Notes for 2016-11-28

Logistics

1. Today is the last lecture of new material – review on Wednesday.
2. Course evaluations posted today — remember to complete them!
3. Final exam will be posted Wednesday, 11/30; due 12/14.
4. Office hours this week: Tuesday 10-11, Wednesday 1:30-2:30. I will be away Thursday-Friday.

1 Lanczos and Arnoldi eigensolvers

The standard ingredients in all the subspace methods we have described so far are a choice of an approximation subspace (usually a Krylov subspace) and a method for choosing an approximation from the space. In the most common methods for large-scale eigensolvers, one uses a Krylov subspace together with a Bubnov-Galerkin condition for choosing approximate eigenpairs; that is, we choose $v \in \mathcal{V}$ such that

$$r = (A - \hat{\lambda} I)\hat{v} \perp \mathcal{V}.$$

In the symmetric case, this is equivalent to finding a constrained stationary point of the Rayleigh quotient (i.e. a $\hat{v}$ such that directional derivatives of $\rho_A(\hat{v})$ are zero for any direction in the space). This approximation scheme is known as the *Rayleigh-Ritz* method, and approximate eigenvectors and eigenvalues obtained in this way are often called *Ritz vectors* and *Ritz values*.

In the Lanczos method for the symmetric eigenvalue problem, we compute the Lanczos decomposition

$$AQ_m = Q_m T_m + \beta_m q_{m+1} e_m^T,$$

and use it to compute the residual relation for an approximate pair $(\mu, Q_m y)$ by

$$r = (A - \mu I)Q_m y = Q_m(T_m - \mu I)y + \beta_m q_{m+1} e_m^T y.$$

The condition $Q_m^T r = 0$ gives us the *projected* problem

$$(T_m - \mu I)y = 0;$$

if we satisfy this condition, we have

$$r = \beta_m q_{m+1} y_m$$

and the residual norm (in the 2-norm) is $|\beta_m y_m|$. Generalizing, if we compute the eigendecomposition

$$T_m = Y \Theta Y^T,$$

we have the collected approximations $Z = Q_m Y_m$ with residuals

$$\|Az_k - z_k\theta_k\|_2 = |\beta_m||e_m^T y_k|.$$

This is useful because, as we discussed before, in the symmetric case a small residual error implies a small distance to the closest eigenvalue. This is also useful because the residual error can be computed with no further matrix operations — we need only to look at quantities that we would already compute in the process of obtaining the tridiangonal coefficients and the corresponding Ritz values.

The Arnoldi method for computing approximate eigenpairs similarly uses the Galerkin condition together with the Arnoldi decomposition to express an approximate partial Schur form. From the decomposition

$$AQ_m = Q_m H_m + h_{m+1,m} q_{m+1} e_m^T,$$

we write a subspace residual for $(Q_m Y, T)$ as

$$R = AQ_m Y - Q_m YT = Q_m(H_m Y - YT) + h_{m+1,m} q_{m+1} e_m^T.$$

Forcing $Q_m^T R = 0$ gives the projected problem

$$H_m Y = YT,$$

i.e. we seek a Schur decomposition of the (already Hessenberg) matrix H_m.

There are three main issues with the Lanczos and Arnoldi methods that we need to address in practical situations.

1. We must deal with forward instability, particularly in the case of the Lanczos method. Unless we are careful to maintain orthogonality between the computed Lanczos basis vectors, the method derails. The result is not that we get bad approximate eigenpairs; indeed, the forward instability is intimately tied to the very thing we want, which is convergence of eigenpairs. The real problem is that we get the same eigenpairs over and over again, a phenomenon known as "ghost" eigenvalue approximations. We deal with this issue by careful re-orthogonalization (selective or complete).

2. Because of the cost of storing a Krylov basis and maintaining its orthogonality, we typically only want to approximate a few eigenpairs at a time.

3. The Krylov subspace generated by A and some random start vector contains iterates of the power method applied to any A (or to $A - \sigma I$ for any shift σ — the Krylov subspace is shift-invariant). This is at least as good as power iteration for approximating the extremal parts of the spectrum, and we can use the same Chebyshev-based games we discussed before to give concrete (though typically pessimistic) convergence bounds. But if eigenvalues cluster, or if we are interested in eigenvalues that are not at the edge of the spectrum, then the convergence in theory and in practice can be painfully slow.

We address these issues with two basic techniques, both of which we have already seen in other contexts: *spectral transformation* and *restarting*.

2 Spectral transformation

We have dealt with the notion of spectral transformation before, when we discussed the power iteration. The idea of spectral transformation is to work not with A, but with some rational $f(A)$ where f maps the eigenvalues of interest to the outside of the spectrum. Usually f is a rational function; common examples include

- *Shift-invert*: $f(z) = (z - \sigma)^{-1}$. Favors eigenvalues close to the shift σ.
- *Cayley*: $f(z) = (\sigma - z)(\sigma + z)^{-1}$. This maps the left half plane to the interior of the unit circle and the right half plane to the exterior; it is commonly used in stability analysis.

- *Polynomial*: Just what it sounds like.

In general, the shifted linear solves needed to carry out rational spectral transformations (e.g. shift-invert and Cayley) must be computed to rather high accuracy. Hence, we favor sparse direct methods. An alternate approach, similar to what we say when we briefly considered flexible GMRES, is to break out of the confines of using a Krylov subspace; the most popular variant here is the *Jacobi-Davidson* method.

3 Restarting

When discussing GMRES, we said that to keep storage under control, we typically *restart* after a few steps. The same technique works for solving eigenvalue problems, but requires more care. In particular, when solving a linear system, it made sense to restart with a whole new Krylov subspace. For eigenvalue problems, *implicit* restarting is the norm.

The earliest versions of implicit restarting followed the strategy:

1. Build an initial Arnoldi decomposition
$$AQ_m = Q_m H_m + \beta_m q_{m+1} e_m^T.$$
2. Do several steps of shifted QR iteration on the projected matrix H_m to get a new decomposition
$$A\tilde{Q}_m = \tilde{Q}_m \tilde{H}_m + \tilde{\beta_m} \tilde{q}_{m+1} e_m^T.$$
3. "Cut back" to a basis consisting of the first p vectors of $\tilde{Q}_m$.

The "filter and cut back" approach of the implicitly restarted Arnoldi method involves some technical difficulties, but there is good software available (the ARPACK code of Lehoucq and Sorensen). This is the basis for the `eigs` code in MATLAB.

A simpler method was introduced in 2002 by Pete Stewart, the *Krylov-Schur* method. The Krylov-Schur approach rests on a more general decomposition than Arnoldi, a so-called *Krylov decomposition*

$$AU_k = U_k B_k + u_{k+1} b_{k+1}^*;$$

if the columns of U_k are orthonormal, we call this an orthonormal decomposition. The idea of the Krylov-Schur method is to compute a Schur decomposition of B_k, then sort the Schur decomposition to move the unwanted Ritz values to the "end," where they can be purged by truncating the decomposition. This approach avoids some of the technical issues in previous implicit restarting methods that maintained an Arnoldi decomposition throughout.

3.1 Jacobi-Davidson

The *Jacobi-Davidson* iteration is an alternative subspace-based large-scale eigenvalue solver that does not use Krylov subspaces. Instead, one builds a subspace via steps of an inexact Newton iteration on the eigenvalue equation. Given an approximate eigenpair (θ, u) where θ is the Rayleigh quotient, we seek a correction $s \perp u$ so that

$$A(u+s) = \lambda(u+s).$$

Rewriting this in terms of $r = (A - \theta I)u$, we have for any approximate $\tilde{\lambda}$ to the desired eigenvalue

$$(A - \tilde{\lambda} I)s = -r + (\lambda - \theta)u + (\lambda - \tilde{\lambda})s.$$

Using the desiderata that $u^* s = 0$ and the fact that $(I - uu^*)u = 0$, we obtain the correction equation

$$(I - uu^*)(A - \tilde{\lambda} I)(I - uu^*)s = -r, \quad \text{where } s \perp u.$$

The method proceeds by at each step solving the correction equation approximately and extending the subspace by a new direction s. One then seeks an approximate eigenpair from within the subspace.

In addition to a proper choice of subspaces, one needs a method to extract approximate eigenvectors and eigenvalues. This is particularly important for approximating interior eigenvalues, as the standard Rayleigh-Ritz approach may give poor results there. One possible method is to use the *refined Ritz* vector, which is obtained by minimizing the residual over all candidate eigenvectors associated with an approximate eigenvalue $\tilde{\lambda}$. The refined Ritz vector may then be plugged into the Rayleigh quotient to obtain a new eigenvector. Another method is the *harmonic Rayleigh-Ritz* approach, which for eigenvalues near a target τ employs the condition

$$(A - \tau I)^{-1}\tilde{u} - (\tilde{\theta} - \tau)^{-1}\tilde{u} \perp \mathcal{V}.$$

We again usually use the Rayleigh quotient from the harmonic Ritz vector rather than the harmonic Ritz value $\tilde{\theta}$.