Fast Iterative Solution Methods for the Incompressible Navier–Stokes Equations

Michele Benzi
Scuola Normale Superiore, Pisa

SCUOLA
NORMALE
SUPERIORE

Iterative Solution of Large-Scale Saddle-Point Problems
Cortona, 9–20 May 2022

Outline

- The problem
- Linearizations
- Weak form and finite element discretization
- Linear systems of saddle point type
- Krylov subspace methods
- Preconditioning
- Augmented Lagrangian formulation
- Convergence analysis
- Relaxed Dimensional Factorization
- Numerical examples
- A large scale problem from hemodynamics
- Conclusions

The problem

Consider the primitive variables formulation of the incompressible Navier–Stokes equations:

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} & \text{on } \Omega \times (0, T], \\ \operatorname{div} \mathbf{u} = 0 & \text{on } \Omega \times [0, T], \\ \mathbf{u} = \mathbf{g} & \text{on } \partial\Omega \times [0, T], \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) & \text{on } \Omega \end{cases}$$

on a Lipschitz domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, where $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ and $p = p(\mathbf{x}, t)$ are the unknown velocity and pressure fields, ν is the kinematic viscosity (inversely proportional to the Reynolds number, Re) and $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{u}_0$ are given functions.

More generally, both Dirichlet and Neumann boundary conditions may be prescribed on different parts of $\partial\Omega$.

The problem (cont.)

The N–S equations are the fundamental model governing the flow of an incompressible, viscous, Newtonian fluid and are widely used in scientific, biomedical, and industrial applications, besides having been the subject of intensive mathematical investigations for many years.

They were first derived by Claude-Louis Navier and independently by George Gabriel Stokes.

C.-L. Navier, *Mémoire sur le lois du mouvement des fluides*, Mémoires de l'Académie Royale des Sciences, VI (1823), pp. 389–416.

G. G. Stokes, *On the theories of the internal friction of fluids in motion, and of the equilibrium and motion of elastic solids*, Transactions of the Cambridge Philosophical Society, VIII (1846), pp. 287–305.

The problem (cont.)

For $\nu = 0$ the Navier–Stokes equations reduce to the Euler equations for an ideal inviscid fluid (1755).

In 2D, the global existence and uniqueness of smooth solutions of the Navier–Stokes (and Euler) equations has been known for a long time (Prodi, Serrin, J.-L. Lions,...).

In 3D, as is well known, the global well-posedness of the N-S equations for arbitrarily "large" data remains open (it's one of the seven "Millennium Prize Problems" proposed by the Clay Mathematics Institute).

Partial results are known: global existence and uniqueness for "small" initial data $\mathbf{u}_0$, local existence and uniqueness for arbitrary initial data, existence (but not uniqueness) of global *weak* solutions (Leray), etc.

The problem (cont.)

Complications also arise for the stationary Navier–Stokes equations:

$$\begin{cases} -\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{on } \Omega, \\ \operatorname{div}\mathbf{u} &= 0 \quad \text{on } \Omega, \\ \mathbf{u} &= \mathbf{g} \quad \text{on } \partial\Omega, \end{cases}$$

For small ν (i.e., large Reynolds numbers), the problem becomes convection-dominated and the solution can be expected to exhibit difficult-to-capture boundary layers; moreover, uniqueness (of weak solutions) may fail to hold unless the forcing term $\mathbf{f}$ satisfies a condition of the form

$$\|\mathbf{f}\|_{-1} := \sup \frac{\langle \mathbf{f}, \mathbf{u}\rangle}{\|\nabla\mathbf{u}\|} \leq \frac{\nu^2}{\Gamma_*}. \tag{1}$$

Here the supremum is taken over all divergence-free vector fields $\mathbf{u} \neq 0$ with components in $H_0^1(\Omega)$ and Γ_* is the best possible constant for which

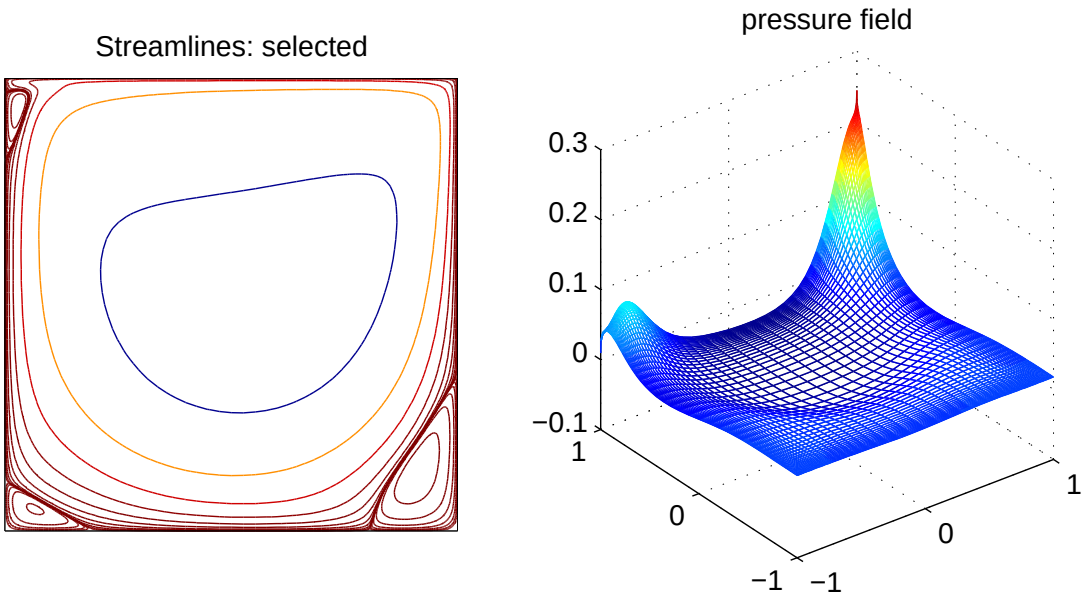
$$c(\mathbf{z},\mathbf{u},\mathbf{v}) := \int_\Omega (\mathbf{z}\cdot\nabla\mathbf{u})\cdot\mathbf{v}\,d\mathbf{x} \leq \Gamma\|\nabla\mathbf{z}\|\|\nabla\mathbf{u}\|\|\nabla\mathbf{v}\|$$

holds.

Example 1: lid driven cavity problem

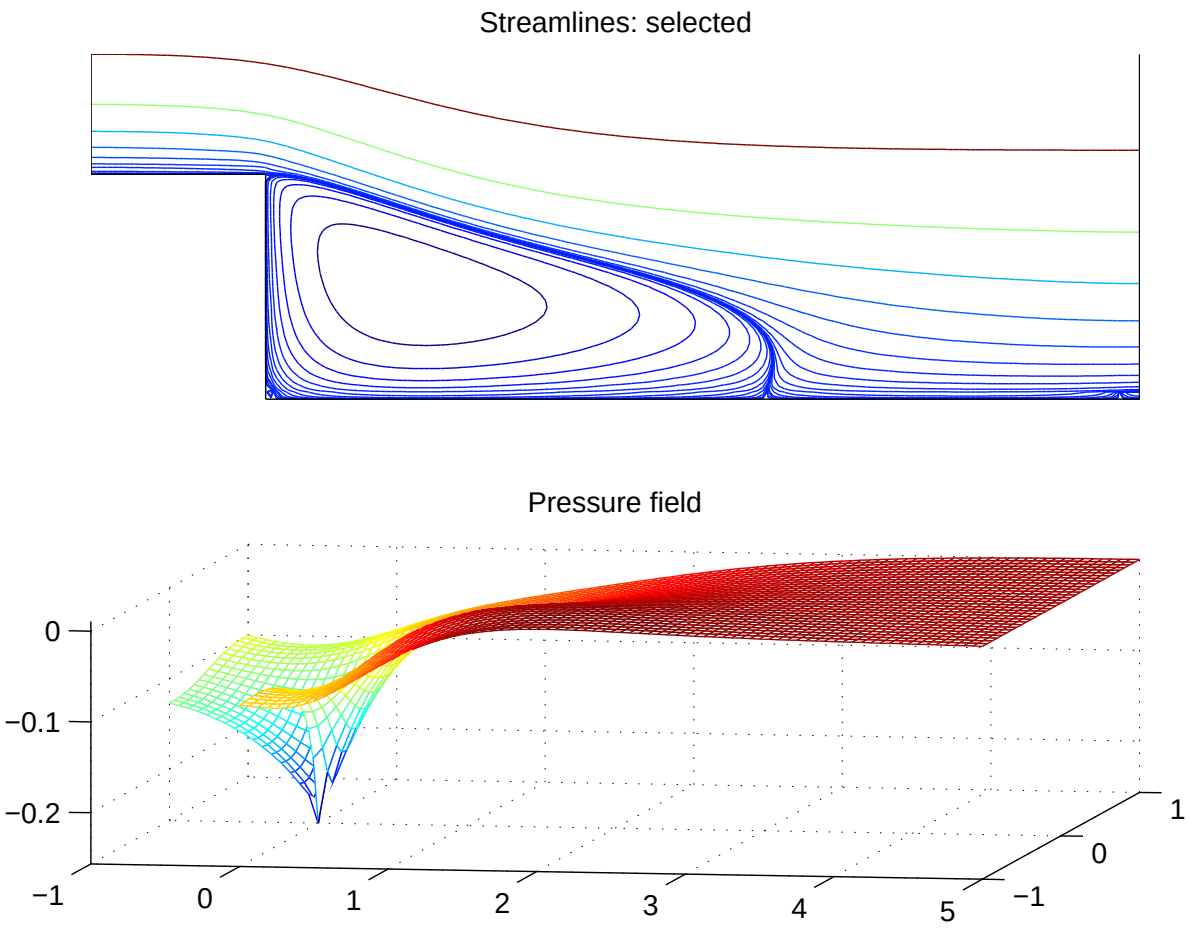
In the lid driven cavity problem, the flow is enclosed in a square with $u_1 = 1 - x^4, u_2 = 0$ on the top to represent the moving lid.

Figure: Lid driven cavity (Q2-Q1, $\nu = 0.001$, stretched 128×128 grid)



Example 2: backward facing step test problem

Figure: Backward facing step problem (Q2-Q1, $\nu = 0.005$, uniform 64×192 grid)



The problem (cont.)

Closed-form solutions of the N-S equations are known only in a few simple cases (e.g., laminar flow in a pipe).

The numerical solution of the Navier–Stokes equations requires the following main steps:

1. Linearization;
2. Space and time discretization;
3. Solution of the discrete (algebraic) problem.

Different linearization strategies are in use, depending on the value of the viscosity ν (equivalently, on the Reynolds number).

Note: We will only consider problems in the sub-critical Reynolds number regime (e.g., laminar flow, such as blood flow in arteries), but many of the techniques here described are also useful in certain turbulence models.

Linearizations

For highly viscous fluids (low Reynolds numbers), the nonlinear convective terms can be dropped, leading to the Stokes problem:

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + \nabla p &= \mathbf{f} & \text{on } \Omega \times (0, T], \\ \operatorname{div} \mathbf{u} &= 0 & \text{on } \Omega \times [0, T], \\ \mathbf{u} &= \mathbf{g} & \text{on } \partial\Omega \times [0, T], \\ \mathbf{u}(\mathbf{x}, 0) &= \mathbf{u}_0(\mathbf{x}) & \text{on } \Omega \end{cases}$$

These equations are appropriate for low-speed, "creeping" flow, or for describing the motion of narrowly confined fluids. This is a well-studied problem, with many results and optimal iterative solvers available.

D. J. Silvester and A. J. Wathen, *SIAM J. Numer. Anal.*, 31 (1994), pp. 1352–1367.

Linearizations (cont.)

For moderate or high Reynolds numbers, the Stokes problem is a poor approximation. Linearization requires replacing the nonlinear problem with a sequence of linear problems the solutions of which converge, under appropriate conditions, to the solution of the original nonlinear problem.

Two main linearization techniques exist: Newton's method and Picard's iteration. Moreover, one can choose to first discretize, then linearize or to first linearize, then discretize. Here we choose to first linearize, then discretize.

Newton's method is quadratically convergent, but it requires the initial guess $\mathbf{u}^{(0)}$ to be sufficiently close to the solution. For a steady problem, the radius of the ball of convergence is typically proportional to ν.

Picard iteration, on the other hand, converges at a linear rate but is globally convergent provided that the standard uniqueness condition (1) is satisfied. The two methods are often combined in practice.

Linearizations (cont.)

Picard's iteration is simply a fixed-point iteration in function space. Combined with an implicit time-stepping scheme (the simplest one being backward Euler), it leads to a sequence of linear systems of PDEs of the form

$$\begin{cases} \alpha \mathbf{u}^{(k+1)} - \nu \Delta \mathbf{u}^{(k+1)} + (\mathbf{u}^{(k)} \cdot \nabla)\mathbf{u}^{(k+1)} + \nabla p^{(k+1)} = \mathbf{f}^{(k)} & \text{on } \Omega, \\ \operatorname{div} \mathbf{u}^{(k+1)} = 0 & \text{on } \Omega, \\ \mathbf{u}^{(k+1)} = \mathbf{g} & \text{on } \partial\Omega, \end{cases}$$

($k = 0, 1, 2, \ldots$), where $\alpha = O((\Delta t)^{-1})$, with Δt the time step. For steady problems, $\alpha = 0$. In this case usually $\mathbf{u}^{(0)} = 0$, so that the first step consists in the solution of a Stokes problem.

The typical number of Picard iterations needed to converge to the nonlinear solution is usually between 5 and 20, depending on the size of α and ν. Convergence is slower for $\alpha = 0$ and small ν.

Linearizations (cont.)

Assume $\alpha = 0$. Writing $\mathbf{u}$ for $\mathbf{u}^{(k+1)}$ and $\mathbf{w}$, $\mathbf{f}$ for $\mathbf{u}^{(k)}$, $\mathbf{f}^{(k)}$ we obtain the steady Oseen problem:

$$\begin{cases} -\nu\Delta\mathbf{u} + (\mathbf{w}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{on } \Omega, \\ \operatorname{div}\mathbf{u} &= 0 \quad \text{on } \Omega, \\ \mathbf{u} &= \mathbf{g} \quad \text{on } \partial\Omega, \end{cases}$$

where $\mathbf{w}$, the "wind", satisfies $\operatorname{div}\mathbf{w} = 0$.

The rest of the talk will focus mainly on the numerical solution of this problem, since this is where all the numerical difficulties are.

For ease of exposition, we will assume $\mathbf{g} = 0$ unless otherwise stated.

Weak form and finite element discretization

Let $\langle \cdot, \cdot \rangle$ denote the L^2-inner product (we use the same notation for both vector and scalar functions), and denote by $L^2_0(\Omega)$ the subspace of $L^2(\Omega)$ consisting of all functions p with $\langle p, 1 \rangle = 0$.

Also, we write $\mathbf{H}^1_0(\Omega)$ for $(H^1_0(\Omega))^d$.

Weak formulation: Given $\mathbf{f} \in \mathbf{H}^{-1}(\Omega)$, find $(\mathbf{u}, p) \in \mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$ such that

$$\alpha\langle \mathbf{u}, \mathbf{v} \rangle + \nu\langle \nabla \mathbf{u}, \nabla \mathbf{v} \rangle + \langle (\mathbf{w} \cdot \nabla)\, \mathbf{u}, \mathbf{v} \rangle - \langle p, \operatorname{div} \mathbf{v} \rangle = \langle \mathbf{f}, \mathbf{v} \rangle, \quad \mathbf{v} \in \mathbf{H}^1_0(\Omega),$$

$$\langle q, \operatorname{div} \mathbf{u} \rangle = 0, \quad q \in L^2_0(\Omega).$$

The weak form of the standard Oseen problem is obtained for $\alpha = 0$ (steady case).

Weak form and finite element discretization (cont.)

Consider now finite-dimensional subspaces $\mathbf{V}_h \subset \mathbf{H}_0^1(\Omega)$ and $Q_h \subset L_0^2(\Omega)$.

Here h denotes the mesh width, or spatial discretization parameter, and the subspaces' dimensions tend to infinity as $h \to 0$.

Weak discrete formulation: Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that

$$\alpha\langle \mathbf{u}_h, \mathbf{v}_h\rangle + \nu\langle \nabla \mathbf{u}_h, \nabla \mathbf{v}_h\rangle + \langle (\mathbf{w}_h \cdot \nabla)\, \mathbf{u}_h, \mathbf{v}_h\rangle - \langle p_h, \operatorname{div} \mathbf{v}_h\rangle = \langle \mathbf{f}, \mathbf{v}_h\rangle, \quad \mathbf{v}_h \in \mathbf{V}_h,$$

$$\langle q_h, \operatorname{div} \mathbf{u}_h\rangle = 0, \quad q_h \in Q_h.$$

This discrete problem is known as the Galerkin formulation. In the finite element method, $\mathbf{V}_h$ and Q_h are subspaces spanned by polynomials of low degree, locally supported on the cells (triangles, tetrahedra, etc.) arising from the subdivision of Ω into small elements.

Weak form and finite element discretization (cont.)

The choice of the velocity and pressure finite element spaces $\mathbf{V}_h$ and Q_h is a delicate matter.

In order to have a stable discretization, the two subspaces must satisfy a compatibility condition known as the discrete inf-sup or LBB condition (named after Ladyzhenskaya, Babuška, and Brezzi):

There exists a constant $\gamma_0 > 0$ (independent of h) such that

$$\inf_{q_h \neq const.} \sup_{\mathbf{v}_h \neq 0} \frac{|\langle q_h, \operatorname{div} \mathbf{v}_h \rangle|}{\|\nabla \mathbf{v}\| \|q_h\|} \geq \gamma_0. \tag{2}$$

A pair $\mathbf{V}_h$, Q_h is inf-sup stable if it satisfies (2). Apparently natural choices of finite element spaces, such as equal degree interpolation for velocity and pressure, are not inf-sup stable. For example, both the $\mathbf{P}_1$-$\mathbf{P}_1$ and the $\mathbf{Q}_1$-$\mathbf{Q}_1$ elements, which are the simplest globally continuous approximations, are not stable.

Weak form and finite element discretization (cont.)

Unstable finite element pairs can, however, be stabilized, at the price of slightly relaxing the incompressibility condition.

See for example

H. Elman, D. Silvester and A. Wathen, *Finite Elements and Fast Iterative Solvers. With Applications in Incompressible Fluid Dynamics*, 2nd Ed., Oxford University Press, 2014.

Weak form and finite element discretization (cont.)

A typical error estimate for FEM approximations is the following.

Theorem: Let $\{\mathcal{T}_h\}$ be a regular family of triangulations of Ω consisting of simplices, such that every $T \in \mathcal{T}_h$ has at least one vertex in the interior of Ω. Then the Taylor–Hood finite element pair $\mathbf{P}_k$-$\mathbf{P}_{k-1}$ with $k \geq 2$ is LBB stable.

Moreover, if the solution to the Oseen problem $(\mathbf{u}, p)$ belongs to $\mathbf{H}^m(\Omega) \times H^{m-1}(\Omega)$ with $m \geq 2$, then for $2 \leq m \leq k+1$ the following holds:

$$\|\mathbf{u} - \mathbf{u}_h\|_{H^1} + \|p - p_h\|_{L^2} \leq C\, h^{m-1}(|\mathbf{u}|_m + |p|_{m-1}),$$

with C a constant independent of h and of $(\mathbf{u}, p)$.

Linear systems of saddle point type

Discretization of Oseen's problem using finite elements leads to large, sparse linear algebraic systems of the form

$$\begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

Systems of this form are called (stabilized) saddle point problems. Here A is a block diagonal matrix with d blocks ($d = 2$ or 3), each of which is a discrete convection-diffusion operator.

The block B^T is a discrete gradient, and B a discrete (negative) divergence operator.

The matrix C has small norm, and is zero for a stable finite element pair.

The entries of the vectors u, f, and p contain the coefficients of the linear combinations expressing $\mathbf{u}_h$, $\mathbf{f}_h$, and p_h with respect to the basis functions spanning the spaces $\mathbf{V}_h$ and Q_h, respectively.

Linear systems of saddle point type (cont.)

In the case of the Stokes problem, $A = A^T$ and each diagonal block is a discretization of the diffusion operator $-\nu\Delta$. For a stable finite element pair, the system is symmetric and takes the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

This system also arises from the method of Lagrange multipliers for finding the minimum of the quadratic (energy) function

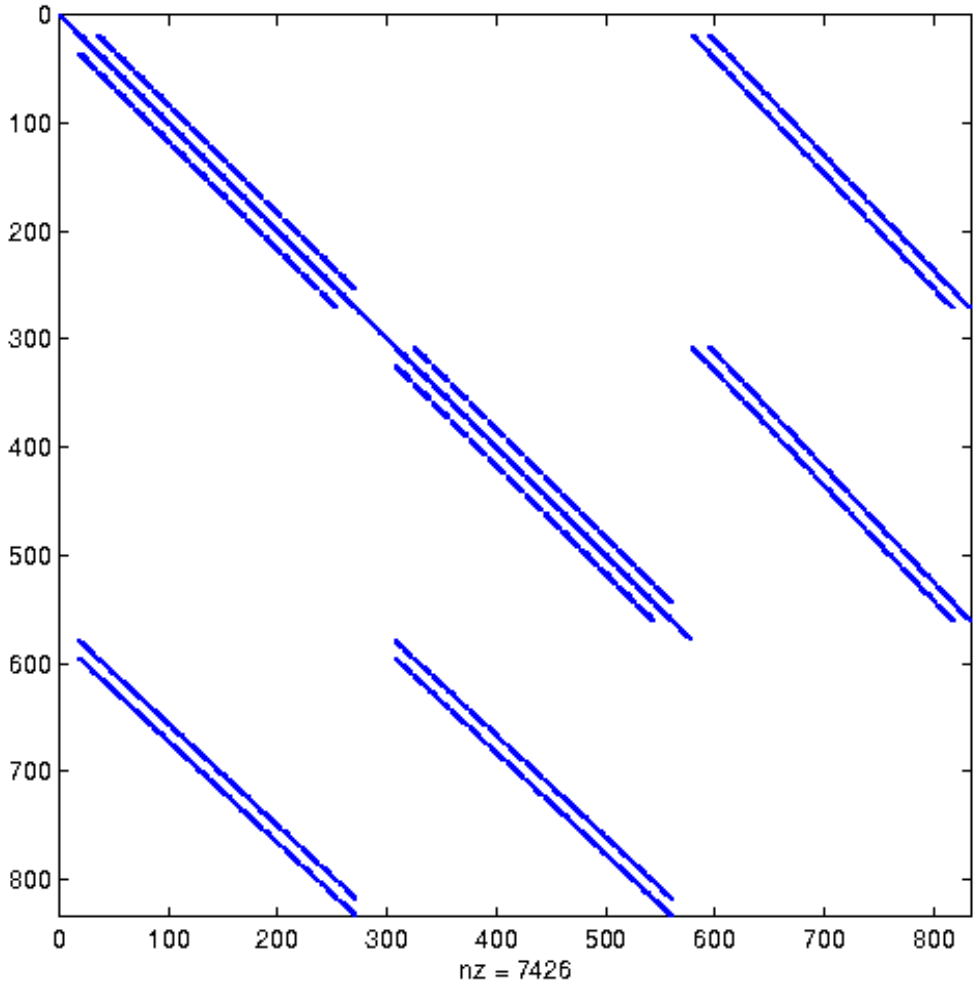
$$\frac{1}{2}\langle Au, u\rangle - \langle f, u\rangle$$

subject to the constraint $Bu = 0$. Hence, the pressure has the meaning of a Lagrange multiplier.

When $A \neq A^T$ we no longer have a genuine saddle point problem, but the same terminology is used.

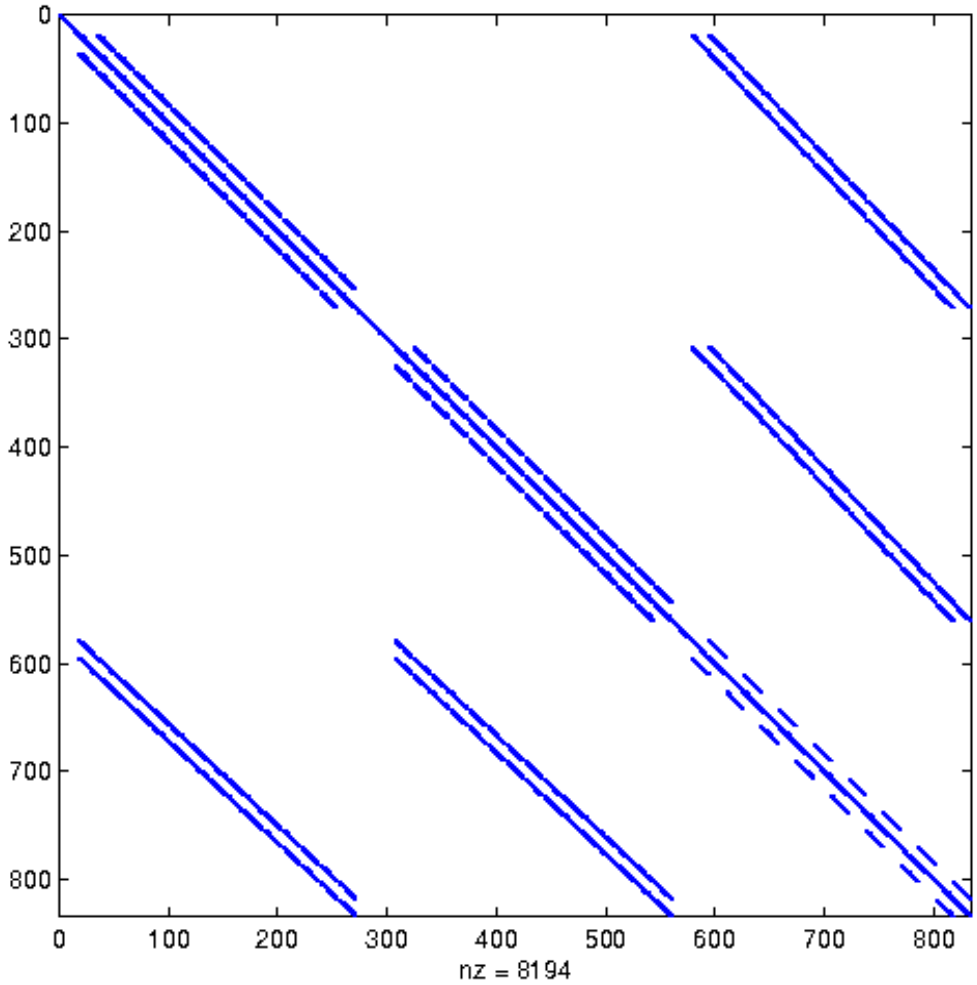
M. B., G. Golub, and J. Liesen, *Acta Numerica*, 14 (2005), pp. 1–137.

Linear systems of saddle point type (cont.)



Without stabilization ($C = O$)

Linear systems of saddle point type (cont.)



With stabilization ($C \neq O$)

Solution of the linear algebraic system

Linear systems of saddle point type arising from incompressible flow problems can be very challenging to solve, especially in the steady case, for small values of the viscosity ν and on stretched meshes.

Direct methods based on the factorization of the coefficient matrix

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \quad or \quad \mathcal{A} = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

are generally not suitable for large scale problems, especially in the 3D case.

Instead, iterative methods must be used.

Solution of the linear algebraic system (cont.)

When solving linear systems of the form $A_h x_h = b_h$ arising from the discretization of PDEs, the goal is to develop iterative methods that are optimal, in the sense that

1. The rate of convergence is independent of h (the mesh size);
2. The cost per iteration scales linearly in the number n of unknowns.

Clearly, with an optimal method it is possible to approximate the solution within a prescribed accuracy with a cost that scales like $O(n)$, for $n \to \infty$ (that is, for $h \to 0$).

Ideally, the method should also be robust with respect to variations in the problem parameters (for example, the viscosity or other PDE coefficients).

Another desideratum is parallel scalability.

Krylov subspace methods

Suppose x_0 is an initial guess for the solution of a linear system $Ax = b$, and let $r_0 = b - Ax_0$ be the corresponding residual.

Krylov subspace methods are iterative (approximation) schemes whose kth iterate x_k satisfies

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots$$

where

$$\mathcal{K}_k(A, r_0) \equiv \text{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\}$$

denotes the *kth Krylov subspace generated by A and* r_0. The Krylov subspaces form a finite chain

$$\mathcal{K}_1(A, r_0) \subset \mathcal{K}_2(A, r_0) \subset \cdots \subset \mathcal{K}_m(A, r_0) = \cdots = \mathcal{K}_n(A, r_0).$$

Note that the elements of $\mathcal{K}_k(A, r_0)$ are of the form $p_k(A)r_0$, where p_k ia a polynomial of degree $k - 1$.

Krylov subspace methods (cont.)

For $k \leq m$, the kth Krylov subspace $\mathcal{K}_k(A, r_0)$ has dimension k. Thus, there are k degrees of freedom in the choice of the iterate x_k.

Uniquely defined iterates are obtained by imposing k constraints, in the form of orthogonality of the kth residual r_k with respect to a prescribed k-dimensional subspace $\mathcal{C}_k$:

$$r_k = b - Ax_k \in r_0 + A\mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{C}_k.$$

This is known as a *Petrov–Galerkin condition*, or simply *Galerkin condition* when $\mathcal{C}_k \equiv \mathcal{K}_k(A, r_0)$.

Different choices of the constraint subspace $\mathcal{C}_k$ lead to different types of Krylov subspace methods.

Def.: If A is SPD, the *A-norm* of $x \in \mathbb{R}^n$ is $\|x\|_A := (\langle Ax, x\rangle)^{\frac{1}{2}}$.

Krylov subspace methods (cont.)

Theorem (Saad): Assume $\dim \mathcal{K}_k(A, r_0) = k$, and let $x^* = A^{-1}b$.

1. If A is SPD and $\mathcal{C}_k = \mathcal{K}_k(A, r_0)$, then x_k is uniquely defined and satisfies

$$\|e_k\|_A \equiv \|x^* - x_k\|_A = \min_{z \in x_0 + \mathcal{K}_k(A, r_0)} \|x^* - z\|_A = \min_{p \in \Pi_k} \|p(A)e_0\|_A.$$

2. If A is nonsingular and $\mathcal{C}_k = A\mathcal{K}_k(A, r_0)$, then x_k is uniquely defined and satisfies

$$\|r_k\|_2 \equiv \|b - Ax_k\|_2 = \min_{z \in x_0 + \mathcal{K}_k(A, r_0)} \|b - Az\|_2 = \min_{p \in \Pi_k} \|p(A)r_0\|_2.$$

Here Π_k denotes the set of all polynomials of degree at most $k - 1$ such that $p(0) = 1$.

The conjugate gradient (CG) method is of Type 1, while minimal residual methods (like MINRES and full GMRES) are of Type 2.

Minimal residual methods

Since the coefficient matrix in the discrete Stokes and Oseen problems is not SPD, we are restricted to using minimal residual methods like MINRES (Paige and Saunders, 1975) or GMRES (Saad and Schultz, 1986).

In a nutshell, this method generates successive approximations x_k to the solution which minimize the Euclidean norm of the residual $r_k = b - Ax_k$ over the (nested) Krylov subspaces of increasing dimension. Note that $\|b - Ax_k\|_2$ cannot increase from one step to the next.

These methods are efficient provided that good approximations x_k to x^* can be obtained from a Krylov subspace of dimension $k \ll n$, where n is the size of the linear system.

Minimal residual methods (cont.)

Like all Krylov subspace methods, MINRES and GMRES are projection methods; the original problem is projected onto a subspace of lower dimension (the Krylov subspace) in which the solution of the residual minimization problem can be easily found, either by solving a small linear system or by solving a small least squares problem.

As is well known, computing projections onto a subspace is greatly facilitated if an orthonormal basis for the subspace is known. This is also desirable for numerical stability reasons.

An orthonormal basis for a Krylov subspace can be efficiently constructed using the Arnoldi process; in the symmetric case, this is known as the Lanczos process (Arnoldi, 1951; Lanczos, 1952). Both of these are efficient implementations of the classical Gram–Schmidt process.

Arnoldi's process

For arbitrary $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$, $v \neq 0$, the Arnoldi process is:

- Set $v_1 = v/\|v\|_2$;
- For $j = 1, \ldots, m$ do:
  - $h_{i,j} = \langle Av_i, v_j \rangle$ for $i = 1, 2, \ldots, j$
  - $w_j = Av_j - \sum_{i=1}^{j} h_{i,j} v_i$
  - $h_{j+1,j} = \|w_j\|_2$
  - If $h_{j+1,j} = 0$ then STOP;
  - $v_{j+1} = w_j/\|w_j\|_2$

Arnoldi's process (cont.)

Remarks:

(i) If the algorithm does not stop before the mth step, the Arnoldi vectors $\{v_1, \ldots, v_m\}$ form an ONB for the Krylov subspace $\mathcal{K}_m(A, v)$.

(ii) At each step j, the Arnoldi process requires one matrix-vector product, $j + 1$ inner products, and j linked triads.

(iii) At each step the algorithm computes Av_j and then orthonormalizes it against all previously computed v_j's.

Arnoldi's process (cont.)

Define $V_m = [v_1, \ldots, v_m] \in \mathbb{R}^{n \times m}$. Introducing the $(m+1) \times m$ matrix $\widehat{H}_m = [h_{ij}]$ and the $m \times m$ upper Hessenberg matrix H_m obtained by deleting the last row of $\widehat{H}_m$, the following Arnoldi relations hold:

$$AV_m = V_m H_m + w_m e_m^T = V_{m+1} \widehat{H}_m$$

$$V_m^T A V_m = H_m$$

Hence, the $m \times m$ matrix H_m is precisely the projected matrix $V_m^T A V_m$. If $A = A^T$, then $H_m = H_m^T = T_m$ is a tridiagonal matrix, and the Arnoldi process becomes the Lanczos process, which is much cheaper in terms of both operations and storage.

Thus, the Lanczos process consists of a three-term recurrence, with constant operation count and storage costs per step, whereas the Arnoldi process has increasing costs for increasing j. This is the main difference between MINRES and GMRES.

GMRES

The GMRES algorithm for solving $Ax = b$ with A nonsingular is as follows:

- Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $v_1 = r_0/\beta$.
- Run m steps of the Arnoldi process on A and v_1.
- Let $\widehat{H}_m = [h_{i,j}]$ where $1 \leq i \leq m+1$ and $1 \leq j \leq m$.
- Solve $\|\beta e_1 - \widehat{H}_m y_m\|_2 = \min$ for y_m and let $x_m = x_0 + V_m y_m$.

Remarks:

(i) The least squares problem can be efficiently solved via the QR factorization $\widehat{H}_m = Q_m R_m$ using Givens rotations.

(ii) The modulus of the last component of $g_m := Q_m^T(\beta e_1)$ is equal to $\|b - Ax_m\|_2$ in exact arithmetic, hence we can monitor convergence after each step.

(iii) A breakdown in the Arnoldi process ($h_{j+1,j} = 0$) means the projection is exact ($A = V_j H_j V_j^T$) and the exact solution has been found.

Convergence of GMRES

The convergence of minimal residual methods for $Ax = b$ can be analyzed if we assume that A is diagonalizable: $A = XDX^{-1}$, with D diagonal. We denote the spectrum of A by $\sigma(A)$.

After k steps we have

$$\|r_k\|_2 = \|b - Ax_k\|_2 = \min_{p \in \Pi_k} \|p(A)r_0\|_2 = \min_{p \in \Pi_k} \|Xp(D)X^{-1}r_0\|_2$$

$$\leq \|X\|_2\|X^{-1}\|_2\|r_0\|_2 \min_{p \in \Pi_k} \|p(D)\|_2 = \kappa_2(X)\|r_0\|_2 \min_{p \in \Pi_k} \max_{\lambda \in \sigma(A)} |p(\lambda)|,$$

where we have set $\kappa_2(X) = \|X\|_2\|X^{-1}\|_2$. Hence, the residual at step k satisfies

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \kappa_2(X) \min_{p \in \Pi_k} \max_{\lambda \in \sigma(A)} |p(\lambda)|, \quad k = 1, 2, \ldots.$$

Note that the bound is only useful if the right-hand side is < 1.

Convergence of GMRES (cont.)

If A is normal ($AA^T = A^T A$), in particular symmetric, then X can be taken to be orthogonal and therefore $\kappa_2(X) = 1$. In this case the bound becomes

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \min_{p \in \Pi_k} \max_{\lambda \in \sigma(A)} |p(\lambda)|, \quad k = 1, 2, \ldots.$$

It has been shown that this bound is sharp. Even if A is not normal, the bound can still yield useful information if $\kappa_2(X)$ is not too large.

Hence, in this case the eigenvalues of A are descriptive of the convergence behavior.

Convergence of GMRES (cont.)

On the other hand, if $\kappa_2(X) \gg 1$ the bound is effectively useless, since the right-hand side is often > 1.

This does not mean that converge will be slow: only that the bound is not descriptive of the actual convergence behavior. In other words, the bound can be very far from being sharp in the highly non-normal case.

Furthermore, or general (non-normal) matrices the eigenvalues alone are not sufficient to describe the convergence behavior (examples by Greenbaum, Ptàk, and Strakoš).

Restarted GMRES

Because GMRES needs to explicitly build and store an orthonormal basis for the Krylov subspace $\mathcal{K}_m(A, r_0)$, its storage and arithmetic costs increase with m. Unless convergence is fast, it is often necessary to periodically "restart" the algorithm, say every m steps, using the current approximation x_m as the new starting vector.

This variant, denoted GMRES(m), is widely used in practice; the choice of the restart parameter m is often dictated by the available memory. Typical values are $m = 20$ or $m = 30$, but values as small as $m = 5$ and as large as $m = 50$ are also used.

Restarting can drastically alter the convergence properties of GMRES, since the global optimality of the method is lost. The method may even fail to converge: perennial stagnation is possible, for any $m < n$.

Restarted GMRES (cont.)

Convergence of restarted GMRES is assured, for any m, if A is positive definite, thanks to the following result due to Elman:

Theorem: Assume $H = \frac{1}{2}(A + A^T)$ is SPD, and let $\sigma = \|A\|_2$ and $\mu = \lambda_{\min}(H) > 0$. Then the residuals in GMRES(m) satisfy

$$\|r_{k+1}\|_2 \leq \left(1 - \frac{\mu^2}{\sigma^2}\right) \|r_k\|_2.$$

It follows that if A is positive definite, GMRES(m) converges for all $m \geq 1$.

Note that convergence can be quite slow if $\mu \approx 0$ and $\sigma = O(1)$.

A clear limitation of this result is that it may be difficult to prove that $H = \frac{1}{2}(A + A^T)$ is positive definite, particularly when preconditioning is being used, and to estimate μ.

The Faber–Manteuffel Theorem

We have seen so far that there exist Krylov subspace methods with optimality properties, such CG, MINRES and GMRES. Some of these methods are based on short recurrences (CG, MINRES), others require long recurrences (GMRES). Truncating or restarting the recurrences in GMRES leads to loss of global optimality.

Hence, in the Hermitian case ($A = A^*$) there are optimal algorithms based on short recurrences. Methods with these desirable properties have also been developed for skew-Hermitian matrices, shifted skew-symmetric matrices, and shifted Hermitian matrices of the form $A = C + z\,I$ with $C = C^*$ and $z \in \mathbb{C}$.

However, in spite of much research, no such methods have been found for general matrices.

The Faber–Manteuffel Theorem (cont.)

In the early 1980s, Gene Golub offered a reward to anyone who could find such a method, or prove that it cannot exist. In other words, the problem is:

Completely characterize the class of matrices A for which an optimal method based on short recurrences exist.

The challenge was taken up by V. Faber and T. Manteuffel. In 1984 they proved a deep theorem that in essence states that apart from trivial cases, the only matrices for which such methods can be defined are the Hermitian ones and "shifted and rotated skew-Hermitian matrices" of the form $A = \mathrm{e}^{i\theta}(\rho I + B)$ where $\theta, \rho \in \mathbb{R}$ and $B = -B^*$. This class includes all previously known cases.

Note that all such matrices have spectra that are contained in a straight line segment in $\mathbb{C}$.

Other Krylov subspace methods

While the Faber–Manteuffel Theorem put an end to the search for optimal short-recurrence Krylov methods for general non-Hermitian systems, several new Krylov subspace methods were introduced in the early 1990s for solving such systems.

These methods give up or relax the global optimality requirement and make use of coupled 2-term or 3-term recurrences.

The most successful methods are *hybrid methods*, obtained by combining algorithms based on the nonsymmetric Lanczos process with local residual minimization, and *quasi-minimal residual methods*, obtained by relaxing the strict residual minimization property.

Other Krylov subspace methods (cont.)

For non-Hermitian linear systems, the most succesful alternatives to GMRES are:

- Bi-CGStab (van der Vorst, 1991)
- QMR and its variants TFQMR and SQMR (Freund & Nachtigal, 1991-1993)

All these methods have lower storage requirements than GMRES. In terms of convergence rates, Bi-CGStab is often competitive with restarted GMRES; SQMR is especially attractive for solving symmetric indefinite linear systems if a symmetric indefinite preconditioner is to be used.

It should be kept in mind that each iteration of Bi-CGStab and TFQMR requires two matrix-vector multiplies with A and two applications of the preconditioner.

Other Krylov subspace methods (cont.)

A weakness of these methods is the possibility of breakdowns in the underlying Lanczos process. Look-ahead techniques exist to avoid breakdowns, at the cost of more complicated coding and additional work.

Another weakness is the total lack of theoretical basis for the convergence of these methods. The absence of optimality properties makes it virtually impossible to analyze their convergence properties.

It is our experience, moreover, that when a good preconditioner is available, the differences between the various nonsymmetric Krylov iterations tend to disappear.

For this reason, in the last 25 years or so much more effort has been put in developing effective preconditioners than in research on the Krylov methods themselves.

Preconditioning

When GMRES is applied to the solution of systems obtained from the linearization and discretization of the Navier–Stokes equations, the convergence is hopelessly slow.

Moreover, the rate of convergence deteriorates rapidly as $h \to 0$, especially for small viscosities.

The answer to this problem is to use preconditioning.

To precondition a linear system of equations means to transform it into one with more favorable properties for iterative solvers like GMRES. At the same time, the cost of this transformation should be small (no more than $O(n)$ operations if optimality is desired).

Preconditioned GMRES

The right-preconditioned GMRES algorithm for solving $Ax = b$ with A nonsingular and preconditioner M also nonsingular reads:

- Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $v_1 = r_0/\beta$.
- Run m steps of the Arnoldi process on AM^{-1} and v_1.
- Let $\widehat{H}_m = [h_{i,j}]$ where $1 \leq i \leq m+1$ and $1 \leq j \leq m$.
- Solve $\|\beta e_1 - \widehat{H}_m y_m\|_2 = \min$ for y_m and let $x_m = x_0 + M^{-1} V_m y_m$.

Remark: Right preconditioning preserves the 2-norm of the residual, which is the quantity being minimized at each step over the Krylov subspace. In contrast, with left preconditioning the quantity being minimized is the preconditioned residual $\|M^{-1}(b - Ax_m)\|_2$. The same preconditioner M can lead to very different convergence behavior depending on which side it is applied to.

Flexible GMRES

It happens frequently that the preconditioner M changes in the course of an iterative process. For example, the preconditioner may be computed adaptively; or, more often, the application of the preconditioner requires itself one or more iterative processes. This is the case of nested iterations, and of block preconditioners where the blocks are handled iteratively.

Unless the inner iterations consist of a fixed number of steps of a stationary iteration, the preconditioner will change from one (outer) iteration to the next.

This is the case if the inner iteration consists of a Krylov method, or if the inner iteration is stopped on the basis of a prescribed residual norm reduction.

Flexible GMRES (cont.)

When we have a variable preconditioner, we can no longer talk of Krylov subspace methods. Methods like GMRES cannot be used with a variable preconditioner.

Nevertheless, so-called flexible methods have been developed, which allow for variable preconditioners $M_1, M_2, \ldots$ These methods are very popular in a wide variety of applications.

The price to pay for the added flexibility is an increase in storage costs.

For example, in flexible GMRES (FGMRES; Saad, 1993) one needs to store not just the basis vectors $v_1, v_2, \ldots, v_m$ produced by the Arnoldi process, but also the preconditioned vectors

$$z_1 = M_1^{-1} v_1, \; z_2 = M_2^{-1} v_2, \ldots, \; z_m = M_m^{-1} v_m \, .$$

Note that only right preconditioning is allowed here.

FGMRES

- Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $v_1 = r_0/\beta$.
- For $j = 1, \ldots, m$ do:
  - Compute $z_j = M_j^{-1} v_j$
  - Compute $w = Az_j$
  - Compute $h_{i,j} = \langle w, v_j \rangle$ for $i = 1, 2, \ldots, j$
  - Compute $w_j = Av_j - \sum_{i=1}^{j} h_{i,j} v_i$
  - Compute $h_{j+1,j} = \|w_j\|_2$
  - If $h_{j+1,j} = 0$ then STOP;
  - Compute $v_{j+1} = w_j / \|w_j\|_2$
  - End do
- Let $\widehat{H}_m = [h_{i,j}]$ where $1 \leq i \leq m+1$ and $1 \leq j \leq m$
- Let $Z_m = [z_1, z_2, \ldots, z_m]$
- Solve $\|\beta e_1 - \widehat{H}_m y_m\|_2 = \min$ for y_m and let $x_m = x_0 + Z_m y_m$.

Remark: The 2-norm of the residual is now minimized over the shifted subspace $S_m = x_0 + \operatorname{span}\{z_1, \ldots, z_m\}$.

Preconditioning (cont.)

For symmetric problems, the preconditioned system should have all (or most) of its eigenvalues clustered away from zero. Indeed, in this case there is a polynomial p of low degree such that $p(0) = 1$ and $p(\lambda) \approx 0$ for $\lambda \in \sigma(A)$.

For general systems it is trickier to specify the type of properties required for fast convergence. Eigenvalues clustered near 1 usually help, but this is not enough to establish desirable properties like h-independent convergence.

However, rigorous results can be obtained in some cases by studying the concept of field-of-values equivalence, see below.

Preconditioning (cont.)

Consider again linear systems of the form $A_h x_h = b_h$, where the problem dimension tends to infinity as $h \to 0$.

Informally, a preconditioner for A_h is a nonsingular matrix P_h such that the preconditioned system $P_h^{-1} A_h x_h = P_h^{-1} b_h$ is "easier" to solve by an iterative method (like CG or GMRES).

It should be emphasized that neither $P_h^{-1} A_h$ nor P_h^{-1} need to be formed explicitly. Instead, the requisite matrix-vector products with $P_h^{-1} A_h$ can be carried out by performing matrix-vector products with A_h and solution of linear systems involving P_h.

Note the delicate trade-off between fast convergence and the computational costs associated with the application of the preconditioning operator P_h^{-1}.

Preconditioning (cont.)

If the preconditioned matrix $P_h^{-1}A_h$ is close to the identity I_h in some norm, then convergence of a Krylov method will typically be quite fast. Indeed, the preconditioned matrix would have all its eigenvalues near 1 and would be nearly normal.

Another favorable situation would be to choose P_h such that $P_h^{-1}A_h$ has minimum polynomial of low degree, since the degree of this polynomial is an upper bound for the dimension of the Krylov subspace generated by $P_h^{-1}A_h$ and $P_h^{-1}r_0$.

However, these properties are difficult to achieve in practice.

Preconditioning (cont.)

Definition: Two families of SPD matrices $\{A_h\}$, $\{P_h\}$ are said to be spectrally equivalent if there exist positive constants α and β, independent of h, such that

$$\alpha \leq \frac{\langle A_h x, x \rangle}{\langle P_h x, x \rangle} \leq \beta, \quad \forall x \neq 0.$$

Note that this is equivalent to requiring that the eigenvalues of the preconditioned matrices $P_h^{-1} A_h$ are all contained in the interval $[\alpha, \beta]$, for all $h > 0$.

In other terms, the condition numbers $\kappa_2(P_h^{-1} A_h)$ are uniformly bounded:

$$\kappa_2(P_h^{-1} A_h) \leq \frac{\beta}{\alpha}, \quad \forall h.$$

This guaranteees the h-independent convergence of Krylov methods like CG or MINRES. The smaller the ratio $\frac{\beta}{\alpha}$, the faster the convergence.

Preconditioning (cont.)

The key to extending this result to general (nonsymmetric) problems is the notion of field-of-values equivalence of two families $\{A_h\}$, $\{P_h\}$.

Def.: Let $a_h(\cdot,\cdot)$ be positive definite, symmetric bilinear forms, having coercivity and continuity constants independent of h. Then $\{A_h\}$, $\{P_h\}$ are said to be a_h-field-of-values-equivalent if there exist positive constants α and β, independent of h, s. t.

$$\alpha \leq \frac{a_h(P_h^{-1}A_h\,x, x)}{a_h(x,x)} \leq \beta, \quad \forall x \neq 0.$$

Note that f.o.v.-equivalence reduces to spectral equivalence when $\{A_h\}$, $\{P_h\}$ are SPD and $a_h(u,v)$ is the standard inner product.

If $\{A_h\}$, $\{P_h\}$ are f.o.v-equivalent, the f.o.v's of the matrices $P_h^{-1}A_h$ are all contained in a compact region of the open right half-plane $Re(z) > 0$ independent of h. Therefore, GMRES will converge at an h-independent rate.

Preconditioning (cont.)

Mesh-independent preconditioners for solving the (steady) Oseen problem were first developed around 2000 by Elman and by Kay, Loghin and Wathen.

Unfortunately, these preconditioners are fairly sensitive to the viscosity; indeed, for small ν the imaginary part of the eigenvalues of the preconditioned matrices grows like $O(\nu^{-1})$.

In the following table we show iteration counts for solving two 2D Oseen problems, one easy (constant convective term) and the other harder (rotating vortex wind). An inf-sup stable finite element pair is used.

Preconditioning (cont.)

mesh size h	viscosity ν				
	1	0.1	0.01	0.001	0.0001
constant wind					
1/16	12	16	24	34	80
1/32	10	16	24	28	92
1/64	10	14	24	32	66
1/128	10	12	26	36	58
rotating vortex					
1/16	8	12	40	188	–
1/32	8	12	40	378	–
1/64	6	12	40	> 400	–
1/128	6	10	44	> 400	–

Results for Kay-Loghin-Wathen preconditioner.
(Stopping criterion: $\|\mathbf{b} - \mathcal{A}\mathbf{x}_k\|_2 < 10^{-6}\|\mathbf{b}\|_2$).

Augmented Lagrangian formulation

Consider the augmented Lagrangian formulation given by

$$\begin{pmatrix} A + \gamma B^T W^{-1} B & B^T \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{3}$$

where $\gamma > 0$ and W is SPD. Note that this is equivalent to the original problem, since $Bu = 0$. It can be interpreted as a form of "grad-div stabilization" of the Navier–Stokes equations. The addition of a SPD term to the (1,1)-block tends to move part of the spectrum to the right.

Letting $A_\gamma := A + \gamma B^T W^{-1} B$, we can rewrite (3) as

$$\begin{pmatrix} A_\gamma & B^T \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \quad \text{or} \quad \widehat{\mathcal{A}}\,\mathbf{x} = \hat{\mathbf{b}}. \tag{4}$$

Augmented Lagrangian formulation (cont.)

The following block triangular preconditioner for (4)

$$\mathcal{P} = \begin{pmatrix} A_\gamma & B^T \\ O & \widehat{S} \end{pmatrix}, \quad \widehat{S}^{-1} = -\nu \widehat{M}_p^{-1} - \gamma W^{-1}. \tag{5}$$

was first proposed in M. B. and M. A. Olshanskii, *SIAM J. Sci. Comput.*, 28 (2006), pp. 2095–2113.

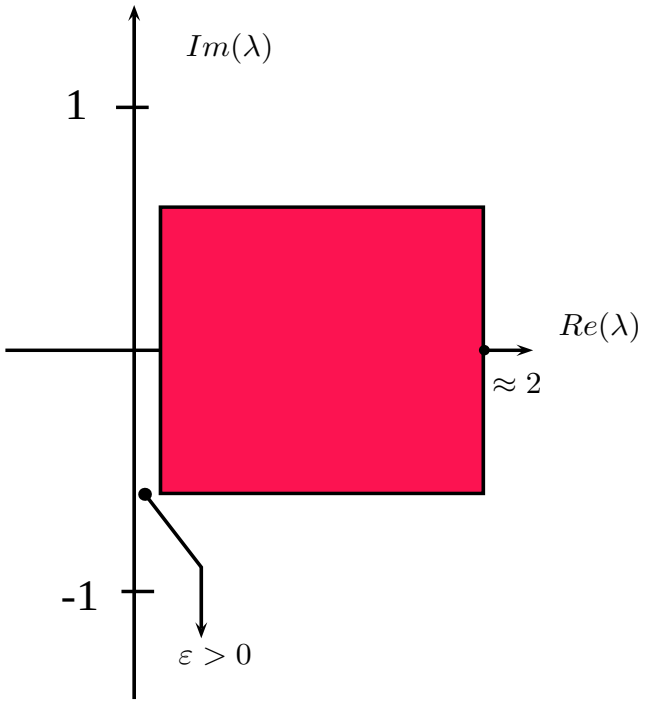
Here $\widehat{M}_p$ is any SPD matrix that is spectrally equivalent to the pressure mass matrix M_p (which is the Gramian associated with the basis functions for the pressure space Q_h: $(M_p)_{ij} = \langle \phi_i, \phi_j \rangle$).

For example, we can take $\widehat{M}_p = \mathrm{diag}(M_p)$ (Wathen).

Application of the action of the preconditioner $\mathcal{P}^{-1}$ at each step of GMRES requires applying $\widehat{S}^{-1} = -\nu \widehat{M}_p^{-1} - \gamma W^{-1}$ (easy!) and solving a sparse linear system with matrix A_γ (non-trivial!).

Convergence analysis

Theorem (B./Olshanskii, 2006): Let $W = M_p$, $n = \dim(\mathbf{V}_h)$, and $m = \dim(Q_h)$. The matrix $\mathcal{P}^{-1}\widehat{\mathcal{A}}$ has the eigenvalue $\lambda = 1$ of multiplicity n; all the remaining m eigenvalues are contained in a rectangle in the right half plane with sides independent of the mesh size h, and bounded away from 0. Moreover, for $\gamma = O(\nu^{-1})$ the rectangle does not depend on ν. When $\gamma \to \infty$, all the eigenvalues tend to 1.



Convergence analysis (cont.)

A stronger (and much more difficult) result is the following.

Theorem (B./Olshanskii, 2011): Let $\gamma = \|(BA^{-1}B^T)^{-1}M_p\|_{M_p}$. If $\nu < 1$, the residual norms in GMRES with the AL preconditioner satisfy

$$\|\hat{\mathbf{b}} - \widehat{\mathcal{A}}\mathbf{x}_k\| \leq q^k \|\hat{\mathbf{b}} - \widehat{\mathcal{A}}\mathbf{x}_0\|,$$

where $q < 1$ is independent of problem parameters h, ν (and Δt in the unsteady case).

This theorem is proved by constructing a suitable bilinear form $a_h(\cdot,\cdot)$ such that $\mathcal{P}$ and $\widehat{\mathcal{A}}$ are a_h-f.o.v-equivalent for all h, ν and Δt. This implies the above uniform convergence result for GMRES.

Convergence analysis (cont.)

Numerical experiments confirm that GMRES combined with the AL preconditioner is both h- and ν-independent, provided that the action of $\mathcal{P}^{-1}$ on a vector is computed exactly, or at least to high accuracy. Often using $\gamma = 1$ is sufficient.

A more practical option is to apply the preconditioner inexactly (which can be done in $O(n)$ work). Under appropriate conditions, the resulting preconditioner is still h-independent, and only mildly sensitive to the value of ν. However, a more careful choice of γ is now necessary. Local Fourier analysis can be used here.

M. B., M. A. Olshanskii, and Z. Wang, *Int. J. Numer. Methods Fluids*, 66 (2011), 486–508.

M. B. and Z. Wang, *SIAM J. Sci. Comput.*, 33 (2011), 2761–2784.

M. B. and M. A. Olshanskii, *SIAM J. Numer. Anal.*, 49 (2011), 770–788.

Numerical examples

mesh size h	viscosity ν				
	1.	0.1	0.01	10^{-3}	10^{-4}
	parameter γ				
	1.	1.	1.	0.1	0.02
constant wind					
1/16	6	6	7	8	24
1/32	7	6	10	8	22
1/64	7	6	8	7	19
1/128	7	6	9	9	18
rotating vortex					
1/16	6	6	7	13	25
1/32	5	6	9	11	32
1/64	4	5	10	11	37
1/128	4	4	10	12	34

Results for inexact Augmented Lagrangian preconditioner.
($\hat{A}_\gamma^{-1}$ is one W(1,1)-cycle of multigrid.)

Eigenvalues of preconditioned matrices

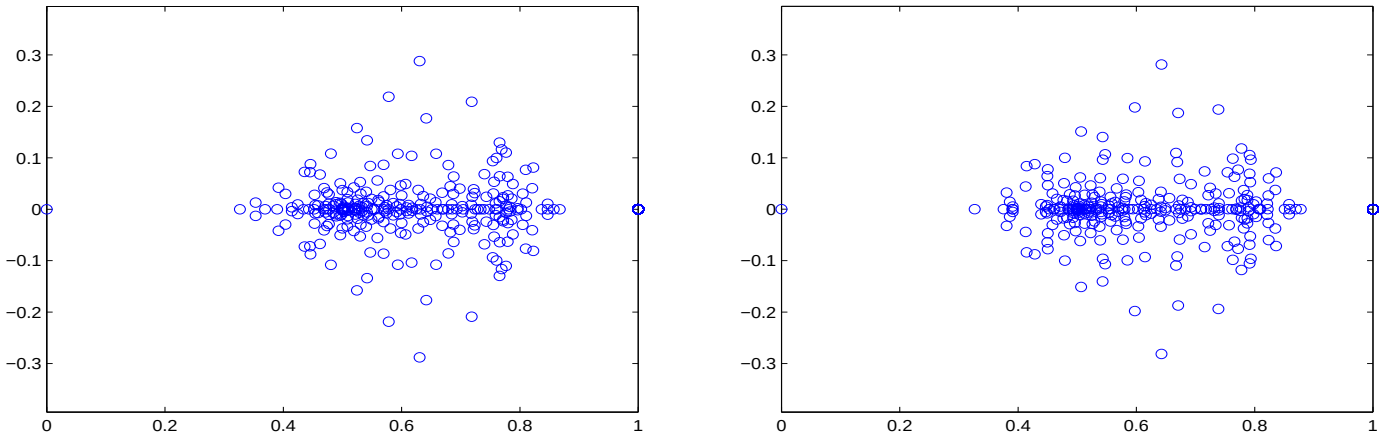


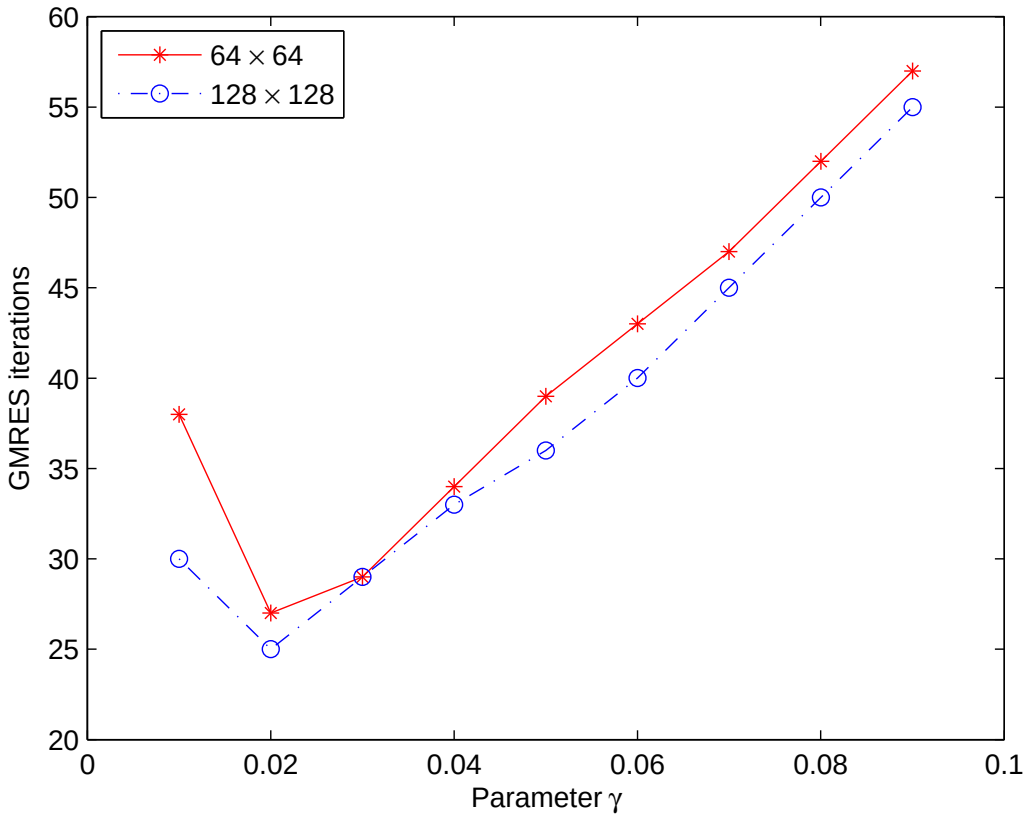
Figure: Plots of the eigenvalues of the preconditioned Oseen matrix (lid driven cavity, Q2-Q1, 32×32 uniform grid, $\nu = 0.01$). Left: with optimal γ. Right: with γ chosen by local Fourier analysis.

The two values of γ are very close: 0.050 vs. 0.056.

The eigenvalue $\lambda = 1$ has multiplicity n (for all γ).

Iteration counts with various values of γ

Figure: GMRES iterations with modified AL preconditioner (driven cavity, Q2-Q1, two different grids, $\nu = 0.001$)



Extensions and parallel performance

Very recently, P. Farrell et al. have presented results with a 3D parallel implementation of the AL preconditioner. Tests on standard benchmark problems with up to more than 10^9 DoFs showed excellent (weak) scalability going from 48 to 24576 MPI processes, with scaling efficiencies around 80%.

Furthermore, J. Moulin et al. have also reported on a parallel implementation of the (modified) AL preconditioner based on open source packages (FreeFEM, PETSc, SLEPc) and observed 80% parallel efficiency in the solution of 3D linear stability analysis problems (flow around plates), on 256 to 2048 processes.

P. Farrell, L. Mitchell and F. Wechsung, *SIAM J. Sci. Comput.*, 41 (2019), pp. A3073–A3096.

J. Moulin, P. Jolivet and O. Marquet, *Computer Meth. Appl. Mech. Engrg.*, 351 (2019), pp. 718–743.

Dimensional Splitting preconditioner

When Picard linearization is used, the coefficient matrix A (in 2D) ca be written as

$$\mathcal{A} = \begin{pmatrix} A_1 & 0 & B_1^T \\ 0 & A_2 & B_2^T \\ -B_1 & -B_2 & 0 \end{pmatrix},$$

with $A_1 \in \mathbb{R}^{n_1 \times n_1}$, $A_2 \in \mathbb{R}^{n_2 \times n_2}$ and $B_i \in \mathbb{R}^{m \times n_i}$, $i = 1, 2$.
The dimensional splitting of A is given by

$$\mathcal{A} = \begin{pmatrix} A_1 & 0 & B_1^T \\ 0 & 0 & 0 \\ -B_1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & A_2 & B_2^T \\ 0 & -B_2 & 0 \end{pmatrix} = H_1 + H_2.$$

Dimensional Splitting preconditioner (cont.)

The DS preconditioner is defined as $\mathcal{P} = \frac{1}{\alpha}(H_1 + \alpha I)(H_2 + \alpha I)$, or

$$\mathcal{P} = \frac{1}{\alpha} \begin{pmatrix} A_1 + \alpha I & 0 & B_1^T \\ 0 & \alpha I & 0 \\ -B_1 & 0 & \alpha I \end{pmatrix} \begin{pmatrix} \alpha I & 0 & 0 \\ 0 & A_2 + \alpha I & B_2^T \\ 0 & -B_2 & \alpha I \end{pmatrix}$$

$$= \begin{pmatrix} \alpha I + A_1 & -\frac{1}{\alpha} B_1^T B_2 & B_1^T \\ 0 & \alpha I + A_2 & B_2^T \\ -B_1 & -B_2 & \alpha I \end{pmatrix}.$$

Here $\alpha > 0$ is a parameter.

Dimensional Splitting preconditioner

- The difference between the preconditioner and the coefficient matrix is

$$\mathcal{P} - \mathcal{A} = \begin{pmatrix} \alpha I & -\frac{1}{\alpha} B_1^T B_2 & 0 \\ 0 & \alpha I & 0 \\ 0 & 0 & \alpha I \end{pmatrix}$$

- As $\alpha \to 0$, the diagonal entries vanish, but the off-diagonal block blows up
- Choice of α requires a trade-off
- Proposed in M. B. and X.-P. Guo, *Appl. Numer. Math.*, 61 (2011), pp. 66–76.

RDF preconditioner

- An improved variant of the DS preconditioner may be constructed as follows:

$$\mathcal{M} = \frac{1}{\alpha} \begin{pmatrix} A_1 & 0 & B_1^T \\ 0 & \alpha I & 0 \\ -B_1 & 0 & \alpha I \end{pmatrix} \begin{pmatrix} \alpha I & 0 & 0 \\ 0 & A_2 & B_2^T \\ 0 & -B_2 & \alpha I \end{pmatrix}$$

$$= \begin{pmatrix} A_1 & -\frac{1}{\alpha} B_1^T B_2 & B_1^T \\ 0 & A_2 & B_2^T \\ -B_1 & -B_2 & \alpha I \end{pmatrix}.$$

- This is the Relaxed Dimensional Factorization preconditioner (RDF).
- Proposed in M. B., M. Ng, Q. Niu and Z. Wang, *J. Comput. Phys.*, 230 (2011), pp. 6185–6202.

RDF preconditioner (cont.)

- The difference between this preconditioner and the coefficient matrix is

$$\mathcal{M} - \mathcal{A} = \begin{pmatrix} 0 & -\frac{1}{\alpha} B_1^T B_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \alpha I \end{pmatrix}$$

- This suggests a better performance of RDF, since more blocks are zero and the remaining nonzero blocks are the same as with DS.

Eigenvalues of the RDF preconditioned matrix

Theorem 1. The preconditioned matrix $\mathcal{A}\mathcal{M}^{-1}$ has an eigenvalue at $\lambda = 1$ with multiplicity $n_1 + n_2$. The remaining m eigenvalues are the eigenvalues μ_i of the matrix

$$Z_\alpha = \alpha^{-1}(S_1 + S_2) - 2\alpha^{-2} S_1 S_2,$$

where

$$S_1 = B_1(A_1 + \alpha^{-1} B_1^T B_1)^{-1} B_1^T$$

and

$$S_2 = B_2(A_2 + \alpha^{-1} B_2^T B_2)^{-1} B_2^T.$$

Eigenvalues of the RDF preconditioned matrix (cont.)

Theorem 2. The eigenvalues μ_i of Z_α are of the form

$$\mu_i = \frac{\alpha\lambda_i}{1+\alpha\lambda_i}, \quad i = 1:m,$$

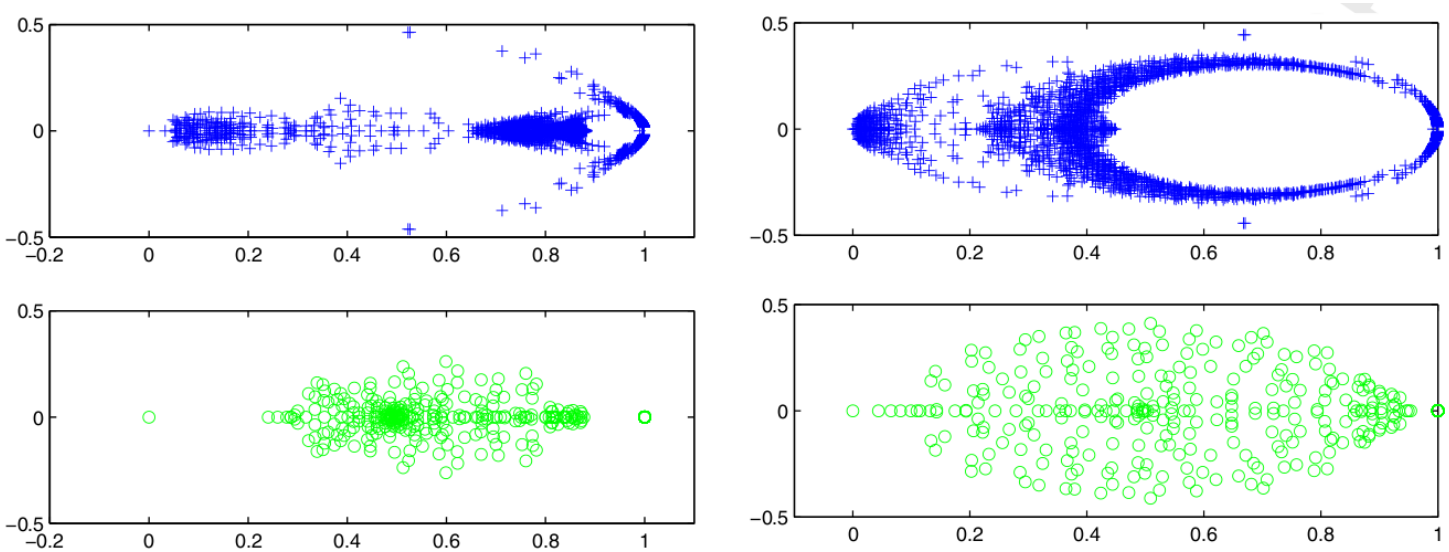
where the λ_i's satisfy the generalized eigenproblem

$$BA^{-1}B^T\phi_i = \lambda_i(\alpha^2 I + \hat{S}_1\hat{S}_2)\phi_i,$$

with $\hat{S}_k = B_k A_k^{-1} B_k^T$ $(k = 1, 2)$.

Eigenvalues of the RDF preconditioned matrix (cont.)

Eigenvalues of the preconditioned Oseen matrix on a 32×32 grid with Q2-Q1 finite elements. Top: DS preconditioner, bottom: RDF preconditioner. Left: $\nu = 0.01$, Right: $\nu = 0.001$.



Eigenvalues of the preconditioned matrix (cont.)

- Theorem 2 can be used to estimate the magnitude of the eigenvalues $\lambda \neq 1$.
- It can be shown that they go to zero like $O(\alpha)$ for $\alpha \to 0^+$ and like $O(\alpha^{-1})$ for $\alpha \to \infty$.
- In practice, diagonal scaling is applied to $\mathcal{A}$ before forming the RDF preconditioner. This significantly improves performance.

Applying the RDF preconditioner

Let $\hat{A}_i := A_i + \alpha^{-1} B_i^T B_i$ $(i = 1, 2)$. The RDF preconditioner can be factored as

$$\mathcal{M} = \begin{pmatrix} A_1 & 0 & B_1^T/\alpha \\ 0 & I & 0 \\ -B_1 & 0 & I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & A_2 & B_2^T \\ 0 & -B_2 & \alpha I \end{pmatrix} =$$

$$\begin{pmatrix} I & 0 & B_1^T/\alpha \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} \hat{A}_1 & 0 & 0 \\ 0 & I & 0 \\ -B_1 & 0 & I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & \hat{A}_2 & B_2^T \\ 0 & 0 & \alpha I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -B_2/\alpha & I \end{pmatrix}$$

so only subsystems involving $\hat{A}_1$ and $\hat{A}_2$ need to be solved. Moreover, these solves can be done inexactly.

Note: 3D case is analogous.

Estimating the optimal α

- This is done using Fourier analysis
- We make the usual simplifying assumptions:
  - the PDE problem has constant coefficients
  - the problem is defined on the unit square/cube with periodic boundary conditions
  - the grid is uniform
  - the matrices A_1, A_2, B_1, B_2 are all the same size and commute

Estimating the optimal α (cont.)

- Under these assumptions, A_1, A_2, B_1, B_2 are all diagonalized by the discrete Fourier transform; we denote their generic eigenvalues by a_1, a_2, b_1, b_2.
- From this we obtain that Z_α is also diagonalized by the discrete Fourier transform and the eigenvalues of Z_α are given by

$$\lambda(\alpha) = (s_1 + s_2)/\alpha - 2s_1 s_2/\alpha^2$$

$$\text{with} \quad s_1 = \frac{b_1^2}{a_1 + b_1^2/\alpha} \text{ and } s_2 = \frac{b_2^2}{a_2 + b_2^2/\alpha}.$$

Estimating the optimal α (cont.)

- For a finite difference discretization, we have

$$a_1 = a_2 = \nu(2 - e^{i2\pi h\theta} - e^{i2\pi h\theta}) + h(e^{i2\pi h\theta} - e^{-i2\pi h\theta}), \quad \theta = 1 : K$$

and

$$b_1 = b_2 = h(1 - e^{-i2\pi h\theta}), \quad \theta = 1 : K.$$

- Following Theorem 1, we try to maximize clustering of the eigenvalues of Z_α around 1.
- To this end, we find the α that minimizes the average distance of the eigenvalues of Z_α from 1. Note that this is an off-line computation.
- Numerical experiments show that performance is not overly sensitive w.r.t. α

Numerical experiments on model problems

- Mostly on a steady 2D lid-driven cavity discretized by Q2-Q1 or Q2-P1 finite elements using the MATLAB package `IFISS`
- Viscosity values $\nu = 0.1$, $\nu = 0.01$, $\nu = 0.001$
- Experiments are performed on 16×16, 32×32, 64×64 and 128×128 (uniform and stretched) grids
- The preconditioner is applied as a right preconditioner with restarted GMRES and maximum subspace dimension 20
- The linear solver stops when the relative residual drops below 10^{-6}

Leaky lid-driven cavity

Top: $\nu = 0.1$, middle: $\nu = 0.01$, bottom: $\nu = 0.001$:

Grid	DS optimal		RDF optimal		RDF FA estimate	
	α_{opt}	its	α_{opt}	its	α_F	its
16×16	0.03	14	0.05	11	0.07	12
32×32	0.009	14	0.01	12	0.025	12
64×64	0.002	14	0.005	11	0.008	12
128×128	0.0005	14	0.001	11	0.002	12

Grid	DS optimal		RDF optimal		RDF FA estimate	
	α_{opt}	its	α_{opt}	its	α_F	its
16×16	0.17	20	0.30	12	0.12	16
32×32	0.06	22	0.10	12	0.057	15
64×64	0.015	21	0.025	11	0.024	11
128×128	0.005	19	0.007	10	0.01	11

Grid	DS optimal		RDF optimal		RDF FA estimate	
	α_{opt}	its	α_{opt}	its	α_F	its
16×16	0.40	35	0.50	22	0.13	52
32×32	0.18	38	0.20	31	0.063	60
64×64	0.07	37	0.05	27	0.032	31
128×128	0.02	35	0.03	24	0.015	28

Numerical results with LifeV

- Finite elements library written in C++ (80'000 lines)
- LGPL license
- Used in the Mathcard European project (http://mathcard.eu).

`LifeV` relies on several external libraries:

- `ParMetis`/`Metis` for parallel mesh partitioning;
- `Trilinos` (10.8) for matrix and vector parallel distribution, for parallel solvers, and for parallel preconditioners;
- `Boost`, `SuiteSparse` (`UMFPACK`), `HDF5`.

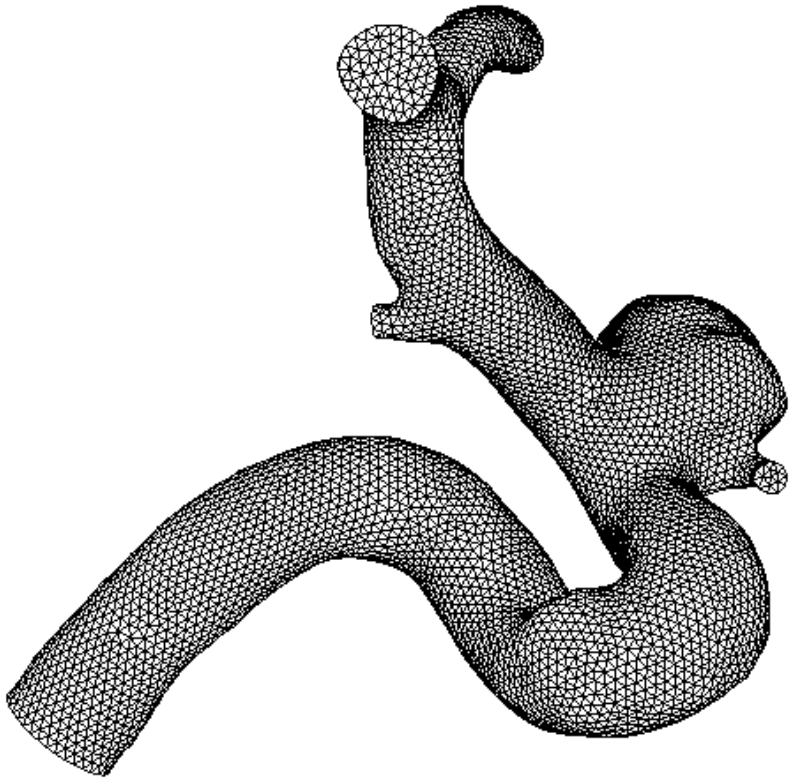
Numerical results with LifeV (cont.)

cadmos
center for advanced modeling science

The simulations were run on the CADMOS IBM Blue Gene/P at EPFL, Lausanne, Switzerland.

4 racks, one row, wired as a 16x16x16 3D torus
4096 quad-core nodes, PowerPC 450, 850 MHz
Energy efficient, water cooled
56 Tflops peak, 46 Tflops LINPACK
16 TB of memory (4 GB per compute node)
1 PB of disk space, GPFS parallel file system
OS Linux SuSE SLES 10

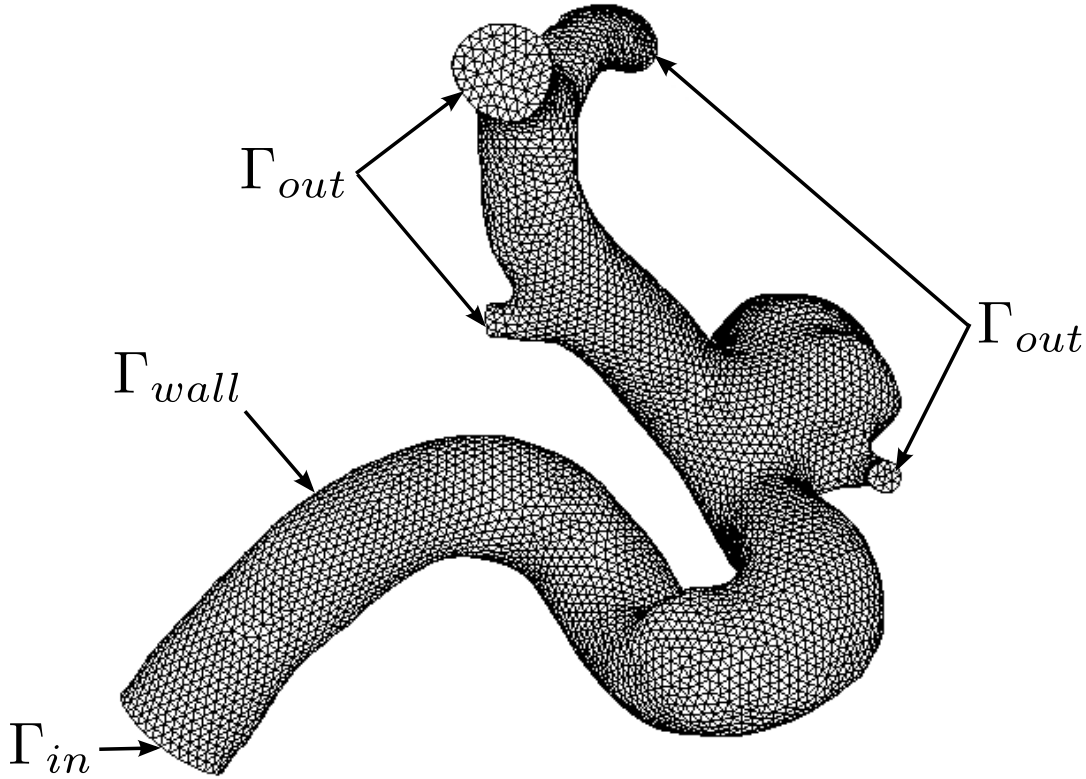
Hemodynamic problem (simulation of cerebral aneurysm)



$\mathbf{P}_2$-$\mathbf{P}_1$ finite elements (tetrahedral mesh)

Mesh	Velocity DoFs	Pressure DoFs	h_{min}	h_{max}	h_{av}
Medium	4,557,963	199,031	0.005	0.051	0.018
Fine	35,604,675	1,519,321	0.0026	0.0277	0.0097

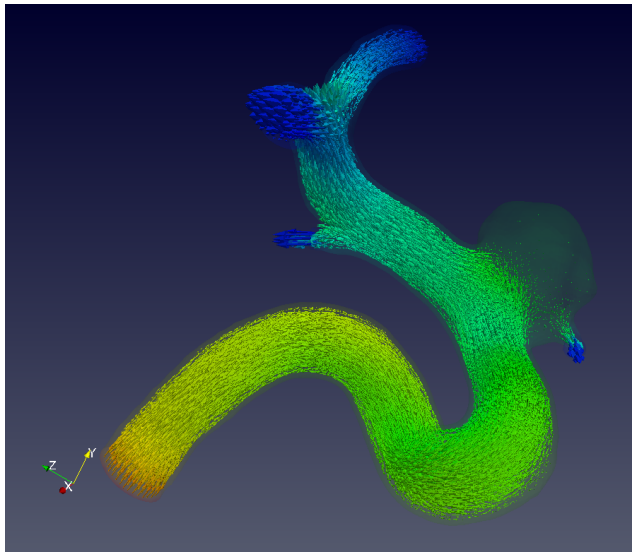
Hemodynamic problem (cont.)

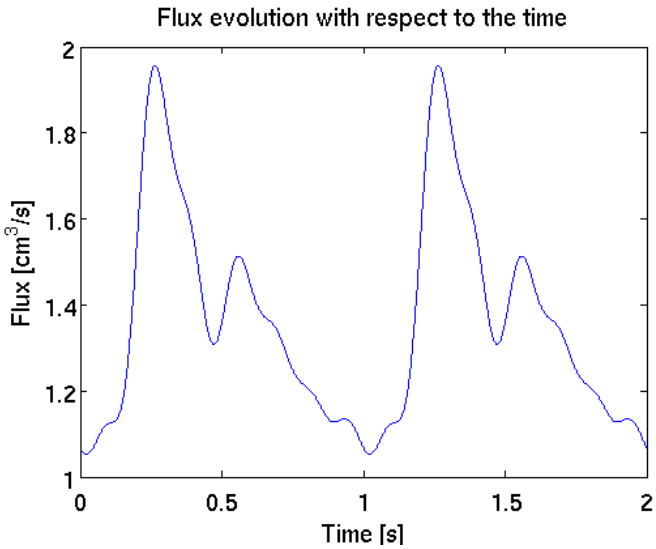


Boundary conditions:

$$\mathbf{u} = 0 \quad \text{on} \quad \Gamma_{wall}, \qquad \mathbf{u} = \varphi_{flux}\mathbf{n} \quad \text{on} \quad \Gamma_{in}, \qquad \nu\frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} = 0 \quad \text{on} \quad \Gamma_{out}.$$

Hemodynamic problem (cont.)





We impose the following inflow:

$$\varphi_{flux}(t) = a_0 + \sum_{i=1}^{7} a_k \cos\left(\frac{2\pi kt}{T}\right) + b_k \sin\left(\frac{2\pi kt}{T}\right)$$

Baek, Jayaraman, Richardson, Karniadakis. Flow instability and wall shear stress variation in intracranial aneurysms. *J R Soc Interface*, 2010.

Hemodynamic problem (cont.)

Other parameter settings:

- Unsteady problem, $\Delta t = 10^{-3}$
- Blood viscosity: $\nu = 0.035$ cm$^2/s$
- Nonlinear relative residual tolerance: 10^{-6}
- Four Picard iterations needed
- Outer (F)GMRES relative residual tolerance: 10^{-6}
- Inner GMRES relative residual tolerance: 0.05
- Inner GMRES sub-iterations preconditioned with `ML` (Trilinos)
- Relaxed Dimensional Factorization (RDF) preconditioner

M. B., S. Deparis, G. Grandperrin, and A. Quarteroni, *Comput. Methods Appl. Mech. Engrg.*, 300 (2016), pp. 129–145.

Hemodynamic problem (cont.)

Mesh	Cores	Average iteration count	Time
Medium	128	24.5	208.0
	256	23.3	37.8
	512	23.2	17.5
	1024	23.2	8.6
	2048	23.3	5.7
Fine	1024	23.2	106.4
	2048	22.3	44.2
	4096	21.6	20.2
	8192	21.7	11.1

Table: Aneurysm test case: preconditioned (F)GMRES iterations

Conclusions

- When implemented in combination with multigrid and domain decomposition methods, the Augmented Lagrangian approach results in robust and scalable preconditioners
- Both stable and stabilized discretizations can be accommodated
- Stretched grids do not pose any difficulties to the AL approach
- Techniques to obtain good estimates of the optimal parameter γ have been developed
- AL approach is currently the state-of-the-art for low-viscosity steady problems
- Parallel code now being developed by groups in UK, France
- Effective for other problens like buoyancy driven flow, coupled Darcy–Stokes problem, etc.
- RDF approach also effective, but less developed so far

References

1. M. Benzi, G. H. Golub and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1–137.
2. M. Benzi and M. A. Olshanskii, *An augmented Lagrangian-based approach to the Oseen problem*, SIAM J. Sci. Comput., 28 (2006), pp. 2095–2113.
3. M. Benzi, M. A. Olshanskii and Z. Wang, *Modified augmented Lagrangian preconditioners for the incompressible Navier–Stokes equations*, Intern. J. Numer. Meth. Fluids, 66 (2011), pp. 6185–6202.
4. M. Benzi and Z. Wang, *Analysis of augmented Lagrangian-based preconditioners for the steady incompressible Navier–Stokes equations*, SIAM J. Sci. Comput., 33 (2011), pp. 2761–2784.
5. M. Benzi and M. A. Olshanskii, *Field-of-values analysis of augmented Lagrangian preconditioners for the linearized Navier–Stokes problem*, SIAM J. Numer. Anal., 49 (2011), pp. 770–788.
6. M. Benzi, S. Deparis, G. Grandperrin, and A. Quarteroni, *Parameter estimates for the Relaxed Dimensional Factorization preconditioner and applications to hemodynamics*, Comput. Methods Appl. Mech. Engrg., 300 (2016), pp. 129–145.