Convergence Rate Estimates for the Conjugate Gradient Method

Igor Kaporin

Parallel Computing Lab.,
Dept. Applied Optimization Problems
A. A. Dorodnicyn Computing Center
Russian Academy of Sciences, Moscow, RUSSIA

The Rome-Moscow School of Matrix Methods
Moscow Part: September 10-15, 2012

Outline of the talk

- CG as the Optimum Krylov Subspace Method
- Spectral bound convergence rate estimates
- New estimates via the K-condition number
- The Preconditioned CG method
- Preconditioning via K-optimization

CG as the Optimum Krylov Subspace Method (1)

Consider a system of linear algebraic equations

$$Ax = b, \qquad x \in \mathsf{R}^n, \quad b \in \mathsf{R}^n, \qquad A^T = A > 0,$$

where A is large, sparse, and not well-conditioned.

The CG approximations x_k to the solution x of the linear system are constructed from the initail residual $r_0 = b - Ax_0$ in the form

$$x_k = x_0 + r_0\alpha_1^{(k)} + \ldots + A^{k-1}r_0\alpha_k^{(k)},$$

where the scalar coefficients are chosen such that $\{\alpha_1^{(k)}, \ldots, \alpha_k^{(k)}\} = \arg\min \|x - x_k\|_A$.

CG as the Optimum Krylov Subspace Method (2)

The CG algorithm is as follows:

$$
\begin{aligned}
& r_0 = b - Ax_0, \\
& p_0 = r_0; \\
& \textbf{for} \quad i = 0, 1, \ldots : \\
& \qquad \alpha_i = r_i^T r_i / p_i^T A p_i, \\
& \qquad x_{i+1} = x_i + p_i \alpha_i, \\
& \qquad r_{i+1} = r_i - A p_i \alpha_i, \\
& \qquad \beta_i = r_{i+1}^T r_{i+1} / r_i^T r_i, \\
& \qquad p_{i+1} = r_{i+1} + p_i \beta_i.
\end{aligned}
$$

Note that $r_k = b - Ax_k = \pi_k(A) r_0$.

CG as the Optimum Krylov Subspace Method (3)

Therefore, one has

$$x - x_k \equiv A^{-1}\pi_k(A)r_0, \qquad \pi_k(t) = 1 - \sum_{i=1}^{k} \alpha_k^{(i)} t^i,$$

and the solution is well approximated even for $k \ll n$, a sufficient condition for which is $\|\pi_k(A)\| \ll 1$.
Indeed, by the optimality of $\pi_k(A)$ one has

$$\|x - x_k\|_A = \|\pi_k(A)r_0\|_{A^{-1}} \le \|\pi_k(A)\|\|r_0\|_{A^{-1}}$$

$$= \|\pi_k(A)\|\|x - x_0\|_A \le \|\tilde{\pi}(A)\|\|x - x_0\|_A,$$

where $\tilde{\pi}_k(\cdot)$ is **any** polynomial such that

$$\deg \tilde{\pi}_k \le k, \qquad \tilde{\pi}_k(0) = 1.$$

CG as the Optimum Krylov Subspace Method (4)

Since A is SPD, it holds (e.g. by the spectral decomposition of A)

$$\|\tilde{\pi}_k(A)\| = \max_{i=1,\ldots,n} |\tilde{\pi}_k(\lambda_i)|,$$

where $\lambda_i = \lambda_i(A) > 0$ are the eigenvalues of A numbered in the nondecreasing order.

Using different particular choices of $\tilde{\pi}_k(\cdot)$, one can construct various CG convergence estimates of the type

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \max_{\lambda=\lambda_i} |\tilde{\pi}_k(\lambda)| \leq \varphi(\lambda_1, \ldots, \lambda_n)$$

the right-hand side of which is anyway dependent on the spectrum of A.

Spectral bound convergence rate estimates (1)

The well-known standard result follows from

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \max_{\lambda=\lambda_i} |\tilde{\pi}_k(\lambda)| \leq \max_{\lambda_1 \leq \lambda \leq \lambda_n} |\tilde{\pi}_k(\lambda)|,$$

where $\tilde{\pi}_k$ is expressed via a properly translated and scaled kth degree Chebyshev polynomial $\mathbf{T}_k(\cdot)$ of the 1st kind, which yields

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq 1 \Big/ \mathbf{T}_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right) < 2 \exp\left(-2k\sqrt{\frac{\lambda_1}{\lambda_n}}\right).$$

Recall that

$$\mathbf{T}_k(z) = \frac{1}{2}\left(\left(z + \sqrt{z^2 - 1}\right)^k + \left(z - \sqrt{z^2 - 1}\right)^k\right)$$

Spectral bound convergence rate estimates (2)

This estimate readily yields an iteration number bound for the CG to converge with the relative precision ε:

$$k \leq \left\lceil \frac{1}{2}\sqrt{\frac{\lambda_n}{\lambda_1}} \log \frac{2}{\varepsilon} \right\rceil$$

Note that in practice, an *a priori* estimation of the spectral bounds (more precisely, a control of their ratio) may be impossible to perform.

However, *a posteriori* estimates of λ_1 and λ_n are readily available from α_i and β_i generated by the CG algorithm.

Spectral bound convergence rate estimates (3)

Thus we have the estimate

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} < 2 \exp\left(\frac{-2k}{\sqrt{\mathbf{C}(A)}}\right),$$

where

$$\mathbf{C}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1$$

is the *spectral condition number* of an SPD matrix A. In this case, for some $\sigma > 0$,

$$A \to \sigma I \qquad \text{iff} \qquad \mathbf{C}(A) \to 1$$

New estimates via the K-condition number (1)

In the presented theory, a key role plays the matrix functional $\mathbf{K}(A)$ defined as

$$\mathbf{K}(A) = (n^{-1}\mathrm{trace}A)^n / \det A$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n \Big/ \prod_{i=1}^{n}\lambda_i = \mathbf{K}(\lambda_1,\ldots,\lambda_n)$$

The latter holds by the well-known property

$$\mathrm{trace}A = \sum_{i=1}^{n}\lambda_i, \qquad \det A = \prod_{i=1}^{n}\lambda_i.$$

New estimates via the K-condition number (2)

In a complete analogy with the spectral condition number, for the K-condition number it holds ($\sigma > 0$)

$$A \to \sigma I \qquad \text{iff} \qquad \mathbf{K}(A) \to 1$$

when A is an SPD matrix. This is nothing but the *Arithmertic-Geometric Mean (AGM)* inequality written for $\{\lambda_1, \ldots, \lambda_n\}$.

First we demonstrate an elementary proof of a (rather rough but instructive) estimate for the decrease of the $\|x - x_k\|_A$ in the CG method.

New estimates via the K-condition number (3)

Theorem 1. [Kaporin,Axelsson'00] *Let A be SPD. Then for any even k satisfying*

$$2\log_2 \mathbf{K}(A) < k < n$$

it holds

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} < \left(\mathbf{K}(A)^{2/k} - 1\right)^{k/2}.$$

Note: this bound is not precise.

New estimates via the K-condition number (4)

Proof. Let $k = 2m$. Using the general estimate with

$$\tilde{\pi}_k(t) = \prod_{i=1}^{m} (1 - t/\lambda_i)(1 - t/\lambda_{n+1-i}),$$

one readily gets

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \max_{i=1,\ldots,n} |\tilde{\pi}_k(\lambda_i)| \leq \max_{i=m+1,\ldots,n-m} |\tilde{\pi}_k(\lambda_i)|$$

$$\leq \prod_{i=1}^{m} \max_{\lambda_{m+1} \leq t \leq \lambda_{n-m}} |(1 - t/\lambda_i)(1 - t/\lambda_{n+1-i})|$$

$$= \prod_{i=1}^{m} \left(\mathbf{K}(\lambda_i, \lambda_{n+1-i}) - 1 \right).$$

New estimates via the K-condition number (5)

Using an obvious consequence of the AGM inequality,

$$\left(\prod_{i=1}^{m}\theta_i\right)^{1/m}+\left(\prod_{i=1}^{m}(1-\theta_i)\right)^{1/m}\leq 1,\qquad 0<\theta_i<1,$$

with $\theta_i = 1/\mathbf{K}(\lambda_i, \lambda_{n+1-i})$ we obtain

$$\frac{\|x-x_k\|_A}{\|x-x_0\|_A}\leq\left(\left(\prod_{i=1}^{m}\mathbf{K}(\lambda_i,\lambda_{n+1-i})\right)^{1/m}-1\right)^{m},$$

and it only remains to prove that

$$\prod_{i=1}^{m}\mathbf{K}(\lambda_i,\lambda_{n+1-i})\leq\mathbf{K}(A)$$

New estimates via the K-condition number (6)

The latter also follows the AGM inequality:

$$\mathbf{K}(A)\left(\prod_{i=1}^{n}\lambda_i\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n$$

$$=\left(\frac{1}{n}\left(\sum_{i=1}^{m}\frac{\lambda_i+\lambda_{n+1-i}}{2}+\sum_{i=1}^{m}\frac{\lambda_i+\lambda_{n+1-i}}{2}+\sum_{i=m+1}^{n-m}\lambda_i\right)\right)^n$$

$$\geq\left(\prod_{i=1}^{m}\left(\frac{\lambda_i+\lambda_{n+1-i}}{2}\right)^2\right)\left(\prod_{i=m+1}^{n-m}\lambda_i\right)$$

$$=\left(\prod_{i=1}^{m}\mathbf{K}(\lambda_i,\lambda_{n+1-i})\right)\left(\prod_{i=1}^{n}\lambda_i\right)$$

Q.E.D.

New estimates via the K-condition number (7)

The corresponding iteration number bound is

$$k < \left\lceil \frac{4 \log \mathbf{K}(A) + 3 \log(\varepsilon^{-1})}{\log\left(4 + \log_{\mathbf{K}(A)}(\varepsilon^{-1})\right)} \right\rceil .$$

For a given $0 < \varepsilon \ll 1$, this condition yields

$$\|x - x_k\|_A \leq \varepsilon \|x - x_0\|_A$$

This can be shown setting $t = \mathbf{K}(A)^{2/k}$ in the inequality

$$t - 1 \leq (\sigma - 1)^{\sigma - 1} \left(\frac{t}{\sigma}\right)^{\sigma}, \qquad \sigma > 1, \qquad t > 1.$$

Hence we establish a CG iteration number bound whicn grows sublinearly with respect to $\log \frac{1}{\varepsilon}$.

New estimates via the K-condition number (8)

Concerning the roughness of the above estimate: an unimprovable (best possible) bound was found, but it relates to reduction of another error norm:

Theorem 2. [Kaporin'92,94] *Let A be SPD. Then for any even k satisfying*

$$1 \leq k < n$$

it holds

$$\frac{\|r_k\|}{\|r_0\|} \leq \left(\mathbf{K}(A)^{1/k} - 1\right)^{k/2}.$$

This means nearly 2 times reduction of the above iteration number bound.

New estimates via the K-condition number (9)

An example is shown in the Figure, where we consider a matrix A of the order $n = 50$ with the prescribed eigenvalues $\lambda_{n+1-j}(M) = n^2 + 1 - j^2$.

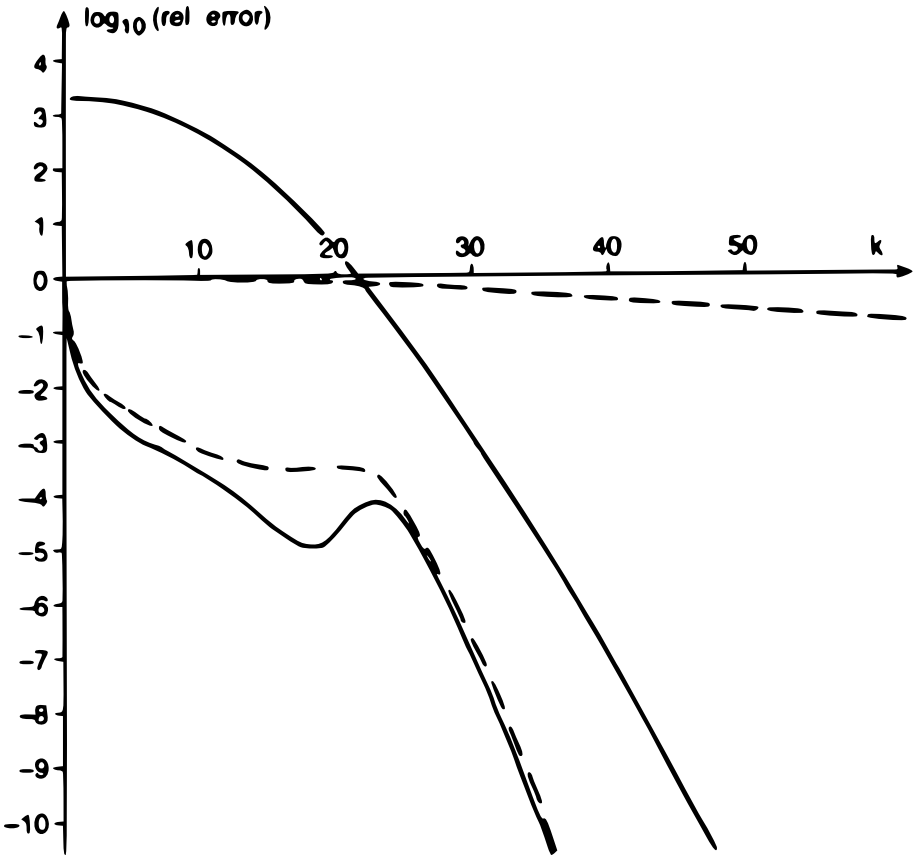
With the dashed line the A^{-1}-norm of the residual and its upper bound via **C**(A) are shown,

while the solid line corresponds to the Euclidean norm of the residual and its estimate via **K**(A).

It is quite clear that the solid lines behave much more similar to each other than the two dashed lines.

Note that the considered example has isolated smaller eigenvalues and clusterized largest eigenvalues, which is exactly the class of eigenvalue distributions which we prefere to deal with.

New estimates via the K-condition number (10)



The A^{-1}- and H-norms of the residuals and their upper bounds vrs. the CG iteration number

The Preconditioned CG method (1)

Obviously, if some condition number of the matrix A is very large, then the CG method may require a huge number of iterations to converge (especially in computer arithmetics) - despite of its optimality.

To overcome this drawback, a very simple but powerful idea of **preconditioning** is applied.

Namely, let us substitute $x = G^T y$, where $\det G \neq 0$, and solve the *preconditioned* linear system

$$GAG^T y = Gb,$$

by the same CG method. This time, we will have the *preconditioned* matrix $M = GAG^T$ instead of A in every formula above!

The Preconditioned CG method (2)

Using appropriate substitutions, and denoting $H = G^T G$, we readily obtain the *preconditioned* CG method:

$$
\begin{aligned}
& r_0 = b - Ax_0, \\
& p_0 = Hr_0; \\
& \textbf{for} \quad i = 0, 1, \ldots : \\
& \qquad \alpha_i = r_i^T H r_i / p_i^T A p_i, \\
& \qquad x_{i+1} = x_i + p_i \alpha_i, \\
& \qquad r_{i+1} = r_i - A p_i \alpha_i, \\
& \qquad \beta_i = r_{i+1}^T H r_{i+1} / r_i^T H r_i, \\
& \qquad p_{i+1} = H r_{i+1} + p_i \beta_i
\end{aligned}
$$

Here we have $r_k = b - Ax_k = \pi_k(AH) r_0$.

The Preconditioned CG method (3)

Both the above iteration number bounds are (nearly) proportional to

$$\sqrt{\mathbf{C}(HA)}$$

in the standard (spectral bounds based) CG theory, or

$$\log \mathbf{K}(HA)$$

when using the nth power of the arithmetic-to-geometric mean ratio for the spectrum of HA (i.e., the K-condition number) for the same purposes.
Hence, the central problem in the PCG theory is:

using easy-to-multiply by a vector matrices H, reduce the condition number of HA - as low as possible

The Preconditioned CG method (3)

The $\mathbf{K}$-optimization vrs. the $\mathbf{C}$-optimization:

- the (second) estimate via $\mathbf{K}$ is as sharp as the one via $\mathbf{C}$;
- the estimates via $\mathbf{K}$ reflect the superlinear convergence of CG, while the one via $\mathbf{C}$ does not;
- generally, the $\mathbf{K}(HA)$-optimization can be feasible, while $\mathbf{C}(HA)$-optimization may not;
- $\mathbf{K}(HA)$-optimization tends to clusterize the spectrum of HA near the largest e.v.'s of HA, while $\mathbf{C}(HA)$-optimization may not.

Preconditioning via K-optimization (1)

A simplest possible example is the preconditioning by a symmetric diagonal scaling (i.e., $G = D$)

$$M = DAD.$$

It can be shown that the K-optimality is attained when

$$D = (\mathrm{Diag}(A))^{-1/2}.$$

That is,

$$D = \arg\min_{\mathcal{D}} \mathbf{K}(\mathcal{D}A\mathcal{D}),$$

where the minimum is taken over the set of all SPD diagonal matrices $\mathcal{D}$.

Preconditioning via K-optimization (2)

Indeed, denoting $\delta_i = (A)_{ii}(\mathcal{D})_{ii}^2$, one has

$$\mathbf{K}(\mathcal{D}A\mathcal{D}) = \frac{\left(n^{-1}\sum_{i=1}^n \delta_i\right)^n}{\det DAD \prod_{i=1}^n \delta_i} \geq \frac{1}{\det DAD} = \mathbf{K}(DAD),$$

where D is defined above. The AGM inequality shows that the equality in the latter estimate is attained iff $\delta_i = \delta > 0$.

Finally, under a natural restriction $\delta = 1$, one gets the required equality $\mathcal{D} = D$.

Preconditioning via K-optimization (3)

It must be stressed that this *(so-called Jacobi)* scaling *is not optimum* in the sense of the spectral condition number $\mathbf{C}(DAD)$ for arbitrary SPD matrices. The exception are some special cases, e.g., *consistently ordered* matrices.

An example: the Toeplitz SPD tridiagonal matrix T=[-1, 2, -1] is consistently ordered and therefore is $\mathbf{C}$-optimally scaled. Therefore, the inverse of it has the same property since for any SPD A it always holds $\mathrm{C}(A) = \mathrm{C}(A^{-1})$.

At the same time, T^{-1} has non-constant diagonal and therefore is not $\mathbf{K}$-optimally scaled.

Preconditioning via K-optimization (4)

In a similar way, it can be shown that the Block Jacobi preconditioning for an SPD matrix is K-optimum over the set of all block-diagonal matrices with prescribed structure.

However, we prefer to consider this structure later as a particular case of a more general construction.

References

- **[Kaporin'92]** Explicitly preconditioned conjugate gradient method for the solution of nonsymmetric linear systems. Int. J. Computer Math., **40**, 169–187, 1992.

- **[Kaporin'94]** New convergence results and preconditioning strategies for the conjugate gradient method. Numer. Linear Algebra with Appls., **1**, no.2, 179–210, 1994.

- **[Kaporin,Axelsson'00]** On the sublinear and superlinear rate of convergence of conjugate gradient methods. Numerical Algorithms, **25**, 1–22, 2000.