57 The Implicitly Restarted Arnoldi Method

D. C. Sorensen
Rice University

The implicitly restarted Arnoldi method (IRAM) [Sor92] is a variant of Arnoldi's method for computing a selected subset of eigenvalues and corresponding eigenvectors for large matrices. Implicit restarting is a synthesis of the implicitly shifted QR iteration and the Arnoldi process that effectively limits the dimension of the Krylov subspace required to obtain good approximations to desired eigenvalues. The space is repeatedly expanded and contracted with each new Krylov subspace generated by an updated starting vector obtained by implicit application of a matrix polynomial to the old starting vector. This process is designed to filter out undesirable components in the starting vector in a way that enables convergence to the desired invariant subspace. This method has been implemented and is freely available as ARPACK. The MATLAB® function `eigs` is based upon ARPACK. Use of this software is described in Chapter 94.

In this chapter, all matrices, vectors, and scalars are complex and the algorithms are phrased in terms of complex arithmetic. However, when the matrix (or matrix pair) happens to be real then the computations may be organized so that only real arithmetic is required. Multiplication of a vector $\mathbf{x}$ by a scalar λ is denoted by $\mathbf{x}\lambda$ so that the eigenvector–eigenvalue relation is $A\mathbf{x} = \mathbf{x}\lambda$. This convention provides for direct generalizations to the more general invariant subspace relations $AX = XH$, where X is an $n \times k$ matrix and H is a $k \times k$ matrix with $k < n$. More detailed discussion of all facts and definitions may be found in the overview article [Sor02].

57.1 Krylov Subspace Projection

The classic power method is the simplest way to compute the dominant eigenvalue and corresponding eigenvector of a large matrix. Krylov subspace projection provides a way to extract additional eigen-information from the power method iteration by considering all possible linear combinations of the sequence of vectors produced by the power method.

Definitions:

The best approximate eigenvectors and corresponding eigenvalues are extracted from the **Krylov subspace**

$$\mathcal{K}_k(A, \mathbf{v}) := \text{span}\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^{k-1}\mathbf{v}\}.$$

The approximate eigenpairs are constructed through a Galerkin condition. An approximate eigenvector $\mathbf{x} \in \mathcal{S}$ is called a **Ritz vector** with corresponding **Ritz value** θ if the **Galerkin condition**

$$\mathbf{w}^*(A\mathbf{x} - \mathbf{x}\theta) = 0, \quad \text{for all} \quad \mathbf{w} \in \mathcal{K}_k(A, \mathbf{v})$$

is satisfied.

Facts: [Sor92], [Sor02]

1. Every $\mathbf{w} \in \mathcal{K}_k$ is of the form $\mathbf{w} = \phi(A)\mathbf{v}_1$ for some polynomial ϕ of degree less than k and $\mathcal{K}_{j-1} \subset \mathcal{K}_j$ for $j = 2, 3, \ldots, k$.
2. If a sequence of orthogonal bases $V_k = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$ has been constructed with $\mathcal{K}_k = \text{range}(V_k)$ and $V_k^* V_k = I_k$, then a new basis vector $\mathbf{v}_{k+1}$ is obtained by the **projection formulas**

$$\begin{aligned} \mathbf{h}_k &= V_k^* A\mathbf{v}_k, \\ \mathbf{f}_k &= A\mathbf{v}_k - V_k\mathbf{h}_k, \\ \mathbf{v}_{k+1} &= \mathbf{f}_k / \|\mathbf{f}_k\|_2. \end{aligned}$$

   The vector $\mathbf{h}_k$ is constructed to achieve $V_k^*\mathbf{f}_k = 0$ so that $\mathbf{v}_{k+1}$ is a vector of unit length that is orthogonal to the columns of V_k.
3. The columns of $V_{k+1} = [V_k, \mathbf{v}_{k+1}]$ provide an orthonormal basis for $\mathcal{K}_{k+1}(A, \mathbf{v}_1)$.
4. The basis vectors are of the form $\mathbf{v}_j = \phi_{j-1}(A)\mathbf{v}_1$, where ϕ_{j-1} is a polynomial of degree $j-1$ for each $j = 1, 2, \ldots, k+1$.
5. This construction fails when $\mathbf{f}_k = 0$, but then

$$AV_k = V_k H_k,$$

   where $H_k = V_k^* A V_k = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k]$ (with a slight abuse of notation). This "good breakdown" happens precisely when $\mathcal{K}_k$ is an invariant subspace of A. Hence, $\sigma(H_k) \subset \sigma(A)$.

57.2 The Arnoldi Factorization

The projection formulas given above result in the fundamental Arnoldi method for constructing an orthonormal basis for $\mathcal{K}_k$.

Definitions:

The relations between the matrix A, the basis matrix V_k, and the residual vector $\mathbf{f}_k$ may be concisely expressed as

$$AV_k = V_k H_k + \mathbf{f}_k \mathbf{e}_k^*,$$

where $V_k \in \mathbb{C}^{n \times k}$ has orthonormal columns, $V_k^*\mathbf{f}_k = 0$, and $H_k = V_k^* A V_k$ is a $k \times k$ upper Hessenberg matrix with nonnegative subdiagonal elements.

The above expression shall be called a **k-step Arnoldi factorization** of A.

When A is Hermitian, H_k will be real, symmetric, and tridiagonal and then the relation is called a **k-step Lanczos factorization** of A.

The columns of V_k are referred to as **Arnoldi vectors** or **Lanczos vectors**, respectively.

The Hessenberg matrix H_k is called **unreduced** if all subdiagonal elements are nonzero.

Facts: [Sor92], [Sor02]

1. The explicit steps needed to form a k-step Arnoldi factorization are shown in Algorithm 1.

Algorithm 1: k-step Arnoldi factorization. A square matrix A, a nonzero vector $\mathbf{v}$, and a positive integer $k \leq n$ are input.

Output is an $n \times k$ ortho-normal matrix V_k, an upper Hessenberg matrix H_k and a vector $\mathbf{f}_k$ such that $AV_k = V_k H_k + \mathbf{f}_k \mathbf{e}_k^T$.

$\mathbf{v}_1 = \mathbf{v}/\|\mathbf{v}\|_2$;
$\mathbf{w} = A\mathbf{v}_1$; $\alpha_1 = \mathbf{v}_1^* \mathbf{w}$;
$\mathbf{f}_1 \leftarrow \mathbf{w} - \mathbf{v}_1 \alpha_1$;
$V_1 \leftarrow [\mathbf{v}_1]$; $H_1 \leftarrow [\alpha_1]$;
for $j = 1, 2, 3, \ldots k-1$,
  $\beta_j = \|\mathbf{f}_j\|_2$; $\mathbf{v}_{j+1} \leftarrow \mathbf{f}_j/\beta_j$;
  $V_{j+1} \leftarrow [V_j, \mathbf{v}_{j+1}]$;
  $\hat{H}_j \leftarrow \begin{bmatrix} H_j \\ \beta_j \mathbf{e}_j^* \end{bmatrix}$;
  $\mathbf{w} \leftarrow A\mathbf{v}_{j+1}$;
  $\mathbf{h} \leftarrow V_{j+1}^* \mathbf{w}$;
  $\mathbf{f}_{j+1} \leftarrow \mathbf{w} - V_{j+1}\mathbf{h}$;
  $H_{j+1} \leftarrow [\hat{H}_j, \mathbf{h}]$;
end

2. Ritz pairs satisfying the Galerkin condition (see Section 57.1) are derived from the eigenpairs of the small projected matrix H_k. If $H_k \mathbf{y} = \mathbf{y}\theta$ with $\|\mathbf{y}\|_2 = 1$, then the vector $\mathbf{x} = V_k \mathbf{y}$ is a vector of unit norm that satisfies

$$\|A\mathbf{x} - \mathbf{x}\theta\|_2 = \|(AV_k - V_k H_k)\mathbf{y}\|_2 = |\beta_k \mathbf{e}_k^* \mathbf{y}|,$$

where $\beta_k = \|\mathbf{f}_k\|_2$.

3. If $(\mathbf{x}, \theta)$ is a Ritz pair constructed as shown in Fact 2, then

$$\theta = \mathbf{y}^* H_k \mathbf{y} = (V_k \mathbf{y})^* A (V_k \mathbf{y}) = \mathbf{x}^* A \mathbf{x}$$

is always a Rayleigh quotient (assuming $\|\mathbf{y}\|_2 = 1$).

4. The Rayleigh quotient residual $\mathbf{r}(\mathbf{x}) := A\mathbf{x} - \mathbf{x}\theta$ satisfies $\|\mathbf{r}(\mathbf{x})\|_2 = |\beta_k \mathbf{e}_k^* \mathbf{y}|$. When A is Hermitian, this relation provides computable rigorous bounds on the accuracy of the approximate eigenvalues [Par80]. When A is non-Hermitian, one needs additional sensitivity information. Nonnormality effects may corrupt the accuracy. In exact arithmetic, these Ritz pairs are eigenpairs of A whenever $\mathbf{f}_k = 0$. However, even with a very small residual these may be far from actual eigenvalues when A is highly nonnormal.

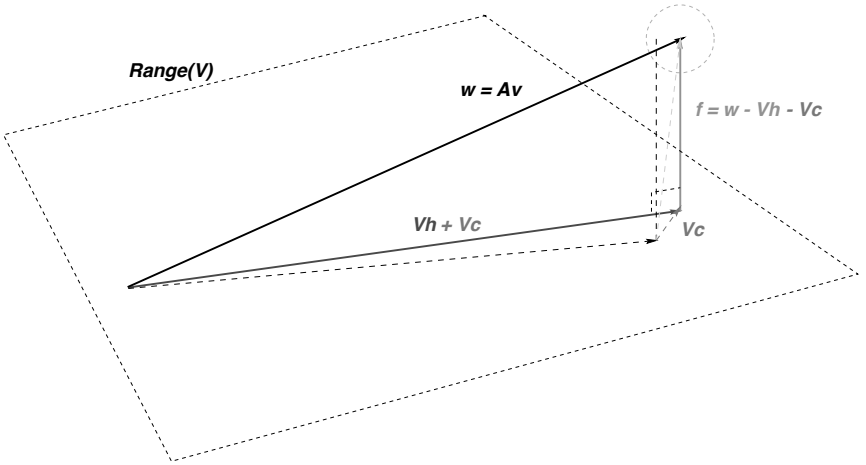

FIGURE 57.1 DGKS Correction.

5. The orthogonalization process is based upon the *classical Gram–Schmidt* (CGS) scheme. This process is notoriously unstable and will fail miserably in this application without modification. The iterative refinement technique proposed by Daniel, Gragg, Kaufman, and Stewart (DGKS) [DGK76] provides an excellent way to construct a vector $\mathbf{f}_{j+1}$ that is numerically orthogonal to V_{j+1}. It amounts to computing a correction
$$\mathbf{c} = V_{j+1}^{*}\mathbf{f}_{j+1}; \quad \mathbf{f}_{j+1} \leftarrow \mathbf{f}_{j+1} - V_{j+1}\mathbf{c}; \quad \mathbf{h} \leftarrow \mathbf{h} + \mathbf{c};$$
just after computing $\mathbf{f}_{j+1}$ if necessary, i.e., when $\mathbf{f}_{j+1}$ is not sufficiently orthogonal to the columns of V_{j+1}. This formulation is crucial to both accuracy and performance. It provides numerically orthogonal basis vectors and it may be implemented using the Level 2 BLAS operation _GEMV [DDH88]. This provides a significant performance advantage on virtually every platform from workstation to supercomputer.
6. The *modified Gram–Schmidt* (MGS) process will generally fail to produce orthogonal vectors and cannot be implemented with Level 2 BLAS in this setting. ARPACK relies on a restarting scheme wherein the goal is to reach a state of dependence in order to obtain $\mathbf{f}_k = 0$. MGS is completely inappropriate for this situation, but the CGS with DGKS correction performs beautifully.
7. Failure to maintain orthogonality leads to numerical difficulties in the Lanczos/Arnoldi process. Loss of orthogonality typically results in the presence of spurious copies of the approximate eigenvalue.

Examples:

1. Figure 57.1 illustrates how the DGKS mechanism works. When the vector $\mathbf{w} = A\mathbf{v}$ is nearly in the range(V), then the projection $V\mathbf{h}$ is possibly inaccurate, but vector $= \mathbf{w} - \mathbf{Vh}$ is not close to range(V) and can be safely orthogonalized to compute the correction $\mathbf{c}$ accurately. The corrected vector $\mathbf{f} \leftarrow \mathbf{f} - V\mathbf{c}$ will be numerically orthogonal to the columns of V in almost all cases. Additional corrections might be necessary in very unusual cases.

57.3 Restarting the Arnoldi Process

The number of Arnoldi steps required to calculate eigenvalues of interest to a specified accuracy cannot be pre-determined. Usually, eigen-information of interest does not appear until k gets very large. In Figure 57.2 the distribution in the complex plane of the Ritz

values (shown in grey dots) is compared with the spectrum (shown as +s). The original matrix is a normally distributed random matrix of order 200 and the Ritz values are from a ($k = 50$)-step Arnoldi factorization. Eigenvalues at the extremes of the spectrum of A are clearly better approximated than the interior eigenvalues.

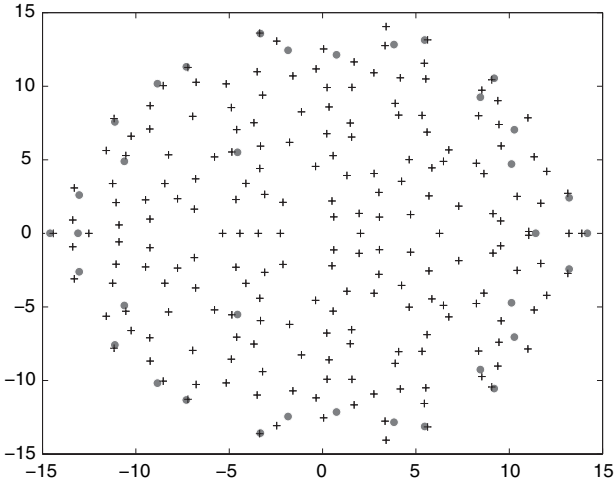

FIGURE 57.2 Typical distribution of Ritz values.

For large problems, it is intractable to compute and store a numerically orthogonal basis set V_k for large k. Storage requirements are $\mathcal{O}(n \cdot k)$ and arithmetic costs are $\mathcal{O}(n \cdot k^2)$ flops to compute the basis vectors plus $\mathcal{O}(k^3)$ flops to compute the eigensystem of H_k. Thus, restarting schemes have been developed that iteratively replace the starting vector $\mathbf{v}_1$ with an "improved" starting vector $\mathbf{v}_1^+$ and then compute a new Arnoldi factorization of fixed length k to limit the costs. Beyond this, there is an interest in forcing $\mathbf{f}_k = 0$ and, thus, producing an invariant subspace. However, this is useful only if the spectrum $\sigma(H_k)$ has the desired properties.

The structure of $\mathbf{f}_k$ suggests the restarting strategy. The goal will be to iteratively force $\mathbf{v}_1$ to be a linear combination of eigenvectors of interest.

Facts: [Sor92], [Sor02]

1. If $\mathbf{v} = \sum_{j=1}^{k} \mathbf{q}_j \gamma_j$ where $A\mathbf{q}_j = \mathbf{q}_j \lambda_j$ and
$$AV = VH + \mathbf{f}\mathbf{e}_k^T$$
is a k-step Arnoldi factorization with unreduced H, then $\mathbf{f} = 0$ and $\sigma(H) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$.
2. Since $\mathbf{v}_1$ determines the subspace $\mathcal{K}_k$, this vector must be constructed to select the eigenvalues of interest. The starting vector must be forced to become a linear combination of eigenvectors that span the desired invariant subspace. There is a necessary and sufficient condition for $\mathbf{f}$ to vanish that involves Schur vectors and does not require diagonalizability.

57.4 Polynomial Restarting

Polynomial restarting strategies replace $\mathbf{v}_1$ by

$$\mathbf{v}_1 \leftarrow \psi(A)\mathbf{v}_1,$$

where ψ is a polynomial constructed to damp unwanted components from the starting vector. If $\mathbf{v}_1 = \sum_{j=1}^{n} \mathbf{q}_j \gamma_j$ where $A\mathbf{q}_j = \mathbf{q}_j \lambda_j$, then

$$\mathbf{v}_1^+ = \psi(A)\mathbf{v}_1 = \sum_{j=1}^{n} \mathbf{q}_j \gamma_j \psi(\lambda_j),$$

where the polynomial ψ has also been normalized to give $\|\mathbf{v}_1\|_2 = 1$. Motivated by the structure of $\mathbf{f}_k$, the idea is to force the starting vector to be closer and closer to an invariant subspace by constructing ψ so that $|\psi(\lambda)|$ is as small as possible on a region containing the unwanted eigenvalues.

An iteration is defined by repeatedly restarting until the updated Arnoldi factorization eventually contains the desired eigenspace. An explicit scheme for restarting was proposed by Saad in [Saa92]. One of the more successful choices is to use Chebyshev polynomials in order to damp unwanted eigenvector components.

Definitions:

The polynomial ψ is sometimes called a **filter polynomial**, which may also be specified by its roots.

The roots of the filter polynomial may also be referred to as **shifts**. This terminology refers to their usage in an implicitly shifted QR-iteration.

One straightforward choice of shifts is to find the eigenvalues θ_j of the projected matrix H and sort these into two sets according to a given criterion: the wanted set $\mathcal{W} = \{\theta_j : j = 1, 2, \ldots, k\}$ and the unwanted set $\mathcal{U} = \{\theta_j : j = k+1, k+2, \ldots, k+p\}$. Then one specifies the polynomial ψ as the polynomial with these unwanted Ritz values as it roots. This choice of roots, called **exact shifts**, was suggested in [Sor92].

Facts: [Sor92], [Sor02]

1. Morgan [Mor96] found a remarkable property of this strategy. If exact shifts are used to define $\psi(\tau) = \prod_{j=k+1}^{k+p}(\tau - \theta_j)$ and if $\hat{\mathbf{q}}_j$ denotes a Ritz vector of unit length corresponding to θ_j, then the Krylov space generated by $\mathbf{v}_1^+ = \psi(A)\mathbf{v}_1$ satisfies

   $$\mathcal{K}_m(A, \mathbf{v}_1^+) = Span\{\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_k, A\hat{\mathbf{q}}_j, A^2\hat{\mathbf{q}}_j, \ldots, A^p\hat{\mathbf{q}}_j\},$$

   for any $j = 1, 2, \ldots, k$. Thus, polynomial restarting with exact shifts will generate a new subspace that contains all of the possible choices of updated staring vector consisting of linear combinations of the wanted Ritz vectors.
2. Exact shifts tend to perform remarkably well in practice and have been adopted as the shift selection of choice in ARPACK when no other information is available. However, there are many other possibilities such as the use of Leja points for certain containment regions or intervals [BCR96].

57.5 Implicit Restarting

There are a number of schemes used to implement polynomial restarting. We shall focus on an implicit restarting scheme.

Definitions:

A straightforward way to implement polynomial restarting is to explicitly construct the starting vector $\mathbf{v}_1^+ = \psi(A)\mathbf{v}_1$ by applying $\psi(A)$ through a sequence of matrix-vector products. This is called **explicit restarting**.

A more efficient and numerically stable alternative is **implicit restarting**. This technique applies a sequence of implicitly shifted QR steps to an m-step Arnoldi or Lanczos factorization to obtain a truncated form of the implicitly shifted QR-iteration.

On convergence, the IRAM iteration (see Algorithm 2) gives an orthonormal matrix V_k and an upper Hessenberg matrix H_k such that $AV_k \approx V_k H_k$.

If $H_k Q_k = Q_k R_k$ is a Shur decompositon of H_k, then we call $\hat{V}_k \equiv V_k Q_k$ a **Schur basis** for the Krylov subspace $\mathcal{K}_k(A, \mathbf{v}_1)$.

Note that if $AV_k = V_k H_k$ exactly, then $\hat{V}_k$ would form the leading k columns of a unitary matrix $\hat{V}$ and R_k would form the leading $k \times k$ block of an upper triangular matrix R, where $A\hat{V} = \hat{V}R$ is a complete Schur decomposition. We refer to this as a **partial Schur decomposition** of A.

Algorithm 2: IRAM iteration
Input is an $n \times k$ ortho-normal matrix V_k, an upper Hessenberg matrix H_k, and a vector $\mathbf{f}_k$ such that $AV_k = V_k H_k + \mathbf{f}_k \mathbf{e}_k^T$.
Output is an $n \times k$ ortho-normal matrix V_k, an upper triangular matrix H_k such that $AV_k = V_k H_k$.

repeat until convergence,
  Beginning with the k-step factorization,
  apply p additional steps of the Arnoldi process
  to compute an $m = k + p$ step Arnoldi factorization
    $AV_m = V_m H_m + \mathbf{f}_m \mathbf{e}_m^*$.
  Compute $\sigma(H_m)$ and select p shifts $\mu_1, \mu_2, ...\mu_p$;
  $Q = I_m$;
  for $j = 1, 2, ..., p$,
    Factor $[Q_j, R_j] = \mathrm{q}r(H_m - \mu_j I)$;
    $H_m \leftarrow Q_j^* H_m Q_j$;
    $Q \leftarrow QQ_j$;
  end
  $\hat{\beta}_k = H_m(k+1, k)$; $\sigma_k = Q(m, k)$;
  $\mathbf{f}_k \leftarrow \mathbf{v}_{k+1}\hat{\beta}_k + \mathbf{f}_m \sigma_k$;
  $V_k \leftarrow V_m Q(:, 1:k)$; $H_k \leftarrow H_m(1:k, 1:k)$;
end

Facts: [Sor92], [Sor02]

1. Implicit restarting avoids numerical difficulties and storage problems normally associated with Arnoldi and Lanczos processes. The algorithm is capable of computing a few (k) eigenvalues with user specified features such as largest real part or largest magnitude using $2nk + \mathcal{O}(k^2)$ storage. The computed Schur basis vectors for the desired k-dimensional eigenspace are numerically orthogonal to working precision.
2. Desired eigen-information from a high-dimensional Krylov space is continually compressed into a fixed size k-dimensional subspace through an implicitly shifted QR mechanism. An Arnoldi factorization of length $m = k + p$,
$$AV_m = V_m H_m + \mathbf{f}_m \mathbf{e}_m^*,$$
is compressed to a factorization of length k that retains the eigen-information of interest. Then the factorization is expanded once more to m-steps and the compression process is repeated.

3. QR steps are used to apply p linear polynomial factors $A - \mu_j I$ implicitly to the starting vector $\mathbf{v}_1$. The first stage of this shift process results in
$$AV_m^+ = V_m^+ H_m^+ + \mathbf{f}_m \mathbf{e}_m^* Q,$$
where $V_m^+ = V_m Q$, $H_m^+ = Q^* H_m Q$, and $Q = Q_1 Q_2 \cdots Q_p$. Each Q_j is the orthogonal matrix associated with implicit application of the shift $\mu_j = \theta_{k+j}$. Since each of the matrices Q_j is Hessenberg, it turns out that the first $k-1$ entries of the vector $\mathbf{e}_m^* Q$ are zero (i.e., $\mathbf{e}_m^* Q = [\sigma \mathbf{e}_k^T, \hat{\mathbf{q}}^*]$). Hence, the leading k columns remain in an Arnoldi relation and provide an updated k-step Arnoldi factorization
$$AV_k^+ = V_k^+ H_k^+ + \mathbf{f}_k^+ \mathbf{e}_k^*,$$
with an updated residual of the form $\mathbf{f}_k^+ = V_m^+ \mathbf{e}_{k+1} \hat{\beta}_k + \mathbf{f}_m \sigma$. Using this as a starting point, it is possible to apply p additional steps of the Arnoldi process to return to the original m-step form.
4. Virtually any explicit polynomial restarting scheme can be applied with implicit restarting, but considerable success has been obtained with exact shifts. Exact shifts result in H_k^+ having the k wanted Ritz values as its spectrum. As convergence takes place, the subdiagonals of H_k tend to zero and the most desired eigenvalue approximations appear as eigenvalues of the leading $k \times k$ block of R as a partial Schur decomposition of A. The basis vectors V_k tend to numerically orthogonal Schur vectors.
5. The basic IRAM iteration is shown in Algorithm 2.

Examples:

1. The expansion and contraction process of the IRAM iteration is visualized in Figure 57.3.

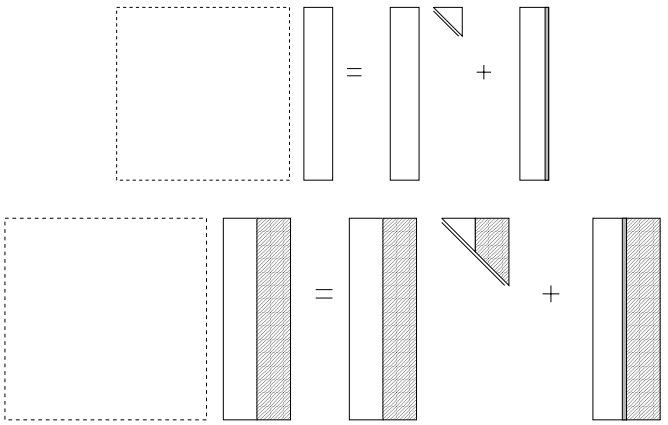

FIGURE 57.3 Visualization of IRAM.

57.6 Convergence of IRAM

IRAM converges linearly. An intuitive explanation follows. If $\mathbf{v}_1$ is expressed as a linear combination of eigenvectors $\{\mathbf{q}_j\}$ of A, then
$$\mathbf{v}_1 = \sum_{j=1}^{n} \mathbf{q}_j \gamma_j \Rightarrow \psi(A)\mathbf{v}_1 = \sum_{j=1}^{n} \mathbf{q}_j \psi(\lambda_j)\gamma_j.$$

Applying the same polynomial (i.e., using the same shifts) repeatedly for ℓ iterations will result in the j-th original expansion coefficient being attenuated by a factor

$$\left(\frac{\psi(\lambda_j)}{\psi(\lambda_1)}\right)^{\ell},$$

where the eigenvalues have been ordered according to decreasing values of $|\psi(\lambda_j)|$. The leading k eigenvalues become dominant in this expansion and the remaining eigenvalues become less and less significant as the iteration proceeds. Hence, the starting vector $\mathbf{v}_1$ is forced into an invariant subspace as desired. The adaptive choice of ψ provided with the exact shift mechanism further enhances the isolation of the wanted components in this expansion. Hence, the wanted eigenvalues are approximated ever better as the iteration proceeds. Making this heuristic argument precise has turned out to be quite difficult. Some fairly sophisticated analysis is required to understand convergence of these methods.

57.7 Convergence in Gap: Distance to a Subspace

To fully discuss convergence we need some notion of nearness of subspaces. When nonnormality is present or when eigenvalues are clustered, the distance between the computed subspace and the desired subspace is a better measure of success than distance between eigenvalues. The subspaces carry uniquely defined Ritz values with them, but these can be very sensitive to perturbations in the nonnormal setting.

Definitions:

A notion of distance that is useful in our setting is the **containment gap** between the subspaces $\mathcal{W}$ and $\mathcal{V}$:

$$\delta(\mathcal{W},\mathcal{V}) := \max_{\mathbf{w}\in\mathcal{W}} \min_{\mathbf{v}\in\mathcal{V}} \frac{\|\mathbf{w}-\mathbf{v}\|_2}{\|\mathbf{w}\|_2}.$$

Note: $\delta(\mathcal{W},\mathcal{V})$ is the sine of the largest canonical angle between $\mathcal{W}$ and the closest subspace of $\mathcal{V}$ with the same dimension as $\mathcal{W}$.

In keeping with the terminology developed in [BER04] and [BES05], $\mathcal{X}_g$ shall be the invariant subspace of A associated with the so called "good" eigenvalues (the desired eigenvalues) and $\mathcal{X}_b$ is the complementary subspace. $\mathbf{P}_g$ and $\mathbf{P}_b$ are the spectral projectors with respect to these spaces.

It is desirable to have **convergence in gap** for the Krylov method, meaning

$$\delta(\mathcal{K}_m(A,\mathbf{v}_1^{(\ell)}),\mathcal{X}_g) \to 0.$$

Fundamental quantities required to study convergence.

1. **Minimal polynomial for $\mathcal{X}_g$:**

   $$a_g := \textit{minimal polynomial of } A \textit{ with respect to } \mathbf{P}_g\mathbf{v}_1,$$

   which is the monic polynomial of least degree s.t. $a_g(A)\mathbf{P}_g\mathbf{v}_1 = \mathbf{0}$.
2. **Nonnormality constant $\kappa(\Omega)$:** The smallest positive number s.t.

   $$\|f(A)\,\Pi_{\mathcal{U}}\|_2 \;\le\; \kappa(\Omega) \max_{z\in\Omega} |f(z)|$$

   uniformly for all functions f analytic on Ω. This constant and its historical origins are discussed in detail in [BER04].
3. **ε-pseudospectrum of A:**

   $$\Lambda_\varepsilon(A) := \{z \in \mathbf{C} : \|(zI-A)^{-1}\|_2 \ge \varepsilon^{-1}\}.$$

Facts: [BER04], [BES05]

1. Two fundamental convergence questions:
   (a) What is the gap $\delta(\mathcal{U}_g, \mathcal{K}_k(A, \mathbf{v}_1))$ as k increases?
   (b) How does $\delta(\mathcal{U}_g, \mathcal{K}_m(A, \widehat{\mathbf{v}}_1))$ depend on $\widehat{\mathbf{v}}_1 = \Phi(A)\mathbf{v}_1$, and how can we optimize the asymptotic behavior?

   Key ingredients to convergence behavior are the nonnormality of A and the distribution of $\mathbf{v}_1$ w. r. t. $\mathcal{U}_g$. The goal of restarting is to attain the unrestarted iteration performance, but within *restricted subspace dimensions.*
2. *Convergence with no restarts:* In [BES05], it is shown that

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(A, \mathbf{v}_1)) \leq C_o C_b \min_{p \in \mathcal{P}_{\ell-2m}} \max_{z \in \Omega_b} \big|1 - a_g(z)p(z)\big|,$$

   where the compact set $\Omega_g \subseteq \mathbf{C} \setminus \Omega_b$ contains all the good eigenvalues.

$$C_o := \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|\psi(A)\mathbf{P}_b\mathbf{v}_1\|_2}{\|\psi(A)\mathbf{P}_g\mathbf{v}_1\|_2}, \quad C_b := \kappa(\Omega_b).$$

3. Rate of convergence estimates are obtained from complex approximation theory. Construct conformal map $\mathcal{G}$ taking the exterior of Ω_b to the exterior of the unit disk with $\mathcal{G}(\infty) = \infty$ and $\mathcal{G}'(\infty) > 0$. Define $\rho := \left(\min_{j=1,\ldots,L} |\mathcal{G}(\lambda_j)|\right)^{-1}$. Then (Gaier, Walsh)

$$\limsup_{k \to \infty} \min_{p \in \mathcal{P}_k} \max_{z \in \Omega_b} \left|\frac{1}{a_g(z)} - p(z)\right|^{1/k} = \rho.$$

   The image of $\{|z| = \rho^{-1}\}$ is a curve $\mathcal{C} := \mathcal{G}^{-1}(\{|z| = \rho^{-1}\})$ around Ω_b. This critical curve passes through a good eigenvalue "closest to" Ω_b. The curve contains at least one good eigenvalue, with all bad and no good eigenvalues in its interior.
4. Convergence with the exact shift strategy has not yet been fully analyzed. However, convergence rates have been established for restarts with asymptotically optimal points. These are the Fejér, Fekete, or Leja points for Ω_b. In [BES05], computational experiments are shown that indicate that exact shifts behave very much like optimal points for certain regions bounded by pseudo-spectral level curves or lemniscates.
5. Let Ψ_M interpolate $1/a_g(z)$ at the M restart shifts:

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(A, \widehat{\mathbf{v}}_1)) \leq C_o C_g \max_{z \in \Omega_b} |1 - \Psi_M(z)a_g(z)| \leq C_o\, C_g\, C_r\, r^M$$

   for any $r > \rho$ (see [Gai87], [FR89]). Here, $\widehat{\mathbf{v}}_1 = \Phi(A)\mathbf{v}_1$, where Φ is the aggregate restart polynomial (its roots are all the implicit restart shifts that have been applied). The subspace dimension is $\ell = 2m$, the restart degree is m, and the aggregate degree is $M = \nu m$.

57.8 The Generalized Eigenproblem

In many applications, the generalized eigenproblem $A\mathbf{x} = M\mathbf{x}\lambda$ arises naturally. A typical setting is a finite element discretization of a continuous problem where the matrix M arises from inner products of basis functions. In this case, M is symmetric and positive (semi) definite, and for some algorithms this property is a necessary condition. Generally, algorithms are based upon transforming the generalized problem to a standard problem.

57.9 Krylov Methods with Spectral Transformations

Definitions:

A very successful scheme for converting the generalized problem to a standard problem that is amenable to a Krylov or a subspace iteration method is to use the **spectral transformation** suggested by Ericsson and Ruhe [ER80],

$$(A - \sigma M)^{-1} M x = \mathbf{x}\nu.$$

Facts: [Sor92], [Sor02]

1. An eigenvector $\mathbf{x}$ of the spectral transformation is also an eigenvector of the original problem $A\mathbf{x} = M\mathbf{x}\lambda$, with the corresponding eigenvalue given by $\lambda = \sigma + \frac{1}{\nu}$.
2. There is generally rapid convergence to eigenvalues near the shift σ because they are transformed to extremal well-separated eigenvalues. Perhaps an even more influential aspect of this transformation is that eigenvalues far from σ are damped (mapped near zero).
3. One strategy is to choose σ to be a point in the complex plane that is near eigenvalues of interest and then compute the eigenvalues ν of largest magnitude of the spectral transformation matrix. It is not necessary to have σ extremely close to an eigenvalue. This transformation, together with the implicit restarting technique, is usually adequate for computing a significant number of eigenvalues near σ.
4. Even when $M = I$, one generally must use the shift-invert spectral transformation to find interior eigenvalues. The extreme eigenvalues of the transformed operator A_σ are generally large and well separated from the rest of the spectrum. The eigenvalues ν of largest magnitude will transform back to eigenvalues λ of the original A that are in a disk about the point σ. This is illustrated in Figure 57.4, where the + symbols are the eigenvalues of A and the circled ones are the computed eigenvalues in the disk (dashed circle) centered at the point σ.

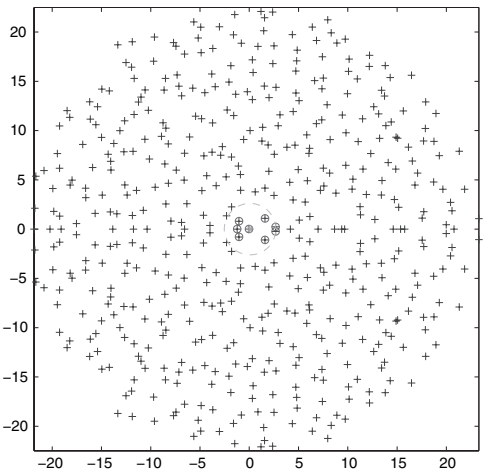

FIGURE 57.4 Eigenvalues from shift-invert.

5. With shift-invert, the Arnoldi process is applied to the matrix $A_\sigma := (A - \sigma M)^{-1} M$. Whenever a matrix-vector product $\mathbf{w} \leftarrow A_\sigma \mathbf{v}$ is required, the following steps are performed:
   (a) $\mathbf{z} = M\mathbf{v}$,

(b) Solve $(A - \sigma M)\mathbf{w} = \mathbf{z}$ for $\mathbf{w}$.

The matrix $A - \sigma M$ is factored initially with a sparse direct **LU**-decomposition or in a symmetric indefinite factorization and this single factorization is used repeatedly to apply the matrix operator A_σ as required.

6. The scheme is modified to preserve symmetry when A and M are both symmetric and M is positive (semi)definite. One can utilize a weighted M (semi)inner product in the Lanczos/Arnoldi process [ER80], [GLS94], [MS97]. This amounts to replacing the computation of $\mathbf{h} \leftarrow V_{j+1}^* \mathbf{w}$ and $\beta_j = \|\mathbf{f}_j\|_2$ with $\mathbf{h} \leftarrow V_{j+1}^* M\mathbf{w}$ and $\beta_j = \sqrt{\mathbf{f}_j^* M \mathbf{f}_j}$, respectively, in the Arnoldi process described in Algorithm 1.
7. The matrix operator A_σ is self-adjoint with respect to this (semi)inner product, i.e., $\langle A_\sigma \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A_\sigma \mathbf{y}\rangle$ for all vectors $\mathbf{x}, \mathbf{y}$, where $\langle \mathbf{w}, \mathbf{v}\rangle := \sqrt{\mathbf{w}^* M \mathbf{v}}$. This implies that the projected Hessenberg matrix H is actually symmetric and tridiagonal and the standard three-term Lanczos recurrence is recovered with this inner product.
8. There is a subtle aspect to this approach when M is singular. The most pathological case, when $\text{null}(A) \cap \text{null}(M) \neq \{0\}$, is not treated here. However, when M is singular there may be infinite eigenvalues of the pair (A, M) and the presence of these can introduce large perturbations to the computed Ritz values and vectors. To avoid these difficulties, a purging operation has been suggested by Ericsson and Ruhe [ER80]. If $\mathbf{x} = V\mathbf{y}$ with $H\mathbf{y} = \mathbf{y}\theta$, then

$$A_\sigma \mathbf{x} = VH\mathbf{y} + \mathbf{f}\mathbf{e}_k^T \mathbf{y} = \mathbf{x}\theta + \mathbf{f}\mathbf{e}_k^T \mathbf{y}.$$

Replacing the $\mathbf{x}$ with the improved eigenvector approximation $\mathbf{x} \leftarrow (\mathbf{x} + \frac{1}{\theta}\mathbf{f}\mathbf{e}_k^T \mathbf{y})$ and renormalizing has the effect of purging undesirable components without requiring any additional matrix vector products with A_σ.
9. The residual error of the purged vector $\mathbf{x}$ with respect to the original problem is

$$\|A\mathbf{x} - M\mathbf{x}\lambda\|_2 = \|M\mathbf{f}\|_2 \frac{|\mathbf{e}_k^T \mathbf{y}|}{|\theta|^2},$$

where $\lambda = \sigma + 1/\theta$. Since $|\theta|$ is usually quite large under the spectral transformation, this new residual is generally considerably smaller than the original.

References

[BCR96] J. Baglama, D. Calvetti, and L. Reichel. Iterative methods for the computation of a few eigenvalues of a large symmetric matrix. *BIT*, 36(3): 400–440, 1996.

[BER04] C.A. Beattie, M. Embree, and J. Rossi. Convergence of restarted Krylov subspaces to invariant subspaces. *SIAM J. Matrix Anal. Appl.*, 25: 1074–1109, 2004.

[BES05] C.A. Beattie, M. Embree, and D.C. Sorensen. Convergence of polynomial restart Krylov methods for eigenvalue computation. *SIAM Rev.*, 47(3): 492–515, 2005.

[DGK76] J. Daniel, W.B. Gragg, L. Kaufman, and G.W. Stewart. Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization. *Math. Comp.*, 30: 772–795, 1976.

[DDH88] J.J. Dongarra, J. DuCroz, S. Hammarling, and R. Hanson. An extended set of Fortran basic linear algebra subprograms. *ACM Trans. Math. Softw.*, 14: 1–17, 1988.

[ER80] T. Ericsson and A. Ruhe. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Math. Comp.*, 35(152): 1251–1268, 1980.

[FR89] B. Fischer and L. Reichel. Newton interpolation in Fejér and Chebyshev points. *Math. Comp.*, 53: 265–278, 1989.

[Gai87] D. Gaier. *Lectures on Complex Approximation.* Birkhäuser, Boston, 1987.

[GLS94] R.G. Grimes, J.G. Lewis, and H.D. Simon. A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems. *SIAM J. Matrix Anal. Appl.*, 15(1): 228–272, 1994.

[LSY98] R. Lehoucq, D.C. Sorensen, and C. Yang. *ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM Publications, Philadelphia, 1998. (Software available at: http://www.caam.rice.edu/software/ARPACK.)

[MS97] K. Meerbergen and A. Spence. Implicitly restarted Arnoldi with purification for the shift–invert transformation. *Math. Comp.*, 218: 667–689, 1997.

[Mor96] R.B. Morgan. On restarting the Arnoldi method for large nonsymmetric eigenvalue problems. *Math. Comp.*, 65: 1213–1230, 1996.

[Par80] B.N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice-Hall, Upper Saddle River, NJ, 1980.

[Saa92] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, Manchester, UK, 1992.

[Sor92] D.C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM J. Matrix Anal. Appl.*, 13: 357–385, 1992.

[Sor02] D.C. Sorensen. Numerical methods for large eigenvalue problems. *Acta Numerica*, 11: 519–584, 2002.