Report no. 07/07

Preconditioning and convergence in the right norm

Andy Wathen
Oxford University Computing Laboratory, Numerical Analysis Group,
Wolfson Building, Parks Road, Oxford OX1 3QD, U.K.,
andy.wathen@comlab.ox.ac.uk

The convergence of numerical approximations to the solutions of differential equations is a key aspect of Numerical Analysis and Scientific Computing. Iterative solution methods for the systems of linear(ised) equations which often result are also underpinned by analyses of convergence. In the function space setting, it is widely appreciated that there are appropriate ways in which to assess convergence and it is well-known that different norms are not equivalent. In the finite dimensional linear algebra setting, however, all norms are equivalent and little attention is often payed to the norms used.

In this paper, we highlight this consideration in the context of preconditioning for minimum residual methods (MINRES and GMRES/GCR/ ORTHOMIN) and argue that even in the linear algebra setting there is a 'right' norm in which to consider convergence: stopping an iteration which is rapidly converging in an irrelevant or highly scaled norm at some tolerance level may still give a poor answer.

Oxford University Computing Laboratory
Numerical Analysis Group
Wolfson Building
Parks Road
Oxford, England OX1 3QD

May, 2007

1 Introduction

There are two important concepts of convergence in Numerical Analysis:

- convergence of an approximation u_h as $h \to 0$ to a desired function u
- convergence of a sequence of vector iterates $\{\mathbf{u}^{(k)}\}$ as $k \to \infty$ to a desired vector $\mathbf{u}$.

When solving elliptic (and other) differential equations with numerical methods both of these concepts usually come into play: one requires a fine enough mesh so that adequate accuracy can be achieved *and* a sufficiently rapidly convergent iterative solver so that not too many iterations are required to get close enough to the exact solution of the linear(ized) equations which result from the approximation scheme. More generally one requires an approximation space in which true solutions can be accurately enough represented and an iterative solver which requires little work per iteration and few iterations to achieve a good enough approximation of the exact solution vector.

Suppose that $u \in X$ is a function that we want to find - the exact solution of a differential equations for example - and a (conforming) Finite Element approximation is employed so that

$$u_h = \sum_j \mathbf{u}_j \phi_j(\mathbf{x}) \in X_h \subset X \tag{1.1}$$

is an approximating function which we will compute. The vector of coefficients $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)^T$ will be the discrete (nodal) values of the approximating function (solution) with the usual definition of (Lagrange) finite element basis $\phi_j(\mathbf{x}_i) = \delta_{i,j}$ where $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are the nodal positions. Solution of the linear(ized) equation system which derives from the relevant approximation method:

$$\mathcal{A}\mathbf{u} = \mathbf{f}$$

will then yield the coefficient vector and hence the finite element solution via (1.1). When the dimension of this system is large an iterative method would usually be employed which would yield vector iterates

$$\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)}, \ldots, \mathbf{u}_n^{(k)})^T, \quad k = 1, 2, \ldots$$

from some starting guess $\mathbf{u}^{(0)}$. The errors committed are then the approximation error $\|u - u_h\|$ and the iteration error $\|\mathbf{u} - \mathbf{u}^{(k)}\|$ each measured in an appropriate norm.

In order to compare these errors it is useful to introduce the iterate functions defined for each k by

$$u_h^{(k)} = \sum_j \mathbf{u}_j^{(k)} \phi_j(\mathbf{x}) \in X_h.$$

The actual error at iteration k will therefore be

$$\|u - u_h^{(k)}\|. \tag{1.2}$$

Now consider some possible choices of norm in (1.2). For the case $X = L_2(\Omega)$ we have for any $v_h = \sum_j \mathbf{v}_j \phi_j(\mathbf{x}) \in X_h$

$$\|v_h\|^2_{L_2(\Omega)} = \int_\Omega v_h^2 = \sum_j \mathbf{v}_j \sum_i \mathbf{v}_i \int_\Omega \phi_i \phi_j = \mathbf{v}^T Q \mathbf{v} = \|\mathbf{v}\|^2_Q \tag{1.3}$$

where Q is the Gram matrix (the mass matrix) given by

$$Q = \{q_{i,j} : i,j = 1, \ldots, n\}, \;\; q_{i,j} = \int_\Omega \phi_i \phi_j. \tag{1.4}$$

For second order differential equations, the Sobolev space

$$\mathcal{H}^1(\Omega) = \left\{ w : \Omega \to \mathbb{R} | w, \frac{\partial w}{\partial x}, \frac{\partial w}{\partial y}, \frac{\partial w}{\partial z} \in Ł_2(\Omega) \right\} = X$$

and the related energy norm $\|v\|^2_e = \|\nabla v\|^2_{L_2(\Omega)} = \int_\Omega \nabla v \cdot \nabla v$ are important and for any $v_h = \sum_j \mathbf{v}_j \phi_j(\mathbf{x}) \in X_h$ we have

$$\|\nabla v_h\|^2_{L_2(\Omega)} = \int_\Omega \nabla v_h \cdot \nabla v_h = \sum_j \mathbf{v}_j \sum_i \mathbf{v}_i \int_\Omega \nabla \phi_i \cdot \nabla \phi_j = \mathbf{v}^T A \mathbf{v} = \|\mathbf{v}\|^2_A \tag{1.5}$$

where A is the discrete Laplacian matrix (the stiffness matrix) given by

$$A = \{a_{i,j} : i,j = 1, \ldots, n\}, \;\; a_{i,j} = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j. \tag{1.6}$$

We use the notation Q and A only for the matrices defined by (1.4) and (1.6) respectively throughout this paper.

Whichever of these or other norms is used, a simple use of the triangle inequality gives

$$\|u - u_h^{(k)}\| \le \|u - u_h\| + \|u_h - u_h^{(k)}\|.$$

Thus for example for the energy norm

$$\begin{aligned} \|\nabla(u - u_h^{(k)})\|_{L_2(\Omega)} &\le \|\nabla(u - u_h)\|_{L_2(\Omega)} + \|\nabla(u_h - u_h^{(k)})\|_{L_2(\Omega)} \\ &= \|\nabla(u - u_h)\|_{L_2(\Omega)} + \|\mathbf{u} - \mathbf{u}^{(k)}\|_A \end{aligned} \tag{1.7}$$

because of (1.5) above. (If the Galerkin method is used then we have more precisely that

$$\|\nabla(u - u_h^{(k)})\|^2_{L_2(\Omega)} = \|\nabla(u - u_h)\|^2_{L_2(\Omega)} + \|\mathbf{u} - \mathbf{u}^{(k)}\|^2_A$$

because of Galerkin orthogonality: $\int_\Omega \nabla(u - u_h) \cdot \nabla v_h = 0$ for every $v_h \in X_h$). The key observation here is that the actual error $\|u - u_h^{(k)}\|$ is bounded by the sum of the approximation error $\|\nabla(u - u_h)\|_{L_2(\Omega)}$ and the iteration error $\|\mathbf{u} - \mathbf{u}^{(k)}\|_A$. More specifically, for any PDE problem for which $\| \cdot \|_e$ is the natural norm for the problem, the norm $\| \cdot \|_A$ is identified as the appropriate norm for the linear algebra. For any error estimate (a priori or a posteriori) which bounds $\|\nabla(u-u_h)\|_{L_2(\Omega)}$, a stopping criterion should be chosen so that the iteration error $\|\mathbf{u}-\mathbf{u}^{(k)}\|_A$ is comparable but smaller (typically an order of magnitude less). The same points apply to other norms: for example if $\|\cdot\|_{L_2(\Omega)}$ is the natural norm for a problem, then $\|\cdot\|_Q$ is identified as the appropriate norm for the linear algebra.

Such interconnection between approximation error and iteration error has been used in connection with iterative stopping criteria [1] and is more broadly understood (see for example [2]). In [2] chapter 2, for example, it is highlighted why the Conjugate Gradient method and the Multigrid method are ideally suited for second order self-adjoint elliptic boundary value problems because they each monotonically reduce the iteration error precisely in $\|\cdot\|_A$.

In the next section we show how preconditioning affects this balance between approximation and iteration error. In particular for widely used iterative methods which minimise the residual (MINRES [3] for symmetric indefinite systems and GMRES/GCR/ORTHOMIN[4]/[5]/[6] for nonsymmetric systems) we show that care is needed to avoid selection of preconditioners which apparently give rapid convergence, but which in fact merely distort the relevant norm so that poor solutions are achieved for all but extremely small convergence tolerences. The following section is devoted to examples which highlight the issue addressed here and our brief conclusions follow. We begin with a simple discussion of preconditioning and iterative methods.

2 Preconditioning and Krylov subspace methods

For $\mathcal{A}$ which is symmetric and positive definite, the Conjugate Gradient (CG) method for solving $\mathcal{A}\mathbf{u}=\mathbf{f}$ from a starting vector $\mathbf{u}^{(0)}$ with corresponding residual $\mathbf{r}^{(0)}=\mathbf{f}-\mathcal{A}\mathbf{u}^{(0)}$ computes for each k the iterate $\mathbf{u}^{(k)}$ in the shifted (affine) Krylov subspace

$$\mathbf{u}^{(0)}+\mathcal{K}_k(\mathcal{A},\mathbf{r}^{(0)})=\mathbf{u}^{(0)}+\operatorname{span}\{\mathbf{r}^{(0)},\mathcal{A}\mathbf{r}^{(0)},\mathcal{A}^2\mathbf{r}^{(0)},\ldots,\mathcal{A}^{k-1}\mathbf{r}^{(0)}\}$$

for which the error $\mathbf{u}-\mathbf{u}^{(k)}$ is minimal in $\|\cdot\|_{\mathcal{A}}$. With a symmetric and positive definite preconditioner, $\mathcal{M}$ the CG method solves the equivalent system

$$H^{-1}\mathcal{A}H^{-T}\mathbf{v}=H^{-1}\mathbf{f},\quad H^T\mathbf{u}=\mathbf{v}$$

where $\mathcal{M}=HH^T$. It is therefore easily seen that for this system the (preconditioned) CG method will minimise $H^T(\mathbf{u}-\mathbf{u}^{(k)})$ in $\|\cdot\|_{H^{-1}\mathcal{A}H^{-T}}$. That is

$$(\mathbf{u}-\mathbf{u}^{(k)})^THH^{-1}\mathcal{A}H^{-T}H^T(\mathbf{u}-\mathbf{u}^{(k)})=(\mathbf{u}-\mathbf{u}^{(k)})^T\mathcal{A}(\mathbf{u}-\mathbf{u}^{(k)})=\|\mathbf{u}-\mathbf{u}^{(k)}\|_A^2$$

is the quantity minimised whatever (symmetric and positive definite) preconditioner is employed. It is emphasised that this quantity is independent of the preconditioner used. Of course the preconditioner affects the Krylov subspaces so that with preconditioning the iterates lie in

$$\mathbf{u}^{(0)}+\operatorname{span}\{\mathcal{M}^{-1}\mathbf{r}^{(0)},\mathcal{M}^{-1}\mathcal{A}\mathcal{M}^{-1}\mathbf{r}^{(0)},(\mathcal{M}^{-1}\mathcal{A})^2\mathcal{M}^{-1}\mathbf{r}^{(0)},\ldots,(\mathcal{M}^{-1}A)^{k-1}\mathcal{M}^{-1}\mathbf{r}^{(0)}\}; \tag{2.1}$$

faster convergence with preconditioning will occur if $\mathbf{u}$ is better approximated from these spaces than from $\mathbf{u}^{(0)} + \mathcal{K}_k(\mathcal{A}, \mathbf{r}^{(0)})$ for each k. We emphasise that this description is purely formal – H is not needed in practice, only $\mathcal{M}$.

For minimum residual methods however, the situation is different. Consider first MINRES which is used when $\mathcal{A}$ is symmetric but not necessarily positive definite. Here a positive definite preconditioner is required in order to preserve symmetry in the preconditioned system. Thus, as described above for CG, one formally considers solving the symmetric system $H^{-1}\mathcal{A}H^{-T}\mathbf{v} = H^{-1}\mathbf{f}$ where $H^T\mathbf{u} = \mathbf{v}$. Here instead though it is the Euclidean norm of the residual $\|H^{-1}(\mathbf{f} - \mathcal{A}H^{-T}\mathbf{v})\|_2 = \|H^{-1}(\mathbf{f} - \mathcal{A}\mathbf{u})\|_2$ (perhaps $\|H^{-1}(\mathbf{f} - \mathcal{A}\mathbf{u})\|_I$ is a better notation here) which is minimised for $\mathbf{u}^{(k)}$ in the shifted Krylov subspace (2.1). Thus for the real residuals $\mathbf{r}^{(k)} = \mathbf{f} - \mathcal{A}\mathbf{u}^{(k)}$ of the original system it is

$$\|\mathbf{r}^{(k)}\|_{H^{-T}H^{-1}} = \|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}} = \|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathcal{A}\mathcal{M}^{-1}\mathcal{A}} \tag{2.2}$$

which is minimised in the preconditioned case. Thus if it were possible that $\mathcal{M} \simeq \mathcal{A}$ so that $\mathcal{M}$ would then be a very good preconditioner for $\mathcal{A}$ we would have $\mathcal{A}\mathcal{M}^{-1}\mathcal{A} \simeq \mathcal{A}$ and convergence of preconditioned MINRES would occur in a norm similar to $\|\cdot\|_{\mathcal{A}}$. There are two dificulties here: for indefinite $\mathcal{A}$, $\|\cdot\|_{\mathcal{A}}$ does not define a norm and even for indefinite matrices $\mathcal{A}$, $\mathcal{M}$ has to be positive definite for use with MINRES. The definition of an appropriate norm for an indefinite problem has therefore to be identified in this case. For the important class of indefinite matrices of saddle-point form, however, there does exist a natural choice of norm and a class of positive definite preconditioners which give convergence in this right norm – see example 2 below. For weaker preconditioners such as $\mathcal{M} = diag(\mathcal{A})$, $\|\cdot\|_{\mathcal{A}\mathcal{M}^{-1}\mathcal{A}}$ becomes almost the $\mathcal{A}^2$ norm which may or may not be so desirable. With MINRES, as with CG, only $\mathcal{M}$ is required, not H in practical computation (see eg. [2]).

Similar consideration applies even if $\mathcal{A}$ is nonsymmetric in the unusual situation that a symmetric and positive definite preconditioner is employed in a centered way; in this case explicit knowledge of H - for example an incomplete Cholesky factor of the symmetric part - would be needed. Usually, however, when $\mathcal{A}$ is nonsymmetric, a preconditioner (whether it be nonsymmetric or symmetric) is used on the right giving the right preconditioned system

$$\mathcal{A}\mathcal{M}^{-1}\mathbf{v} = \mathbf{f}, \quad \mathcal{M}\mathbf{u} = \mathbf{v}$$

or on the left giving the left preconditioned system

$$\mathcal{M}^{-1}A\mathbf{u} = \mathcal{M}^{-1}\mathbf{f}.$$

Another possibility is to have $\mathcal{M} = H_L H_R$ and solve

$$H_L^{-1}\mathcal{A}H_R^{-1}\mathbf{v} = H_L^{-1}\mathbf{f}, \quad H_R\mathbf{u} = \mathbf{v}$$

in which case explicit knowledge of the factors H_L, H_R (or rather routines for the effective application of H_L^{-1}, H_R^{-1} to known vectors) is required. Since this case covers the more common right and left preconditioning cases with $H_L = I$ or $H_R = I$ respectively we continue only with it. A minimum residual method such as GMRES, GCR or ORTHOMIN will compute iterates $\mathbf{u}^{(k)}$ in (2.1) which minimise

$$\|\mathbf{r}^{(k)}\|_{H_L^{-T}H_L^{-1}}.$$

It is often argued that right preconditioning is therefore to be prefered because it is then the Euclidean norm of the residual of the original system which is minimised. We would argue that this may be a reasonable choice in the situation where the origin of the linear system gives no guidance but that the real question is whether

$$\|\mathbf{r}^{(k)}\|_2 = \|\mathcal{A}(\mathbf{u} - \mathbf{u}^{(k)})\|_2 = \|\mathbf{u} - \mathbf{u}^{(k)}\|_{\mathcal{A}^T\mathcal{A}}$$

is the appropriate norm in which to measure the error vector; in the case of some PDE problems we contest that it is not - see the examples in the next section!

We comment that there are few, but important, examples where nonsymmetric preconditioners $\mathcal{M}$ can yield symmetric and positive definite preconditioned systems $\mathcal{M}^{-1}\mathcal{A}$ or $\mathcal{A}\mathcal{M}^{-1}$ when $\mathcal{A}$ itself is not symmetric and positive definite ([7], [8]). Further, if $\mathcal{A}$ is symmetric and $\mathcal{M}$ is symmetric and positive definite, $\mathcal{M}^{-1}\mathcal{A}$ is self-adjoint (symmetric) in the non-standard inner product defined by $\langle\cdot,\cdot\rangle_{\mathcal{M}}$ where $\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{M}}^2 = \mathbf{x}^T\mathcal{M}y$ and if $\mathcal{M}$ is symmetric and $\mathcal{A}$ is symmetric and positive definite then $\mathcal{A}\mathcal{M}^{-1}$ is self-adjoint in the non-standard inner product $\langle\cdot,\cdot\rangle_{\mathcal{A}^{-1}}$ defined analogously.

3 Examples

In this section some simple expository examples are presented.

3.1 Example 1.

The first example is a well-conditioned symmetric and positive definite matrix, the nice matrix $\mathcal{A} \in \mathbb{R}^{21\times 21}$ obtainable in `matlab` via the command `gallery('wathen',2,2)`.

Table 1:

	CG	MINRES	CG	MINRES
k	$\|\mathbf{u}-\mathbf{u}^{(k)}\|_Q$	$\|\mathbf{u}-\mathbf{u}^{(k)}\|_Q$	$\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$	$\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$
1	19.835	19.835	25.159	25.159
2	13.671	13.871	11.123	10.173
3	7.401	8.963	8.853	6.678
4	4.512	5.304	5.009	4.007
5	2.982	3.176	2.306	1.998
6	1.021	1.118	0.963	0.876

It is precisely a finite element mass matrix for an 8-node serendipity element computed here for a 2×2 grid on the unit square and thus to be consistent with our nomenclature above we denote it by $\mathcal{A} = Q$. The diagonal of the matrix provides a preconditioner, $\mathcal{M}$, for which the precise analytic bounds $\frac{1}{4} \leq \lambda \leq \frac{9}{2}$ on the eigenvalues λ of the preconditioned matrix have been established [9]. Most iterative methods will converge rapidly for this matrix with this preconditioner; we employ it here simply to show how the convergence of CG and MINRES compare in terms of $\|\mathbf{u} - \mathbf{u}^{(k)}\|_Q$ and $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$. The exact solution is chosen to be the vector of ones and $\mathbf{f}$ is constructed accordingly. The smallest diagonal entry of $\mathcal{M}$ is approximately 0.508 and the largest diagonal entry is approximately 50.579. The initial vector is random. The results for the first six iterations are shown in Table 1.

A natural context in which this algebraic problem arises is in the best L_2 approximation (L_2 projection) of a function $u \in X = L_2(\Omega)$ from the 21-dimensional subspace X_h of piecewise polynomials defined by the 4 square 8-node seredipity finite elements of side length $\frac{1}{2}$ on $\Omega = [0,1] \times [0,1]$. That is piecewise bi-quadratic bivariate polynomials without the quartic term x^2y^2. In this context the overall error is

$$\begin{aligned} \|u - u_h^{(k)}\|^2_{L_2(\Omega)} &= \|u - u_h + u_h - u_h^{(k)}\|^2_{L_2(\Omega)} \\ &= \|u - u_h\|^2_{L_2(\Omega)} + \|u_h - u_h^{(k)}\|^2_{L_2(\Omega)} \\ &= \|u - u_h\|^2_{L_2(\Omega)} + \|\mathbf{u} - \mathbf{u}^{(k)}\|^2_Q \end{aligned} \tag{3.1}$$

where the equality in the second line follows because the best approximation u_h is defined by the orthogonality of $u - u_h$ to (every function in) X_h and the final line because of (1.3). Similarly to above it is therefore the CG method which is reducing exactly the right quantity $\|\mathbf{u} - \mathbf{u}^{(k)}\|_Q$ in this case, though the differences in the different norm quantities are not large in this nice example. Given an estimate of the approximation error, an appropriate stopping criteron for the iteration is indicated by (3.1).

3.2 Example 2.

Problems with constraints lead to saddle-point systems – an important class of symmetric (and nonsymmetric) indefinite matrices. The general structure is

$$\begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \tag{3.2}$$

where F may either be symmetric (giving the classical saddle-point system) or nonsymmetric (giving a generalized saddle-point system). For a comprehensive survey on solution methods for saddle-point systems see [10].

One of the more important PDE examples is the Stokes problem:

$$\begin{aligned} \nu\nabla^2 u + \nabla p &= f \\ \nabla \cdot u &= 0, \end{aligned}$$

see for example [2] chapters 5 and 6. This problem arises as the most common model for the slow flow of an incompressible fluid. This problem is self-adjoint and most discretisations including conforming mixed finite elements in any domain $\Omega \subset \mathbb{R}^d$ lead to a symmetric matrix block F which is in the usual case a $d \times d$ block diagonal matrix with diagonal blocks which are just discrete Lapacians given by (1.6). We thus write $F = \mathbf{A}$ in this situation to denote a $d \times d$ block diagonal matrix with diagonal blocks each equal to A. The full incompressible Navier-Stokes equations are used in situations were fluid inertia is important and then the diagonal blocks in F typically become matrices representing discrete advection-diffusion operators and are hence non-symmetric – again see [2] for a complete description. Whichever of these problems is being considered, the weak form of these problems are defined for velocity vectors $u \in \mathcal{H}^1(\Omega)^d$ and scalar pressures $p \in \L_2(\Omega)$ and it is in these spaces that natural error estimates of the form

$$\begin{aligned} \|\nabla(u-u_h)\|_{L_2(\Omega)} &+ \|p-p_h\|_{L_2(\Omega)} \\ &\le C(\inf_{v_h \in X_h} \|\nabla(u-v_h)\|_{L_2(\Omega)} + \inf_{q_h \in M_h} \|p-q_h\|_{L_2(\Omega)}) \\ &\le C\,h^2\,\left(\|D^3 u\|_{L_2(\Omega)} + \|D^2 p\|_{L_2(\Omega)}\right), \end{aligned}$$

are found. Thus directly from (1.5) and (1.3), the natural norm for the linear algebra is

$$\|\mathbf{u}-\mathbf{u}^{(k)}\|_{\mathbf{A}} + \|\mathbf{p}-\mathbf{p}^{(k)}\|_Q;$$

if we write the Stokes saddle-point system as $\mathcal{A}\mathbf{x} = \mathbf{b}$, that is the right norm for iterative convergence is $\|\cdot\|_E$ where

$$E = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & Q \end{bmatrix}.$$

Thus when iterates $\mathbf{x}^{(k)}$ are computed with some iterative method, as shown in [2] page 291, it is desirable to get convergence of

$$\|\mathbf{x}-\mathbf{x}^{(k)}\|_E^2 = \|\mathbf{r}^{(k)}\|_{\mathcal{A}^{-1}E\mathcal{A}^{-1}} = \|\mathbf{r}^{(k)}\|_{(\mathcal{A}E^{-1}\mathcal{A})^{-1}}$$

where $\mathbf{r}^{(k)}$ are the corresponding residuals.

Comparing this with (2.2), the ‘ideal’ preconditioner

$$\mathcal{M} = \mathcal{A}E^{-1}\mathcal{A} = \begin{bmatrix} \mathbf{A} + B^T Q^{-1} B & B^T \\ B & B\mathbf{A}^{-1}B^T \end{bmatrix}$$

for the Stokes problem is identified. Of course this is not a practical choice since solution of linear systems involving $\mathcal{M}$ is at least as difficult as solving systems with $\mathcal{A}$! Nevertheless, it leads to practical block diagonal preconditioners which give norms equivalent up to small constants which are independent of problem parameters – see [2], theorem 6.9.

We comment that similar considerations arise whenever mixed finite element approximations are used (see[11]).

3.3 Example 3.

Our final example also involves the saddle point system (3.2). In a number of situations, notably when F has a significant null space (see for example [12]), it has been suggested to employ an Augmented Lagrangian formulation; that is to solve

$$\begin{bmatrix} F + B^TWB & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} + B^TW\mathbf{g} \\ \mathbf{g} \end{bmatrix} \tag{3.3}$$

for some matrix W ([13],[14]) or to use a preconditioner based on an Augmented Lagrangian [12], [15]. In this context a preconditioner which is frequently suggested is the block triangular matrix

$$\begin{bmatrix} F + \gamma B^TWB & B^T \\ 0 & -\gamma^{-1}W^{-1} \end{bmatrix} \tag{3.4}$$

or the related block diagonal matrix

$$\begin{bmatrix} F + \gamma B^TWB & 0 \\ 0 & \pm\gamma^{-1}W^{-1} \end{bmatrix} \tag{3.5}$$

where W is taken to be a simple matrix in either case; for example $W = I$ which we shall consider here. This is a common choice even though it can be a poor choice depending on the application: for example for the Stokes problem, Benzi and Olshanskii [14] argue why the inverse of the mass matrix Q or the inverse of the diagonal of Q (using the results of [9]) give the right scaling and are thus the appropriate choice for W. Indeed this choice gives slightly better numerical results than those we present below for $W = I$, but they are not so different and the choice of γ remains key.

For reasons of practicality, the leading (1,1) block in (3.4),(3.5) is often approximated, a multigrid cycle being the choice in [14] for example, however since such considerations are motivated by the exact form of (3.4) or (3.5), we consider only this exact form here and use a direct solver for the blocks in the preconditioner. It is a considerable challenge in practice to find a robust iterative solver for the (1,1) block which works effectively over a range of values of γ.

To fix ideas we address the particular situation in which the saddle point matrix and preconditioner are given by

$$\mathcal{A} = \begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix}, \quad \mathcal{M} = \begin{bmatrix} F + \gamma B^TB & 0 \\ 0 & \gamma^{-1}I \end{bmatrix}. \tag{3.6}$$

A number of variants of this precise pairing have been proposed, but this is sufficient for our purpose and similar structures exist with triangular preconditioners and/or when augmentation of $\mathcal{A}$ is also used (see the references above). Note that invertibility of $\mathcal{A}$ requires that B be of full rank and additionally a sufficient condition for invertibility is that $F + F^T$ is positive semi-definite and $ker(F + F^T) \cap ker(B) = \{0\}$ and we shall assume this.

Firstly we consider the case when F (and thus $\mathcal{A}$ and $\mathcal{M}$ also) are symmetric. It is a simple task to calculate the eigenvalues of $\mathcal{M}^{-1}\mathcal{A}$ on which MINRES convergence depends: see [12], theorem 2.2. There turn out to be eigenvalues of 1, of -1 and the remaining eigenvalues are λ satisfying

$$\lambda = -\frac{\gamma\mu}{\gamma\mu + 1}$$

where μ are the non-zero eigenvalues of $\mu F\mathbf{v} = B^T B\mathbf{v}$. Thus for large values of γ these eigenvalues cluster at -1 and rapid MINRES convergence should result: indeed only two MINRES iterations should be required to compute the exact solution as $\gamma \to \infty$. Unfortunately, the sensitivity of the iteration to values of the augmentation parameter γ is somewhat hidden: it can significantly distort the norm in which MINRES converges.

Note that $(F + \gamma B^T B)^{-1}$ is a matrix with entries which do not grow with increasing values of γ. It is evident therefore from (2.2) that the residuals are minimised in the matrix norm defined by the matrix

$$\mathcal{M}^{-1} = \begin{bmatrix} (F + \gamma B^T B)^{-1} & 0 \\ 0 & \gamma I \end{bmatrix}$$

which heavily emphasises satisfying the second equation $B\mathbf{u} = \mathbf{g}$ in the saddle point system for large γ. Similar distortion can be seen also in terms of the error which is minimised–see again (2.2)–in the matrix norm defined by

$$\mathcal{A}\mathcal{M}^{-1}\mathcal{A} = \begin{bmatrix} F(F + \gamma B^T B)^{-1}F + \gamma B^T B & F(F + \gamma B^T B)^{-1}B^T \\ B(F + \gamma B^T B)^{-1}F & B(F + \gamma B^T B)^{-1}B^T \end{bmatrix}.$$

Since it is only the term $\gamma B^T B$ in the (1,1) block which grows with increasing values of γ, it can be expected that poorer accuracy will be observed in the second component, $\mathbf{p}$, with the above preconditioner.

In order to illustrate, we present the results for just one set of computations using a saddle point system coming from mixed finite element approximation of the Stokes problem generated by the IFISS software ([16]). The specific problem is approximation of flow over a backwards facing step using the Q2-Q1 (Taylor-Hood) mixed finite element pair on a 16×48 grid. MINRES is used as the iterative method with the coefficient matrix $\mathcal{A}$ and preconditioner $\mathcal{M}$ as above. We present results for the first 8 iterates for the values $\gamma = 10^3$, $\gamma = 1$ and $\gamma = 10^{-3}$ in the tables below.
Tabulated are the residual of the combined vector $[\mathbf{u}, \mathbf{p}]^T$ in $\|\cdot\|_{\mathcal{M}^{-1}}$ which is the minimised quantity, the errors of the velocity ($\mathbf{u}$) components in $\|\cdot\|_A$ and the pressure ($\mathbf{p}$) components in $\|\cdot\|_Q$ which are the natural norms for the problem as well as the Euclidean norm of $B\mathbf{u}$.

It can be seen that convergence is quickest for the largest value $\gamma = 10^3$, however this convergence is actually measured. The quantity $\|B\mathbf{u}\|$ reduces very rapidly (in the context of this problem $B\mathbf{u} = 0$ is the discrete statement of conservation of mass, thus the quantity $\|B\mathbf{u}\|$ is a measure of by how much the mass is not conserved), however, the pressure error is two orders of magnitude greater at all iterations and one order of magnitude greater than the velocity error. For $\gamma = 1$ the residual norm and $\|B\mathbf{u}\|$ reduce quickly even though the velocity and pressure errors remain reasonably large.

Table 2: Stokes Augmented Lagrangian Preconditioning: $\gamma = 10^3$

k	$\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$	$\|\mathbf{u}-\mathbf{u}^{(k)}\|_A$	$\|\mathbf{p}-\mathbf{p}^{(k)}\|_Q$	$\|B\mathbf{u}^{(k)}\|_I$
1	102.8321	72.6571	15.3214	2.3662
2	87.1623	36.9047	196.2536	2.2998
3	3.2999	1.5564	18.8032	0.0787
4	0.6591	0.4421	5.3614	0.0164
5	0.1172	0.0604	0.5790	0.0028
6	0.0168	0.0104	0.0614	0.0004
7	0.0035	0.0020	0.0158	0.0001
8	0.0004	0.0002	0.0014	0.0000

Table 3: Stokes Augmented Lagrangian Preconditioning: $\gamma = 1$

k	$\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$	$\|\mathbf{u}-\mathbf{u}^{(k)}\|_A$	$\|\mathbf{p}-\mathbf{p}^{(k)}\|_Q$	$\|B\mathbf{u}^{(k)}\|_I$
1	72.6282	72.6571	15.3214	2.3662
2	0.3941	3.6037	15.3369	0.3885
3	0.1862	2.8639	14.4748	0.1847
4	0.1294	2.5406	13.6403	0.1288
5	0.1024	2.2466	12.5859	0.1021
6	0.0784	1.8421	10.8597	0.0782
7	0.0603	1.4981	9.1325	0.0601
8	0.0481	1.2501	7.6958	0.0480

Table 4: Stokes Augmented Lagrangian Preconditioning: $\gamma = 10^{-3}$

k	$\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$	$\|\mathbf{u}-\mathbf{u}^{(k)}\|_A$	$\|\mathbf{p}-\mathbf{p}^{(k)}\|_Q$	$\|B\mathbf{u}^{(k)}\|_I$
1	72.6288	72.6571	15.3214	2.3662
2	0.0126	3.6369	15.3214	0.3986
3	0.0061	2.8878	14.4901	0.1919
4	0.0042	2.5648	13.6945	0.1323
5	0.0042	2.5647	13.6943	0.1323
6	0.0033	2.2773	12.6883	0.1043
7	0.0026	1.8920	11.0749	0.0810
8	0.0020	1.5415	9.3567	0.0625

For $\gamma = 10^{-3}$, the situation is even more distorted with rapid residual reduction even though velocity errors remain $\mathcal{O}(1)$ and pressure errors remain $\mathcal{O}(10)$.

Disparity in the reduction of the various tabulated quantities could lead to larger errors than expected in all cases if inappropriate stopping criteria are used.

In the case above when F is nonsymmetric, GMRES with right preconditioning at least gives residual reduction and correspondingly error reduction in norms which do not depend on γ, whereas left preconditioning would give residual resuction in $\| \cdot \|_{\mathcal{M}^{-2}}$ and thus correspondingly error reduction in the matrix norm defined by the matrix

$$\left[\begin{array}{cc} F(F+\gamma B^TB)^{-2}F+\gamma^2B^TB & F(F+\gamma B^TB)^{-2}B^T \\ B(F+\gamma B^TB)^{-2}F & B(F+\gamma B^TB)^{-2}B^T \end{array} \right]$$

for which clearly similar but even more extreme issues arise than for the centred preconditioner considered for MINRES above. If augmentation was used also in $\mathcal{A}$ then even with right preconditioning, GMRES would give residual and error reduction in norms dependent on the augmentation parameter γ.

A number of variants of this preconditioning approach are described in the literature, but we expect similar behaviour for extreme values of any parameter having an analogous role to γ.

4 Conclusions

There are situations where it is important to consider the effect of preconditioning on convergence; whilst preconditioning can lead to fewer iterations - indeed this is exactly why preconditioning is usually used - inaccurate solutions may be obtained with inappropriate convergence criteria in situations where preconditioning leads to highly distorted norms.

Acknowledgement I am grateful to an anonymous referee for comments which improved this manuscript and to Ke Chen for his patience.

References

[1] Arioli, M. , 2004, A Stopping Criterion for the Conjugate Gradient Algorithm in a Finite Element Method Framework. *Numer. Math.* **97**, 1–24.

[2] Elman, H.C., Silvester, D.J. and Wathen, A.J. ,2005, Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics. Oxford University Press, Oxford.

[3] Paige, C.C. and Saunders, M.A., 1975, Solution of sparse indefinite systems of linear equations, *SIAM J. Num. Anal.* **12**, 617–629.

[4] Saad, Y. and Schultz, M.H., 1986, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.* **7**, 856–869.

[5] Axelsson, O., 1994, Iterative Solution Methods, Cambridge University Press, New York.

[6] Vinsome, P.K.W., 1976, Orthomin, an iterative method for solving sparse sets of linear equations, in Proceedings of the 4th Symposium on Reservoir Simulation, Society of Petroleum Engineers of AIME, 149–159.

[7] Bramble, J. and Pasciak, J., 1988, A preconditioning technique for indefinite systems resulting from mixed approximation of elliptic problems, *Math. Comput.* **50**, 1–17.

[8] Klawonn, A., 1998, Block-triangular preconditioners for saddle point problems with a penalty term, *SIAM J. Sci. Comput.* **19**, 172–184.

[9] Wathen, A.J., 1987, Realistic eigenvalue bounds for the Galerkin mass matrix, *IMA J. Numer. Anal.* **7**,449–457.

[10] Benzi, M, Golub, G.H. and Liesen, J., 2005, Numerical solution of saddle point problems, *Acta Numer.* **14**, 1–137.

[11] Arioli, M. and Loghin, D., 2006, Stopping criteria for mixed finite element problems, *Electr. Trans. Numer. Anal.* submitted

[12] Greif, C. and Schötzau, D., 2006, Preconditioners for saddle point linear systems with highly singular (1,1) blocks, *Electronic Trans. Numer. Anal.* **22**, 114–121.

[13] Golub, G.H. and Greif, C., 2003, On solving block-structured indefinite linear systems, *SIAM J. Sci. Comput.* **24**, 2076–2092.

[14] Benzi, M. and Olshanskii, M.A., 2006, An Augmented Lagrangian-based approach to the Oseen problem, *SIAM J. Sci. Comput.* **28**, 2095–2113.

[15] de Niet, A.C. and Wubs, F.W., Two preconditioners for saddle point problems in fluid flows, *Int. J. Numer. Methods Fluids* (to appear)

[16] Elman, H.C., Ramage, A. and Silvester. D.J., IFISS: a Matlab toolbox for modelling incompressible flow, *ACM Trans. Math. Software* (to appear)