MATHEMATICS OF COMPUTATION
VOLUME 59, NUMBER 199
JULY 1992, PAGES 1–20

THE EFFECT OF NUMERICAL QUADRATURE IN THE p-VERSION OF THE FINITE ELEMENT METHOD

UDAY BANERJEE AND MANIL SURI

ABSTRACT. We investigate the use of numerical quadrature in the p-version of the finite element method. We describe a set of minimal conditions that the quadrature rules should satisfy, for various types of elements. Under sufficient assumptions of smoothness, we establish optimality of the asymptotic rate of convergence. Some computational results are presented, which illustrate under what conditions overintegration may be useful.

1. INTRODUCTION

The accuracy in terms of the asymptotic rate of convergence of finite element calculations is affected by many factors. One of these is the accuracy of quadrature schemes employed to calculate the components of the stiffness (and mass) matrix and the load vector. The classical finite element method, called the h-version, increases accuracy by refining the mesh while keeping fixed the degree p of piecewise polynomials used. For this, the dependence of the convergence rate on the accuracy of quadrature used has been thoroughly investigated (see, for instance, [4, 5]). The basic rule to ensure optimal (asymptotic) convergence is to use an integration scheme for which the asymptotic order of the error between u^h (the approximation using exact integration) and $\tilde{u}^h$ (the approximation using numerical integration) is no worse than that for the error between u^h and u (the exact solution). Suppose we consider a second-order elliptic problem. Then for piecewise polynomials of degree p on a quasiuniform family of triangular meshes, for example, this is achieved by ensuring that the quadrature scheme is exact for all polynomials of degree $\leq 2p-2$ on every triangle.

The last decade has seen the rise in popularity of two new versions of the finite element method, the p- and $(h-p)$-versions. In the p-version, the mesh is kept fixed and the degree p of polynomials used is increased for accuracy. The $(h-p)$-version changes both h and p. Although there are currently several commercial codes available that implement the p- and $(h-p)$-versions (for instance MSC/PROBE, FIESTA, and the research code STRIPE), the problem

Received by the editor April 30, 1991.
1991 *Mathematics Subject Classification.* Primary 65D30, 65N30.
The work of the first author was supported in part by the Office of Naval Research under Naval Research Grant N00014-90-J-1030 and by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant #AFOSR 89-0252.
The work of the second author was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant #AFOSR 89-0252.

0025-5718/92 $1.00 + $.25 per page

of what quadrature scheme to use has been treated only in an ad hoc manner. In this paper, we investigate in a systematic manner the effect of quadrature error on the accuracy of the finite element computations when various types of rectilinear and curved elements are used.

Our first goal is to list some minimal requirements (analogous to the h-version rule mentioned above) on the quadrature scheme to ensure existence, uniqueness, and convergence of the approximations. These may be found in §3, for various quadrature schemes used for the p-version. Our main goal is to establish under sufficient conditions (also listed in §3) that the optimal rate of convergence is achieved even under the effect of quadrature error. We deal with the case that the input data are smooth and the curvilinear elements used are obtained through sufficiently smooth mappings. For this case, we show that the asymptotic rate is preserved, essentially without any overintegration being required, both for smooth and singular solutions (§5). Section 6 contains some computational examples which illustrate our results. In this section, we also briefly investigate the effectivity of overintegration when distorted elements are used.

Let us mention that there are some references [3, 13] that address the problem of numerical integration as applied to the spectral element method (which is related to the p-version). Our approach is broader, since it is not restricted to tensor-product elements (see Remark 5.1).

2. The p-version for the model problem

Let Ω be a curvilinear polygonal domain in R^n, $n = 1, 2, 3$, and consider the problem

$$\begin{aligned} Lu &= f \quad \text{on } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where

$$Lu = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right).$$

The variational form of this problem is

$$\text{(2.1)} \qquad \begin{cases} u \in H_0^1(\Omega), \\ a(u, v) = (f, v) \quad \forall v \in H_0^1(\Omega), \end{cases}$$

where

$$a(u, v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx, \qquad (f, v) = \int_\Omega fv\,dx.$$

Here, we use the standard notation for Sobolev spaces $W^{\alpha,\beta}(\Omega)$, $1 \le \beta \le \infty$, $H^\alpha(\Omega) = W^{\alpha,2}(\Omega)$, and $H_0^1(\Omega) = \{u \in H^1(\Omega): u = 0 \text{ on } \partial\Omega\}$. We will also use the usual notation $\|\cdot\|_{\alpha,\beta,\Omega}$ and $|\cdot|_{\alpha,\beta,\Omega}$ to denote the corresponding norms and seminorms, with the subscript β being dropped for $\beta = 2$.

We will assume that the functions $a_{ij}(x)$ satisfy $a_{ij} = a_{ji}$ and

$$\text{(2.2)} \qquad \|a_{ij}\|_{\alpha,\Omega} \le A_1$$

for a fixed constant A_1. Here, α will be a sufficiently large number, $\alpha > \frac{n}{2}$. Also, the operator L is assumed to be uniformly elliptic, i.e.,

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \kappa \sum_{i=1}^{n} \xi_i^2 \quad \forall \xi \in R^n, \forall x \in \overline{\Omega} \tag{2.3}$$

for some $\kappa > 0$.

Several regularity results are well known for the above problem (2.1). For the case that $\partial\Omega$ is smooth (C^∞), the usual shift theorems hold, and we have, for $k \geq 2$,

$$\|u\|_{k,\Omega} \leq C\|f\|_{k-2,\Omega}. \tag{2.4}$$

For $\partial\Omega$ nonsmooth, (2.4) will only hold for $k \leq k_0$, k_0 being determined by the domain.

When $\Omega \subset R^2$ is a curvilinear polygon, the solution consists of two parts, a smooth component u_1 for which (2.4) holds and a component u_2 which arises from the singularities at the corners of the domain. More precisely, if V denotes the set of vertices v of Ω, we have [10]

$$u = u_1 + u_2, \qquad u_1, u_2 \in H_0^1(\Omega), \quad u_2 = \sum_{v \in V} u_v \chi_v, \tag{2.5}$$

where χ_v is a C^∞ (or sufficiently smooth) cutoff function with support near v, and

$$u_v = \sum_{l=1}^{M} \sum_{k=0}^{N(l)} d_{kl} |\log r|^k r^{\alpha_l} \phi_{kl}(\theta), \tag{2.6}$$

where $\alpha_l > 0$, $\alpha_{l+1} > \alpha_l$, $N(l) \geq 0$, $\phi_{kl}(\theta)$ is a C^∞ (or sufficiently smooth) function, and (r, θ) are polar coordinates with origin at v. In the three-dimensional case, the situation is similar, but more complicated.

To approximate the solution of problem (2.1) by the finite element method (FEM), we consider a *fixed* triangulation τ of $\overline{\Omega}$ by elements K_i. These will be line segments in R^1, triangles and quadrilaterals in R^2, and tetrahedra and parallelepipeds in R^3. We will also consider the corresponding curvilinear elements. The intersection $K_i \cap K_j$ will be either empty, a common vertex, an entire side, or an entire face of K_i and K_j. We assume that every corner vertex of the domain $\overline{\Omega}$ is also a vertex of some K_i.

For each type of element under consideration, we define a corresponding reference element. Accordingly, with $I = [-1, 1]$, we define the reference interval, the reference square, and the reference cube by I, $Q = I^2$, and $C = I^3$, respectively. Also, we define by T the reference triangle with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$, and by B the reference tetrahedron with vertices $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$.

Remark 2.1. We could consider other 3-d elements as well, like wedges and pyramids, for which results analogous to ours may be obtained. Some such elements are in use in 3-d p-version codes like STRIPE.

We assume that for each $K \in \tau$, there exists an invertible map F_K such that $K = F_K(\hat{K})$, where $\hat{K}$ is the reference element corresponding to K, i.e., $\hat{K} = I, T, Q, B$, or C. This mapping then establishes the correspondence

$(\hat{v}: \widehat{K} \to R) \leftrightarrow (v = \hat{v} \circ F_K^{-1}: K \to R)$ between the functions defined on $\widehat{K}$ and K. In this paper, we will deal with the case that F_K and F_K^{-1} are sufficiently smooth and the Jacobian J_K (assumed to be positive everywhere) is bounded below, away from zero. Accordingly, we assume that there exists a constant A_2 such that, for all $K \in \tau$,

$$\|F_K\|_{j,\infty,\widehat{K}}, \|F_K^{-1}\|_{j,\infty,K} \le A_2, \qquad 0 \le j \le M, \tag{2.7}$$

$$\|J_K\|_{j,\infty,\widehat{K}}, \|J_K^{-1}\|_{J,\infty,K} \le A_2, \qquad 0 \le j \le M-1, \tag{2.8}$$

where $J_K^{-1} = (J_K)^{-1}$ is the Jacobian of F_K^{-1} and where $M \ge 1$ is large enough. In essence, assuming M is large ensures that the elements used are not distorted in any significant way, so that the convergence analysis proceeds similarly to the case that the F_K are affine. We remark that for nonsmooth mappings, (2.7)–(2.8) may still hold, but the constant A_2 may be very large (see the numerical example in §6).

Using (2.7)–(2.8) and the argument of Theorem 4.3.2 of [4], we see that for any $\beta \in [1, \infty]$ and l, $0 \le l \le M$, one has $\hat{v} \in W^{l,\beta}(\widehat{K}) \Leftrightarrow v \in W^{l,\beta}(K)$ with

$$C_1\|v\|_{l,\beta,K} \le \|\hat{v}\|_{l,\beta,\widehat{K}} \le C_2\|v\|_{l,\beta,K}, \tag{2.9}$$

where the constants C_1 and C_2 are independent of v and $\hat{v}$ but depend on l, β, and A_2. Moreover, the norms in (2.9) may be replaced by seminorms when $l = 0, 1$.

Now let $\{U_p(\widehat{K})\}$, $p = 1, 2, \ldots$, be a sequence of polynomial spaces defined on the reference element $\widehat{K}$ such that

$$w_1 \in U_p,\ w_2 \in U_q \Rightarrow w_1 w_2 \in U_{p+q}. \tag{2.10}$$

We then define finite-dimensional spaces for $p = 1, 2, 3, \ldots$ by

$$S_p = \{v \in C^0(\Omega),\ v|_{K_i} \circ F_{K_i} \in U_p(\widehat{K}_i)\ \forall K_i \in \tau\},$$

where $\widehat{K}_i$ is the reference element corresponding to K_i and $U_p(\widehat{K}_i)$ is the corresponding polynomial space. (Note that more than one type of space may be in use for the same type of reference element.) We also define $S_{p,0} = S_p \cap H_0^1(\Omega)$. Then the p-version of the FEM to approximate the solution of (2.1) is given by

$$\begin{cases} u_p \in S_{p,0}, \\ a(u_p, v) = (f, v) \quad \forall v \in S_{p,0}. \end{cases} \tag{2.11}$$

It is well known that for (2.11) one has

$$\|u - u_p\|_{1,\Omega} \le C \inf_{v \in S_{p,0}} \|u - v\|_{1,\Omega}. \tag{2.12}$$

Let us define the H_0^1 projection operator $P_p^1: H_0^1(\Omega) \to S_{p,0}$ by

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega \nabla(P_p^1 u) \cdot \nabla v \quad \forall v \in S_{p,0}. \tag{2.13}$$

Then, similarly to (2.12), we have

$$\|u - P_p^1 u\|_{1,\Omega} \le C \inf_{v \in S_{p,0}} \|u - v\|_{1,\Omega}. \tag{2.14}$$

To bound the right sides of (2.12) and (2.14), one needs subspaces with certain approximability properties. Let us describe the subspaces we consider.

For $\hat{K} = I$, Q, or C, we can take $U_p(\hat{K}) = Q_p(\hat{K})$, the set of polynomials of degree $\leq p$ in each variable. For $\hat{K} = T$ or B, we take $U_p(\hat{K}) = \mathscr{P}_p(\hat{K})$, the set of polynomials of total degree $\leq p$. (Note that for $\hat{K} = I$, $Q_p(I) = \mathscr{P}_p(I)$.) For $\hat{K} = Q \subset R^2$, the choice $U_p(\hat{K}) = Q'_p(\hat{K})$ (serendipity elements) given by

$$Q'_p = \mathscr{P}_p \oplus \{x_1^p x_2, x_1 x_2^p\}$$

is also used in practice. However, it does not satisfy (2.10). We therefore treat this choice by noting that $Q'_{p-1} \subset \mathscr{P}_p$. We now add to our list of possible U_p's the choice $U_p(\hat{K}) = \mathscr{P}_p(\hat{K})$ for $\hat{K} = Q$ $(p > 1)$. With the above choices of $U_p(\hat{K})$, we have the following lemma.

Lemma 2.1. *For each integer $l \geq 0$, there exists a sequence of projections $\Pi_p^l : H^l(\hat{K}) \to U_p(\hat{K})$, $p = 1, 2, 3, \ldots$, such that*

$$\Pi_p^l v = v \quad \forall v \in U_p(\hat{K}),$$
$$\|v - \Pi_p^l v\|_{s,\hat{K}} \leq C p^{-(r-s)} \|v\|_{r,\hat{K}}, \qquad 0 \leq s \leq l \leq r,$$

where C is a constant independent of p and v but dependent upon r and l.

Proof. The lemma has been proven for the 1-d case in [2]. In 2-d, it follows from Lemma 3.1 of [15] for the cases $U_p = Q_p(Q)$ and $\mathscr{P}_p(T)$. This proof generalizes easily to the 3-d case (as well as to $U_p = Q'_p(Q)$). □

Remark 2.2. The restriction of l being an integer is not necessary. This has been proven in [2] for the 1-d case.

The above lemma with $l = 1$ is used in the proof of the following theorem.

Theorem 2.1. *Let $\{U_p(\hat{K})\}$ be a sequence of polynomial spaces described above and let $\{S_{p,0}\}$ be the corresponding spaces on Ω. Then the sequence of projections $P_p^1 : H_0^1(\Omega) \to S_{p,0}(\Omega)$, $p = 1, 2, 3, \ldots$, defined by* (2.13), *satisfies*

$$\|v - P_p^1 v\|_{1,\Omega} \leq C p^{-(r-1)} \|v\|_{r,\Omega}, \qquad 1 \leq r,$$

with C a constant independent of p and v but dependent upon r and the constant A_2.

Proof. The two-dimensional case has been proven in [1], the argument from which can be generalized to the 3-dimensional case. A different proof of the n-dimensional case may be found in [7], the result being optimal up to arbitrary $\varepsilon > 0$. □

For the case that the solution $u \in H^k(\Omega)$, we see by (2.12) and Theorem 2.1 that

$$\|u - u_p\|_{1,\Omega} \leq C p^{-(k-1)} \|u\|_{k,\Omega}. \tag{2.15}$$

Here, the constant C will depend upon A_2. For nonsmooth mappings, C can be quite large, which effectively means that the above rate of convergence may not be observed for practically chosen discretization levels in such cases.

For nonsmooth domains, the rate of convergence is dominated instead by singularities of the form u_v in (2.6). In the 2-d case, assuming that $u \equiv u_v$, we have the estimate [1]:

$$\inf_{v \in S_{p,0}} \|u - v\|_{1,\Omega} \leq C |\log p|^{\gamma_1} p^{-2\alpha_1}, \tag{2.16}$$

where $\gamma_1 = N(1)$ and where the constant C depends on d_{kl} but is independent of p. Note that (2.16) gives twice the rate of convergence that can be derived using Theorem 2.1.

Remark 2.3. Let us mention that if u is very smooth, the rate of convergence can be stronger than that in (2.15). For a wide class of C^∞ functions, one obtains exponential convergence (see, for example, [11]).

3. Quadrature rules

The p-version introduced in the previous section, and the related results for it, assume that all integrations have been performed exactly. In practice, however, the integrals in the terms $a(u_p, v)$ and (f, v) in (2.1) are usually evaluated numerically. We consider families of quadrature rules $\{R_p\}$ defined on the master element $\hat{K}$ by

$$\int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x} \sim \sum_{l=1}^{L_p} \hat{\omega}_l^p \hat{v}(\hat{b}_l^p). \tag{3.1}$$

This results in a quadrature rule R_p^K over each $K \in \tau$ given by

$$\int_K v(x)\, dx \sim \sum_{l=1}^{L_p} \omega_{l,K}^p v(b_{l,K}^p), \tag{3.2}$$

where $\omega_{l,K}^p = J_K(\hat{b}_l^p)\hat{\omega}_l^p$ and $b_{l,K}^p = F_K(\hat{b}_l^p)$. Now, if the various integrals in (2.11) are evaluated by quadrature rules, then instead of solving problem (2.11), one solves

$$\begin{cases} \tilde{u}_p \in S_{p,0}, \\ a_p(\tilde{u}_p, v) = (f, v)_p \quad \forall v \in S_{p,0}, \end{cases} \tag{3.3}$$

where

$$a_p(u, v) = \sum_{K \in \tau} a_{p,K}(u, v) = \sum_{K \in \tau} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} \left(a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \right)(b_{l,K}), \tag{3.4}$$

$$(f, v)_p = \sum_{K \in \tau} (f, v)_{p,K} = \sum_{K \in \tau} \sum_{l=1}^{L} \omega_{l,K}(fv)(b_{l,K}), \tag{3.5}$$

where the dependence on p is understood in $\omega_{l,K}$, $b_{l,K}$, and L. We remark that, although we have used the same quadrature rules to evaluate all the integrals, different quadrature rules may be used to evaluate different integrals in (2.11).

For the master element $\hat{K}$, let the associated polynomial space be $U_p(\hat{K})$. Then we restrict our attention to families of quadrature rules $\{R_p\}$ that satisfy the following four assumptions:

(A) $\hat{\omega}_l^p > 0$ and $\hat{b}_l^p \in \hat{K}$.

(B1) There exists a constant C_1 independent of p and $\hat{v}$ such that

$$\sum_{l=1}^{L_p} \hat{\omega}_l^p \hat{v}^2(\hat{b}_l^p) \le C_1 \|\hat{v}\|_{0,\hat{K}}^2, \qquad \hat{v} \in U_p(\hat{K}).$$

(B2) There exists a constant C_2 independent of p and $\hat{v}$ such that

$$\sum_{l=1}^{L_p} \hat{\omega}_l^p \hat{v}^2(\hat{b}_l^p) \ge C_2 \|\hat{v}\|_{0,\hat{K}}^2, \qquad \hat{v} \in \tilde{U}_p(\hat{K}),$$

where $\tilde{U}_p(\hat{K}) = \{\partial \hat{v}/\partial \hat{x}_i,\ 1 \le i \le n \mid \hat{v} \in U_p(\hat{K})\} \subset U_p(\hat{K})$.

(B3) R_p is exact for all $\hat{v} \in U_m(\hat{K})$ with $m = m(p) \ge m_0(p)$.

Obviously, (B1)–(B2) follow from (B3) when $m \ge 2p$, so that, when this happens, we may replace (B1), (B2), and (B3) by the single assumption:

(B) R_p is exact for all $\hat{v} \in U_m(\hat{K})$ with $m \ge 2p$.

However, (B) is not necessary for (B1)–(B2) to hold (see examples below). In fact, it is easy to construct examples where (B1)–(B2) hold but (B) (or (B3)) is not satisfied for any m. The conditions (A) and (B2) guarantee "V_h ellipticity" in the sense of [4], while (B1) and (B3) will be needed to derive our error estimates. Below, we list the minimum value of $m_0(p)$ for some rules.

Remark 3.1. We see from (B2) that as $p \to \infty$, we must have $L_p \to \infty$. In fact, in 1-d, (B2) implies we must have $L_p \ge p$, since otherwise we may construct a $\hat{v} \in \tilde{U}_p(\hat{K}) = P_{p-1}(\hat{K})$ with roots at $\{\hat{b}_l^p\}$ for which (B2) is violated. This is in contrast to the h-version, where a fixed (composite) rule is used as $h \to 0$. In practice, however, p can be increased only up to a fixed value (for example, $p \le 8$ in MSC/PROBE), so that one could use a fixed rule of sufficiently high precision.

Let us describe some commonly used rules and see if they satisfy our assumptions.

Newton-Cotes rules. These rules (which can be used for the h-version for small p) are known to violate the positivity of the weights in (A) when p is large, and moreover have low degree of precision, i.e., m in (B3) is small. Hence we do not discuss them here.

Gaussian rules for $\mathcal{P}_p(I)$. We consider Gauss-Legendre and Gauss-Lobatto rules, both of which satisfy (A) (see [6]). If L_p points are used, then these rules are, respectively, exact for polynomials of degree $\le 2L_p - 1$ and $2L_p - 3$. Hence, condition (B) is satisfied if, respectively, $L_p \ge p + 1$ or $L_p \ge p + 2$.

To see if this requirement may be relaxed, we first consider Gauss-Legendre rules. Let $L_p = p$ (by Remark 3.1, this is the minimum possible). Then (B2) obviously holds, with $C_2 = 1$. Also, by [6, p. 75], we have for some $-1 < \xi < 1$,

$$\|\hat{v}\|_{0,\hat{K}}^2 - \sum_{l=1}^{p} \hat{w}_l^p \hat{v}^2(\hat{b}_l^p) = \frac{2^{2p+1}(p!)^4}{(2p+1)[(2p)!]^3} \frac{d^{2p}\hat{v}^2(\xi)}{dx^{2p}},$$

which is positive, since the $(2p)$th derivative of $\hat{v}^2$ $(\hat{v} \in \mathcal{P}_p(I))$ is a positive constant on I. Hence, (B1) holds with $C_1 = 1$. Therefore, the choice $L_p = p$ will (minimally) satisfy the required assumptions, with $m = m_0(p) = 2p - 1$ in (B3). Note that this is exactly the same minimum L_p that would work in the case of the h-version (see Theorem 4.1.6 of [4]).

Next, for Gauss-Lobatto quadrature, it has been shown in equation (3.9) of [3] that

$$\|\hat{v}\|^2_{0,\hat{K}} \le \sum_{l=1}^{p+1} \hat{\omega}_l^p \hat{v}^2(\hat{b}_l^p) \le C\|\hat{v}\|^2_{0,\hat{K}} \tag{3.6}$$

for all $\hat{v} \in U_p(\hat{K})$. Hence, the choice $L_p = p+1$ satisfies (B1), (B2), and (B3) (with $m = m_0(p) = 2p-1$), though (B) does not hold.

Gaussian rules for $Q_p(Q)$, $Q_p(C)$. Given an N_p-point rule on I, we may construct the corresponding tensor product rule on I^n which will have $(N_p)^n$ points. When applied to $Q_p(I^n)$, these rules once again will be exact for the same degree as the 1-d case, so that the minimum N_p required for (B) to hold is once again $p+1$ and $p+2$ for Legendre and Lobatto rules, respectively.

We now examine the minimum N_p required for (B1)–(B3) instead of (B) to hold. We observe that

$$\tilde{U}_p(I^n) = \tilde{Q}_p(I^n) = Q_p(I^n) \cap \mathscr{P}_{np-1}(I^n).$$

The dimension of $\tilde{U}_p(I^n)$ is therefore $(p+1)^n - 1$. A necessary condition for (B2) to hold is unisolvency of the $(N_p)^n$ points for $\tilde{U}_p(I^n)$, which gives $N_p \ge p+1$ for $n > 1$. This shows that unlike the 1-d case, we cannot relax the number of Legendre points to less than $N_p = p+1$. This is exactly the same requirement needed in the h-version when $n > 1$ (see p. 205 of [4]). For this choice, (B3) will hold with the minimal $m_0(p) = 2p+1$. (As noted above, this means that (B) holds.)

For Gauss-Lobatto rules, it is shown in [3] that (3.6) holds for any n (with $(p+1)$ replaced by $(p+1)^n$ in the summation). Hence, we get the minimum requirement for (B1)–(B3) to be $N_p = p+1$, with $m_0(p) = 2p-1$.

Rules for $Q'_p(Q)/\mathscr{P}_p(Q)$. Since $Q'_p(Q) \subset \mathscr{P}_{p+1}(Q)$ and $Q'_p(Q), \mathscr{P}_p(Q) \subset Q_p(Q)$, we can, of course, use the rules for $Q_p(Q)$. However, this is not efficient, since a number of extra points are used, without raising m in (B)'. In [9], the problem of minimizing the number of quadrature points for $\mathscr{P}_p(Q)$ is analyzed and symmetrical rules are derived for m up to 21. Using the results of [9], we may now obtain the minimum L_p required for the rule to be exact up to degree m (≤ 21), i.e., for condition (B) to hold. Note that theoretical results guaranteeing positivity of the weights or giving error estimates are not available.

Rules for $\mathscr{P}_p(T)$. Various rules have been derived in this case, a survey of which may be found in [12]. Similar to the rules for $\mathscr{P}_p(Q)$, theoretical results are not fully developed, so that we use condition (B) instead of (B1)–(B3). In [8], the problem of deriving symmetrical rules with the minimum L_p for (B) to hold has been investigated for m up to 20. These are significantly more economical than Gaussian product rules which are derived by mapping the tensor product rule for I^n onto T and have been used in MSC/PROBE.

4. Preliminary results

We assume that (A) and (B1)–(B3) hold. Let us define, for any $K \in \tau$ or for $K = \hat{K}$,

$$E_K(u) = \int_K u\,dx - \sum_{l=1}^{L} \omega_{l,K} u(b_{l,K})$$

and

$$E_K(u, v) = (u, v)_K - (u, v)_{p,K} = E_K(uv),$$

where

$$(u, v)_K = \int_K uv\, dx = \int_{\widehat{K}} J_K(\hat{x})\hat{u}(\hat{x})\hat{v}(\hat{x})\, d\hat{x}$$

and $(u, v)_{p,K}$ is as defined in (3.5). Then

$$E_K(u, v) = E_{\widehat{K}}(J_K\hat{u}, \hat{v}). \tag{4.1}$$

Also, by (B3),

$$E_{\widehat{K}}(\hat{u}, \hat{v}) = 0 \quad \forall \hat{u} \in U_{m-p}(\widehat{K}),\ \hat{v} \in U_p(\widehat{K}). \tag{4.2}$$

Let $[x]$ denote the integer part of x. We begin by proving a technical lemma, using Lemma 2.1.

Lemma 4.1. *Let $\widehat{K} \subseteq R^n$, $n = 1, 2,$ or 3. Let $\gamma = [\frac{n}{2} + 1]$. Let $u \in H^s(\widehat{K})$ with $s \geq \gamma$. Then the projection Π_p^γ in Lemma 2.1 satisfies*

$$\|u - \Pi_p^\gamma u\|_{0,\infty,\widehat{K}} \leq Cp^{-(s-n/2)}\|u\|_{s,\widehat{K}}.$$

Proof. We note that by an interpolation result from [3], for $0 < \varepsilon \leq \frac{1}{2}$,

$$\|u - \Pi_p^\gamma u\|_{0,\infty,\widehat{K}} \leq C\|u - \Pi_p^\gamma u\|^{1/2}_{n/2+\varepsilon,\widehat{K}}\|u - \Pi_p^\gamma u\|^{1/2}_{n/2-\varepsilon,\widehat{K}}. \tag{4.3}$$

Also, by Lemma 2.1, for $0 \leq r \leq \gamma \leq s$,

$$\|u - \Pi_p^\gamma u\|_{r,\widehat{K}} \leq Cp^{-(s-r)}\|u\|_{s,\widehat{K}}. \tag{4.4}$$

The lemma follows by using (4.4) with $r = \frac{n}{2} + \varepsilon$ and $r = \frac{n}{2} - \varepsilon$ in (4.3). □

Remark 4.1. By Remark 2.2, we may relax the restriction on γ to $\gamma > \frac{n}{2}$.

Lemma 4.2. *Let $\phi, \psi \in U_p(\widehat{K})$ and $c \in L_\infty(K)$. Then*

$$|E_{\widehat{K}}(c\phi, \psi)| \leq C\{\|c - \bar{c}\|_{0,\infty,\widehat{K}}\|\phi\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}} + \|\bar{c}\|_{0,\infty,\widehat{K}}\|\phi - w_1\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}}\}$$

for all $\bar{c} \in U_q(\widehat{K})$ and $w_1 \in U_{m-p-q}(\widehat{K})$, where $C > 0$ is independent of p and q, and $m - p - q > 0$.

Proof. We have

$$\begin{aligned} |E_{\widehat{K}}(c\phi, \psi)| &= |(c\phi, \psi)_{\widehat{K}} - (c\phi, \psi)_{p,\widehat{K}}| \\ &\leq |((c - \bar{c})\phi, \psi)_{\widehat{K}}| + |E_{\widehat{K}}(\bar{c}\phi, \psi)| + |((c - \bar{c})\phi, \psi)_{p,\widehat{K}}| \end{aligned} \tag{4.5}$$

for any $\bar{c} \in U_q(\widehat{K})$. Now for $w_1 \in U_{m-p-q}(\widehat{K})$, we have $\bar{c}w_1 \in U_{m-p}(\widehat{K})$, so that by (4.2),

$$|E_{\widehat{K}}(\bar{c}\phi, \psi)| = |(\bar{c}(\phi - w_1), \psi)_{\widehat{K}} - (\bar{c}(\phi - w_1), \psi)_{p,\widehat{K}}|. \tag{4.6}$$

Next, using (B2), we have

$$\begin{aligned} &|(\bar{c}(\phi - w_1), \psi)_{p,\widehat{K}}| \\ &\quad = \left|\sum_{l=1}^{L} \hat{\omega}_l(\bar{c}(\phi - w_1)\psi)(\hat{b}_l)\right| \leq \|\bar{c}\|_{0,\infty,\widehat{K}} \sum_{l=1}^{L} \hat{\omega}_l|((\phi - w_1)\psi)(\hat{b}_l)| \\ &\quad \leq \|\bar{c}\|_{0,\infty,\widehat{K}} \left(\sum_{l=1}^{L} \hat{\omega}_l(\phi - w_1)^2(\hat{b}_l)\right)^{1/2} \left(\sum_{l=1}^{L} \hat{\omega}_l\psi^2(\hat{b}_l)\right)^{1/2} \\ &\quad \leq C\|\bar{c}\|_{0,\infty,\widehat{K}}\|\phi - w_1\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}}. \end{aligned} \tag{4.7}$$

We may similarly bound

$$|((c-\overline{c})\phi, \psi)_{p,\widehat{K}}| \le C\|c-\overline{c}\|_{0,\infty,\widehat{K}}\|\phi\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}}. \tag{4.8}$$

Next, by Schwarz's inequality,

$$|(\overline{c}(\phi - w_1), \psi)_{\widehat{K}}| \le \|\overline{c}\|_{0,\infty,\widehat{K}}\|\phi - w_1\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}}. \tag{4.9}$$

Similarly,

$$|((c-\overline{c})\phi, \psi)_{\widehat{K}}| \le \|c-\overline{c}\|_{0,\infty,\widehat{K}}\|\phi\|_{0,\widehat{K}}\|\psi\|_{0,\widehat{K}}. \tag{4.10}$$

The lemma follows from (4.5)–(4.10). □

Lemma 4.3. *Let* $\Omega \subseteq R^n$ *and* $\gamma = [\frac{n}{2} + 1]$ *be as in Lemma* 4.1. *Then for* $f \in H^s(K)$, $s \ge \gamma$, *and* $v \in U_p(K)$,

$$|E_K(f, v)| \le C(m-p)^{-(s-n/2)}\|f\|_{s,K}\|v\|_{0,K}$$

if the mapping F_K *satisfies* (2.7) *with* $M \ge s+1$.

Proof. We have for $K \in \tau$, by (4.1) and (4.2),

$$E_K(f, v) = E_{\widehat{K}}(J_K\hat{f}, \hat{v}) = E_{\widehat{K}}(J_K\hat{f} - \hat{w}, \hat{v})$$

for any $\hat{w} \in U_{m-p}(\widehat{K})$. Hence,

$$|E_K(f, v)| \le \|J_K\hat{f} - \hat{w}\|_{0,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}} + \sum_{l=1}^{L}\hat{\omega}_l|(J_K\hat{f} - \hat{w})\hat{v}(\hat{b}_l)|. \tag{4.11}$$

Using (B2), we have

$$\begin{aligned}\sum_{l=1}^{L}\hat{\omega}_l|(J_K\hat{f} - \hat{w})\hat{v}(\hat{b}_l)| &\le \left(\sum_{l=1}^{L}\hat{\omega}_l(J_K\hat{f} - \hat{w})^2(\hat{b}_l)\right)^{1/2}\left(\sum_{l=1}^{L}\hat{\omega}_l\hat{v}^2(\hat{b}_l)\right)^{1/2}\\ &\le C\|J_K\hat{f} - \hat{w}\|_{0,\infty,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}}.\end{aligned}$$

Taking $\hat{w} = \Pi_p^{\gamma}(J_K\hat{f})$ and using Lemma 4.1, we get

$$\sum_{l=1}^{L}\hat{\omega}_l|(J_K\hat{f} - \hat{w})\hat{v}(\hat{b}_l)| \le C(m-p)^{-(s-n/2)}\|J_K\hat{f}\|_{s,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}}. \tag{4.12}$$

Also, by Lemma 2.1,

$$\|J_K\hat{f} - \hat{w}\|_{0,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}} \le C(m-p)^{-s}\|J_K\hat{f}\|_{s,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}}, \tag{4.13}$$

so that by (4.11)–(4.13), we have

$$\begin{aligned}|E_K(f, v)| &\le C(m-p)^{-(s-n/2)}\|J_K\hat{f}\|_{s,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}}\\ &\le C(m-p)^{-(s-n/2)}\|J_K\|_{s,\infty,\widehat{K}}\|\hat{f}\|_{s,\widehat{K}}\|\hat{v}\|_{0,\widehat{K}}.\end{aligned}$$

The lemma follows, using (2.8) and (2.9). □

5. Convergence estimates

We now analyze the effect of numerical integration on the problem (2.1). Throughout this section, we will refer to the solutions u, u_p, and $\tilde{u}_p$ of problems (2.1), (2.11), and (3.3), respectively. We begin by proving a lemma which

ensures the so-called "V_h-ellipticity" of the form $a_p(\cdot, \cdot)$ under the assumptions on the numerical quadrature rule. We will use the notation $Dp(x)$ to denote the row vector $(\frac{\partial p}{\partial x_1}(x), \ldots, \frac{\partial p}{\partial x_n}(x))$ for any function $p: R^n \to R$. Also, for $P: R^n \to R^n$, $DP(x)$ will denote the Jacobian matrix of P at x and $\|DP(x)\|$ the corresponding Euclidean matrix norm.

Lemma 5.1. *Let the mappings F_K satisfy* (2.7) *and* (2.8) *with $M \geq 1$. Then there exists a constant $C > 0$ such that*

$$C|v|_{1,\Omega}^2 \leq a_p(v, v) \quad \forall v \in S_{p,0},$$

with C independent of p.

Proof. Let $v \in S_{p,0}$. For any $K \in \tau$, let $v_K = v|_K$, so that $\hat{v}_K \in U_p(\widehat{K})$. Using (2.3), we have (noting that $\omega_{l,K} > 0$ by (A)),

$$\begin{aligned} a_{p,K}(v_K, v_K) &= \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} \left(a_{ij} \frac{\partial v_K}{\partial x_i} \frac{\partial v_K}{\partial x_j} \right)(b_{l,K}) \\ &\geq \kappa \sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} \left(\frac{\partial v_K}{\partial x_i} \right)^2 (b_{l,K}) \\ &\geq \kappa \sum_{l=1}^{L} \frac{\hat{\omega}_l J_K(\hat{b}_l)}{\|DF_K(\hat{b}_l)\|^2} \sum_{i=1}^{n} \left(\frac{\partial \hat{v}_K}{\partial \hat{x}_i} \right)^2 (\hat{b}_l), \end{aligned}$$

as in equation (4.4.26) of [4]. This gives

$$a_{p,K}(v_K, v_K) \geq \frac{C}{|J_K^{-1}|_{0,\infty,K} |F_K|_{1,\infty,\widehat{K}}^2} \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} \left(\frac{\partial \hat{v}_K}{\partial \hat{x}_i} \right)^2 (\hat{b}_l).$$

Using (B2) together with (2.7)–(2.9), we obtain

$$a_{p,K}(v_K, v_K) \geq \frac{C|\hat{v}_K|_{1,\widehat{K}}^2}{|J_K^{-1}|_{0,\infty,K} |F_K|_{1,\infty,\widehat{K}}^2} \geq C|v_K|_{1,K}^2.$$

Hence,

$$a_p(v, v) = \sum_{K \in \tau} a_{p,K}(v_K, v_K) \geq C \sum_{K \in \tau} |v|_{1,K}^2 = C|v|_{1,\Omega}^2,$$

which is the desired result. □

Let us define the $n \times n$ matrix $A = A(x) = [a_{ij}(x)]$. Then we see that

$$E^K = \sum_{i,j=1}^{n} E_K \left(a_{ij} \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_j} \right) = E_K((Du) A (Dv)^\top).$$

Noting that

$$Dw(x) = D\hat{w}(\hat{x}) DF_K^{-1}(x)$$

for any $w \leftrightarrow \hat{w}$, we see that

$$\begin{aligned} E^K &= E_{\widehat{K}}(J_K(D\hat{u} DF_K^{-1}) A (D\hat{v} DF_K^{-1})^\top) \\ &= E_{\widehat{K}}((D\hat{u}) B^K (D\hat{v})^\top) = \sum_{i,j=1}^{n} E_{\widehat{K}} \left(b_{ij}^K \frac{\partial \hat{u}}{\partial \hat{x}_i}, \frac{\partial \hat{v}}{\partial \hat{x}_j} \right), \end{aligned} \tag{5.1}$$

where

$$B^K = [b_{ij}^K] = J_K(DF_K^{-1})A(DF_K^{-1})^\top . \tag{5.2}$$

Denote $B = \{B^K,\ K \in \tau\} = \{[b_{ij}^K],\ K \in \tau\}$. We define

$$\beta_{l,t}(B) = \max_{i,j,K} \|b_{ij}^K\|_{l,t,\widehat{K}}, \tag{5.3}$$

with the subscript t being dropped when $t = 2$.

Lemma 5.2. *Let* $\overline{B}^K = [\overline{b}_{ij}^K]$, $\overline{b}_{ij}^K \in U_q(\widehat{K})$. *Then*

$$\|\tilde{u}_p - u_p\|_{1,\Omega} \le C\left\{ \frac{|\sum_{K\in\tau} E_K(f,\, \tilde{u}_p - u_p)|}{\|\tilde{u}_p - u_p\|_{1,\Omega}} + \beta_{0,\infty}(B - \overline{B})\|u\|_{1,\Omega} \right.$$
$$\left. + \beta_{0,\infty}(\overline{B})(\|u - u_p\|_{1,\Omega} + \|u - P_r^1 u\|_{1,\Omega}) \right\},$$

where $r = m - p - q$ *and* $P_r^1 u$ *is defined by* (2.13).

Proof. Let $v \in S_{p,0}$. Then from (2.1), (2.11), and (3.3), we get

$$a_p(\tilde{u}_p - u_p,\, v) = a(u_p,\, v) - a_p(u_p,\, v) + (f,\, v)_p - (f,\, v)$$
$$= \sum_{K\in\tau} \left\{ \sum_{i,j=1}^n E_K\left(a_{ij}\frac{\partial u_p}{\partial x_i},\, \frac{\partial v}{\partial x_j}\right) - E_K(f,\, v) \right\}.$$

Using (5.1), we have

$$E = \left|\sum_{K\in\tau} E^K\right| = \left|\sum_{K\in\tau}\sum_{i,j=1}^n E_K\left(a_{ij}\frac{\partial u_p}{\partial x_i},\, \frac{\partial v}{\partial x_j}\right)\right|$$
$$= \left|\sum_{K\in\tau}\sum_{i,j=1}^n E_{\widehat{K}}\left(b_{ij}^K\frac{\partial \hat{u}_p}{\partial \hat{x}_i},\, \frac{\partial \hat{v}}{\partial \hat{x}_j}\right)\right|.$$

Applying Lemma 4.2, we obtain

$$E \le C\sum_{K\in\tau}\sum_{i,j=1}^n \left\{ \|b_{ij}^K - \overline{b}_{ij}^K\|_{0,\infty,\widehat{K}} \left\|\frac{\partial \hat{u}_p}{\partial \hat{x}_i}\right\|_{0,\widehat{K}} \left\|\frac{\partial \hat{v}}{\partial \hat{x}_j}\right\|_{0,\widehat{K}} \right.$$
$$\left. + \|\overline{b}_{ij}^K\|_{0,\infty,\widehat{K}} \left\|\frac{\partial \hat{u}_p}{\partial \hat{x}_i} - \frac{\partial}{\partial \hat{x}_i}\widehat{P_r^1 u_p}\right\|_{0,\widehat{K}} \left\|\frac{\partial \hat{v}}{\partial \hat{x}_j}\right\|_{0,\widehat{K}} \right\},$$

where we have taken $w_1 = \frac{\partial}{\partial \hat{x}_i}\widehat{P_r^1 u_p}$, with P_r^1 as defined in (2.13). This gives, using (2.9),

$$E \le C\{\beta_{0,\infty}(B - \overline{B})\|u_p\|_{1,\Omega}\|v\|_{1,\Omega} + \beta_{0,\infty}(\overline{B})\|u_p - P_r^1 u_p\|_{1,\Omega}\|v\|_{1,\Omega}\}. \tag{5.4}$$

Now, using the boundedness of the P_r^1 projection, we have

$$\|u_p - P_r^1 u_p\|_{1,\Omega} \le \|u - u_p\|_{1,\Omega} + \|u - P_r^1 u\|_{1,\Omega} + \|P_r^1(u - u_p)\|_{1,\Omega}$$
$$\le C(\|u - u_p\|_{1,\Omega} + \|u - P_r^1 u\|_{1,\Omega}). \tag{5.5}$$

Also,

$$\|u_p\|_{1,\Omega} \le C\|u\|_{1,\Omega}. \tag{5.6}$$

Combining the above, we get

$$a_p(\tilde{u}_p - u_p, v) \le C\Big\{ \Big| \sum_{K\in\tau} E_K(f, v) \Big| + \beta_{0,\infty}(B - \overline{B}) \|u\|_{1,\Omega} \|v\|_{1,\Omega} + \beta_{0,\infty}(\overline{B})(\|u - u_p\|_{1,\Omega} + \|u - P_r^1 u\|_{1,\Omega}) \Big\} \|v\|_{1,\Omega}. \tag{5.7}$$

The lemma follows by putting $v = \tilde{u}_p - u_p$ in (5.7) and using the "V_h-ellipticity" of Lemma 5.1. □

We now prove two theorems which show that the rate of convergence, using the p-version, essentially remains unchanged under the effect of numerical quadrature.

Theorem 5.1. *Let $f \in H^s(\Omega)$ with $s > \frac{n}{2}$, and let the solution u of (2.1) be in $H^k(\Omega)$, $k > 1$. Let $b_{ij}^K \in H^l(\widehat{K})$ for each i, j, K, with $l > \frac{n}{2}$. Let $\tilde{u}_p$ denote the solution of (3.3), with the quadrature rule satisfying (A) and (B1)–(B3), where $r = \min\{p, m - p - q\} > 0$ in (B3). Then*

$$\|u - \tilde{u}_p\|_{1,\Omega} \le C\{(m-p)^{-(s-n/2)} \|f\|_{s,\Omega} + q^{-(l-n/2)} \beta_l(B) \|u\|_{1,\Omega} + r^{-(k-1)}(\beta_{0,\infty}(B) + q^{-(l-n/2)} \beta_l(B)) \|u\|_{k,\Omega}\}, \tag{5.8}$$

where the constant C is independent of u, m, p, and q.

Proof. By the triangle inequality, we have

$$\|u - \tilde{u}_p\|_{1,\Omega} \le C\|u - u_p\|_{1,\Omega} + \|u_p - \tilde{u}_p\|_{1,\Omega}. \tag{5.9}$$

Using Lemma 5.2, we get

$$\|u_p - \tilde{u}_p\|_{1,\Omega} \le C\{E_1 + E_2 + E_3\}, \tag{5.10}$$

where

$$E_1 = \frac{|\sum_{K\in\tau} E_K(f, \tilde{u}_p - u_p)|}{\|\tilde{u}_p - u_p\|_{1,\Omega}} \le C(m-p)^{-(s-n/2)} \|f\|_{s,\Omega}, \tag{5.11}$$

by Lemma 4.3. Also,

$$E_2 = \max_{i,j,K} \|b_{ij}^K - \overline{b}_{ij}^K\|_{0,\infty,\widehat{K}} \|u\|_{1,\Omega}.$$

Taking $\overline{b}_{ij}^K = \Pi_q^\gamma b_{ij}^K$ as in Lemma 4.1, we get

$$E_2 \le C q^{-(l-n/2)} \beta_l(B) \|u\|_{1,\Omega}. \tag{5.12}$$

Finally, using the above bound for $\beta_{0,\infty}(B - \overline{B})$, we find

$$\begin{aligned} E_3 &= \beta_{0,\infty}(\overline{B})(\|u - u_p\|_{1,\Omega} + \|u - P_r^1 u\|_{1,\Omega}) \\ &\le C(\beta_{0,\infty}(B) + q^{-(l-n/2)} \beta_l(B)) \inf_{v\in S_{r,0}} \|u - v\|_{1,\Omega}, \end{aligned} \tag{5.13}$$

where we have used (2.12) and (2.14) to bound $\|u - u_p\|_{1,\Omega}$ and $\|u - P_r^1 u\|_{1,\Omega}$, respectively. We can now bound the infimum in (5.13) by $Cr^{-(k-1)}\|u\|_{k,\Omega}$. The theorem follows by (5.9)–(5.13) and (2.14). □

Corollary 5.1. *Let f and u be as above. Suppose the $a_{ij} \in H^{\alpha}(\Omega)$ satisfy (2.2) and the mappings F_K satisfy (2.7)–(2.8) with $M \geq 1$ such that $l = \min\{\alpha, M\} > \frac{n}{2}$. Then*

$$\|u - \tilde{u}_p\|_{1,\Omega} = O((m-p)^{-(s-n/2)} + q^{-(l-n/2)} + r^{-(k-1)}). \tag{5.14}$$

Proof. Using (2.2), (2.7), (2.8) together with the definition (5.2) of B, we note that

$$\beta_l(B),\ \beta_{0,\infty}(B) < C,$$

where C is a constant depending on A_1 and A_2. The assertion (5.14) then follows from (5.8). □

Let us now use (5.14) to investigate the rate of convergence for the case that f is very smooth (which occurs frequently in practice). We assume that $m \approx 2p$ or larger, which, as seen in §3, holds for practically chosen quadrature rules. Then we see that

$$(m-p)^{-(s-n/2)} \approx p^{-(s-n/2)} = O(p^{-(k-1)})$$

in (5.14), provided $s - \frac{n}{2} \geq k - 1$. Similarly, if the mappings F_K and the coefficients $a_{ij}(x)$ are sufficiently smooth, we may assume $l - \frac{n}{2} \geq k - 1$. Taking $q = \frac{p}{2}$ (for instance) then gives

$$q^{-(l-n/2)} = \left(\frac{p}{2}\right)^{-(l-n/2)} = O\left(\left(\frac{p}{2}\right)^{-(k-1)}\right),$$

$$r^{-(k-1)} = O\left(\left(\frac{p}{2}\right)^{-(k-1)}\right),$$

so that using (5.14), we obtain

$$\|u - \tilde{u}_p\|_{1,\Omega} = O\left(\left(\frac{p}{2}\right)^{-(k-1)}\right) = O(p^{-(k-1)}),$$

i.e., the asymptotic order of convergence achievable by exact integration (equation (2.15)) is preserved when numerical quadrature with $m \approx 2p$ is used.

In the previous analysis, one can also take $q = p^{\varepsilon}$, where $\varepsilon = (k-1)/(l-\frac{n}{2})$. Then, with $m = 2p + q$, we have

$$q^{-(l-n/2)} = p^{-(k-1)}, \qquad r^{-(k-1)} = p^{-(k-1)},$$

so that when ε is small, there is no deterioration in convergence (a possible overintegration by one point may be required to ensure $m \geq 2p+q$). Note that, if a_{ij} and J_K are constant functions, instead of (5.14), we get the estimate

$$\|u - \tilde{u}_p\|_{1,\Omega} = O((m-p)^{-(s-n/2)} + (m-p)^{-(k-1)}). \tag{5.15}$$

If l is small, then the above analysis suggests that overintegration by an amount large enough to ensure that $m \geq 2p + p^{(k-1)/(l-n/2)}$ would be sufficient to preserve the error bound. Indeed, for the case that the lack of smoothness lies in the coefficients $a_{ij}(x)$, the numerical examples in [13] show that overintegration does reestablish the expected accuracy (though m does not have to be that large). If, however, the Jacobian is nonsmooth, then in general, the error of best approximation also deteriorates, and this cannot improve with overintegration. In §6, we consider some examples of nonsmooth mappings and investigate how

they respond to overintegration. Let us mention that even if the asymptotic rate of convergence $Cp^{-(k-1)}$ is regained, the constant C (which depends on A_1 and A_2) may be quite large. Let us also remark that the amount of smoothness we require on f in the above estimates is probably excessive (see, e.g., [16], where sharp conditions on the smoothness required on f (in the context of the h-version) are derived).

Suppose now that $\Omega \subset R^2$ is a polygon such that the rate of convergence for u_p (using exact integration) is governed by the singularities, i.e., (2.16) holds. Then, if f, a_{ij}, and F_K are smooth enough, we have the following theorem, which once again shows that the rate of convergence of $\|u-u_p\|_{1,\Omega}$ is preserved.

Theorem 5.2. *Let the conditions of Theorem* 5.1 *hold, except that* u, *the solution of* (2.1), *is of the form* u_v *in* (2.6). *Then, with* $\tilde{r} = \min\{p, m-p-q\}$,

$$\|u - \tilde{u}_p\|_{1,\Omega} = O((m-p)^{-(s-n/2)} + q^{-(l-n/2)} + |\log \tilde{r}|^{\gamma_1} \tilde{r}^{-2\alpha_1}).$$

Proof. The proof is identical to that of Theorem 5.1, except that the term E_3 is now bounded using (2.16) instead. □

Once again, it is observed that if $m \approx 2p$ and if f and b_{ij}^K are smooth enough, we can show

$$\|u - \tilde{u}_p\|_{1,\Omega} \leq C|\log p|^{\gamma_1} p^{-2\alpha_1}, \tag{5.16}$$

i.e., the asymptotic rate of convergence from (2.16) is preserved. Let us also remark that if u is very smooth (as in Remark 2.3) and the error $\|u - \tilde{u}_p\|_{1,\Omega}$ is exponential (say), then this will again be reflected in the error $\|u - \tilde{u}_p\|_{1,\Omega}$, provided the first two terms in (5.14) are sufficiently small.

Remark 5.1. Theorem 5.1, when applied to Gauss-Lobatto quadrature over rectangular meshes (in R^n), gives an estimate similar to the central result in [13]. The analysis in [13] (and in [3]) depends strongly upon a sharp estimate of the interpolation error at Gaussian quadrature points. Consequently, it is not readily applicable to quadrature rules for which such interpolation estimates have not been developed (for the rules in [8, 9], for example), or to non-Gaussian rules. Moreover, the approach in [13] assumes tensor product elements (as used in spectral element methods) and does not carry over in an obvious way to elements like triangles and tetrahedra, which are common in finite elements. Finally, estimates like (5.16), showing the doubling of the convergence rate for singular solutions, do not easily follow. Our approach does not require an estimate of interpolation error at quadrature points and hence can be applied to these situations as well.

6. Numerical results

In this section, we present the results of some numerical computations related to the one-dimensional model problem

$$-u''(x) = f(x), \quad 0 < x < 1, \qquad u(0) = u(1) = 0. \tag{6.1}$$

We put (6.1) into the variational form (2.1) and then employ the p-version, using a single element on $(0, 1)$ which is the image of the reference element $(-1, 1)$ under the mapping $x = F(\xi)$. Obviously, a smooth affine choice exists for the mapping F in this one-dimensional case. However, by choosing F to

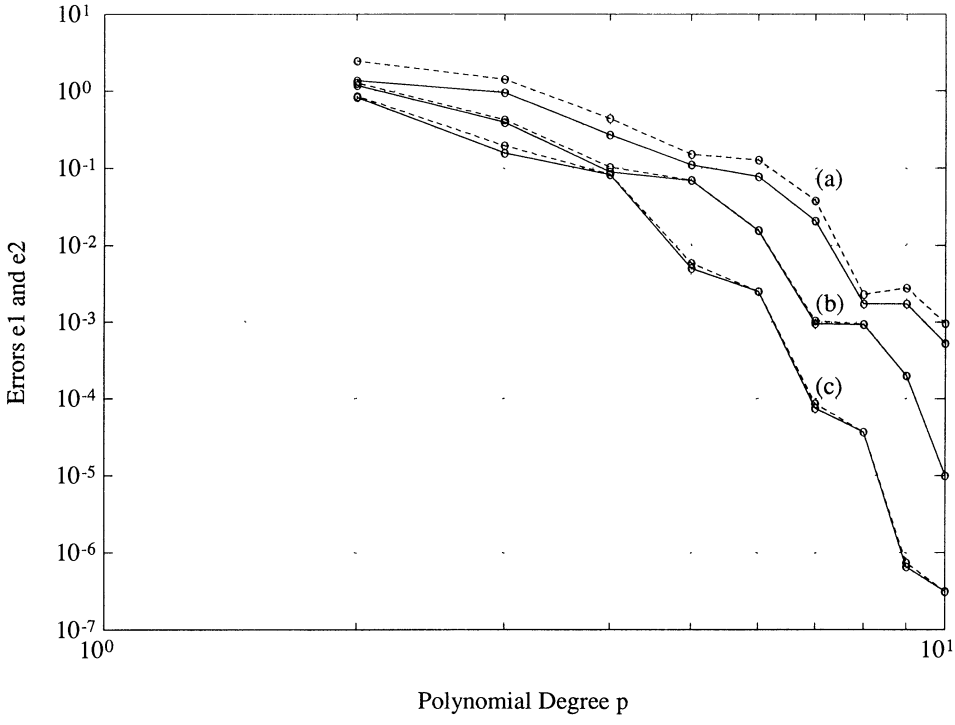

FIGURE 1. The errors e_1 and e_2 for various F. (a) quadratic, $\varepsilon = 0.01$; (b) quadratic, $\varepsilon = 1.0$; (c) affine.

be a nonlinear mapping (and varying its smoothness), we can model the case of higher dimensions, where the use of more complicated mappings is usually unavoidable. We take F to be

$$F(\xi) = \frac{(1+\xi+\varepsilon)^\alpha - \varepsilon^\alpha}{(2+\varepsilon)^\alpha - \varepsilon^\alpha}. \tag{6.2}$$

For $\alpha = 1$, this gives an affine mapping. For $\alpha \neq 1$, we obtain a nonlinear mapping whose smoothness depends on the parameter ε. For ε close to 0, the inverse of the Jacobian will be very large at $\xi = -1$, giving a nonsmooth mapping.

Suppose that f is chosen so that the exact solution is given by

$$u(x) = x \sin \pi x. \tag{6.3}$$

Let $e_1 = \|u - u_p\|_1$ be the error when u_p is the finite element solution obtained through exact integration (approximated here by a 4000-point composite Simpson's rule). Let $e_2 = \|u - \tilde{u}_p\|_1$ be the error, where $\tilde{u}_p$ is the finite element solution obtained by using the p-point Gauss-Legendre rule (the minimal possible) for the stiffness matrix (the load vector being calculated exactly). In Figure 1, we have plotted e_1 (solid lines) and e_2 (broken lines) on a log-log scale against p for three choices of the mapping F.

First, the curves (c) represent the case where F is affine ($\alpha = 1$). As expected, it is observed that e_1 and e_2 are essentially identical. This agrees very well with the results of §5, and shows that p-point quadrature for the stiffness

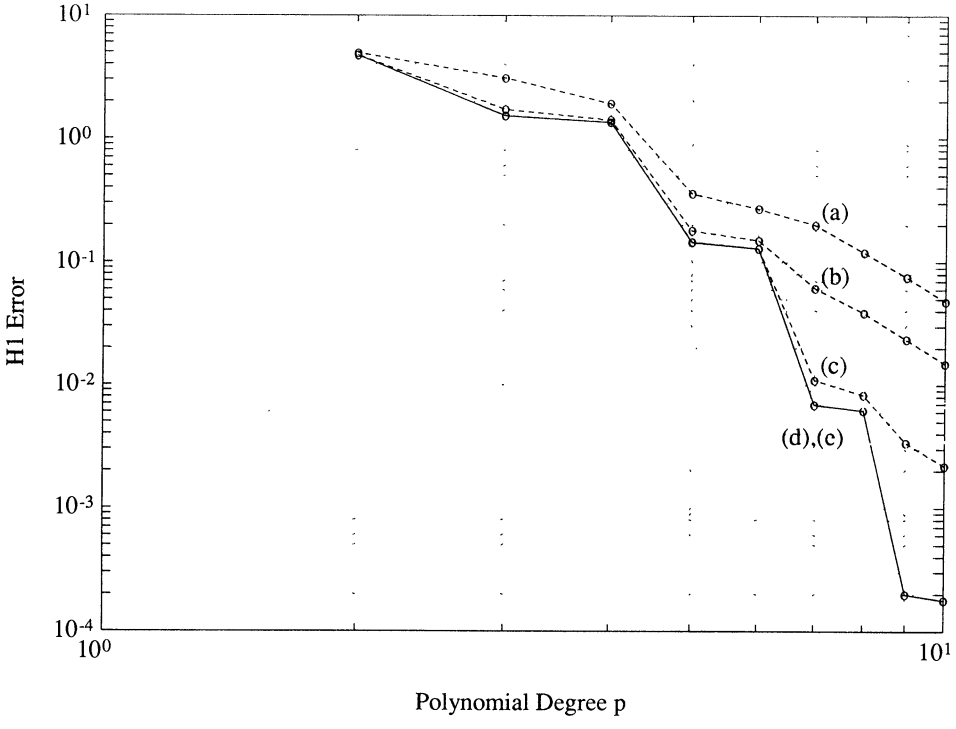

FIGURE 2. The effect of overintegration when $\alpha = 1.8$ and $\varepsilon = 0.1$. (a) p points; (b) $p+1$ points; (c) $p+3$ points; (d) $2p$ points; (e) exact.

matrix is sufficient to preserve the error. Notice that the convergence rate is exponential, since the exact solution is analytic (see Remark 2.3).

In (b), a smooth quadratic mapping is chosen with $\alpha = 2$ and $\varepsilon = 1.0$. It is observed that once again, e_1 and e_2 are extremely close, as expected by Theorem 5.1. In fact, similar to the case when F is affine, no overintegration is required to make e_2 of the same order as e_1. This shows that when the mapping is smooth, we can take q very small in (5.14) (so that we essentially get (5.15)), and the rate of convergence will still be preserved. Notice, however, that the magnitude of the error has become larger when we compare e_1 for the quadratic case to that for the affine case. This is due to the fact that using a quadratic F has led to a deterioration of the approximability of the trial functions, a phenomenon that is unrelated to the accuracy of the quadrature scheme employed.

As ε is decreased (i.e., the mapping is made less smooth), two effects occur. The dominant effect, seen from curves (a) (where $\alpha = 2$ and $\varepsilon = 0.01$), is that the "exact" error e_1 becomes worse, owing to the degradation of the approximability. However, in addition, the error e_2 deteriorates even further, and a shift occurs between the graphs of e_1 and e_2. Overintegration will now help in reducing the difference between e_2 and e_1 (which grows larger as ε decreases), but will obviously not help in decreasing e_1.

The above example shows that when the mapping is nonsmooth, the use of overintegration may only have a limited effect in decreasing the error. This is

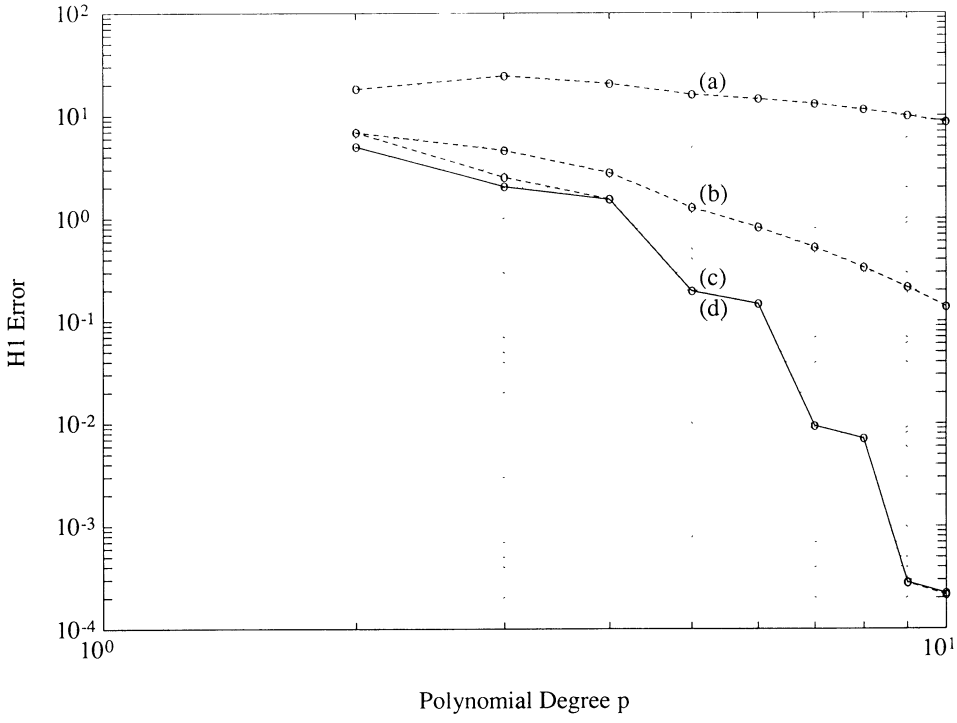

FIGURE 3. The effect of overintegration when $\alpha = 2$ and $\varepsilon = 0.01$. (a) p points; (b) $2p$ points; (c) p^2 points; (d) exact.

because the deterioration in the approximability properties of the underlying subspaces may have a more serious effect, independent of the quadrature used. However, there are situations where it is the lack of accuracy of the quadrature, rather than the approximability, which plays the dominant role. As an example, we consider the case where f is chosen so that instead of (6.3), the exact solution is given by

$$u(x) = \sin\{\pi[(cx + \varepsilon^\alpha)^{1/\alpha} - (1 + \varepsilon)]\}, \tag{6.4}$$

where $c = (2 + \varepsilon)^\alpha - \varepsilon^\alpha$. We take F to be given by (6.2) again. Using this combination of F and u, we see that

$$u(F(\xi)) = \sin \pi\xi . \tag{6.5}$$

Hence, $u(F(\xi))$ is very smooth on $(-1, 1)$ and can therefore easily be approximated by the trial functions. In effect, this means that approximability is not an issue here.

In Figure 2, we have plotted the results of various computations, using $\alpha = 1.8$ and $\varepsilon = 0.1$. The solid line (e) once again represents e_1, the error when exact integration is used, while the broken lines (a), (b), and (c) represent the errors with Gauss-Legendre quadrature using respectively p, $p + 1$, and $p + 3$ points on the stiffness matrix (the load vector being calculated exactly once more). It is observed that there is a significant loss in the convergence rate, owing to the quadrature rule being insufficiently accurate. Using $2p$ points,

however, restores the rate of convergence, as is observed from curve (d), which is identical to curve (e).

A question that arises from the above calculation is whether there is a quadrature rule that would be accurate enough to work uniformly for all mappings (i.e., all α and ε). The answer seems to be no, as observed from Figure 3. Here, we have taken $\alpha = 2$ and $\varepsilon = 0.01$. The solid line (d) represents exact integration, for which the error is essentially identical to the previous case, since the exact solution is once again given by (6.4). Curves (a), (b), and (c), respectively, represent quadrature with p, $2p$, and p^2 points. It is seen now that practically no convergence is observed with p points, and to recover the exponential rate of convergence, p^2 points are needed. This number will increase further as ε is made smaller.

We note that in Remarks 3.2 and 3.3 of [13], it was stated that a loss of regularity in $F(\xi)$ induces a loss of regularity in $u(F(\xi))$, so that overintegration will not help in these cases. However, our example above shows that a nonsmooth $F(\xi)$ can do just the opposite as well: it can *increase* the smoothness of $u(F(\xi))$. In fact, this is the idea used to treat singularities by the well-known "quarter-point mapping" in the h-version and by the "method of auxiliary mappings" (see, e.g., [14]) in the p-version. In such situations, as is illustrated by the above example, overintegration can be particularly useful in dealing with the effect of nonsmooth mappings.

BIBLIOGRAPHY

1. I. Babuška and M. Suri, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal. **24** (1987), 750–776.
2. I. Babuška, B. Guo, and M. Suri, *Implementation of nonhomogeneous Dirichlet boundary conditions in the p-version of the finite element method*, Impact Comput. Sci. Engrg. **1** (1989), 36–63.
3. C. Canuto and A. Quarteroni, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp. **38** (1982), 67–86.
4. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, Amsterdam, 1978.
5. P. G. Ciarlet and P.-A. Raviart, *The combined effect of curved boundaries and numerical integration in isoparametric finite element methods*, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 409–474.
6. P. J. Davis and P. Rabinowitz, *Methods of numerical integration*, 2nd ed., Academic Press, New York, 1975.
7. M. R. Dorr, *The approximation theory for the p-version of the finite element method*, SIAM J. Numer. Anal. **21** (1984), 1181–1207.
8. D. Dunavant, *High degree efficient symmetrical Gaussian quadrature rules for the triangle*, Internat. J. Numer. Methods Engrg. **21** (1985), 1129–1148.
9. ——, *Economical symmetrical quadrature rules for complete polynomials over a square domain*, Internat. J. Numer. Methods Engrg. **21** (1985), 1777-1784.
10. V. A. Kondrat'ev, *Boundary-value problems for elliptic equations in domains with conic or corner points*, Trudy Moskov. Mat. Obshch. **16** (1967), 209–292; English transl., Trans. Moscow Math. Soc. **16** (1967), 227–313.
11. W. Gui and I. Babuška, *The h, p and h – p versions of the finite element method in one dimension. Part* I. *The error analysis of the p-version*, Numer. Math. **49** (1986), 577–612.
12. J. N. Lyness, QUG2-*integration over a triangle*, Technical Memo #13, Math. and Comp. Sci. Div., Argonne National Lab., 1983.

13. Y. Maday and E. M. Ronquist, *Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries*, Comput. Methods Appl. Mech. Engrg. **80** (1990), 91–115.
14. H.-S. Oh and I. Babuška, *The p-version of the finite element method for the elliptic boundary value problems with interfaces*, Comput. Methods Appl. Mech. Engrg. (1992) (in press).
15. M. Suri, *The p-version of the finite element method for elliptic equations of order* $2l$, RAIRO Modél. Math. Anal. Numér. **24** (1990), 265–304.
16. L. B. Wahlbin, *Maximum norm estimates in the finite element methods with isoparametric quadratic elements and numerical integration*, RAIRO Anal. Numér. **12** (1978), 173–202.

Department of Mathematics, Syracuse University, Syracuse, New York 13244-1150
E-mail address: Banerjee@sunrise.bitnet

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228
E-mail address: Suri@umbc1.umbc.edu