Preconditioning and the CG method in the context of solving PDEs

Zdeněk Strakoš
Nečas Center for Mathematical Modeling
Charles University in Prague and Czech Academy of Sciences
http://www.karlin.mff.cuni.cz/~strakos

Symposium in honour of Martin Gutknecht, Lausanne 2014
R. C. Kirby, SIREV (2010):

“We examine condition numbers, preconditioners and iterative methods for FEM discretization of coercive PDEs in the context of the solvability result, the Lax-Milgram lemma.

Moreover, useful insight is gained as to the relationship between Hilbert space and matrix condition numbers, and translating Hilbert space fixed point iterations into matrix computations provides new ways of motivating and explaining some classic iteration schemes. [ ... ] This paper is [ ... ] intending to bridge the functional analysis techniques common in finite elements and the linear algebra community.”
“The main focus will be on an abstract approach to the construction of preconditioners for symmetric linear systems in a Hilbert space setting [ ... ] The discussion of preconditioned Krylov space methods for the continuous systems will be a starting point for a corresponding discrete theory.

By using this characterization it can be established that the conjugate gradient method converges [ ... ] with a rate which can be bounded by the condition number [ ... ] However, if the operator has a few eigenvalues far away from the rest of the spectrum, then the estimate is not sharp. In fact, a few ‘bad eigenvalues’ will have almost no effect on the asymptotic convergence of the method.”

“To preserve sparsity, the arising system is normally solved using an iterative solution method, commonly a preconditioned conjugate gradient method [ ... ] the rate of convergence depends in general on a generalized condition number of the preconditioned operator [ ... ]

- if the two operators (original and preconditioner) are equivalent then the corresponding PCG method provides mesh independent linear convergence [ ... ]

- if the two operators (original and preconditioner) are compact-equivalent then the corresponding PCG method provides mesh independent superlinear convergence.”
R. Hiptmair, CMA (2006):

“There is a continuous operator equation posed in infinite-dimensional spaces that underlies the linear system of equations [... ] awareness of this connection is key to devising efficient solution strategies for the linear systems.

Operator preconditioning is a very general recipe [... ]. It is simple to apply, but may not be particularly efficient, because in case of the [condition number] bound of Theorem 2.1 is too large, the operator preconditioning offers no hint how to improve the preconditioner. Hence, operator preconditioner may often achieve [...] the much-vaunted mesh independence of the preconditioner, but it may not perform satisfactorily on a given mesh.”

“For a fixed $h$, using a preconditioning strategy based on an equivalent operator may not be superior to classical methods [ ... ] Equivalence alone is not sufficient for a good preconditioning strategy. One must also choose an equivalent operator for which the bound is small.

There is no flaw in the analysis, only a flaw in the conclusions drawn from the analysis [ ... ] asymptotic estimates ignore the constant multiplier. Methods with similar asymptotic work estimates may behave quite differently in practice.”
Functional analysis and iterative methods

Operator preconditioning:

Klawonn (1995, 1996); Arnold, Falk, and Winther (1997, 1997); Steinbach and Wendland (1998); Mc Lean and Tran (1997); Christiansen and Nédélec (2000, 2000); Powell and Silvester (2003); Elman, Silvester, and Wathen (2005); Hiptmair (2006); Axelsson and Karátson (2009); Mardal and Winther (2011); Kirby (2011); Zulehner (2011); Preconditioning Conference 2013, Oxford; ...

Related ideas can be found also in Faber, Manteuffel and Parter (1990) with references to D’Yakonov (1961) and Gunn(1964, 1965).

CG in Hilbert spaces:

Hayes (1954); Daniel (1967, 1967); ... , Fortuna (1979); Ernst (2000); Axelsson and Karatson (2002); Glowinski (2003); .... ; Zulehner (2011); Günnel, Herzog, and Sachs (2012); ... ; Vorobyev (1958, 1965)
Computational cost of finding sufficiently accurate approximation to the exact solution heavily depends on

- the underlying real world problem,
- the mathematical model,
- on its discretization.

Construction and analysis of computational algorithms should respect that.

Evaluation of accuracy and of the computational cost must take into account algebraic errors, including rounding errors.
Outline

1. Basic settings
2. CG in Hilbert spaces
3. Spectral theory and the moment problem formulation
4. Galerkin discretization and matrix CG
5. Algebraic preconditioning as the transformation of the discretization basis
6. Distribution of errors, a-posteriori analysis and stopping criteria
7. Conclusions
1 Basic setting

Let $V$ be a real infinite dimensional Hilbert space with the inner product $(\cdot, \cdot)_V : V \times V \rightarrow \mathbb{R}$, the associated norm $\| \cdot \|_V$, $V^\#$ be the dual space of bounded (continuous) linear functionals on $V$ with the duality pairing

$$\langle \cdot, \cdot \rangle : V^\# \times V \rightarrow \mathbb{R}.$$ 

For each $f \in V^\#$ there exists a unique $\tau f \in V$ such that

$$\langle f, v \rangle = (\tau f, v)_V \quad \text{for all} \quad v \in V.$$ 

In this way the inner product $(\cdot, \cdot)_V$ determines the Riesz map

$$\tau : V^\# \rightarrow V.$$
Let $a(\cdot, \cdot) = V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form. For a fixed $u \in V$ we can see $a(u, \cdot)$ as the bounded linear functional on $V$ and we can write it as

$$\mathcal{A}u \equiv a(u, \cdot) \in V^\#, \quad \text{i.e.},$$

$$\langle \mathcal{A}u, v \rangle = a(u, v) \quad \text{for all } v \in V.$$

This defines the bounded and coercive operator

$$\mathcal{A} : V \to V^\#, \quad \inf_{u \in V, \|u\|_V = 1} \langle \mathcal{A}u, u \rangle = \alpha > 0, \quad \|\mathcal{A}\| = C.$$

The Lax-Milgram theorem ensures that for any $b \in V^\#$ there exist a unique solution $x \in V$ of the following problem

$$a(x, v) = \langle b, v \rangle \quad \text{for all } v \in V.$$
1 Transformed weak formulation

Equivalently,
\[ \langle Ax - b, v \rangle = 0 \quad \text{for all } v \in V , \]
which can be written as the equation in \( V^\# \),
\[ Ax = b , \quad A : V \to V^\# , \quad x \in V , \quad b \in V^\# . \]

Using the Riesz map,
\[ (\tau Ax - \tau b, v)_V = 0 \quad \text{for all } v \in V . \]

Clearly, the Riesz map \( \tau \) can be interpreted as transformation of the original problem \( Ax = b \) in \( V^\# \) into the equation in \( V \)
\[ \tau Ax = \tau b , \quad \tau A : V \to V , \quad x \in V , \quad \tau b \in V , \]
which is commonly (and inaccurately) called preconditioning.
1 Optimal preconditioning for $a(\cdot, \cdot)$ symmetric

If $a(\cdot, \cdot)$ is (in addition) symmetric, then $\mathcal{A}$ is self adjoint wrt to the duality pairing. Consider the special choice of the inner product

$$(u, v)_V = (u, v)_a \equiv a(u, v) = \langle \mathcal{A}u, v \rangle \quad \text{for all } u, v \in V.$$ 

then the definition of the associated Riesz map $\tau_a$ gives

$$(u, v)_a = \langle \mathcal{A}u, v \rangle = (\tau_a(\mathcal{A}u), v)_a \quad \text{for all } u, v \in V.$$ 

Therefore

$$\tau_a(\mathcal{A}u) = u, \quad \text{for all } u \in V, \text{ i.e. } \quad \tau_a = \mathcal{A}^{-1} : V^\# \to V,$$

and $\tau A x = \tau b$ will reduce to $x = \mathcal{A}^{-1}b$. Therefore the motivation of preconditioning is to take, in general, the inner product $(\cdot, \cdot)_V$ as close as possible to $(\cdot, \cdot)_a$. **Computational cost tradeoff.**
2 CG in Hilbert spaces

As \( \tau A : V \to V \), we can form for \( g \in V \) the Krylov sequence

\[
g, \tau Ag, (\tau A)^2 g, \ldots \quad \text{in } V
\]

and define Krylov subspace methods in the Hilbert space operator setting. Here we will do CG. Our goal is to construct a method for solving the (operator) equation

\[
Ax = b, \quad x \in V, \quad b \in V^\#
\]

such that with \( r_0 = b - Ax_0 \in V^\# \) the approximations \( x_n \) to the solution \( x \), \( n = 1, 2, \ldots \) belong to the Krylov manifolds in \( V \)

\[
x_n \in x_0 + K_n(\tau A, \tau r_0) \\
\subseteq x_0 + \text{span}\{\tau r_0, \tau A(\tau r_0), (\tau A)^2(\tau r_0), \ldots, (\tau A)^{n-1}(\tau r_0)\}
\]
2 Optimality conditions

and minimize

$$\|x - x_n\|_a = \min_{z \in x_0 + K_n} \|x - z\|_a.$$  

That is equivalent to the Galerkin orthogonality condition

$$\langle b - Ax_n, w \rangle = \langle r_n, w \rangle = 0 \quad \text{for all} \quad w \in K_n \equiv K_n(\tau A, \tau r_0),$$

and to finding the minimum

$$F(x_n) = \min_{z \in x_0 + K_n} F(z), \quad F(z) = \frac{1}{2} \langle Az, z \rangle - \langle b, z \rangle, \quad z \in V.$$  

Starting from $p_0 = \tau r_0 \in V$, we will construct a sequence of direction vectors $p_0, p_1, \ldots$, and a sequence of scalars $\alpha_0, \alpha_1, \ldots$ such that

$$x_n = x_{n-1} + \alpha_{n-1}p_{n-1}, \quad n = 1, 2, \ldots$$
2 Local and global minimization

Here $\alpha_{n-1}$ ensures the minimization of $F(z)$ along the line $z(\alpha) = x_{n-1} + \alpha p_{n-1}$. If we moreover satisfy

$$K_n = \text{span}\{p_0, p_1, \ldots, p_{n-1}\} \quad \text{and} \quad p_i \perp a p_j, \ i \neq j,$$

then the one-dimensional line minimizations along the given $n$ individual lines will be equivalent to minimization of the functional $F(z)$ over the whole $n$ dimensional manifold $x_0 + K_n$. Then

$$x - x_0 = \sum_{\ell=0}^{n-1} \alpha_{\ell} p_{\ell} + (x - x_n)$$

represents the orthogonal expansion of the initial error $x - x_0$ along the direction vectors $p_0, \ldots, p_{n-1}$ with $(x - x_n) \perp a K_n$. 
2 Search vectors (Hackbusch (1994))

It is important to realize that the steepest descent direction in minimizing $F(z)$ depends on the inner product $(\cdot, \cdot)_V$. It minimizes the directional derivative,

$$\delta F(x_\ell; d_\ell) = \lim_{\nu \to 0} \frac{F(x_\ell + \nu d_\ell) - F(x_\ell)}{\nu} = -\langle r_\ell, d_\ell \rangle.$$ 

Here $r_\ell = b - A x_\ell \in V^\#$ while $d_\ell \in V$. In order to determine $d_\ell$, we use the Riesz map $\tau$ associated with the inner product $(\cdot, \cdot)_V$, which immediately gives

$$-\langle r_\ell, d_\ell \rangle = - (\tau r_\ell, d_\ell)_V \quad \text{and} \quad d_\ell = \tau r_\ell.$$

The choice

$$p_n = \tau r_n + \beta_n p_{n-1}$$

with the condition $p_n \perp_a p_{n-1}$ will naturally deliver the global orthogonality (and, consequently, the global minimization).
2 CG in Hilbert spaces

\[ r_0 = b - Ax_0 \in V^\#, \quad p_0 = \tau r_0 \in V \]

For \( n = 1, 2, \ldots, n_{\text{max}} \)

\[ \alpha_{n-1} = \frac{\langle r_{n-1}, \tau r_{n-1} \rangle}{\langle Ap_{n-1}, p_{n-1} \rangle} = \frac{(\tau r_{n-1}, \tau r_{n-1})_V}{(\tau Ap_{n-1}, p_{n-1})_V} \]

\[ x_n = x_{n-1} + \alpha_{n-1} p_{n-1}, \quad \text{stop when the stopping criterion is satisfied} \]

\[ r_n = r_{n-1} - \alpha_{n-1} Ap_{n-1} \]

\[ \beta_n = \frac{\langle r_n, \tau r_n \rangle}{\langle r_{n-1}, \tau r_{n-1} \rangle} = \frac{(\tau r_n, \tau r_n)_V}{(\tau r_{n-1}, \tau r_{n-1})_V} \]

\[ p_n = \tau r_n + \beta_n p_{n-1} \]

End
3 Spectral theory and the moment problem

Using the orthogonal projection $E_n$ onto $K_n$ with respect to the inner product $(\cdot, \cdot)_V$, consider the orthogonally restricted operator

$$\tau A_n : K_n \to K_n, \quad \tau A_n \equiv E_n (\tau A) E_n,$$

by formulating the following equalities

$$\tau A_n (\tau r_0) \equiv \tau A (\tau r_0),$$

$$(\tau A_n)^2 \tau r_0 = \tau A_n (\tau A (\tau r_0)) \equiv (\tau A)^2 \tau r_0,$$

$$\vdots$$

$$(\tau A_n)^{n-1} \tau r_0 = \tau A_n ((\tau A)^{n-2} \tau r_0) \equiv (\tau A)^{n-1} \tau r_0,$$

$$(\tau A_n)^n \tau r_0 = \tau A_n ((\tau A)^{n-1} \tau r_0) \equiv E_n (\tau A)^n \tau r_0.$$
3 Spectral theory and the moment problem

The $n$-dimensional approximation $\tau A_n$ of $\tau A$ matches the first $2n$ moments

$$
((\tau A_n)^{\ell} \tau r_0, \tau r_0)_V = ((\tau A)^{\ell} \tau r_0, \tau r_0)_V, \quad \ell = 0, 1, \ldots, 2n - 1.
$$

Denote symbolically $Q_n = (q_1, \ldots, q_n)$ a matrix composed of the columns $q_1, \ldots, q_n$ forming an orthonormal basis of $K_n$ determined by the Lanczos process

$$
\tau A Q_n = Q_n T_n + \delta_{n+1} q_{n+1} e_n^T
$$

with $q_1 = \tau r_0 / \|\tau r_0\|_V$. We get $(\tau A_n)^{\ell} = Q_n T_n^{\ell} Q_n^*$, $\ell = 0, 1, \ldots$ and the matching moments condition

$$
e_1^* T_n^{\ell} e_1 = q_1^* (\tau A)^{\ell} q_1, \quad \ell = 0, 1, \ldots, 2n - 1,
$$
3 Matrix CG representation of the first $n$ steps

\[
T_n = \begin{pmatrix}
\gamma_1 & \delta_2 \\
\delta_2 & \ddots & \ddots \\
\ddots & \ddots & \ddots \\
\ddots & \ddots & \ddots & \delta_n \\
\delta_n & \gamma_n \end{pmatrix}
\]

is the Jacobi matrix of the orthogonalization coefficients and the CG method is formulated by

\[
T_n y_n = \| \tau r_0 \|_V e_1, \quad x_n = x_0 + Q_n y_n, \quad x_n \in V.
\]
3 Spectral moment problem

Since $\tau A$ is bounded and self-adjoint, its spectral decomposition is written using the Riemann-Stieltjes integral as

$$
\tau A = \int_{\lambda_L}^{\lambda_U} \lambda \, d\mathcal{E}_\lambda,
$$

The spectral function $\mathcal{E}_\lambda$ of $\tau A$ represents a family of orthogonal projections which is

- non-decreasing, i.e., if $\mu > \nu$ , then the subspace onto which $\mathcal{E}_\mu$ projects contains the subspace into which $\mathcal{E}_\nu$ projects;
- $\mathcal{E}_{\lambda_L} = 0$, $\mathcal{E}_{\lambda_U} = I$;
- $\mathcal{E}_\lambda$ is right continuous, i.e. $\lim_{\lambda' \to \lambda^+} \mathcal{E}_{\lambda'} = \mathcal{E}_\lambda$.

The values of $\lambda$ where $\mathcal{E}_\lambda$ increases by jumps represent the eigenvalues of $\tau A$, $\tau A z = \lambda z$, $z \in V$. 

3 Spectral moment problem

For the (finite) Jacobi matrix $T_n$ we can analogously write

$$T_n = \sum_{j=1}^{n} \theta_{\ell}^{(n)} s_{\ell}^{(n)}, \quad \lambda_L < \theta_1^{(n)} < \theta_2^{(n)} < \cdots < \theta_n^{(n)} < \lambda_U,$$

and the operator moment problem turns into

$$\int_{\lambda_L}^{\lambda_U} \lambda^\ell \, d\omega(\lambda) = \sum_{j=1}^{n} \left\{ \theta_j^{(n)} \right\}^\ell \omega_j^{(n)}, \quad \ell = 0, 1, \ldots, 2n - 1,$$

where $d\omega(\lambda) = q_1^* dE_\lambda q_1$ represents the Riemann-Stieltjes distribution function associated with $\tau A$ and $q_1$. The distribution function $\omega^{(n)}(\lambda)$ approximates $\omega(\lambda)$ in the sense of the $n$th Gauss-Christoffel quadrature; Gauss (1814), Jacobi (1826), Christoffel (1858).
3 Spectral moment problem

Using $f(\lambda) = \lambda^{-1}$ gives the equation for the energy norm of the CG error

$$\int \lambda^{-1} d\omega(\lambda) = \sum_{i=1}^{n} \omega_i^{(n)} \left( \theta_i^{(n)} \right)^{-1} + \frac{\|x - x_n\|^2_a}{\|\tau r_0\|^2_V}$$

Condition number bounds should always be checked against this CG - Gauss-Christoffel quadrature equivalence.

This is of particular importance when the large outlying eigenvalues are present. Then we either use CG (short recurrences) and we cannot apply the commonly presented bounds assuming exact arithmetic, or we reorthogonalize, but then we do not use CG with short recurrences;

Consider an $N$-dimensional subspace $V_h \subset V$ with the duality pairing, the inner product and the Riesz map as above. Then the restriction to $V_h$ gives an approximation $x_h \in V_h$ to $x \in V$,

$$a(x_h, v) = \langle b, v \rangle \quad \text{for all } v \in V_h.$$ 

As above, the bilinear form $a(\cdot, \cdot) : V_h \times V_h \to \mathbb{R}$ defines the operator $A_h : V_h \to V_h^\#$ such that

$$\langle A_h x_h - b, v \rangle = 0 \quad \text{for all } v \in V_h.$$ 

With restricting $b$ to $V_h$, i.e. $\langle b_h, v \rangle \equiv \langle b, v \rangle$ for all $v \in V_h$, we get the operator form

$$A_h x_h = b_h, \quad x_h \in V_h, \quad b_h \in V_h^\#, \quad A_h : V_h \to V_h^\#.$$
Let \( \Phi_h = (\phi_1^{(h)}, \ldots, \phi_N^{(h)}) \) be the basis of \( V_h \), \( \Phi_h^\# = (\phi_1^{(h)}\# , \ldots , \phi_N^{(h)}\#) \) the basis of its dual \( V_h^{\#} \). Using the coordinates in \( \Phi_h \) and \( \Phi_h^\# \),

\[
\langle \phi^\#, v \rangle \rightarrow v^\ast z ,
\]

\[
(u, v)_V \rightarrow v^\ast M u , \quad (M_{ij}) = ((\phi_j, \phi_i)_V)_{i,j=1,\ldots,N} ,
\]

\[
\tau \rightarrow M^{-1} ,
\]

\[
A_h \rightarrow A , \quad (A_{ij}) = (a(\phi_j, \phi_i))_{i,j=1,\ldots,N} = (\langle A\phi_j, \phi_i \rangle)_{i,j=1,\ldots,N} ,
\]

\[
b \rightarrow b ,
\]

we get with \( x_n = \Phi_h x_n , \ p_n = \Phi_h p_n , \ r_n = \Phi_h^\# r_n \).
4 4 FEM Galerkin discretization and matrix CG

\[ r_0 = b - Ax_0, \quad \text{solve} \quad Mz_0 = r_0, \; p_0 = z_0 \]

For \( n = 1, \ldots, n_{\text{max}} \)

\[ \alpha_{n-1} = \frac{z_{n-1}^* r_{n-1}}{p_{n-1}^* Ap_{n-1}} \]
\[ x_n = x_{n-1} + \alpha_{n-1} p_{n-1}, \quad \text{stop when the stopping criterion is satisfied} \]
\[ r_n = r_{n-1} - \alpha_{n-1} Ap_{n-1} \]
\[ z_n = M^{-1} r_n, \quad \text{solve for} \; z_n \]
\[ \beta_n = \frac{z_n^* r_n}{z_{n-1}^* r_{n-1}} \]
\[ p_n = z_n + \beta_n p_{n-1} \]

End
Algebraic preconditioning with the preconditioner $M$ can be viewed as the unpreconditioned CG (with $M = I$) applied to

$$Bw = c$$

with

$$B = L_h^{-1}AL_h^{-*}, \quad c = L_h^{-1}b, \quad x = L_h^{-*}w, \quad M_h = L_hL_h^*.$$  

**Observation:**

The associated Hilbert space formulation of CG in $V_h$ corresponds to the transformation of the bases

$$\Phi_t = \Phi_h L_h^{-*}, \quad \Phi_t^\# = \Phi_h^\# L_h^*.$$
5 Algebraic preconditioning - main point

\[ B \equiv (B_{ij}) = \left( \langle A \phi_j^{(t)} , \phi_i^{(t)} \rangle \right)_{i,j=1,...,N} = (a(\phi_j^{(t)} , \phi_i^{(t)}))_{i,j=1,...,N} , \]

where

\[ \phi_{\ell}^{(t)} = \Phi_h (L_{-h}^* e_{\ell}) , \quad \ell = 1, \ldots, N \]

and the right hand side

\[ c = \Phi_h^# L_{h}^* b . \]

Please recall the hierarchical bases preconditioning

The locality of supports and sparsity is on purpose lost.
Knupp and Salari, 2003:

“There may be incomplete iterative convergence (IICE) or round-off-error that is polluting the results. If the code uses an iterative solver, then one must be sure that the iterative stopping criteria is sufficiently tight so that the numerical and discrete solutions are close to one another. Usually in order-verification tests, one sets the iterative stopping criterion to just above the level of machine precision to circumvent this possibility.”

In solving tough problems this can not be afforded.

How to measure the algebraic error?
6 Local discretisation and global computation

Discrete (piecewise polynomial) FEM approximation

\[ x_h = \Phi_h x_n. \]

- If the algebraic solution \( x_h \) is known exactly, then the accuracy of approximating the solution \( x \) by \( x_h \) over the given domain is fully determined by the local approximation property of the piecewise polynomial FEM space.

- However, apart from trivial cases, \( x_n \) that supply the global information is not known exactly. Then

\[
\begin{align*}
 x - x_h(n) &= x - x_h + x_h - x_h(n) \\
 &\quad \text{total error} \quad \text{discretisation error} \quad \text{algebraic error}
\end{align*}
\]
6 Local discretization, adaptivity
6 Global norm of the error (Poisson problem)

Using the Galerkin orthogonality, we get up to a small inaccuracy proportional to machine precision

$$\| \nabla (x - x_h^{(n)}) \|^2 = \| \nabla (x - x_h) \|^2 + \| \nabla (x_h - x_h^{(n)}) \|^2$$

$$= \| \nabla (x - x_h) \|^2 + \| x - x_n \|_A^2 .$$

What is the distribution of the algebraic error in the functional space ?
7 L-shape model, Papež, Liesen, S (2013)

Exact solution $x$ (left) and the discretisation error $x - x_h$ (right).
Algebraic error $x_h - x_h^{(n)}$ (left) and the total error $x - x_h^{(n)}$ (right). Here

$$\| \nabla (x - x_h) \| > 0.1 \| x - x_n \|_A.$$
8 Conclusions


“With the often noted tremendous increases in computer speed and memory, and with the less often acknowledged but equally powerful increases in algorithmic accuracy and efficiency, a natural question suggest itself. What are we doing with the new computer power? with the new GUI and other set-up advances? with the new algorithms? What *should* we do? ... Get the right answer.”

This requires to consider modelling, discretisation, analysis, and computation tightly coupled parts of a *single solution process* and to cautiously analyze all necessary simplifications.

*Analysis is much needed.*
References

- J. Málek and Z.S., Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs. SIAM Spotlight Series, SIAM (2014), in print
All the best, Martin!