The Numerical Solution of the Large Scale Eigenvalue Problem with Orthogonal Projection Methods.

R. B. Lehoucq

Mathematics and Computer Science Division,

Argonne National Laboratory

lehoucq@mcs.anl.gov

http://www.mcs.anl.gov/home/lehoucq/index.html

Problem

Let $\mathbf{A}$ be a matrix of order n. Find $k \ll n$ solutions to the matrix eigenvalue problem

$$\mathbf{A}\mathbf{x} = \mathbf{x}\lambda,$$

Interested in methods that only require matrix vector products with $\mathbf{A}$.

No **dense similarity transformations** as in EISPACK and LAPACK for they **destroy the sparsity** of $\mathbf{A}$ and **require $\mathcal{O}(n^2)$ storage.** These are implementations of the QR algorithm.

Definition: Large is something for which the full QR algorithm is prohibitive, both in storage and/or efficiency.

Orthogonal Projection Methods

Two spaces $\mathcal{V}$ and $\mathcal{W}$

- Choose a "Ritz" pair $(\theta, \mathbf{x})$ such that

$$\mathbf{x} \in \mathcal{V}$$

$$\mathbf{A}\mathbf{x} - \theta\mathbf{x} \perp \mathcal{W} \qquad \text{(Petrov–Galerkin Condition)}$$

- $\mathcal{V}$ and $\mathcal{W}$ are expanded in some fashion.
- Our interest is $\mathcal{V} = \mathcal{W}$ or an Orthogonal Projection method.

Orthogonal projection methods use unitary transformations and thus we can design backward stable algorithms implementing them.

Bai, Day & Ye have a non-Hermitian block Lanczos method for $\mathcal{V} \neq \mathcal{W}$.

Outline

- The QR algorithm
- Review of some Large Scale methods
- A Partial or Truncated QR Algorithm.
- More challenging problems.

Implicitly restarting the Arnoldi method is nothing more than a partial or truncated QR algorithm

QR Algorithm

Let $\mathbf{A}^{+} \leftarrow \mathbf{A}$. and $\mathbf{W} = \mathbf{I}$.

- Let $\mathbf{QR} \leftarrow \mathbf{A}^{+} - \tau_i \mathbf{I}$ be a QR factorization.
- $\mathbf{A}^{+} \leftarrow \mathbf{RQ} + \tau_i \mathbf{I}$ and $\mathbf{W} \leftarrow \mathbf{WQ}$.

Let $\mathcal{P}(\lambda) \equiv (\lambda - \tau_1) \cdots (\lambda - \tau_N)$. If we take the QR factorization $\mathcal{P}(\mathbf{A}) = \mathbf{QR}$, then

$$\mathbf{W} = \mathbf{Q} \qquad \text{and} \qquad \mathbf{Q}^T \mathbf{A} \mathbf{Q} = \mathbf{A}^{+}.$$

The QR algorithm is simultaneous iteration in disguise. The polynomial $\mathcal{P}(\mathbf{A})$ is **neither formed nor applied.** The unitary matrix $\mathbf{W}$ is the key quantity. (Kahan & Parlett (1968), Stewart (1973) and Watkins (1982))

Schur Decomposition

$$\mathbf{AZ} = \mathbf{Z} \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1,n} \\ 0 & T_{22} & \cdots & T_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{n,n} \end{bmatrix}$$

- $\mathbf{Z}^H\mathbf{Z} = \mathbf{I}$, The columns of $\mathbf{Z}$ are the Schur vectors.
- Each $T_{i,i}$ is an eigenvalue.
- If $\mathbf{A}$ is Hermitian, then $\mathbf{T}$ is diagonal and $\mathbf{Z}$ are the eigenvectors.
- The Schur decomposition is always defined. It's unique given an ordering along the diagonal and up to multiplicities.
- Can design backward stable algorithms to compute it.

QR Algorithm, continued

- The shifts τ_i are chosen to get a quadratic rate of convergence.(Cubic rate for Hermitian $\mathbf{A}$.)
- $\mathbf{A}$ is usually initially transformed to upper Hessenberg form. One iteration of QR algorithm is now $\mathcal{O}(n^2)$ work—down from $\mathcal{O}(n^3)$ on the untransformed $\mathbf{A}$.
- Can re-order $\mathbf{T}$ so that the k eigenvalues of interest are located in $\mathbf{T}_k$ (the leading principle matrix of order k).

Since we're only interested in k eigenvalues, why not just only compute $\mathbf{T}_k$, the leading principle matrix of order k of $\mathbf{T}_n$? And thus remove the $\mathcal{O}(n^2)$ storage requirements.

How about a partial or truncated QR algorithm ?

Review of some Large Scale Methods

Power Method with a vector $\mathbf{u}$

$$\mathbf{u},\ \mathbf{Au},\ \cdots\ ,\ \mathbf{A}^{m-1}\mathbf{u}$$

Sequence converges to $\mathbf{A}$'s eigenvector associated with eigenvalue of largest magnitude.

Simultaneous iteration with a martrix $\mathbf{U}$ of b vectors

$$\mathbf{U},\ \mathbf{AU},\ \cdots\ ,\ \mathbf{A}^{m-1}\mathbf{U}$$

Sequence converges to $\mathbf{A}$'s eigenspace associated with the b eigenvalues of largest magnitude.

Block Arnoldi Reduction of Length m

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_m \mathbf{H}_m + \mathbf{F}_m \mathbf{E}_m^T$$

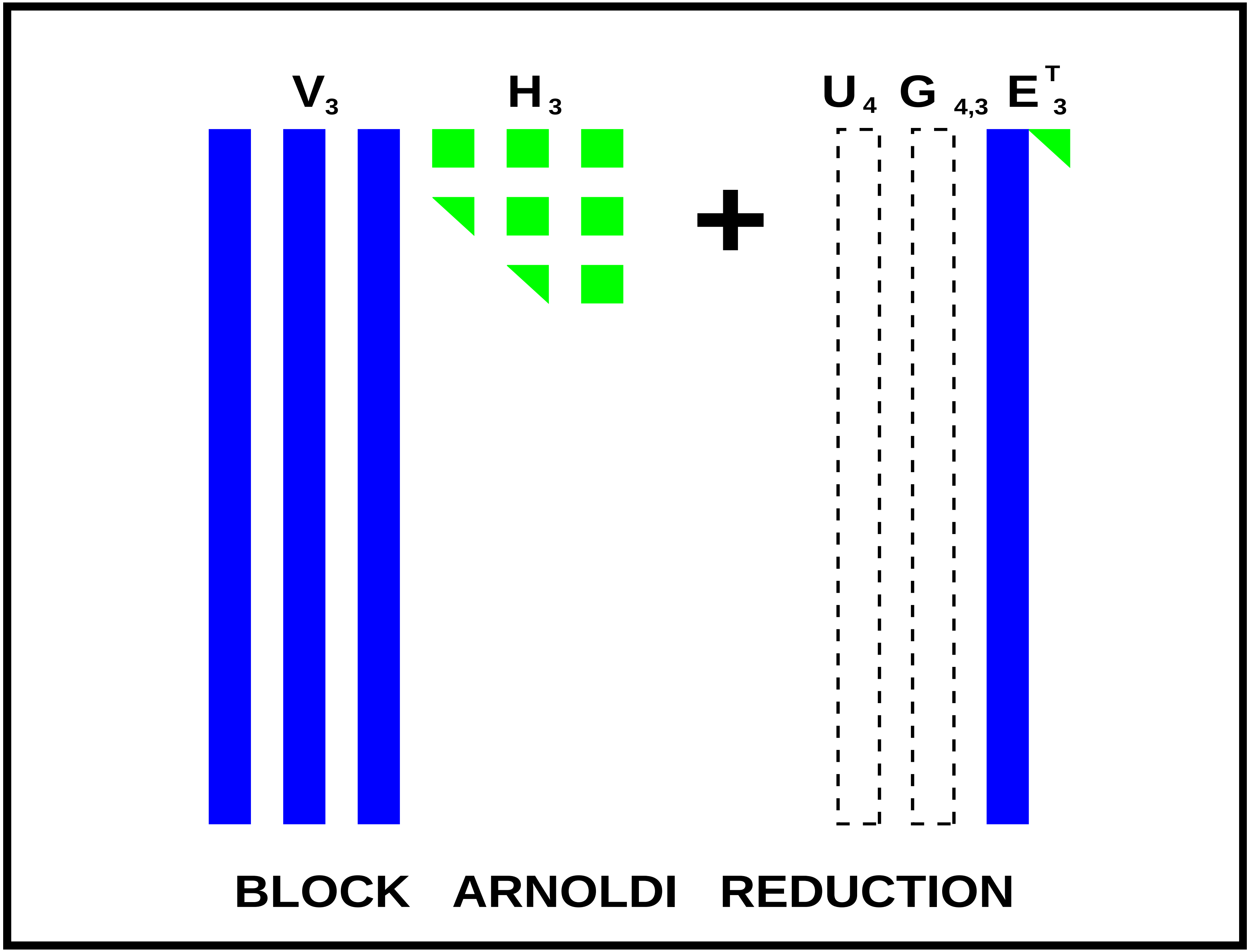
- $\mathbf{V}_m^H \mathbf{A} \mathbf{V}_m = \mathbf{H}_m$ is a band Hessenberg matrix of order $b \cdot m$.
- $\mathbf{V}_m^H \mathbf{V}_m = \mathbf{I}_{bm}$, and $\mathbf{V}_m^H \mathbf{F}_m = \mathbf{0}$.
- $\mathcal{K}_m(\mathbf{A}, \mathbf{U}_1) \equiv \{\mathbf{U}_1, \mathbf{A}\mathbf{U}_1, \cdots, \mathbf{A}^{m-1}\mathbf{U}_1\} = \mathcal{R}(\mathbf{V}_m)$.
- $\mathbf{H}_m$ is band tridiagonal when $\mathbf{A}$ is Hermitian.

Block Arnoldi Reduction of Length m, cont.

$$\mathbf{A}\,\mathbf{V}_m = \begin{bmatrix} \mathbf{U}_1 & \cdots & \mathbf{U}_m \end{bmatrix} \begin{bmatrix} \mathbf{G}_{1,1} & \cdots & \cdots & \mathbf{G}_{1,m} \\ \mathbf{G}_{2,1} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{G}_{m,m-1} & \mathbf{G}_{m,m} \end{bmatrix}$$

$$+\mathbf{U}_{m+1}\mathbf{G}_{m+1,m}\mathbf{E}_m^T$$

- $\mathbf{U}_{m+1}\mathbf{G}_{m+1,m} = \mathbf{A}\mathbf{U}_m - \sum_{i=1}^{m} \mathbf{U}_i\mathbf{G}_{i,m}$ where $\mathbf{G}_{i,m} = \mathbf{U}_i^H\mathbf{A}\mathbf{U}_m$.
- The columns of $\mathbf{U}_{m+1}$ are an unitary basis for the orthonormal projection of $\mathbf{A}^m\mathbf{U}_1$ onto $\mathbf{V}_m$.



V_3
H_3
+
U_4 $G_{4,3}$ E_3^T
BLOCK ARNOLDI REDUCTION

Computing Eigenvalues

Partition the eigenvalues of $\mathbf{H}_m$ so that

$$\{\underbrace{\theta_1, \ldots, \theta_k}_{\text{wanted}}\} \cup \{\underbrace{\theta_{k+1}, \ldots, \theta_{bm}}_{\text{unwanted}}\}$$

Let $1 \leq i \leq bm$ & $\mathbf{H}_m \mathbf{s}_i = \theta_i \mathbf{s}_i$ so that

$$\|\mathbf{A}\mathbf{V}_m \mathbf{s}_i - \mathbf{V}_m \mathbf{s}_i \theta_i\| = \|\mathbf{F}_m \mathbf{E}_m^T \mathbf{s}_i\|$$

should be small for the k desired eigenvalues.

The extremal, well–separated eigenvalues of $\mathbf{A}$ are well approximated by the θ_i for m large. (Kaniel & Paige, Saad).

A truncated QR algorithm (restarting) pushes the interesting directions to the leading portion of the reduction. **Compute a partial Schur decomposition.**

Restarting an Arnoldi Reduction

- Define an iterative procedure by building out the compressed block reduction to a length m one. Restarting retains the good information. **Fixed Storage Requirements.**
- Karush (1951) introduced restarting for the Lanczos method, Saad (1980) for the Arnoldi method.
- Sorensen (1992) introduced implicit restarting (for a blocksize of one) as a more robust and efficient way to restart the Arnoldi/Lanczos process. (Implementation in ARPACK).

When $m = 1$, we have the power method for $b = 1$ and subspace iteration otherwise. The Schur (Rayleigh-Ritz) projection step is a restarting procedure.

Truncated QR Algorithm

Let $\psi_p(\lambda) = \prod_{i=1}^{p}(\lambda - \tau_i)$ where $1 \le p < m$ and let $r = m - p$.

Let $\psi_p(\mathbf{H}) \begin{bmatrix} \mathbf{E}_1 & \cdots & \mathbf{E}_r \end{bmatrix} = \mathbf{QR}$.

Suppose we run p steps of the implicitly shifted QR algorithm on $\mathbf{H}$ with $\tau_1, \ldots, \tau_p$ resulting in $\mathbf{H}_m\mathbf{W} = \mathbf{W}\mathbf{H}_m^+$.

Recall that $\mathbf{W} \begin{bmatrix} \mathbf{E}_1 & \cdots & \mathbf{E}_r \end{bmatrix} = \mathbf{Q}$ for $r = 1, 2, \ldots, \text{ceiling}(n/b)$.

Theorem Let $\mathbf{V}_r^+ = \mathbf{V}_m\mathbf{Q}$. Given a length m block Arnoldi reduction

$$\psi_p(\mathbf{A})\mathbf{V}_r = \mathbf{V}_r^+\mathbf{R}$$

and

$$\mathbf{A}\mathbf{V}_r^+ = \mathbf{V}_r^+\mathbf{H}_r^+ + \mathbf{F}_r^+\mathbf{E}_r^+.$$

Truncated QR Algorithm, cont.

- Connects subspace iteration with the a partial or truncated QR algorithm.
- No need for an explicit restart from scratch or applying a polynomial filter in $\mathbf{A}$. **Unnecessary matrix vector products are avoided.**
- A truncated QR algorithm is backward stable. Computes the exact partial Schur decomposition for a nearby matrix.
- Can develop high-quality software because of the connection with the QR algorithm (ARPACK is an example).

Computing a Partial Schur Decomposition

Incrementally build a partial Schur form

$$\|\mathbf{A}\mathbf{Z}_j - \mathbf{Z}_j\mathbf{T}_j\| = \mathcal{O}(\|\mathbf{F}_m\|\|\mathbf{E}_m^T\mathbf{s}_i\|)$$

for $i = 1, \ldots, j$ (the converged Ritz pairs).

Based on the Locking technique in L. & Sorensen (1996) (extended to block Arnoldi methods).

$\mathbf{T}_j$ is upper triangular of order j containing the j eigenvalues of interest and the columns of $\mathbf{Z}_j$ are an approximate Schur basis.

Deflation within Block Krylov Subspaces

Suppose $\mathbf{A}\mathbf{Z}_j = \mathbf{Z}_j\mathbf{T}_j$, ($\mathbf{Z}_j$ not blocked) then,

$$\mathbf{A}\begin{bmatrix} \mathbf{Z}_j & \mathbf{V}_m \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_j & \mathbf{V}_m \end{bmatrix}\begin{bmatrix} \mathbf{T}_j & \mathbf{M}_j \\ \mathbf{0} & \mathbf{H}_m \end{bmatrix} + \mathbf{F}_m\mathbf{E}_{m+\tilde{j}}^T.$$

Clean form of selective orthogonalization (Parlett & Scott).

Always work with a Schur basis.

Truncated QR Algorithm, cont.

The differences with QR algorithm are that

- Can't use the same shifting strategy unless $mb \geq n$. So the best rate of convergence is superlinear.
- We use a block Arnoldi reduction to partially transform $\mathbf{A}$ into band Hessenberg form—only need matrix vector products.

Increasing b (blocksize) does a better job of getting clusters and/or multiplicities. **Increase the convergence tolerance.** Can also be more efficient since $\mathbf{A}$ is applied to a group of vectors.
Amortize the cost of applying A.

Convergence Theory

Let $1 \le r \le m$.

$$\mathbf{A}\begin{bmatrix} \mathbf{V}_r & \mathbf{W} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_r & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{H}_r & \mathbf{M}_r \\ \mathbf{G}_{r+1,r}\mathbf{E}_r^T & \mathbf{C}_r \end{bmatrix}$$

At what rate does $\|\mathbf{G}_{r+1,r}\|$ go to zero given a shift strategy ?

$\|\mathbf{G}_{r+1,r}\| \to 0$ with asymptotic rate $\frac{\|\mathbf{M}_r\|}{\mathrm{sep}(\mathbf{H}_r,\mathbf{C}_r)}\left|\frac{\lambda_{r+1}}{\lambda_r}\right|$
See Watkins & Elsner 1991.

(sep$(\mathbf{H}_r, \mathbf{C}_r)$ is the smallest singular value of the linear transformation $\mathbf{H}_r\mathbf{X} - \mathbf{X}\mathbf{C}_r$ and can be much smaller then the minumum difference between the eigenvalues.)

Best strategy is an adaptive one that varies, r, b, m to compute the k eigenvalues.

Practical Implementation

- An analysis of an implicitly restarting Arnoldi iteration (L. Thesis 1995). **How to go (is it possible ?) to production quality software**.
- Deflation strategies. (L. & Sorensen SIMAX 1996)
- See Preprint MCS-P648-0297, Argonne National Laboratory, *Truncated QR Algorithms and the Solution of Large-Scale Eigenvalue Problems* for more details.
- Block Formulation *Implementation of an Implicitly Restarted Block Arnoldi Method*, (with K. J. Maschhoff) Preprint MCS-P649-0297, Argonne National Laboratory.

Some After thoughts

Implictly restarted block Arnoldi methods will do better than subspace iteration because they themselves are subspace iteration in disguise.

The caveat is that they are much trickier to implement.

For many problems, the cost of maintaining orthogonality of the Arnoldi basis is less than 10 % of the total computation cost.

More Challenging Eigenvalue Problems

Previous methods only require the **application** of a matrix vector product. Sufficient for the well separated extremal eigenvalues.

- **What if we're interested in the interior eigenvalues ?**
- **What about a generalized eigenvalue problem $\mathbf{Ax} = \mathbf{Bx}\lambda$?**

Use a rational spectral transformation $\mathcal{F}(\mathbf{A}, \mathbf{B})$ such as shift-invert $(\mathbf{A} - \sigma\mathbf{B})^{-1}\mathbf{B}$ or generalized Cayley $(\mathbf{A} - \sigma\mathbf{B})^{-1}(\mathbf{A} - \mu\mathbf{B})$.

More Challenging Eigenvalue Problems, cont.

- Run block Arnoldi on $\mathcal{F}(\mathbf{A}, \mathbf{B})$ and solve the resulting linear system of equations using a direct method or high accuracy linear solve.
- Expand $\mathcal{V}$ by suitably projecting an approximate solution to

$$\mathbf{Mv} = \mathbf{Ax} - \mathbf{Bx}\theta$$

  where $\mathbf{M} \approx (\mathbf{A} - \sigma\mathbf{B})^{-1}$.

Examples include Davidson's method, Jacobi–Davidson (Sleijpen & Van der Vorst), Inexact Cayley (Meerbergen & Roose) Inexact rational Krylov (Lehoucq & Meerbergen), Rational Krylov (Ruhe), TRQ (Sorensen & Yang), methods by Knyazev.