

A Stopping Criterion for the Conjugate Gradient Algorithm in a Finite Element Method Framework

Mario Arioli¹

ABSTRACT

The Conjugate Gradient method has always been successfully used in solving the symmetric and positive definite systems obtained by the finite element approximation of self-adjoint elliptic partial differential equations. Taking into account recent results by Golub and Meurant (1997), Meurant (1997), Meurant (1999*a*), and Strakoš and Tichy (2002) which make it possible to approximate the energy norm of the error during the conjugate gradient iterative process, we adapt the stopping criterion introduced by Arioli, Noulard and Russo (2001). Moreover, we show that the use of efficient preconditioners does not require to change the energy norm used by the stopping criterion. Finally, we present the results of several numerical tests that experimentally validate the effectiveness of our stopping criterion.

Keywords: Conjugate Gradient Method, Stopping Criteria, Finite Element Method.

AMS(MOS) subject classifications: 65F10, 65N30, 65F50.

Current reports available by anonymous ftp to <ftp.numerical.rl.ac.uk> in directory pub/reports.

¹ M.Arioli@rl.ac.uk, Rutherford Appleton Laboratory,

Computational Science and Engineering Department

Atlas Centre

Rutherford Appleton Laboratory

Oxon OX11 0QX

December 19, 2002

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 2 | An elementary Finite Element framework | 2 |
| 3 | The stopping criterion for the conjugate gradient method | 5 |
| 4 | Numerical experiments | 10 |
| 4.1 | L-shape test problems | 10 |
| 4.1.1 | Problem 1 | 12 |
| 4.1.2 | Problem 2 | 13 |
| 4.2 | Three dimensional test problems | 13 |
| 4.3 | Error upper bounds and practical choices of d | 13 |
| 5 | Conclusions | 20 |

1 Introduction

In this paper, we combine linear algebra techniques with finite element techniques to obtain a reliable stopping criterion for the conjugate gradient algorithm. The finite element method approximates the weak form of an self adjoint, coercive elliptic partial differential equation defined within a Hilbert space by a linear system of equations

$$Ax = b$$

where $A \in \mathbb{R}^{N \times N}$ is symmetric and positive definite and $b \in \mathbb{R}^N$. The conjugate gradient method is a very effective iterative algorithm for solving these linear systems. In particular, using the conjugate gradient algorithm, we will compute the information which is necessary to evaluate the energy norm of the difference between the solution of the continuous problem, and the approximate solution obtained when we stop the iterations by our criterion. Owing to the close relationship between the conjugate gradient method behaviour and the variational properties of finite element methods, we will focus only on stopping criteria based on the energy norm: $\|y\|_A^2 = y^T A y$. Moreover, our experiments give very good evidence that the usual stopping criterion based on the Euclidean norm of the residual $b - Ax$ can be totally unsatisfactory and frequently misleading.

Recently, several authors have proposed rules that compute error bounds for the conjugate gradient method (Ashby, Holst, Manteuffel and Saylor 2001, Axelsson and Kaporin 2001, Calvetti, Morigi, Reichel and Sgallari 2000, Calvetti, Morigi, Reichel and Sgallari 2001, Golub and Meurant 1997, Golub and Strakoš 1994, Meurant 1997, Meurant 1999*a*, Meurant 1999*b*, Strakoš and Tichy 2002). Some of these rules compute estimate of the error in Euclidean norm and others compute estimates related to the energy norm. In their historical paper, Hestenes and Stiefel (1952) proposed a method for the estimate of the error energy norm that use the values computed during the conjugate gradient method. Strakoš and Tichy, (2002), studied the relations between the estimates proposed by Hestenes and Stiefel (1952), Golub and Meurant (1997), Golub and Strakoš (1994), and Meurant (1997), (1999*a*), (1999*b*) and they proved that the Hestenes-Stiefel estimate is numerically stable.

We shall first summarise the principal properties of the finite-element method in Section 2. Then, in Section 3, we will use the recent results of Arioli et al. (2001), Golub and Meurant (1997), Meurant (1997), (1999*a*), and Strakoš and Tichy (2002) to build reliable stopping criteria and to analyse their properties. Finally, in Section 4, we will present the numerical experiments we performed on a selected ill-conditioned test problem, and, in Section 5, we will present our conclusions.

For the sake of simplicity, we will mainly focus on the 2D case. The results can be easily extended to the 3D case. In the paper, we will use the following notation.

Let Ω be a simply connected bounded polygonal domain in \mathbb{R}^2 , defined by a closed curve Γ .

In the following, we will denote by $\mathcal{D}(\Omega)$ the space of all infinitely differentiable functions with compact support in Ω and by

$$D^\alpha \mathbf{u} = \frac{\partial^{|\alpha|} \mathbf{u}}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2}, \quad \alpha = (\alpha_1, \alpha_2) \in \mathbf{N}^2, \quad |\alpha| = \alpha_1 + \alpha_2, \quad \mathbf{x} = (x_1, x_2).$$

Furthermore, we will denote by $H^s(\Omega)$ (Grisvard 1992) the space of all distributions \mathbf{u} defined in Ω that satisfy the following properties

- $D^\alpha \mathbf{u} \in L^2(\Omega)$ for $|\alpha| \leq m$ when $s = m$ is a non-negative integer,
- $\mathbf{u} \in H^m(\Omega)$ and

$$|\mathbf{u}|_{s,\Omega} = \sum_{|\alpha|=m} \int \int_{\Omega \times \Omega} \frac{|D^\alpha \mathbf{u}(\mathbf{x}) - D^\alpha \mathbf{u}(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|^{2+2\sigma}} d\mathbf{x}d\mathbf{y} < +\infty,$$

where $m = \lfloor s \rfloor$ and $\sigma = s - \lfloor s \rfloor$ when $s \in \mathbb{R}_+ \setminus \mathbb{N}$.

We define the Hilbert norm on $H^s(\Omega)$ by

$$\|\mathbf{u}\|_{s,\Omega} = \left(\sum_{|\alpha| \leq m} \int_{\Omega} |D^\alpha \mathbf{u}(\mathbf{x})|^2 d\mathbf{x} \right)^{1/2},$$

in the first case, and, by

$$\|\mathbf{u}\|_{s,\Omega} = \left(\|\mathbf{u}\|_{m,\Omega}^2 + \sum_{|\alpha|=m} \int \int_{\Omega \times \Omega} \frac{|D^\alpha \mathbf{u}(\mathbf{x}) - D^\alpha \mathbf{u}(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|^{2+2\sigma}} d\mathbf{x}d\mathbf{y} \right)^{1/2},$$

in the second case. Finally, we will denote by $H_0^s(\Omega)$ the closure of $\mathcal{D}(\Omega)$ in $H^s(\Omega)$, and by $H^{-s}(\Omega)$ the dual space of $H_0^s(\Omega)$.

2 An elementary Finite Element framework

Let

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathfrak{K}(\mathbf{x}) \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\mathbf{x}, \quad \forall \mathbf{u}, \mathbf{v} \in H_0^1(\Omega) \quad (1)$$

be a continuous and coercive bilinear form:

$\forall \mathbf{u}, \mathbf{v} \in H_0^1(\Omega)$, $\exists \gamma \in \mathbb{R}_+$ and $\exists M \in \mathbb{R}_+$ such that

$$\gamma \|\mathbf{u}\|_{1,\Omega}^2 \leq a(\mathbf{u}, \mathbf{u}) \quad (2)$$

$$a(\mathbf{u}, \mathbf{v}) \leq M \|\mathbf{u}\|_{1,\Omega} \|\mathbf{v}\|_{1,\Omega}, \quad (3)$$

and $L(\mathbf{v}) = \int_{\Omega} f \mathbf{v} d\mathbf{x}$ be a continuous linear functional, $L(\mathbf{v}) \in H^{-1}(\Omega)$. Using the hypotheses stated above the problem

$$\begin{cases} \text{Find } \mathbf{u} \in H_0^1(\Omega) \text{ such that} \\ a(\mathbf{u}, \mathbf{v}) = L(\mathbf{v}), \quad \forall \mathbf{v} \in H_0^1(\Omega), \end{cases} \quad (4)$$

has a unique solution. A finite element approximation of problem (4) with the use of continuous piecewise linear elements can be briefly described as follows. Let \mathcal{T}_h be a family of triangulations of Ω , i.e. each \mathcal{T}_h is a set of disjoint triangles $\{T\}$ which covers Ω in such a way that no vertex of any triangle lies in the interior of an edge of another triangle. Let $h = \max_{T \in \mathcal{T}_h} \text{diameter}(T)$, we assume that \mathcal{T}_h is *regular* in the sense of Ciarlet (1978, page 132), i.e. triangles do not

degenerate as $h \rightarrow 0$. Moreover, we assume that each triangle cannot have more than one edge lying on Γ .

Consider then the space:

$$V_h = \{\mathbf{v}_h(\mathbf{x}) : \Omega \rightarrow \mathbb{R}, \mathbf{v}_h(\mathbf{x})|_T \text{ is linear } \forall T \in \mathcal{T}_h, \mathbf{v}_h|_\Gamma = 0\}.$$

The space V_h is the piecewise linear space (we refer to Ciarlet, 1978, for a detailed analysis), and $V_h \subset H_0^1(\Omega)$. We will now describe the usual finite element basis for V_h . Let $\{P_j\}_{j=1,\dots,N}$ be the set of internal vertices of \mathcal{T}_h (i.e. we exclude the vertices lying on Γ). Then for all j , $1 \leq j \leq N$, we define the function $\phi_j \in V_h$ by

$$\phi_j(P_i) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and then we extend it linearly on each triangle T . It is easy to show that $\{\phi_j\}_{j=1,\dots,N}$ is a basis for V_h ; hence, $\dim V_h = N$. Every $v_h \in V_h$ is a linear combination of the functions belonging to the basis

$$\mathbf{v}_h(\mathbf{x}) = \sum_{j=1}^N v_j \phi_j(\mathbf{x}).$$

Therefore, the approximated problem that we want to solve, will be

$$\begin{cases} \text{Find } \mathbf{u}_h \in V_h \text{ such that} \\ a(\mathbf{u}_h, \mathbf{v}_h) = L(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h, \end{cases} \quad (5)$$

Hereafter, for the sake of simplicity, we assume that all the integrals are computed exactly. The approximated problem (5) is equivalent to the following system of linear equations:

$$Au = b \quad (6)$$

where A and b are defined as follows

$$\begin{aligned} A_{ij} &= a(\phi_i, \phi_j), \\ b_i &= L(\phi_i). \end{aligned}$$

If we use an iterative method, at each step k we will have a vector $u^{(k)} \in \mathbb{R}^N$, which in turn identifies a function $\mathbf{u}_h^{(k)}(\mathbf{x}) = \sum_{i=1}^N u_i^{(k)} \phi_i(\mathbf{x})$, $\mathbf{u}_h^{(k)}(\mathbf{x}) \in V_h$, and a residual $R_h^{(k)} \in V_h'$ (V_h' is the topological dual space of V_h) which is defined by

$$R_h^{(k)}(v_h) = a(\mathbf{u}_h^{(k)}, \mathbf{v}_h) - L(\mathbf{v}_h), \quad \forall \mathbf{v}_h \in V_h.$$

Arioli et al. (2001) propose a stopping criterion based on the evaluation of the dual norm of $R_h^{(k)}(\mathbf{v}_h)$:

$$\|R_h^{(k)}(\mathbf{v}_h)\|_{H^{-1}},$$

which is appropriate when the bilinear form is non symmetric.

Here however, we want to take full advantage of the following relation:

$$v^T Av = a(\mathbf{v}_h(\mathbf{x}), \mathbf{v}_h(\mathbf{x})), \quad \forall \mathbf{v}_h \in V_h \quad (7)$$

and of the fact that the bilinear form $a(\cdot, \cdot)$ induces on $H_0^1(\Omega)$ the norm: $\|\cdot\|_a = (a(\cdot, \cdot))^{1/2}$ equivalent to $\|\cdot\|_{1,\Omega}$. Since, Ciarlet (1978, page 105),

$$\|\mathbf{u} - \mathbf{u}_h\|_a \leq \|\mathbf{u} - \mathbf{v}_h\|_a, \quad \forall \mathbf{v}_h \in V_h$$

we have that

$$\|\mathbf{u} - \mathbf{u}_h\|_a \leq \|\mathbf{u} - \Pi_h \mathbf{u}\|_a, \quad (8)$$

$$\|\mathbf{u}_h\|_a \leq \|\mathbf{u}\|_a + \|\mathbf{u} - \mathbf{u}_h\|_a, \quad (9)$$

where $\Pi_h \mathbf{u} \in V_h$ denotes a suitable interpolation of \mathbf{u} (Ciarlet 1978).

Let $\mathbf{u}_h^* \in V_h$ be a function such that:

$$\|\mathbf{u}_h - \mathbf{u}_h^*\|_a^2 = a(\mathbf{u}_h - \mathbf{u}_h^*, \mathbf{u}_h - \mathbf{u}_h^*) \leq h^{2t} a(\mathbf{u}_h, \mathbf{u}_h) = h^{2t} \|\mathbf{u}_h\|_a^2. \quad (10)$$

Therefore, using (9), we can give the following estimate of the error $\|\mathbf{u} - \mathbf{u}_h^*\|_a$:

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h^*\|_a &\leq \|\mathbf{u} - \mathbf{u}_h\|_a + \|\mathbf{u}_h - \mathbf{u}_h^*\|_a \\ &\leq \|\mathbf{u} - \mathbf{u}_h\|_a + h^t \|\mathbf{u}_h\|_a \\ &\leq \|\mathbf{u} - \mathbf{u}_h\|_a + h^t (\|\mathbf{u}\|_a + \|\mathbf{u} - \mathbf{u}_h\|_a) \\ &\leq h^t \|\mathbf{u}\|_a + \|\mathbf{u} - \mathbf{u}_h\|_a (1 + h^t). \end{aligned}$$

Finally, assuming that $h < 1$ and $t > 0$, we have

$$\|\mathbf{u} - \mathbf{u}_h^*\|_a \leq h^t \|\mathbf{u}\|_a + 2 \|\mathbf{u} - \mathbf{u}_h\|_a. \quad (11)$$

In the following, the function \mathbf{u}_h^* will be $\mathbf{u}_h^{(k)}$ the function identified by the last iteration of the conjugate gradient method. The value of the parameter t is related to the threshold we will use in our stopping criterion and will depend on the regularity of the solution. In the present paper, we do not assume that the domain Ω is convex. In particular, in some of our numerical tests Ω has an L-shape. Therefore, the solution $\mathbf{u} \in H^s(\Omega) \cap H_0^1(\Omega)$, with $1 < s < \frac{5}{3}$ (Grisvard 1992). Under this regularity of \mathbf{u} , it is possible to prove (Babuska 1971, Babuska 1972, Dupont and Scott 1980) that

$$\|\mathbf{u} - \Pi_h \mathbf{u}\|_{1,\Omega} \leq Ch^{s-1} |\mathbf{u}|_{s,\Omega}, \quad (12)$$

with C independent from h and \mathbf{u} . Therefore, from (2), (3), (11), (8), and (12), we have that

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h^*\|_{1,\Omega} &\leq 2C \sqrt{\frac{M}{\gamma}} h^{s-1} (h^{t-s+1} \|\mathbf{u}\|_{1,\Omega} + |\mathbf{u}|_{s,\Omega}) \\ &\leq C_1 \sqrt{\frac{M}{\gamma}} h^{s-1} \|\mathbf{u}\|_{s,\Omega}, \end{aligned}$$

with C_1 independent from h and \mathbf{u} .

Remark 2.1 Let us assume that \mathfrak{K} is a piecewise constant function and \mathcal{T}_h is such that $\mathfrak{K}|_T = \text{const}$, $\forall T \in \mathcal{T}_h$. It is possible to express the bilinear form $a(\cdot, \cdot)$ as follows

$$a(\mathbf{u}, \mathbf{u}) = \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T |\nabla \mathbf{u}|_{0,T}^2, \quad \forall \mathbf{u} \in H_0^1(\Omega). \quad (13)$$

Therefore, the error $\|\mathbf{u} - \mathbf{u}_h\|_a$ can be more precisely bounded as follows (Dupont and Scott 1980)

$$\|\mathbf{u} - \mathbf{u}_h\|_a^2 = \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T |\nabla(\mathbf{u} - \mathbf{u}_h)|_{0,T}^2 \quad (14)$$

$$\leq C_0 \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T h_T^{2(s-1)} |\mathbf{u}|_{s,T}^2, \quad (15)$$

where h_T is the diameter of the triangle T and C_0 is independent from h and \mathbf{u} . Let h_{\min} be the smallest diameter of a triangle: $h_{\min} = \min_{T \in \mathcal{T}_h} \text{diameter}(T)$. If $h^{-1} h_{\min} \geq c_1 h^p$, $c_1 \leq 1$ independent from h and h_{\min} , and $p \geq 0$, we have that

$$\begin{aligned} c_1^{2(s-1)} h^{2(p+1)(s-1)} \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T |\mathbf{u}|_{s,T}^2 &\leq \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T h_T^{2(s-1)} |\mathbf{u}|_{s,T}^2 \\ &\leq h^{2(s-1)} \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T |\mathbf{u}|_{s,T}^2. \end{aligned}$$

In particular, if $p > 0$ we have a non uniform mesh. If we consider the exponent t in (10) to be greater than $(p+1)(s-1)$, we assume that c_1 is not too small (i.e. $c_1 > 0.25$), and we are in favourable situation

$$\|\mathbf{u} - \mathbf{u}_h\|_a^2 \approx \sum_{T \in \mathcal{T}_h} \mathfrak{K}_T h_T^{2(s-1)} |\mathbf{u}|_{s,T}^2,$$

then we can reasonably expect that

$$\|\mathbf{u} - \mathbf{u}_h\|_a \approx \|\mathbf{u} - \mathbf{u}_h^*\|_a.$$

3 The stopping criterion for the conjugate gradient method

When using an iterative method for solving the linear system (6), we normally incorporate a stopping criterion based on the a-posteriori component-wise or norm-wise backward error theory (Arioli, Duff and Ruiz 1992, Higham 1996). If we use the conjugate gradient method, it is quite natural to have a stopping criterion which takes advantage of the minimization property of this method. At each step k the conjugate gradient method minimizes the energy norm of the error $\delta u^{(k)} = u - u^{(k)}$ (u solution of (6)) on a Krylov space $u^{(0)} + \mathcal{K}_k$ (Greenbaum 1997):

$$\min_{u^{(k)} \in u^{(0)} + \mathcal{K}_k} \delta u^{(k)T} A \delta u^{(k)}. \quad (16)$$

The space \mathbb{R}^N with the norm

$$\|y\|_A = (y^T A y)^{(1/2)}$$

induces on its dual space the dual norm

$$\|f\|_{A^{-1}} = (f^T A^{-1} f)^{1/2}.$$

Let $r^{(k)} = b - Au^{(k)}$ denote the residual at step k . Therefore, the value $\|\delta u^{(k)}\|_A$ will be equal to the dual norm of the residual $\|r^{(k)}\|_{A^{-1}}$. Moreover, from (7), we have that

$$\|u^{(k)}\|_A = \|\mathbf{u}_h^{(k)}(\mathbf{x})\|_a \quad \forall k.$$

Therefore, a stopping criterion such as the following:

$$\text{IF } \|Au^{(k)} - b\|_{A^{-1}} \leq \eta \|b\|_{A^{-1}} \text{ THEN STOP ,} \quad (17)$$

with $\eta < 1$ an a-priori threshold fixed by the user, will guarantee (Arioli et al. 2001) that a $u^{(k)}$ which satisfies it, is the solution of the perturbed linear system:

$$\begin{aligned} Au^{(k)} &= b - r^{(k)}, \\ \|r^{(k)}\|_{A^{-1}} &\leq \eta \|b\|_{A^{-1}}. \end{aligned}$$

Moreover, we have for $\mathbf{u}_h = \sum_{j=1}^N u_j \phi_j$, solution of (5), and $\mathbf{u}_h^{(k)} = \sum_{j=1}^N u_j^{(k)} \phi_j$ that

$$\|\mathbf{u}_h - \mathbf{u}_h^{(k)}\|_a = \|u - u^{(k)}\|_A = \|r^{(k)}\|_{A^{-1}} \leq \eta \|b\|_{A^{-1}} = \eta \|u\|_A = \eta \|\mathbf{u}_h\|_a.$$

The choice of η will depend on the properties of the problem that we want to solve, and, in the practical cases, η can be frequently much larger than ε , the roundoff unit of the computer finite precision arithmetic. From (10) and (12), a reasonable choice for η would be:

$$\eta = h^{s-1} \quad \text{or} \quad \eta = h^2.$$

Frequently, it is easier to have the area of each triangle for a given mesh instead of h . This is indeed the case in some of our experiments. Therefore, a practical choice could be:

$$\eta = \left(\max_{T_j \in \mathcal{T}_h} \int_{T_j} 1 d\mathbf{x} \right)^{1/2} \approx h.$$

First of all, we need to add, within the conjugate gradient algorithm, some tool for estimating the value $e_A^{(k)} = r^{(k)T} A^{-1} r^{(k)}$ at each step k . This can be achieved using Gauss quadrature rules as proposed by Golub and Meurant (1997) or using the rule presented by Hestenes and Stiefel in their original paper (1952). Golub and Meurant (1997) present three different quadrature families: Gauss, Gauss-Lobatto and Gauss-Radau. The Gauss quadrature does not require any a-priori knowledge of the smallest and the biggest eigenvalues and computes a lower bound of $e_A^{(k)}$. The other two quadrature rules compute both a lower and an upper bound using the extremes eigenvalues of A . In particular, the Hestenes-Stiefel rule computes a lower bound ξ_k for $e_A^{(k)}$ that is equal to the bound computed by the Gauss rule proposed by Golub and Meurant (1997). Moreover, Strakoš and Tichy (2002) proved that the Hestenes-Stiefel rule is numerically stable when finite precision arithmetic is used. Owing the better stability properties and the independence of a-priori estimates of the spectrum, we choose the Hestenes-Stiefel rule. The Hestenes-Stiefel rule can be cheaply computed using the quantities already computed during

the conjugate gradient process. The conjugate gradient iterates satisfy the following relations (Greenbaum 1997, Meurant 1999a):

$$\begin{aligned} u^{(k)} &= u^{(k-1)} + \alpha_{k-1} p^{(k-1)}, & \alpha_{k-1} &= \frac{r^{(k-1)T} r^{(k-1)}}{p^{(k-1)T} A p^{(k-1)}}, \\ r^{(k)} &= r^{(k-1)} - \alpha_{k-1} A p^{(k-1)}, \\ p^{(k)} &= r^{(k)} + \beta_{k-1} p^{(k-1)}, & \beta_{k-1} &= \frac{r^{(k)T} r^{(k)}}{r^{(k-1)T} r^{(k-1)}}, \end{aligned}$$

where $u^{(0)} = 0$ and $r^{(0)} = p^{(0)} = b$. The quantity α_{k-1} gives the step-size on the direction $p^{(k-1)}$ during the conjugate gradient algorithm. Therefore, in exact arithmetic, we have that the final value

$$u = \sum_{j=1}^N \alpha_j p^{(j)},$$

and taking into account that

$$p^{(j)T} A p^{(i)} = 0, \quad i \neq j,$$

the energy norm of the error is

$$\|\delta u^{(k)}\|_A^2 = e_A^{(k)} = \sum_{j=k+1}^N \alpha_j r^{(j)T} r^{(j)}. \quad (18)$$

Under the assumption that $e_A^{(k+d)} \ll e_A^{(k)}$, where the integer d denotes a suitable delay, the Hestenes-Stiefel estimate ξ_k will be then computed by the formula

$$\xi_k = \sum_{j=k+1}^{k+d} \alpha_j r^{(j)T} r^{(j)}.$$

(Golub and Meurant 1997) indicated $d = 10$ as a successful compromise, and numerical experiments support this conclusion ((Golub and Meurant 1997, Arioli and Baldini 2001)). In Section 4, we will indicate that the cheaper choice $d = 5$ can be reliable if the solution \mathbf{u} of (4) is reasonably regular, and we will experimentally compare several choices for the value of d when the matrix A is very ill conditioned and the solution \mathbf{u} is only continuous. In this latter case, we must choose a larger value for d .

Finally, we must estimate $b^T A^{-1} b$. It follows from (16) that

$$r^{(k)T} v = 0, \quad \forall v \in \mathcal{K}_k.$$

Therefore, we have

$$\begin{aligned} \delta u^{(k)T} A \delta u^{(k)} &= u^T A u + u^{(k)T} A u^{(k)} - u^{(k)T} A u \\ &= u^T A u + u^{(k)T} A u^{(k)} - 2b^T u^{(k)} \\ &= u^T A u - (u^{(0)} + v^{(k)})^T r^{(k)} - b^T u^{(k)} \\ &= u^T A u - u^{(0)T} r^{(k)} - b^T u^{(k)} \\ &= u^T A u - u^{(k)T} r^{(0)} - b^T u^{(0)}. \end{aligned}$$

Then, it follows that

$$u^T Au \geq u^{(k)T} r^{(0)} + b^T u^{(0)}, \quad (19)$$

and the right-hand side will converge monotonically to $\|u\|_A^2$. Taking into account (19), we could replace $\|b\|_{A^{-1}}$ with its lower bound at the step k of the conjugate gradient method. Therefore, we can substitute (17) with:

$$\text{IF } \xi_k \leq \eta^2(u^{(k)T} r^{(0)} + b^T u^{(0)}) \quad \text{THEN STOP.} \quad (20)$$

Introducing a preconditioner, we want to speed up the convergence rate of the conjugate gradient method but this will change the matrix and, therefore, the energy norm. However, we still want to estimate $e_A^{(k)}$. Nonetheless, we can prove that the energy norm of the preconditioned problem is equal to $e_A^{(k)}$.

Let us assume that we symmetrically precondition the linear system (6) by the non singular matrix U . We obtain the equivalent system

$$U^{-T} AU^{-1} y = U^{-T} b, \quad (21)$$

where $y = Uu$. If we directly apply the conjugate gradient method to (21), the iterates satisfy the following relations (Greenbaum 1997, Meurant 1999a):

$$\begin{aligned} y^{(k)} &= y^{(k-1)} + \alpha_{k-1} \hat{p}^{(k-1)}, & \alpha_{k-1} &= \frac{\hat{r}^{(k-1)T} \hat{r}^{(k-1)}}{\hat{p}^{(k-1)T} U^{-T} AU^{-1} \hat{p}^{(k-1)}}, \\ \hat{r}^{(k)} &= \hat{r}^{(k-1)} - \alpha_{k-1} U^{-T} AU^{-1} \hat{p}^{(k-1)}, \\ \hat{p}^{(k)} &= \hat{r}^{(k)} + \beta_{k-1} \hat{p}^{(k-1)}, & \beta_{k-1} &= \frac{\hat{r}^{(k)T} \hat{r}^{(k)}}{\hat{r}^{(k-1)T} \hat{r}^{(k-1)}}, \end{aligned}$$

where $y^{(0)} = 0$ and $\hat{r}^{(0)} = \hat{p}^{(0)} = U^{-T} b$. Moreover, in exact arithmetic, we have that $\hat{r}^{(k)} = U^{-T} b - U^{-T} AU^{-1} y^{(k)}$, and, therefore, defining $u^{(k)} = U^{-1} y^{(k)}$, we have that

$$\hat{r}^{(k)} = U^{-T} (b - Au^{(k)}) = U^{-T} r^{(k)}.$$

Then, we have that

$$\begin{aligned} \|\hat{r}^{(k)}\|_{(U^{-T} AU^{-1})^{-1}}^2 &= \hat{r}^{(k)T} U A^{-1} U^T \hat{r}^{(k)} \\ &= \|r\|_{A^{-1}}^2. \end{aligned}$$

Finally, if we define $p^{(k)} = U^{-1} \hat{p}^{(k)}$, and $M = U^T U$, we obtain the variant of the preconditioned conjugate gradient algorithm, which incorporates the proposed stopping criterion with a suitable choice of d described in Figure 1.

Remark 3.1 *The value of the exponent t in formula (10) can be higher than the one suggested by (12) which depends on the regularity of the solution. This is the case when super-convergence in the nodes of the mesh occurs and we know that the values in the mesh nodes are very accurate (Wahlbin 1995, Babuska, Strouboulis, Upadhyay and Gangaraj 2001).*

```

Preconditioned Conjugate Gradient Algorithm (PCG)
Given an initial guess  $u^{(0)}$ , compute  $r^{(0)} = b - Au^{(0)}$ , and solve
 $Mz^{(0)} = r^{(0)}$ . Set  $p^{(0)} = z^{(0)}$ ,  $\beta_0 = 0$ ,  $\alpha_{-1} = 1$ ,  $\rho_0 = b^T u^{(0)}$ , and
 $\xi_0 = \infty$ .

 $k = 0$ 
while  $\xi_k > \eta^2(\rho_0 + r^{(0)T} u^{(k)})$  do
   $k = k + 1$ ;
   $\chi_k = r^{(k-1)T} z^{(k-1)}$ ;
   $\alpha_{k-1} = \frac{r^{(k-1)T} z^{(k-1)}}{p^{(k-1)T} Ap^{(k-1)}}$ ;
   $\psi_k = \alpha_{k-1} \chi_k$ ;
   $u^{(k)} = u^{(k-1)} + \alpha_{k-1} p^{(k-1)}$ ;
   $r^{(k)} = r^{(k-1)} - \alpha_{k-1} Ap^{(k-1)}$ ;
  Solve  $Mz^{(k)} = r^{(k)}$ ;
   $\beta_k = \frac{r^{(k)T} z^{(k)}}{r^{(k-1)T} z^{(k-1)}}$ ;
   $p^k = z^k + \beta_k p^{(k-1)}$ ;
  if  $k > d$  then
     $\xi_k = \sum_{j=k-d+1}^k \psi_j$ ;
  else
     $\xi_k = \xi_{k-1}$ ;
  endif
end while.

```

Figure 1: Preconditioned Conjugate Gradient Algorithm (PCG)

Remark 3.2 *The conjugate gradient algorithm convergence rate can be estimated by the Chebyshev polynomials (Greenbaum 1997), and we have that*

$$\frac{\|u - u^{(k)}\|_A}{\|u\|_A} \leq 2 \left[\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right]^k,$$

where $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$ is the condition number of A . Therefore, the number k^* needed to guarantee that

$$\|u - u^{(k^*)}\|_A \leq \eta \|u\|_A,$$

can be estimated by

$$k^* \approx \lceil \log \eta \rceil \frac{\sqrt{\kappa(A)}}{2}, \quad (22)$$

and, choosing $\eta = h$,

$$k^* \approx \frac{1}{2} \lceil \log h \rceil \sqrt{\kappa(A)} \approx \frac{1}{2} \sqrt{\kappa(A)} \log N.$$

Finally, we observe that, if $\mathfrak{K} = 1$ in Ω , $\kappa(A) \approx h^{-2}$. Therefore, for this Poisson like equation, we can reach convergence when

$$k^* \approx \frac{1}{2} |\log h| h^{-1} \approx \frac{1}{2} \sqrt{N} \log N.$$

This bound is very pessimistic, and, in our experiments, we reach convergence much before. Nonetheless, these bounds present a theoretical interest when we look at 3D problems. The previous theory still holds without major changes and we have:

$$k^* \approx \frac{1}{2} |\log h| h^{-1} \approx \frac{1}{2} N^{\frac{1}{3}} \log N.$$

This upper bound of the max number of steps needed to have convergence, combined with the cost of a step of the conjugate gradient method, gives an estimate of the global computational complexity which is asymptotically better than the one for a direct solver.

Remark 3.3 The effect of rounding errors on the Gauss quadrature calculations has been analysed by Golub and Strakoš (1994). More precisely, Golub and Strakoš (1994) proved that the total accuracy of the Gauss quadrature, computed by a finite precision arithmetic conjugate gradient algorithm, is related to the energy norm of the error of the conjugate gradient process itself.

Strakoš and Tichy (2002) proved that the Hestenes-Stiefel rule is mathematically equivalent to the Gauss quadrature and, moreover, that is numerically stable. In our numerical experiments, we tested both these rules and compute the relative error between them. The numerical tests show that both the rules give the same results within an error of order machine precision times $\kappa(U^{-T}AU^{-1})$.

4 Numerical experiments

We generated two test problem classes using **FEMLAB**[©] under **Matlab**[©]. The first test problem class is define on a L-shape domain Ω of \mathbb{R}^2 . The second test problem class is defined on the cube $[0, 1] \times [0, 1] \times [0, 1]$. In both the classes, we chose boundary condition zero and, in the conjugate gradient algorithm, the starting point $u^{(0)} = 0$. Finally, in all the figures, the estimate of the energy norm stops d steps before the final iteration because of the choice of our stopping criterion, and the values in the legends of the figures are:

- $\|Au^{(k)} - b\|_2 / \|b\|_2$, the value of the residual at step k is computed using A ;
- $\|\delta u\|_A / \|u\|_A = \|u - u^{(k)}\|_A / \|u\|_A$ energy norm of the algebraic error;
- $\|\delta \mathbf{u}\|_a / \|\mathbf{u}\|_a = \|\mathbf{u} - \mathbf{u}_h^{(k)}\|_a / \|\mathbf{u}\|_a = \left(a(\mathbf{u}, \mathbf{u}) - b^T u^{(k)} \right)^{1/2}$ error in energy between the solution of (4) and current solution at step k .

4.1 L-shape test problems

In Fig. 2, we plot the geometry of the domain Ω . In problem (4), we choose the functional $L(\mathbf{v}) = \int_{\Omega} 10\mathbf{v}d\mathbf{x}$, $\forall \mathbf{v} \in H_0^1(\Omega)$, and in the bilinear form (1), the function $\mathfrak{K}(\mathbf{x}) \in L^\infty(\Omega)$ takes

different values in each subdomain. In the first test problem within this class we have:

$$\mathfrak{K}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \Omega \setminus \{\Omega_1 \cup \Omega_2 \cup \Omega_3\}, \\ 10^{-6} & \mathbf{x} \in \Omega_1, \\ 10^{-4} & \mathbf{x} \in \Omega_2, \\ 10^{-2} & \mathbf{x} \in \Omega_3. \end{cases}$$

For the second problem, we have:

$$\mathfrak{K}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \Omega \setminus \{\Omega_1 \cup \Omega_2 \cup \Omega_3\}, \\ 10^6 & \mathbf{x} \in \Omega_1, \\ 10^4 & \mathbf{x} \in \Omega_2, \\ 10^2 & \mathbf{x} \in \Omega_3. \end{cases}$$

Using **FEMLAB**[©], we generated a mesh where the largest triangle has an area of $3.44305 \times$

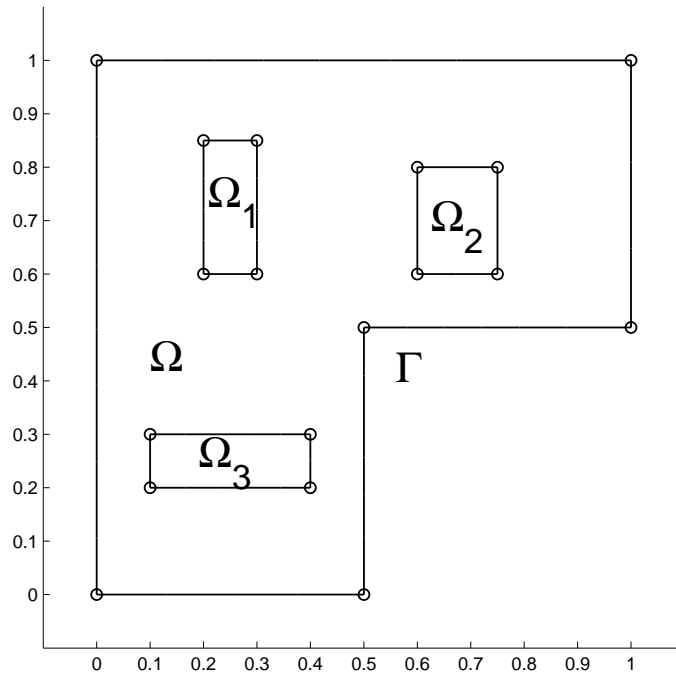


Figure 2: Geometry of the domain Ω .

10^{-5} , and we refined the mesh around the corner (minimum triangle area 7.90545×10^{-7}), therefore, the resulting linear system (6) has 59807 triangles, 30190 nodes, and 29619 degrees of freedom.

In the preconditioned conjugate gradient algorithm, we chose $\eta^2 = 3.44305 \times 10^{-5}$ which is the max area of the triangles. Moreover, we used three kinds of preconditioners: the classical Jacobi diagonal matrix, $M = \text{diag}(A)$, the incomplete Cholesky decomposition of A with zero fill-in (Greenbaum 1997, Meurant 1999a), and the incomplete Cholesky decomposition of A with drop tolerance 10^{-2} (Greenbaum 1997, Meurant 1999a). Using the incomplete Cholesky decompositions, we computed the upper triangular matrix U such that $M = U^T U$. In Table 1, we report on the values of the condition numbers $\kappa(A)$ and $\kappa(M^{-1}A)$ for both problems, and for the Jacobi and incomplete Cholesky with zero fill-in. The condition numbers of the precon-

| M | Problem 1 | Problem 2 |
|------------------|------------------|---------------------|
| I | $3.6 \cdot 10^8$ | $1.8 \cdot 10^{10}$ |
| Jacobi | $2.4 \cdot 10^4$ | $1.5 \cdot 10^9$ |
| Inc. Cholesky(0) | $7.2 \cdot 10^3$ | $4.3 \cdot 10^8$ |

Table 1: Estimates for $\kappa(M^{-1}A)$.

ditioned matrices $M^{-1}A$ for the second problem are still very high, and only the incomplete Cholesky preconditioner with drop tolerance 10^{-2} is an effective choice. Unfortunately, we could not compute the condition number of the preconditioned matrix for lack of memory.

Finally, we assume that the solution u computed by a direct solver applied to (6) is exact, and we assume that the energy norm of the solution on the finer mesh with ≈ 500000 degree of freedom is a good approximation of the energy norm of \mathbf{u} solution of the continuous problem (4) for both Problem 1 and Problem 2. Therefore, by this approximate value $\mathfrak{E}(\mathbf{u})$, we estimated the error at step k :

$$\frac{\|\delta \mathbf{u}\|_a}{\|\mathbf{u}\|_a} = \left(1 - \frac{a(\mathbf{u}_h^{(k)}, \mathbf{u}_h^{(k)})}{\mathfrak{E}(\mathbf{u})}\right)^{1/2} \approx \|\mathbf{u}_h^{(k)}(\mathbf{x}) - \mathbf{u}(\mathbf{x})\|_a / \|\mathbf{u}(\mathbf{x})\|_a.$$

In the experiments, $u^{(k)}$ is the computed value at iteration k of the conjugate gradient algorithm.

We compare the behaviour of

$$\frac{\|u - u^{(k)}\|_A}{\|u\|_A} = \frac{\|Au^{(k)} - b\|_{A^{-1}}}{\|b\|_{A^{-1}}},$$

with the corresponding estimate $\xi_k / (b^T u^{(0)} + r^{(0)T} u^{(k)})$, and the value of $\|Au^{(k)} - b\|_2 / \|b\|_2$. Moreover, we plot the values at each step k of $\|\delta \mathbf{u}\|_a / \|\mathbf{u}\|_a$.

The stopping criteria normally used are based on the values of $\|Au^{(k)} - b\|_2 / \|b\|_2$ (Arioli et al. 1992). In the practice, the conjugate gradient algorithm is stopped when $\|Au^{(k)} - b\|_2 / \|b\|_2 \leq \sqrt{\varepsilon}$.

4.1.1 Problem 1

In Fig. 3 and Fig. 4, respectively for the Jacobi and the incomplete Cholesky decomposition preconditioners and for $d = 5$, we present the history of convergence for Problem 1. During the

initial iterations of the conjugate gradient algorithm, the ratio between $\|u^{(k)}\|_A$ and $\|u\|_A$ is relatively large, as can be seen from Fig. 5 relatively to the Jacobi preconditioner. Nevertheless, the ratio value quickly stabilises itself close to 1. We obtained similar plots for the incomplete Cholesky preconditioner.

4.1.2 Problem 2

Problem 2 is harder to solve. Both Jacobi and Incomplete Cholesky without fill-in failed for small values of d . In Fig. 6, we plot the estimates relative to several values of d for the Jacobi preconditioner. Only when $d \geq 90$, the oscillations were smaller than η and, then, the algorithm stopped with an accurate solution. In Fig. 7 and Fig. 8, we present the convergence history for the cases relative to the incomplete Cholesky with drop tolerance 10^{-2} and $d = 10$ and $d = 20$ respectively. In these cases, the good preconditioner allows to choose a small value for d . Nonetheless, the convergence is not particularly fast and we can see in Fig. 9 that the ratio between the lower bound (19) and $\|u\|_A^2$ stagnates. Finally, in Fig. 10, we forced the large value of $d = 160$ when using the incomplete Cholesky preconditioner with drop tolerance 10^{-2} . We point out that in this case the $\|Au^{(k)} - b\|_2/\|b\|_2$ does not go under the value $\sqrt{\epsilon}$. Therefore, in this case, the criterion based on the Euclidean norm of the residual gives a misleading information about the iterative process.

4.2 Three dimensional test problems

The second test problems class has been built on the unitary cube choosing $\mathfrak{K}(\mathbf{x}) = 1$ in problem (4). The function f has been computed analytically from the exact solution:

$$\mathbf{u}(\mathbf{x}) = -(x^2 - x)(y^2 - y)(z^2 - z) + \frac{1000}{(5\pi)^2} \sin(5\pi x) \sin(5\pi y) \sin(5\pi z).$$

For this Problem 3, we generated by **FEMLAB**[©] a mesh of 190656 Tetrahedrons with 36365 nodes and 32226 degrees of freedom. The volume of the largest element is $\mathfrak{a}_{max} = 2.72601 \times 10^{-5}$ and the volume of the smallest element is 5.69×10^{-7} . We have chosen $\eta^2 = (\mathfrak{a}_{max})^{2/3} = 9 \times 10^{-4}$. In Fig. 11 and Fig. 12, respectively for Jacobi and the incomplete Cholesky with zero fill-in, we present the convergence history for $d = 5$. Even if the solution is highly oscillatory, the convergence is very fast and the final solution is close to the exact solution of (6) computed by a direct solver. The final error $\|u - u^{(k)}\|_2/\|u\|_2$ is 1.4×10^{-2} for the Jacobi preconditioner and 7×10^{-3} for the incomplete Cholesky preconditioner.

4.3 Error upper bounds and practical choices of d

Up to now, we have mainly focused on the lower bound estimate provided by the Hestenes and Stiefel algorithm. Nonetheless, there are two algorithms that provide upper bounds of the energy norm of the error and that deserve further discussion.

The upper bound based on the Gauss-Radau integration formula introduced by Golub and Meurant (1997) and utilised in the preconditioned conjugate gradient algorithm by Meurant (1999b), requires the knowledge of a lower bound of the smallest eigenvalue λ_{min} of the preconditioned matrix $M^{-1}A$. In our experiments this was a serious drawback: we were able to compute an estimate of λ_{min} only for the Jacobi preconditioner cases. Moreover, the cost for computing of the estimate has been quite high. We point out that loose approximations produce

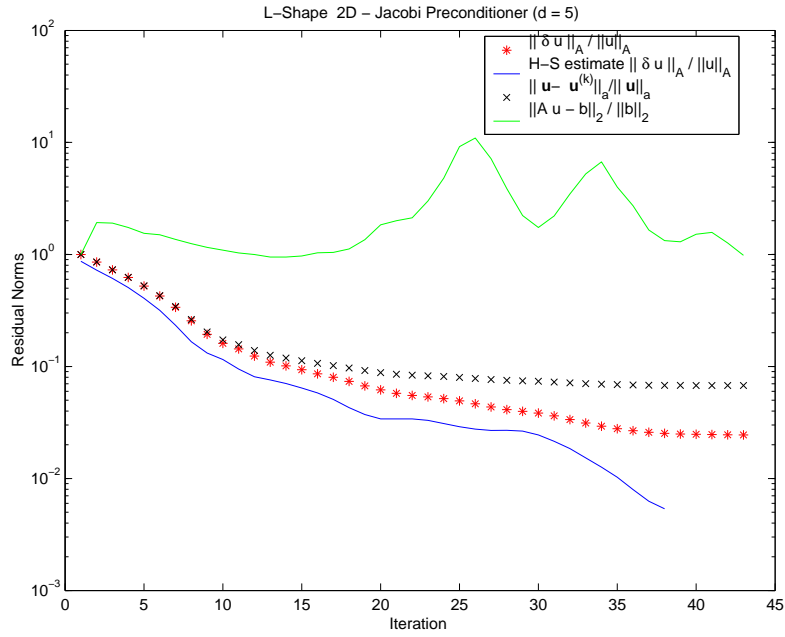


Figure 3: Behaviour of the norms of the residual for the Jacobi preconditioner in Problem 1.

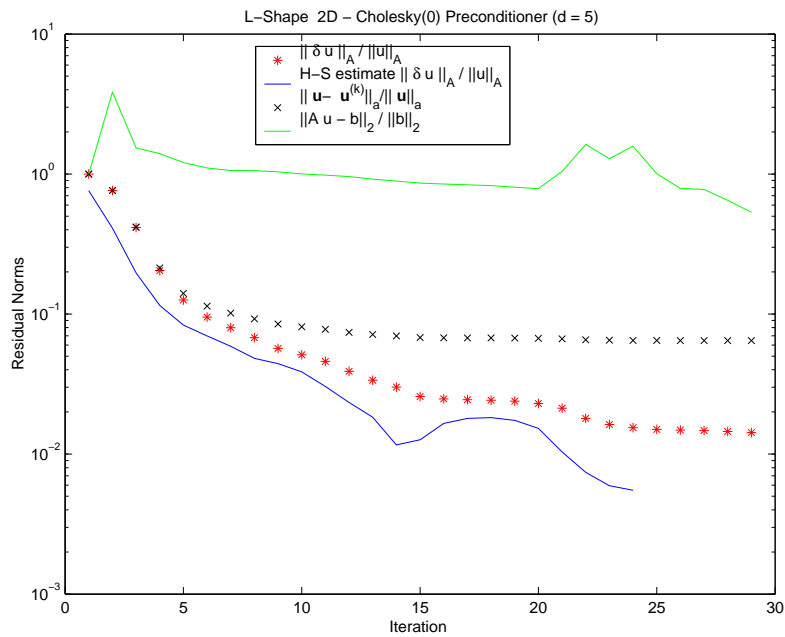


Figure 4: Behaviour of the norms of the residual for the incomplete Cholesky preconditioner in Problem 1.

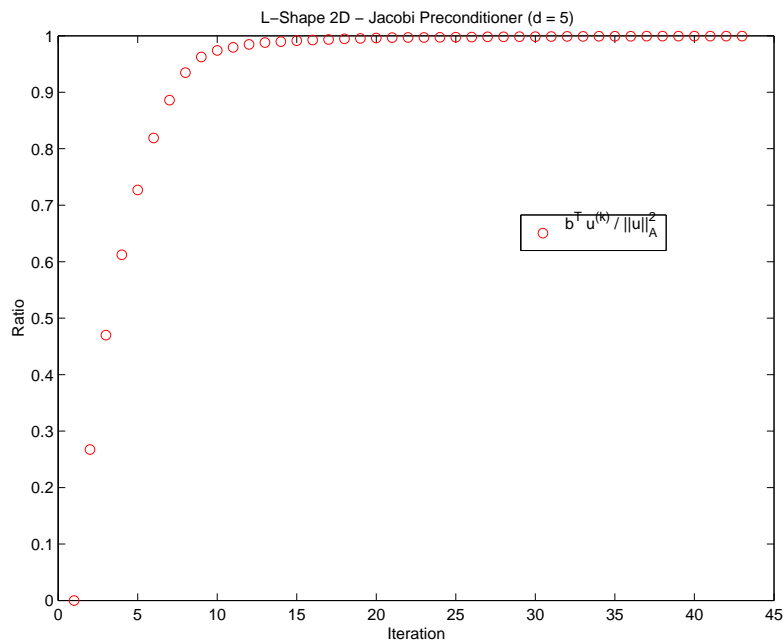


Figure 5: Ratio $b^T u^{(k)} / ||u||_A^2$ for the Jacobi preconditioner in Problem 1.

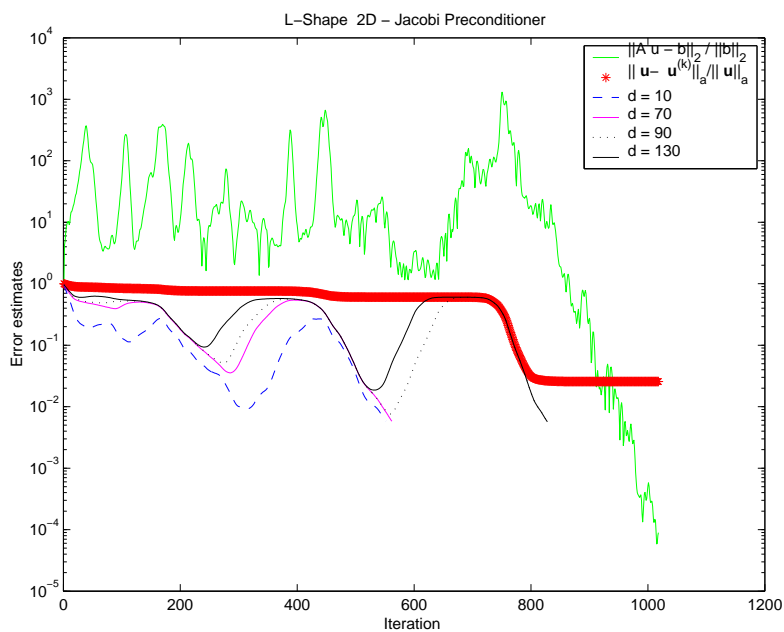


Figure 6: Comparison of several estimates of the energy error for $d = 10, 70, 90, 130$ in Problem 2.

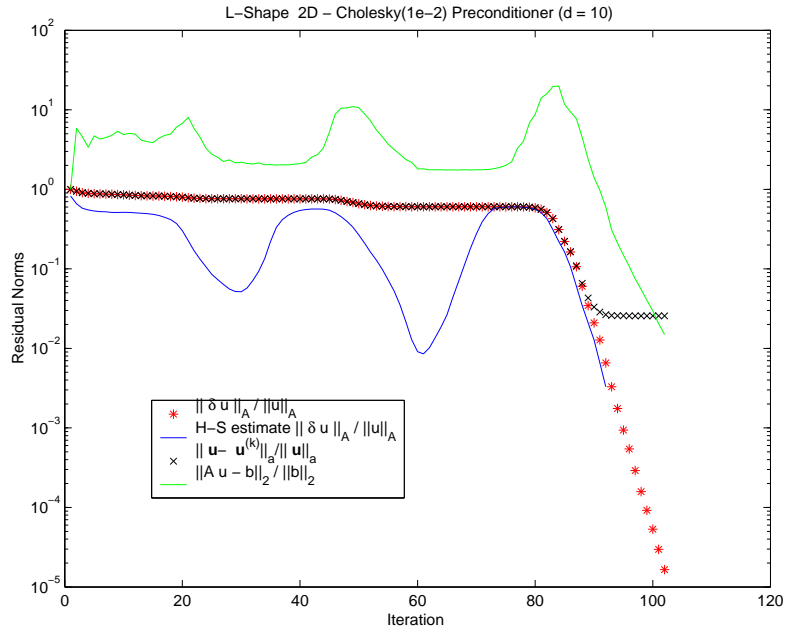


Figure 7: Behaviour of the norms of the residual for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 10$ in Problem 2.

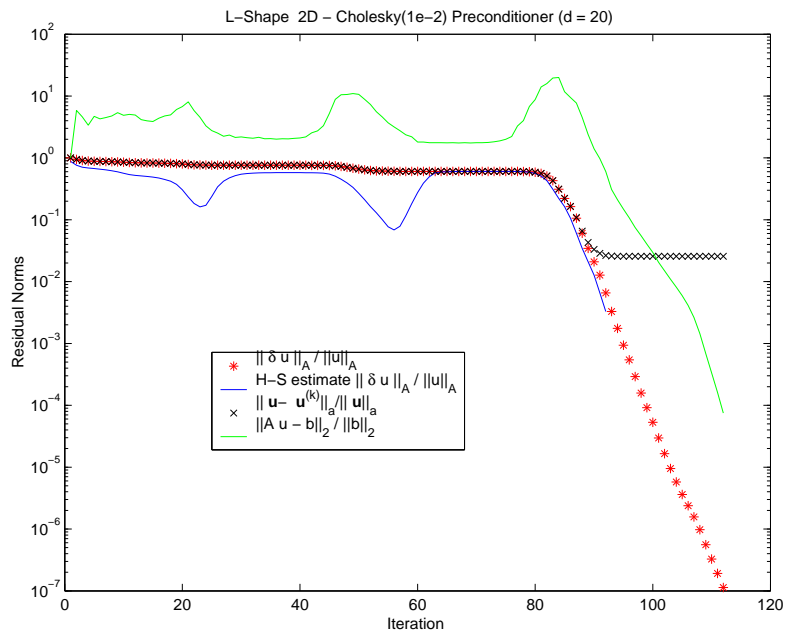


Figure 8: Behaviour of the norms of the residual for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 20$ in Problem 2.

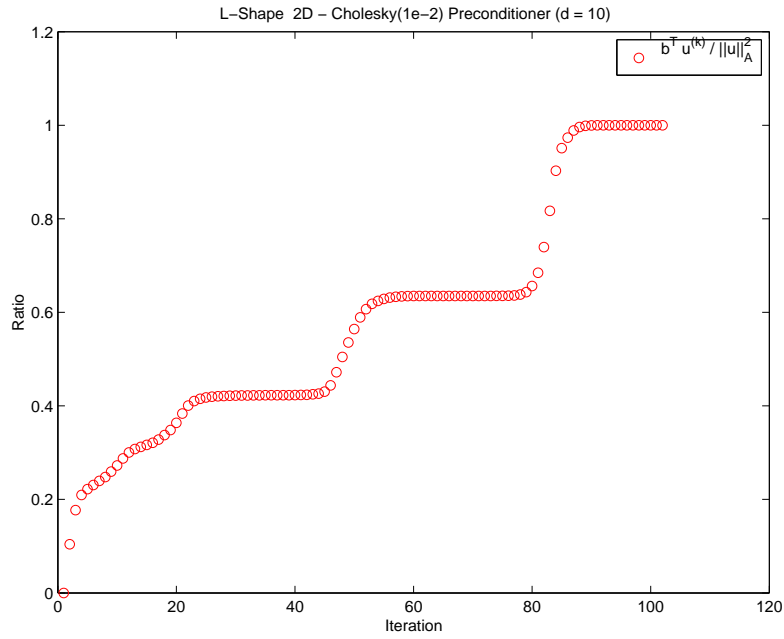


Figure 9: Ratio $b^T u^{(k)} / ||u||_A^2$ for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 10$ in Problem 2.

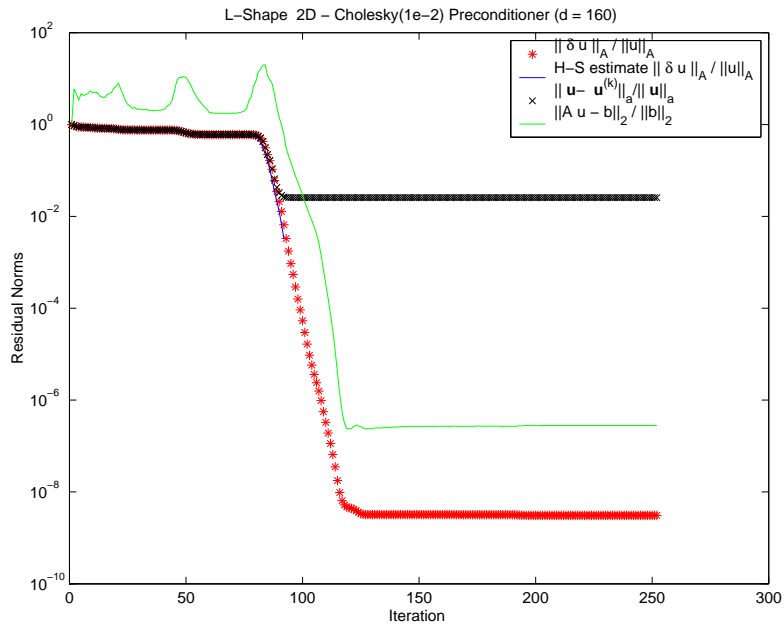


Figure 10: Behaviour of the norms of the residual for the incomplete Cholesky preconditioner with drop tolerance 10^{-2} and $d = 160$ in Problem 2.

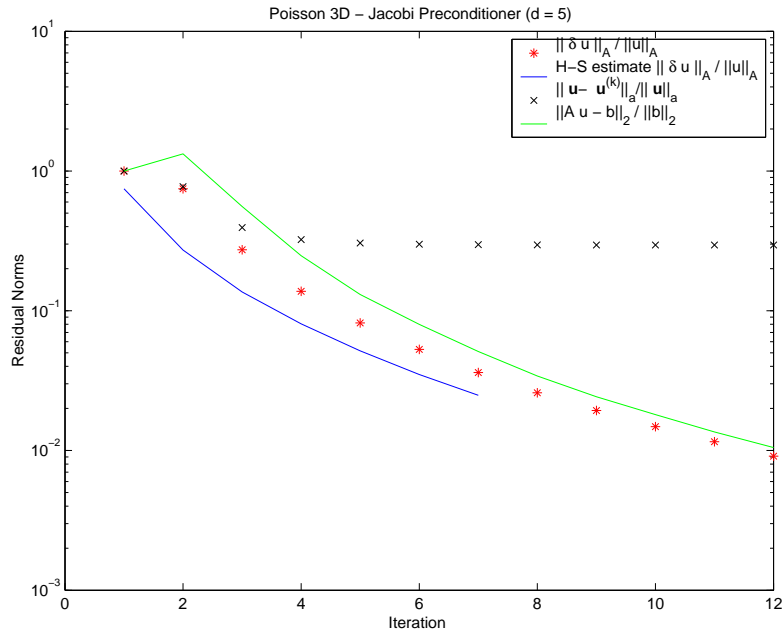


Figure 11: Behaviour of the norms of the residual for the Jacobi preconditioner with $d = 5$ in Problem 3.

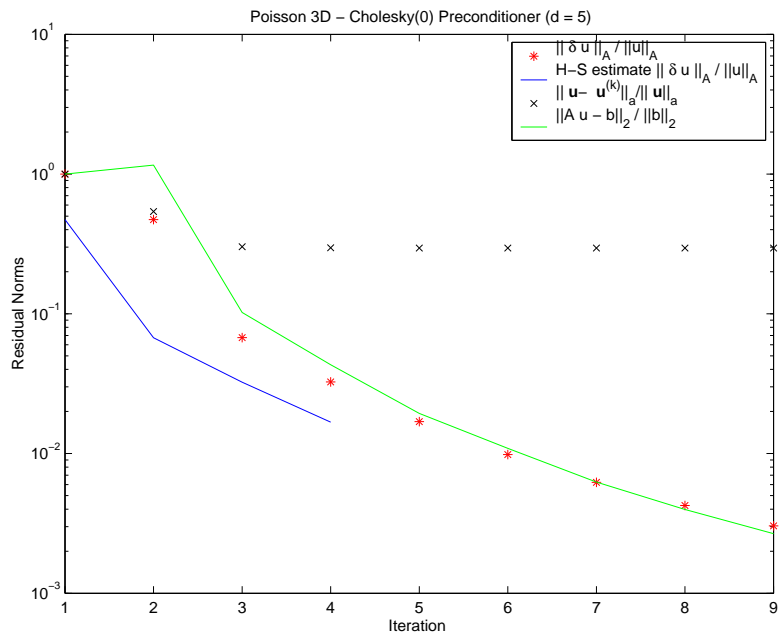


Figure 12: Behaviour of the norms of the residual for the incomplete Cholesky preconditioner with zero fill-in and $d = 5$ in Problem 3.

very conservative upper bounds that can be of some use only when the convergence has been already obtained. Usually, we noticed in our experiments that a stopping criterion based on the Gauss-Radau approximation would have request a number of steps significantly greater than the stopping criterion based on the Hestenes-Stiefel lower bound approximation.

An other upper bound that can be used in estimating the energy norm of the error is the one computed by the anti-Gauss algorithm (Calvetti et al. 2000, Calvetti et al. 2001). The anti-Gauss approach does not need any estimate of λ_{min} : this is a very attractive property, however, the method does not always produce a bound. The algorithm computes the upper bound by means of the difference between the (1,1) entries of the inverses of two tridiagonal matrices T_k and \check{T}_k of size k . Calvetti et al. (2000), (2001), proved that

$$\|\delta u\|_A^2 = \|b\|_{M^{-1}}^2 \left[(\check{T}_k^{-1})_{11} - (T_k^{-1})_{11} \right].$$

The tridiagonal matrix T_k is the matrix of the Lanczos method and it can be easily computed during the conjugate gradient algorithm. Let ω_i and γ_i , $i = 1, \dots, k$, be respectively the diagonal and the super(sub)-diagonal entries of T_k . Then, their values can be computed by means of the values of α_i and β_i (see Figure 1) as

$$\omega_i = \frac{1}{\alpha_{i-1}} + \frac{\beta_{i-1}}{\alpha_{i-2}}, \quad \gamma_i = \frac{\sqrt{\beta_i}}{\alpha_{i-1}}.$$

The tridiagonal matrix \check{T}_k is obtained by the tridiagonal matrix T by a correction of rank one and a symmetric scaling:

$$\check{T}_k = \mathcal{D}(T_k - \frac{\omega_k}{2} e_k e_k^T) \mathcal{D},$$

where $\mathcal{D} = \text{diag}(1, \dots, 1, \sqrt{2})$ is a diagonal matrix of size k and e_k is the k -th column of the $k \times k$ identity matrix. Because $\mathcal{D}e_1 = e_1$, by using the Sherman-Morrison formula, we can deduce the following recursive formula for the anti-Gauss estimate ξ_k of $\|\delta u_k\|_A^2$ at step k ,

$$\begin{aligned} \zeta_1 &= \mu_1 = \frac{1}{\omega_1}, \\ \zeta_k &= \frac{1}{\omega_k - \gamma_{k-1}^2 \zeta_{k-1}}, \quad \mu_k = -\gamma_{k-1} \zeta_k \mu_{k-1}, \\ \check{\xi}_k &= \|b\|_{M^{-1}}^2 \frac{\mu_k^2 \omega_k}{2 - \omega_k \zeta_k}. \end{aligned}$$

The matrix \check{T}_k is symmetric but there is not guaranty that it will be positive definite, and that the (1,1) entry of the inverse will be either positive or bigger than $(T_k^{-1})_{11}$. In our experiments this phenomenon appeared frequently making the overall estimate process very fragile. For test problem 2 when using Jacobi preconditioner, the value of $\check{\xi}_k$ was negative in 36% of the iterations. In Figure 13, we compare the results of the anti-Gauss strategy with the Gauss-Radau and the Hestenes-Stiefel ones, when the choice $d = 150$ is made, for the test problem 2 with the Jacobi preconditioner. In plotting the anti-Gauss estimate, we eliminated the points with negative values for $\check{\xi}_k$. We point out that the anti-Gauss formula would have arrested the conjugate gradient method at the wrong iteration and does not give the desired upper bound until the convergence is already reached.

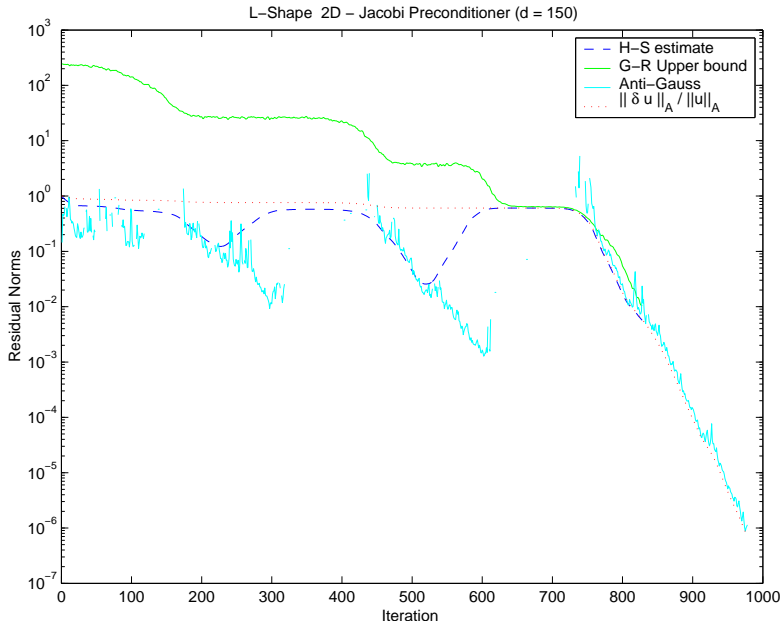


Figure 13: Anti-Gauss estimate versus Hestenes-Stiefel and the Gauss-Radau upper bound for the Jacobi preconditioner with $d = 150$ in Problem 2.

In Figure 6, it is shown how difficult it can be to identify a delay parameter d for the Hestenes-Stiefel estimate that would be reliable. The oscillatory behaviour of the estimates is very strong when despite the preconditioner the convergence is poor and the graph of energy norm of the error versus iterates stagnates. An *a priori* correct choice of d can be as difficult as the determination of a lower bound of λ_{min} for the Gauss-Radau method. Taking into account that $\|\delta u_k\|_A$ decreases monotonically, a possible heuristic for a dynamic choice of d can rely on comparison of the two consecutive values ξ_k and ξ_{k+1} . Every time $\xi_{k+1} > \tau \xi_k$, where $\tau > 1$ is a threshold parameter, we increase the value of d by a fixed amount id . In Figure 14, we report the result of a such strategy when $\tau = 1.01$ and $id = 20$. This crude strategy increased the value of d from the initial value $d = 10$ upto a final value $d = 110$ and the stopping criterion halted the conjugate gradient method at the correct iteration. We reserve to investigate more on this and other heuristics in the future.

5 Conclusions

In this paper, we try to bridge the finite element method with the linear algebra aspects of the conjugate gradient method. The aim was to extract useful information regarding the accuracy of the computed function from which we would approximate the true solution of the original partial differential equation.

We gave evidence that the proposed stopping criteria (17) and (20) are cheap and capable of stopping the conjugate gradient method when the function $\mathbf{u}_h^*(\mathbf{x})$ is a reasonable approximation of $\mathbf{u}(\mathbf{x})$.

Finally, because of the independence of the energy norm from the choice of the preconditioner,

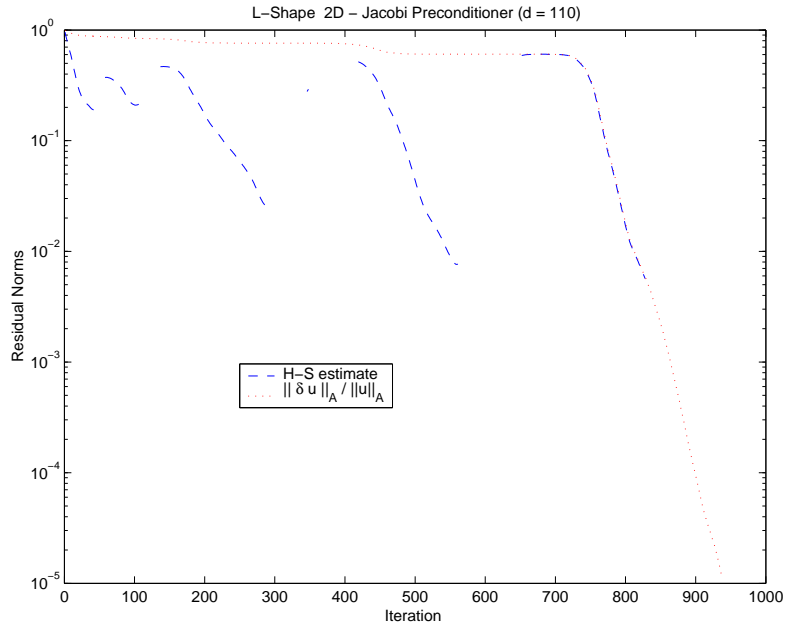


Figure 14: Adaptive strategy for the dynamical choice of d in Problem 2 for the Jacobi preconditioner (initial value of d is 10, $id = 20$ and $\tau = 1.01$).

tioner, the stopping criterion (20) would be the appropriate measurement tool for evaluating preconditioner performance in accelerating the convergence of the conjugate gradient method.

Acknowledgment

The author would like to thank Franco Brezzi and Zdenek Strakoš for the useful and valuable discussions.

References

- Arioli, M. and Baldini, L. (2001), ‘A backward error analysis of a null space algorithm in sparse quadratic programming’, *SIAM J. Matrix Anal. Appl.* **23**, 425–442.
- Arioli, M., Duff, I. S. and Ruiz, D. (1992), ‘Stopping criteria for iterative solvers’, *SIAM J. Matrix Anal. Appl.* **13**, 138–144.
- Arioli, M., Noulard, E. and Russo, A. (2001), ‘Stopping criteria for iterative methods: Applications to PDE’s’, *CALCOLO* **38**, 97–112.
- Ashby, S. F., Holst, M. J., Manteuffel, T. A. and Saylor, P. E. (2001), ‘The Role of the Inner Product in Stopping Criteria for Conjugate Gradient Iterations’, *BIT* **41**(1), 26–52.

- Axelsson, O. and Kaporin, I. (2001), ‘Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations’, *Journal of Numerical Linear Algebra with Applications* **8**, 265–286.
- Babuska, I. (1971), ‘Error-bounds for finite element method’, *Numer. Math.* **16**, 322–333.
- Babuska, I. (1972), ‘A finite element scheme for domains with corners’, *Numer. Math.* **20**, 1–21.
- Babuska, I., Strouboulis, T., Upadhyay, C. S. and Gangaraj, S. K. (2001), ‘Computer-based proof of the existence of superconvergence points in the finite element method; superconvergence of the derivatives in finite element solutions of Laplace’s, Poisson’s, and the elasticity equations’, *Numer. Methods Partial Diff Equations* **21**, 643–665.
- Calvetti, D., Morigi, S., Reichel, L. and Sgallari, F. (2000), ‘Computable error bounds and estimates for the conjugate gradient method’, *Numerical Algorithms* **25**, 75–88.
- Calvetti, D., Morigi, S., Reichel, L. and Sgallari, F. (2001), ‘An iterative method with error estimators’, *J. Comp. Appl. Math.* **127**, 93–119.
- Ciarlet, P. (1978), *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, The Netherlands.
- Dupont, T. and Scott, R. (1980), ‘Polynomial approximation of function in sobolev spaces’, *Math. Comp.* **34**, 441–463.
- Golub, G. and Meurant, G. (1997), ‘Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods’, *BIT* **37**, 687–705.
- Golub, G. and Strakoš, Z. (1994), ‘Estimates in quadratic formulas’, *Numerical Algorithms* **8**, 241–268.
- Greenbaum, A. (1997), *Iterative Methods for Solving Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- Grisvard, P. (1992), *Singularity in Boundary Value problems*, Masson & Springer-Verlag, Paris, France.
- Hestenes, M. and Stiefel, E. (1952), ‘Methods of conjugate gradients for solving linear systems’, *J. Res. Nat. Bur. Standards* **49**, 409–436.
- Higham, N. J. (1996), *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- Meurant, G. (1997), ‘The computation of bounds for the norm of the error in the conjugate gradient algorithm’, *Numerical Algorithms* **16**, 77–87.
- Meurant, G. (1999a), *Computer Solution of Large Linear Systems*, Vol. 28 of *Studies in Mathematics and its Application*, Elsevier/North-Holland, Amsterdam, The Netherlands.
- Meurant, G. (1999b), ‘Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm’, *Numerical Algorithms* **22**, 353–365.

- Strakoš, Z. and Tichy, P. (2002), 'On error estimation by conjugate gradient method and why it works in finite precision computations', *Electronic Transactions on Numerical Analysis*.
- Wahlbin, L. B. (1995), *Superconvergence in Galerkin finite element methods*, Vol. 1605 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin.