Lecture 36. The Lanczos Iteration

In the last three lectures we considered Krylov subspace iterations for nonhermitian matrix problems. We shall return to nonhermitian problems in Lecture 39, for there is more to this subject than Arnoldi and GMRES. But first, in this and the following two lectures, we specialize to the hermitian case, where a major simplification takes place.

Three-Term Recurrence

The Lanczos iteration is the Arnoldi iteration specialized to the case where A is hermitian. For simplicity of notation, we shall go a step further and assume, here and in the next two lectures, that A is real and symmetric.

Let us consider what happens to the Arnoldi process in this special case. Of course, all of the equations of Lectures 33 and 34 still apply, and in each formula we can replace * by ^{T}. The first thing we notice is that it follows from (33.12) that the Ritz matrix H_n is symmetric. Therefore its eigenvalues, the Ritz values or Lanczos estimates (33.10), are also real. This seems natural enough, since the eigenvalues of A are real.

The second thing we notice is more dramatic. Since H_n is both symmetric and Hessenberg, it is tridiagonal. This means that in the inner loop of the Arnoldi iteration (Algorithm 33.1), the limits 1 to n can be replaced by $n-1$ to n. Thus instead of the $(n+1)$-term recurrence (33.4) at step n, the Lanczos iteration involves just a three-term recurrence. The result is that each step of the Lanczos iteration is much cheaper than the corresponding step of the

Arnoldi iteration. In Lecture 38 we shall see that analogously, for solving $Ax = b$, each step of the conjugate gradient iteration is much cheaper than the corresponding step of GMRES.

The fact that H_n is tridiagonal is so important that it is worth reviewing how it arises from the symmetry of A. The key equation is (33.12), which we can write entry-wise for real matrices A, H_n, and Q_n as

$$h_{ij} = q_i^T A q_j. \tag{36.1}$$

This implies that $h_{ij} = 0$ for $i > j + 1$, since $Aq_j \in \langle q_1, q_2, ..., q_{j+1} \rangle$ and the Krylov vectors are orthogonal. Taking the transpose gives

$$h_{ij} = q_j^T A^T q_i. \tag{36.2}$$

If $A = A^T$, then $h_{ij} = 0$ for $j > i + 1$ by the same reasoning as before. This simple argument leading to a three-term recurrence relation applies to arbitrary self-adjoint operators, not just to matrices.

The Lanczos Iteration

Since a symmetric tridiagonal matrix contains only two distinct vectors, it is customary to replace the generic notation a_{ij} by new variables. Let us write $\alpha_n = h_{nn}$ and $\beta_n = h_{n+1,n} = h_{n,n+1}$. Then H_n becomes

$$T_n = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{36.3}$$

In this notation Algorithm 33.1 takes the following form.

Algorithm 36.1. Lanczos Iteration

```
$\beta_0 = 0,\ \ q_0 = 0,\ \ b =$ arbitrary, $\ q_1 = b/\|b\|$
for $n = 1, 2, 3, \ldots$
    $v = Aq_n$
    $\alpha_n = q_n^T v$
    $v = v - \beta_{n-1} q_{n-1} - \alpha_n q_n$
    $\beta_n = \|v\|$
    $q_{n+1} = v/\beta_n$
```

Each step consists of a matrix-vector multiplication, an inner product, and a couple of vector operations. If A has enough sparsity or other structure that matrix-vector products can be computed cheaply, then such an iteration can be applied without too much difficulty to problems of dimensions in the tens or hundreds of thousands.

The following theorem summarizes some of the properties of the Lanczos iteration (when carried out in exact arithmetic, of course, as with all such theorems in this book). Nothing here is new; these are restatements in the new notation of the results of Theorems 33.1 and 34.1 for the Arnoldi iteration.

Theorem 36.1. *The matrices Q_n of vectors q_n generated by the Lanczos iteration are reduced QR factors of the Krylov matrix* (33.6),

$$K_n = Q_n R_n. \tag{36.4}$$

The tridiagonal matrices T_n are the corresponding projections

$$T_n = Q_n^* A Q_n, \tag{36.5}$$

and the successive iterates are related by the formula

$$AQ_n = Q_{n+1}\tilde{T}_n, \tag{36.6}$$

which we can write in the form of a three-term recurrence at step n,

$$Aq_n = \beta_{n-1}q_{n-1} + \alpha_n q_n + \beta_n q_{n+1}. \tag{36.7}$$

As long as the Lanczos iteration does not break down (i.e., K_n is of full rank n), the characteristic polynomial of T_n is the unique polynomial $p^n \in P^n$ that solves the Arnoldi/Lanczos approximation problem (34.3), *i.e., that achieves*

$$\| p^n(A)b \| = \textit{minimum}. \tag{36.8}$$

Lanczos and Electric Charge Distributions

In practice, the Lanczos iteration is used to compute eigenvalues of large symmetric matrices just as the Arnoldi iteration is used for nonsymmetric matrices (Lecture 34). At each step n, or at occasional steps, the eigenvalues of the growing tridiagonal matrix T_n are determined by standard methods. These are the Ritz values or "Lanczos estimates" (33.10) for the given matrix A and starting vector q_1. Often some of these numbers are observed to converge geometrically to certain limits, which can then be expected to be eigenvalues of A.

As with the Arnoldi iteration, it is the outlying eigenvalues of A that are most often obtained first. This assertion can be made more precise by the following rule of thumb:

> If the eigenvalues of A are more evenly spaced than Chebyshev points, then the Lanczos iteration will tend to find outliers.

Here is what this statement means. Suppose the m eigenvalues $\{\lambda_j\}$ of A are spread reasonably densely around an interval on the real axis. Since the Lanczos iteration is scale- and translation-invariant (Theorem 34.2), we can assume without loss of generality that this interval is $[-1,1]$. The m *Chebyshev points* in $[-1,1]$ are defined by the formula

$$x_j = \cos\theta_j, \qquad \theta_j = \frac{(j-\frac{1}{2})\pi}{m}, \qquad 1 \le j \le m. \tag{36.9}$$

The exact definition is not important; what matters is that these points cluster quadratically near the endpoints, with the spacing between points $O(m^{-1})$ in the interior and $O(m^{-2})$ near ±1. The rule of thumb asserts that if the eigenvalues $\{\lambda_j\}$ of A are more evenly distributed than this—less clustered at the endpoints—then the Ritz values computed by a Lanczos iteration will tend to converge to the outlying eigenvalues first. In particular, an approximately uniform eigenvalue distribution will produce rapid convergence towards outliers. Conversely, if the eigenvalues of A are more than quadratically clustered at the endpoints—a situation not so common in practice—then we can expect convergence to some of the "inliers."

These observations can be given a physical interpretation. Consider m point charges free to move about the interval $[-1,1]$. Assume that the repulsive force between charges located at x_j and x_k is proportional to $|x_j - x_k|^{-1}$. (For electric charges in 3D the force would be $|x_j - x_k|^{-2}$, but this becomes $|x_j - x_k|^{-1}$ in 2D, where we can view each point as the intersection of an infinite line in 3D with the plane.) Let these charges distribute themselves in a minimal-energy equilibrium in $[-1,1]$. Then this minimal-energy distribution and the Chebyshev distribution are approximately the same, and in the limit $m \to \infty$, they both converge to a limiting continuous charge density distribution proportional to $(1-x^2)^{-1/2}$.

Think of the eigenvalues of A as point charges. If they are distributed approximately in a minimal-energy configuration in an interval, then the Lanczos iteration will be useless; there will be little convergence before step $n = m$. If the distribution is very different from this, however, then there is likely to be rapid convergence to some eigenvalues, namely, the eigenvalues in regions where there is "too little charge" in the sense that if the points were free to move, more would tend to cluster here. The rule of thumb can now be restated:

> The Lanczos iteration tends to converge to eigenvalues in regions of "too little charge" for an equilibrium distribution.

The explanation of this observation depends on the connection (36.8) of the Lanczos iteration with polynomial approximation. Some of the details are worked out in Exercise 36.2.

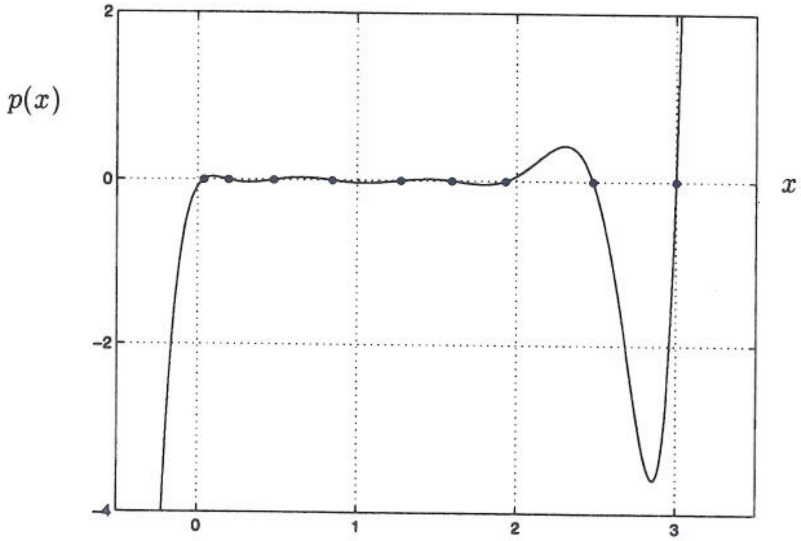

Figure 36.1. *Plot of the Lanczos polynomial at step* 9 *of the Lanczos iteration for the matrix* (36.10). *The roots are the Ritz values or "Lanczos eigenvalue estimates." The polynomial is small throughout* $[0,2] \cup \{2.5\} \cup \{3.0\}$. *To achieve this, it must place one root near* 2.5 *and another very near* 3.0.

Example

The convergence of the Lanczos iteration is best illustrated by a numerical example. Let A be the 203×203 matrix

$$A = \operatorname{diag}(0, .01, .02, \ldots, 1.99, 2, \;\; 2.5, \;\; 3.0). \tag{36.10}$$

The spectrum of A consists of a dense collection of eigenvalues throughout $[0,2]$ together with two outliers, 2.5 and 3.0. We carry out a Lanczos iteration beginning with a random starting vector q_1.

Figure 36.1 shows the Ritz values and the associated Lanczos polynomial at step $n = 9$. Seven of the Ritz values lie in $[0,2]$, and the polynomial is uniformly small on that interval; the beginnings of a tendency for the Ritz values to cluster near the endpoints can be detected. The other two Ritz values lie near the eigenvalues at 2.5 and 3.0. The leading three Ritz values are

$$1.93, \quad 2.48, \quad 2.999962.$$

Evidently we have little accuracy in the lower eigenvalues but five-digit accuracy in the leading one. A plot like this gives an idea of why outliers tend to be estimated accurately. The graph of $p(x)$ is so steep for $x \approx 3$ that if $p(3)$ is to be small, there must be a root of p very close to 3. This steepness of the graph is related to the presence of "too little charge" near this point. If the

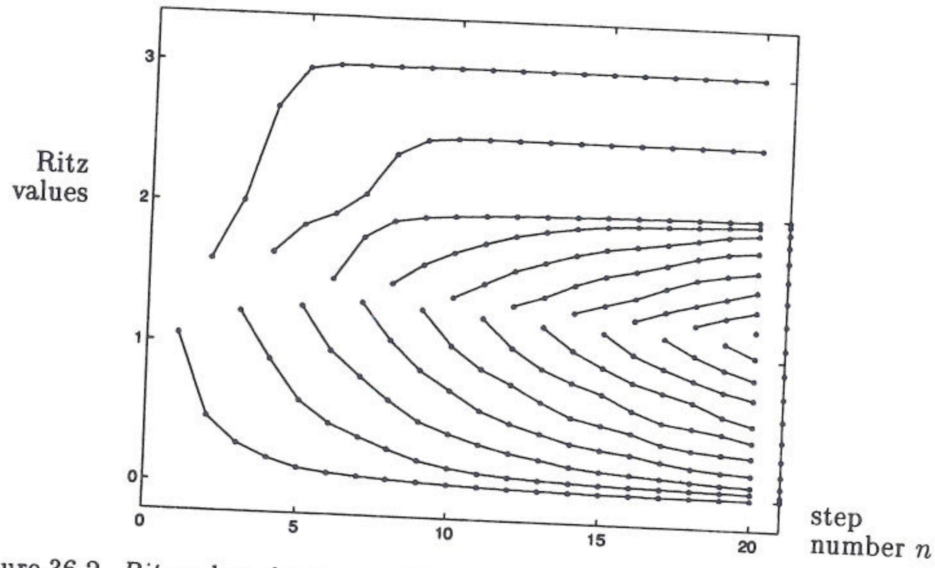

Figure 36.2. *Ritz values for the first* 20 *steps of the Lanczos iteration applied to the same matrix. The convergence to the eigenvalues* 2.5 *and* 3.0 *is geometric. Little useful convergence to individual eigenvalues occurs in the* [0, 2] *part of the spectrum. Instead, the Ritz values in* [0, 2] *approximate Chebyshev points in that interval, marked by dots on the right-hand boundary.*

charges were free to move about [0, 3] to minimize energy, more points would cluster near $x = 3$, and $p(x)$ would not be so steep there.

At step 20 the leading three Ritz values are

$$1.9906, \quad 2.499999999987, \quad 3.00000000000000.$$

Now we have about fifteen digits of accuracy in the leading eigenvalue and twelve digits in the second. A plot of $p(x)$ would be correspondingly steep near the points 2.5 and 3.0. Note that convergence to the third eigenvalue is also beginning to occur, a reflection of the fact that the eigenvalues in [0, 2] are distributed evenly rather than in a Chebyshev distribution.

An "aerial view" of the convergence process appears in Figure 36.2, which shows the Ritz values for all steps from $n = 1$ to $n = 20$. Each vertical slice of this plot corresponds to the Ritz values at one iteration; the lines connecting the dots help the eye follow what is going on but have no precise meaning. The plot shows pronounced convergence to the leading eigenvalue after about $n = 5$ and to the next one around $n = 10$. In the interval [0, 2] containing the other eigenvalues, they show a density of Ritz values approximately proportional to $(1 - x^2)^{-1/2}$, with very clear bunching at endpoints.

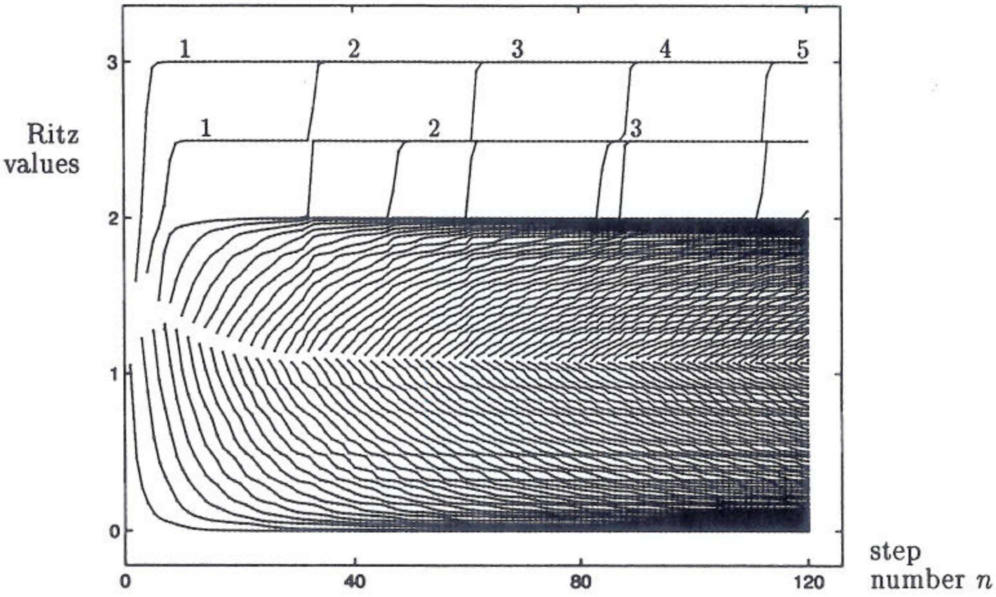

Figure 36.3. *Continuation to* 120 *steps of the Lanczos iteration. The numbers indicate multiplicities of the Ritz values. Note the appearance of four "ghost" copies of the eigenvalue* 3.0 *and two "ghost" copies of the eigenvalue* 2.5.

Rounding Errors and "Ghost" Eigenvalues

Rounding errors have a complex effect on the Lanczos iteration and, indeed, on all iterations of numerical linear algebra based on three-term recurrence relations. The source of the difficulty is easily identified. In an iteration based on an n-term recurrence relation, such as Arnoldi or GMRES, the vectors $q_1, q_2, q_3, \dots$ are forced to be orthogonal by explicit Gram–Schmidt operations. Three-term recurrences like Lanczos and conjugate gradients, however, depend upon orthogonality of the vectors $\{q_j\}$ to arise "automatically" from a mathematical identity. In practice, such identities are not accurately preserved in the presence of rounding errors, and after a number of iterations, orthogonality is lost.

The loss of orthogonality in practical Lanczos iterations sounds wholly bad, but the situation is more subtle than that. As it happens, loss of orthogonality is connected closely with the convergence of Ritz values to eigenvalues of A. A great deal is known about this subject, though not as much as one might like; we shall not give details.

Because of complexities like these, no straightforward theorem is known to the effect that the Lanczos or conjugate gradient iterations is stable in the sense defined in this book. Nonetheless, these iterations are extraordinarily useful in practice. Figure 36.3 gives an idea of the way in which instability is often manifested in practice without preventing the iteration from being useful. The figure is a repetition of Figure 36.2, but for 120 instead of 20 steps of the iteration. Everything looks as expected until around step 30, when a second copy of the eigenvalue 3.0 appears among the Ritz values. A third copy appears around step 60, a fourth copy around step 90, and so on. Meanwhile, additional copies of the eigenvalue 2.5 also appear around step 40 and 80 and (just beginning to be visible) 120. These extra Ritz values are known as "ghost" eigenvalues, and they have nothing to do with the actual multiplicities of the corresponding eigenvalues of A.

A rigorous analysis of the phenomenon of ghost eigenvalues is complicated. Intuitive explanations, however, are not hard to devise. One idea is that in the presence of rounding errors, one should think of each eigenvalue of A not as a point but as a small interval of size roughly $O(\epsilon_{\text{machine}}\|A\|)$; ghost eigenvalues arise from the need for $p(z)$ to be small not just at the exact eigenvalues but throughout these small intervals. Another, rather different explanation is that convergence of a Ritz value to an eigenvalue of A annihilates the corresponding eigenvector component in the vector being operated upon; but in the presence of rounding errors, random noise must be expected to excite that component slightly again. After sufficiently many iterations, this previously annihilated component will have been amplified enough that another Ritz value is needed to annihilate it again—and then again, and again.

Both of these explanations capture some of the truth about the behavior of the Lanczos iteration in floating point arithmetic. The second one has perhaps more quantitative accuracy.

Exercises

36.1. In Lecture 27 it was pointed out that the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{m\times m}$ are the stationary values of the Rayleigh quotient $r(x) = (x^TAx)/(x^Tx)$ for $x \in \mathbb{R}^m$. Show that the Ritz values at step n of the Lanczos iteration are the stationary values of $r(x)$ if x is restricted to $\mathcal{K}_n$.

36.2. Consider a polynomial $p \in P^n$, i.e., $p(z) = \prod_{k=1}^n (z - z_k)$ for some $z_k \in \mathbb{C}$.

(a) Write $\log|p(z)|$ as a sum of n terms corresponding to the points z_k.

(b) Explain why the term involving z_k can be interpreted as the potential corresponding to a negative unit point charge located at z_k, if charges repel in inverse proportion to their separation. Thus $\log|p(z)|$ can be viewed as the potential at z induced by n point charges.

(c) Replacing each charge -1 by $-1/n$ and taking the limit $n \to \infty$, we get a continuous charge density distribution $\mu(\zeta)$ with integral -1, which we can expect to be related to the limiting density of zeros of polynomials $p \in P^n$ as

$n \to \infty$. Write an integral representing the potential $\varphi(z)$ corresponding to $\mu(\zeta)$, and explain its connection to $|p(z)|$.

(d) Let S be a closed, bounded subset of $\mathbb{C}$ with no isolated points. Suppose we seek a distribution $\mu(z)$ with support in S that minimizes $\max_{z\in S}\varphi(z)$. Give an argument (not rigorous) for why such a $\mu(z)$ should satisfy $\varphi(z) =$ constant throughout S. Explain why this means that the "charges" are in equilibrium, experiencing no net forces. In other words, S is like a 2D electrical conductor on which a quantity -1 of charge has distributed itself freely. Except for an additive constant, $\varphi(z)$ is the *Green's function* for S.

(e) As a step toward explaining the rule of thumb of p. 279, suppose that A is a real symmetric matrix with spectrum densely distributed in $[a,b]\cup\{c\}\cup[d,e]$ for $a < b < c < d < e$. Thus (b,d) is a region of "too little charge" for the set $S = [a,e]$. Explain why rapid convergence of a Ritz value to c can be expected, and estimate the rate of convergence in terms of the equilibrium potential $\varphi(z)$ associated with the set $S' = [a,b] \cup [c,d]$.

36.3. Let A be the 1000 × 1000 symmetric matrix whose entries are all zero except for $a_{ij} = \sqrt{i}$ on the diagonal, $a_{ij} = 1$ on the sub- and superdiagonals, and $a_{ij} = 1$ on the 100th sub- and superdiagonals, i.e., for $|i - j| = 100$. Determine the smallest eigenvalue of A to six digits of accuracy by the Lanczos iteration.

36.4. As a special case of the Arnoldi lemniscates of Lecture 34, "Lanczos lemniscates" can be employed to illustrate the convergence of the Lanczos iteration. Find a way to modify your program of Exercise 36.3 to plot the Lanczos lemniscates at each step. Your method need not be elegant, efficient, or numerically robust. Produce plots of Lanczos lemniscates at steps $n = 1, 2, \ldots, 12$ for the example of Figure 36.2 and for an example of your own choosing.