The Idea Behind Krylov Methods

Ilse C.F. Ipsen and Carl D. Meyer

1 Introduction

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 can be 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 a 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.

Our discussion demonstrates that eigenvalues play a central role when it comes to ensuring existence and uniqueness of Krylov solutions; they are not merely an artifact of convergence analyses.

2 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? Perhaps the matrix A exists only implicitly as a subroutine that, when given a vector v, returns Av.

In this case you may want to use a Krylov method. Krylov methods are used in numerical as well as in symbolic computation [7], [14]. Since there is no universally agreed upon definition, we say here that a *Krylov method* solves $Ax = b$ by repeatedly performing matrix-vector multiplications involving A [18, §6.1] (this excludes methods like Lanczos biorthogonalization, QMR, and biconjugate gradient methods that also require matrix-vector multiplications involving the conjugate transpose A^*).

Starting with an initial guess x_0, a Krylov method bootstraps its way up (we hope!) to ever more accurate approximations x_k to a desired solution. In iteration k a Krylov method produces an approximate solution x_k from a *Krylov space generated by a vector c*,

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

A popular choice is $c = b$ (because one can obtain convergence estimates, and because there is often no other problem-dependent guess) and $x_0 = 0$ (we deal with a nonzero x_0 in §9). That's why we restrict ourselves to Krylov spaces $\mathcal{K}_k(A,\, b)$ that are generated by the right-hand side b of a linear system $Ax = b$.

Let's look at a specific example.

3 An Example of a Krylov Method

The generalized minimal residual method (GMRES) was published by Saad and Schultz in 1986 [19]. 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

$$\min_{z \in \mathcal{K}_k(A,\, b)} \|b - Az\| \tag{1}$$

in the Euclidean norm $\|\cdot\|$. 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's method recursively builds an orthonormal basis for $\mathcal{K}_{j+1}(A,\, b)$ from an orthonormal basis for $\mathcal{K}_j(A,\, b)$ by orthogonalizing the vector Av_j from $\mathcal{K}_{j+1}(A,\, b)$ against the previous space $\mathcal{K}_j(A,\, b)$. That is,

$$\hat{v}_{j+1} = Av_j - \left(h_{1j}v_1 + \cdots + h_{jj}v_j\right), \tag{2}$$

where $h_{ij} = v_i^* A v_j$ and * denotes the conjugate transpose. 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, $V_j = \left(\, v_1 \ldots v_j \,\right)$, we get the decomposition associated with Arnoldi's method:

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

where H_j is an upper Hessenberg matrix of size $(j+1) \times j$ (an upper triangular matrix with an additional off-diagonal below the diagonal).

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 $x_0 = 0, \quad v_1 = b/\beta, \quad V_1 = v_1.$

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: $V_{k+1} = \begin{pmatrix} V_k & v_{k+1} \end{pmatrix}, \qquad H_k = \begin{pmatrix} H_{k-1} & h_k \\ 0 & \|\hat{v}_{k+1}\| \end{pmatrix},$
   where the first column in H_k is omitted when $k = 1$.
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 when it produces 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. Arnoldi's method implies $AV_s = V_s \hat{H}_s$. This means the columns of V_s span an invariant subspace of A and the eigenvalues of $\hat{H}_s$ are eigenvalues of A. Since A has no zero eigenvalues, neither does $\hat{H}_s$. Thus $\hat{H}_s$ is nonsingular, and the least squares problem reduces to a nonsingular linear system $\hat{H}_s y_s = \beta e_1$. From $AV_s = V_s \hat{H}_s$ follows

$$AV_s y_s = V_s \hat{H}_s y_s = \beta V_s e_1 = b,$$

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 a space of dimension n can accommodate at most n linearly independent vectors.

Therefore, GMRES works properly (our discussion is restricted to exact arithmetic; we ignore the effects of floating point arithmetic).

In practice a Krylov method like GMRES is not run to completion but is terminated prematurely as soon as an iterate is deemed to be good enough. This may mean that the residual norm $\|Ax_k - b\|$ is sufficiently small or that some other convergence criterion is satisfied. In order to retain our focus on the common features of Krylov methods, we assume until Section 9 that they are always run to completion.

4 Questions

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

Like GMRES, these methods tend to provide acceptable solutions in a number of iterations much less than the order of A. Just how few iterations are required depends on the eigenvalues (or pseudo eigenvalues [17]) 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, those who are not experts may not readily see the common ground shared by 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 $\mathcal{K}_k(A, b)$ often a good space from which to construct an approximate solution?

   At first sight Krylov methods did not strike us as a natural way to solve linear systems. In contrast to factorization-based methods, like Gaussian elimination, Krylov methods must expend extra effort to solve a system whose number of equations differs 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 affect the sensitivity of a linear system. Moreover, the number of zero singular values determines the dimension of the space containing all b for which $Ax = b$ has a solution.

3. Why do Krylov methods often do so well for Hermitian 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 matrix?

Strategy If we can show that the solution to $Ax = b$ has a 'natural' representation as a member of a Krylov space $\mathcal{K}_k(A, b)$, then we can understand why one would construct approximations to x from this space. If the dimension of $\mathcal{K}_k(A, b)$ is small then a Krylov method has an opportunity to find x in few iterations. This is why we select as our gauge for convergence the dimension of the smallest Krylov space $\mathcal{K}_k(A, b)$ containing x. If this dimension is small, we have a plausible reason to expect rapid convergence (in practice, convergence may be judged not only by the number of iterations but also by some estimate for error reduction).

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 $\mathcal{K}_k(A, b)$ we 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 the Krylov space $\mathcal{K}_n(A, b)$.

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

5 The Idea

The minimal polynomial $q(t)$ of A is the unique monic polynomial of minimal degree such that $q(A) = 0$. It is constructed from the eigenvalues of A as follows. If the distinct eigenvalues of A are $\lambda_1, \ldots, \lambda_d$ and if λ_j has 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.

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

$$0 = q(A) = \alpha_0 I + \alpha_1 A + \cdots + \alpha_m A^m,$$

where I is the identity matrix and $\alpha_0 \neq 0$, 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 (see the remark concerning 'the minimal polynomial of b' in Section 10). If the minimal polynomial has low degree then the Krylov space containing the solution is small, and a Krylov method has an opportunity to converge fast.

Example Theorem 1 suggests that a Krylov space should have maximal dimension 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

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

Saad and Schultz [19] have shown that the maximal number of iterations in GMRES does not exceed the degree of the minimal polynomial of A.

Summary At this point we have answered Questions 1 and 2 in Section 4 for non-singular matrices: The space $\mathcal{K}_k(A,\, b)$ is a good space from which to construct approximate solutions for a non-singular linear system $Ax = b$ because it is intimately tied to the inverse of the matrix. Eigenvalues are important for Krylov methods because the dimension of the solution space is determined by the degree of the minimal polynomial of the matrix.

To complete the answer to Questions 1 and 2 we need to look at singular matrices. Although singular systems are not as abundant in practice as nonsingular systems, they do occur [2, Chapt. 7], and we cannot take for granted existence and uniqueness of a solution in $\mathcal{K}_n(A,\, b)$.

6 Why Are Singular Systems Different?

Suppose a linear system has a singular coefficient matrix. Even if a solution exists, it may not lie in the Krylov space $\mathcal{K}_n(A,\, b)$. 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 an i such that $N^i = 0$ but $N^{i-1} \neq 0$. Suppose a 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

$$c = Nx = \xi_0 Nc + \cdots + \xi_{i-2} N^{i-1} c \qquad \text{and} \qquad (I - \xi_0 N - \cdots - \xi_{i-2} N^{i-1}) c = 0.$$

But the matrix in parentheses is nonsingular. Its eigenvalues are all equal to one, because the sum of the terms containing N is nilpotent. Consequently, $c = 0$. In other words, a solution to a nilpotent system with nonzero right-hand side cannot lie in the Krylov space $\mathcal{K}_n(A,\, b)$.

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 must restrain b by somehow keeping it away from the 'nilpotent part' of A.

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 $R(\cdot)$ and $N(\cdot)$ denote range and nullspace. Let's assume that A is a Jordan matrix with all zero eigenvalues at the bottom. Then the space decomposition induces the matrix decomposition

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

where C is nonsingular and N is nilpotent of index i.

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

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

Partitioning the vectors conformally with the matrix,

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

gives

$$x_1 = \sum_{j=0}^{p} \alpha_j C^j b_1 \qquad \text{and} \qquad x_2 = \sum_{j=0}^{p} \alpha_j N^j b_2.$$

But $Ax = b$ implies $Nx_2 = b_2$, so $N(\sum_{j=0}^{p} \alpha_j N^j b_2) = b_2$ and

$$(I - \textstyle\sum_{j=0}^{p} \alpha_j N^{j+1})b_2 = 0.$$

The matrix in parentheses is nonsingular and $b_2 = 0$. Thus $b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \in R(A^i)$. 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

$$b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix}, \qquad \text{and} \qquad x = \begin{pmatrix} C^{-1}b_1 \\ 0 \end{pmatrix}$$

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 Section 5 to the matrix C. The minimal polynomial for C has degree $m - i$, and there is a polynomial $p(x)$ of degree $m - i - 1$ such that $C^{-1} = p(C)$. Substituting this polynomial into the expression for x gives

$$x = \begin{pmatrix} C^{-1}b_1 \\ 0 \end{pmatrix} = \begin{pmatrix} p(C) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ 0 \end{pmatrix} = \begin{pmatrix} p(C) & 0 \\ 0 & p(N) \end{pmatrix} b = p(A)b \, \in \mathcal{K}_{m-i}(A, b).$$

Therefore $b \in R(A^i)$ guarantees the existence of a Krylov solution. The proof is only slightly more complicated, but not very different, when A is not a Jordan matrix.

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 is 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, $i = 1$ and the condition on b reduces to the familiar consistency condition $b \in R(A)$. This occurs, for example, 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 largest Krylov space for the singular case has shrunk. Its dimension is smaller by i than the degree of the minimal polynomial. The index of the zero eigenvalue affects the dimension of the search space as well as the dimension of the space of right-hand sides that admit a solution in $\mathcal{K}_n(A, b)$. In particular, as the defectiveness of the zero eigenvalue grows the search space shrinks, and so does the space of desirable right-hand sides. This answers Question 2 in Section 4: Eigenvalues are important for Krylov methods because the index of the zero eigenvalue affects the existence of a solution in $\mathcal{K}_n(A, b)$.

We still have to answer Question 1 for singular matrices: Why is a Krylov space a good space from which to construct an approximate solution for $Ax = b$? In the non-singular case we argued that A^{-1} can be expressed as a polynomial in A and is therefore intimately tied to $\mathcal{K}_k(A, b)$. But now we don't have an inverse. So let's look for a suitable pseudo-inverse. 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 [5, Section 7.5]. So let's give the Drazin inverse a try.

7 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

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

see [6], [5, Section 7.5]. If 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, this one can be answered by decomposing the Drazin inverse conformably with the Jordan form of A. If

$$A = X \begin{pmatrix} C & 0 \\ 0 & N \end{pmatrix} X^{-1}, \quad \text{then} \quad A^D = X \begin{pmatrix} C^{-1} & 0 \\ 0 & 0 \end{pmatrix} X^{-1} \tag{6}$$

where C and N are the same as in (4). 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 solution in the Krylov space $\mathcal{K}_n(A, b)$.*

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 in $\mathcal{K}_{m-i}(A, b)$! To see this, assume for simplicity that A is a Jordan matrix (4). 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 $b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \in R(A^i)$ to conclude

$$y = \sum_{j=0}^{r} \alpha_j \begin{pmatrix} C^j b_1 \\ 0 \end{pmatrix}.$$

But $Ay = 0$ implies $C \left[\sum_{j=0}^{r} \alpha_j C^j b_1 \right] = 0$. Since C is nonsingular, $\sum_{j=0}^{r} \alpha_j C^j b_1 = 0$. Hence $y = 0$. Therefore the Drazin inverse solution is the unique Krylov solution. The proof is slightly more complicated, but not very different, when A is not a Jordan matrix.

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 $x = A^D b \in \mathcal{K}_{m-i}(A, b)$. If $b \notin R(A^i)$ then $Ax = b$ does not have a solution in the Krylov space $\mathcal{K}_n(A, b)$.*

Finally we have answered Question 2 in Section 4 for singular matrices: A Krylov space $\mathcal{K}_k(A, b)$ is a good space from which to construct an approximate solution to a singular system $Ax = b$ because when it is large enough it contains a unique pseudo-inverse solution (provided b lies in $R(A^i)$).

8 The Grand Finale

Combining all our results gives a complete statement about Krylov solutions in $\mathcal{K}_n(A, b)$.

Summary: 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 in $\mathcal{K}_n(A, b)$ if and only if $b \in R(A^i)$.
- When a Krylov solution exists, it is unique and is 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.

9 Krylov Methods in Practice

The preceding discussion does not completely explain the popularity of Krylov methods. In practice, it is not good enough to know that the dimension of a Krylov space is bounded by n, because n can be very large and the dimension of the search space can be equal to n. For example, matrices stored in finite precision arithmetic tend to be non-singular with distinct eigenvalues, resulting in a search space of maximal dimension. For large linear systems it is not practical to execute anywhere near n iterations. As a consequence, Krylov algorithms are used as iterative methods. This means that 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 ensure that a small number of iterations delivers an approximate solution that is reasonably accurate.

Statement (5) provides the clue. Suppose it were possible to find a nonsingular matrix M that makes MA diagonalizable with only a few distinct eigenvalues. Then we would expect to find a solution to $MAx = Mb$ in a Krylov space of small dimension. Premultiplying (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 vs. 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 may be able to 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, we would expect t iterations of a Krylov method to produce a reasonably accurate approximation. 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 [4] illustrate this.

Constructing good preconditioners and then proving that 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.

10 Parting Remarks

Hermitian Matrices. We give only a cursory answer to Question 3 in Section 4: why Krylov methods often work well for Hermitian matrices. First, a consistent linear system $Ax = b$ with Hermitian coefficient matrix A always has a Krylov solution. Second, the eigenvector matrix of a Hermitian matrix may be chosen to be unitary, hence it is well-conditioned. If the Hermitian matrix A is also positive-definite, the number of iterations required to produce a satisfactory solution tends to be small.

Another reason is efficiency. Take GMRES, for instance. When A is Hermitian, $V_j^* A V_j$ is also Hermitian and H_j is tridiagonal. Hence the operation count of a GMRES iteration is 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 matrix.

The Minimal Polynomial of b. If we had replaced the minimal polynomial of the matrix A by the minimal polynomial of the right-hand side b in Section 5, we would have got the precise value for the dimension of the Krylov space containing x [13, Section 1.5], [9, p 155]. The minimal polynomial $q_b(t)$ of b accounts for a possible relation between A and b. It divides the minimal polynomial $q(t)$ of A and it annihilates b: $q_b(A)b = 0$. If p is the degree of $q_b(t)$ then $x \in \mathcal{K}_p(A, b)$, where p can be much smaller than the degree of $q(t)$.

A Nonzero Initial Guess. Many Krylov methods express the iterates as $x_k = x_0 + p_k$, where x_0 (not necessarily zero) is an *initial guess* and p_k is a *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 solving $Ax = b$, 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.

Further Reading. There is a vast literature on Krylov methods for solving nonsingular linear systems. We mention only the books by Axelson [1], Golub and van Loan [12], Kelley [15], and Saad [18]; and the survey paper by Freund, Nachtigal, and Golub [10]. They contain many references for further study. Our ideas about the use of the Drazin inverse for the solution of singular systems originated in work by Meyer and Plemmons [16] and Campbell and Meyer [5]. Related results can be found in papers by Eiermann, Marek, and Niethammer [8], Freund and Hochbruck [11], and Brown and Walker [3].

Acknowledgements

We thank Tim Kelley, Michele Benzi, and in particular Stan Eisenstat for helpful discussions. The work of the first author was supported in part by NSF grant CCR-9400921. The work of the second author was supported in part by NSF grant CCR-9413309.

References

[1] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, Cambridge, 1994.

[2] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* SIAM Classics In Applied Mathematics. SIAM, Philadelphia, 1994.

[3] P.N. Brown and H.F. Walker. GMRES on (nearly) singular systems. *SIAM J. Matrix Anal. Appl.*, 18 (1997), pp. 37–51.

[4] S.L. Campbell, I.C.F. Ipsen, C.T. Kelley, and C.D. Meyer. GMRES and the minimal polynomial. *BIT*, 36 (1996), pp. 664–675.

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

[6] M.P. Drazin. Pseudoinverses in associate rings and semigroups. *Amer. Math. Monthly*, 65 (1968), pp. 506–514.

[7] W. Eberly and E. Kaltofen. On randomized Lanczos algorithms. In W. Küchlin, editor, *Proc. Internat. Symp. Symbolic Algebraic Comput. ISSAC '97.* ACM Press, New York, 1997, pp. 176–183.

[8] M. Eiermann, I. Marek, and W. Niethammer. On the solution of singular linear systems of algebraic equations by semiiterative methods. *Numer. Math.*, 53 (1988), pp. 265–283.

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

[10] R.W. Freund, G.H. Golub, and N.M. Nachtigal. Iterative solution of linear systems. In *Acta Numerica 1992.* Cambridge University Press, 1992, pp. 57–100.

[11] R.W. Freund and M. Hochbruck. On the use of two QMR algorithms for solving singular systems and applications in Markov chain modeling. *Num. Linear Algebra Appl.*, 1 (1994), pp. 403–420.

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

[13] A.S. Householder. *The Theory of Matrices in Numerical Analysis.* Dover, New York, 1964.

[14] E. Kaltofen and A. Lobo. Distributed matrix-free solution of large sparse linear systems over finite fields. In A.M. Tentner, editor, *Proc. High Performance Computing '96.* Simulation Councils, Inc., San Diego, 1996, pp. 244–247.

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

[16] C.D. Meyer and R.J. Plemmons. Convergent powers of a matrix with applications to iterative methods for singular linear systems. *SIAM J. Numer. Anal.*, 14 (1977), pp. 699–705.

[17] N.M. Nachtigal, S.C. Reddy, and L.N. Trefethen. How fast are nonsymmetric matrix iterations? *SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 778–795.

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

[19] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Sci. Stat. Comput.*, 7 (1986), pp. 856–869.

Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205, USA
ipsen@math.ncsu.edu, meyer@math.ncsu.edu