The Idea Behind Krylov Methods

Ilse C.F. Ipsen* Carl D. Meyer†

February 3, 1997

Technical Report CRSC-TR97-3
Center for Research in Scientific Computation
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA

Abstract

We explain why Krylov methods make sense, and why it is natural to represent a solution to a linear system as a member of a Krylov space.

In particular we show that the solution to a nonsingular linear system $Ax = b$ lies in a Krylov space whose dimension is the degree of the minimal polynomial of A. Therefore, if the minimal polynomial of A has low degree then the space in which a Krylov method searches for the solution is small. In this case a Krylov method has the opportunity to converge fast.

When the matrix is singular, however, Krylov methods can fail. Even if the linear system does have a solution, it may not lie in a Krylov space. In this case we describe the class of right-hand sides for which a solution lies in a Krylov space. As it happens, there is only a single solution that lies in a Krylov space, and it can be obtained from the Drazin inverse.

*Center for Research In Scientific Computation, Department of Mathematics, North Carolina State University, P. O. Box 8205, Raleigh, NC 27695-8205 (ipsen@math.ncsu.edu). The research of this author was supported in part by NSF grant CCR-9400921.

†Center for Research In Scientific Computation, Department of Mathematics, North Carolina State University, P. O. Box 8205, Raleigh, NC 27695-8205 (meyer@math.ncsu.edu). The research of this author was supported in part by NSF grant CCR-9413309.

1 Why Krylov Methods?

How do you solve a system of linear equations $Ax = b$ when your coefficient matrix A is large and sparse (i.e. contains many zero entries)? What if the order n of the matrix is so large that you cannot afford to spend about n^3 operations to solve the system by Gaussian elimination? Or what if you do not have direct access to the matrix? Say the matrix A exists only implicitly as a subroutine which, when given a vector v, returns Av?

In this case you may want to use a Krylov method. A Krylov method solves $Ax = b$ by repeatedly performing matrix vector multiplications involving A. Starting with an initial guess x_0, it bootstraps its way up to (hopefully) ever more accurate approximations x_k to the desired solution. Suppose we choose $x_0 = 0$ as our initial guess (we deal with a non-zero x_0 in §13). In iteration k a Krylov method produces an approximate solution x_k from the *Krylov space*

$$\mathcal{K}_k(A,\, b) \equiv \operatorname{span}\{b, Ab, \ldots, A^{k-1}b\}.$$

Let's look at a specific example.

2 An Example of a Krylov Method

The generalized minimal residual method (GMRES) was published by Saad and Schultz in 1986 [SS86]. In iteration $k \geq 1$ GMRES picks the 'best' solution x_k from the Krylov space $\mathcal{K}_k(A,\, b)$. 'Best' means that the residual is as small as possible over $\mathcal{K}_k(A,\, b)$; i.e. x_k solves the least squares problem

$$\min_{z \in \mathcal{K}_k(A,\, b)} \|b - Az\| \qquad (\| \star \| \text{ is the Euclidean norm}). \tag{1}$$

GMRES solves this least squares problem by constructing an orthonormal basis $\{v_1, v_2, \ldots, v_k\}$ for $\mathcal{K}_k(A,\, b)$ using *Arnoldi's method,* which is a version of the Gram–Schmidt procedure tailored to Krylov spaces. Starting with the normalized right-hand side $v_1 = b/\|b\|$ as a basis for $\mathcal{K}_1(A,\, b)$, Arnoldi recursively builds an orthonormal basis for $\mathcal{K}_{j+1}(A,\, b)$ from an orthonormal basis for $\mathcal{K}_j(A,\, b)$ as follows. It orthogonalizes the vector Av_j from $\mathcal{K}_{j+1}(A,\, b)$ against the previous space $\mathcal{K}_j(A,\, b)$. That is,

$$\hat{v}_{j+1} = Av_j - \left(h_{1j}v_1 + \cdots + h_{jj}v_j\right), \qquad \text{where} \quad h_{ij} = v_i^* A v_j \quad (* \text{ is conjugate transpose}). \tag{2}$$

The new basis vector is

$$v_{j+1} = \hat{v}_{j+1}/\|\hat{v}_{j+1}\|.$$

If we collect the orthonormal basis vectors for $\mathcal{K}_j(A,\, b)$ in a matrix, $V_j = (\, v_1 \ldots v_j \,)$, we get the decomposition associated with Arnoldi's method:

$$AV_j = V_{j+1}H_j,$$

where H_j is a Hessenberg matrix of size $(j+1) \times j$.

In the context of the least squares problem (1) this means: If $z \in \mathcal{K}_k(A, b)$, then $z = V_k y$ for some y, so

$$Az = AV_k y = V_{k+1} H_k y \quad \text{and} \quad b = \beta v_1 = \beta V_{k+1} e_1$$

where $\beta = \|b\|$ and e_1 is the first column of the identity matrix. The least squares problem in iteration k of GMRES reduces to

$$\min_{z \in \mathcal{K}_k(A,b)} \|b - Az\| = \min_y \|\beta e_1 - H_k y\|.$$

Thus GMRES proceeds as follows.

Iteration 0**:** Initialize $x_0 = 0, \quad v_1 = b/\beta, \quad V_1 = v_1, \quad H_0 = 0.$

Iteration $k \geq 1$**:**

1. Orthogonalize: $\hat{v}_{k+1} = Av_k - V_k h_k$ where $h_k = V_k^* A v_k$
2. Normalize: $v_{k+1} = \hat{v}_{k+1} / \|\hat{v}_{k+1}\|$
3. Update: $V_{k+1} = \begin{pmatrix} V_k & v_{k+1} \end{pmatrix}, \qquad H_k = \begin{pmatrix} H_{k-1} & h_k \\ 0 & \|\hat{v}_{k+1}\| \end{pmatrix}$
4. Solve the least squares problem $\min_y \|\beta e_1 - H_k y\|$, and call the solution y_k.
5. The approximate solution is $x_k = V_k y_k$.

Why does GMRES do what it is supposed to do? GMRES stops as soon as it has produced a zero vector. Let s be the first index for which $\hat{v}_{s+1} = 0$. If $s = 0$ then clearly $b = 0$, and $x_0 = 0$. In this case GMRES has found the solution to $Ax = b$.

If $s > 0$ then the last row of H_s is zero. Let $\hat{H}_s$ be H_s without its last row. It can be shown that $\hat{H}_s$ is nonsingular. Hence the least squares problem reduces to a nonsingular linear system $H_s y_s = \beta e_1$. From $AV_s = V_s H_s$ follows

$$AV_s y_s = V_s H_s y_s = \beta V_s e_1 = b,$$

and $x_s = V_s y_s$ is the solution to $Ax = b$. Again, GMRES has found the solution. Note that s cannot exceed n because we cannot have more than n linearly independent vectors for a $n \times n$ matrix A.

Therefore, GMRES works properly.

3 Questions

There is no shortage of Krylov methods. The big names include, among others: conjugate gradient; conjugate residual; Lanczos biorthogonalization; quasi-minimal residual (QMR); biconjugate gradient; and A-conjugate direction methods.

Like GMRES, all of these methods tend to provide acceptable solutions in far fewer than n iterations, n being the order of A. Just how few iterations are required depends on the eigenvalues of A, and the nature of this dependence is crucial for understanding Krylov methods. But because the existing literature tends to concentrate on particular details of specific methods, it is not easy to see the common ground shared by all Krylov methods. This was our motivation for writing this article. Here are some of the general questions that occurred to us when we tried to understand Krylov methods.

1. Why is $\mathcal{K}_k(A, b)$ a good space from which to construct an approximate solution?

   (At first sight Krylov space methods did not strike us as a natural way to solve linear systems. They don't work, for instance, when the number of equations is different from the number of unknowns.)

2. Why are eigenvalues important for Krylov methods?

   (We would have expected the action to evolve around the singular values, because they are the ones that usually matter when it comes to linear system solution.)

3. Why do Krylov methods tend to converge faster for Hermitian, or real symmetric matrices?

   (After all, we just want to represent b as a linear combination of columns of A. Why should it matter that the columns belong to a Hermitian or symmetric matrix?)

4 Answers

If we can show that the solution to $Ax = b$ has a 'natural' representation as a member of a Krylov space, then we can understand why one would construct approximations to x from a Krylov space. Moreover, if x lies in a Krylov space of small dimension, a Krylov method would have the opportunity to find x in few iterations. This means the dimension of the smallest Krylov space harboring x is our gauge for convergence. If this space is small, we have a plausible reason to expect rapid convergence.

Our strategy is to begin with nonsingular matrices. We use the minimal polynomial of the coefficient matrix A to express A^{-1} in terms of powers of A. This casts the solution $x = A^{-1}b$ automatically as a member of a Krylov space. The dimension of this space is the degree of the minimal polynomial of A.

Next we consider linear systems whose coefficient matrix A is singular. To be assured of a solution that lies in a Krylov space, we have to confine the right-hand side b to the 'nonsingular part' of A and keep it away from the 'nilpotent part'. As a result, the dimension of the Krylov space shrinks: It is the degree of the minimal polynomial of A minus the index of the zero eigenvalue. It also turns out that there is only a single solution that lies in a Krylov space.

Our discussion is restricted to exact arithmetic; we ignore finite precision effects such as rounding errors.

5 The Minimal Polynomial of a Matrix

The minimal polynomial $q(t)$ of A is defined as the unique monic polynomial of minimal degree such that $q(A) = 0$, and it provides an economical way to represent a matrix in terms of its eigenvalues. If A has d distinct eigenvalues λ_j of index m_j (the size of a largest Jordan block associated with λ_j), then the sum of all indices is

$$m \equiv \sum_{j=1}^{d} m_j, \quad \text{and} \quad q(t) = \prod_{j=1}^{d} (t - \lambda_j)^{m_j}. \tag{3}$$

For example, the matrix

$$\begin{pmatrix} 3 & 1 & & \\ & 3 & & \\ & & 4 & \\ & & & 4 \end{pmatrix},$$

has an eigenvalue 3 of index 2 and an eigenvalue 4 of index 1, so $m = 3$ and $q(t) = (t-3)^2\,(t-4)$. When A is diagonalizable, m is the number of distinct eigenvalues of A. When A is a Jordan block of order n, then $m = n$.

It's clear from (3) that if we write

$$q(t) = \sum_{j=0}^{m} \alpha_j t^j,$$

then the constant term is $\alpha_0 = \prod_{j=1}^{d} \lambda_j^{m_j}$. Therefore $\alpha_0 \neq 0$ if and only if A is nonsingular. This observation will come in handy in the next section.

6 The Idea

Using the minimal polynomial to represent the inverse of a nonsingular matrix A in terms of powers of A is at the heart of the issue. Since $\alpha_0 \neq 0$ in

$$0 = q(A) = \alpha_0 I + \alpha_1 A + \cdots + \alpha_m A^m \qquad (I \text{ is the identity matrix}),$$

it follows that

$$A^{-1} = \frac{1}{\alpha_0} \sum_{j=0}^{m-1} \alpha_j A^j.$$

Consequently, the smaller the degree of the minimal polynomial the shorter the description for A^{-1}. This description of A^{-1} portrays $x = A^{-1}b$ immediately as a member of a Krylov space:

Theorem 1 *If the minimal polynomial of the nonsingular matrix A has degree m, then the solution to $Ax = b$ lies in the space $\mathcal{K}_m(A, b)$.*

Therefore, in the absence of any information about b, we have to assume that the dimension of the smallest Krylov space containing x is m, the degree of the minimal polynomial of A. If the minimal polynomial has low degree then the Krylov space containing the solution is small, and a Krylov method has the opportunity to converge fast.

7 An Extreme Example

The results in the previous section suggest that we should expect the maximal dimension from a Krylov space when the matrix is a nonsingular Jordan block, because in this case the minimal polynomial has maximal degree. Let's find out what GMRES does with $Ax = b$ when

$$A = \begin{pmatrix} 2 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Suppose A has order n, and denote the columns of the identity matrix of order n by $e_1 \dots e_n$. Then $b = e_n$.

Iteration 0: $v_1 = b = e_n$.

Iteration 1: $h_{11} = v_1^* A v_1 = e_n^* A e_n = 2$ and

$$v_2 = \hat{v}_2 = (A - h_{11} I) v_1 = (A - 2I) e_n = e_{n-1}.$$

Iteration 2:

$$h_{12} = v_1^* A v_2 = e_n^* A e_{n-1} = 0, \qquad h_{22} = v_2^* A v_2 = e_{n-1}^* A e_{n-1} = 2$$

and

$$v_3 = \hat{v}_3 = (A - h_{22} I) v_2 = (A - 2I) e_{n-1} = e_{n-2}.$$

Now it becomes clear that the orthonormal basis vectors v_i are going to run through all the columns of the identity matrix before finally ending up with a zero vector at the last possible moment.

Iteration n:

$$h_{1,n} = \cdots = h_{n-1,n} = 0, \qquad h_{n,n} = v_n^* A v_n = e_1^* A e_1 = 2$$

and

$$v_{n+1} = \hat{v}_{n+1} = (A - h_{n,n} I) v_n = (A - 2I) e_1 = 0.$$

Indeed, Saad and Schultz have shown that the maximal number of iterations in GMRES does not exceed the degree of the minimal polynomial of A.

Now that we have understood the situation for nonsingular matrices, let's look at singular matrices.

8 What's Different About Singular Systems?

Suppose a linear system has a singular coefficient matrix. Even if a solution exists, it may not lie in a Krylov space. The following example illustrates this.

Let $Nx = c$ be a consistent linear system, where N is a nilpotent matrix and $c \neq 0$. This means there is a number i such that $N^i = 0$ but $N^{i-1} \neq 0$. Suppose the solution to $Nx = c$ is a linear combination of Krylov vectors, i.e. $x = \xi_0 c + \xi_1 Nc + \cdots \xi_{i-1} N^{i-1} c$. Then

$$c = Nx = \xi_0 Nc + \cdots + \xi_{i-2} N^{i-1} c$$

and

$$(I - \xi_0 N - \cdots - \xi_{i-2} N^{i-1})c = 0.$$

But the matrix in parentheses is nonsingular. Its eigenvalues are all equal to one, because the terms containing N make up a nilpotent matrix. Consequently, $c = 0$. In other words, a solution to a nilpotent system with non-zero right-hand side cannot lie in a Krylov space.

This observation is important because it suggests that if we want the solution to a general square system $Ax = b$ to lie in a Krylov space we are going to have to restrain b by somehow keeping it away from the 'nilpotent part' of A.

9 Exactly When Do Krylov Solutions Exist?

The trick is to decompose the space into

$$\mathcal{C}^n = R(A^i) \oplus N(A^i),$$

where i is the index of the zero eigenvalue of $A \in \mathcal{C}^{n \times n}$, and where $R(\star)$ and $N(\star)$ denote range and nullspace. This space decomposition in turn induces a matrix decomposition

$$A = X \begin{pmatrix} C & 0 \\ 0 & N \end{pmatrix} X^{-1}, \tag{4}$$

where C is nonsingular, and N is nilpotent of index i. This decomposition is basically a coarse version of a Jordan decomposition.

Now suppose that $Ax = b$ has a Krylov solution

$$x = \sum_{j=0}^{p} \alpha_j A^j b = \sum_{j=0}^{p} \alpha_j X \begin{pmatrix} C^j & 0 \\ 0 & N^j \end{pmatrix} X^{-1} b.$$

Setting $y = X^{-1} x = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ and $z = X^{-1} b = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ gives

$$y_1 = \sum_{j=0}^{p} \alpha_j C^j z_1, \qquad y_2 = \sum_{j=0}^{p} \alpha_j N^j z_2.$$

But $Ax = b$ implies $Ny_2 = z_2$, hence $N(\sum_{j=0}^{p} \alpha_j N^j z_2) = z_2$, and

$$(I - \textstyle\sum_{j=0}^{p} \alpha_j N^{j+1}) z_2 = 0.$$

Like in §8, the matrix in parentheses is nonsingular, and $z_2 = 0$. Thus $X^{-1}b = \begin{pmatrix} z_1 \\ 0 \end{pmatrix}$, and $b \in R(A^i)$. Therefore the existence of a Krylov solution forces b into $R(A^i)$.

It turns out that the converse is also true. If we start with $b \in R(A^i)$, then $X^{-1}b = \begin{pmatrix} z_1 \\ 0 \end{pmatrix}$ for some z_1. Hence

$$x = X \begin{pmatrix} C^{-1} z_1 \\ 0 \end{pmatrix}$$

is a solution to $Ax = b$. Since we have confined the right-hand side to the 'nonsingular part' of A, we can apply the idea of §6 to C. Since the minimal polynomial for C has degree $m - i$, there is a polynomial $p(x)$ of degree $m - i - 1$ such that $C^{-1} = p(C)$. Substituting this into the expression for x gives

$$\begin{aligned} x &= X \begin{pmatrix} C^{-1} z_1 \\ 0 \end{pmatrix} = X \begin{pmatrix} p(C) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ 0 \end{pmatrix} = X \begin{pmatrix} p(C) & 0 \\ 0 & p(N) \end{pmatrix} X^{-1} b \\ &= p(A)b \ \in \mathcal{K}_{m-i}(A, b). \end{aligned}$$

Therefore $b \in R(A^i)$ guarantees the existence of a Krylov solution.

The following theorem summarizes our findings so far.

Theorem 2 (Existence of a Krylov Solution) *A square linear system $Ax = b$ has a Krylov solution if and only if $b \in R(A^i)$, where i is the index of the zero eigenvalue of A.*

In other words, a linear system has a Krylov solution if and only if the right-hand side is kept away from the 'nilpotent part' of the matrix and confined to the 'nonsingular part'.

In the special case when A is nonsingular, $i = 0$, and the condition on b is vacuous. When A has a non-defective zero eigenvalue, then $i = 1$, and the condition on b reduces to the familiar consistency condition $b \in R(A)$. This occurs, for instance, when A is diagonalizable. In this case a consistent system $Ax = b$ has a solution

$$x \in \begin{cases} \mathcal{K}_{d-1}(A, b) & \text{if } A \text{ is singular}, \\ \mathcal{K}_d(A, b) & \text{if } A \text{ is nonsingular}, \end{cases} \tag{5}$$

where d is the number of distinct eigenvalues of A.

Compared to the nonsingular case, the Krylov space for the singular case has shrunk. It's dimension is by i smaller than the degree of the minimal polynomial. Thus, the search space shrinks as the defectiveness of the zero eigenvalue grows. As a trade-off, though, the selection of desirable right-hand sides diminishes as well.

How about the number of possible Krylov solutions? We don't have any idea yet how many there can be. To answer this question, we need a compact representation of solutions to a linear system with confined right-hand side. Pseudoinverses are often useful in this context, and the first thing that comes to mind is the Moore-Penrose inverse of A. But this isn't going to work because the Moore-Penrose inverse generally cannot be expressed as a polynomial in A [CM79, Section 7.5]. So let's give the Drazin inverse a try.

10 The Drazin Inverse Comes to the Rescue

If A has a zero eigenvalue with index i then the Drazin inverse of A is defined as the unique matrix A^D that satisfies [Dra68], [CM79, Section 7.5]

$$A^D A A^D = A^D, \qquad A^D A = A A^D, \qquad A^{i+1} A^D = A^i.$$

In the special case where A is nonsingular, $i = 0$ and the Drazin inverse is the ordinary inverse, $A^D = A^{-1}$.

Let's first establish the circumstances under which the Drazin inverse is useful for representing solutions of linear systems. That is, when is $A^D b$ a solution to $Ax = b$? Like most other questions concerning the Drazin inverse, we can easily answer this one by decomposing the Drazin inverse conformably with the decomposition (4) induced by the index of the zero eigenvalue,

$$A^D = X \begin{pmatrix} C^{-1} & 0 \\ 0 & 0 \end{pmatrix} X^{-1}. \tag{6}$$

Because AA^D is the projector onto $R(A^i)$ along $N(A^i)$, we conclude that $AA^D b = b$ if and only if $b \in R(A^i)$. The following lemma sums up the state of affairs at this point.

Lemma 1 *The following statements are equivalent.*

- *$A^D b$ is a solution of $Ax = b$.*
- *$b \in R(A^i)$, where i is the index of the zero eigenvalue of A.*
- *$Ax = b$ has a Krylov solution.*

Now the only piece missing in the puzzle is the connection between Krylov solutions and the Drazin inverse. Suppose $b \in R(A^i)$, and proceed as in the previous section. The minimal polynomial for C has degree $m - i$, so there is a polynomial $p(x)$ of degree $m - i - 1$ such that $C^{-1} = p(C)$. Then (6) and Lemma 1 imply

$$\begin{aligned} A^D b &= X \begin{pmatrix} C^{-1} & 0 \\ 0 & 0 \end{pmatrix} X^{-1} b = X \begin{pmatrix} p(C) & 0 \\ 0 & 0 \end{pmatrix} X^{-1} b = X \begin{pmatrix} p(C) & 0 \\ 0 & p(N) \end{pmatrix} X^{-1} b \\ &= p(A) b \in \mathcal{K}_{m-i}(A, b). \end{aligned}$$

Therefore the Drazin inverse solution $A^D b$ is a Krylov solution!

Moreover, it's the only Krylov solution! To see this, observe that each solution of $Ax = b$ can be expressed as $x = A^D b + y$ for some $y \in N(A)$. Consequently, if x lies in a Krylov space then so does y. Write $y = \sum_{j=0}^{r} \alpha_j A^j b$, and use the fact that $b \in R(A^i)$ to conclude

$$X^{-1}y = \sum_{j=0}^{r} \alpha_j X^{-1} A^j X X^{-1} b = \sum_{j=0}^{r} \alpha_j \begin{pmatrix} C^j z_1 \\ 0 \end{pmatrix}.$$

But $Ay = 0$ implies $X^{-1}AXX^{-1}y = 0$, and $C\left[\sum_{j=0}^{r} \alpha_j C^j z_1\right] = 0$. Since C is nonsingular, $\sum_{j=0}^{r} \alpha_j C^j z_1 = 0$. This implies $X^{-1}y = 0$, and ultimately $y = 0$. Therefore the Drazin inverse solution is the unique Krylov solution. This means we have proved the following statement.

Theorem 3 (Uniqueness of the Krylov Solution) *Let m be the degree of the minimal polynomial for A, and let i be the index of the zero eigenvalue of A.*

If $b \in R(A^i)$, then the linear system $Ax = b$ has a unique Krylov solution, which can be expressed as

$$x = A^D b \in \mathcal{K}_{m-i}(A, b).$$

If $b \notin R(A^i)$ then $Ax = b$ does not have a Krylov solution.

11 The Grand Finale

Combining all our results gives a complete statement about Krylov solutions.

Summary 1 *Let m be the degree of the minimal polynomial for $A \in \mathcal{C}^{n \times n}$, and let i be the index of the zero eigenvalue of A.*

- *The linear system $Ax = b$ has a Krylov solution if and only if $b \in R(A^i)$.*
- *When a Krylov solution exists, it is unique and equal to the Drazin inverse solution*

$$x = A^D b \in \mathcal{K}_{m-i}(A, b).$$

- *Every consistent system $Ax = b$ with diagonalizable coefficient matrix A has a Krylov solution*

$$x = A^D b \in \begin{cases} \mathcal{K}_{d-1}(A, b) & \text{if } A \text{ is singular,} \\ \mathcal{K}_d(A, b) & \text{if } A \text{ is nonsingular,} \end{cases}$$

where d is the number of distinct eigenvalues of A.

12 The Other Half of the Krylov Story

The preceding discussion does not completely explain the popularity of Krylov methods. As we have seen, the dimension of a Krylov space containing a solution to a linear system cannot exceed the order n of the matrix. This means that the space in which we search for a solution has dimension at most n.

At first sight, this looks like good news because we need not iterate indefinitely to solve the system. But in practice n can be very large and the dimension of the search space is often equal to n. This occurs frequently. For example, (5) implies that this is the case when A has distinct eigenvalues. For large problems it is therefore not practical to execute anywhere near n iterations. As a consequence, Krylov algorithms are used as iterative methods. This means, they are prematurely terminated, long before all n iterations have been completed. The other half of the story revolves around the issue of how to insure that a small number of iterations delivers an approximate solution that is reasonably accurate.

Statement (5) provides the clue. If we can multiply $Ax = b$ by a nonsingular matrix M so that the coefficient matrix in $MAx = Mb$ is diagonalizable with only a few distinct eigenvalues, then a solution can be found in a Krylov space of small dimension. The process of pre- or postmultiplying the linear system to reduce the number of iterations in a Krylov method is called 'preconditioning'.

Of course, there is a delicate trade-off between reduction of search space versus the cost of obtaining the preconditioner M. Consider, for example, the extreme case $M = A^{-1}$. The search space is minimal (it has dimension one), but the construction of the preconditioner is as expensive as the solution of the original system, so we have gained nothing.

Although a diagonalizable MA with few distinct eigenvalues may not be cheap to come by, one can often exploit the structure of the underlying physical problem to construct preconditioners that deliver a diagonalizable MA whose eigenvalues fall into a few clusters, say t of them. If the diameters of the clusters are small enough, then MA behaves numerically like a matrix with t distinct eigenvalues. As a result, t iterations of a Krylov method tend to produce reasonably accurate approximations. While the intuition is simple, rigorous arguments are not always easy to establish. Different algorithms require different techniques, and this has been the focus of much work. The ideas for GMRES in [CIKM96] illustrate this.

Constructing good preconditioners and then proving they actually work as advertised is the other half of the Krylov story, and this continues to be an active area of research in numerical analysis.

13 Remarks

There are a couple of things we still need to discuss.

1. If one replaces the minimal polynomial of the matrix A by the minimal polynomial of the right-hand side b [Hou64, Section 1.5], [Fad59, p 155] one gets the precise value for the dimension of the Krylov space harboring x.

2. Many Krylov methods express the iterates as $x_k = x_0 + p_k$, where x_0 is a (not necessarily zero) initial guess and p_k is a so-called direction vector.

   We retain the context of the preceding discussion by incorporating the initial guess into the right-hand side, $r_0 \equiv b - Ax_0$. Instead of $Ax = b$, we solve $Ap = r_0$ and recover the solution from $x = x_0 + p$. Thus r_0 replaces b, p replaces x, and p_k replaces x_k.

3. When A is Hermitian (or real symmetric), the matrix $V_j^* A V_j$ in GMRES is also Hermitian (or real symmetric) and the matrix H_j is tridiagonal. Hence the operation count of a GMRES iteration is fixed and independent of the iteration number. Therefore the cost of t GMRES iterations is proportional to the cost of only t matrix vector products.

   Like GMRES, many other Krylov methods are equally cheap when applied to a Hermitian (or real symmetric) matrix. If, in addition, the matrix is also positive-definite, the number of iterations required to produce a reasonably accurate solution tends to be especially small.

4. There is an incredible amount of literature on Krylov methods for solving linear systems. We only mention the books by Axelson [Axe94], Golub and van Loan [Gv89], Kelley [Kel95] and Saad [Saa96]; and the survey paper by Freund, Nachtigal and Golub [FGN92].

Acknowledgements

We thank Stan Eisenstat and Tim Kelley for helpful discussions.

References

[Axe94] O. Axelson. *Iterative Solution Methods.* Cambridge University Press, Cambridge, 1994.

[CIKM96] S.L. Campbell, I.C.F. Ipsen, C.T. Kelley, and C.D. Meyer. GMRES and the minimal polynomial. *BIT*, 36(4):664–75, 1996.

[CM79] S.L. Campbell and C.D. Meyer. *Generalized Inverses of Linear Transformations.* Dover Publications, New York, 1979.

[Dra68] M.P. Drazin. Pseudoinverses in associate rings and semigroups. *Amer. Math. Monthly*, 65:506–14, 1968.

[Fad59] V.N. Faddeeva. *Computational Methods of Linear Algebra.* Dover, New York, NY, USA, 1959.

[FGN92] R.W. Freund, G.H. Golub, and N.M. Nachtigal. *Iterative Solution of Linear Systems*, pages 57–100. Cambridge University Press, 1992.

[Gv89] G.H. Golub and C.F. van Loan. *Matrix Computations.* The Johns Hopkins Press, Baltimore, second edition, 1989.

[Hou64] A.S. Householder. *The Theory of Matrices in Numerical Analysis.* Dover Publications, 1964.

[Kel95] C.T. Kelley. *Iterative Methods for Linear and Nonlinear Equations.* SIAM, Philadelphia, 1995.

[Saa96] Y. Saad. *Iterative Methods for Sparse Linear Systems.* PWS Publishing Company, Boston, 1996.

[SS86] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Sci. Stat. Comput.*, 7(3):856–69, 1986.