Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 227 (2008) 2411–2429
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Direct Minimization of the least-squares spectral element functional – Part I: Direct solver

Wijnand Hoitinga [a], Roel de Groot [b], Marcel Kwakkel [a], Marc Gerritsma [a,*]

[a] Department of Aerospace Engineering, Delft University of Technology, 2629 HS Delft, The Netherlands
[b] TNO Institute of Applied Physics, Flow and Structural Dynamics, P.O. Box 155, 2600 AD Delft, The Netherlands

Received 11 June 2007; received in revised form 4 October 2007; accepted 22 October 2007
Available online 7 November 2007

Abstract

This paper describes an equivalent but improved least-squares formulation for the numerical approximation of the solution of partial differential equations. Instead of using variational analysis to impose the conditions for minimizing the residual, the residuals are minimized directly, thus leading to a method we will denote by Direct Minimization (DM). DM circumvents setting up the normal equations which consists of matrix–matrix multiplications. Matrix–matrix multiplications are expensive, may lead to loss of accuracy and destroy the sparsity pattern present in the original system. The condition number of the DM formulation is the square root of the condition number which would be obtained if variational analysis was employed. An element-by-element procedure will be presented which allows for parallelization of DM. A computational comparison between DM and the conventional least-squares formulation based on variational analysis will be presented.

Keywords: Least-squares formulation; Spectral element method; QR-decomposition; Incompressible Navier–Stokes equations

1. Introduction

The least-squares formulation has proven to be an interesting alternative to Galerkin-type weak formulations for solving systems of partial differential equations. The least-squares method is based on minimization of a functional of the residual, which is equivalent to the error in a suitably chosen norm. Necessary conditions for a minimizer may be obtained by applying variational analysis which leads, for well-posed problems, to a symmetric positive definite (SPD) system of algebraic equations. These systems are therefore amenable to well-established iterative methods such as pre-conditioned conjugate gradient methods.

Least-squares formulations have been applied to a variety of scientific problems, amongst others: fluid dynamics, [16], electro-magnetism, [22], sound propagation, [20,21], viscoelastic flows, [10]. In [10] Direct Minimization was already used in the discrete formulation.

* Corresponding author.
E-mail address: M.I.Gerritsma@TUDelft.nl (M. Gerritsma).

0021-9991/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.10.022

Application of the least-squares method to finite element discretizations leads to the least-squares finite element method (LSFEM), see [3–6,8,15,16]. LSFEM combines geometric flexibility of the finite element approach with the desirable minimization properties of the least-squares formulation, thus circumventing compatibility constraints and saddle-point problems such as encountered in Galerkin type formulations for mixed methods.

If, instead of low order finite element approximations, high order orthogonal polynomials are used to approximate the exact solution, the method is called the *Least-Squares Spectral Element Method* (LSQSEM). Exponential convergence with polynomial enrichment will take place if the underlying exact solution is sufficiently smooth. The combination of the finite element approach (geometric flexibility), spectral methods (high order accuracy) and the least-squares formulation (SPD systems) has, for example, been investigated by Proot and Gerritsma [11,26–28] and Pontaza and Reddy, [24,25].

Despite the attractive features of the least-squares spectral element method, this approach leads to algebraic systems with relatively high condition numbers compared to least-squares finite element methods and conventional finite/spectral element methods based on the Galerkin formulation. Spectral methods are known to produce algebraic systems with high condition numbers, [9], and the direct application of the conventional least-squares formulation may also lead to higher condition numbers. The combination of both techniques is therefore prone to lead to excessively high condition numbers.

There are several ways to make the condition number more manageable. The first step that is usually employed in the least-squares formulation is to rewrite the system of governing equations as an equivalent first order system, [4–6,15,16]. Transformation to a first order system necessarily introduces new variables which is often considered an additional cost. However, the newly introduced variables in general represent physical quantities of interest, such as vorticity or stresses. These physically relevant variables will then be directly approximated in the least-squares formulation.

A second step is to develop suitable pre-conditioners for the resulting algebraic system. A variety of approaches is available, from Jacobi pre-conditioning in the Conjugate Gradient solver, [24,25] to Sobolev pre-conditioning, [17,18].

In this first paper an alternative method will be described which minimizes the residual directly in contrast to the conventional least-squares formulation where one employs variational analysis to set up the weak formulation. The resulting condition number is only the square root of the condition number that would be obtained if the conventional least-squares method had been used. In addition, the new method circumvents a costly matrix–matrix multiplication thus avoiding loss of precision and fill-in in the stiffness matrix. This paper is restricted to direct solvers; the extension to pre-conditioned iterative Direct Minimization methods is discussed in the second paper on Direct Minimization.

This paper is organized as follows: In Section 2 the least-squares formulation will be described succinctly. In Section 3 the spectral element discretization will be given and in Section 4 the conventional least-squares formulation will be presented. In Section 5 Direct Minimization will be discussed. Section 6 addresses solution procedures for over-determined linear systems. A direct comparison between the conventional least-squares formulation and Direct Minimization for a simple test problems will be given in Section 7. This paper ends with the conclusions in Section 8.

2. Least-squares formulation based on variational analysis

Consider the following abstract boundary value problem

$$\mathcal{L}(u) = f \quad \text{in} \quad \Omega, \tag{2.1}$$

$$\mathcal{R}(u) = g \quad \text{on} \quad \Gamma \subset \partial\Omega, \tag{2.2}$$

where $\Omega \in \mathbb{R}^d$ is bounded, $\mathcal{L}$ is a linear first order partial differential operator and $\mathcal{R}$ a trace operator. Both operators act on a scalar or vector unknown u defined over the domain Ω. The right hand side of (2.1) is supplemented with a given forcing function f. The right hand side of (2.2) can be set to zero without loss of generality. We therefore take in what follows $g = 0$.

Assume that we can find a function space X, which contains the exact solution to (2.1) and elements satisfying the boundary conditions $\mathcal{R}(u) = 0$, and a function space Y such that there $\exists C_1,\ C_2 > 0$

$$C_1\|u\|_X \leqslant \|\mathcal{L}(u)\|_Y \leqslant C_2\|u\|_X, \quad \forall u \in X, \tag{2.3}$$

which states that $\mathcal{L}$ is a continuous mapping from the function space X onto Y possessing a continuous inverse.

Second and higher order partial differential equations are first transformed into an equivalent first order system. The two main reasons to rewrite higher order PDEs into equivalent first order systems are: to reduce the condition number of the resulting algebraic system and to mitigate continuity requirements between neighboring elements in the discrete approximation.

If the right hand side function f is measurable in the Y-norm and the exact solution u_{ex} of the differential equation is contained in the function space X, the residual norm associated with any approximation u^N is equivalent to the error measured in the X-norm, since by (2.3) and the linearity of $\mathcal{L}$ we have constants C_1, $C_2 > 0$ such that

$$C_1\|u^N - u_{\text{ex}}\|_X \leqslant \|\mathcal{L}(u^N) - f\|_Y \leqslant C_2\|u^N - u_{\text{ex}}\|_X. \tag{2.4}$$

Since $\|\cdot\|_X$ and $\|\mathcal{L}(\cdot)\|_Y$ define equivalent norms for elements $u \in X$ we approximate the exact solution in the X-norm by minimizing the residual in the Y-norm. Therefore we introduce the functional

$$\mathcal{I}(u) = \frac{1}{2}\|\mathcal{L}u - f\|_Y^2. \tag{2.5}$$

Minimizing (2.5) is equivalent to solving the abstract problem given by equations (2.1) and (2.2). The Euler-Lagrange equations for the minimization of $\mathcal{I}(u)$ are given by

$$\lim_{t\to 0}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}(u + tv) = 2\int_\Omega (\mathcal{L}v)^{\mathrm{T}}(\mathcal{L}u - f)\,\mathrm{d}\Omega = 0 \quad \forall v \in X. \tag{2.6}$$

The least-squares method can therefore be stated as
Find $u \in X$ such that

$$\mathcal{B}(u, v) = F(v) \quad \forall v \in X, \tag{2.7}$$

where

$$\mathcal{B}(u, v) := (\mathcal{L}u, \mathcal{L}v)_Y, \tag{2.8}$$

$$F(v) := (f, \mathcal{L}v)_Y, \tag{2.9}$$

where $\mathcal{B}$ a symmetric, bilinear form. If the coercivity constant $C_1 > 0$, this bilinear form is positive definite.

In practice we choose for the function space $Y(\Omega) = L^2(\Omega)$. Other norms such as $H(\text{div}, \Omega)$, [23,29,30], are certainly feasible in the above described framework, but the evaluation and continuity requirements for Sobolev space $H^k(\Omega)$, $k > 0$, is not easy, especially for fractional Sobolev spaces. In this paper we will evaluate the residuals in the L^2-norm.

3. Spectral elements

Instead of seeking the minimizer over the infinite dimensional space X we restrict our search to a conforming subspace $X^h \subset X$ by performing a domain decomposition where the solution within each sub-domain is expanded with respect to a polynomial basis. The domain Ω is sub-divided into K non-overlapping quadrilateral sub-domains Ω^k

$$\Omega = \bigcup_{k=1}^{K}\Omega^k, \quad \overset{\circ}{\Omega}{}^k \cap \overset{\circ}{\Omega}{}^l = \emptyset,\ k \neq l. \tag{3.1}$$

Each sub-domain is mapped onto the unit cube $[-1,1]^d$, where $d = \dim(\Omega)$. Within this unit cube the unknown function is approximated by polynomials. In this paper a spectral element method based on *Legendre* polynomials, $L_k(x)$ over the interval $[-1, 1]$, is employed, [7,9,19]. Define the Gauss–Lobatto–Legendre (GLL) nodes by the zeroes of the polynomial

$$(1-x^2)L'_N(x) \tag{3.2}$$

and the Lagrange polynomials, $h_i(x)$ through these GLL-points, x_i, by

$$h_i(x) = \frac{1}{N(N+1)}\frac{(x^2-1)L'_N(x)}{L_N(x_i)(x-x_i)} \quad \text{for } i = 0,\ldots,N, \tag{3.3}$$

where $L'_N(x)$ denotes the derivative of the Nth Legendre polynomial.

We can expand the approximate solution in each sub-domain in terms of a truncated series of these Lagrangian basis functions, which for $d = 2$ yields

$$u^N(x,y) = \sum_{i=0}^{N}\sum_{j=0}^{N}\hat{u}_{ij}h_i(x)h_j(y), \tag{3.4}$$

where the $\hat{u}_{ij}$'s are to be determined by the least-squares method. Since we have converted a general higher order PDE to an equivalent first order system, C^0-continuity suffices to patch the solutions on the individual subdomains together.

The integrals appearing in the least-squares formulation, (2.7) are approximated by Gauss–Lobatto quadrature

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=0}^{N} f(x_i)w_i, \tag{3.5}$$

where w_i are the GLL weights given by

$$w_i = \frac{2}{N(N+1)}\frac{1}{L_N^2(x_i)}, \quad i = 0,\ldots,N. \tag{3.6}$$

Since we use only first order PDE's and use GLL-integration to approximate the integrals we need to evaluate the derivative of the Lagrangian function at the GLL-points

$$\left.\frac{\mathrm{d}h_j}{\mathrm{d}x}\right|_{x_i} = d_{ij} = \begin{cases} -\frac{N(N+1)}{4} & i = j = 0 \\ \frac{L_N(x_i)}{L_N(x_j)}\frac{1}{x_i - x_j} & 0 \leqslant i \neq j \leqslant N \\ 0 & 1 \leqslant i = j \leqslant N \\ \frac{N(N+1)}{4} & i = j = N \end{cases}. \tag{3.7}$$

The extension to multidimensional problems is performed by using tensor products.

Inserting the finite dimensional approximation and its derivatives in (2.7) and evaluating the integrals using (3.7) leads to an algebraic system for the unknowns $\hat{u}_{ij}$.

4. Conventional least-squares finite element method

In Sections 2 and 3 the approach for the conventional or variational least-squares formulation is described. This approach can be summarized by

$$(\mathcal{L}(u), \mathcal{L}(v)) = (f, \mathcal{L}(v)) \Longleftrightarrow \int_\Omega \mathcal{L}(u)\mathcal{L}(v)\mathrm{d}\Omega = \int_\Omega f\mathcal{L}(v)\mathrm{d}\Omega, \quad \forall v \in X(\Omega). \tag{4.1}$$

The next step consists of domain decomposition where the integration over Ω is written as the sum of the integrals over the sub-domains Ω^k, $k = 1,\ldots,K$,

$$\sum_k \int_{\Omega^k} \mathcal{L}(u)\mathcal{L}(v)\mathrm{d}\Omega^k = \sum_k \int_{\Omega^k} f\mathcal{L}(v)\mathrm{d}\Omega^k \quad \forall v \in X(\Omega^k). \tag{4.2}$$

Then we insert the finite dimensional approximation for each element $u^{N,k} = \sum_i u_i^{N,k}\phi_i(x)$, where the ϕ_i are basis functions, which span the finite dimensional subspace in Ω^k

$$\sum_k \left[\sum_i u_i^{N,k} \int_{\Omega^k} \mathcal{L}(\phi_i)\mathcal{L}(\phi_j)\mathrm{d}\Omega^k\right] = \sum_k \int_{\Omega^k} f\mathcal{L}(\phi_j)\mathrm{d}\Omega^k \quad \forall\phi_j,\ \ j = 1,\ldots,N. \tag{4.3}$$

It suffices in (4.3) to take $v = \phi_j$, because $\mathcal{L}$ is assumed to be a linear operator and since any arbitrary v in the finite dimensional subspace is a linear combination of these basis functions.

Inserting the Gauss–Lobatto integration then gives

$$\sum_k \left[\sum_i u_i^{N,k} \sum_p \mathcal{L}(\phi_i)(\boldsymbol{x}_p)\mathcal{L}(\phi_j)(\boldsymbol{x}_p)\boldsymbol{w}_p\right] = \sum_k \sum_p f(\boldsymbol{x}_p)\mathcal{L}(\phi_j)(\boldsymbol{x}_p)\boldsymbol{w}_p \quad \forall\phi_j,\ \ j = 1,\ldots,N. \tag{4.4}$$

Here x_p denote the Gauss–Lobatto points and w_p the Gauss–Lobatto weights defined by (3.2) and (3.6), respectively. Note that in the multi-dimensional case $\boldsymbol{x}_p$ is a vector, ϕ_i is a tensor product and $\boldsymbol{w}_p$ is the product of the Gauss–Lobatto weights in each direction separately. We now define in each element the matrix $\boldsymbol{A}^k$ by

$$(\boldsymbol{A}^k)_{pi} = \mathcal{L}(\phi_i)(x_p), \tag{4.5}$$

i.e. the matrix coefficient denotes the application of the differential operator to the ith basis function evaluated at the pth Gauss–Lobatto point. Furthermore we introduce the diagonal weight matrix $\boldsymbol{W}^k$ by

$$(\boldsymbol{W}^k)_{pp} = \boldsymbol{w}_p. \tag{4.6}$$

The discretized least-squares problem (4.4) can be written as

$$\sum_k \left[(\boldsymbol{A}^k)^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}^k\right]u^{N,k} = \sum_k \left[(\boldsymbol{A}^k)^{\mathrm{T}}\boldsymbol{W}\boldsymbol{F}\right], \tag{4.7}$$

where the vector $\boldsymbol{F}$ contains the elements $(\boldsymbol{F})_p = f(\boldsymbol{x}_p)$. The system of algebraic equations obtained in this way, i.e. using variational analysis, is called the *normal equations*. The normal equations reflect on a discrete level the symmetry that was already mentioned at the continuous level, (2.7). Note that Gauss–Lobatto integration may be performed on a finer grid than the grid on which the unknowns are defined. In this case the matrix $\boldsymbol{A}^k$ is non-square, i.e. there are more rows than columns in the matrix. The resulting normal equations, however, deliver a square, positive definite matrix which possesses a unique solution.

There are various reasons to look for ways to circumvent the use of normal equations. First of all, suppose that at elemental level the matrix $\boldsymbol{A}^k$ possesses a well-defined sparsity structure that could be utilized in storing or solving the system, then this sparsity structure will vanish after pre-multiplication with its transpose. Due to fill-in the number of non-zero entries in the global system may increase significantly.

Furthermore, if the polynomial degree is high, the number of equations to be satisfied and the number of degrees of freedom is high and we use Gauss–Lobatto integration which is at least as high as the representation of the unknowns, the size of the matrices $\boldsymbol{A}^k$ is therefore significant. But the matrix multiplication to obtain the element matrix in (4.7) then becomes very expensive. In fact Proot, [28], has shown that in some calculations setting-up of the discrete system may take up to 50% of the total CPU time. If we can avoid this step, then this will considerably speed-up the algorithm.

Not only is the matrix multiplication time consuming, it may also lead to loss of information. Consider for instance a matrix of the following form:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 0 & \epsilon \end{pmatrix}, \quad \epsilon \ll 1, \tag{4.8}$$

where $\epsilon \neq 0$ and can be represented by the machine. Pre-multiplication of this matrix with its transpose gives

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon^2 \end{pmatrix} \approx \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{4.9}$$

if ϵ^2 cannot be represented by the machine accuracy. So ill-conditioned systems may become singular when a least-squares approach is employed. Such is the case, for example, in Navier–Stokes equations at high Reynolds numbers. If we introduce the vorticity as auxiliary variable to reduce the second order partial differential operator to an equivalent first order operator, [16], we obtain for steady two-dimensional flow

$$\begin{pmatrix} 0 & \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & 0 \\ \frac{\partial}{\partial x} & u\frac{\partial}{\partial x}+v\frac{\partial}{\partial y} & 0 & \epsilon\frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & 0 & u\frac{\partial}{\partial x}+v\frac{\partial}{\partial y} & -\epsilon\frac{\partial}{\partial x} \\ 0 & -\frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 1 \end{pmatrix} \begin{pmatrix} p \\ u \\ v \\ \omega \end{pmatrix} = \begin{pmatrix} 0 \\ f_x \\ f_y \\ 0 \end{pmatrix}, \tag{4.10}$$

where $\epsilon = 1/Re, \omega = \vec{\nabla} \times \vec{u}$ is the vorticity, p denotes the pressure, (u, v) the velocity vector and (f_x, f_y) the body force vector. In practical applications the Reynolds number is of the order $Re = 10^6$–10^8. Pre-multiplication by the transpose operator may obscure the full contribution of the ϵ-terms when the contribution of ϵ^2-terms are lost due to the finite precision of our machines.

Even if the contribution of small terms is not lost entirely, setting up the normal equations may reduce the number of significant digits in the computation.

5. Direct Minimization – LSQSEM–DM

In the previous section we argued that setting up the normal equations is costly, leads to loss of precision and increases the condition number of the resulting system compared to the system which would have been obtained without pre-multiplication with the transpose of the partial differential operator. We therefore introduce an equivalent, but improved formulation which circumvents the above mentioned disadvantages. This can be done by avoiding the variational analysis as was done in (2.6). In order to avoid variational analysis we start with the original minimization problem (2.5)

Find $u \in X(\Omega)$ which minimizes the functional $\mathcal{I}(u) = \frac{1}{2}\|\mathcal{L}u - f\|^2_{Y(\Omega)}$. (5.1)

Since we decompose the computational domain Ω into a union of non-overlapping sub-domains Ω^k, $k = 1, \ldots, K$, we can also write this as

Find all $u^k \in X(\Omega^k)$ which minimize the functional

$$\mathcal{I}(u^1, \ldots, u^K) = \sum_{k=1}^{K} \|\mathcal{L}u^k - f\|^2_{Y(\Omega^k)}. \tag{5.2}$$

Now in each domain Ω^k we are going to restrict our search to a finite dimensional subspace $X^N(\Omega^k) \subset X(\Omega^k)$ using the spectral approximation given by (3.4)

Find all $u^{N,k} \in X^N(\Omega^k)$ which minimize the functional

$$\mathcal{I}(u^{N,1}, \ldots, u^{N,K}) = \sum_{k=1}^{K} \|\mathcal{L}u^{N,k} - f\|^2_{Y(\Omega^k)}. \tag{5.3}$$

Next we introduce numerical quadrature to evaluate the integrals which constitute the L^2-norm. This gives

Find all $u^{N,k} \in X^N(\Omega^k)$ which minimize the functional

$$\mathcal{I}(u^{N,1}, \ldots, u^{N,K}) \approx \sum_{p=0}^{N^k_{\mathrm{int}}} \sum_{k=1}^{K} \left(\mathcal{L}u^{N,k} - f\right)^2 \Big|_{x_p} w_p, \tag{5.4}$$

where N^k_{int} denotes the number of integration points in element k. Introducing our matrix notation (4.5) and (4.6) this can be written as

Find all $u^{N,k} \in X^h(\Omega^k)$ which minimize the functional

$$\sum_{k=1}^{K} \left(\boldsymbol{A}^k u^{N,k} - \boldsymbol{F}^k\right)^{\mathrm{T}} \boldsymbol{W}^k \left(\boldsymbol{A}^k u^{N,k} - \boldsymbol{F}^k\right) = \sum_{k=1}^{K} \| \sqrt{\boldsymbol{W}^k} \left(\boldsymbol{A}^k u^{N,k} - \boldsymbol{F}^k\right) \|^2. \tag{5.5}$$

So the procedure of domain decomposition, insertion of an approximate solution and the use of numerical integration has converted the minimization in the function space $L^2(\Omega)$ to a minimization problem in Euclidean space: Find the finite dimensional vector $u = (u^1, \ldots, u^K)^{\mathrm{T}} \in \mathbb{R}^n$ such that the norm in $\mathbb{R}^m$ given by (5.5) is minimized. If $m = n$, i.e. the number of unknowns in the global system equals the number of equations, the use of the weight matrix $\boldsymbol{W}^k$ is inconsequential and the problem reduces to a collocation method evaluated in the GLL-points, [12–14], given by

$$\sum_{k=1}^{K} \left(\boldsymbol{A}^k u^{N,k} - \boldsymbol{F}^k\right) = 0. \tag{5.6}$$

In case $m > n$, we have more equations than unknowns and the solution which minimizes the residual norm of the overdetermined system is given by

$$\sum_{k=1}^{K} \sqrt{\boldsymbol{W}^k} \boldsymbol{A}^k u^k = \sum_{k=1}^{K} \sqrt{\boldsymbol{W}^k} \boldsymbol{F}^k. \tag{5.7}$$

Let us for convenience introduce the following notation $\boldsymbol{B} = \sum_{k=1}^{K} \sqrt{\boldsymbol{W}^k} \boldsymbol{A}^k$ and $\boldsymbol{G} = \sum_{k=1}^{K} \sqrt{\boldsymbol{W}^k} \boldsymbol{F}^k$. Then we have the following Theorem, [1,2]

Theorem. *Let $\boldsymbol{B} \in \mathbb{R}^{m,n}$ and $\boldsymbol{G} \in \mathbb{R}^m$, then the following 2 statements are equivalent*:

- *Determine the vector $u \in \mathbb{R}^n$ which minimize the Euclidean norm $\|\boldsymbol{B}u - \boldsymbol{G}\|^2$.*
- *Determine the vector $u \in \mathbb{R}^n$ such that the residual $\boldsymbol{R} = \boldsymbol{G} - \boldsymbol{B}u \in \mathcal{N}(\boldsymbol{B}^{\mathrm{T}})$.*

Proof. Let $\boldsymbol{R}_u$ denote the residual obtained by $\boldsymbol{R}_u = \boldsymbol{G} - \boldsymbol{B}u$. Assume that $\boldsymbol{B}^{\mathrm{T}} \boldsymbol{R}_u = 0$, then for any v we have that

$$\boldsymbol{R}_v = \boldsymbol{G} - \boldsymbol{B}v = \boldsymbol{R}_u + \boldsymbol{B}(u - v). \tag{5.8}$$

Then the norm of $\boldsymbol{R}_v$ is larger than the norm of $\boldsymbol{R}_u$, because

$$\begin{aligned} \|\boldsymbol{R}_v\|^2 &= \|\boldsymbol{R}_u + \boldsymbol{B}(u - v)\|^2 = \|\boldsymbol{R}_u\|^2 + 2(u - v)^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \boldsymbol{R}_u + \|\boldsymbol{B}(u - v)\|^2 = \|\boldsymbol{R}_u\|^2 + \|\boldsymbol{B}(u - v)\|^2 \\ &\geqslant \|\boldsymbol{R}_u\|^2. \end{aligned} \tag{5.9}$$

So if $\boldsymbol{B}^{\mathrm{T}} \boldsymbol{R}_u = 0$ then u minimizes the residual in the Euclidean norm.
If on the other hand $\boldsymbol{B}^{\mathrm{T}} \boldsymbol{R}_u = z \neq 0$ then we can find a $v = u + \epsilon z$ for which we have

$$\|\boldsymbol{R}_v\|^2 = \|\boldsymbol{R}_u - \epsilon \boldsymbol{B} z\|^2 = \|\boldsymbol{R}_u\|^2 - 2\epsilon \|z\|^2 + \epsilon^2 \|\boldsymbol{B} z\|^2 < \|\boldsymbol{R}_u\|^2 \quad \text{for sufficiently small } \epsilon \tag{5.10}$$

So if $\boldsymbol{B}^{\mathrm{T}} \boldsymbol{R}_u \neq 0$, then u is not a minimizer of the over-determined system of equations. □

The above Theorem shows that finding the minimizer of the overdetermined system (5.7) is equal to imposing

$$\begin{aligned} &\left(\sum_{k=1}^{K} \sqrt{\boldsymbol{W}^k} \boldsymbol{A}^k\right)^{\mathrm{T}} \left(\sqrt{\boldsymbol{W}^k} \left(\boldsymbol{A}^k u^k - \boldsymbol{F}^k\right)\right) = 0 \\ &\qquad \Longleftrightarrow \\ &\sum_{k=1}^{K} (\boldsymbol{A}^k)^{\mathrm{T}} \boldsymbol{W}^k (\boldsymbol{A}^k) u = \sum_{k=1}^{K} (\boldsymbol{A}^k)^{\mathrm{T}} \boldsymbol{W}^k \boldsymbol{F}^k, \end{aligned} \tag{5.11}$$

which is the same equation that we obtained using variational analysis. Therefore, Direct Minimization given by (5.7) is equivalent to (4.7) as a result of the Theorem.

However note that (5.7) is more appealing to use than (4.7). Since no pre-multiplication is employed we do not lose the sparsity pattern of the matrix $\boldsymbol{A}^k$ and we prevent fill-in in the global matrix. Bear in mind that $\boldsymbol{W}^k$ is a diagonal matrix and so is its square root. Pre-multiplication with a diagonal matrix amounts to row-scaling which does not affect the sparsity.

The pre-multiplication with the transpose of $\boldsymbol{A}^k$ is a costly operation that is circumvented by using (5.7) instead of (4.7).

Furthermore, the number of significant digits in any small parameter in the original system of equation is retained by Direct Minimization, while is has been shown that this cannot be guaranteed in case we use variational analysis.

6. Solving the over-determined system

The a priori estimates (2.3) guarantee that a unique minimizer exists. This is equivalent to saying: the global matrix $\boldsymbol{B}$ has full rank. Despite the fact that the system possesses a unique solution we cannot invert the global matrix $\boldsymbol{B}$ in case $m > n$, i.e. if we have more equations than unknowns. One way of dealing with such a situation is to use the *pseudo-inverse*[1]. We will introduce the pseudo-inverse here to demonstrate the effect of application of variational analysis on the condition number of the resulting system.

Every matrix $\boldsymbol{B} \in \mathbb{R}^{m,n}$ can be written using a so-called *singular value decomposition*

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}, \tag{6.1}$$

where $\boldsymbol{D}$ is an $n \times n$ diagonal matrix, $\boldsymbol{U}$ is an $m \times n$ orthogonal matrix a $\boldsymbol{V}$ is an $n \times n$ orthogonal matrix, i.e.

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I} \in \mathbb{R}^{n,n}. \tag{6.2}$$

If the matrix $\boldsymbol{B}$ is of full rank, the diagonal entries in $\boldsymbol{D}$ are non-zero. These diagonal elements are called the *singular values* of the matrix $\boldsymbol{B}$. In this case the pseudo-inverse of $\boldsymbol{B}$ is defined as

$$\boldsymbol{B}^{\text{pseudo-inv}} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^{\mathrm{T}} \in \mathbb{R}^{n,m}. \tag{6.3}$$

It is straightforward to show that $\boldsymbol{B}^{\text{pseudo-inv}}\boldsymbol{B} = I \in \mathbb{R}^{n,n}$. So if we have an over-determined system $\boldsymbol{B}x = f$ the solution is formally given by $x = \boldsymbol{B}^{\text{pseudo-inv}}f$. Using the fact that

$$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathrm{T}} \Rightarrow (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^{\mathrm{T}}, \tag{6.4}$$

gives

$$(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{B}^{\text{pseudo-inv}}. \tag{6.5}$$

So the solution of the over-determined system $\boldsymbol{B}x = f$ is given by

$$x = \boldsymbol{B}^{\text{pseudo-inv}}f = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}f, \tag{6.6}$$

which solution is also obtained from the normal equations

$$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}x = \boldsymbol{B}^{\mathrm{T}}f. \tag{6.7}$$

So the pseudo-inverse also minimizes the residual.

Having established the relation between the singular value decomposition, the pseudo inverse and the original minimization problem, we will now define the condition number for the non-square matrix $\boldsymbol{B}$.

Definition. The condition number of a non-square matrix of full rank, cond ($\boldsymbol{B}$), is defined as the ratio of the largest absolute singular value over the smallest absolute singular value

[1] The pseudo-inverse is also known as the *Moore-Penrose inverse*.

$$\mathrm{cond}(\boldsymbol{B}) = \frac{|d_i|_{\max}}{|d_j|_{\min}}. \tag{6.8}$$

As a consequence of this definition, the condition number of $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ is given by

$$\mathrm{cond}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}) = \frac{|d_i|^2_{\max}}{|d_j|^2_{\min}} = \mathrm{cond}(\boldsymbol{B})^2. \tag{6.9}$$

So the condition number of the normal equations is the square of the condition number obtained by using Direct Minimization.

Jiang, [16], and Bochev and Gunzburger, [4], demonstrate that for the Stokes and Navier–Stokes equations, one should always convert the higher order system of partial differential equations into a system of first order differential equations in order to keep the condition number under control. If in addition to this operation, we also use Direct Minimization, we can reduce the condition number even further. Jiang, [16], has demonstrated that by writing the Poisson problem as an equivalent first order system, the condition number of the resulting linear system of equations scales with mesh refinement in the same way as the system that would be obtained by application of the Galerkin method. Therefore, Direct Minimization performs even better than the Galerkin formulation with respect to the condition number.

In case no reduction to a first order system is possible, since the governing equations are already posed as a first order system, least-squares produces higher condition numbers than the Galerkin formulation. In this case the Direct Minimization approach produces condition numbers comparable to the Galerkin formulation.

Although the singular value decomposition is a useful analytical tool in comparing condition numbers, we will now describe the use of orthogonal methods to solve the over-determined system of equations. The use of orthogonal methods is particularly efficient for spectral methods, where the element matrices are quite large.

6.1. Global QR

Let us return to our global system of algebraic equation given by

$$\boldsymbol{B}u = G \Longleftrightarrow \text{Find } u \text{ which minimizes} \|Bu - G\|^2, \tag{6.10}$$

where $\boldsymbol{B} \in \mathbb{R}^{m,n}, u \in \mathbb{R}^n$ and $G \in \mathbb{R}^m$. Now for any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{m,m}$ we have

$$\|\boldsymbol{Q}(\boldsymbol{B}u - G)\|^2 = \|\boldsymbol{B}u - G\|^2, \tag{6.11}$$

since the Euclidean norm is invariant under orthogonal transformations.

We now decompose the $m \times n$ matrix $\boldsymbol{B}$ in a $\boldsymbol{QR}$-decomposition, $\boldsymbol{B} = \boldsymbol{QR}$, where $\boldsymbol{Q}$ is an orthogonal $m \times m$-matrix and $\boldsymbol{R}$ is an $m \times n$ upper-triangular matrix. The $\boldsymbol{R}$ matrix can be written as

$$\boldsymbol{R} = \begin{pmatrix} \tilde{\boldsymbol{R}} \\ \boldsymbol{0} \end{pmatrix}, \tag{6.12}$$

where $\tilde{\boldsymbol{R}}$ is an upper-triangular $n \times n$ matrix with non-zero diagonal entries, and $\boldsymbol{0}$ is an $(m-n) \times n$ matrix with zero entries. With this decomposition we have

$$\|\boldsymbol{B}u - G\|^2 = \|\boldsymbol{Q}^{\mathrm{T}}(\boldsymbol{B}u - G)\|^2 = \|\boldsymbol{R}u - \boldsymbol{Q}^{\mathrm{T}}G\|^2 = \left\| \begin{pmatrix} \tilde{\boldsymbol{R}} \\ \boldsymbol{0} \end{pmatrix} u - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\|^2 = \|\tilde{\boldsymbol{R}}u - c_1\|^2 + \|c_2\|^2, \tag{6.13}$$

where c_1 is a n-vector and c_2 is an $m-n$-vector. With this decomposition, minimizing the Euclidean norm is straightforward. The second term, $\|c_2\|^2$, in (6.13) cannot be minimized. The only term that can be made small – zero in fact – is the first term on the right hand side of (6.13). So we have for the least-squares solution

$$u_{LS} = \tilde{\boldsymbol{R}}^{-1}c_1, \tag{6.14}$$

which is just a back-substitution for the upper-triangular matrix $\tilde{\boldsymbol{R}}$. An approximation to the L^2-norm of the residual is given by the second term, $\|c_2\|^2$, and this value is available without solving for u_{LS}.

Note again, that when exact arithmetic is used the minimizer u_{LS} is equal to the least-squares solution obtained by the conventional least-squares formulation which is obtained by applying variational analysis and solving the normal equations.

As can be seen from the above derivation, Direct Minimization avoids the costly matrix–matrix multiplication to set up the global system matrix, and therefore does not alter the sparsity pattern of the original matrix $\boldsymbol{A}$ and the condition number is the square root of the condition number that would be obtained by employing conventional least-squares.

6.2. Schur numbering and block-QR

Although the Direct Minimization approach described in the previous subsection has many desirable features, the application of a QR decomposition is much more expensive than applying a Choleski decomposition to the normal equations.

By applying a suitable numbering of the global degrees of freedom the following matrix structure can be obtained, see also [1]

$$\begin{bmatrix} \boldsymbol{A}_1 & & & \boldsymbol{E}_1 \\ & \ddots & & \vdots \\ & & \boldsymbol{A}_K & \boldsymbol{E}_K \end{bmatrix} \begin{pmatrix} u^1_{\text{int}} \\ \vdots \\ u^K_{\text{int}} \\ u_{if} \end{pmatrix} = \begin{pmatrix} F^1 \\ \vdots \\ F^K \end{pmatrix}, \tag{6.15}$$

where the matrices $\boldsymbol{A}_i$ are associated with the internal degrees of freedom in each spectral element Ω_k, $k = 1, \ldots, K$, the matrices $\boldsymbol{E}_i$ are associated with degrees of freedom shared by multiple elements, u^j_{int} denote the degrees of freedom in element j not shared by another element and u_{if} denote the unknowns at the element interfaces. Since we have a C^0 polynomial basis and use Lagrangian basis functions, these elements are associated with interface unknowns. Such a structure of the global system matrix is obtained when a global Schur numbering of the unknowns is used, [28].

The global matrix (6.15) consists of a global assembly of the local element matrices of the form

$$\boldsymbol{B}_i = [\, \boldsymbol{A}_i \qquad \boldsymbol{E}_i \,]. \tag{6.16}$$

The block-QR algorithm is now applied to all the sub-matrices $\boldsymbol{A}_i = \boldsymbol{Q}_i \boldsymbol{R}_i$. Pre-multiplication with $\boldsymbol{Q}_i^{\mathrm{T}}$ then gives

$$\boldsymbol{Q}_i^{\mathrm{T}} [\, \boldsymbol{A}_i \qquad \boldsymbol{E}_i \,] u = \boldsymbol{Q}_i^{\mathrm{T}} F_i \Longleftrightarrow \begin{bmatrix} \tilde{\boldsymbol{R}}_i & \boldsymbol{S}_i \\ \boldsymbol{0} & \boldsymbol{T}_i \end{bmatrix} \begin{pmatrix} u_{\text{int}} \\ u_{if} \end{pmatrix} = \begin{pmatrix} c_i \\ d_i \end{pmatrix}. \tag{6.17}$$

This operation can be performed for each element separately and in parallel. Especially for spectral methods, the matrices $\boldsymbol{A}_i$ are of reasonable size to justify a separate QR-decomposition. For least-squares finite element methods, the matrices $\boldsymbol{A}_i$ are generally too small, but then the same procedure can be applied to patches of elements.

Having performed the QR-decomposition for all spectral elements, we assemble all matrices $\boldsymbol{T}_i$ in a large global matrix $\boldsymbol{T}$

$$\boldsymbol{T} u_{if} = d_{\text{global}} \Longleftrightarrow \begin{bmatrix} \boldsymbol{T}_1 \\ \vdots \\ \boldsymbol{T}_K \end{bmatrix} u_{if} = \begin{pmatrix} d_1 \\ \vdots \\ d_K \end{pmatrix}. \tag{6.18}$$

This – generally – overdetermined system can again be solved with a QR-decomposition, which gives

$$\boldsymbol{Q}_{if}^{\mathrm{T}} \boldsymbol{T} u_{if} = \boldsymbol{Q}_{if}^{\mathrm{T}} d_{\text{global}} \Longleftrightarrow \begin{bmatrix} \boldsymbol{R}_T \\ \boldsymbol{0} \end{bmatrix} u_{if} = \begin{pmatrix} D_1 \\ D_2 \end{pmatrix}. \tag{6.19}$$

This system can now be solved for u_{if} by back-substitution. When the interface solution is known, the elemental internal degrees of freedom follow from (6.17):

$$\tilde{\boldsymbol{R}}_i u_{\text{int}} = c_i - \boldsymbol{S}_i u_{if}. \tag{6.20}$$

This back-substitution can be performed parallel again for each spectral element.

7. Numerical results

Despite the fact that Direct Minimization is equivalent to the conventional least-squares approach, the method is more stable when applied to problems with a high condition number. In order to show the difference between the conventional LS formulation and Direct Minimization method the two algorithms will be applied a Poisson equation and the incompressible Navier–Stokes equations.

7.1. The Poisson equation

In this section a sample problem is presented which consists of a modified Poisson equation given by

$$\kappa\Delta\phi = f(x,y), \quad (x,y) \in [-1,1]^2, \tag{7.1}$$

where

$$f(x,y) = -2\kappa \sin x \sin y. \tag{7.2}$$

The solution in this case is obviously independent of the parameter κ, but the condition number of the resulting system will strongly depend on κ.

In order to apply the least-squares formulation which allows for a C^0 finite element approximation, the governing equation needs to rewritten as an equivalent first order system

$$\boldsymbol{u} - \nabla\phi = 0, \tag{7.3}$$

$$\kappa\nabla \cdot \boldsymbol{u} = f. \tag{7.4}$$

Note that there are other equivalent first order systems possible with improved stability estimates, but this model problem is only introduced to compare formulations.

Following Jiang, [15], it is easy to show that this problem is well-posed

$$\kappa^2 C\left(\|\phi\|^2_{H^1(\Omega)} + \|\boldsymbol{u}\|^2_{H(\text{div};\Omega)}\right) \leqslant \|\boldsymbol{u} - \nabla\phi\|^2_{L^2(\Omega)} + \|\kappa\nabla\cdot\boldsymbol{u}\|_{L^2(\Omega)} \leqslant \|\phi\|^2_{H^1(\Omega)} + \|\boldsymbol{u}\|^2_{H(\text{div};\Omega)}, \tag{7.5}$$

for $\kappa \leqslant 1$. So the coercivity constant scales with κ^2 and therefore the condition number is bounded by κ^{-2}. We therefore expect to see differences between the conventional least-squares formulation and the Direct Minimization as proposed in this paper for $\kappa \ll 1$. For $\kappa = O(1)$, however, both formulations are expected to give similar results. In order to assess the improved stability of Direct Minimization the artificially ill-conditioned system is solved on a 5×5 grid with polynomial degree $N = 5$.

Fig. 1 (left) shows a plot of the solution of the Poisson equation obtained by the conventional least-squares formulation with $\kappa = 1$. Fig. 1 (right) gives the solution obtained by Direct Minimization for $\kappa = 1$. The results are indistinguishable. This follows from the observation that both methods are equivalent if exact arithmetic is used.

In Fig. 2 results for the case $\kappa = 10^{-5}$ are presented, where the differences between the conventional least-squares formulation and Direct Minimization become apparent. The conventional least-squares formulation is unable to approximate the exact solution sufficiently accurate due to the loss of precision, whereas Direct Minimization still approximates the solution sufficiently accurate.

The L^2-error for both the conventional least-squares formulation and Direct Minimization versus the parameter κ is depicted in Fig. 3. The conventional least-squares formulation (red line) approximates the solution rather well for a κ up to 10^{-4} after which the error grows dramatically. Direct Minimization is capable of approximating the solution up to a κ of $O(10^{-11})$. Note that the solution is independent of κ.

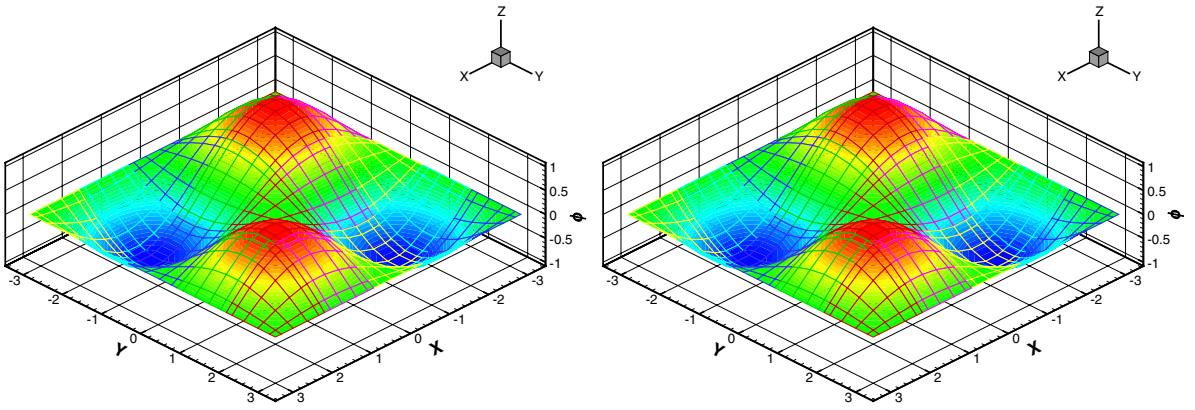


Fig. 1. Solution for $\kappa = 1$ obtained by the conventional least-squares formulation (left) and the result obtained by Direct Minimization (right).

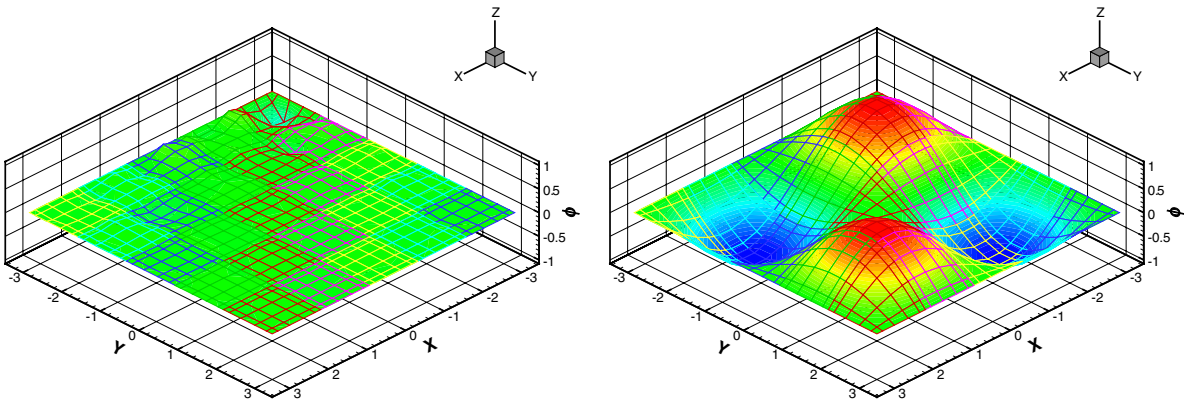


Fig. 2. Solution for $\kappa = 10^{-5}$ obtained by the conventional least-squares formulation (left) and the result obtained by Direct Minimization (right).

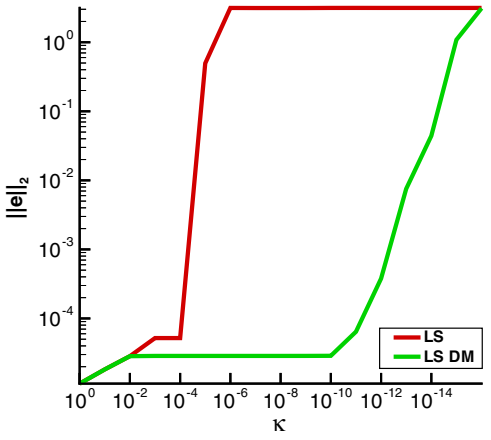


Fig. 3. The L^2-error of the conventional least-squares solution (red) and the solution obtained by Direct Minimization (green).

Table 1 shows the condition number of the conventional least-squares and Direct Minimization approach versus polynomial degree in the spectral element method. The condition number for DM is obtained from (6.8) and the condition number of LS is the absolute value of the largest over the smallest eigenvalue. One observes that the condition number of Direct Minimization is approximately the square root of the condition number associated with the conventional least-squares formulation.

Table 2 shows the growth of the condition number as a function of the number of elements for two polynomial degrees. One observes that, especially for high order methods which employ much higher polynomial degrees than $N = 4$, the difference in condition number between the conventional least-squares method and Direct Minimization grows very fast.

7.2. The incompressible Navier–Stokes equations

In Section 4 it was argued that the use of the conventional least-squares formulation will have detrimental effects for high Reynolds numbers, which may be circumvented or postponed by using Direct Minimization.

In order to verify this statement we need a test problem where only the influence of the small parameter – in this case $1/Re$ – in the discrete system is compared. In general for high Reynolds number flows the whole dynamics of the flow also changes from two-dimensional to three-dimensional flow, from steady to unsteady flow and/or from laminar to turbulent flow. In order to assess only the influence of the small coefficient $1/Re$ in the system, a simple periodic Couette flow has been chosen.

The Couette flow comprises a two-dimensional flow between two parallel moving plates. The height of the channel is $h = 1$ and the length is set to L. The lower plate moves with a velocity u_{lower} and the upper plate moves with a velocity u_{upper}. An adverse pressure gradient of $\Delta p = p_{\text{outlet}} - p_{\text{inlet}} = 2/Re$ is imposed over the length of the channel. With this choice of pressure gradient, the velocity field becomes independent of the Reynolds number

$$u(x,y) = u(y) = y(y-h)/L + u_{\text{lower}} + y(u_{\text{upper}} - u_{\text{lower}})/h, \tag{7.6}$$

$$v(x,y) = 0, \tag{7.7}$$

and

$$p(x,y) = p(x) = C + \frac{x}{L}\Delta p. \tag{7.8}$$

Since this solution is quadratic, the exact solution can be represented with polynomials of degree $N \geqslant 2$. The Navier–Stokes equations are written as a first order system as given by (4.10).

Table 1
Comparison between the condition numbers obtained from Direct Minimization (DM) and the conventional least-squares method (LS) as a function of the polynomial degree

N	DM	LS
2	5.934	35.215
3	11.914	142.056
4	20.228	409.369
5	30.646	936.737

Table 2
Condition number as a function of the number of elements for Direct Minimization (DM) and the conventional least-squares method (LS)

	$P = 2$		$P = 4$	
K	DM	LS	DM	LS
4	4.430	19.635	13.394	197.635
9	5.935	35.222	20.230	409.635
16	7.688	59.111	27.060	732.635
25	9.534	90.891	33.879	1147.786

For the numerical test we take two spectral elements in the y-direction and two elements per unit length in the x-direction. For the first exercise we take $L = 10$ and therefore use a 20×2-grid as show in Fig. 4 and a polynomial degree $N = 3$. The following boundary conditions have been used:

$$u_{lower} = 1 \quad \text{and} \quad u_{\text{upper}} = 2.$$

Despite the fact that the exact solution can be represented by polynomials of degree $N \geqslant 2$, the error will grow as a function of the Reynolds number due to loss of precision. The L^2-error as a function of the Reynolds number in the x-component of the velocity is displayed in Fig. 5, for $Re = 1 - 10^9$. From this figure we see that the L^2-error in the solution grows as Re^2 for the conventional least-squares formulation and as Re^1 for Direct Minimization.

This difference in error growth can be directly attributed to the condition numbers for both methods. In Table 3 the condition numbers for the conventional least-squares method and Direct Minimization are

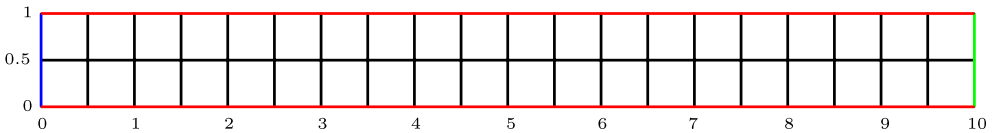


Fig. 4. Grid used for Couette flow; plates (red), inlet (blue) and outlet (green).

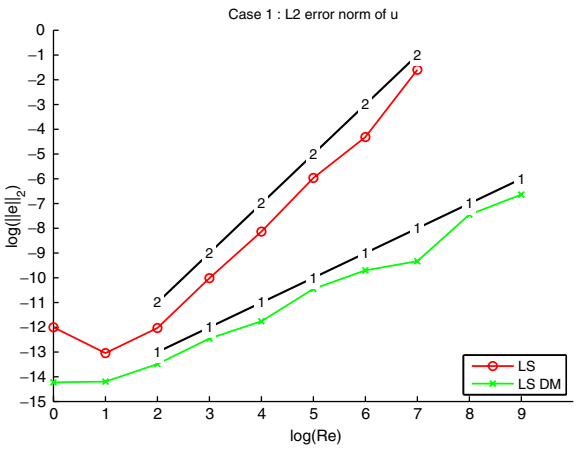


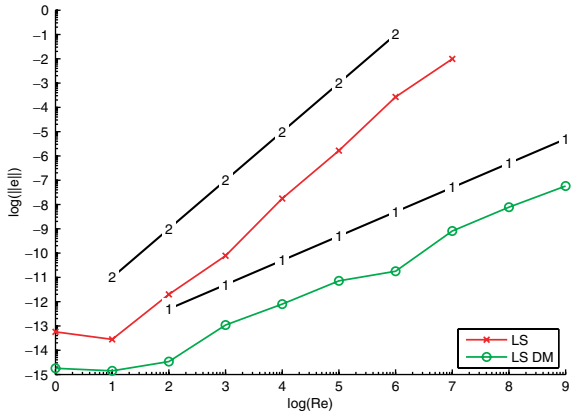
Fig. 5. The L^2-error in u as a function of the Reynolds number. Conventional least-squares (red) and the solution obtained by Direct Minimization (green).

Table 3
Comparison between the condition numbers obtained from Direct Minimization (DM) and the conventional least-squares method (LS) as a function of Re

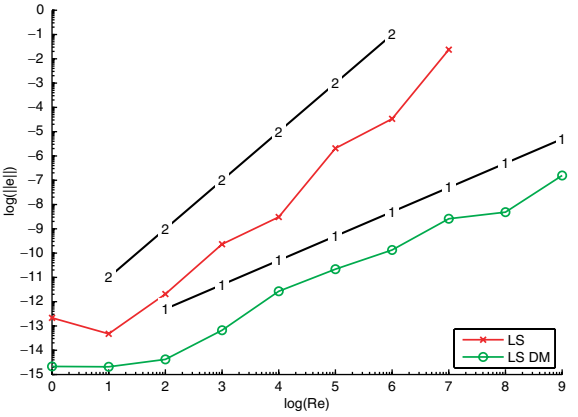
Re	DM	LS
10^0	9.57E2	9.15E2
10^1	1.99E2	3.97E4
10^2	7.08E2	5.03E5
10^3	7.08E3	5.01E7
10^4	7.08E4	5.01E9
10^5	7.08E5	5.01E11
10^6	7.08E6	5.01E13
10^7	7.08E7	–

compared. The condition number of the conventional least-squares formulation is the square of the condition number of Direct Minimization for a given Reynolds number.

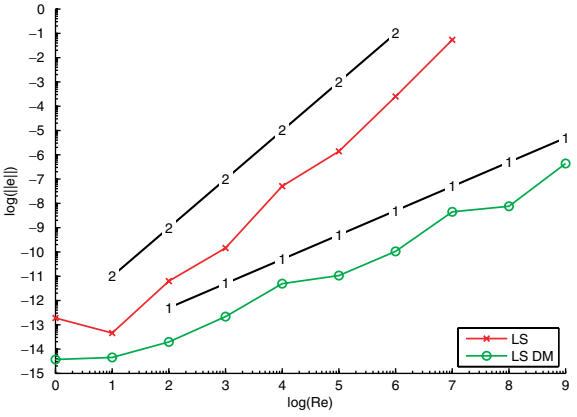
A similar trend is observed if we vary the polynomial degree, see Fig. 6. Note that the black lines which indicate the growth rate of the error are at the same position in all sub-figures of Fig. 6. This



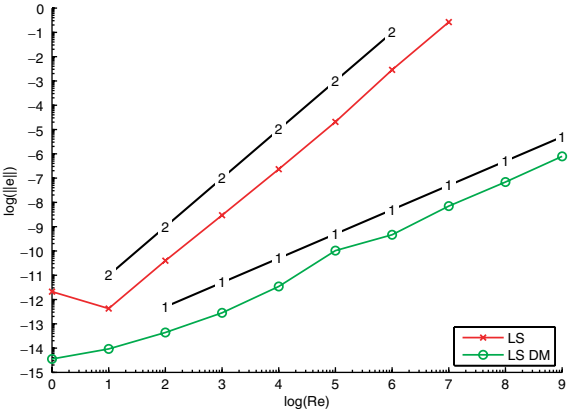
(a) $N = 2$



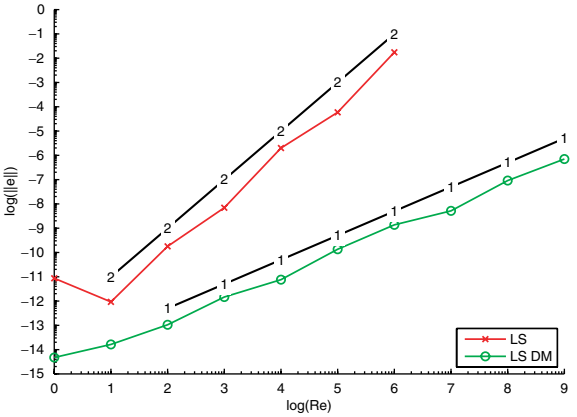
(b) $N = 3$



(c) $N = 4$



(d) $N = 6$



(e) $N = 8$

Fig. 6. The L^2-error of the variable u with a fixed $L = 5$ and variable N.

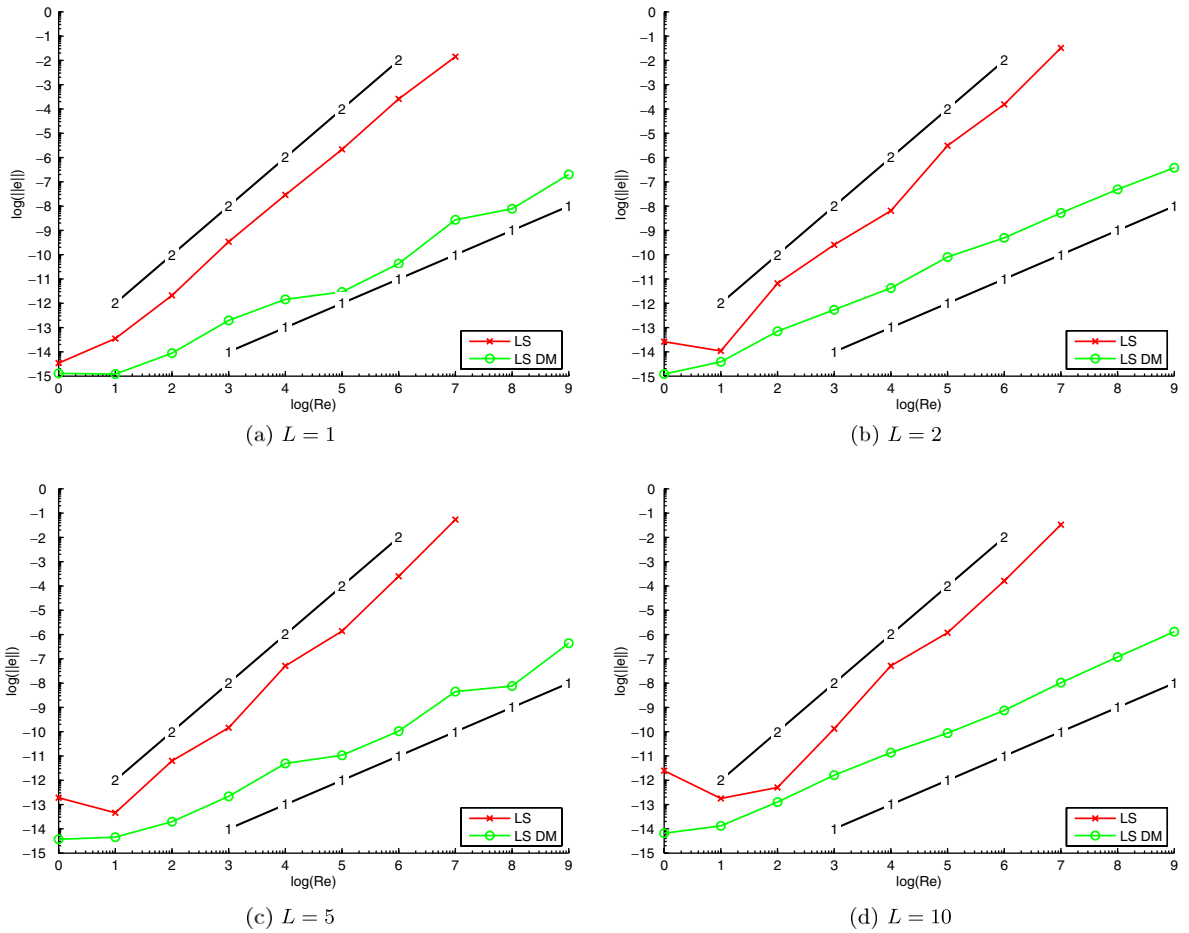
indicates that the condition number also grows as a function of the polynomial degree as was mentioned previously.

The growth of the L^2-error as function of the channel length L and the number of elements is displayed in Fig. 7.

It was argued in this paper that it is advantageous to avoid the costly matrix–matrix multiplication in the conventional least-squares formulation. However, the use of a QR-decomposition is generally more time consuming than the use of a Choleski decomposition for the symmetric, positive definite systems which need to be solved when the conventional least-squares method is used.

In Fig. 8 the CPU times for both methods are compared for various polynomial degrees. One observes that in the case of this very simple Couette flow, the least-squares formulation is faster for Reynolds numbers up to 10^6. The number of non-linear Newton iterations is equal to two for both formulations until the error in conventional LS becomes too big, in which case the number of non-linear iterations and consequently the CPU time increases significantly. A similar trend is observed in Fig. 9, where the CPU time is shown as a function of the Reynolds number for various channel lengths, L.

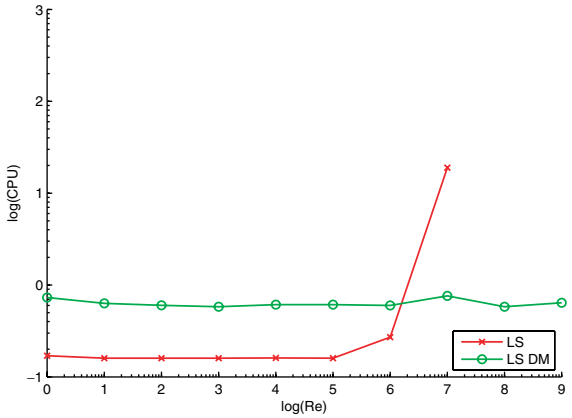
One of the reasons that Direct Minimization as described in this paper is less competitive with the conventional least-squares approach in terms of CPU time, is the use of the expensive full QR-decomposition. This is a direct consequence of the fact that we use a direct solver in this paper and this requires that we need both the Q matrix and the R matrix. If an iterative method would have been used, only the R matrix is required. This R-matrix then acts as a pre-conditioner and does not need to be computed every



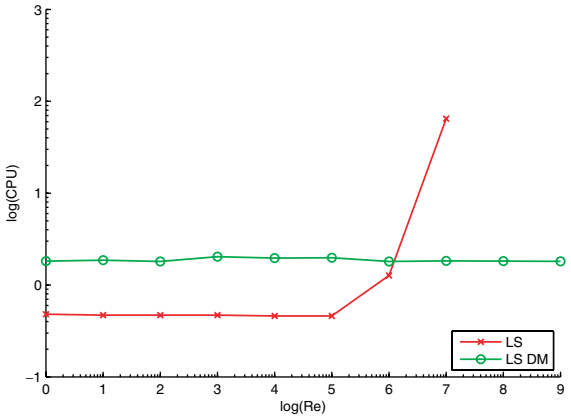
(a) $L = 1$ (b) $L = 2$

(c) $L = 5$ (d) $L = 10$

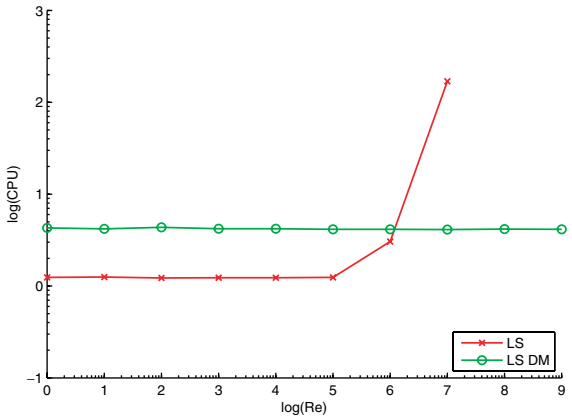
Fig. 7. The L^2-error of the variable u for a fixed $P = 4$ and variable L.



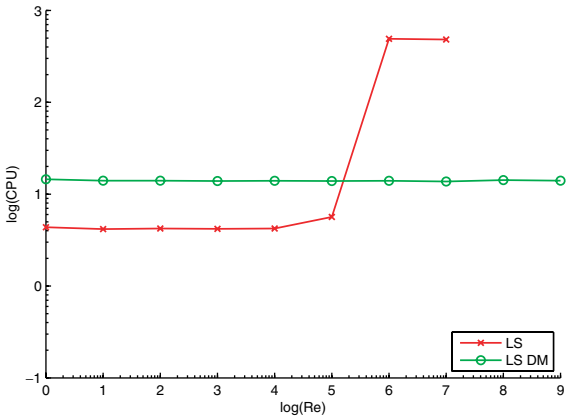
(a) $N = 2$



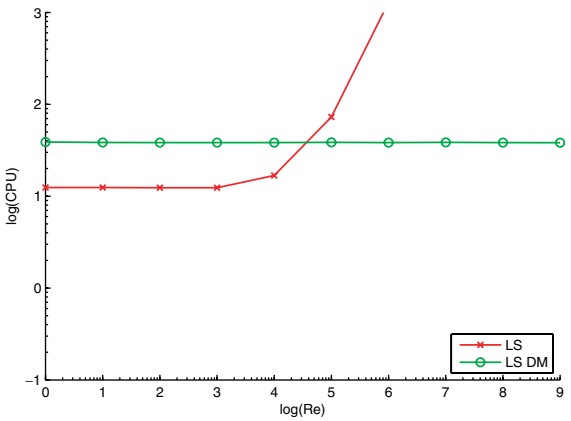
(b) $N = 3$



(c) $N = 4$



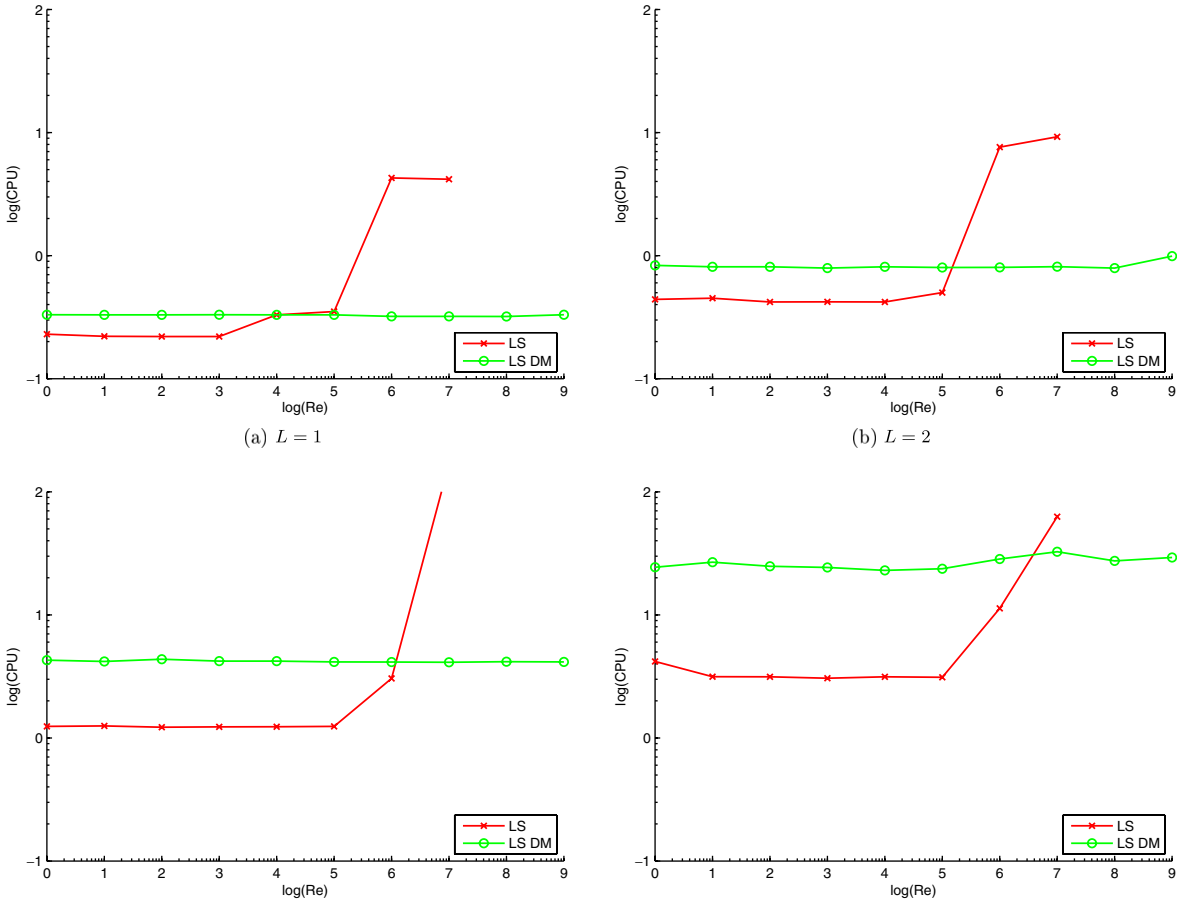
(d) $N = 6$



(e) $N = 8$

Fig. 8. The CPU time for a fixed $L = 5$ and variable N.

iteration. Furthermore, when we apply Direct Minimization for an iterative solver, there is no need to set up the global matrix and the sparsity pattern that was mentioned in this paper can then be fully employed. The use of iterative methods to solve overdetermined algebraic systems will be addressed in Part II of this paper.



(a) $L = 1$ (b) $L = 2$

(c) $L = 5$ (d) $L = 10$

Fig. 9. The CPU time for a fixed $N = 4$ and variable L.

8. Conclusions

By directly minimizing the residual in a suitably chosen norm, without variational analysis, a new computational solution procedure has been found for the least-squares method. Direct Minimization does not require the costly setup of the normal equations and thereby avoids the squaring of the condition number compared to conventional LSQSEM. Using a two dimensional model problems, it is demonstrated that for well-conditioned systems Direct Minimization performs equally well as the conventional least-squares method in terms of accuracy, but performs better than the conventional least-squares method for ill-conditioned problems.

The reduction in condition number grows with polynomial degree and number of elements. Direct Minimization is therefore the advocated least-squares method when a large number of high order spectral elements are used.

Although costly matrix–matrix multiplications are avoided in the use of Direct Minimization, plots of the CPU time versus the Reynolds number indicate that for the small sample problems presented in this paper, the conventional least-squares approach is still faster. This is a consequence of the fact that in the direct solver, described in this paper, both the Q matrix and the R matrix need to be computed.

References

[1] Å. Björck, Least squares methods, in: P.G. Ciarlet, J.L. Lions (Eds.), Handbook of Numerical Analysis, Solution of Equations in R^n. Part 1, vol. I, Elsevier, North Holland, Amsterdam, 1990, pp. 466–647.

[2] Å. Björck, The calculation of linear least squares problems, Acta Numer. (2004) 1–53.
[3] P. B. Bochev, Finite element methods based on least-squares and modified variational principles. Technical report, POSTECH, 2001.
[4] P.B. Bochev, M.D. Gunzburger, Finite element methods of least-squares type, SIAM Rev. 40 (4) (1998) 789–837.
[5] Z. Cai, R. Lazarov, T.A. Manteuffel, S.F. McCormick, First-order system least-squares for second-order partial differential equations: Part I 31 (6) (1994) 1785–1799.
[6] Z. Cai, T.A. Manteuffel, S.F. McCormick, First-order system least-squares for second-order partial differential equations: Part II. 34 (2) (1997) 425–454.
[7] A. Canuto, C. Hussaini, M.Y. Quarteroni, T.A. Zang, Spectral Methods in Fluid Dynamics. Springer Series in Computational Physics (1988).
[8] G.F. Carey, B.-N. Jiang, Least-squares finite element method and preconditioned conjugate gradient solution 24 (1987) 1283–1296.
[9] M.O. Deville, P.F. Fisher, E.H. Mund, High-Order Methods for Incompressible Fluid Flow, Cambridge University Press, 2002.
[10] M. Gerritsma, Direct minimization of the discontinuous least-squares spectral element method for viscoelastic fluids, J. Sci. Comput. 27 (1–3) (2006) 245–256.
[11] M.I. Gerritsma, M.M.J. Proot, Analysis of a discontinuous least squares spectral element method, J. Sci. Comput. 17 (1–3) (2002) 323–333.
[12] W. Heinrichs, Least-squares spectral collocation for the Navier–Stokes equations, J. Sci. Comput. 21 (2004) 81–90.
[13] W. Heinrichs, Spectral collocation schemes on the unit disc, J. Comput. Phys. 199 (2004) 66–86.
[14] W. Heinrichs, An adaptive least-squares spectral element collocation method with triangular elements for the incompressible Navier–Stokes equations, J. Eng. Math. 56 (2006) 337–350.
[15] B.-N. Jiang, On the least-squares method, Comput. Meth. Appl. Mech. Engrg. 152 (1–2) (1998) 239–257.
[16] B.N. Jiang, The Least Squares Finite Element Method – Theory and Applications in Computational Fluid Dynamics and Electromagnetics, Springer, 1998.
[17] W.B. Richardson Jr, Sobolev preconditioning for the Poisson–Boltzmann equation, Comput. Meth. Appl. Mech. Engrg. 181 (2000) 425–436.
[18] W.B. Richardson Jr., Steepest descent using smoothed gradients, Appl. Math. Comp. 112 (2000) 241–254.
[19] G.E.M. Karniadakis, S.J. Sherwin, Spectral h/p Element Methods for CFD, Oxford University Press, 1999.
[20] W.H. Lin, A least-squares spectral element method for sound propagation in acoustic ducts, J. Acoust. Soc. Am. 104 (5) (1998) 3111–3114.
[21] W.H. Lin, Least-squares legendre spectral element solution to sound propagation problems, J. Acoust. Soc. Am. 109 (2) (2001) 465–474.
[22] G. Mazzarella, F. Maggio, C. Pitzianti, Least-squares spectral element method for the 2D Maxwell equations in frequency domain, Int. J. Numer. Model. 17 (2004).
[23] J.P. Pontaza, A least-squares finite element formulation for incompressible flows with improved velocity–pressure coupling, J. Comput. Phys. 217 (2) (2006) 563–588.
[24] J.P. Pontaza, J.N. Reddy, Spectral/hp least squares finite element formulation for the Navier–Stokes equation, J. Comput. Phys. 190 (2) (2003) 523–549.
[25] J.P. Pontaza, J.N. Reddy, Space-time coupled spectral/*hp* least squares finite element formulation for the incompressible Navier–Stokes equation, J. Comput. Phys. 197 (2) (2004) 418–459.
[26] M.M.J. Proot, M.I. Gerritsma, A least-squares spectral element formulation for the Stokes problem, J. Sci. Comput. 17 (1–3) (2002) 311–322.
[27] M.M.J. Proot, M.I. Gerritsma, Mass- and momentum conservation of the least-squares spectral element method for the Stokes problem, J. Sci. Comput. 27 (1–3) (2006) 389–401.
[28] M.M.J. Proot. The Least-Squares Spectral Element Method – Theory, Implementation and Application to Incompressible Flows. PhD thesis, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands, 2003.
[29] H. De Sterck, T.J. Manteuffel, S.F. McCormick, Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs, SIAM J. Sci. Comput. 26 (2004) 31–54.
[30] H. De Sterck, T.J. Manteuffel, S.F. McCormick, Numerical conservation properties of h(div)-conforming least-squares finite element methods for the Burgers equation, SIAM J. Sci. Comput. 26 (5) (2005) 1573–1597.