 0$ in a finite number of steps.

THEOREM 3.2. *Suppose that $G(x^*) = 0$ and $x^* \in \mathrm{cl}(S)$. If all $V \in \partial|_S G(x^*)$ are nonsingular and G is semismooth at x^*, then there exists a neighborhood $N(x^*)$ of x^* such that for any $x^0 \in N(x^*) \cap S$ the relative generalized Newton method* I *either stops in a finite number of steps with some $x^k \notin S$ or generates an infinite sequence $\{x^k\} \in N(x^*) \cap S$ and the whole sequence converges Q-superlinearly to x^*. Moreover, if G is strongly semismooth at x^*, then the rate of convergence is Q-quadratic.*

Proof. By using Lemma 3.1, there exist a neighborhood $N(x^*)$ of x^* and a positive number κ such that for any $x \in N(x^*) \cap S$, all $V \in \partial|_S G(x)$ are nonsingular and

$$\|V^{-1}\| \le \kappa. \tag{14}$$

Since G is semismooth at x^*, by shrinking $N(x^*)$ if necessary, we have for all $x \in N(x^*) \cap S$ and $V \in \partial|_S G(x)$,

$$\|G(x) - G(x^*) - V(x - x^*)\| \le \frac{1}{2\kappa}\|x - x^*\|. \tag{15}$$

By using (14) and (15), we have for $k = 0, 1, \ldots$ that

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \|x^k - V_k^{-1} G(x^k) - x^*\| \\
&= \|V_k^{-1}[G(x^k) - G(x^*) - V_k(x - x^*)]\| \\
&\le \|V_k^{-1}\|\|G(x^k) - G(x^*) - V_k(x - x^*)\| \\
&\le \tfrac{1}{2}\|x^k - x^*\|,
\end{aligned}$$

which implies that if (13) does not stop at some step with $x^k \notin S$, then $\{x^k\} \in N(x^*) \cap S$ and the whole sequence converges to x^* linearly.

Next, suppose that (13) does not stop at some step with $x^k \notin S$. Since G is semismooth at x^* and $x^k \to x^*$, we have

$$G(x^k) - G(x^*) - V_k(x^k - x^*) = o(\|x^k - x^*\|),$$

which, together with (13), implies that

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \|x^k - V_k^{-1} G(x^k) - x^*\| \\
&= \|V_k^{-1}[G(x^k) - G(x^*) - V_k(x - x^*)]\| \\
&= O(\|G(x^k) - G(x^*) - V_k(x - x^*)\|) \\
&= o(\|x^k - x^*\|).
\end{aligned}$$

This proves the superlinear convergence of $\{x^k\}$.

By the above argument, we can see that if G is strongly semismooth at x^*, then (13) either stops in finitely many steps with some $x^k \notin S$ or generates an infinite

sequence $\{x^k\} \in N(x^*) \cap S$ and the whole sequence converges Q-quadratically to x^*. This completes the proof. □

The proof of Theorem 3.2 might serve as an example to show the simplicity of the analysis of Newton's method by using the concept of (strong) semismoothness. Parallel to the definition of $\partial_B G(x)$ and $\partial_Q G(x)$, we define

$$\partial_B|_S G(x) := \{V|\ V \text{ is a limit of } V_i \in \partial_B G(y_i),\ y_i \in S,\ y_i \to x\}$$

and

$$\partial_Q|_S G(x) := \{V|\ V \text{ is a limit of } V_i \in \partial_Q G(y_i),\ y_i \in S,\ y_i \to x\}.$$

Similar to Lemma 3.1, we have the following lemma.

LEMMA 3.3. *Let G be Lipschitz continuous near x. Then we have the following:*

(a) *$\partial_B|_S G(x)$ and $\partial_Q|_S G(x)$ are compact subsets of $\partial_B G(x)$ and $\partial_Q G(x)$, respectively.*

(b) *$\partial_B|_S G(x) = \partial_B G(x)$ and $\partial_Q|_S G(x) = \partial_Q G(x)$, if x lies in the interior part of S; $\partial_B|_S G(x) = \partial_Q|_S G(x) = \emptyset$ if $(x + \varepsilon B) \cap S = \emptyset$ for some $\varepsilon > 0$; both $\partial_B|_S G(x)$ and $\partial_Q|_S G(x)$ are nonempty if $x \in \mathrm{cl}(S)$, the closure of S.*

(c) *$\partial_B|_S G(\cdot)$ and $\partial_Q|_S G(\cdot)$ are upper semicontinuous at x.*

Analogously, we define the second and third relative generalized Newton methods.

Relative generalized Newton method II (III). Given $x^0 \in \Re^n$, for $k = 0, 1, \ldots$, and $x^k \in S$,

$$x^{k+1} = x^k - V_k^{-1} G(x^k), \tag{16}$$

where $V_k \in \partial_B|_S G(x^k)$ ($V_k \in \partial_Q|_S G(x^k)$ in method III).

The following theorem can be similarly proved by using Lemma 3.3 and the approach of proving Theorems 3.2. We omit the details.

THEOREM 3.4. *Suppose that $G(x^*) = 0$ and $x^* \in \mathrm{cl}(S)$. If all $V \in \partial_B|_S G(x^*)$ ($V \in \partial_Q|_S G(x^*)$ in method* III*) are nonsingular and G is semismooth at x^* (Φ is semismooth at $\Psi(x^*)$ in method* III*), then there exists a neighborhood $N(x^*)$ of x^* such that for any $x^0 \in N(x^*) \cap S$ the relative generalized Newton methods* II *and* III *either stop in a finite number of steps with some $x^k \notin S$ or generate an infinite sequence $\{x^k\} \in N(x^*) \cap S$ and the whole sequence converges Q-superlinearly to x^*. Moreover, if G (Φ in method* III*) is strongly semismooth at x^* (at $\Psi(x^*)$ and Ψ' is Lipschitz continuous around x^* in method* III*), then the rate of convergence is Q-quadratic.*

4. Strong semismoothness of eigenvalues. As a building block for applying relative generalized Newton methods, we shall prove the strong semismoothness of eigenvalues of symmetric matrices in this section. Suppose $X \in \mathcal{S}$. Then, there exists an orthogonal matrix $Q \in \mathcal{O}$ such that X satisfies

$$Q^T X Q = \Lambda(X) := \mathrm{diag}\,(\lambda_1(X), \ldots, \lambda_n(X)), \tag{17}$$

where $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$.

We define a "configuration vector" K to distinguish different eigenvalues. Let

$$K := \{k_0, k_1, \ldots, k_l\} \tag{18}$$

with $1 = k_0 < k_1 < \cdots < k_l = n + 1$ such that there is a change of eigenvalues at k_i. Namely for $t = 1, \ldots, l$,

$$\lambda_s(X) = \lambda_{k_{t-1}}(X), \qquad s \in [k_{t-1}, k_t - 1], \tag{19}$$

where we use the simple notation $[k_{t-1}, k_t - 1]$ to represent the index set $\{k_{t-1}, k_{t-1} + 1, \ldots, k_t - 1\}$.

Let $H \in \mathcal{S}$ and let P (depending on H) be an orthogonal matrix such that

$$P^T(\Lambda(X) + H)P = \Lambda(Y) := \text{ diag}\,(\lambda_1(Y), \ldots, \lambda_n(Y)), \tag{20}$$

where $\lambda_1(Y) \geq \cdots \geq \lambda_n(Y)$ and $Y := \Lambda(X) + H$.

After the above preparation, we can state the following result, which was essentially proved in the derivation of Lemma 4.2 of [24].

LEMMA 4.1. *For any $H \in \mathcal{S}$ and $H \to 0$, we have*

$$P_{ij} = O(\|H\|), \qquad i, j = 1, \ldots, n, \ (i, j) \notin \bigcup_{t=1}^{l} \{[k_{t-1}, k_t - 1] \times [k_{t-1}, k_t - 1]\}. \tag{21}$$

Proof. It has been proved in the proof of Lemma 4.2 of [24] that (21) is true for any $H \in \mathcal{S}$ such that $\Lambda(X) + H$ is nonsingular and $H \to 0$.

Next, we prove that (21) is also true for the case that $\Lambda(X) + H$ is singular and $H \to 0$. It is easy to check that the conclusion of this lemma holds if $H = 0$. Hence, we can assume $H \neq 0$. Define

$$\lambda_{\min}(|Y|) = \min_{\lambda_i(Y) \neq 0} |\lambda_i(Y)| \quad \text{and} \quad \tilde{\Lambda} = \text{ diag}\,(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n),$$

where $|Y| := (Y^2)^{\frac{1}{2}}$ and for $i = 1, \ldots, n$

$$\tilde{\lambda}_i = \begin{cases} \lambda_i(Y) & \text{if } \lambda_i(Y) \neq 0, \\ \lambda_{\min}(|Y|) \min\{\frac{1}{2}, ||H||^2\} & \text{otherwise.} \end{cases}$$

Denote

$$\tilde{H} = P\tilde{\Lambda}P^T - \Lambda(X).$$

Hence, $P^T[\Lambda(X) + \tilde{H}]P = \tilde{\Lambda}$ is nonsingular. By noting the fact $\tilde{H} = H + O(\|H\|^2)$, it follows that (21) also holds for the case that $\Lambda(X) + H$ is singular and $H \to 0$. This completes the proof. ∎

Define a "truncated" matrix $W \in \Re^{n \times n}$ as follows:

$$W_{ij} = \begin{cases} P_{ij} & \text{if } (i, j) \in \bigcup_{t=1}^{l} \{[k_{t-1}, k_t - 1] \times [k_{t-1}, k_t - 1]\}, \\ 0 & \text{otherwise,} \end{cases} \qquad i, j = 1, \ldots, n. \tag{22}$$

Hence, from Lemma 4.1, we know that for any $H \to 0$,

$$W = P + O(\|H\|). \tag{23}$$

It is noted, however, that W may not be an orthogonal matrix but has a block-diagonal structure with each block corresponding to a set of identical eigenvalues of X. That is,

$$W = \text{ diag}\,(W_1, \ldots, W_l),$$

where

$$W_t = \left(P_{ij}\right)_{i,j=k_{t-1}}^{k_t - 1}, \qquad \text{for } t = 1, \ldots, l.$$

Since $P \in \mathcal{O}$, by using Lemma 4.1 and (22), for $t = 1, \ldots, l$ and $i, j = 1, \ldots, k_t - k_{t-1}$, we have for any $H \to 0$,

$$\|(W_t)_{\cdot j}\|^2 = 1 + O(\|H\|^2) \quad \text{and} \quad \langle (W_t)_{\cdot j}, (W_t)_{\cdot i} \rangle = O(\|H\|^2), \qquad i \neq j. \tag{24}$$

It is obvious from (24) that for any $H \in \mathcal{S}$ sufficiently close to 0 the columns of W_t are independent because

$$\sum_j \beta_j (W_t)_{\cdot j} = 0 \;\Rightarrow\; \beta_j[1 + O(\|H\|^2)] = O(\|H\|^2) \;\Rightarrow\; \beta_j = 0 \quad \forall\, j.$$

For each $t = 1, \ldots, l$, let $\tilde{P}_t$ be a matrix of the same order of W_t and be obtained by applying the Gram–Schmidt orthogonalization algorithm to each W_t; i.e., for $j = 1, \ldots, k_t - k_{t-1}$, let

$$(\tilde{W}_t)_{\cdot j} = (W_t)_{\cdot j} - \sum_{i=1}^{j-1} \langle (\tilde{P}_t)_{\cdot i}, (W_t)_{\cdot j} \rangle (\tilde{P}_t)_{\cdot i} \;\text{ and }\; (\tilde{P}_t)_{\cdot j} = (\tilde{W}_t)_{\cdot j} / \|(\tilde{W}_t)_{\cdot j}\|. \tag{25}$$

By (24) and (25), for $i, j = 1, \ldots, k_t - k_{t-1}$, $t = 1, \ldots, l$, we have for any $H \to 0$ that

$$\|(\tilde{P}_t)_{\cdot j}\|^2 = 1, \;\; (\tilde{P}_t)_{\cdot j} = (W_t)_{\cdot j} + O(\|H\|^2) \quad \text{and} \quad \langle (\tilde{P}_t)_{\cdot j}, (\tilde{P}_t)_{\cdot i} \rangle = 0, \quad i \neq j. \tag{26}$$

Denote

$$\tilde{P} = \operatorname{diag}(\tilde{P}_1, \ldots, \tilde{P}_l). \tag{27}$$

Then, we have the following lemma.

LEMMA 4.2. *For any $H \in \mathcal{S}$ sufficiently small, the matrix $\tilde{P}$ defined by* (27) *and* (25) *is an orthogonal matrix and satisfies*

$$\tilde{P}^T \Lambda(X) \tilde{P} = \Lambda(X). \tag{28}$$

Furthermore, for any $H \to 0$,

$$P = \tilde{P} + O(\|H\|). \tag{29}$$

Proof. By (26), we know that each $\tilde{P}_t$, $t = 1, \ldots, l$, is an orthogonal matrix. Since $\lambda_{k_{t-1}}(X) = \cdots = \lambda_{k_t - 1}(X)$, $t = 1, \ldots, l$, we have

$$\tilde{P}_t^{\,T} \operatorname{diag}(\lambda_{k_{t-1}}(X), \ldots, \lambda_{k_t-1}(X)) \tilde{P}_t = \operatorname{diag}(\lambda_{k_{t-1}}(X), \ldots, \lambda_{k_t-1}(X)).$$

Hence, $\tilde{P}$ is an orthogonal matrix and satisfies (28). By using (23) and (26), we directly obtain (29). This completes the proof. ∎

For any $\Delta X \in \mathcal{S}$, let $U \in \mathcal{O}$ (depending on X and ΔX) be any orthogonal matrix such that

$$U^T(X + \Delta X)U = \Lambda(X + \Delta X) := \operatorname{diag}(\lambda_1(X + \Delta X), \ldots, \lambda_n(X + \Delta X)), \tag{30}$$

where $\lambda_1(X + \Delta X) \geq \cdots \geq \lambda_n(X + \Delta X)$.

By using the above lemma, we have the following result.

LEMMA 4.3. *For any $\Delta X \in \mathcal{S}$ sufficiently small and U satisfying* (30)*, there exists a $V \in \mathcal{O}$ such that*

$$V^T X V = \Lambda(X) \quad \text{and} \quad U = V + O(\|\Delta X\|). \tag{31}$$

Proof. Let $P = Q^T U$ and $H = Q^T \Delta X Q$, where Q is defined in (17). Then, by Lemma 4.2, for any such defined P, there exists $\tilde{P} \in \mathcal{O}$ such that

$$\tilde{P}^T \Lambda(X) \tilde{P} = \Lambda(X)$$

and

$$P = \tilde{P} + O(\|H\|) = \tilde{P} + O(\|\Delta X\|).$$

Let $V = Q\tilde{P}$. Then $V \in \mathcal{O}$,

$$V^T X V = \tilde{P}^T Q^T X Q \tilde{P} = \tilde{P}^T \Lambda(X) \tilde{P} = \Lambda(X),$$

and for any $\Delta X \to 0$

$$U = V + O(\|\Delta X\|).$$

This completes the proof. ∎

A similar result to Lemma 4.3 has also been proved in [4] based on a so-called $\sin(\Theta)$ theorem in [21, Theorem 3.4]. The proof provided here is due to a direct comparison between entries of P and $\tilde{P}$ and it indeed furnishes an algorithm for computing V.

One direct result of Lemma 4.3 is that the (normalized) eigenvectors of symmetric matrices, though not continuous, are upper Lipschitz continuous. To see this, for any $Z \in \mathcal{S}$, let

$$\mathcal{U}(Z) := \{U \in \mathcal{O} |\ U^T Z U \text{ is diagonal}\},$$

and let

$$\mathcal{E} := \{M \in \mathcal{S} \mid\ |M_{i,j}| \leq 1,\ i, j = 1, \ldots, n\}.$$

PROPOSITION 4.4. *For any $X \in \mathcal{S}$, there exists a constant $\mu > 0$ such that*

$$\mathcal{U}(X + \Delta X) \subseteq \mathcal{U}(X) + \mu ||\Delta X|| \mathcal{E} \tag{32}$$

for all ΔX sufficiently small.

Proof. For any $U \in \mathcal{U}(X + \Delta X)$, there exists a diagonal matrix $D(X + \Delta X)$ such that

$$U^T (X + \Delta X) U = D(X + \Delta X).$$

Let $R \in \Re^{n \times n}$ be a permutation matrix such that

$$R D(X + \Delta X) R^T = \Lambda(X + \Delta X)$$

with $\lambda_1(X + \Delta X) \geq \cdots \geq \lambda_n(X + \Delta X)$. Let $\tilde{U} = U R^T$. Then we obtain $\tilde{U}^T (X + \Delta X) \tilde{U} = \Lambda(X + \Delta X)$. Hence, by Lemma 4.3, there exists a $\tilde{V} \in \mathcal{O}$ such that $\tilde{V}^T X \tilde{V} = \Lambda(X)$ and

$$\tilde{U} = \tilde{V} + O(\|\Delta X\|),$$

i.e.,

$$U = \tilde{V} R + O(\|\Delta X\|)$$

because $R^T = R^{-1}$ and $\|R^T\| = \sqrt{n}$. Let $V = \tilde{V}R$. Then

$$V^TV = R^T\tilde{V}^T\tilde{V}R = R^TR = I, \quad \text{and} \quad V^TXV = (\tilde{V}R)^TX\tilde{V}R = R^T\Lambda(X)R$$

is a diagonal matrix. Hence, we have proved $V \in \mathcal{U}(X)$ and for $\Delta X \to 0$

$$U = V + O(\|\Delta X\|).$$

This implies that there exists a $\mu > 0$ such that (32) holds. □

In section 1 we have seen from an example of a two by two matrix that the eigenvalues are not differentiable if X has multiple eigenvalues. This can be easily extended to the general case: $\Lambda(\cdot)$ is not differentiable at X if X has multiple eigenvalues. On the other hand, by [26, pp. 66–68] and [25, Theorem 2.3] we know that if X has distinct eigenvalues, then $\Lambda(\cdot)$ is analytic in a neighborhood of X. Hence, we have the following lemma.

LEMMA 4.5. *$\Lambda(\cdot)$ is analytic in a neighborhood of X if and only if $X \in \mathcal{S}$ has distinct eigenvalues.*

Next, we cite a useful formula for the derivative of $\Lambda(X)$ when $X \in \mathcal{S}$ has distinct eigenvalues.

LEMMA 4.6 (see [21, p. 185, Corollary 2.4]). *For any $X \in \mathcal{S}$, if X has distinct eigenvalues, then $\Lambda(\cdot)$ is continuously differentiable at X and for any $\Delta X \in S$*

$$\lambda_i'(X)(\Delta X) = q_i(X)^T\Delta X q_i(X), \quad i = 1, \ldots, n. \tag{33}$$

For any $X \in \mathcal{S}$, let $Q(X) \in \mathcal{O}$ be such that $Q(X)^TXQ(X) = \Lambda(X)$ with $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$. Define

$$q_i(X) = (Q(X))_{\cdot i}, \quad i = 1, \ldots, n.$$

The following result is our main theorem of this section.

THEOREM 4.7. *$\Lambda(\cdot)$ is a strongly semismooth function.*

Proof. By Ky Fan [8] and Mifflin [14], $\Lambda(\cdot)$ is a semismooth function. Thus, we only have to prove (8) with $p = 1$. Let $D_\Lambda = \{Y \in \mathcal{S} |\ Y \text{ has distinct eigenvalues}\}$. By Lemma 4.5, D_Λ is the subset of $\mathcal{S}$ on which Λ is continuously differentiable. Clearly, D_Λ is dense in $\mathcal{S}$.

Suppose that $X_0 \in \mathcal{S}$ is a given matrix. For any $X \in D_\Lambda$, denote $\Delta X = X - X_0$. For $i = 1, \ldots, n$ from $Xq_i(X) = \lambda_i(X)q_i(X)$, we have

$$q_i(X)^TX_0q_i(X) + q_i(X)^T\Delta X q_i(X) = \lambda_i(X). \tag{34}$$

By Lemma 4.3, there exists a $\mu > 0$ such that for any X sufficiently close to X_0 there exists a matrix $Q(X_0) \in \mathcal{O}$ (depending on the choice of X) such that $Q(X_0)^TX_0Q(X_0) = \Lambda(X_0)$ and

$$\|q_i(X) - q_i(X_0)\| \leq \mu||X - X_0||, \tag{35}$$

where $q_i(X_0) := (Q(X_0))_{\cdot i}$, $i = 1, \ldots, n$. Hence, from (34), (35), and the local

Lipschitz continuity of $\Lambda(\cdot)$, for $i = 1, \ldots, n$, $X \in D_\Lambda$, and $\Delta X \to 0$, we have

$$\begin{aligned}
\lambda_i(X) &= q_i(X)^T \left[X_0 q_i(X_0) + X_0(q_i(X) - q_i(X_0))\right] + q_i(X)^T \Delta X q_i(X) \\
&= \lambda_i(X_0) q_i(X)^T q_i(X_0) + q_i(X)^T X \left[q_i(X) - q_i(X_0)\right] + O(\|\Delta X\|^2) \\
&\quad + q_i(X)^T \Delta X q_i(X) \\
&= \lambda_i(X_0) q_i(X)^T q_i(X_0) + \lambda_i(X) q_i(X)^T \left[q_i(X) - q_i(X_0)\right] \\
&\quad + q_i(X)^T \Delta X q_i(X) + O(\|\Delta X\|^2) \\
&= \lambda_i(X_0) q_i(X)^T q_i(X_0) + \left[\lambda_i(X_0) + O(\|\Delta X\|)\right] q_i(X)^T \left[q_i(X) - q_i(X_0)\right] \\
&\quad + q_i(X)^T \Delta X q_i(X) + O(\|\Delta X\|^2) \\
&= \lambda_i(X_0) q_i(X)^T q_i(X) + q_i(X)^T \Delta X q_i(X) + O(\|\Delta X\|^2) \\
&= \lambda_i(X_0) + q_i(X)^T \Delta X q_i(X) + O(\|\Delta X\|^2), \qquad (36)
\end{aligned}$$

which, according to Lemma 4.6, implies

$$\lambda_i(X) - \lambda_i(X_0) - \lambda_i'(X)(\Delta X) = O(\|\Delta X\|)^2, \quad i = 1, \ldots, n.$$

This, together with Theorem 2.2, implies that for $X \to X_0$ and $V \in \partial\Lambda(X)$,

$$\Lambda(X) - \Lambda(X_0) - V(X - X_0) = O(\|X - X_0\|^2).$$

Hence, (8), and therefore the strong semismoothness of $\Lambda(\cdot)$, is proved. □

5. Newton's method for inverse eigenvalue problems. In this section, we shall show how the strong semismoothness of eigenvalues of symmetric matrices can be used to analyze the quadratic convergence of Newton's method for solving IEPs. Unless stated otherwise, $A : \Re^n \to \mathcal{S}$ is assumed to be continuously differentiable everywhere and $F : \Re^n \to \Re^n$ is defined by (2), i.e.,

$$F(c) = \begin{bmatrix} \lambda_1(A(c)) - \lambda_1^* \\ \vdots \\ \lambda_n(A(c)) - \lambda_n^* \end{bmatrix},$$

where $\{\lambda^*\}_{i=1}^n$ are given n numbers and arranged in the decreasing order. Then the IEP is equivalent to finding $c^* \in \Re^n$ such that $F(c^*) = 0$.

For any $c \in \Re^n$, let $\mathcal{Q}(c) \subseteq \mathcal{O}$ be a subset of $\Re^{n\times n}$ such that for any $Q(c) \in \mathcal{Q}(c)$ we have

$$Q(c)^T A(c) Q(c) = \Lambda(A(c))$$

with $\lambda_1(A(c)) \geq \cdots \geq \lambda_n(A(c))$. For any $Q(c) \in \mathcal{Q}(c)$, define

$$q_i(c) = (Q(c))_{\cdot i}, \qquad i = 1, \ldots, n.$$

Let $\partial A(c)/\partial c_j$ be the partial derivative of $A(c)$ with respect to c_j, $j = 1, \ldots, n$. Then for any $c \in \Re^n$

$$\partial_Q F(c) = \partial_B \Lambda(A(c))(A'(c))$$

is well defined. By using Lemmas 4.5 and 4.6 and [24, Theorem 2.5], we have the following result.

PROPOSITION 5.1.

(a) *For any $c \in \Re^n$, $V \in \partial_Q F(c)$ if and only if there exists a $Q(c) \in \mathcal{Q}(c)$ such that*

$$V_{i\cdot} = \left[q_i(c)^T(\partial A(c)/\partial c_1)q_i(c), \ldots, q_i(c)^T(\partial A(c)/\partial c_n)q_i(c)\right]. \tag{37}$$

(b) *If $c \in \Re^n$ is such that $A(c)$ has distinct eigenvalues, then F is continuously differentiable at c and for any $Q(c) \in \mathcal{Q}(c)$*

$$F_i'(c) = \left[q_i(c)^T(\partial A(c)/\partial c_1)q_i(c), \ldots, q_i(c)^T(\partial A(c)/\partial c_n)q_i(c)\right]. \tag{38}$$

Hence, according to Proposition 5.1, a generalized Newton method for solving the IEP can be described as follows.

ALGORITHM 5.1 (a generalized Newton method).

Step 0. *Choose a starting point value c^0. $k := 0$.*

Step 1. *Compute a $Q(c^k) \in \mathcal{Q}(c^k)$ and form $V_k \in \partial_Q F(c^k)$ according to Proposition* 5.1.

Step 2. *Set $c^{k+1} := c^k + \Delta c^k$, where Δc^k is computed by $F(c^k) + V_k \Delta c^k = 0$.*

Step 3. *Replace k by $k+1$ and go to Step* 1.

In the above generalized Newton method, at the kth step one needs to compute eigenvectors $Q(c^k)$ and eigenvalues $\Lambda(A(c^k))$. Once they are computed, $F(c^k)$ and $V^k \in \partial_Q F(c^k)$ can be formulated easily. If $A(c)$ takes form (1) and at each step $A(c^k)$ has distinct eigenvalues, Algorithm 5.1 reduces to the Newton method considered by many authors, e.g., see [15, 10] and references therein.

THEOREM 5.2. *Suppose that F is defined by* (2) *and $F(c^*) = 0$. If all $V \in \partial_Q F(c^*)$ are nonsingular and A' is Lipschitz continuous around c^*, then there exists a neighborhood $N(c^*)$ of c^* such that for any $c^0 \in N(c^*)$ Algorithm* 5.1 *is well defined and the iterates $\{c^k\}$ converge to c^* Q-quadratically.*

Proof. From Theorem 4.7, we know that $\Lambda(\cdot)$ is strongly semismooth everywhere. Hence, by Theorem 2.4 we obtain the conclusion of this theorem. □

Theorem 5.2 contains a very general convergence result for the quadratic convergence of Newton's method for solving IEPs. However, the nonsingularity assumption on $\partial_Q F(c^*)$ is too strong for IEPs when $A(c^*)$ has multiple eigenvalues. To relax this condition, let $S \subseteq \Re^n$ be defined by

$$S = \{c \in \Re^n |\ A(c) \text{ has distinct eigenvalues}\}. \tag{39}$$

Then, by Lemma 4.5 and Proposition 5.1 for any $c \in S$, $F(\cdot)$ is continuously differentiable at c and

$$\partial_B F(c) = \partial_Q F(c) = \partial F(c) = \{F'(c)\}.$$

THEOREM 5.3. *Suppose that F is defined by* (2), *$F(c^*) = 0$, and S is defined by* (39). *If* (i) *for each k, $c^k \in S$ and $c^* \in \mathrm{cl}S$;* (ii) *all $V \in \partial_B|_S F(c^*)$ are nonsingular; and* (iii) *A' is Lipschitz continuous around c^*, then there exists a neighborhood $N(c^*)$ of c^* such that, for any $c^0 \in N(c^*)$, Algorithm* 5.1 *is well defined and the iterates $\{c^k\}$ converge to c^* Q-quadratically.*

Proof. By using Theorems 3.4 and 4.7, we obtain this theorem. □

In Theorem 5.3, we need only the nonsingularity of $\partial_B|_S F(c^*)$ rather than $\partial_Q F(c^*)$. The price to pay is that all the iterates must stay in S, where S is defined by (39).

Since $\Re^n \backslash S$ is usually a null set, this condition is reasonable for IEPs [10, pp. 647–648]. For illustration, let us consider the following IEP with

$$F(c) = \begin{bmatrix} \lambda_1(A(c)) - \lambda_1^* \\ \lambda_2(A(c)) - \lambda_2^* \end{bmatrix},$$

$$A(c) = c_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

and $\lambda_1^* = \lambda_2^* = 1$. Then, $\lambda_1(A(c)) = c_1 + |c_2|$, $\lambda_2(A(c)) = c_1 - |c_2|$, and $S = \{c \in \Re^2 |\ c_2 \neq 0\}$. The function F has a unique solution at $c^* = (1, 0)$. Note that $A(c^*)$ has a multiple eigenvalues at c^* and

$$\partial_B|_S F(c^*) = \left\{ \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right\}.$$

Therefore, all $V \in \partial_B|_S F(c^*)$ are nonsingular.

It was probably Nocedal and Overton [15] who first discussed the quadratic convergence of Newton's method for solving IEPs with multiple eigenvalues. In their proof, a theorem of Rellich [19] on analytic matrix functions was invoked. In [10], by using the eigenprojector, Friedland, Nocedal, and Overton presented a different elegant proof on the quadratic convergence of Newton's method for solving IEPs with multiple eigenvalues. The latter did not use Rellich's theorem. Our results in this paper could be thought of as a generalization of their method I by explicitly exploring the strong semismoothness of the eigenvalue functions.

Before we finish this section, let us consider the generalized inverse eigenvalue problem (GIEP). Let $C : \Re^n \to \mathcal{S}$ and $D : \Re^n \to \mathcal{S}$ be continuously differentiable and $D(c)$ be positive definite whenever $c \in \Omega$, an open subset of $\Re^n$. Given n real numbers $\{\lambda_i^*\}_{i=1}^n$, which are arranged in the decreasing order $\lambda_1^* \geq \cdots \geq \lambda_n^*$, the GIEP is to find a vector $c^* \in \Omega$ such that the symmetric generalized eigenvalue problem $C(c^*)x = \lambda D(c^*)x$ has the prescribed eigenvalues $\lambda_1^*, \ldots, \lambda_n^*$. If $D(c) \equiv I$, then the GIEP is the IEP considered above. It is readily seen that the GIEP can be converted into the form of solving $F(c) = 0$ with

$$F(c) = \begin{bmatrix} \lambda_1(A(c)) - \lambda_1^* \\ \vdots \\ \lambda_n(A(c)) - \lambda_n^* \end{bmatrix}, \quad c \in \Omega, \tag{40}$$

where $A(c) = D(c)^{-\frac{1}{2}} C(c) D(c)^{-\frac{1}{2}}$.

Dai and Lancaster [7] and Dai [6] considered a special case of the GIEP, i.e., $C(c)$ and $D(c)$ are defined by

$$C(c) = C_0 + \sum_{i=1}^n c_i C_i, \qquad D(c) = D_0 + \sum_{i=1}^n c_i D_i, \tag{41}$$

where $C_0, C_1, \ldots, C_n, D_0, D_1, \ldots D_n \in \mathcal{S}$ and $D(c)$ is positive definite whenever $c \in \Omega$.

When $C(c)$ and $D(c)$ take the form (41), Dai and Lancaster [7] proposed the following Newton method for solving the GIEP.

ALGORITHM 5.2 (a Newton method of Dai and Lancaster [7]).

Step 0. *Choose a starting point value* c^0. $k := 0$.

Step 1. *Compute* $C(c^k) = C_0 + \sum_{j=1}^n c_j^k C_j$, $D(c^k) = D_0 + \sum_{j=1}^n c_j^k D_j$.
Step 2. *Set* $c^{k+1} := c^k + \Delta c^k$, *where* Δc^k *is computed by* $F(c^k) + F'(c^k)\Delta c^k = 0$.
Step 3. *Replace* k *by* $k+1$ *and go to Step* 1.

The following theorem gives an affirmative answer to a conjecture made in [7, p. 11] on the quadratic convergence of Algorithm 5.2, which was supported by numerical experiments.

THEOREM 5.4. *Suppose that* $c^* \in \Omega$ *such that* $F(c^*) = 0$. *If* (i) *for each* k, $A(c^k)$ *has distinct eigenvalues and* $F'(c^k)$ *is invertible; and* (ii) $\limsup_{k\to\infty} \|F'(c^k)^{-1}\| < \infty$, *then there exists a neighborhood* $N(c^*)$ *of* c^* *such that for any* $c^0 \in N(c^*)$ *the iterates* $\{c^k\}$ *generated by Algorithm* 5.2 *converge to* c^* Q*-quadratically.*

Proof. Since $A(c^k)$ has distinct eigenvalues, F is continuously differentiable at c^k. Note that Algorithm 5.2 is a special case of Algorithm 5.1. By using Theorems 4.7 and 3.4 with $S = \{c^0, c^1, \ldots\}$, we get the conclusion of the theorem. □

6. Summary and possible future research topics. In this paper we review basic concepts of semismoothness and Newton's method for semismooth equations. We show the strong semismoothness of eigenvalues of symmetric matrices and demonstrate how this result can be used to provide a unified analysis for the quadratic convergence of the Newton-type methods for IEPs and GIEPs.

We feel that several topics could be further investigated. First, it would be interesting to look at the strong semismoothness of the functions arising from other IEPs, e.g., the least square IEPs [10]. Second, we could develop nonsmooth quasi-Newton [22] methods, rather than Newton's method, for IEPs and GIEPs. Chan and Tseng [3] provided such an approach for IEPs with distinct eigenvalues. The problem is still unsolved in the case of multiple eigenvalues. Third, it is desirable to have a "smoothing" version of the Newton method discussed in this paper; namely, we find a parameterized function $H(\varepsilon, x)$ for a strongly semismooth function $F(x)$ such that $H(\varepsilon, y) \to F(x)$ as $(\varepsilon, y) \to (0^+, x)$ and that $H(\varepsilon, x)$ is differentiable for $\varepsilon \neq 0$. It is proved in [23] that any nonsmooth function has approximate smoothing functions, but the proof does not give any concrete smoothing functions for IEPs. It is then interesting to ask what smoothing function could be used for IEPs.

Acknowledgment. The authors are grateful to the referees for their very constructive comments.

REFERENCES

[1] R. BHATIA, *Matrix Analysis*, Springer–Verlag, New York, 1997.
[2] F. W. BIEGLER-KÖNIG, *A Newton iteration process for inverse eigenvalue problems*, Numer. Math., 37 (1981), pp. 349–354.
[3] R. H. CHAN, S.-F. XU, AND H.-M. ZHOU, *On the convergence rate of a quasi-Newton method for inverse eigenvalue problems*, SIAM J. Numer. Anal., 36 (1999), pp. 436–441.
[4] X. CHEN AND P. TSENG, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., to appear.
[5] F. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[6] H. DAI, *An algorithm for symmetric generalized inverse eigenvalue problems*, Linear Algebra Appl., 296 (1999), pp. 79–98.
[7] H. DAI AND P. LANCASTER, *Newton's method for a generalized inverse eigenvalue problem*, Numer. Linear Algebra Appl., 4 (1997), pp. 1–21.
[8] K. FAN, *On a theorem of Weyl concerning eigenvalues of linear transformations*. I., Proc. Nat. Acad. Sci. U.S.A., 35 (1949), pp. 652–655.
[9] A. FISCHER, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Program., 76 (1997), pp. 513–532.

[10] S. Friedland, J. Nocedal, and M. L. Overton, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24 (1987), pp. 634–667.
[11] P. Lancaster, *On eigenvalues of matrices dependent on a parameter*, Numer. Math., 6 (1964), pp. 377–387.
[12] A. S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24.
[13] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.
[14] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[15] J. Nocedal and M. L. Overton, *Numerical methods for solving inverse eigenvalue problems*, in Lecture Notes in Math. 1005, V. Pereya and A. Reinoza, eds., Springer–Verlag, New York, Berlin, 1983, pp. 212–226.
[16] F. Potra, L. Qi, and D. Sun, *Secant methods for semismooth equations*, Numer. Math., 80 (1998), pp. 305–304.
[17] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[18] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[19] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, London, Paris, 1969.
[20] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
[21] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[22] D. Sun and J. Han, *Newton and quasi-Newton methods for a class of nonsmooth equations and related problems*, SIAM J. Optim., 7 (1997), pp. 463–480.
[23] D. Sun and L. Qi, *Solving variational inequality problems via smoothing-nonsmooth reformulations*, J. Comput. Appl. Math., 129 (2001), pp. 37–62.
[24] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[25] J.-G. Sun, *Eigenvalues and eigenvectors of a matrix dependent on several parameters*, J. Comput. Math., 3 (1985), pp. 351–364.
[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
[27] S.-F. Xu, *An Introduction to Inverse Algebraic Eigenvalue Problems*, Beijing University Press, Beijing, 1998.

SIAM J. NUMER. ANAL.
Vol. 40, No. 6, pp. 2368–2394

A MODEL FOR TWO COUPLED TURBULENT FLUIDS PART II: NUMERICAL ANALYSIS OF A SPECTRAL DISCRETIZATION*

C. BERNARDI†, T. CHACÓN REBOLLO‡, R. LEWANDOWSKI§, AND F. MURAT†

Abstract. We consider a system of equations that models the stationary flow of two immiscible turbulent fluids on adjacent subdomains. The equations are coupled by nonlinear boundary conditions on the interface which is here a fixed given surface. We propose a spectral discretization of this problem and perform its numerical analysis. The convergence of the method is proven in the two-dimensional case, together with optimal error estimates.

Key words. turbulent fluids, numerical analysis, spectral discretization

AMS subject classifications. 65N35, 76D99

PII. S0036142901385829

1. Introduction.

In this paper, we are interested in the numerical analysis of the spectral discretization of a model for two stationary turbulent fluids coupled by boundary conditions on the interface:

$$
\begin{cases}
-\operatorname{div}\big(\alpha_i(k_i)\,\nabla \boldsymbol{u}_i\big) + \mathbf{grad}\, p_i = \boldsymbol{f}_i & \text{in } \Omega_i,\ 1 \le i \le 2,\\
\operatorname{div} \boldsymbol{u}_i = 0 & \text{in } \Omega_i,\ 1 \le i \le 2,\\
-\operatorname{div}\big(\gamma_i(k_i)\,\nabla k_i\big) = \alpha_i(k_i)\,|\nabla \boldsymbol{u}_i|^2 & \text{in } \Omega_i,\ 1 \le i \le 2,\\
\boldsymbol{u}_i = \mathbf{0} & \text{on } \Gamma_i,\ 1 \le i \le 2,\\
k_i = 0 & \text{on } \Gamma_i,\ 1 \le i \le 2,\\
\alpha_i(k_i)\,\partial_{n_i} \boldsymbol{u}_i - p_i\,\boldsymbol{n}_i + (\boldsymbol{u}_i - \boldsymbol{u}_j)\,|\boldsymbol{u}_i - \boldsymbol{u}_j| = \mathbf{0} & \text{on } \Gamma,\ 1 \le i \ne j \le 2,\\
k_i = |\boldsymbol{u}_1 - \boldsymbol{u}_2|^2 & \text{on } \Gamma,\ 1 \le i \le 2,
\end{cases}
\tag{1.1}
$$

where each triple $(\boldsymbol{u}_i, k_i, p_i)$ is defined in the domain Ω_i, $1 \le i \le 2$. The vector field $\boldsymbol{u}_i$ represents the velocity of a turbulent fluid in Ω_i, p_i represents its pressure, and k_i represents its turbulent kinetic energy (TKE in what follows). The domains Ω_i are two- or three-dimensional bounded open sets with common interface Γ, while each Γ_i stands for $\partial\Omega_i \setminus \Gamma$.

*Received by the editors March 2, 2001; accepted for publication (in revised form) June 28, 2002; published electronically February 6, 2003.
http://www.siam.org/journals/sinum/40-6/38582.html

†Laboratoire Jacques-Louis Lions (U.M.R. 7598), C.N.R.S. & Université Pierre et Marie Curie, boîte 187, 4 place Jussieu, 75252 Paris Cedex 05, France (bernardi@ann.jussieu.fr, murat@ann.jussieu.fr). The research of the fourth author was partially supported by Iberdrola Visiting Professors Programme.

‡Departamento de Ecuaciones Diferenciales y Análisis Numerico, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain (chacon@numer.us.es). The research of this author was partially supported by Spanish Government grant REN2000–1168–C02–01.

§IRMAR (U.M.R. 6625), Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex 03, France (lewandow@maths.univ-rennes1.fr).

System (1.1) is motivated by the coupling of two turbulent fluids F_i, $i = 1$ and 2, which appears in the framework ocean/atmosphere or in the case of two layers of a stratified fluid (see, e.g., [16, Chaps. 1 and 3] or [18]). Note that, in these situations, the operator $-\mathrm{div}\left(\alpha_i(k_i)\,\nabla\cdot\right)$ in (1.1) should be replaced by a different one, derived from the deformation rate tensor [11, sect. 2]. However, this change leads to more technical proofs, involving additional Korn-type inequalities, and we prefer to avoid it for simplicity of the presentation. These fluids F_i are coupled through the interface condition on their common boundary Γ, which is supposed to be fixed. Indeed, we assume that the so-called "rigid lid hypothesis" holds, which is standard in geophysics and oceanography. According to this assumption, Γ is a fixed mean interface and in fact the values of $\boldsymbol{u}_i$, p_i, and k_i on Γ are mean values of the velocity, pressure, and TKE. This law characterizes mean momentum exchanges between the fluids (see [16, Chap. 1] and [1]), and it is derived in a rather different way from standard wall laws [21] (but the mathematical formulation is rather similar): the turbulent mixed layer of the two turbulent fluids is modelled by the sixth and seventh lines in (1.1) which summarize the information related to a realistic interface ocean/atmosphere (see, e.g., [16, sect. 1.4] for more details about this model). Slightly more realistic models, obtained, for instance, by adding the convection term $\boldsymbol{u}_i \cdot \nabla \boldsymbol{u}_i$ in the first line of problem (1.1) and/or the dissipative term $-\frac{1}{L}\, k_i^{\frac{3}{2}}$ (where L represents the mixing length) in the right-hand side of the third line of this problem, can also be considered. Since their analysis relies on exactly the same arguments as for problem (1.1), we skip these further terms for brevity.

The analysis of problem (1.1) is performed in [3] for two- or three-dimensional domains Ω_i which are either convex or of class $\mathsf{C}^{1,1}$. In that paper, an equivalent variational formulation of problem (1.1) is written, where the equations for the TKE are taken in the transposition sense (see [23] and [17, Chap. 2, sect. 6] for the definition of a solution by tranposition). Indeed, due to the lack of regularity of the right-hand side in the third line of (1.1) which belongs only to $L^1(\Omega_i)$, a standard formulation cannot be used here. However, the present formulation by transposition allows one to derive a priori estimates. Next the existence of a solution is proved. The uniqueness of smooth solutions is also established under some rather restrictive assumptions on the parameters and the data, and some regularity properties of the solutions are derived when the domains Ω_i are two-dimensional rectangles. Note, moreover, that the transposition formulation of the equations on the TKE is equivalent to the standard variational one when the solution is sufficiently smooth. We also refer to [2] for a slightly different proof of the existence result.

In the present paper, we are interested in the spectral discretization of problem (1.1), which relies on the approximation by high-degree polynomials. For simplicity, we consider only the key geometry where the domains are rectangles or rectangular parallelepipeds. However, in order to take into account the possible anisotropy of the flows which can be induced by the large aspect ratios of the domains, we use different degrees of polynomials with respect to the horizontal and vertical variables. We propose a discrete problem which, as usual for spectral methods [7, Chap. III], relies on the variational formulation of the equations for the velocity, the pressure, and also the TKE: it combines a conforming approximation in these spaces of polynomials with the use of numerical integration relying on tensorized Gauss–Lobatto formulas.

As standard for nonlinear systems, the numerical analysis of the discrete problem is performed via the discrete implicit function theorem of Brezzi, Rappaz, and Raviart [10]. As for the continuous problem, the main difficulty is due to the lack of regularity of the right-hand sides in the discrete TKE equations, and, as far as we

know, the numerical analysis of problems with data in L^1 has been performed only in a few works (see [14], [13], and [12]). Thanks to the Brezzi–Rappaz–Raviart theory, in the two-dimensional case, we derive the existence of a solution of the discrete problem in a neighborhood of a nonsingular exact solution under some reasonable assumptions on its regularity. We also prove the convergence of the method, together with optimal error estimates. The same properties hold in the three-dimensional case; however, we think that the assumptions that are needed to prove them are no longer reasonable. A different analysis, leading to weaker convergence results, is under consideration.

To conclude, we propose an algorithm for solving the discrete problem. Its convergence is currently checked via numerical experiments and is likely at least for small variations of the functions α_i and γ_i.

The numerical analysis of the finite element discretization of system (1.1) is under consideration, and its convergence seems to be likely in the two- and three-dimensional cases under realistic assumptions.

An outline of the paper is as follows.

- In section 2, we recall from [3] the variational formulation and the main properties of problem (1.1). We also write a different formulation in view of the discretization.
- In section 3, we describe the choice of the approximation spaces and the discrete problem. We also write a different and equivalent formulation of this problem, which is needed for its analysis.
- Section 4 is devoted to the numerical analysis of the discrete linear Laplace and Stokes problems with variable coefficients that are involved in the discretization.
- In section 5, we perform the numerical analysis of the coupled system. We prove the existence of a solution and derive error estimates.
- In section 6, we propose some conclusions and present a numerical algorithm for solving the discrete problem in the two-dimensional case.

2. Main properties of the continuous problem. In what follows, Ω_1 and Ω_2 stand for disjoint bounded domains in $\mathbb{R}^d$, $d = 2$ or 3, which are either convex or of class $\mathtt{C}^{1,1}$. The generic point in $\mathbb{R}^2$ (resp., in $\mathbb{R}^3$) is denoted by $\boldsymbol{x} = (x, z)$ (resp., $\boldsymbol{x} = (x, y, z)$). We assume for simplicity that the interface $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ coincides with the intersection of both $\overline{\Omega}_1$ and $\overline{\Omega}_2$ with the hyperplane $z = 0$, while Ω_1 and Ω_2 are contained in the half-spaces $z > 0$ and $z < 0$, respectively. We denote by Γ_i the part of the boundary $\partial\Omega_i \backslash \Gamma$. It must be noted that, in a number of practical situations, the vertical heights of the Ω_i are much smaller than their horizontal diameters.

Throughout the paper, we assume that the functions α_i and γ_i, $1 \le i \le 2$, are continuous and bounded on $\mathbb{R}$, and are continuously differentiable with bounded derivatives. Moreover, we assume that there exists a positive constant ν such that, for $1 \le i \le 2$,

$$\forall k \in \mathbb{R}, \qquad \alpha_i(k) \ge \nu \quad \text{and} \quad \gamma_i(k) \ge \nu. \tag{2.1}$$

We now write a variational formulation of problem (1.1). Next we recall its properties. Finally, we write another formulation of it that relies on the introduction of the Stokes and Laplace operators.

The variational formulation. Throughout the paper, we use the spaces $L^p(\Omega_i)$, $1 \le p \le \infty$, and the Sobolev spaces $H^s(\Omega_i)$ and $H^s_0(\Omega_i)$ for any real number s, provided with the standard norm $\|\cdot\|_{H^s(\Omega_i)}$ and seminorm $|\cdot|_{H^s(\Omega_i)}$, together with their analogues on Γ. We also need the special space $H^{\frac{1}{2}}_{00}(\Gamma)$, defined, e.g., in [17, Chap. 1, Thm. 11.7].

For $1 \le i \le 2$, we introduce the spaces

$$X_i = \big\{ \boldsymbol{v}_i \in H^1(\Omega_i)^d;\ \boldsymbol{v}_i = \boldsymbol{0} \text{ on } \Gamma_i \big\}. \tag{2.2}$$

For reasons explained in [3, sect. 2], we also define the functions G_i, $1 \le i \le 2$, by

$$G_i(k) = \int_0^k \gamma_i(\kappa)\, d\kappa. \tag{2.3}$$

Problem (1.1) can be written (see [2] and [3]) as the following variational system of two coupled problems:

Find $(\boldsymbol{u}_i, p_i)$ *in* $X_i \times L^2(\Omega_i)$, $1 \le i \le 2$, *such that, for* $1 \le i \ne j \le 2$,

$$\begin{aligned} &\forall \boldsymbol{v}_i \in X_i, \quad \int_{\Omega_i} \alpha_i(k_i)\, \nabla \boldsymbol{u}_i \,:\, \nabla \boldsymbol{v}_i \, d\boldsymbol{x} - \int_{\Omega_i} p_i (\operatorname{div} \boldsymbol{v}_i)\, d\boldsymbol{x} \\ &\qquad + \int_\Gamma |\boldsymbol{u}_i - \boldsymbol{u}_j|\, (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{v}_i \, d\tau = \int_{\Omega_i} \boldsymbol{f}_i \cdot \boldsymbol{v}_i \, d\boldsymbol{x}, \\ &\forall q_i \in L^2(\Omega_i), \quad - \int_{\Omega_i} q_i (\operatorname{div} \boldsymbol{u}_i)\, d\boldsymbol{x} = 0; \end{aligned} \tag{2.4}$$

Find k_i *in* $L^2(\Omega_i)$, $1 \le i \le 2$, *such that, for* $1 \le i \le 2$,

$$\begin{aligned} &\forall \varphi_i \in H^2(\Omega_i) \cap H_0^1(\Omega_i), \\ &- \int_{\Omega_i} G_i(k_i)\, \Delta \varphi_i \, d\boldsymbol{x} = - \int_\Gamma G_i(|\boldsymbol{u}_1 - \boldsymbol{u}_2|^2)\, \partial_{n_i} \varphi_i \, d\tau \\ &\qquad + \int_{\Omega_i} \alpha_i(k_i)\, |\nabla \boldsymbol{u}_i|^2\, \varphi_i \, d\boldsymbol{x}. \end{aligned} \tag{2.5}$$

Note that the equations for the velocities and the pressure are of standard variational type and involve the bilinear forms, for $1 \le i \le 2$,

$$a_i(t_i; \boldsymbol{u}_i, \boldsymbol{v}_i) = \int_{\Omega_i} \alpha_i(t_i)\, \nabla \boldsymbol{u}_i \,:\, \nabla \boldsymbol{v}_i \, d\boldsymbol{x}, \quad b_i(\boldsymbol{v}_i, q_i) = - \int_{\Omega_i} q_i (\operatorname{div} \boldsymbol{v}_i)\, d\boldsymbol{x}. \tag{2.6}$$

However, the equation on the TKE is formulated in the transposition sense of Stampacchia [23] and of Lions and Magenes [17, Chap. 2, sect. 6].

As standard for the Stokes problem, we consider the kernel

$$V_i = \big\{ \boldsymbol{v}_i \in X_i;\ \operatorname{div} \boldsymbol{v}_i = 0 \text{ in } \Omega_i \big\},$$

and we observe that, for each solution $(\boldsymbol{u}_i, p_i)$ of problem (2.4), the velocity $\boldsymbol{u}_i$ is a solution of the following problem:

Find $\boldsymbol{u}_i$ *in* V_i, $1 \le i \le 2$, *such that, for* $1 \le i \ne j \le 2$,

$$\begin{aligned} &\forall \boldsymbol{v}_i \in V_i, \quad \int_{\Omega_i} \alpha_i(k_i)\, \nabla \boldsymbol{u}_i \,:\, \nabla \boldsymbol{v}_i \, d\boldsymbol{x} \\ &\qquad + \int_\Gamma |\boldsymbol{u}_i - \boldsymbol{u}_j|\, (\boldsymbol{u}_i - \boldsymbol{u}_j) \cdot \boldsymbol{v}_i \, d\tau = \int_{\Omega_i} \boldsymbol{f}_i \cdot \boldsymbol{v}_i \, d\boldsymbol{x}. \end{aligned} \tag{2.7}$$

Conversely, we recall from [3, Lem. 3.1] that, for $1 \le i \le 2$, there exists a positive constant β_i such that the following inf-sup condition holds:

$$\forall q_i \in L^2(\Omega_i), \quad \sup_{\boldsymbol{v}_i \in X_i} \frac{b_i(\boldsymbol{v}_i, q_i)}{\|\boldsymbol{v}_i\|_{H^1(\Omega_i)^d}} \ge \beta_i \, \|q_i\|_{L^2(\Omega_i)}. \tag{2.8}$$

This yields, for any solution $\boldsymbol{u}_i$ of problem (2.7), the existence of a unique function p_i in $L^2(\Omega_i)$ such that the pair $(\boldsymbol{u}_i, p_i)$ is a solution of problem (2.4). So, for the next results, we work with the simpler system (2.5)–(2.7).

Main properties. We first recall from [3, Lems. 3.3 and 4.2] the following a priori estimates: for any $\boldsymbol{f}_i$ in $L^2(\Omega_i)^d$, $1 \le i \le 2$, every solution $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ of problem (2.7) satisfies

$$\|\boldsymbol{u}_1\|_{H^1(\Omega_1)^d} + \|\boldsymbol{u}_2\|_{H^1(\Omega_2)^d} \le \frac{c}{\nu}\left(\|\boldsymbol{f}_1\|_{L^2(\Omega_1)^d} + \|\boldsymbol{f}_2\|_{L^2(\Omega_2)^d}\right), \tag{2.9}$$

and, for any real number s, $0 \le s < \frac{1}{2}$, and for $1 \le i \le 2$, every solution ℓ_i of problem (2.5) satisfies

$$\|\ell_i\|_{H^s(\Omega_i)} \le c_s\left(\|\boldsymbol{u}_1\|^2_{H^1(\Omega_1)^d} + \|\boldsymbol{u}_2\|^2_{H^1(\Omega_2)^d}\right). \tag{2.10}$$

The constants c and c_s depend on the geometry of Ω, on ν, and on the maximal value of the α_i and γ_i; moreover, the constant c_s depends on s.

Using these estimates, an existence result is proved in [3, Cor. 5.3]. We only state it.

THEOREM 2.1. *For any $\boldsymbol{f}_i$ in $L^2(\Omega_i)^d$, $1 \le i \le 2$, system (2.4)–(2.5) has a solution $(\tilde{U}_1, \tilde{U}_2)$ with each $\tilde{U}_i = (\boldsymbol{u}_i, p_i, k_i)$ in $X_i \times L^2(\Omega_i) \times L^2(\Omega_i)$. Moreover, each function k_i, $i = 1$ and 2, is nonnegative and belongs to $H^s(\Omega_i)$ for all $s < \frac{1}{2}$.*

In contrast, the uniqueness result in [3] (see also [11] for a similar result) is rather disappointing. It states that, if system (2.4)–(2.5) admits a solution $(\boldsymbol{u}_i, p_i, k_i)_{1\le i\le 2}$ such that each $\boldsymbol{u}_i$ belongs to $W^{1,p}(\Omega_i)^d$ for some $p > 2d$, and if its norm in this space is small enough with respect to the relative variation of the α_i, then this solution $(\boldsymbol{u}_i, p_i, k_i)_{1\le i\le 2}$ is unique. So our idea is to give up making any uniqueness assumption for the analysis of the discretization.

Finally, let us recall the regularity property of the solution which is proved in [3, Thm. 7.5] when the domains Ω_1 and Ω_2 are two-dimensional rectangles: let $(\tilde{U}_1, \tilde{U}_2)$ be any solution of system (2.4)–(2.5), with $\tilde{U}_i = (\boldsymbol{u}_i, p_i, k_i)$, such that $\boldsymbol{u}_i$, $i = 1$ and 2, belong to $H^{s_-}(\Omega_i)^2$ for some $s_- > 1$; then, this solution satisfies

$$\tilde{U}_i \in H^s(\Omega_i)^d \times H^{s-1}(\Omega_i) \times H^s(\Omega_i), \qquad i = 1 \text{ and } 2,$$

for all $s \le s_0 \simeq 1.5946$, where the value of s_0 is derived from [20, Cor. 4.2]. So the following assumption seems reasonable, especially in dimension $d = 2$.

Hypothesis 2.2. System (2.4) and (2.5) admits a solution $(\tilde{U}_1^*, \tilde{U}_2^*)$ such that each $\tilde{U}_i^*$, $1 \le i \le 2$, belongs to $H^{s^*}(\Omega_i)^d \times H^{s^*-1}(\Omega_i) \times H^{s^*}(\Omega_i)$ for some $s^* > \frac{d}{2}$.

Remark 2.3. Assume that the functions $\boldsymbol{u}_i$, $i = 1$ and 2, belong to $H^s(\Omega_i)^d$, for some $s > \frac{d}{2}$. If a solution k_i of problem (2.5) belongs to $H^1(\Omega_i)$, then it satisfies the more standard formulation of this problem:

Find k_i in $H^1(\Omega_i)$, $1 \le i \le 2$, with

$$k_i = 0 \quad \text{on } \Gamma_i \qquad \text{and} \qquad k_i = |\boldsymbol{u}_1 - \boldsymbol{u}_2|^2 \quad \text{on } \Gamma,$$

such that, for $1 \le i \le 2$,

$$\forall \varphi_i \in H^1_0(\Omega_i), \quad c_i(k_i; k_i, \varphi_i) = \int_{\Omega_i} \alpha_i(k_i)\,|\nabla \boldsymbol{u}_i|^2\,\varphi_i\,d\boldsymbol{x}, \tag{2.11}$$

where the bilinear form $c_i(t_i; \cdot, \cdot)$ is defined by

$$c_i(t_i; \ell_i, \varphi_i) = \int_{\Omega_i} \gamma_i(t_i)\,\nabla \ell_i \cdot \nabla \varphi_i\,d\boldsymbol{x}. \tag{2.12}$$

The discretization below relies on this last formulation. However, for technical reasons, we consider in what follows its extension to the case where k_i is sought for in $H^{1-\varepsilon}(\Omega_i)$ and φ_i runs through $H^{1+\varepsilon}(\Omega_i)$ for a small positive ε.

Another presentation. For $1 \le i \le 2$, we first introduce a generalized Laplace operator, which we still denote by $\mathcal{L}_i$ for simplicity: for a fixed t_i in $L^1(\Omega_i)$, the operator $\mathcal{L}_i(t_i)$ associates with any g_i in $L^1(\Omega_i)$ and λ_i in $L^2(\Gamma)$ the solution $k_i = \mathcal{L}_i(t_i)(g_i, \lambda_i)$ in $H^s(\Omega_i)$, $s < \frac{1}{2}$, defined by transposition, of the problem

$$\begin{cases} -\operatorname{div}\left(\gamma_i(t_i)\,\nabla k_i\right) = g_i & \text{in } \Omega_i, \\ k_i = 0 & \text{on } \Gamma_i, \\ k_i = \lambda_i & \text{on } \Gamma,\ 1 \le i \le 2. \end{cases} \tag{2.13}$$

The existence and uniqueness of this solution are checked for instance in [3, sect. 4]. Similarly, we introduce the Stokes operator $\mathcal{S}_i$: for a fixed t_i in $L^1(\Omega_i)$, the operator $\mathcal{S}_i(t_i)$ associates with any $\boldsymbol{g}_i$ in the dual space of X_i and $\boldsymbol{\lambda}_i$ in the dual space of $H^{\frac{1}{2}}_{00}(\Gamma)$ the solution $\boldsymbol{u}_i = \mathcal{S}_i(t_i)(\boldsymbol{g}_i, \boldsymbol{\lambda}_i)$ in V_i of the Stokes problem

$$\begin{cases} -\operatorname{div}\left(\alpha_i(t_i)\,\nabla \boldsymbol{u}_i\right) + \mathbf{grad}\, p_i = \boldsymbol{g}_i & \text{in } \Omega_i, \\ \operatorname{div} \boldsymbol{u}_i = 0 & \text{in } \Omega_i, \\ \boldsymbol{u}_i = \mathbf{0} & \text{on } \Gamma_i, \\ \alpha_i(t_i)\,\partial_{n_i}\boldsymbol{u}_i - p_i\,\boldsymbol{n}_i = \boldsymbol{\lambda}_i & \text{on } \Gamma. \end{cases} \tag{2.14}$$

Next it is readily checked that problem (1.1) can be written as

$$\begin{pmatrix} \boldsymbol{u}_1 \\ k_1 \\ \boldsymbol{u}_2 \\ k_2 \end{pmatrix} + \begin{pmatrix} \mathcal{S}_1(k_1) & 0 & 0 & 0 \\ 0 & \mathcal{L}_1(k_1) & 0 & 0 \\ 0 & 0 & \mathcal{S}_2(k_2) & 0 \\ 0 & 0 & 0 & \mathcal{L}_2(k_2) \end{pmatrix} \begin{pmatrix} (-\boldsymbol{f}_1, \boldsymbol{\lambda}_1(\boldsymbol{u}_1, \boldsymbol{u}_2)) \\ (-g_1(k_1, \boldsymbol{u}_1), \lambda(\boldsymbol{u}_1, \boldsymbol{u}_2)) \\ (-\boldsymbol{f}_2, \boldsymbol{\lambda}_2(\boldsymbol{u}_1, \boldsymbol{u}_2)) \\ (-g_2(k_2, \boldsymbol{u}_2), \lambda(\boldsymbol{u}_1, \boldsymbol{u}_2)) \end{pmatrix} = 0, \tag{2.15}$$

with

$$\boldsymbol{\lambda}_i(\boldsymbol{u}_1, \boldsymbol{u}_2) = (\boldsymbol{u}_i - \boldsymbol{u}_j)\,|\boldsymbol{u}_i - \boldsymbol{u}_j|, \quad g_i(k_i, \boldsymbol{u}_i) = \alpha_i(k_i)\,|\nabla \boldsymbol{u}_i|^2, \qquad \lambda(\boldsymbol{u}_1, \boldsymbol{u}_2) = -|\boldsymbol{u}_1 - \boldsymbol{u}_2|^2. \tag{2.16}$$

Let $\mathcal{T}(k_1, k_2)$ denote the diagonal matrix made of the operators $\mathcal{S}_i(k_i)$ and $\mathcal{L}_i(k_i)$ that appears in (2.15), and let $\mathcal{G}(U_1, U_2)$ stand for the last vector in this formula. For technical reasons, we introduce a small parameter ε, $0 < \varepsilon < \frac{1}{2}$, and we consider the spaces

$$\mathcal{X}_i = X_i \times H^{1-\varepsilon}(\Omega_i), \qquad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2. \tag{2.17}$$

Then problem (1.1) is equivalent to finding a solution (U_1, U_2) in $\mathcal{X}$ of the equation

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} + \mathcal{T}(k_1, k_2)\,\mathcal{G}(U_1, U_2) = 0. \tag{2.18}$$

In view of [10], we work with a solution (U_1^*, U_2^*) of (2.18) which satisfies the following hypothesis, as usual for the discretization of nonlinear problems.

Hypothesis 2.4. The solution (U_1^*, U_2^*), of system (2.5)–(2.7), with each $U_i^* = (\boldsymbol{u}_i^*, k_i^*)$, is such that the operator

$$\mathrm{Id} + D\mathcal{T}(k_1^*, k_2^*)\mathcal{G}(U_1^*, U_2^*) + \mathcal{T}(k_1^*, k_2^*)D\mathcal{G}(U_1^*, U_2^*) \tag{2.19}$$

(where D stands for the differential operator) is an isomorphism of $\mathcal{X}$.

The idea is that the conditions for the global uniqueness of the solution (U_1^*, U_2^*), if they exist, are most often too restrictive (see [3, Thm. 6.3]). Hypothesis 2.4 ensures only the local uniqueness of the solution, which is much weaker. Indeed the analogous assumption for the standard Navier–Stokes equations is often used for the numerical analysis of the discretization and is not at all restrictive. Note that Hypothesis 2.4 is equivalent to the well-posedness of the linearized system for any data $(\boldsymbol{g}_i, \boldsymbol{\lambda}_i)$ in the dual space of $X_i \times H_{00}^{\frac{1}{2}}(\Gamma)$ and (g_i, λ_i) in $H^{-1-\varepsilon}(\Omega_i) \times H^{\frac{1}{2}-\varepsilon}(\Gamma)$:

Find $(\boldsymbol{w}_i, r_i)$ *in* $X_i \times L^2(\Omega_i)$, $1 \le i \le 2$, *such that, for* $1 \le i \ne j \le 2$,

$$\begin{aligned}
&\forall \boldsymbol{v}_i \in X_i,\\
&\int_{\Omega_i} \alpha_i(k_i^*)\,\nabla \boldsymbol{w}_i : \nabla \boldsymbol{v}_i\,d\boldsymbol{x} + \int_{\Omega_i} \alpha_i'(k_i^*)\ell_i\,\nabla \boldsymbol{u}_i^* : \nabla \boldsymbol{v}_i\,d\boldsymbol{x} - \int_{\Omega_i} r_i(\operatorname{div} \boldsymbol{v}_i)\,d\boldsymbol{x}\\
&+ \int_\Gamma \frac{(\boldsymbol{u}_i^* - \boldsymbol{u}_j^*)\cdot(\boldsymbol{w}_i - \boldsymbol{w}_j)}{|\boldsymbol{u}_i^* - \boldsymbol{u}_j^*|}\,(\boldsymbol{u}_i^* - \boldsymbol{u}_j^*)\cdot \boldsymbol{v}_i\,d\tau + \int_\Gamma |\boldsymbol{u}_i^* - \boldsymbol{u}_j^*|\,(\boldsymbol{w}_i - \boldsymbol{w}_j)\cdot \boldsymbol{v}_i\,d\tau\\
&\qquad = \int_{\Omega_i} \boldsymbol{g}_i \cdot \boldsymbol{v}_i\,d\boldsymbol{x} + \int_\Gamma \boldsymbol{\lambda}_i \cdot \boldsymbol{v}_i\,d\tau,
\end{aligned} \tag{2.20}$$

$$\forall q_i \in L^2(\Omega_i), \quad -\int_{\Omega_i} q_i(\operatorname{div} \boldsymbol{w}_i)\,d\boldsymbol{x} = 0;$$

Find ℓ_i *in* $H^{1-\varepsilon}(\Omega_i)$, $1 \le i \le 2$, *with*

$$\ell_i = 0 \quad \text{on } \Gamma_i \qquad \text{and} \qquad \ell_i = \lambda_i + 2\,(\boldsymbol{u}_1^* - \boldsymbol{u}_2^*)\cdot(\boldsymbol{w}_1 - \boldsymbol{w}_2) \quad \text{on } \Gamma,$$

such that, for $1 \le i \le 2$,

$$\begin{aligned}
&\forall \varphi_i \in H_0^{1+\varepsilon}(\Omega_i),\\
&\int_{\Omega_i} \gamma_i(k_i^*)\nabla \ell_i \cdot \nabla \varphi_i\,d\boldsymbol{x} + \int_{\Omega_i} \gamma_i'(k_i^*)\ell_i\,\nabla k_i^* \cdot \nabla \varphi_i\,d\boldsymbol{x} = \int_{\Omega_i} g_i\,\varphi_i\,d\boldsymbol{x}\\
&\qquad + 2\int_{\Omega_i} \alpha_i(k_i^*)\,\nabla \boldsymbol{u}_i^* : \nabla \boldsymbol{w}_i\,\varphi_i\,d\boldsymbol{x} + \int_{\Omega_i} \alpha_i'(k_i^*)\ell_i\,|\nabla \boldsymbol{u}_i^*|^2\,\varphi_i\,d\boldsymbol{x}.
\end{aligned} \tag{2.21}$$

Even if this is nothing but a linear problem, writing it is rather technical.

In what follows, we always assume that Hypotheses 2.2 and 2.4 hold. These assumptions are nearly realistic and seem necessary for proving the convergence of any type of discretization.

3. Description of the discrete problem. From now on, we assume that the Ω_i are rectangles in the case $d = 2$, and rectangular parallelepipeds in the case $d = 3$. More precisely, as illustrated in Figure 1, by an appropriate scaling, we take Ω_1 (resp., Ω_2) equal to $]-1, 1[^{d-1} \times]0, h_1[$ (resp., $]-1, 1[^{d-1} \times]-h_2, 0[$), where the h_i are positive real numbers. As already said, the h_i are often small in practical situations.

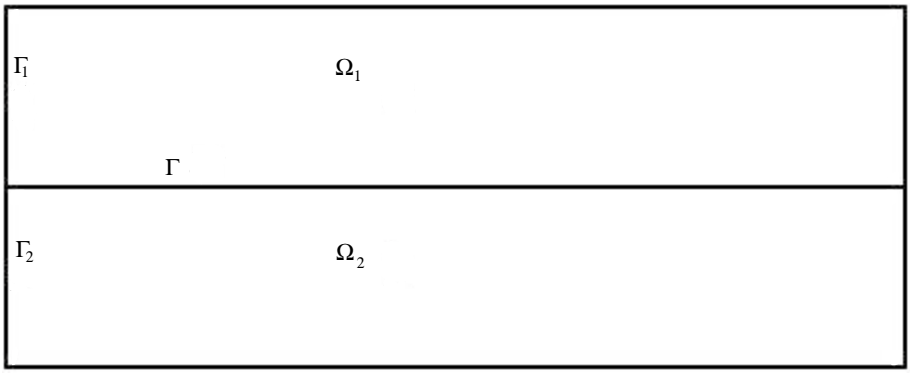


FIG. 1.

We first describe the discrete problem. Second, as for the continuous problem, we write it in a different form, in order to apply the theory of Brezzi, Rappaz, and Raviart [10] for its numerical analysis.

The discrete problem. For each pair of nonnegative integers (k, n), we introduce the space $\mathbb{P}_{k,n}(\Omega_i)$ of restrictions to Ω_i of polynomials with degree $\leq k$ with respect to x (and also to y in the case $d = 3$) and with degree $\leq n$ with respect to z. We denote by $\mathbb{P}_k(\Gamma)$ the space of restrictions to Γ of polynomials with degree $\leq k$ with respect to each tangential variable. We fix a 4-tuple $\delta = (K_1, N_1, K_2, N_2)$ of positive integers, in order to define the discrete spaces of velocities and turbulent energies,

$$X_{i\delta} = \mathbb{P}_{K_i,N_i}(\Omega_i)^d \cap X_i, \qquad Y_{i\delta} = \mathbb{P}_{K_i,N_i}(\Omega_i) \cap H_0^1(\Omega_i). \tag{3.1}$$

As for the standard Stokes problem, two different choices exist for the discrete spaces of pressures $M_{i\delta}$, namely

$$M_{i\delta}^1 = \mathbb{P}_{K_i-2,N_i-2}(\Omega_i) \quad \text{and} \quad M_{i\delta}^2 = \mathbb{P}_{K_i-2,N_i-2}(\Omega_i) \cap \mathbb{P}_{[\lambda K_i],[\lambda N_i]}(\Omega_i), \tag{3.2}$$

for a parameter λ, $0 < \lambda < 1$, where the brackets $[\cdot]$ denote the integral part.

We denote by $(L_n)_{n\geq 0}$ the orthogonal basis of $L^2(-1, 1)$ made by the Legendre polynomials. Each L_n has degree n and satisfies $L_n(1) = 1$. For any positive integer n, let ξ_j^n and ρ_j^n, $0 \leq j \leq n$, be the nodes (in increasing order) and weights of the Gauss–Lobatto formula on $]-1, 1[$, which is exact on all polynomials with degree $\leq 2n - 1$. We recall that ξ_0^n (resp., ξ_n^n) is equal to -1 (resp., 1), that the ξ_j^n, $1 \leq j \leq n-1$, are the zeros of L_n', and that the ρ_j^n are given by

$$\rho_j^n = \frac{2}{n(n+1)\, L_n^2(\xi_j^n)}, \qquad 0 \leq j \leq n. \tag{3.3}$$

For simplicity, we denote by x_{ik} and ρ_{ik}, $0 \leq k \leq K_i$, the nodes $\xi_k^{K_i}$ and weights $\rho_k^{K_i}$. In the z-direction, we set, for $0 \leq j \leq N_i$,

$$z_{ij} = \frac{h_i}{2}\left((-1)^{i+1} + \xi_j^{N_i}\right) \quad \text{and} \quad \omega_{ij} = \frac{h_i}{2}\,\rho_j^{N_i}.$$

We introduce the grids

$$\Xi_{i\delta} = \begin{cases} \left\{(x_{ik}, z_{ij});\ 0 \leq k \leq K_i,\ 0 \leq j \leq N_i\right\} & \text{in the case } d = 2, \\ \left\{(x_{ik}, x_{i\ell}, z_{ij});\ 0 \leq k, \ell \leq K_i,\ 0 \leq j \leq N_i\right\} & \text{in the case } d = 3, \end{cases}$$

and we denote by $\mathcal{I}_{i\delta}$ the Lagrange interpolation operator on the grid $\Xi_{i\delta}$ with values in $\mathbb{P}_{K_i,N_i}(\Omega_i)$. Two different grids are then defined on the interface Γ: we denote by $\mathcal{I}_{i\delta}^{\Gamma}$ the Lagrange interpolation operator on the grid $\Xi_{i\delta}\cap\Gamma$ with values in $\mathbb{P}_{K_i}(\Gamma)$.

Finally, we introduce the discrete product, for all functions u and v continuous on $\overline{\Omega}_i$,

$$\big(u,v\big)_{i\delta}=\begin{cases}\sum_{k=0}^{K_i}\sum_{j=0}^{N_i}u(x_{ik},z_{ij})v(x_{ik},z_{ij})\,\rho_{ik}\omega_{ij} & \text{in the case } d=2,\\ \sum_{k=0}^{K_i}\sum_{\ell=0}^{K_i}\sum_{j=0}^{N_i}u(x_{ik},x_{i\ell},z_{ij})v(x_{ik},x_{i\ell},z_{ij})\,\rho_{ik}\rho_{i\ell}\omega_{ij} & \text{in the case } d=3,\end{cases}$$

and its analogue on Γ

$$\big(u,v\big)_{i\delta}^{\Gamma}=\begin{cases}\sum_{k=0}^{K_i}u(x_{ik})v(x_{ik})\,\rho_{ik} & \text{in the case } d=2,\\ \sum_{k=0}^{K_i}\sum_{\ell=0}^{K_i}u(x_{ik},x_{i\ell})v(x_{ik},x_{i\ell})\,\rho_{ik}\rho_{i\ell} & \text{in the case } d=3.\end{cases}$$

We fix an operator $\Pi_{i\delta}^{\Gamma}$ from $H_{00}^{\frac12}(\Gamma)$ into $\mathbb{P}_{K_i}(\Gamma)\cap H_{00}^{\frac12}(\Gamma)$ which will be made precise later on. We are now in a position to state the discrete problem associated with problem (1.1). It reads as follows:

Find $(\tilde U_{1\delta},\tilde U_{2\delta})$, *with each* $\tilde U_{i\delta}=(\boldsymbol{u}_{i\delta},p_{i\delta},k_{i\delta})$ *in* $X_{i\delta}\times M_{i\delta}\times\mathbb{P}_{K_i,N_i}(\Omega_i)$, *such that, for* $1\le i\ne j\le 2$,

$$k_{i\delta}=0\quad\text{on }\Gamma_i\quad\text{and}\quad k_{i\delta}=\Pi_{i\delta}^{\Gamma}(|\boldsymbol{u}_{1\delta}-\boldsymbol{u}_{2\delta}|^2)\quad\text{on }\Gamma,\tag{3.4}$$

and

$$\begin{aligned}&\forall\boldsymbol{v}_{i\delta}\in X_{i\delta},\\ &\quad a_{i\delta}(k_{i\delta};\boldsymbol{u}_{i\delta},\boldsymbol{v}_{i\delta})+b_{i\delta}(\boldsymbol{v}_{i\delta},p_{i\delta})+\big(\,|\boldsymbol{u}_{i\delta}-\boldsymbol{u}_{j\delta}|\,(\boldsymbol{u}_{i\delta}-\boldsymbol{u}_{j\delta}),\boldsymbol{v}_{i\delta}\big)_{i\delta}^{\Gamma}=\big(\boldsymbol{f}_i,\boldsymbol{v}_{i\delta}\big)_{i\delta},\\ &\forall q_{i\delta}\in M_{i\delta},\quad b_{i\delta}(\boldsymbol{u}_{i\delta},q_{i\delta})=0,\\ &\forall\varphi_{i\delta}\in Y_{i\delta},\quad c_{i\delta}(k_{i\delta};k_{i\delta},\varphi_{i\delta})=\big(\alpha_i(k_{i\delta})\,|\nabla\boldsymbol{u}_{i\delta}|^2,\varphi_{i\delta}\big)_{i\delta},\end{aligned}\tag{3.5}$$

where, for any continuous function t_i, the bilinear forms $a_{i\delta}(t_i;\cdot,\cdot)$, $b_{i\delta}(\cdot,\cdot)$, and $c_{i\delta}(t_i;\cdot,\cdot)$ are now defined by

$$\begin{aligned}a_{i\delta}(t_i;\boldsymbol{u}_{i\delta},\boldsymbol{v}_{i\delta})=\big(\alpha_i(t_i)\,\nabla\boldsymbol{u}_{i\delta},\nabla\boldsymbol{v}_{i\delta}\big)_{i\delta},\quad b_{i\delta}(\boldsymbol{v}_{i\delta},q_{i\delta})=-\big(q_{i\delta},\operatorname{div}\boldsymbol{v}_{i\delta}\big)_{i\delta},\\ c_{i\delta}(t_i;k_{i\delta},\varphi_{i\delta})=\big(\gamma_i(t_i)\,\nabla k_{i\delta},\nabla\varphi_{i\delta}\big)_{i\delta}.\end{aligned}\tag{3.6}$$

Remark 3.1. A natural choice of operator $\Pi_{i\delta}^{\Gamma}$ would be the Lagrange interpolation operator $\mathcal{I}_{i\delta}^{\Gamma}$. However, for $K_1\ne K_2$, since two different discrete products are defined on the interface Γ, the trace of $\boldsymbol{u}_{1\delta}$ on Γ must be re-interpolated on the nodes of $\Xi_{2\delta}\cap\Gamma$ and conversely. Moreover, the convergence of the interpolate of a function φ toward this function in $H_{00}^{\frac12}(\Gamma)$ or even in $H^{\frac12-\varepsilon}(\Gamma)$ would require too much regularity of the function φ; see [7, sect. 14]. Other choices of operator $\Pi_{i\delta}^{\Gamma}$, such as orthogonal projection operators, are possible but seem more expensive to implement.

Remark 3.2. For the choices $M_{i\delta}^1$ and $M_{i\delta}^2$ of discrete pressure spaces introduced in (3.2), and thanks to the exactness property of the quadrature formula, each $b_{i\delta}(\cdot,\cdot)$ can be replaced by $b_i(\cdot,\cdot)$ in formulation (3.5).

The numerical analysis of system (3.4)–(3.5) is rather technical. However, we begin with the same simplication as for the continuous problem. For $i=1$ and 2, we introduce the discrete kernel

$$V_{i\delta}=\big\{\boldsymbol{v}_{i\delta}\in X_{i\delta};\ \forall q_{i\delta}\in M_{i\delta},\ b_{i\delta}(\boldsymbol{v}_{i\delta},q_{i\delta})=0\big\}.$$

Note that, for the two choices $M_{i\delta} = M^1_{i\delta}$ and $M_{i\delta} = M^2_{i\delta}$ proposed in (3.2), $V_{i\delta}$ is not contained in V_i, i.e., is not made of exactly divergence-free polynomials. It is readily checked with this definition that, for each pair $(\tilde{U}_{1\delta}, \tilde{U}_{2\delta})$ solution of system (3.4)–(3.5), the reduced pair $(U_{1\delta}, U_{2\delta})$ of discrete velocities and discrete turbulent energies is a solution of the following system:

Find $(U_{1\delta}, U_{2\delta})$, *with each* $U_{i\delta} = (\boldsymbol{u}_{i\delta}, k_{i\delta})$ *in* $V_{i\delta} \times \mathbb{P}_{K_i,N_i}(\Omega_i)$,
satisfying (3.4) *and such that, for* $1 \le i \ne j \le 2$,

$$
\begin{aligned}
&\forall \boldsymbol{v}_{i\delta} \in V_{i\delta}, \quad a_{i\delta}(k_{i\delta}; \boldsymbol{u}_{i\delta}, \boldsymbol{v}_{i\delta}) + \big(|\boldsymbol{u}_{i\delta} - \boldsymbol{u}_{j\delta}| \, (\boldsymbol{u}_{i\delta} - \boldsymbol{u}_{j\delta}), \boldsymbol{v}_{i\delta} \big)^{\Gamma}_{i\delta} = \big(\boldsymbol{f}_i, \boldsymbol{v}_{i\delta} \big)_{i\delta}, \\
&\forall \varphi_{i\delta} \in Y_{i\delta}, \quad c_{i\delta}(k_{i\delta}; k_{i\delta}, \varphi_{i\delta}) = \big(\alpha_i(k_{i\delta}) \, |\nabla \boldsymbol{u}_{i\delta}|^2, \varphi_{i\delta} \big)_{i\delta}.
\end{aligned} \tag{3.7}
$$

The converse property relies on a discrete inf-sup condition, which is derived in two steps, relying on the arguments in [8] and [9], respectively. For a while, let $\tilde{M}^m_{i\delta}$ stand for the subspace of $M^m_{i\delta}$ made of polynomials with a null integral on Ω_i.

LEMMA 3.3. *For* $i = 1$ *and* 2, *and for the discrete spaces* $\tilde{M}^m_{i\delta}$, $m = 1$ *and* 2, *there exists a constant* $\tilde{\beta}^m_{i\delta} > 0$ *such that*

$$
\forall q_{i\delta} \in \tilde{M}^m_{i\delta}, \qquad \sup_{\boldsymbol{v}_{i\delta} \in X_{i\delta} \cap H^1_0(\Omega_i)^d} \frac{b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta})}{\|\boldsymbol{v}_{i\delta}\|_{H^1(\Omega_i)^d}} \ge \tilde{\beta}^m_{i\delta} \, \|q_{i\delta}\|_{L^2(\Omega_i)}. \tag{3.8}
$$

Moreover, these constants $\tilde{\beta}^m_i$ *satisfy, for* $i = 1$ *and* 2,

$$
\tilde{\beta}^1_i \ge c \, K_i^{\frac{2-d}{2}} \inf\{K_i^{-\frac{1}{2}}, N_i^{-\frac{1}{2}}\} \qquad \textit{and} \qquad \tilde{\beta}^2_i \ge c. \tag{3.9}
$$

Proof. Since any $q_{i\delta}$ in $\tilde{M}_{i\delta}$ has a null integral on Ω_i, there exists [15, Chap. I, Cor. 2.4] a function $\boldsymbol{v}_i$ in $H^1_0(\Omega_i)^d$ such that

$$
\operatorname{div} \boldsymbol{v}_i = q_{i\delta} \quad \text{in } \Omega_i \quad \text{and} \quad \|\boldsymbol{v}_i\|_{H^1(\Omega_i)^d} \le c \, \|q_{i\delta}\|_{L^2(\Omega_i)}.
$$

Next, we recall from [8, Lems. 3.2 and 3.3] that, for any μ, $0 < \mu < 1$, there exists an operator π^μ_n from $H^1_0(-1,1)$ onto $\mathbb{P}_{[(1+\mu)n]}(-1,1) \cap H^1_0(-1,1)$ which preserves all polynomials in $\mathbb{P}_{n-1}(-1,1)$ and satisfies, for all φ in $H^1_0(-1,1)$,

$$
\|(\pi^\mu_n \varphi)'\|_{L^2(-1,1)} \le \|\varphi'\|_{L^2(-1,1)} \quad \text{and} \quad \|\pi^\mu_n \varphi\|_{L^2(-1,1)} \le c \, \mu^{-\frac{1}{2}} \, \|\varphi\|_{L^2(-1,1)}.
$$

The idea consists of choosing the operator $\pi^m_{K_i}$ in the x- or y-direction equal to $\pi^{\mu(K_i)}_M$, with

$$
(1 + \mu(K_i))M = K_i \quad \text{and} \quad M = \begin{cases} K_i - 1 & \text{if } m = 1, \\ [\frac{1+\lambda}{2} \, K_i] & \text{if } m = 2 \end{cases}
$$

(recall that λ is introduced in (3.2)), and denoting them by $\pi^{m(x)}_{K_i}$ and $\pi^{m(y)}_{K_i}$, respectively. Similarly, the operator $\pi^m_{N_i}$ in the z-direction is equal to $\pi^{\mu(N_i)}_M$, with

$$
(1 + \mu(N_i))M = N_i \quad \text{and} \quad M = \begin{cases} N_i - 1 & \text{if } m = 1, \\ [\frac{1+\lambda}{2} \, N_i] & \text{if } m = 2, \end{cases}
$$

and denoted by $\pi^{m(z)}_{N_i}$. Next we set

$$
\boldsymbol{v}^s_{i\delta} = \begin{cases} \pi^{m(x)}_{K_i} \circ \pi^{m(z)}_{N_i} \boldsymbol{v}_i & \text{in the case } d = 2, \\ \pi^{m(x)}_{K_i} \circ \pi^{m(y)}_{K_i} \circ \pi^{m(z)}_{N_i} \boldsymbol{v}_i & \text{in the case } d = 3. \end{cases}
$$

From the properties of these operators, it is readily checked that

$$b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta}) = -\int_{\Omega_i} q_{i\delta}\,(\operatorname{div}\boldsymbol{v}_i)\,d\boldsymbol{x} = \int_{\Omega_{i\delta}} (q_{i\delta})^2\,d\boldsymbol{x},$$

while the norm of $\boldsymbol{v}_{i\delta}$ in $H^1_0(\Omega)^d$ is bounded by

$$\|\boldsymbol{v}_{i\delta}\|_{H^1(\Omega_i)^d} \le c\,\sup\{\mu(K_i)^{\frac{2-d}{2}}\mu(N_i)^{-\frac12}, \mu(K_i)^{\frac{1-d}{2}}\}\,\|\boldsymbol{v}_i\|_{H^1(\Omega_i)^d}.$$

Evaluating the quantities $\mu(K_i)$ and $\mu(N_i)$ as a function of N_i or K_i for $m = 1$ or 2 leads to the desired result.

LEMMA 3.4. *For $i = 1$ and 2, and for the discrete spaces $M^m_{i\delta}$, $m = 1$ and 2, defined in* (3.2)*, there exists a constant $\beta^m_{i\delta} > 0$ such that*

$$\forall q_{i\delta} \in M^m_{i\delta}, \quad \sup_{\boldsymbol{v}_{i\delta}\in X_{i\delta}} \frac{b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta})}{\|\boldsymbol{v}_{i\delta}\|_{H^1(\Omega_i)^d}} \ge \beta^m_{i\delta}\,\|q_{i\delta}\|_{L^2(\Omega_i)}. \tag{3.10}$$

Moreover, these constants $\beta^m_{i\delta}$, $i = 1$ and 2, satisfy (3.9).

Proof. Any function $q_{i\delta}$ in $M^m_{i\delta}$ admits the expansion

$$q_{i\delta} = \tilde q_{i\delta} + \overline{q}_{i\delta}, \qquad \text{with} \quad \overline{q}_{i\delta} = \frac{1}{2^{d-1}\,h_i}\int_{\Omega_i} q_{i\delta}(\boldsymbol{x})\,d\boldsymbol{x}.$$

Since the function $\tilde q_{i\delta}$ belongs to $\tilde M^m_{i\delta}$, it follows from Lemma 3.3 that there exists a $\tilde{\boldsymbol{v}}_{i\delta}$ in $X_{i\delta}\cap H^1_0(\Omega_i)^d$ such that

$$b_{i\delta}(\tilde{\boldsymbol{v}}_{i\delta}, \tilde q_{i\delta}) = \|\tilde q_{i\delta}\|^2_{L^2(\Omega_i)} \quad \text{and} \quad \|\tilde{\boldsymbol{v}}_{i\delta}\|_{H^1(\Omega_i)^d} \le \frac{1}{\tilde\beta^m_{i\delta}}\,\|\tilde q_{i\delta}\|_{L^2(\Omega_i)}.$$

On the other hand, the function $\overline{\boldsymbol{v}}_{i\delta}$ equal to $(0, v_{i\delta})$ in dimension $d = 2$, and to $(0, 0, v_{i\delta})$ in dimension $d = 3$, with

$$v_{i\delta} = \begin{cases} \big(L_0(x) - L_2(x)\big)\,\big((-1)^{i+1}h_i - z\big)\,\overline{q}_{i\delta} & \text{in dimension } d = 2,\\ \big(L_0(x) - L_2(x)\big)\,\big(L_0(y) - L_2(y)\big)\,\big((-1)^{i+1}h_i - z\big)\,\overline{q}_{i\delta} & \text{in dimension } d = 3,\end{cases}$$

belongs to $X_{i\delta}$ and satisfies, for a fixed constant c_0,

$$b_{i\delta}(\overline{\boldsymbol{v}}_{i\delta}, \overline{q}_{i\delta}) = \|\overline{q}_{i\delta}\|^2_{L^2(\Omega_i)} \quad \text{and} \quad \|\overline{\boldsymbol{v}}_{i\delta}\|_{H^1(\Omega_i)^d} \le c_0\,\|\overline{q}_{i\delta}\|_{L^2(\Omega_i)}.$$

Next we take $\boldsymbol{v}_{i\delta}$ equal to $\tilde{\boldsymbol{v}}_{i\delta} + \lambda\,\overline{\boldsymbol{v}}_{i\delta}$ for a fixed constant λ. Indeed, it follows by integration by parts that $b_{i\delta}(\tilde{\boldsymbol{v}}_{i\delta}, \overline{q}_{i\delta})$ vanishes so that

$$b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta}) = b_{i\delta}(\tilde{\boldsymbol{v}}_{i\delta}, \tilde q_{i\delta}) + \lambda\,b_{i\delta}(\overline{\boldsymbol{v}}_{i\delta}, \overline{q}_{i\delta}) + \lambda\,b_{i\delta}(\overline{\boldsymbol{v}}_{i\delta}, \tilde q_{i\delta}).$$

The previous properties, together with the continuity of $b_{i\delta}(\cdot,\cdot)$ (which coincides with $b_i(\cdot,\cdot)$ everywhere in the previous equation), yield

$$\begin{aligned} b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta}) &\ge \|\tilde q_{i\delta}\|^2_{L^2(\Omega_i)} + \lambda\,\|\overline{q}_{i\delta}\|^2_{L^2(\Omega_i)} - c\lambda\,\|\overline{\boldsymbol{v}}_{i\delta}\|_{H^1(\Omega_i)^d}\,\|\tilde q_{i\delta}\|_{L^2(\Omega_i)}\\ &\ge \|\tilde q_{i\delta}\|^2_{L^2(\Omega_i)} + \lambda\,\|\overline{q}_{i\delta}\|^2_{L^2(\Omega_i)} - cc_0\lambda\,\|\tilde q_{i\delta}\|_{L^2(\Omega_i)}\,\|\overline{q}_{i\delta}\|_{L^2(\Omega_i)}, \end{aligned}$$

whence

$$b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta}) \ge \frac12\|\tilde q_{i\delta}\|^2_{L^2(\Omega_i)} + \lambda\Big(1 - \frac{c^2c_0^2\lambda}{2}\Big)\,\|\overline{q}_{i\delta}\|^2_{L^2(\Omega_i)}.$$

We now choose λ equal to $\frac{1}{c^2 c_0^2}$, which gives (note that $\tilde{q}_{i\delta}$ and $\overline{q}_{i\delta}$ are orthogonal in $L^2(\Omega_i)$)

$$b_{i\delta}(\boldsymbol{v}_{i\delta}, q_{i\delta}) \geq \inf\Big\{\frac{1}{2}, \frac{\lambda}{2}\Big\} \|q_{i\delta}\|^2_{L^2(\Omega_i)}.$$

We also have

$$\|\boldsymbol{v}_{i\delta}\|_{H^1(\Omega_i)^d} \leq \|\tilde{\boldsymbol{v}}_{i\delta}\|_{H^1(\Omega_i)^d} + \lambda \|\overline{\boldsymbol{v}}_{i\delta}\|_{H^1(\Omega_i)^d} \leq \left(\Big(\frac{1}{\tilde{\beta}^m_{i\delta}}\Big)^2 + c_0^2\lambda^2\right)^{\frac{1}{2}} \|q_{i\delta}\|_{L^2(\Omega_i)},$$

which concludes the proof.

Remark 3.5. From the previous proofs, the constants $\beta^m_{i\delta}$ given in (3.9) a priori depend on h_i. However, by using the vertical homothety that maps Ω_i onto the reference square or cube, it is readily checked that these constants satisfy

$$\beta^1_i \geq c\, K_i^{\frac{2-d}{2}} \inf\{h_i\, K_i^{-\frac{1}{2}}, N_i^{-\frac{1}{2}}\} \qquad \text{and} \qquad \beta^2_i \geq c\, h_i, \tag{3.11}$$

where c is now independent of h_i.

So we now work with system (3.4)–(3.7). The first idea consists of writing it in a more appropriate form which is the discrete analogue of (2.18).

Another presentation. For $i = 1$ and 2, we introduce the discrete Laplace operator $\mathcal{L}_{i\delta}$. For a fixed continuous function t_i, the operator $\mathcal{L}_{i\delta}(t_i)$ associates with any g_i in $H^{-1}(\Omega_i)$ and any function λ_i in $H^{\frac{1}{2}}_{00}(\Gamma)$, the solution $k_{i\delta} = \mathcal{L}_{i\delta}(t_i)(g_i, \lambda_i)$ of the following problem:

Find $k_{i\delta}$ in $\mathbb{P}_{K_i,N_i}(\Omega_i)$ such that

$$k_{i\delta} = 0 \quad \text{on } \Gamma_i \quad \text{and} \quad k_{i\delta} = \Pi^{\Gamma}_{i\delta}\lambda_i \quad \text{on } \Gamma, \tag{3.12}$$

and

$$\forall \varphi_{i\delta} \in Y_{i\delta}, \quad c_{i\delta}(t_i; k_{i\delta}, \varphi_{i\delta}) = \int_{\Omega_i} g_i\, \varphi_{i\delta}\, d\boldsymbol{x}. \tag{3.13}$$

It follows from (3.3) that the weights ρ_{ik} and ω_{ij} are positive. When combined with (2.1), this yields that the only solution of (3.12)–(3.13) for $g_i = 0$ and $\lambda_i = 0$ is zero. Hence, since problem (3.12)–(3.13) results into a square linear system, the operator $\mathcal{L}_{i\delta}(t_i)$ is well-defined.

Similarly, we introduce the discrete Stokes operator $\mathcal{S}_{i\delta}$. For a fixed continuous function t_i, the operator $\mathcal{S}_{i\delta}(t_i)$ associates with any $\boldsymbol{g}_i$ in the dual space of X_i and $\boldsymbol{\lambda}_i$ in the dual space of $H^{\frac{1}{2}}_{00}(\Gamma)^d$ the solution $\boldsymbol{u}_{i\delta} = \mathcal{S}_{i\delta}(t_i)(\boldsymbol{g}_i, \boldsymbol{\lambda}_i)$ in $V_{i\delta}$ of the following Stokes problem:

Find $\boldsymbol{u}_{i\delta}$ in $V_{i\delta}$ such that

$$\forall \boldsymbol{v}_{i\delta} \in V_{i\delta}, \quad a_{i\delta}(t_i; \boldsymbol{u}_{i\delta}, \boldsymbol{v}_{i\delta}) = \int_{\Omega_i} \boldsymbol{g}_i \cdot \boldsymbol{v}_{i\delta}\, d\boldsymbol{x} + \int_{\Gamma} \boldsymbol{\lambda}_i \cdot \boldsymbol{v}_{i\delta}\, d\tau. \tag{3.14}$$

There also it follows from (3.3) and (2.1) that the operator $\mathcal{S}_{i\delta}(t_i)$ is well-defined.

Finally, the matrix $\mathcal{T}_\delta(t_1, t_2)$ is defined by

$$\mathcal{T}_\delta(t_1, t_2) = \begin{pmatrix} \mathcal{S}_{1\delta}(t_1) & 0 & 0 & 0 \\ 0 & \mathcal{L}_{1\delta}(t_1) & 0 & 0 \\ 0 & 0 & \mathcal{S}_{2\delta}(t_2) & 0 \\ 0 & 0 & 0 & \mathcal{L}_{2\delta}(t_2) \end{pmatrix}. \tag{3.15}$$

We introduce the vector $\mathcal{G}_\delta(U_{1\delta}, U_{2\delta})$,

$$\mathcal{G}_\delta(U_{1\delta}, U_{2\delta}) = \begin{pmatrix} (-\boldsymbol{f}_{1\delta}, \boldsymbol{\lambda}_{1\delta}(\boldsymbol{u}_{1\delta}, \boldsymbol{u}_{2\delta})) \\ (-g_{1\delta}(k_{1\delta}, \boldsymbol{u}_{1\delta}), \lambda(\boldsymbol{u}_{1\delta}, \boldsymbol{u}_{2\delta})) \\ (-\boldsymbol{f}_{2\delta}, \boldsymbol{\lambda}_{2\delta}(\boldsymbol{u}_{1\delta}, \boldsymbol{u}_{2\delta})) \\ (-g_{2\delta}(k_{2\delta}, \boldsymbol{u}_{2\delta}), \lambda(\boldsymbol{u}_{1\delta}, \boldsymbol{u}_{2\delta})) \end{pmatrix}, \tag{3.16}$$

where the functions $\boldsymbol{f}_{i\delta}$, $g_{i\delta}$, and $\boldsymbol{\lambda}_{i\delta}$ are defined by duality, for smooth enough functions $\boldsymbol{v}_i$ and φ_i (we do not make precise the spaces),

$$\begin{gathered} \langle \boldsymbol{f}_{i\delta}, \boldsymbol{v}_i \rangle = \big(\boldsymbol{f}_i, \boldsymbol{v}_i\big)_{i\delta}, \qquad \big\langle \boldsymbol{\lambda}_{i\delta}(\boldsymbol{u}_{1\delta}, \boldsymbol{u}_{2\delta}), \boldsymbol{v}_i \big\rangle = \big(|\boldsymbol{u}_{i\delta} - \boldsymbol{u}_{j\delta}| \, (\boldsymbol{u}_{i\delta} - \boldsymbol{u}_{j\delta}), \boldsymbol{v}_i \big)_{i\delta}^{\Gamma}, \\ \big\langle g_{i\delta}(k_{i\delta}, \boldsymbol{u}_{i\delta}), \varphi_i \big\rangle = \big(\alpha_i(k_{i\delta}) \, |\nabla \boldsymbol{u}_{i\delta}|^2, \varphi_i\big)_{i\delta}. \end{gathered} \tag{3.17}$$

The quantity $\lambda(\boldsymbol{u}_1, \boldsymbol{u}_2)$ is defined in (2.16).

Thus, it is readily checked that problem (3.4)–(3.7) can be equivalently written

$$U_\delta + \mathcal{T}_\delta(k_{1\delta}, k_{2\delta}) \, \mathcal{G}_\delta(U_{1\delta}, U_{2\delta}) = 0, \qquad \text{with} \quad U_\delta = \begin{pmatrix} \boldsymbol{u}_{1\delta} \\ k_{1\delta} \\ \boldsymbol{u}_{2\delta} \\ k_{2\delta} \end{pmatrix}. \tag{3.18}$$

This formulation is fully appropriate for performing its numerical analysis thanks to the theory of Brezzi, Rappaz, and Raviart [10].

4. The discrete Laplace and Stokes operators. As a first step for the numerical analysis of the discrete problem (3.4)–(3.5), we investigate the properties of the discrete quasi-linear operators $\mathcal{L}_{i\delta}$ and $\mathcal{S}_{i\delta}$; more precisely, we prove stability and error estimates. In all that follows, c, c', and c'' stand for generic constants that may vary from one line to the other but are always independent of δ.

The discrete Laplace operator. For $i = 1$ and 2, and for a fixed continuous function t_i, let us first consider the operator $\mathcal{L}_{i\delta}(t_i)$ defined from problem (3.12)–(3.13). In order to prove its stability, we first recall [7, Form. (13.10)] that, for any polynomial φ_n of degree $\le n$ on $]-1, 1[$,

$$\|\varphi_n\|^2_{L^2(-1,1)} \le \sum_{j=0}^{n} \varphi_n^2(\xi_j^n) \, \rho_j^n \le 3 \, \|\varphi_n\|^2_{L^2(-1,1)}. \tag{4.1}$$

Combined with the boundedness and positivity of γ_i (see (2.1)), this obviously yields some basic properties of the form $c_{i\delta}(\cdot, \cdot)$ that we now state.

LEMMA 4.1. *For any continuous function t_i, the form $c_{i\delta}(t_i; \cdot, \cdot)$ satisfies the following properties of continuity:*

$$\begin{gathered} \forall \psi_{i\delta} \in \mathbb{P}_{K_i, N_i}(\Omega_i), \ \forall \varphi_{i\delta} \in \mathbb{P}_{K_i, N_i}(\Omega_i), \\ c_{i\delta}(t_i; \psi_{i\delta}, \varphi_{i\delta}) \le c \, \|\psi_{i\delta}\|_{H^1(\Omega_i)} \|\varphi_{i\delta}\|_{H^1(\Omega_i)}, \end{gathered} \tag{4.2}$$

and of ellipticity

$$\forall \psi_{i\delta} \in Y_{i\delta}, \quad c_{i\delta}(t_i; \psi_{i\delta}, \psi_{i\delta}) \ge c \, \|\psi_{i\delta}\|^2_{H^1(\Omega_i)}. \tag{4.3}$$

Let R_i be a continuous lifting operator form $H_{00}^{\frac{1}{2}}(\Gamma)$ into $H^1(\Omega_i)$, defined as follows. For any λ in $H_{00}^{\frac{1}{2}}(\Gamma)$, $R_i\lambda$ belongs to $H^1(\Omega_i)$, is equal to λ on Γ, vanishes

on Γ_i, and satisfies (this is proven by using the analogous lifting operator on the unit square or cube), for all $s \geq 1$,

$$\forall \lambda \in H_{\diamond}^{s-\frac{1}{2}}(\Gamma), \quad \|R_i\lambda\|_{H^s(\Omega_i)} \leq c\, h_i^{\frac{1}{2}-s}\, \|\lambda\|_{H_{\diamond}^{s-\frac{1}{2}}(\Gamma)}, \tag{4.4}$$

where $H_{\diamond}^{s-\frac{1}{2}}(\Gamma)$ stands for the intersection $H^{s-\frac{1}{2}}(\Gamma) \cap H_{00}^{\frac{1}{2}}(\Gamma)$, provided with the norm of $H_{00}^{\frac{1}{2}}(\Gamma)$ if s is equal to 1, of $H^{s-\frac{1}{2}}(\Gamma)$ if $s > 1$. A similar operator $R_{i\delta}$, satisfying the same properties, is constructed in [19] and [4], which maps polynomials in $\mathbb{P}_{K_i}(\Gamma)$ vanishing on $\partial\Gamma$ into $\mathbb{P}_{K_i,K_i}(\Omega_i)$. Moreover, this operator satisfies, for all $s \geq 1$,

$$\forall \lambda_\delta \in \mathbb{P}_{K_i}(\Gamma) \cap H_{00}^{\frac{1}{2}}(\Gamma), \quad \|R_{i\delta}\lambda_\delta\|_{H^s(\Omega_i)} \leq c\, h_i^{\frac{1}{2}-s}\, \|\lambda_\delta\|_{H_{\diamond}^{s-\frac{1}{2}}(\Gamma)}. \tag{4.5}$$

However, in order to obtain a lifting operator of the same space of polynomials into $\mathbb{P}_{K_i,N_i}(\Gamma)$, we apply the interpolation operator $\mathcal{I}_{i\delta}$ to $R_{i\delta}\lambda_\delta$ and derive from the stability properties of this operator on polynomials (see [7, Forms. (13.27) and (13.28)]) that

$$\forall \lambda_\delta \in \mathbb{P}_{K_i}(\Gamma) \cap H_{00}^{\frac{1}{2}}(\Gamma), \quad \|\mathcal{I}_{i\delta}R_{i\delta}\lambda_\delta\|_{H^1(\Omega_i)} \leq c\, \sup\Big\{h_i^{-\frac{1}{2}}, \frac{K_i}{N_i}\, h_i^{\frac{1}{2}}\Big\}\, \|\lambda_\delta\|_{H_{00}^{\frac{1}{2}}(\Gamma)}. \tag{4.6}$$

LEMMA 4.2. *For any continuous function t_i, the following stability property holds for any g_i in $H^{-1}(\Omega_i)$ and any continuous function λ_i on Γ:*

$$\|\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)} \leq c\left(\|g_i\|_{H^{-1}(\Omega_i)} + \sup\Big\{h_i^{-\frac{1}{2}}, \frac{K_i}{N_i}\, h_i^{\frac{1}{2}}\Big\}\, \|\Pi_{i\delta}^{\Gamma}\lambda_i\|_{H_{00}^{\frac{1}{2}}(\Gamma)}\right). \tag{4.7}$$

Proof. The function $k_{i\delta}^0 = \mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i) - \mathcal{I}_{i\delta}R_{i\delta}\Pi_{i\delta}^{\Gamma}\lambda_i$ belongs to $Y_{i\delta}$ so that applying the ellipticity property (4.3) leads to

$$c\,\|k_{i\delta}^0\|_{H^1(\Omega_i)}^2 \leq c_{i\delta}(t_i; k_{i\delta}^0, k_{i\delta}^0) = \int_{\Omega_i} g_i\, k_{i\delta}^0\, d\boldsymbol{x} - c_{i\delta}(t_i; \mathcal{I}_{i\delta}R_{i\delta}\Pi_{i\delta}^{\Gamma}\lambda_i, k_{i\delta}^0).$$

The continuity property (4.2) gives

$$\|k_{i\delta}^0\|_{H^1(\Omega_i)} \leq c\,\big(\|g_i\|_{H^{-1}(\Omega_i)} + \|\mathcal{I}_{i\delta}R_{i\delta}\Pi_{i\delta}^{\Gamma}\lambda_i\|_{H^1(\Omega_i)}\big),$$

whence, by a triangle inequality,

$$\|\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)} \leq c\,\big(\|g_i\|_{H^{-1}(\Omega_i)} + \|\mathcal{I}_{i\delta}R_{i\delta}\Pi_{i\delta}^{\Gamma}\lambda_i\|_{H^1(\Omega_i)}\big).$$

The desired estimate then follows from (4.6).

Remark 4.3. Note from the previous proof that estimate (4.7) can be replaced by

$$\|\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)} \leq c\left(\|g_i\|_{Y_{i\delta}'} + \sup\Big\{h_i^{-\frac{1}{2}}, \frac{K_i}{N_i}\, h_i^{\frac{1}{2}}\Big\}\, \|\Pi_{i\delta}^{\Gamma}\lambda_i\|_{H_{00}^{\frac{1}{2}}(\Gamma)}\right), \tag{4.8}$$

where the dual norm $\|\cdot\|_{Y_{i\delta}'}$ is defined in a trivial way by

$$\|g_i\|_{Y_{i\delta}'} = \sup_{\varphi_{i\delta}\in Y_{i\delta}} \frac{\int_{\Omega_i} g_i\varphi_{i\delta}\, d\boldsymbol{x}}{\|\varphi_{i\delta}\|_{H^1(\Omega_i)}}.$$

This modified estimate is needed later on.

Next we define the integers K_i' and N_i' as the integral parts of $\frac{K_i-1}{2}$ and $\frac{N_i-1}{2}$, respectively. For technical reasons, we introduce the modified parameter $\delta' = (K_1', N_1', K_2', N_2')$.

LEMMA 4.4. *For any continuous function t_i, the following error estimate holds for any g_i in $H^{-1}(\Omega_i)$ and any continuous function λ_i in $H^{\frac12}_{00}(\Gamma)$ such that $\mathcal{L}_i(t_i)(g_i,\lambda_i)$ belongs to $H^s(\Omega_i)$, $s>1$,*

$$
\begin{aligned}
&\|(\mathcal{L}_i-\mathcal{L}_{i\delta})(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)}\\
&\le c\,\sup\Big\{h_i^{-\frac12},\frac{K_i}{N_i}\,h_i^{\frac12}\Big\}\,\Big((K_i^{1-s}+h_i^{s-1}\,N_i^{1-s})\,h_i^{\frac12-s}\,\|\mathcal{L}_i(t_i)(g_i,\lambda_i)\|_{H^s(\Omega_i)}\\
&+\|\lambda_i-\Pi^\Gamma_{i\delta}\lambda_i\|_{H^{\frac12}_{00}(\Gamma)}\\
&+\inf_{\gamma_{i\delta'}\in\mathbb{P}_{K_i',N_i'}(\Omega_i)}\|\gamma_i(t_i)-\gamma_{i\delta'}\|_{L^\infty(\Omega_i)}\|\mathcal{L}_i(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)}\Big).
\end{aligned}
\tag{4.9}
$$

Proof. We set $k_i=\mathcal{L}_i(t_i)(g_i,\lambda_i)$, $k_{i\delta}=\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)$. The proof is performed in several steps.

(1) We introduce an approximation $\lambda_{i\delta'}$ of λ_i in $\mathbb{P}_{K_i'}(\Gamma)$ which vanishes on $\partial\Gamma$ and we take $k_i'=k_i-R_i(\lambda_i-\lambda_{i\delta'})$. It follows from (4.4) that

$$
\|k_i-k_i'\|_{H^1(\Omega_i)}\le c\,h_i^{-\frac12}\,\|\lambda_i-\lambda_{i\delta'}\|_{H^{\frac12}_{00}(\Gamma)}
$$

and also that

$$
\|k_i'\|_{H^s(\Omega_i)}\le\|k_i\|_{H^s(\Omega_i)}+c\,h_i^{\frac12-s}\,\big(\|\lambda_i\|_{H^{s-\frac12}(\Gamma)}+\|\lambda_{i\delta'}\|_{H^{s-\frac12}(\Gamma)}\big).
$$

Next we set $k_{i\delta}'=\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_{i\delta'})$ and we deduce from Lemma 4.2 that

$$
\|k_{i\delta}-k_{i\delta}'\|_{H^1(\Omega_i)}\le c\,\sup\Big\{h_i^{-\frac12},\frac{K_i}{N_i}\,h_i^{\frac12}\Big\}\,\Big(\|\lambda_i-\lambda_{i\delta'}\|_{H^{\frac12}_{00}(\Gamma)}+\|\lambda_i-\Pi^\Gamma_{i\delta}\lambda_i\|_{H^{\frac12}_{00}(\Gamma)}\Big).
$$

Thanks to the triangle inequality

$$
\|k_i-k_{i\delta}\|_{H^1(\Omega_i)}\le\|k_i-k_i'\|_{H^1(\Omega_i)}+\|k_i'-k_{i\delta}'\|_{H^1(\Omega_i)}+\|k_{i\delta}-k_{i\delta}'\|_{H^1(\Omega_i)},
$$

it remains to estimate $\|k_i'-k_{i\delta}'\|_{H^1(\Omega_i)}$.

(2) The functions $k_i^0=k_i'-R_{i\delta'}\lambda_{i\delta'}$ and $k_{i\delta}^0=k_{i\delta}'-\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}$ belong to $H_0^1(\Omega_i)$ and $Y_{i\delta}$, respectively, and satisfy

$$
\begin{aligned}
&\forall\varphi_i\in H_0^1(\Omega_i),\\
&c_i(t_i;k_i^0,\varphi_i)=\int_{\Omega_i}g_i\varphi_i\,d\boldsymbol{x}-c_i(t_i;R_{i\delta'}\lambda_{i\delta'},\varphi_i)-c_i(t_i;R_i(\lambda_i-\lambda_{i\delta'}),\varphi_i),\\
&\forall\varphi_{i\delta}\in Y_{i\delta},\quad c_{i\delta}(t_i;k_{i\delta}^0,\varphi_{i\delta})=\int_{\Omega_i}g_i\varphi_{i\delta}\,d\boldsymbol{x}-c_{i\delta}(t_i;\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'},\varphi_{i\delta}).
\end{aligned}
\tag{4.10}
$$

So, denoting by $\varphi_{i\delta'}^0$ the orthogonal projection of k_i^0 onto $Y_{i\delta'}$ for the norm of $H_0^1(\Omega_i)$ and adding the difference of these equations, we deduce from the ellipticity property (4.3) that

$$
\|k_{i\delta}^0-\varphi_{i\delta'}^0\|^2_{H^1(\Omega_i)}\le c\,c_{i\delta}(t_i;k_{i\delta}^0-\varphi_{i\delta'}^0,k_{i\delta}^0-\varphi_{i\delta'}^0)
$$

$$\le c\,\Big(c_i(t_i;k_i^0-\varphi_{i\delta'}^0,k_{i\delta}^0-\varphi_{i\delta'}^0)+c_i(t_i;R_i(\lambda_i-\lambda_{i\delta'}),k_{i\delta}^0-\varphi_{i\delta'}^0)$$
$$+c_i(t_i;(\mathrm{Id}-\mathcal{I}_{i\delta'})R_{i\delta'}\lambda_{i\delta'},k_{i\delta}^0-\varphi_{i\delta'}^0)$$
$$+(c_i-c_{i\delta})(t_i;\varphi_{i\delta'}^0,k_{i\delta}^0-\varphi_{i\delta'}^0)+(c_i-c_{i\delta})(t_i;\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'},k_{i\delta}^0-\varphi_{i\delta'}^0)\Big)$$

Thanks to a triangle inequality, this yields

$$\|k_i'-k_{i\delta}'\|_{H^1(\Omega_i)}$$
$$\le c\,\Bigg(\|k_i^0-\varphi_{i\delta'}^0\|_{H^1(\Omega_i)}+c\,h_i^{-\frac12}\,\|\lambda_i-\lambda_{i\delta'}\|_{H_{00}^{\frac12}(\Gamma)}+\|(\mathrm{Id}-\mathcal{I}_{i\delta'})R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}$$
$$+\sup_{\chi_{i\delta}\in Y_{i\delta}}\frac{\int_{\Omega_i}\gamma_i(t_i)\,\nabla(\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'})\cdot\nabla\chi_{i\delta}\,d\boldsymbol{x}-c_{i\delta}(t_i;\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'},\chi_{i\delta})}{\|\chi_{i\delta}\|_{H^1(\Omega_i)}}\Bigg).$$

(3) In order to evaluate the last term, we observe that, for any $\chi_{i\delta}$ in Y_δ, any $\psi_{i\delta'}$ in $\mathbb{P}_{K_i',N_i'}(\Omega_i)$, and any $\gamma_{i\delta'}$ in $\mathbb{P}_{K_i',N_i'}(\Omega_i)$,

$$\int_{\Omega_i}\gamma_{i\delta'}\,\nabla\psi_{i\delta'}\cdot\nabla\chi_{i\delta}\,d\boldsymbol{x}=(\gamma_{i\delta'}\,\nabla\psi_{i\delta'},\nabla\chi_{i\delta})_{i\delta}.$$

Adding and subtracting this quantity and using the continuity property (4.2) leads to, for any $\chi_{i\delta}$ in $Y_{i\delta}$,

$$\int_{\Omega_i}\gamma_i(t_i)\,\nabla(\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'})\cdot\nabla\chi_{i\delta}\,d\boldsymbol{x}-c_{i\delta}(t_i;\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'},\chi_{i\delta})$$
$$\le c\,\|\gamma_i(t_i)-\gamma_{i\delta'}\|_{L^\infty(\Omega_i)}\|\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}\|\chi_{i\delta}\|_{H^1(\Omega_i)}.$$

Moreover, it follows from the definition of $\varphi_{i\delta'}^0$ that

$$\|\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}\le\|k_i^0\|_{H^1(\Omega_i)}+\|\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}$$
$$\le\|k_i'\|_{H^1(\Omega_i)}+\|R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}+\|\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}.$$

Thanks to (4.5) and (4.6), we obtain

$$\|\varphi_{i\delta'}^0+\mathcal{I}_{i\delta'}R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}\le\|k_i'\|_{H^1(\Omega_i)}+c\sup\Big\{h_i^{-\frac12},\frac{K_i}{N_i}\,h_i^{\frac12}\Big\}\,\|\lambda_{i\delta'}\|_{H_{00}^{\frac12}(\Gamma)}.$$

(4) To conclude, we note that the trace λ_i of k_i belongs to $H^{s-\frac12}(\Gamma)$ and choose the polynomial $\lambda_{i\delta'}$ such that (see [7, Thm. 7.4])

$$\|\lambda_i-\lambda_{i\delta'}\|_{H_{00}^{\frac12}(\Gamma)}\le c\,K_i^{1-s}\,\|\lambda_i\|_{H^{s-\frac12}(\Gamma)},\quad\|\lambda_{i\delta'}\|_{H^{s-\frac12}(\Gamma)}\le c\,\|\lambda_i\|_{H^{s-\frac12}(\Gamma)}.$$

Next it can be observed that, for any polynomial $r_{i\delta'}$ in $\mathbb{P}_{K_i',N_i'}(\Omega_i)$,

$$\|(\mathrm{Id}-\mathcal{I}_{i\delta'})R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}=\|(\mathrm{Id}-\mathcal{I}_{i\delta'})(R_{i\delta'}\lambda_{i\delta'}-r_{i\delta'})\|_{H^1(\Omega_i)}.$$

Using the stability properties of the operator $\mathcal{I}_{i\delta'}$ on polynomials (see [7, Forms. (13.27) and (13.28)]) and taking $r_{i\delta'}$ equal to the orthogonal projection of $R_{i\delta'}\lambda_{i\delta'}$ in $H^1(\Omega_i)$ yields

$$\|(\mathrm{Id}-\mathcal{I}_{i\delta'})R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)}\le c\,\sup\Big\{h_i^{-\frac12},\frac{K_i}{N_i}\,h_i^{\frac12}\Big\}\,(K_i^{1-s}+h_i^{s-1}\,N_i^{1-s})\,\|R_{i\delta'}\lambda_{i\delta'}\|_{H^s(\Omega_i)},$$

whence, from (4.5),

$$\|(\mathrm{Id}-\mathcal{I}_{i\delta'})R_{i\delta'}\lambda_{i\delta'}\|_{H^1(\Omega_i)} \le c\,\sup\Big\{h_i^{-\frac12}, \frac{K_i}{N_i}\,h_i^{\frac12}\Big\}\,(K_i^{1-s}+h_i^{s-1}\,N_i^{1-s})\,h^{\frac12-s}\,\|\lambda_{i\delta'}\|_{H^{s-\frac12}(\Gamma)}.$$

Finally, using the previous estimates also yields

$$\|k_i^0-\varphi_{i\delta'}^0\|_{H^1(\Omega_i)} \le c\,(K_i^{1-s}+h_i^{s-1}N_i^{1-s})\Big(\|k_i'\|_{H^s(\Omega_i)}+h_i^{\frac12-s}\,\|\lambda_{i\delta'}\|_{H^{s-\frac12}(\Gamma)}\Big).$$

To conclude, we observe that $\|\lambda_i\|_{H^{s-\frac12}(\Gamma)}$ is bounded by a constant $\|k_i\|_{H^s(\Omega_i)}$. This ends the proof.

Remark 4.5. The following estimate can be derived by combining [7, Thm. 7.4] with a Gagliardo–Nirenberg inequality: if the function γ_i is of class $\mathtt{C}^{\,m}$ with bounded derivatives of order $\le m$ and if the function t_i belongs to $H^s(\Omega_i)$, $\frac{d}{2} < s \le m$,

$$\inf_{\gamma_{i\delta'}\in\mathbb{P}_{K_i,N_i}(\Omega_i)} \|\gamma_i(t_i)-\gamma_{i\delta'}\,\|_{L^\infty(\Omega_i)} \le c\,(K_i^{\frac d2-s}+h_i^{s-\frac d2}\,N_i^{\frac d2-s})\,\|t_i\|_{H^s(\Omega_i)}. \tag{4.11}$$

Moreover, a more sophisticated argument, using the full regularity of $\mathcal{L}_i(t_i)(g_i,\lambda_i)$ allows us to replace when s is $> \frac{d}{2}$ the last term in (4.9) by the better estimate

$$c\,(K_i^{1-s}+h_i^{s-1}\,N_i^{1-s})\,\|t_i\|_{H^s(\Omega_i)}\|\mathcal{L}_i(t_i)(g_i,\lambda_i)\|_{H^s(\Omega_i)}. \tag{4.12}$$

Remark 4.6. If the function γ_i is differentiable with bounded derivative and if the function t_i belongs to $H^r(\Omega_i)$ for $r>1$, the Aubin–Nitsche duality argument [7, Thm. 15.4] leads to the improved estimate

$$\begin{aligned}\|(\mathcal{L}_i-\mathcal{L}_{i\delta})(t_i)(g_i,\lambda_i)\|_{H^{1-\varepsilon}(\Omega_i)}&\\ \le c\,\Big((K_i^{-\varepsilon}+h_i^{\varepsilon}\,N_i^{-\varepsilon})\|(\mathcal{L}_i-\mathcal{L}_{i\delta})(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)}&\\ +\inf_{\gamma_{i\delta'}\in\mathbb{P}_{K_i',N_i'}(\Omega_i)}\|\gamma_i(t_i)-\gamma_{i\delta'}\,\|_{L^\infty(\Omega_i)}\|\mathcal{L}_i(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)}\Big).&\end{aligned} \tag{4.13}$$

Finally, we investigate the dependency of $\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)$ with respect to t_i.

LEMMA 4.7. *For any continuous functions t_i and t_i', the following stability property holds for any g_i in $H^{-1}(\Omega_i)$ and any continuous function λ_i on Γ:*

$$\begin{aligned}\|\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)-\mathcal{L}_{i\delta}(t_i')(g_i,\lambda_i)\|_{H^1(\Omega_i)}&\\ \le c\,\|\gamma_i(t_i)-\gamma_i(t_i')\|_{L^\infty(\Omega_i)}\|\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)\|_{H^1(\Omega_i)}.&\end{aligned} \tag{4.14}$$

Proof. Setting $k_{i\delta}=\mathcal{L}_{i\delta}(t_i)(g_i,\lambda_i)$ and $k_{i\delta}'=\mathcal{L}_{i\delta}(t_i')(g_i,\lambda_i)$, we observe that the function $k_{i\delta}-k_{i\delta}'$ belongs to $Y_{i\delta}$ and satisfies

$$\forall\varphi_{i\delta}\in Y_{i\delta},\quad \big(\gamma_i(t_i)\,\nabla k_{i\delta},\nabla\varphi_{i\delta}\big)_{i\delta}=\big(\gamma_i(t_i')\,\nabla k_{i\delta}',\nabla\varphi_{i\delta}\big)_{i\delta},$$

whence

$$\begin{aligned}c_{i\delta}(t_i';k_{i\delta}-k_{i\delta}',k_{i\delta}-k_{i\delta}')&=\big(\gamma_i(t_i')\,(\nabla k_{i\delta}-\nabla k_{i\delta}'),(\nabla k_{i\delta}-\nabla k_{i\delta}')\big)_{i\delta}\\ &=-\big(\big(\gamma_i(t_i)-\gamma_i(t_i')\big)\,\nabla k_{i\delta},(\nabla k_{i\delta}-\nabla k_{i\delta}')\big)_{i\delta}.\end{aligned}$$

So the desired estimated follows from the properties of $c_{i\delta}(\cdot;\cdot,\cdot)$; see Lemma 4.1.

The discrete Stokes operator. We now present similar properties for the Stokes operator $\mathcal{S}_{i\delta}$ defined by (3.14); however, we skip the proofs except for the error estimates.

LEMMA 4.8. *For any continuous function t_i, the form $a_{i\delta}(t_i;\cdot,\cdot)$ satisfies the following properties of continuity:*

$$\forall \boldsymbol{u}_{i\delta} \in X_{i\delta},\ \forall \boldsymbol{v}_{i\delta} \in X_{i\delta}, \quad a_{i\delta}(t_i;\boldsymbol{u}_{i\delta},\boldsymbol{v}_{i\delta}) \le c\,\|\boldsymbol{u}_{i\delta}\|_{H^1(\Omega_i)^d}\|\boldsymbol{v}_{i\delta}\|_{H^1(\Omega_i)^d}, \tag{4.15}$$

and of ellipticity:

$$\forall \boldsymbol{v}_{i\delta} \in X_{i\delta}, \quad a_{i\delta}(t_i;\boldsymbol{v}_{i\delta},\boldsymbol{v}_{i\delta}) \ge c\,\|\boldsymbol{v}_{i\delta}\|^2_{H^1(\Omega_i)^d}. \tag{4.16}$$

LEMMA 4.9. *For any continuous function t_i, the following stability property holds for any $\boldsymbol{g}_i$ in $L^2(\Omega_i)^d$ and any $\boldsymbol{\lambda}_i$ in the dual space of $H^{\frac12}_{00}(\Gamma)^d$:*

$$\|\mathcal{S}_{i\delta}(t_i)(\boldsymbol{g}_i,\boldsymbol{\lambda}_i)\|_{H^1(\Omega_i)^d} \le c\,\big(\|\boldsymbol{g}_i\|_{L^2(\Omega_i)^d} + \|\boldsymbol{\lambda}_i\|_{H^{\frac12}_{00}(\Gamma)'^d}\big). \tag{4.17}$$

Remark 4.10. As for the Laplace operator, the norms of $\boldsymbol{g}_i$ and $\boldsymbol{\lambda}_i$ in the right-hand side can be replaced, respectively, by the dual norms of $X_{i\delta}$ (when provided with the norm $\|\cdot\|_{H^1(\Omega_i)^d}$) and of $\mathbb{P}_{K_i}(\Gamma)^d \cap H^{\frac12}_{00}(\Gamma)^d$ (provided with the norm $\|\cdot\|_{H^{\frac12}_{00}(\Gamma)^d}$).

However, the proof of the convergence estimate is slightly different (but simpler).

LEMMA 4.11. *For any continuous function t_i, the following error estimate holds for any $\boldsymbol{g}_i$ in $L^2(\Omega_i)^d$ and any $\boldsymbol{\lambda}_i$ in the dual space of $H^{\frac12}_{00}(\Gamma)^d$:*

$$\begin{aligned}
&\|(\mathcal{S}_i - \mathcal{S}_{i\delta})(t_i)(\boldsymbol{g}_i,\boldsymbol{\lambda}_i)\|_{H^1(\Omega_i)^d} \\
&\quad\le c \inf_{\boldsymbol{w}_{i\delta}\in X_{i\delta'}\cap V_i}\Big(\|\mathcal{S}_i(t_i)(\boldsymbol{g}_i,\boldsymbol{\lambda}_i) - \boldsymbol{w}_{i\delta}\|_{H^1(\Omega_i)^d} \\
&\qquad\qquad + \inf_{\alpha_{i\delta'}\in\mathbb{P}_{K_i',N_i'}(\Omega_i)} \|\alpha_i(t_i) - \alpha_{i\delta'}\|_{L^\infty(\Omega_i)}\|\boldsymbol{w}_{i\delta}\|_{H^1(\Omega_i)^d}\Big).
\end{aligned} \tag{4.18}$$

Proof. Setting $\boldsymbol{u}_i = \mathcal{S}_i(t_i)(\boldsymbol{g}_i,\boldsymbol{\lambda}_i)$ and $\boldsymbol{u}_{i\delta} = \mathcal{S}_{i\delta}(t_i)(\boldsymbol{g}_i,\boldsymbol{\lambda}_i)$, we derive from (4.16) that, for any $\boldsymbol{w}_{i\delta'}$ in $X_{i\delta'}\cap V_i$,

$$\|\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}\|^2_{H^1(\Omega_i)^d} \le c\,a_{i\delta}(t_i;\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}, \boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}).$$

Using (2.14) (in variational form) and (3.14), we derive

$$\|\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}\|^2_{H^1(\Omega_i)^d} \le c\,\big(a_i(t_i;\boldsymbol{u}_i, \boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}) - a_{i\delta}(t_i;\boldsymbol{w}_{i\delta'}, \boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'})\big).$$

Next we deduce from the exactness of the quadrature formula that, for any $\alpha_{i\delta'}$ in $\mathbb{P}_{K_i',N_i'}(\Omega_i)$,

$$\int_{\Omega_i} \alpha_{i\delta'}\,\nabla\boldsymbol{w}_{i\delta'}\,\cdot\,\nabla(\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'})\,d\boldsymbol{x} = (\alpha_{i\delta'}\,\nabla\boldsymbol{w}_{i\delta'}, \nabla(\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}))_{i\delta}.$$

Adding and subtracting this quantity yield

$$\begin{aligned}
\|\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}\|^2_{H^1(\Omega_i)^d} \le c\,\Big(&a_i(t_i;\boldsymbol{u}_i - \boldsymbol{w}_{i\delta'}, \boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}) \\
&+ \int_{\Omega_i} (\alpha_i(t_i) - \alpha_{i\delta'})\,\nabla\boldsymbol{w}_{i\delta'}\,\cdot\,\nabla(\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'})\,d\boldsymbol{x} \\
&- ((\alpha_i(t_i) - \alpha_{i\delta'})\,\nabla\boldsymbol{w}_{i\delta'}, \nabla(\boldsymbol{u}_{i\delta} - \boldsymbol{w}_{i\delta'}))_{i\delta}\Big).
\end{aligned}$$

So the desired estimate follows from the continuity property (4.15), together with a triangle inequality.

Remark 4.12. In dimension $d = 2$, it is easy to evaluate the distance of a function $\boldsymbol{u}_i$ in V_i to $X_{i\delta} \cap V_i$ by introducing the stream function ψ_i such that $\boldsymbol{u}_i = \mathbf{curl}\ \psi_i$. Indeed, the functions $\mathbf{curl}\ \psi_{i\delta}$, where $\psi_{i\delta}$ belongs to $\mathbb{P}_{K_i,N_i}(\Omega_i)$, satisfies the desired boundary conditions and approximates ψ_i in $H^2(\Omega_i)$, belongs to $X_{i\delta} \cap V_i$, and provides a good approximation of $\boldsymbol{u}_i$. The case of dimension $d = 3$ is more complex; however, the right approximation properties have been proved in [22] for smooth functions and extended in [5] to arbitrary functions. So the general result reads as follows: for any function $\boldsymbol{u}_i$ in $H^s(\Omega_i)^d \cap V_i$, $s \geq 1$,

$$\inf_{\boldsymbol{w}_{i\delta} \in X_{i\delta} \cap V_i} \|\mathcal{S}_i(t_i)(\boldsymbol{g}_i, \boldsymbol{\lambda}_i) - \boldsymbol{w}_{i\delta}\|_{H^1(\Omega_i)^d} \leq c\,(K_i^{1-s} + h_i^{s-1}\, N_i^{1-s})\, \|\boldsymbol{u}_i\|_{H^s(\Omega_i)^d}. \tag{4.19}$$

Note also that the last term in (4.18) can be bounded analogously to (4.11) or (4.12).

LEMMA 4.13. *For any continuous functions t_i and t_i', the following stability property holds for any $\boldsymbol{g}_i$ in $L^2(\Omega_i)^d$ and any $\boldsymbol{\lambda}_i$ in the dual space of $H_{00}^{\frac{1}{2}}(\Gamma)^d$:*

$$\begin{aligned} &\|\mathcal{S}_{i\delta}(t_i)(\boldsymbol{g}_i, \boldsymbol{\lambda}_i) - \mathcal{S}_{i\delta}(t_i')(\boldsymbol{g}_i, \boldsymbol{\lambda}_i)\|_{H^1(\Omega_i)^d} \\ &\qquad \leq c\, \|\alpha_i(t_i) - \alpha_i(t_i')\|_{L^\infty(\Omega_i)} \|\mathcal{S}_{i\delta}(t_i)(\boldsymbol{g}_i, \boldsymbol{\lambda}_i)\|_{H^1(\Omega_i)^d}. \end{aligned} \tag{4.20}$$

5. Numerical analysis of the discrete problem. The aim of this section is to prove that, if Hypotheses 2.2 and 2.4 hold, problem (3.4)–(3.5) has a unique solution in a neighborhood of (U_1^*, U_2^*) and that this solution converges to (U_1^*, U_2^*). We also derive optimal error estimates. To this aim, we check the assumptions of the theorem of Brezzi, Rappaz, and Raviart [10] in Propositions 5.5 to 5.7.

From now on, we denote by $\mathcal{X}_{i\delta}$, $i = 1$ and 2, the space $X_{i\delta} \times \mathbb{P}_{K_i,N_i}(\Omega_i)$, provided with the norm of $\mathcal{X}_i$, and by $\mathcal{X}_\delta$ the product $\mathcal{X}_{1\delta} \times \mathcal{X}_{2\delta}$.

In view of (4.6), (4.9), and (4.19), for instance, we decide to take the N_i, $i = 1$ and 2, such that, for a fixed constant κ,

$$\kappa\, h_i\, K_i \leq N_i < \kappa\, h_i\, K_i + 1, \tag{5.1}$$

and we do not any longer take into account the dependency of the constants with respect to the h_i. We also choose an approximation $(U_{1\delta'}^*, U_{2\delta'}^*)$, with $U_{i\delta'}^* = (\boldsymbol{u}_{i\delta'}^*, k_{i\delta'}^*)$, of the solution (U_1^*, U_2^*) in $\prod_{i=1}^2 (X_{i\delta'} \times P_{K_i',N_i'}(\Omega_i))$ which satisfies the following approximation properties for $0 \leq r \leq s^*$, where s^* is introduced in Hypothesis 2.2:

$$\begin{aligned} &\|\boldsymbol{u}_i^* - \boldsymbol{u}_{i\delta'}^*\|_{H^r(\Omega_i)^d} \leq c\, K_i^{r-s^*}\, \|\boldsymbol{u}_i^*\|_{H^{s^*}(\Omega_i)^d}, \\ &\|k_i^* - k_{i\delta'}^*\|_{H^r(\Omega_i)} \leq c\, K_i^{r-s^*}\, \|k_i^*\|_{H^{s^*}(\Omega_i)}. \end{aligned} \tag{5.2}$$

The existence of such an approximation is stated in [7, Thm. 7.4]. Finally, we assume that the functions α_i and γ_i are of class $\mathtt{C}^{\,2}$, with bounded derivatives up to order 2 and also that the operators $\Pi_{i\delta}^\Gamma$ satisfy, for all $s \geq \frac{1}{2}$ (the notation $H_\diamond^{s-\frac{1}{2}}(\Gamma)$ is introduced in (4.4)),

$$\forall \lambda \in H_\diamond^{s-\frac{1}{2}}(\Gamma), \quad \|\lambda - \Pi_{i\delta}^\Gamma \lambda\|_{H_{00}^{\frac{1}{2}}(\Omega_i)} \leq c\, K_i^{1-s}\, \|\lambda\|_{H_\diamond^{s-\frac{1}{2}}(\Gamma)}. \tag{5.3}$$

In a first step, we must prove the analogue of Hypothesis 2.4 for the discrete operator. The proof relies on the expansion

$$\begin{aligned} &\mathrm{Id} + D\mathcal{T}_\delta(k_{1\delta'}^*, k_{2\delta'}^*)\mathcal{G}_\delta(U_{1\delta'}^*, U_{2\delta'}^*) + \mathcal{T}_\delta(k_{1\delta'}^*, k_{2\delta'}^*) D\mathcal{G}_\delta(U_{1\delta'}^*, U_{2\delta'}^*) \\ &\qquad = \mathrm{Id} + D\mathcal{T}(k_1^*, k_2^*)\mathcal{G}(U_1^*, U_2^*) + \mathcal{T}(k_1^*, k_2^*) D\mathcal{G}(U_1^*, U_2^*) + \sum_{j=1}^4 \mathcal{O}_j, \end{aligned}$$

with

$$\begin{aligned}
\mathcal{O}_1 &= -\big(D\mathcal{T}(k_1^*,k_2^*)\mathcal{G}(U_1^*,U_2^*) - D\mathcal{T}_\delta(k_1^*,k_2^*)\mathcal{G}(U_{1\delta'}^*,U_{2\delta'}^*)\big)\\
&\qquad - (\mathcal{T}-\mathcal{T}_\delta)(k_1^*,k_2^*)D\mathcal{G}(U_1^*,U_2^*),\\
\mathcal{O}_2 &= -\big(D\mathcal{T}_\delta(k_1^*,k_2^*) - D\mathcal{T}_\delta(k_{1\delta'}^*,k_{2\delta'}^*)\big)\mathcal{G}(U_{1\delta'}^*,U_{2\delta'}^*)\\
&\qquad - \big(\mathcal{T}_\delta(k_1^*,k_2^*) - \mathcal{T}_\delta(k_{1\delta'}^*,k_{2\delta'}^*)\big)D\mathcal{G}(U_1^*,U_2^*),\\
\mathcal{O}_3 &= -\mathcal{T}_\delta(k_{1\delta'}^*,k_{2\delta'}^*)\big(D\mathcal{G}(U_1^*,U_2^*) - D\mathcal{G}(U_{1\delta'}^*,U_{2\delta'}^*)\big),\\
\mathcal{O}_4 &= -D\mathcal{T}_\delta(k_{1\delta'}^*,k_{2\delta'}^*)\big(\mathcal{G}(U_{1\delta'}^*,U_{2\delta'}^*) - \mathcal{G}_\delta(U_{1\delta'}^*,U_{2\delta'}^*)\big)\\
&\qquad - \mathcal{T}_\delta(k_{1\delta'}^*,k_{2\delta'}^*)\big(D\mathcal{G}(U_{1\delta'}^*,U_{2\delta'}^*) - D\mathcal{G}_\delta(U_{1\delta'}^*,U_{2\delta'}^*)\big).
\end{aligned}$$

So we now prove that each $\mathcal{O}_j$ tends to zero when K_1 and K_2 go to $+\infty$, in the norm of the space $\mathsf{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)$ of linear mappings from $\mathcal{X}_\delta$ into itself. These properties are stated in the following lemmas.

Let us first observe that, for any $W=(W_1,W_2)$ in $\mathcal{X}$, with $W_i=(\boldsymbol{w}_i,m_i)$,

$$D\mathcal{G}(U_1^*,U_2^*)\cdot W = \begin{pmatrix} (\mathbf{0}, D\boldsymbol{\lambda}_1(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2))\\ (-Dg_1(k_1^*,\boldsymbol{u}_1^*).(m_1,\boldsymbol{w}_1), D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2))\\ (\mathbf{0}, D\boldsymbol{\lambda}_2(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2))\\ (-Dg_2(k_2^*,\boldsymbol{u}_2^*).(m_2,\boldsymbol{w}_2), D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2)) \end{pmatrix}, \tag{5.4}$$

with

$$\begin{aligned}
D\boldsymbol{\lambda}_i(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2) &= |\boldsymbol{u}_i^*-\boldsymbol{u}_j^*|\,(\boldsymbol{w}_i-\boldsymbol{w}_j) + \frac{(\boldsymbol{u}_i^*-\boldsymbol{u}_j^*)\cdot(\boldsymbol{w}_i-\boldsymbol{w}_j)}{|\boldsymbol{u}_i^*-\boldsymbol{u}_j^*|}\,(\boldsymbol{u}_i^*-\boldsymbol{u}_j^*),\\
Dg_i(k_i^*,\boldsymbol{u}_i^*).(m_i,\boldsymbol{w}_i) &= 2\,\alpha_i(k_i^*)\,\nabla\boldsymbol{u}_i^*\,\cdot\,\nabla\boldsymbol{w}_i + \alpha_i'(k_i^*)m_i\,|\nabla\boldsymbol{u}_i^*|^2,\\
D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2) &= -2\,(\boldsymbol{u}_1^*-\boldsymbol{u}_2^*)\,\cdot\,(\boldsymbol{w}_1-\boldsymbol{w}_2).
\end{aligned} \tag{5.5}$$

Moreover, we have the following formula (which can be derived from the explicit form (2.20)–(2.21) of the linearized problem):

$$\begin{aligned}
&D\mathcal{T}(k_1^*,k_2^*)\mathcal{G}(U_1^*,U_2^*) + \mathcal{T}(k_1^*,k_2^*)D\mathcal{G}(U_1^*,U_2^*)\\
&\qquad = \mathcal{T}(k_1^*,k_2^*)(D\mathcal{G}(U_1^*,U_2^*) + D\mathcal{H}(U_1^*,U_2^*)),
\end{aligned} \tag{5.6}$$

with

$$D\mathcal{H}(U_1^*,U_2^*).(m_1,m_2) = \begin{pmatrix} (-\mathrm{div}\ (\alpha_1'(k_1^*)m_1\nabla\boldsymbol{u}_1^*),\mathbf{0})\\ (-\mathrm{div}\ (\gamma_1'(k_1^*)m_1\nabla k_1^*),0)\\ (-\mathrm{div}\ (\alpha_2'(k_2^*)m_2\nabla\boldsymbol{u}_2^*),\mathbf{0})\\ (-\mathrm{div}\ (\gamma_1'(k_2^*)m_2\nabla k_2^*),0) \end{pmatrix}.$$

A similar formula would hold for the discrete problem

$$\begin{aligned}
&D\mathcal{T}_\delta(k_1^*,k_2^*)\mathcal{G}(U_1^*,U_2^*) + \mathcal{T}_\delta(k_1^*,k_2^*)D\mathcal{G}(U_1^*,U_2^*)\\
&\qquad = \mathcal{T}_\delta(k_1^*,k_2^*)(D\mathcal{G}(U_1^*,U_2^*) + D\mathcal{H}_\delta(U_1^*,U_2^*)),
\end{aligned} \tag{5.7}$$

where the duality product of the first part of the first and third components with a $\boldsymbol{v}_{i\delta}$ in $V_{i\delta}$, respectively, of the second and fourth components with a $\ell_{i\delta}$ in Y_δ, are given by the formulas

$$(\alpha_i'(k_i^*)m_{i\delta}\,\nabla\boldsymbol{u}_i^*,\nabla\boldsymbol{v}_{i\delta})_{i\delta}, \qquad (\gamma_i'(k_i^*)m_{i\delta}\,\nabla k_i^*,\nabla\ell_{i\delta})_{i\delta}.$$

(Here, the differential operator is applied to a $W_\delta = (W_{1\delta}, W_{2\delta})$ in $\mathcal{X}_\delta$, with each $W_{i\delta}$ equal to $(\boldsymbol{w}_{i\delta}, m_{i\delta})$.)

From now on, we fix ε such that $2\varepsilon < s^* - \frac{d}{2}$.

LEMMA 5.1. *If Hypothesis* 2.2 *is satisfied, the following property holds:*

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|\mathcal{O}_1\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)} = 0. \tag{5.8}$$

Proof. Due to formulas (5.6) and (5.7), we observe that

$$\begin{aligned}\|\mathcal{O}_1\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)} \le{}& \|(\mathcal{T}-\mathcal{T}_\delta)(k_1^*,k_2^*)\big(D\mathcal{G}(U_1^*,U_2^*)+D\mathcal{H}(U_1^*,U_2^*)\big)\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)}\\ &+\|\mathcal{T}_\delta(k_1^*,k_2^*)\big(D\mathcal{H}(U_1^*,U_2^*)-D\mathcal{H}_\delta(U_{1\delta'}^*,U_{2\delta'}^*)\big)\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)}.\end{aligned}$$

(1) To bound the first term, we note that, when $(W_{1\delta}, W_{2\delta})$ runs through the unit sphere of $\mathcal{X}_\delta$, the quantities

$$D\boldsymbol{\lambda}_i(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_{1\delta},\boldsymbol{w}_{2\delta}),\quad Dg_i(k_i^*,\boldsymbol{u}_i^*).(m_{i\delta},\boldsymbol{w}_{i\delta}),\quad \text{and}\quad D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_{1\delta},\boldsymbol{w}_{2\delta})$$

belong to a bounded set in $L^2(\Omega_i)^d$, $H^{-1}(\Omega_i)$, and $H_{00}^{\frac{1}{2}}(\Gamma)$, respectively. Then it can be checked that $\mathcal{S}_i(k_i^*)(\mathbf{0}, D\boldsymbol{\lambda}_i(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_{1\delta},\boldsymbol{w}_{2\delta}))$ remains inside a bounded set of $H^s(\Omega_i)^d$ for some $s>1$. Thanks to (4.18), (4.19), and the analogue of (4.11), this yields the uniform convergence of $(\mathcal{S}_i-\mathcal{S}_{i\delta})(k_i^*)(\mathbf{0}, D\boldsymbol{\lambda}_i(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_{1\delta},\boldsymbol{w}_{2_\delta}))$. The convergence of the other terms, say

$$(\mathcal{L}_i-\mathcal{L}_{i\delta})(k_i^*)(Dg_i(k_i^*,\boldsymbol{u}_i^*).(m_{i\delta},\boldsymbol{w}_{i\delta}), D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_{1\delta},\boldsymbol{w}_{2\delta})),$$

follows from (4.13) together with the fact that both

$$\mathcal{L}_i(Dg_i(k_i^*,\boldsymbol{u}_i^*).(m_i,\boldsymbol{w}_i), D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2))$$

and its discrete analogue

$$\mathcal{L}_{i\delta}(k_i^*)(Dg_i(k_i^*,\boldsymbol{u}_i^*).(m_i,\boldsymbol{w}_i), D\lambda(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*).(\boldsymbol{w}_1,\boldsymbol{w}_2))$$

are bounded in $H^1(\Omega_i)$; see (4.7). So combining these facts yields

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|(\mathcal{T}-\mathcal{T}_\delta)(k_1^*,k_2^*)D\mathcal{G}(U_1^*,U_2^*)\|_{\mathcal{L}\ (\mathcal{X},\mathcal{X})} = 0.$$

(2) Similarly, we observe that, when $(W_{1\delta}, W_{2\delta})$ runs through the unit sphere of $\mathcal{X}_\delta$, the quantities $\alpha_i'(k_i^*)m_{i\delta}\,\nabla\boldsymbol{u}_i^*$ and $\gamma_i'(k_i^*)m_{i\delta}\,\nabla k_i^*$ belong to a bounded set of $H^s(\Omega_i)^{d^2}$ and $H^s(\Omega_i)^d$, for some $s>0$, so that their divergence belongs to $H^{s-1}(\Omega_i)^d$ and $H^{s-1}(\Omega_i)$, respectively. This implies

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|(\mathcal{T}-\mathcal{T}_\delta)(k_1^*,k_2^*)D\mathcal{H}(U_1^*,U_2^*)\|_{\mathcal{L}\ (\mathcal{X},\mathcal{X})} = 0.$$

(3) To prove the convergence of the last term, we observe from (4.8) that it suffices to prove, for any $m_{i\delta}$ in the intersection of $\mathbb{P}_{K_i,N_i}(\Omega_i)$ with the unit sphere of $H^{1-\varepsilon}(\Omega_i)$, the convergence of

$$\sup_{\ell_{i\delta}\in Y_{i\delta}} \frac{\int_{\Omega_i}\gamma_i'(k_i^*)m_{i\delta}\,\nabla k_i^*\,\cdot\,\nabla\ell_{i\delta}\,d\boldsymbol{x} - (\gamma_i'(k_i^*)m_{i\delta}\,\nabla k_{i\delta'}^*, \nabla\ell_{i\delta})_{i\delta}}{\|\ell_{i\delta}\|_{H^1(\Omega_i)}}.$$

Handling the other terms is similar. Denoting by δ'' the 4-tuple $(K_1'', N_1'', K_2'', N_2'')$ with each K_i'' equal to the integral part of $\frac{K_i-1}{4}$ and each N_i'' equal to the integral part of $\frac{N_i-1}{4}$, we derive from the exactness of the quadrature formula for any $\tilde{\gamma}_{i\delta''}$ and $m_{i\delta''}$ in $\mathbb{P}_{K_i'',N_i''}(\Omega_i)$,

$$\int_{\Omega_i} \gamma_{i\delta''} m_{i\delta''} \nabla k^*_{i\delta'} \cdot \nabla \ell_{i\delta}\, d\boldsymbol{x} = (\gamma_{i\delta''} m_{i\delta''} \nabla k^*_{i\delta'}, \nabla \ell_{i\delta})_{i\delta}.$$

By adding and subtracting this line, we prove that the previous quantity is bounded by the sum of the terms

$$\begin{aligned}
&\|\gamma_i'(k_i^*)\|_{L^\infty(\Omega_i)} \|m_{i\delta} \nabla(k_i^* - k^*_{i\delta'})\|_{L^2(\Omega_i)^d},\\
&\quad \|\gamma_i'(k_i^*)\|_{L^\infty(\Omega_i)} \|(m_{i\delta} - m_{i\delta''}) \nabla k^*_{i\delta'}\|_{L^2(\Omega_i)^d},\\
&\quad\quad \|\gamma_i'(k_i^*) - \tilde{\gamma}_{i\delta''}\|_{L^\infty(\Omega_i)} \|m_{i\delta''} \nabla k^*_{i\delta'}\|_{L^2(\Omega_i)^d},\\
&\quad\quad\quad \|\gamma_i'(k_i^*)\|_{L^\infty(\Omega_i)} \|\mathcal{I}_{i\delta}\big((m_{i\delta} - m_{i\delta''}) \nabla k^*_{i\delta'}\big)\|_{L^2(\Omega_i)^d},\\
&\quad\quad\quad\quad \|\gamma_i'(k^*_{i\delta}) - \tilde{\gamma}_{i\delta''}\|_{L^\infty(\Omega_i)} \|\mathcal{I}_{i\delta}\big(m_{i\delta''} \nabla k^*_{i\delta'}\big)\|_{L^2(\Omega_i)^d}.
\end{aligned}$$

Thanks to the choice of ε, the product of two functions is continuous from $H^{1-\varepsilon}(\Omega_i) \times H^{s^*-1-\varepsilon}(\Omega_i)$ into $L^2(\Omega_i)$ so that the uniform convergence of the first term follows from (5.2). Again, thanks to the choice of ε, the product of two functions is continuous from $H^{1-2\varepsilon}(\Omega_i) \times H^{s^*-1}(\Omega_i)$ into $L^2(\Omega_i)$ so that the uniform convergence of the second term is obtained by taking $m_{i\delta''}$ equal to the projection of $m_{i\delta}$ onto $\mathbb{P}_{K_i'',N_i''}(\Omega_i)$ for the scalar product of $H^{1-\varepsilon}(\Omega_i)$ and using the approximation properties of this projection operator; see [7, Thm. 7.4]. Similarly, the convergence of the third term follows from (4.11). The convergence of the last two terms is derived by similar arguments combined with the stability on the operator $\mathcal{I}_{i\delta}$ on polynomials; see [7, Form. (13.28)].

So the proof is complete.

LEMMA 5.2. *If Hypothesis* 2.2 *is satisfied, the following property holds:*

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|\mathcal{O}_2\|_{\mathcal{L}(\mathcal{X}_\delta, \mathcal{X}_\delta)} = 0. \tag{5.9}$$

Proof. From (5.7), we have

$$\|\mathcal{O}_2\|_{\mathcal{L}(\mathcal{X}_\delta,\mathcal{X}_\delta)} = \|\big(\mathcal{T}_\delta(k_1^*, k_2^*) - \mathcal{T}_\delta(k^*_{1\delta'}, k^*_{2\delta'})\big)\big(D\mathcal{G}(U_1^*, U_2^*) + D\mathcal{H}_\delta(U^*_{1\delta'}, U^*_{2\delta'})\big)\|_{\mathcal{L}(\mathcal{X}_\delta,\mathcal{X}_\delta)}.$$

When $(W_{1\delta}, W_{2\delta})$ runs through the unit sphere of $\mathcal{X}_\delta$, the quantity

$$\mathcal{T}_\delta(k_1^*, k_2^*)\big(D\mathcal{G}(U_1^*, U_2^*) + D\mathcal{H}_\delta(U_1^*, U^*_{2\delta'})\big)$$

remains bounded in $\mathcal{X}$ so that the desired convergence result follows from (4.14) and (4.20), combined with (5.2) and the embedding of $H^s(\Omega_i)$ into $L^\infty(\Omega_i)$ for all $s > \frac{d}{2}$.

LEMMA 5.3. *If Hypothesis* 2.2 *is satisfied, the following property holds:*

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|\mathcal{O}_3\|_{\mathcal{L}(\mathcal{X}_\delta, \mathcal{X}_\delta)} = 0. \tag{5.10}$$

Proof. Here, the convergence of each term is a straightforward consequence of (5.2).

LEMMA 5.4. *If Hypothesis* 2.2 *is satisfied and if the data* $\boldsymbol{f}_i$, $1 \le i \le 2$, *belong to* $H^\sigma(\Omega_i)^d$ *for some* $\sigma > \frac{d}{2}$, *the following property holds:*

$$\lim_{K_1\to+\infty, K_2\to+\infty} \|\mathcal{O}_4\|_{\mathcal{L}(\mathcal{X}_\delta, \mathcal{X}_\delta)} = 0. \tag{5.11}$$

Proof. Thanks to (4.8) and its analogue for the Stokes problem, the convergence of the term

$$\|\mathcal{T}_\delta(k^*_{1\delta},k^*_{2\delta})\big(D\mathcal{G}(U^*_{1\delta'},U^*_{2\delta'})-D\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})\big)\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)}$$

is a consequence of the convergence of the terms

$$\sup_{\boldsymbol{w}_{i\delta}\in X_{i\delta}}\frac{\int_{\Omega_i}\alpha_i(k^*_{i\delta})\,\nabla\boldsymbol{u}^*_{i\delta}\,\cdot\,\nabla\boldsymbol{w}_{i\delta}\,d\boldsymbol{x}-(\alpha_i(k^*_{i\delta})\nabla\boldsymbol{u}^*_{i\delta},\nabla\boldsymbol{w}_{i\delta})_{i\delta}}{\|\boldsymbol{w}_{i\delta}\|_{H^1(\Omega_i)^d}},$$
$$\sup_{m_{i\delta}\in Y_{i\delta}}\frac{\int_{\Omega_i}\alpha_i'(k^*_{i\delta})m_{i\delta}\,\nabla\boldsymbol{u}^*_{i\delta}\,\cdot\,\nabla\boldsymbol{u}^*_{i\delta}\,d\boldsymbol{x}-(\alpha_i'(k^*_{i\delta})m_{i\delta}\,\nabla\boldsymbol{u}^*_{i\delta},\nabla\boldsymbol{u}^*_{i\delta})_{i\delta}}{\|m_{i\delta}\|_{H^{1-\varepsilon}(\Omega_i)}}$$

and of their analogues for the Stokes problem. As in the end of the proof of Lemma 4.1, this is obtained by adding and subtracting appropriate terms of lower degree.

Similar arguments yield the convergence of $\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})$ toward $\mathcal{G}(U^*_{1\delta'},U^*_{2\delta'})$, whence the convergence of the second term in $\mathcal{O}_4$.

Combining the results of Lemmas 5.1 to 5.4 leads to the following result.

PROPOSITION 5.5. *If Hypotheses* 2.2 *and* 2.4 *are satisfied and if the data* $\boldsymbol{f}_i$, $1\le i\le 2$, *belong to* $H^\sigma(\Omega_i)^d$ *for some* $\sigma>\frac{d}{2}$, *there exists a constant* K *such that, for* $K_1\ge K$ *and* $K_2\ge K$, *the operator*

$$\mathrm{Id}+D\mathcal{T}_\delta(k^*_{1\delta'},k^*_{2\delta'})\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})+\mathcal{T}_\delta(k^*_{1\delta'},k^*_{2\delta'})D\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'}) \tag{5.12}$$

is an isomorphism of $\mathcal{X}_\delta$. *Moreover, the norm of its inverse is bounded by a constant* γ *independent of* K_1 *and* K_2.

The following proposition states a Lipschitz property for the discrete operator. Since its proof is simpler than for the previous result, we only sketch it.

PROPOSITION 5.6. *The following property holds for all nonnegative real numbers* α *and for any* $(Z_{1\delta},Z_{2\delta})$ *in* $\mathcal{X}_\delta$ *which satisfies* $\|(Z_{1\delta},Z_{2\delta})-(U^*_{1\delta'},U^*_{2\delta'})\|_{\mathcal{X}}\le\alpha$, *with* $Z_{i\delta}=(\boldsymbol{z}_{i\delta},r_{i\delta})$:

$$\begin{aligned}&\|D\mathcal{T}_\delta(k^*_{1\delta'},k^*_{2\delta'})\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})+\mathcal{T}_\delta(k^*_{1\delta'},k^*_{2\delta'})D\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})\\&\quad-D\mathcal{T}_\delta(r_{1\delta},r_{2\delta})\mathcal{G}_\delta(Z_{1\delta},Z_{2\delta})-\mathcal{T}_\delta(r_{1\delta},r_{2\delta})D\mathcal{G}_\delta(Z_{1\delta},Z_{2\delta})\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)}\le c\,\kappa_\delta\,\alpha,\end{aligned} \tag{5.13}$$

where according to the dimension d *the constant* κ_δ *is given by*

$$\kappa_\delta=K_i^{2(d-2)+2\varepsilon}\,(\log K_i)^{\frac{3-d}{2}}. \tag{5.14}$$

Proof. We must bound the four terms

$$\begin{aligned}&\|\mathcal{T}_\delta(r_{1\delta},r_{2\delta})\big(D\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})-D\mathcal{G}_\delta(Z_{1\delta},Z_{2\delta})\big)\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)},\\&\|\big(\mathcal{T}_\delta(k^*_{1\delta'},r^*_{2\delta'})-\mathcal{T}_\delta(r_{1\delta},r_{2\delta})\big)D\mathcal{G}_\delta(U^*_{1\delta'},U^*_{2\delta'})\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)},\\&\|\mathcal{T}_\delta(r_{1\delta},r_{2\delta})\big(D\mathcal{H}_\delta(U^*_{1\delta'},U^*_{2\delta'})-D\mathcal{H}_\delta(Z_{1\delta'},Z_{2\delta'})\big)\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)},\\&\|\big(\mathcal{T}_\delta(k^*_{1\delta'},r^*_{2\delta'})-\mathcal{T}_\delta(r_{1\delta},r_{2\delta})\big)D\mathcal{H}_\delta(U^*_{1\delta'},U^*_{2\delta'})\|_{\mathcal{L}\ (\mathcal{X}_\delta,\mathcal{X}_\delta)}.\end{aligned}$$

For instance, in the first term, we must bound the quantity, for $m_{i\delta}$ running through the unit ball of $Y_{i\delta}$,

$$\sum_{\ell_{i\delta}\in Y_{i\delta}}\frac{\big(\alpha_i'(k^*_{i\delta'})\,m_{i\delta}\,|\nabla(\boldsymbol{u}_{i\delta'}-z_{i\delta})|^2,\ell_{i\delta}\big)_{i\delta}}{\|\ell_{i\delta}\|_{H^1(\Omega_i)}}.$$

Applying the inverse inequalities

$$\forall \varphi_{i\delta} \in \mathbb{P}_{K_i,N_i}(\Omega_i), \quad \|\varphi_{i\delta}\|_{L^\infty(\Omega_i)} \le \begin{cases} c\, K_i^{d-2+2\varepsilon}\, \|\varphi_{i\delta}\|_{H^{1-\varepsilon}(\Omega_i)}, \\ c\, K_i^{d-2}\, (\log K_i)^{\frac{3-d}{2}}\, \|\varphi_{i\delta}\|_{H^1(\Omega_i)}, \end{cases}$$

we have to bound the term, for all $m_{i\delta}$ in $\mathbb{P}_{K_i,N_i}(\Omega_i)$,

$$K_i^{2(d-2)+2\varepsilon}\, (\log K_i)^{\frac{3-d}{2}}\, \|\nabla(\boldsymbol{u}^*_{i\delta'} - \boldsymbol{z}_{i\delta})\|_{L^2(\Omega_i)}\, \|\nabla(\boldsymbol{u}^*_{i\delta'} + \boldsymbol{z}_{i\delta})\|_{L^2(\Omega_i)}.$$

This yields the value of κ_δ. The other terms are simpler; they can be evaluated by similar arguments.

Finally, we must evaluate the quantity

$$\varepsilon_\delta = \left\| \begin{pmatrix} U^*_{1\delta'} \\ U^*_{2\delta'} \end{pmatrix} + \mathcal{T}_\delta(k^*_{1\delta'}, k^*_{2\delta'})\, \mathcal{G}_\delta(U^*_{1\delta'},\ U^*_{2\delta'}) \right\|_{\mathcal{X}}. \tag{5.15}$$

Proposition 5.7. *If Hypothesis* 2.2 *is satisfied and if the data* $\boldsymbol{f}_i$, $1 \le i \le 2$, *belong to* $H^\sigma(\Omega_i)^d$ *for some* $\sigma > \frac{d}{2}$, *the following estimate holds for a constant* C *only depending on the norm of* (U_1^*, U_2^*) *in* $\prod_{i=1}^2 \big(H^{s^*}(\Omega_i)^d \times H^{s^*}(\Omega_i)\big)$*:*

$$\varepsilon_\delta \le C\, (\inf\{K_1, K_2\})^{1-s^*}. \tag{5.16}$$

Proof. By using formulation (2.18), we observe that

$$\begin{aligned} \varepsilon_\delta \le{}& \|U_1^* - U^*_{1\delta}\|_{\mathcal{X}_1} + \|U_2^* - U^*_{2\delta}\|_{\mathcal{X}_2} \\ &+ \|(\mathcal{T} - \mathcal{T}_\delta)(k_1^*, k_2^*)\, \mathcal{G}(U_1^*,\ U_2^*)\|_{\mathcal{X}} + \|\big(\mathcal{T}_\delta(k_1^*, k_2^*) - \mathcal{T}_\delta(k^*_{1\delta'}, k^*_{2\delta'})\big)\, \mathcal{G}(U_1^*,\ U_2^*)\|_{\mathcal{X}} \\ &+ \|\mathcal{T}_\delta(k^*_{1\delta'}, k^*_{2\delta'})\, \big(\mathcal{G}(U_1^*,\ U_2^*) - \mathcal{G}(U^*_{1\delta'},\ U^*_{2\delta'})\big)\|_{\mathcal{X}} \\ &+ \|\mathcal{T}_\delta(k^*_{1\delta'}, k^*_{2\delta'})\, \big(\mathcal{G}(U^*_{1\delta'},\ U^*_{2\delta'}) - \mathcal{G}_\delta(U^*_{1\delta'},\ U^*_{2\delta'})\big)\|_{\mathcal{X}}. \end{aligned}$$

The bound for the first two terms in the right-hand side comes from (5.2), and the bound for the third term is derived from (4.9) and its analogue for the Stokes problem together with the regularity Hypothesis 2.2. The fourth term is easily bounded from (4.14) and (4.20), while estimating the fifth one relies on (4.7) and (4.17), combined with (5.2). Finally, estimating the last term also relies on (4.7) and (4.17) together with the introduction of approximations of $\alpha_i(k_i^*)$ and $\gamma_i(k_i^*)$ in $\mathbb{P}_{K'_i,N'_i}(\Omega_i)$; see (4.12).

We are now in a position to apply the Brezzi–Rappaz–Raviart theorem [10] (see also [15, Chap. IV, Thm. 3.1]).

Theorem 5.8. *If Hypotheses* 2.2 *and* 2.4 *are satisfied with* $s^* > 2(d-2)+1$ *and if the data* $\boldsymbol{f}_i$, $i = 1, 2$, *belong to* $H^\sigma(\Omega_i)^d$ *for some* $\sigma > \sup\{\frac{d}{2}, s^* - 1\}$, *there exist an integer* K_0 *and a constant* λ *such that, for* $K_1 \ge K_0$ *and* $K_2 \ge K_0$, *problem* (3.4)–(3.7) *has a unique solution* $(U_{1\delta}, U_{2\delta})$, *with each* $U_{i\delta}$ *equal to* $(\boldsymbol{u}_{i\delta}, k_{i\delta})$, *in the neighborhood* U *of* $U = (U_1^*, U_2^*)$ *defined as follows:*

$$\mathsf{U} = \Big\{ Z_\delta = (Z_{1\delta}, Z_{2\delta}) \in \mathcal{X}_\delta;\ \|U - Z_\delta\|_{\mathcal{X}} \le \lambda\, \kappa_\delta^{-1} \Big\}. \tag{5.17}$$

Moreover, the following error estimates hold for $i = 1$ *and* 2*:*

$$\|\boldsymbol{u}_i^* - \boldsymbol{u}_{i\delta}\|_{H^1(\Omega_i)^d} + \|k_i^* - k_{i\delta}\|_{L^2(\Omega_i)} \le C_0\, (\inf\{K_1, K_2\})^{1-s^*} \tag{5.18}$$

for a constant C_0 *depending only on the norms of* (U_1^*, U_2^*) *in* $\prod_{i=1}^2 (H^{s^*}(\Omega_i)^d \times H^{s^*}(\Omega_i))$ *and of* $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ *in* $\prod_{i=1}^2 H^\sigma(\Omega_i)^d$.

We conclude with an estimate on the pressure, which is easily derived from the inf-sup condition established in Lemma 3.4.

COROLLARY 5.9. *If the assumptions of Theorem* 5.8 *are satisfied, for the solution* $(U_{1\delta}, U_{2\delta})$ *exhibited in Theorem* 5.8, *there exists a unique pair* $(p_{1\delta}, p_{2\delta})$ *in* $M^m_{1\delta} \times M^m_{2\delta}$ *such that* $(\tilde{U}_{1\delta}, \tilde{U}_{2\delta})$, *with each* $\tilde{U}_{i\delta}$ *equal to* $(\boldsymbol{u}_{i\delta}, p_{i\delta}, k_{i\delta})$, *is a solution of problem* (3.4)–(3.5). *Moreover, the following error estimates hold for* $i = 1$ *and* 2*:*

$$\|p_i^* - p_{i\delta}\|_{L^2(\Omega_i)} \le C_0' \, (\beta_{i\delta}^m)^{-1} \, (\inf\{K_1, K_2\})^{1-s^*} \tag{5.19}$$

for the constants $\beta_{i\delta}^m$ *evaluated in* (3.9) *and a constant* C_0' *depending only on the norms of* $(\tilde{U}_1^*, \tilde{U}_2^*)$ *in* $\prod_{i=1}^2 (H^s(\Omega_i)^d \times H^{s-1}(\Omega_i) \times H^s(\Omega_i))$ *and of* $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ *in* $\prod_{i=1}^2 H^\sigma(\Omega_i)^d$.

6. Conclusions and numerical algorithms. The regularity assumptions in Theorem 5.8 in the three-dimensional case are very unlikely; however, they seem unavoidable. This comes from the fact that the linearized problem (2.20) and (2.21) makes sense only for a smooth solution U^*. Nevertheless, this does not prevent the convergence of numerical experiments.

In contrast, the assumptions of Theorem 5.8 are fully reasonable in the two-dimensional case and, if these assumptions hold, optimal error estimates are derived for the velocity, the kinetic energy, and the pressure for appropriate choices of the spaces $M_{i\delta}$. In this case, the maximal regularity s^* would very likely coincide with the $s_0 \simeq 1.5946$ introduced in section 2 so that the error would be smaller than $C_0 \, (\inf\{K_1, K_2\})^{-0.5946}$. Moreover, in the case of the rectangle, the explicit form of the singular functions associated with the Stokes operator with Dirichlet–Neumann boundary conditions is known [20]. As usual [6], this would lead to double the convergence order: for smooth enough data $\boldsymbol{f}_i$, $i = 1$ and 2, the error would be smaller than $C_0 \, (\inf\{K_1, K_2\})^{-1.1892}$.

The most standard choice of an operator $\Pi_{i\delta}^\Gamma$ satisfying (5.3) would be the orthogonal projection operator $\Pi_{K_i}^{\frac{1}{2}}$ from $H_{00}^{\frac{1}{2}}(\Gamma)$ onto $\mathbb{P}_{K_i}(\Gamma) \cap H_{00}^{\frac{1}{2}}(\Gamma)$; however, computing this operator is not easy. So we take $\Pi_{i\delta}^\Gamma$ equal to $\mathcal{I}_{i\delta}^\Gamma \, \Pi_{K_i}^{\frac{1}{2}}$. Indeed, estimate (5.3) is still satisfied by this operator and, in this case, the boundary conditions at the interface can be written in a very simple way:

$$k_{i\delta} = \mathcal{I}_{i\delta}^\Gamma(|\boldsymbol{u}_{1\delta} - \boldsymbol{u}_{2\delta}|^2) \quad \text{on} \quad \Gamma.$$

Note, moreover, that in the present situation there is no theoretical reason to choose $K_1 \neq K_2$ and that, when K_1 and K_2 coincide, these conditions are still less expensive to enforce. However, for more complex geometries (for instance, if the Ω_i are convex quadrilaterals), different values of K_1 and K_2 can be needed since the regularity properties of the velocities in Ω_1 and Ω_2 are different.

Exactly the same arguments as for Theorem 5.8 prove [10] the convergence of Newton's algorithm for solving the nonlinear problem (3.4)–(3.5), when the initial guess $(U_{1\delta}^0, U_{2\delta}^0)$ belongs to the domain U introduced in (5.17); however, this method seems too expensive for the present problem. Instead of this, for an initial guess $(U_{1\delta}^0, U_{2\delta}^0)$, we propose to solve iteratively the following problem: if the pair $(\tilde{U}_{1\delta}^n, \tilde{U}_{2\delta}^n)$, with $\tilde{U}_{i\delta}^n = (\boldsymbol{u}_{i\delta}^n, p_{i\delta}^n, k_{i\delta}^n)$, is supposed to be known, then we solve the following problem:

Find $(\boldsymbol{u}_{i\delta}^{n+1}, p_{i\delta}^{n+1})$ *in* $X_{i\delta} \times M_{i\delta}$ *such that*

$$\begin{aligned}
&\forall \boldsymbol{v}_{i\delta} \in X_{i\delta},\\
&\quad a_{i\delta}(k_{i\delta}^n; \boldsymbol{u}_{i\delta}^{n+1}, \boldsymbol{v}_{i\delta}) + b_{i\delta}(\boldsymbol{v}_{i\delta}, p_{i\delta}^{n+1})\\
&\qquad\qquad + \left(|\boldsymbol{u}_{i\delta}^n - \boldsymbol{u}_{j\delta}^n| \, (\boldsymbol{u}_{i\delta}^{n+1} - \boldsymbol{u}_{j\delta}^{n+1}), \boldsymbol{v}_{i\delta} \right)_{i\delta}^{\Gamma} = \left(\boldsymbol{f}_i, \boldsymbol{v}_{i\delta} \right)_{i\delta},\\
&\forall q_{i\delta} \in M_{i\delta}, \quad b_{i\delta}(\boldsymbol{u}_{i\delta}^{n+1}, q_{i\delta}) = 0;
\end{aligned} \tag{6.1}$$

Find $k_{i\delta}$ *in* $\mathbb{P}_{K_i,N_i}(\Omega_i)$*, such that*

$$k_{i\delta}^{n+1} = 0 \quad \text{on } \Gamma_i \quad \text{and} \quad k_{i\delta}^{n+1} = \mathcal{I}_{i\delta}^{\Gamma}(|\boldsymbol{u}_{1\delta}^{n+1} - \boldsymbol{u}_{2\delta}^{n+1}|^2) \quad \text{on } \Gamma, \tag{6.2}$$

and

$$\forall \varphi_{i\delta} \in Y_{i\delta}, \quad c_{i\delta}(k_{i\delta}^n; k_{i\delta}^{n+1}, \varphi_{i\delta}) = \left(\alpha_i(k_{i\delta}^n) \, |\nabla \boldsymbol{u}_{i\delta}^{n+1}|^2, \varphi_{i\delta} \right)_{i\delta}. \tag{6.3}$$

Clearly, these two linear problems are well-posed and the sequence $(\tilde{U}_{1\delta}^n, \tilde{U}_{2\delta}^n)_n$ converges. Numerical experiments to check the efficiency of the algorithm are under consideration.

REFERENCES

[1] Y. Achdou and O. Pironneau, *Domain decomposition and wall laws*, C. R. Acad. Sci. Paris Sér. I, 320 (1995), pp. 541–547.

[2] C. Bernardi, T. Chacón Rebollo, R. Lewandowski, and F. Murat, *Existence d'une solution pour un modèle de deux fluides turbulents couplés*, C. R. Acad. Sci. Paris Sér. I, 328 (1999), pp. 993–998.

[3] C. Bernardi, T. Chacón Rebollo, R. Lewandowski, and F. Murat, *A model for two coupled turbulent fluids. Part* I: *Analysis of the system*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XIV, D. Cioranescu and J.-L. Lions, eds., Elsevier, Amsterdam, 2002, pp. 69–102.

[4] C. Bernardi, M. Dauge, and Y. Maday, *Relèvement de traces préservant les polynômes*, C. R. Acad. Sci. Paris Sér. I, 315 (1992), pp. 333–338.

[5] C. Bernardi, M. Dauge, and Y. Maday, *Interpolation of nullspaces for polynomial approximation of divergence-free functions in a cube*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains, Lecture Notes in Pure and Appl. Math. 167, M. Costabel, M. Dauge, and S. Nicaise, eds., Dekker, New York, 1995, pp. 27–46.

[6] C. Bernardi and Y. Maday, *Polynomial approximation of some singular functions*, Appl. Anal., 42 (1991), pp. 1–32.

[7] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis V, P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[8] C. Bernardi and Y. Maday, *Uniform inf-sup conditions for the spectral discretization of the Stokes problem*, Math. Models Methods Appl. Sci., 9 (1999), pp. 395–414.

[9] J.M. Boland and R.A. Nicolaides, *Stability of finite elements under divergence constraints*, SIAM J. Numer. Anal., 20 (1983), pp. 722–731.

[10] F. Brezzi, J. Rappaz, and P.-A. Raviart, *Finite dimensional approximation of nonlinear problems, Part* I: *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1–25.

[11] F. Brossier and R. Lewandowski, *Impact of the variations of the mixing length in a first order turbulent closure system,*, M2AN Modél. Math. Anal. Numér., 36 (2002), pp. 345–372.

[12] J. Casado Diaz, T. Chacón Rebollo, M. Gómez Marmol, V. Girault, and F. Murat, *Numerical approximation of a Laplace equation with* L_1 *data*, submitted.

[13] C.M. Elliott and S. Larsson, *A finite element model for the time-dependent Joule heating problem*, Math. Comput., 64 (1995), pp. 1433–1453.

[14] D.A. French and S.M.F. Garcia, *Finite element approximation of an evolution problem modeling shear band formation*, Comput. Methods Appl. Mech. Engrg., 118 (1994), pp. 153–161.

[15] V. Girault and P.-A. Raviart, *Finite Element Methods for the Navier–Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[16] R. Lewandowski, *Analyse mathématique et océanographie*, Collection Recherches en Mathématiques Appliquées, Masson, Paris, 1997.

[17] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.

[18] J.-L. Lions, R. Temam, and S. Wang, *Models for the coupled atmosphere and ocean*, Comput. Mech. Adv., 1 (1993), pp. 1–120.

[19] Y. Maday, *Relèvement de traces polynomiales et interpolations hilbertiennes entre espaces de polynômes*, C. R. Acad. Sci. Paris Sér. I, 309 (1989), pp. 463–468.

[20] M. Orlt and A.-M. Sändig, *Regularity of viscous Navier–Stokes flows in nonsmooth domains*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains, Lecture Notes in Pure and Appl. Math. 167, M. Costabel, M. Dauge, and S. Nicaise, eds., Dekker, New York, 1995, pp. 185–201.

[21] C. Parès, *Existence, uniqueness and regularity of solution of the equations of a turbulence model for incompressible fluids*, Appl. Anal., 43 (1992), pp. 245–296.

[22] G. Sacchi Landriani and H. Vandeven, *Polynomial approximation of divergence-free functions*, Math. Comput., 52 (1989), pp. 103–130.

[23] G. Stampacchia, *Équations elliptiques du second ordre à coefficients discontinus*, Presses de l'Université de Montréal, Montréal, QB, Canada, 1966.