J. Numer. Math., Vol. 0, No. 0, pp. 1–16 (2006)
 Prepared using jnm.sty *[Version: 02.02.2002 v1.2]*

Some notes on Krylov subspace methods and minimal residuals

E. Tyrtyshnikov*

Abstract — We give a concise introduction to some recent developments of the method of minimal residuals and tendencies in the design and analysis of good preconditioners.

Keywords: Method of minimal residuals, iterative methods, preconditioners, spectral clusters.

1. INTRODUCTION

When solving a linear system $Ax = b$ by a direct method, one can be unsatisfied with accuracy or time or both. A general advice in this case is to use iterations. Appealing advantages read as follows:

- iterations can be *matrix-free*, and
- *time* strongly depends upon *accuracy*.

The ways for iterations and convergence analysis are many and in this short note we do not indend to mention even those that otherwise deserve it (many good references can be found in [4,10,24]). Instead, we focus on one idea which is probably simplest, most free from constraints and most popular in applications. A "geometrical" implementation of this idea was probably first presented by G. Marchuk and Yu. Kuznetsov [12].

Let us build up a sequence of subspaces

$$L_1 \subset L_2 \subset \cdots \subset L_m, \quad \dim L_k = k, \quad k = 1, \ldots, m, \tag{1.1}$$

in the n-dimensional space and define $x_k \in L_k$ as the minimizer of the *residual functional*

$$R(x) = ||b - Ax||_2^2 \tag{1.2}$$

on the manifold $x_0 + L_k$. If m is sufficiently large then for some k we obtain the exact solution $x_* = x_k$ and quit. Here are still the two issues to be specified:

This work was supported by the Russian Foundation for Basic Research grant No. 05-01-00721

*Institute of Numerical Mathematics of the Russian Academy of Sciences, Gubkina Street, 8, Moscow 119991, Russia

- How do we move to L_{k+1} from L_k?
- How do we minimize $R(x)$ on L_k?

Take an arbitrary initial vector x_0, set $r_0 = b - Ax_0$ and consider the following *Krylov subspaces*:

$$L_k = \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}, \quad k = 1, \ldots . \tag{1.3}$$

For unification, let L_0 be zero space.

It is easy to prove that $L_k = L_{k+1}$ if and only if $x_k = x_*$. Let m be the minimal index such that $L_k = L_{k+1}$. Then we consider L_k only for $k \leqslant m$.

In computations with the Krylov subspaces, we construct some bases other than those in the definition (1.3). Let

$$L_k = \operatorname{span}\{p_1, \ldots, p_k\}. \tag{1.4}$$

In fact, we can choose any vectors p_k such that

$$p_k \in L_k, \quad p_k \notin L_{k-1}. \tag{1.5}$$

On the kth iteration we do not change $p_1, \ldots, p_{k-1}$. In practical algorithms, we take a *probe* vector p_k satisfying (1.5) and then modify it to fit one of the following options:

(1) keep $p_1, \ldots, p_k$ orthogonal;

(2) keep $Ap_1, \ldots, Ap_k$ orthogonal.

Option (1) is associated with a somewhat "algebraic" approach to minimization of the residual $r_k = b - Ax_k$. Option (2) provides a "geometric" approach to the same minimization: by the theorem of Pyphagoras, the minimality of length of r_k means that

$$r_k \perp AL_k. \tag{1.6}$$

This is a framework for several algorithms: GMRES [14] (algebraic approach); ORTHOMIN, ORTHODIR [15](geometrical approach). Without the latter names, the geometrical approach was probably first described in [12]. Quite naturally, the framework is called the *method of minimal residuals*. Details of implementation are easy to find elsewhere (perhaps the reader can suggest one more version).

An important case is when A is Hermitian positive definite. Using the same subspaces L_k, now we can pick up $x_k \in L_k$ as the minimizer of the *error functional*

$$E(x) = \frac{1}{2}(A(x - x_*), x - x_*). \tag{1.7}$$

Eventually we come up with the famous *method of conjugate gradients*.

One may rejoin that what we discuss pretends to be iterations in a formal way but is actually a direct method (as it quits with the exact solution). However, a common practice is to check how close x_k comes to x_* (at least by the residual) and stop when a prescribed accuracy is reached. To this end, we need some "convergence analysis" and error estimates.

2. ERROR ESTIMATES

Minimization properties allow us to estimate how $||r_k||_2$ decreases as k grows. However, we *have to* impose some additional assumptions on A: it is not difficult to produce an example of A for which the minimal residual method yields the same values of $||r_k||_2$ till the final step with zero residual.

Despite the above observation, we need not be too pessimistic. Since the kth residual is minimal on the search space, we have

$$\begin{aligned} ||r_k||_2 &\leqslant \min_{\alpha} ||r_{k-1} - \alpha A r_{k-1}||_2 \\ &= |\alpha|^2 (Ar_{k-1}, Ar_{k-1}) - 2\mathrm{re}\,(\alpha (Ar_{k-1}, r_{k-1})) + (r_{k-1}, r_{k-1}). \end{aligned}$$

Assume that α is real. Then the minimum of the right-hand side is attained at

$$\alpha = \frac{\mathrm{re}(Ar_{k-1}, r_{k-1})}{(Ar_{k-1}, Ar_{k-1})},$$

which implies that

$$||r_k||_1 \leqslant \sqrt{1 - \frac{(\mathrm{re}(Ar_{k-1}, r_{k-1})/(r_{k-1}, r_{k-1}))^2}{(Ar_{k-1}, Ar_{k-1})/(r_{k-1}, r_{k-1})}}\, ||r_{k-1}||_2. \tag{2.1}$$

Now consider the following *strong ellipticity* (*coercitivity*) assumption:

$$|\mathrm{re}(Ax, x)| \geqslant \tau (x, x) \quad \forall\, x, \quad \tau > 0, \tag{2.2}$$

and note that

$$(Ar_{k-1}, Ar_{k-1})/(r_{k-1}, r_{k-1}) \leqslant ||A||_2^2.$$

Consequently, we arrive from (2.1) at the *Elman estimate* [6]

$$||r_k||_2 \leqslant \sqrt{1 - \frac{\tau^2}{||A||_2^2}}\, ||r_{k-1}||_2. \tag{2.3}$$

This very form of estimate appeared for the first time in [9] (however, in the context of symmetric positive definite matrices).

The coercitivity condition (2.2) means, in fact, that

$$\mathrm{re}(Ax,x) \geqslant \tau(x,x) \quad \forall\, x, \quad \tau > 0, \tag{2.4}$$

or, alternatively,

$$\mathrm{re}(Ax,x) \leqslant -\tau(x,x) \quad \forall\, x, \quad \tau > 0. \tag{2.5}$$

If (2.2) is fulfilled, then the inequalities of (2.4) and (2.5) may not occur both for some different values of x. This is an obvious corollary from the Toeplitz–Hausdorff theorem on the convexity of the *field of values*, defined as the set of values (Ax,x) for all x on the unit sphere $||x||_2 = 1$.

Let $S(\tau,\sigma)$, $\sigma \geqslant \tau > 0$, be a domain on the complex plane with complex numbers ζ with the two properties:

$$\mathrm{re}\zeta \geqslant \tau, \quad |\zeta| \leqslant \sigma. \tag{2.6}$$

As is readily seen, all the eigenvalues of A lie in $S(\tau,\sigma)$ with $\sigma = ||A||_2$. However, if $\lambda \in S(\tau, ||A||_2)$ for any eigenvalue λ of A, then this alone is not enough to guarantee (2.3). To make it true we should add some extra condition: for example, it is valid if A is a normal matrix.

Remark that (2.3) is an estimate for a one-dimensional minimization along r_{k-1} and should be rough as it does not reflect advantages of the Krylov spaces. Shortly we can see that, at least for normal matrices and many others, a way more optimistic estimate should hold.

2.1. Using polynomials

Since we use the Krylov subspaces, r_k is of the form

$$r_k = f_k(A) r_0, \tag{2.7}$$

where $f_k(\zeta)$ is a degree k polynomial such that $f_k(0) = 1$. For shortness, let us write $f_k \in \mathscr{F}_k$, where $\mathscr{F}_k$ is the set of such polynomials. The minimality of length of r_k can be interpreted in the following way:

$$||r_k||_2 \;\leqslant\; ||f_k(A)||_2 \, ||r_0||_2 \quad \forall\, f_k \in \mathscr{F}_k. \tag{2.8}$$

Let Γ_δ be a boundary of $S(\tau-\delta, \sigma+\delta)$ for some δ such that $0 < \delta < \tau$. Then, for any polynomial $f_k(\zeta)$, one can express $f_k(A)$ via the so-called *resolvent of A*, a matrix function of the form $(A-\zeta I)^{-1}$ of complex variable ζ , as follows:

$$f_k(A) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_\delta} (A-\zeta I)^{-1} \, f_k(\zeta) \, d\zeta. \tag{2.9}$$

Hence,

$$\frac{||r_k||_2}{||r_0||_2} \;\leqslant\; \frac{|\Gamma_\delta|}{2\pi} \;\; T_k(\Gamma_\delta) \;\; R(A,\Gamma_\delta), \tag{2.10}$$

where

$$T_k(\Gamma_\delta) \equiv \min_{f_k \in \mathscr{F}_k} \max_{\zeta \in \Gamma_\delta} |f_k(\zeta)|, \tag{2.11}$$

$$R(A,\Gamma_\delta) \equiv \max_{\zeta \in \Gamma_\delta} ||(A-\zeta I)^{-1}||_2, \tag{2.12}$$

and $|\Gamma_\delta|$ is the length of Γ_δ. Thus, we can separately estimate $|\Gamma_\delta|$ (an easy matter), $T_k(\Gamma_\delta)$ (a fabulous problem of complex analysis and function approximation theory related to the name of P. L. Chebyshev), and $R(A,\Gamma_\delta)$ (difficult in general but quite feasible in many cases of particular interest).

2.2. Resolvent and field of values

Assume that $S(\tau,\sigma)$ contains the *field of values* of A. Denote the latter by $\Phi(A)$. In this case we can prove that (cf. [5])

$$R(A,\Gamma_\delta) \leqslant \frac{1}{\delta}. \tag{2.13}$$

The proof is based on the following general inequality:

$$||(A-\zeta I)^{-1}||_2 \leqslant \frac{1}{d(\zeta,\Phi(A))}, \tag{2.14}$$

where

$$d(\zeta,\Phi(A)) = \min_{\xi \in \Phi} ||\zeta - \xi||_2. \tag{2.15}$$

There exist unit-length vectors x and y with the property

$$(A-\zeta I)^{-1} y = ||(A-\zeta I)^{-1}||_2 \, x.$$

It follows that

$$|(Ax,x) - \zeta(x,x)| = |((A-\zeta I)x,x)| = \frac{|(y,x)|}{||(A-\zeta I)^{-1}||_2} \leqslant \frac{1}{||(A-\zeta I)^{-1}||_2},$$

which obviously proves (2.14) and, hence, (2.13).

2.3. Normal matrices

In the case of normal matrix A, (2.13) emanates directly from the claim that all the eigenvalus of A are located in $S(\tau,\sigma)$. We know that the field of values of a normal matrix coincides with the convex hull of its eigenvalues and take into account that $S(\tau,\sigma)$ is a convex set.

Moreover, if A is normal then it possesses an orthonormal system of eigenvectors. Consequently,

$$A = Q\Lambda Q^{-1},$$

where Q is a unitary matrix ($Q^*Q = I$) and Λ is a diagonal matrix of the eigenvaues of A. Observe that

$$||f_k(A)||_2 = ||Qf_k(\Lambda)Q^{-1}||_2 = ||f_k(\Lambda)||_2,$$

because the spectral norm is unitarily invariant. Assume that all the eigenvalues of A belong to a domain S with boundary Γ. Then,

$$\frac{||r_k||_2}{||r_0||_2} \leqslant \min_{f_k \in \mathscr{F}_k} \max_{\zeta \in S} |f_k(\zeta)| = \min_{f_k \in \mathscr{F}_k} \max_{\zeta \in \Gamma} |f_k(\zeta)|. \tag{2.16}$$

2.4. Asymptotic convergence rate

Consider the following ansatz for the convergence estimate:

$$||r_k||_2 \leqslant cq^k ||r_0||_2, \quad 0 < q < 1, \quad c > 0. \tag{2.17}$$

This can be easily deduced from (2.3) with a specific value for q. However, that value of q is far from the true characteristic of asymptotic convergence rate. A smaller q (with a greater value of c, alas) can arise from a thorough study of the behaviour of T_k in (2.10).

We need not stick to a particular case of the curve Γ_δ and consider $T_k(\Gamma)$ for an arbitrary smooth (or piece-wise smooth) curve Γ that subdivides the complex plane into two open subdomains, one of which is bounded and another one unbounded; denote the latter by Ω. Consider a polynomial

$$p(z) = a\prod_{i=1}^{n}(z - z_i)$$

with the roots z_i inside the bounded domain and define a function

$$g_p(z) = \log |p(z)|^{1/n}.$$

This is a real-valued function of a complex variable $z \in \Omega$. This is also a function of real-valued variables x and y such that $z = x + \mathrm{i}y$. It is easy to verify that

$$\Delta g_p(z) = 0 \ \ \forall\, z \in \Omega,$$

where Δ is the Laplace operator

$$\Delta g_p(z) = \frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2}, \quad z = x + \mathrm{i}y.$$

Besides that, at the infinity

$$g_p(z) = \log |z| + \gamma + o(1), \quad \gamma = \log |a|^{1/n} = \text{const}.$$

It is less trivial to prove that for sufficiently large n the roots can be chose so that $g_p(z) \geqslant 0$ on Γ and, moreover, $g_p(z) \approx 0$ for $z \in \Gamma$ [8]. The proof involves the following boundary value problem:

$$\begin{aligned} &\text{(a)}\ \ \Delta g(z) = 0, && z \in \Omega, \\ &\text{(b)}\ \ g(z) = 0, && z \in \Gamma, \\ &\text{(c)}\ \ g(z) - \log|z| - \gamma = o(1), && z \to \infty. \end{aligned} \tag{2.18}$$

Then, polynomials $p(z)$ are constructed so that $g_p(z) \approx g(z)$ for $z \in \Gamma$. Eventually this leads to the following

Theorem 2.1. *Let $0 \in \Omega$ and $g(z)$ satisfy (2.18). Then*

$$\limsup_{k\to\infty} (T_k)^{1/k} \;=\; \mathrm{e}^{-g(0)}. \tag{2.19}$$

In several particular cases $g(0)$ can be obtained analytically:

- If Γ is a circle of radius $0 < r < a$ with center at a on the real axis. then
$$g(0) = \log \frac{a}{r}. \tag{2.20}$$
- If Γ is an ellipse with center at a on the real axis and half-axes $0 < r_1 < a$ and r_2, then
$$g(0) = \log \frac{\sqrt{a^2 - r_1^2 + r_2^2} + a}{r_1 + r_2}. \tag{2.21}$$

2.5. Recent improvement of the Elman estimate

Theorem 2.2. [2] *Let A be a matrix satisfying (2.4), and let*

$$\sin\beta = \sqrt{1 - \frac{\tau^2}{||A||_2^2}}, \quad \beta \in (0, \pi/2),$$

be the convergence rate factor in the Elman estimate (2.3). Then for the kth relative residual of the method of minimal residuals (GMRES) we have

$$\frac{||r_k||}{||r_0||} \leqslant (2 + 2/\sqrt{3})\,(2 + \gamma_\beta)\,\gamma_\beta^k, \qquad k \geqslant 1, \tag{2.22}$$

where

$$\gamma_\beta := 2\sin\left(\frac{\beta}{4 - 2\beta/\pi}\right) < \sin(\beta).$$

3. SUPERLINEAR CONVERGENCE AND SPECTRAL CLUSTERS

Assume that all the eigenvalues of A lie in a small disc. Then, according to (2.20) and Theorem 2.1, the asymptotic convergence rate for the residuals is the faster the smaller the disc is. However, this fastness might look a bit too abstract as A is of some finite order, n, and we never take k greater than n.

All the same, a small spot of eigenvalues suggests an idea of *spectral cluster*. It does not make any rigorous sense for one matrix and applies only to a sequence of matrices of increasing orders. Thus, let A_n be of order n and consider A_n for a strictly increasing sequence of n. Of course, A_n are associated with one common application.

Let $\mathscr{K}$ be a set on the complex plane and $\mathscr{K}_\varepsilon$ be a larger set (ε-extension of $\mathscr{K}$) including any z such that

$$\inf_{\zeta\in\mathscr{K}} |z-\zeta| \leqslant \varepsilon.$$

Let $\gamma_n(\varepsilon)$ count how many eigenvalues of A_n fall inside $\mathscr{K}_\varepsilon$. We call $\mathscr{K}$ an *eigenvalue cluster* for A_n if

$$\gamma_n(\varepsilon) = o(n) \quad \forall\, \varepsilon > 0. \tag{3.1}$$

A cluster is called *proper* if

$$\gamma_n(\varepsilon) = O(1) \quad \forall\, \varepsilon > 0. \tag{3.2}$$

Now, we consider a sequence of linear systems with coefficient matrices A_n and apply the method of minimal residuals. These residuals certainly depend on n and, hence, c and q in (2.17) must depend on n.

In effect, a spectral cluster at $\mathscr{K}$ allows us to take $q = q(\varepsilon)$ as $T_k(\mathscr{K}_\varepsilon)$. The price we pay is a somewhat larger value of $c = c(\varepsilon, n)$. If a cluster consists of one point, then q can be made arbitrarily small. This is the so-called *superlinear convergence*.

Assume that m eigenvales of A_n are outside $\mathscr{K}$ and denote them by $\lambda_1, \dots, \lambda_m$. Then, for $k > m$, we can take

$$f_k(\zeta) = f_{k-m}(\zeta) \prod_{i=1}^{m} \left(1 - \frac{\zeta}{\lambda_i}\right)$$

and be sure that $f_k \in \mathscr{F}_k$. For simplicity, suppose that matrices A_n are normal. Then, obviuosly,

$$\frac{||r_k||_2}{||r_0||_2} \leqslant \min_{f_{k-m}\in\mathscr{F}_{k-m}} \max_{\zeta\in\mathscr{K}} |f_{k-m}(\zeta)| \left(1 + \frac{r_{\max}}{r_{\min}}\right)^m, \tag{3.3}$$

where r_{min} and r_{max} are the minimal and maximal moduli of the eigenvalues of A_n, respectively.

Roughly speaking, the iterations behave as if all the eigenvalues were in $\mathscr{K}$, the price for having m outsiders being m additional iterations. The latter observation is well known due to the semenal papers [1] and [18] where it was made for the case of

symmetric matrices. The definitions for general and proper clusters were introduced in [19] and then used and studied in many subsequent papers (cf. [16,17,22,23]).

If the convergence is slow for a particular problem, we can try to replace the problem with a better one with better convergence properties. The procedure itself is called *preconditioning*. Ususally it consists in getting from A_n to $A_n P_n^{-1}$ for a suitably chosen *preconditioner* P_n. Then, a good preconditioner has everything to do with the following properties:

(a) AP^{-1} is conditioned uniformly in n;

(b) AP^{-1} has an eigenvalue cluster at unity.

At least for Hermitian positive definite matrices and under some additional assumptions in the general case, property (a) indicates the *linear convergence* and (b) accounts for *superlinear convergence* (cf. [21,23]).

The existence of cluster is related to decompositions of the form [19]

$$A_n = P_n + R_n + E_n, \tag{3.4}$$

where R_n is of *small rank* and E_n is of *small norm*.

Theorem 3.1. [19] *Let A_n and P_n be two sequences such that for any $\varepsilon > 0$ there exists a decomporition (*3.4*) with*

$$||E_n||_F^2 = o(n), \quad \text{rank} R_n = o(n). \tag{3.5}$$

Then the singular values of A_n and P_n have the same clusters. If A_n and P_n are Hermitian then the eigenvalues of A_n and B_n have the same clusters as well.

Thus, as long as we have

$$A_n P_n^{-1} \;=\; I_n + R_n' + E_n'$$

with

$$||E_n'||_F^2 = o(n), \quad \text{rank} R_n' = o(n),$$

it means that this method of preconditioning provides a cluster at 1.

Historically, decompositions (3.4) were considered first in the context of Toeplitz matrices and circulant preconditioners and used, in effect, to prove superlinear convergence properties [3]. However, a new paradigm suggested in [13] is to use (3.4) directly for construction of P_n. It is shown in [13] that this approach results in the best of circulant preconditioners in comparison with all those discussed in the literature. A supporting approximation theory was recently presented in [25].

Concerning the eigenvalue clusters for non-Hermitian matrices, we have a smaller room for two matrix sequences to be considered as "close sequences". In this case Theorem 3.1 modifies as follows.

Theorem 3.2. [20] *Assume that A_n and P_n are diagonalizable via the eigenvector matrices X_n and Y_n, respectively, and let*

$$\mathrm{cond}_2^2 X_n \,\mathrm{cond}_2^2 Y_n \,||A_n - P_n||_F^2 = o(n).$$

Then the eigenvalues of A_n and P_n have the same clusters.

This theorem obviously applies to normal matrices. However, in the nonnormal case we have a substantial result only for special clusters consisting of one point. If the singular values are clustered at zero, then this is generally true also for the eigenvalues [22,23].

Let A_n have the singular values $\sigma_1(A_n) \geqslant ... \geqslant \sigma_n(A_n)$ and the eignevalues $\lambda_i(A_n)$ ordered so that $|\lambda_1(A_n)| \geqslant ... \geqslant |\lambda_n(A_n)|$. Using the Schur theorem and the interlacing properties, one can fastly arrive at the following *Weyl inequalities*:

$$\prod_{i=1}^{l} |\lambda_i(A_n)| \leqslant \prod_{i=1}^{l} \sigma_i(A_n), \quad 1 \leqslant l \leqslant n.$$

Now assume that zero is the singular value cluster for A_n. Define

$$k = k(\delta, n), \quad m = m(\varepsilon, n)$$

in the following way:

$$\sigma_k(A_n) \geqslant \delta > \sigma_{k+1}(A_n) \quad \text{and} \quad |\lambda_m(A_n)| \geqslant \varepsilon > |\lambda_{m+1}(A_n)|.$$

From the Weyl inequalities,

$$\varepsilon^m \leqslant ||A_n||_2^k \; \delta^{m-k} \quad \Rightarrow \quad \left(\frac{\varepsilon}{\delta}\right)^m \leqslant \left(\frac{||A_n||_2}{\delta}\right)^k \tag{3.6}$$

and, evidently,

$$\left(\frac{\varepsilon}{\delta}\right)^{m/n} \leqslant \left(\frac{||A_n||_2}{\delta}\right)^{k/n}. \tag{3.7}$$

In the case of general singular value cluster at zero we have

$$k(\delta, n)/n \to 0.$$

Hence, if $m(\varepsilon, n)/n \not\to 0$, then the left-hand side of (3.7) can be made arbitrarily large by choosing a sufficiently small $\delta > 0$. It implies that the right-hand side cannot be uniformly bounded for all n and all sufficiently small $\delta > 0$.

If the singular value cluster is proper, then $k(\delta, n) \leqslant c(\delta) \;\; \forall n$. Let us fix an arbitrary $\varepsilon > 0$ and admit that $m(\varepsilon, n) \to \infty$ as $n \to \infty$. Then, for any fixed $\delta > 0$ and all sufficiently large n we obtain $m(\varepsilon, n) > k(\delta, n)$. Choosing, for instance, $\delta = \varepsilon/2$, we conclude from (3.6) that $||A_n||_2 \to \infty$. The same reasoning can be applied to a subsequence of $m(\varepsilon, n)$ and the corresponding subsequence of $||A_n||_2$. It means that the unboundedness of $m(\varepsilon, n)$ implies that the norms $||A_n||_2$ are not bounded. Thus, we have proved the following

Theorem 3.3. *Let the singular values of A_n be clustered at zero with the number of δ-distant values $k(\delta,n)$ and assume that for some $c>0$*

$$|\ln||A_n||_2| \leqslant c\frac{n}{k(\delta,n)}$$

for all n and sufficiently small $\delta>0$. Then the eignevalues of A_n are clustered at zero. If the spectral norms of A_n are uniformly bounded in n and the singular value cluster at zero is proper, then the eigenvalue cluster at zero is also proper.

The observation that the singular value cluster at zero implies the eigenvalue cluster at zero was first presented in [22]. The proof was based, as above, on the inequality (3.7). The proper cluster case is based on the preceeding inequality (3.6) and certainly is within the lines of the same proof. However, this case was not mentioned explicitly in [22] and was remarked later in [17].

4. SHORT RECURRENCES

In contrast to the Krylov subspace methods for Hermitian matrices, in the non-Hermitian case we seem to have to store complete bases for the Krylov subspaces. A natural question actively discussed in the late 1970s is whether this can be avoided through some "short recurrences" in the non-Hermitian case.

To begin with, the very question should be specified in mathematical terms. Given a system $Ax=b$ with a nonsingular (non-Hermitian) matrix, we choose an initial guess x_0, find the initial residual r_0, set $p_1=r_0$ (of course, if $r_0\neq 0$) and successively add a new vector to the previously obtained bases $p_1,\ldots,p_k$ in the Krylov subspaces

$$L_k = L(r_0,Ar_0,\ldots,A^{k-1}r_0) = L(p_1,\ldots,p_k)$$

in such a way that the constructed vectors satisfy the *formal A-orthogonality* conditions

$$(Ap_i,p_j)=0,\quad i\neq j,\quad 1\leqslant i,j\leqslant k;\quad (Ap_i,p_i)\neq 0,\ \ 1\leqslant i\leqslant k.$$

As soon as L_k is found, we look for x_k in the form $x_k=x_0+y,\ y\in L_k$. However, we need to sacrifice something in favor of "short recurrences". This will be the minimization property of the residuals $r_k=b-Ax_k$. Instead, we will define y by the *projection property*

$$r_k\perp L_k.$$

It follows that

$$x_k=x_{k-1}+\alpha_k p_k,\quad r_k=r_{k-1}-\alpha_k Ap_k,$$

where α_k is defined by the projection property.

If $r_k = 0$, then the solution is already obtained. If $r_k \neq 0$, then we seek p_{k+1} in the form

$$p_{k+1} = r_k + \gamma_{11} p_1 + \cdots + \gamma_{k1} p_k \quad \Rightarrow \quad \gamma_{jk} = -(r_k, A^* p_j)/(A p_j, p_j).$$

Thus, using a formal A-orthogonal basis $p_1, \ldots, p_k$ in L_k, we can readily find p_{k+1} retaining the formal A-orthogonality properties $(Ap_{k+1}, p_j) = 0,\ 1 \leqslant j \leqslant k$.

Despite the case of Hermitian positive definite matrices, now we *cannot* take it for granted that $(Ap_{k+1}, p_{k+1}) \neq 0$. This is what we should *assume*; in particular, we assume that $(Ar_0, r_0) \neq 0$. If the residuals $r_0, r_1, \ldots, r_{k-1}$ are nonzero and the formal A-orthogonal basis $p_1, \ldots, p_k$ L_k is built up, then let us agree to say that the process *does not break down* at the kth step. If $r_k = 0$ then let us say that the process *quits successfully* at the kth step.

Lemma 4.1. *If the process does not break down at the kth step, then the residuals $r_0, \ldots, r_{k-1}$ produce an orthogonal basis in L_k.*

Proof. Since $r_j \in L_{j+1} \subset L_k$ for $0 \leqslant j \leqslant k-1$ and due to the projection property, we have $r_j \perp r_0, \ldots, r_{j-1}$. $\square$

The question about "short recurrences" can be set up as follows. Let us fix $1 \leqslant s \leqslant n-1$ and suppose that

$$\gamma_{jk} = (r_k, A^* p_j) = 0 \quad \text{for} \quad 1 \leqslant j \leqslant k-s \tag{4.1}$$

whenever the process does not break at the kth step. This means that p_{k+1} is expressed through s last vectors of the Krylov basis via *short recurrences*

$$p_{k+1} \ = \ r_k + \sum_{j=k-s+1}^{k} \gamma_{jk} p_j. \tag{4.2}$$

In order to have (4.2), *what properties should A have*?

To this end, in 1970s V. V. Voevodin proposed to consider matrices with the following property:

$$A^* \ = \ \sum_{j=0}^{s-1} a_j A^j. \tag{4.3}$$

It gets easy to prove that the property (4.3) is sufficient to guarantee (4.1) and, therefore, (4.2).

Lemma 4.2. *The matrix property (*4.3*) implies (*4.1*), the latter being valid for any initial residual $r_0 \neq 0$ with no break-down at the kth step.*

Proof. In line with (4.3), A^*p_j is a linear combination of $p_1, \ldots, p_{j+s}$. From the projection property we deduce that $r_k \perp p_1, \ldots, p_{j+s}$ for $j+s \leqslant k$, which makes (4.1) evident. □

Does the same property (4.3) is necessary for short recurrences? A principal positive answer was first given in [26]. The proof of [26] was astonishingly simple; however, it used an additional assumption that $n \geqslant 2s+1$. Later the necessity of (4.3) was established in [7] by remarkably complicated techniques. Recently, the elementary technique of [26] has got a new life in [11], where the authors found a way to pursue the same lines of proof without assuming that $n \geqslant 2s+1$.

The elementary proof of [26] starts from the following observation.

Lemma 4.3. *Assume that an initial residual $r_0 \neq 0$ is such that the process does not break at the nth step and the equalities (*4.1*) are vaild for all $1 \leqslant k \leqslant n$. Then for some $\alpha_j = \alpha_j(r_0)$ we obtain*

$$A^*r_0 \;=\; \sum_{j=0}^{s-1} \alpha_j A^j r_0.$$

Proof. The claim that the process does not break at the nth step means orthogonality of the residuals $r_0, \ldots, r_{n-1}$ and linear independence of vectors

$$r_0, Ar_0, \ldots, A^{n-1}r_0.$$

The equalities $(Ar_k, p_j) = 0$ for $1 \leqslant j \leqslant k-s$ ensure that $(Ar_k, r_j) = 0$ for $0 \leqslant j \leqslant k-s-1$. Consequently, $A^*r_0 \perp r_k$ for $k \geqslant s-1 \quad \Rightarrow \quad A^*r_0$ is a linear combination of vectors $r_0, \ldots, r_{s-2}$. It follows that A^*r_0 is a linear combination of vectors $r_0, Ar_0, \ldots, A^{s-1}r_0$. □

The final result on short recurrences is formulated as follows.

Theorem 4.1. *Let $1 \leqslant s < n$ and let A be such that the process does not break at the nth step for at least one intitial residual $r_0 \neq 0$. Then, the necessary and sufficient condition for short recurrences (*4.1*) to hold for all initial residuals with the same property is that A possesses the property (*4.3*).*

Proof. The sufficiency of (4.3) is established in Lemma 4.2; thus, it remains to prove necessity. Linear independence of vectors $r_0, Ar_0, \ldots, A^{n-1}r_0$ means that the minimal polynomial for the matrix A is of degree $n \quad \Rightarrow \quad$ any eigenvalue has exactly one Jordan block. Let

$$x = A^l r_0, \quad y = A^m r_0, \qquad 0 \leqslant l < m < l+n. \tag{4.4}$$

Clearly, if the initial residual is equal to x or y, then the process does not break at the nth step. Moreover, for any initial residual of the form $x+\gamma y$ the process can break before the nth step only for a finite number of values of γ (which does not exceed the number of Jordan blocks for A). According to Lemma 4.3, we obtain

$$A^*x=\sum_{j=0}^{s-1}\alpha_jA^jx,\qquad A^*y=\sum_{j=0}^{s-1}\beta_jA^jy,\qquad A^*(x+\gamma y)=\sum_{j=0}^{s-1}\varphi_jA^j(x+\gamma y).$$

Taking into account (4.4), we can write

$$\begin{aligned}A^*(x+\gamma y) &= \sum_{j=0}^{s-1}\varphi_jA^{l+j}x+\gamma\sum_{j=0}^{s-1}\varphi_jA^{m+j}y\\ &= \sum_{j=0}^{s-1}\alpha_jA^{l+j}x+\gamma\sum_{j=0}^{s-1}\beta_jA^{m+j}y.\end{aligned}$$

If $l+s\leqslant m$ and $m+s-l\leqslant n$ then the vectors

$$A^lr_0,...,A^{l+s-1}r_0,\ A^mr_0,...,A^{m+s-1}r_0$$

are part of the basis

$$A^lr_0,...,A^{l+n-1}r_0.$$

Therefore,

$$\alpha_j=\beta_j=\gamma_j,\quad 0\leqslant j\leqslant s-1. \tag{4.5}$$

This is exactly the main point of the proof of [26]. To complete the proof, we consider the sequence of integers

$$0,\ s+1,\ 1,\ s+2,\ 2,\ s+3,\ ...$$

and succesively choose l and m as the pairs of neigboring integers in this sequence. Obviously, if $n\geqslant 2s+1$ then we obtain (4.5) for each of these pairs and eventually deduce that the coefficients α_j do not depend on l if $x=A^lr_0$.

In [11] it is shown how to get rid of the above assumption. It is suggested to take $x=r_0$ and $y=Ar_0$. Then the equation

$$\alpha_0x+\sum_{j=1}^{s-1}(\alpha_j+\gamma\beta_{j-1})A^jx+\beta_{s-1}A^sx=\varphi_0x+\sum_{j=1}^{s-1}(\varphi_j+\gamma\varphi_{j-1})A^jx+\varphi_{s-1}A^sx$$

implies that

$$\varphi_0=\alpha_0;\qquad \varphi_j+\gamma\varphi_{j-1}=\alpha_j+\gamma\beta_{j-1},\ \ 1\leqslant j\leqslant s-1;\qquad \varphi_{s-1}=\beta_{s-1}.$$

Subtract the first equation from the second one multiplied by γ:

$$\varphi_1=\alpha_1+\gamma(\beta_0-\alpha_0).$$

Multiply the resulting equation by γ and subtarct it from the third equation:

$$\varphi_2 = \alpha_2 + \gamma(\beta_1 - \alpha_1) - \gamma^2(\beta_0 - \alpha_0).$$

And so on. In the end we obtain

$$\varphi_{s-1} = \beta_{s-1} = \alpha_{s-1} + \gamma(\beta_{s-2} - \alpha_{s-2}) - \gamma^2(\beta_{s-3} - \alpha_{s-3}) + \cdots + (-1)^s\gamma^{s-2}(\beta_0 - \alpha_0)$$

$$\Rightarrow \quad \sum_{j=0}^{s-2} \gamma^{s-2-j}(\beta_j - \alpha_j)(-1)^{s-j} = 0.$$

The latter equation should hold for infinitely many values of $\gamma \;\Rightarrow\; \alpha_j = \beta_j$ for all $0 \leqslant j \leqslant s-1$. Hence, the equality

$$A^*z = \sum_{j=0}^{s-1} \alpha_j A^j z$$

holds true with the same values of α_j for all vectors $z = x, Ax, \ldots, A^{n-1}x$. Since this is a basis in $\mathbb{C}^n$, we arrive at the matrix equality (4.3) with $a_j = \alpha_j$. $\square$

REFERENCES

1. O. Axelsson and G. Lindskog, The rate of convergence of the conjugate gradient method, *Numer. Math.*, 48 (1986), 499–523.
2. B.Beckermann, S.A.Goreinov, E.E.Tyrtyshnikov, Some remarks on the Elman estimate for GMRES, *SIMAX*, Vol. 27, Issue 3 (2006), 772–778.
3. R. H. Chan, M. K. Ng and A. M. Yip, A Survey of Preconditioners for Ill-Conditioned Toeplitz Systems, Structured Matrices in Mathematics, Computer Science, and Engineering II, *Contemporary Mathematics*, 281 (2001), 175–191 (Ed. V. Olshevsky).
4. P. Concus, G. Golub, D. O'Leary, A generalized conjugate gradient method for non-symmetric systems of linear equations, *Lecture Notes in Economics and Mathematical Systems*, Berlin, Springer, 134 (1976), 56–65.
5. M. Eiermann, Fields of Values and Iterative Methods, *Lin. Alg. Applics* **180** (1993) 167-197.
6. H. C. Elman, *Iterative Methods for Sparse Nonsymmetric Systems of Linear Equations*, PhD Thesis, Yale University, Department of Computer Science, 1982.
7. V. Faber and T. Manteuffel, Necessary and sufficient conditions for the existence of a conjugate gradient method, *SINUM*, Vol. 21, Issue 2 (1984), 352–362.
8. G. M. Goluzin, *Geometrical theory of functions of complex variable*, Nauka, Moscow, 1966.
9. M. A. Krasnoselsky and S. G. Krein, Iterative process with minimal residuals, *Sbornik: Mathematics*, 31 (73), no. 2 (1952), 315–334.
10. Yu. A. Kuznetsov, Method of conjugate gradients, its generalizations and applications, *Computational processes and systems*, Nauka, 1983, pp. 267–301.
11. J. Liesen and P. E. Saylor, Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases, *SINUM*, Vol. 42, Issue 5 (2005), 2148–2158.

12. G. I. Marchuk and Yu. A. Kuznetsov, *Iterative methods and quadratic functionals*, Novosibirsk, 1972.

13. I. Oseledets and E. Tyrtyshnikov, A unifying approach to the construction of circulant preconditioners, *Linear Algebra Appl.*, 2006.

14. Y. Saad and M. H. Schultz, GMRES: A generalized minimal residual algorithm for solving nunsymmetric linear systems, *SIAM J. Sci. Comput.* **7** (1986) 856-869.

15. Y. Saad, *Iterative methods for sparse linear systems*, PWS Publishing, Boston, MA, 1996.

16. S. Serra Capizzano and E. Tyrtyshnikov, Any circulant-like preconditioner for multilevel Toeplitz matrices is not superlinear. *SIAM J. on Matrix Analysis and Appl.* 21 (2): 431–439 (1999).

17. S. Serra-Capizzano, D. Bertaccini and G. H. Golub, How to deduce a proper eigenvalue cluster from a proper singular value cluster in the nonnormal case, *SIMAX*, Vol. 27, Issue 1 (2005), 82–86.

18. A. van der Sluis and H. A. van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.*, 48 (1986), 543–560.

19. E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering, *Linear Algebra Appl.*, 1996, 232: 1–43.

20. E. E. Tyrtyshnikov, *A Brief Introduction to Numerical Analysis*, Birkhauser, Boston, 1997.

21. E. Tyrtyshnikov and R.Chan, Spectral Equivalence and Proper Clusters for Boundary Element Method Matrices, *Int. J. Numer. Meth. Engnr.* 49 (2000), 1211–1224.

22. E. E. Tyrtyshnikov and N. L. Zamarashkin, On eigen and singular value clusters. *Calcolo* 33: 71–78 (1997).

23. E. E. Tyrtyshnikov, N. L. Zamarashkin, and A. Yu. Yeremin, Clusters, preconditioners, convergence, *Linear Algebra Appl.* 263: 25–48 (1997).

24. R. S. Varga, *Matrix iterative analysis*, Englewood Cliffs, M. Y., Prentice Hall, 1962.

25. N. L. Zamarashkin, I. V. Oseledets, E. E. Tyrtyshnikov, Approximation of Toeplitz matrices by sums of circulants and small-rank matrices, *Doklady Mathematics*, Vol. 73, No. 1 (2006), 100–101.

26. V. V. Voevodin and E. E. Tyrtyshnikov, On generalization of conjugate directions method, *Numerical Methods of Linear Algebra*, Moscow University Press, 1981, 3–9. English translation: “http://www.inm.ras.ru/library/Tyrtyshnikov/generalization-cg.pdf”.