2020-02-28

1 Eigenvalue problems

An eigenvalue $\lambda \in \mathbb{C}$ of a matrix $A \in \mathbb{C}^{n \times n}$ is a value for which the equations $Av = v\lambda$ and $w^*A = \lambda w^*$ have nontrivial solutions (the eigenvectors w^* and v). Together, (λ, v) forms an eigenpair and (λ, v, w^*) forms an eigentriple. An eigenvector is a basis for a one-dimensional invariant subspace: that is, A maps anything multiple of v to some other multiple of v. More generally, a matrix $V \in \mathbb{C}^{n \times m}$ spans an invariant subspace if $AV = VL$ for some $L \in \mathbb{C}^{n \times m}$.

Associated with any square A, we can write a matrix Q whose columns form an orthonormal basis for nested invariant subspaces of A; that is, the first k columns of Q form a k-dimensional invariant subspace of A. This structure of nested invariant subspaces gives us that

$$AQ = QT,$$

where T is an upper triangular matrix. The factorization

$$A = QTQ^*$$

is a *Schur factorization*. The next two lectures will be devoted to methods to compute Schur factorizations (or parts of Schur factorizations). The Schur factorization is nearly as versatile as, and is far more numerically stable than, the *Jordan canonical form*

$$AV = VJ.$$

where J is a block diagonal matrices with *Jordan* blocks of the form

$$J_\lambda = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{bmatrix}.$$

The *algebraic multiplicity* of an eigenvalue λ is the number of times it appears on the diagonal of the Jordan form, or the number of times the factor $z - \lambda$ divides the characteristic polynomial $\det(A - zI)$. The *geometric multiplicity* is given by the number of Jordan blocks associated to λ, or by

the dimension of the null space of $(A - \lambda I)$. In general, there is exactly one eigenvector of A for each Jordan block, and the eigenvectors form a basis iff A is diagonalizable – that is, if A has only 1-by-1 Jordan blocks and all geometric and algebraic multiplicities match. The diagonalizable matrices form a dense set in $\mathbb{C}^{n \times n}$, a fact which is often convenient in proofs (since an argument for the diagonalizable case together with a continuity argument often yields a general solution). This fact also explains part of why the Jordan canonical form is annoying for numerical work: if every matrix is an arbitrarily small perturbation of something diagonalizable, then the Jordan form is discontinuous as a function of A! Even among the diagonalizable matrices, though, the eigenvector decomposition is often overrated for computational purposes. Poor conditioning of the eigenvector basis can make diagonalization a numerically unstable business, and most computations that are naively formulated in terms of an eigenvector basis can equally well be formulated in terms of Schur basis.

In *generalized eigenvalue problems*, we ask for nontrivial solutions to

$$(A - \lambda B)v = 0.$$

There are also *nonlinear eigenvalue problems*, which show up in my research but which we will not talk about in class. In addition to these variants on the eigenvalue problem, there are also many different factors that affect the how we choose algorithms. Is the problem...

1. nonsymmetric or symmetric?

2. standard or generalized?

3. to find all eigenvalues or just a few?

4. to compute eigenvectors, invariant subspaces, or just eigenvalues?

For different answers to these questions, there are different "best" choices of algorithm. For the next week or two, we will focus specifically on the problem of computing eigenpairs, invariant subspaces, and Schur forms for nonsymmetric matrices. We will only briefly touch on the special case of the symmetric problem, which has so much more mathematical structure that it is treated almost entirely differently from the nonsymmetric case.

2   The symmetric case

In general, a real matrix can have complex eigenvalues (though in conjugate pairs), and it may or may not have a basis of eigenvectors. In the case of real *symmetric* matrices $(A = A^T)$, we have much more structure: namely,

- All the eigenvalues are real.

- There is a complete *orthonormal* basis of eigenvectors.

To see the former, observe that if (v, λ) is an eigenpair and $\|v\| = 1$ then

$$\lambda = v^* A v = v^* \bar{A} v = \bar{\lambda},$$

which implies that λ is real. To see that eigenvectors associated with different eigenvalues must be orthogonal, note that if (v, λ) and (u, μ) are eigenpairs with $\lambda \neq \mu$, then

$$v^* A u = \begin{cases} (Av)^* u = \lambda v \cdot u \\ v^*(Au) = \mu v \cdot u \end{cases}$$

and the only way for these to be the same is if $v \cdot u = 0$. Combining these two facts about the symmetric eigenvalue problem, we usually write the standard decomposition

$$A = Q \Lambda Q^T$$

where Q is an orthogonal matrix of eigenvectors and Λ is the corresponding diagonal matrix of eigenvalues.

We often use symmetric matrices to represent quadratic forms, and this is one reason why symmetric eigenvalue problems are so common. If $A = Q\Lambda Q^T$, then we can define $w = Q^T v$ to get the expression

$$v^* A v = \sum_{i=1}^{n} w_i^2 \lambda_i.$$

If $\|v\|^2 = 1$ (implying $\|w\|^2 = 1$), then we can see $v^* A v$ is a weighted average of the eigenvalues of A. Hence, the minimum or maximum of $v^* A v$ over all unit length vectors gives the largest and smallest of the eigenvalues of A; and, more generally, the eigenvalues are stationary points of $v^* A v$ subject to the constraint $\|v\|^2 = 1$. Sometimes we prefer to work with all nonzero

vectors rather than vectors with unit length, and hence define the *Rayleigh quotient*

$$\rho_A(v) = \frac{v^* A v}{v^* v};$$

this ratio plays a central role in the theory of the symmetric eigenproblem. In particular, if we differentiate $\rho_A(v)$ with respect to v, we find

$$\delta \rho_A(v) = \frac{2\delta v^*}{v^* v} \left(Av - v\rho_A(v) \right),$$

and the places where all directional derivatives are zero (the stationary points) satisfy the eigenvalue equation

$$Av = v\rho_A(v),$$

where v is the eigenvector and $\rho_A(v)$ is the eigenvalue.

3   Why eigenvalues?

I spend a lot of time thinking about eigenvalue problems. In part, this is because I look for problems that can be solved via eigenvalues. But I might have fewer things to keep me out of trouble if there weren't so many places where eigenvalue analysis is useful! The purpose of this lecture is to tell you about a few applications of eigenvalue analysis, or perhaps to remind you of some applications that you've seen in the past.

3.1   Nonlinear equation solving

The eigenvalues of a matrix are the roots of the characteristic polynomial

$$p(z) = \det(zI - A).$$

One way to compute eigenvalues, then, is to form the characteristic polynomial and run a root-finding routine on it. In practice, this is a terrible idea, if only because the root-finding problem is often far more sensitive than the original eigenvalue problem. But even if sensitivity were not an issue, finding *all* the roots of a polynomial seems like a nontrivial undertaking. Iterations like Newton's method, for example, only converge locally. In fact, the `roots` command in MATLAB computes the roots of a polynomial by finding the

eigenvalues of a corresponding *companion matrix* with the polynomial coefficients on the first row, ones on the first subdiagonal, and zeros elsewhere:

$$C = \begin{bmatrix} c_{d-1} & c_{d-2} & \cdots & c_1 & c_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

The characteristic polynomial for this matrix is precisely

$$\det(zI - C) = z^d + c_{d-1}z^{d-1} + \ldots + c_1 z + c_0.$$

There are some problems that connect to polynomial root finding, and thus to eigenvalue problems, in surprising ways. For example, the problem of finding "optimal" rules for computing integrals numerically (sometimes called Gaussian quadrature rules) boils down to finding the roots of orthogonal polynomials, which can in turn be converted into an eigenvalue problem; see, for example, "Calculation of Gauss Quadrature Rules" by Golub and Welsch (*Mathematics of Computation*, vol 23, 1969).

More generally, eigenvalue problems are one of the few examples I have of a nonlinear equation where I can find *all* solutions in polynomial time! Thus, if I have a hard nonlinear equation to solve, it is very tempting to try to massage it into an eigenvalue problem, or to approximate it by an eigenvalue problem.

3.2   Optimization

As we noted before, the symmetric eigenvalue problem has an interpretation in terms of optimization of a quadratic form over unit length vectors. More generally, one can look at generalized eigenvalue problems in terms of optimization of a ratio of quadratic forms. We now discuss some applications where this interpretation is useful.

Recall that the matrix 2-norm is defined as

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

Taking squares and using the monotonicity of the map $z \to z^2$ for nonnegative arguments, we have

$$\|A\|_2^2 = \max_{\|x\|^2=1} \|Ax\|^2 = \max_{x^T x=1} x^T A^T A x.$$

The x that solves this constrained optimization problem must be a stationary point for the augmented Lagrangian function

$$L(x, \lambda) = x^T A^T A x - \lambda (x^T x - 1),$$

i.e.

$$\nabla_x L(x, \lambda) = 2(A^T A x - \lambda x) = 0$$
$$\nabla_\lambda L(x, \lambda) = x^T x - 1 = 0.$$

These equations say that x is an eigenvector of $A^T A$ with eigenvalue λ. The largest eigenvalue of $A^T A$ is therefore $\|A\|_2^2$.

More generally, if H is any Hermitian matrix, the *Rayleigh quotient*

$$\rho_H(v) = \frac{v^* H v}{v^* v}$$

has stationary points exactly when v is an eigenvector of H. Optimizing the Rayleigh quotient is therefore example of a *non-convex* global optimization problem that I know how to solve in polynomial time. Such examples are rare, and so it is tempting to try to massage other nonconvex optimization problems so that they look like Rayleigh quotient optimization, too.

To give an example of a nonconvex optimization that can be usefully approximated using Rayleigh quotients, consider the superficially unrelated problem of graph bisection. Given an undirected graph G with vertices V and edges $E \subset V \times V$, we want to find a partition of the nodes into two equal-size sets such that few edges go between the sets. That is, we want to write V as a disjoint union $V = V_1 \cup V_2$, $|V_1| = |V_2|$, such that the number of edges cut $|E \cap (V_1 \times V_2)|$ is minimized. Another way to write the same thing is to label each node i in the graph with $x_i \in \{+1, -1\}$, and define V_1 to be all the nodes with label $+1$, V_2 to be all the nodes with label -1. Then the condition that the two sets are the same size is equivalent to

$$\sum_i x_i = 0,$$

and the number of edges cut is

$$\frac{1}{4} \sum_{(i,j) \in E} (x_i - x_j)^2$$

We can rewrite the constraint more concisely as $e^T x = 0$, where e is the vector of all ones; as for the number of edges cut, this is

$$\text{edges cut} = \frac{1}{4} x^T L x$$

where the *graph Laplacian* L has the node degrees on the diagonal and -1 in off-diagonal entry (i, j) iff there is an edge from i to j.

Unsurprisingly, the binary quadratic programming problem

$$\text{minimize } x^T L x \text{ s.t. } e^T x = 0 \text{ and } x \in \{+1, -1\}^n$$

is NP-hard, and we know of no efficient algorithms that are guaranteed to work for this problem in general. On the other hand, we can *relax* the problem to

$$\text{minimize } v^T L v \text{ s.t. } e^T v = 0 \text{ and } \|v\|^2 = n, v \in \mathbb{R}^n,$$

and this problem is an eigenvalue problem: v is the eigenvector associated with the smallest positive eigenvalue of L, and $v^T L v$ is n times the corresponding eigenvalue. Since the constraint in the first problem is strictly stronger than the constraint in the second problem, $n\lambda_2(L)$ is in fact a lower bound on the smallest possible cut size, and the sign pattern of v often provides a partition with a small cut size. This is the heart of *spectral partitioning* methods.

3.3   Dynamics

Eigenvalue problems come naturally out of separation of variables methods, and out of transform methods for the dynamics of discrete or continuous linear time invariant systems, including examples from physics and from probability theory. They allow us to analyze complicated high-dimensional dynamics in terms of simpler, low-dimensional systems. We consider two examples: separation of variables for a free vibration problem, and convergence of a discrete-time Markov chain.

3.3.1   Generalized eigenvalue problems and free vibrations

One of the standard methods for solving differential equations is *separation of variables*. In this approach, we try to write special solutions as a product of simpler functions, and then write the equations that those functions have

to satisfy. As an example, consider a differential equation that describes the free vibrations of a mechanical system:

$$M\ddot{u} + Ku = 0$$

Here $M \in \mathbb{R}^{n \times n}$ is a symmetric positive definite *mass matrix* and $K \in \mathbb{R}^{n \times n}$ is a symmetric *stiffness matrix* (also usually positive definite, but not always). We look for solutions to this system of the form

$$u(t) = u_0 \cos(\omega t),$$

where u_0 is a fixed vector. To have a solution of this form, we must have

$$Ku_0 - \omega^2 M u_0 = 0,$$

i.e. (ω^2, u_0) is an eigenpair for a *generalized* eigenvalue problem. In fact, the eigenvectors for this generalized eigenvalue problem form an M-orthonormal basis for $\mathbb{R}^n$, and so we can write *every* free vibration as a linear combination of these simple "modal" solutions.

3.3.2   Markov chain convergence and the spectral gap

This high-level idea of using the eigenvalue decomposition to understand dynamics is not limited to differential equations, nor to mechanical systems. For example, a *discrete-time Markov chain* on n states is a random process where the state X_{k+1} is a random variable that depends only on the state X_k. The *transition matrix* for the Markov chain is a matrix P where P_{ij} is the (fixed) probability of transitioning to state i from state j, i.e.

$$P_{ij} = P\{X_{k+1} = j | X_k = i\}.$$

Let $\pi^{(k)} \in \mathbb{R}^n$ be the distribution vector at time k, i.e.

$$\pi_i^{(k)} = P\{X_k = i\}.$$

Then we have the recurrence relationship

$$(\pi^{(k+1)})^T = (\pi^{(k)})^T P.$$

In general, this means that

$$(\pi^{(k)})^T = (\pi^{(0)})^T P^k.$$

Now, suppose the transition matrix P is diagonalizable, i.e. $P = V\Lambda V^{-1}$. Then

$$P^k = V\Lambda V^{-1}V\Lambda V^{-1}\ldots V\Lambda V^{-1} = V\Lambda\ldots\Lambda V^{-1} = V\Lambda^k V^{-1},$$

and so

$$(\pi^{(k)})^T = (\pi^{(0)})^T V\Lambda^k V^{-1}.$$

An *ergodic* Markov chain has one eigenvalue at one, and all the other eigenvalues are less than one in modulus. In this case, the row eigenvector associated with the eigenvalue at one can be normalized so that the coefficients are all positive and sum to 1. This normalized row eigenvector $\pi^{(*)}$ represents the *stationary* distribution to which the Markov chain eventually converges. To compute the rate of convergence, one looks at

$$\|(\pi^{(k)} - \pi^{(*)})^T\| = \|(\pi^{(0)} - \pi^{(*)})^T(V\tilde{\Lambda}^k V^{-1})\| \leq \|(\pi^{(0)} - \pi^{(*)})^T\|\,\kappa(V)\|\tilde{\Lambda}\|^k$$

where $\Lambda = \mathrm{diag}(1, \lambda_2, \lambda_3, \ldots)$, $|\lambda_i| \geq |\lambda_{i+1}|$, and $\tilde{\Lambda} = \mathrm{diag}(0, \lambda_2, \lambda_3, \ldots)$. In most reasonable operator norms, $|\tilde{\Lambda}|^k = |\lambda_2|^k$, and so a great deal of the literature on convergence of Markov chains focuses on $1 - |\lambda_2|$, called the *spectral gap*. But note that this bound does not depend on the eigenvalues alone! The condition number of the eigenvector matrix also plays a role, and if $\kappa(V)$ is very large, then it may take a long time indeed before anyone sees the asymptotic behavior reflected by the spectral gap.

3.4   Matrix functions

If $f : \mathbb{C} \to \mathbb{C}$ is analytic[1] in some nice domain containing the eigenvalues of a matrix A, then we can define $f(A)$ by a power series expansion; for example,

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k,$$

and we can write similar expansions for matrix logarithms, square roots, trig functions, and so forth. You have already seen one example of this in the Neumann series expansion for $(I - A)^{-1}$.

---

[1]An analytic function has a convergent power series expansion. These functions have many other nice properties as well, as you might learn if you take a complex analysis class — but a complex analysis class is not a pre-requisite for the course.

Suppose A is diagonalizable, i.e. $A = V\Lambda V^{-1}$. Then $A^k = V\Lambda^k V^{-1}$; and if we have a power series

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

that converges for each eigenvalue, we can write

$$f(A) = \sum_{k=0}^{\infty} c_k A^k = V\left(\sum_{k=0}^{\infty} c_k \Lambda^k\right) V^{-1} = V\left[\mathrm{diag}(f(\lambda_i))_{i=1}^{n}\right] V^{-1}.$$

Hence, $f(A)$ has the same eigenvectors as A, but the eigenvalues are mapped via f. This is the *spectral mapping theorem*, and it is useful both as a way of reasoning about matrix functions and as a method of computing such functions — though in practice, the standard eigenvalue-based method for computing matrix functions (the Schur-Parlett method) does not diagonalize the matrix, but instead uses a Schur decomposition.

3.5   Deductions from eigenvalue distributions

In most of our examples so far, we have considered both the eigenvalues and the eigenvectors. Now let us turn to a simple example where the distribution of eigenvalues can be illuminating.

Let A be the adjacency matrix for a graph, i.e.

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from } i \text{ to } j \\ 0, & \text{otherwise.} \end{cases}$$

Then $(A^k)_{ij}$ is the number of paths of length k from node i to node j. In particular, $(A^k)_{ii}$ is the number of cycles of length k that start and end at node i, and $\mathrm{trace}(A^k)$ is the total number of length k cycles starting from any node. Recalling that the trace of a matrix is the sum of the eigenvalues, and that the eigenvalues of a matrix power are the power of the eigenvalues, we have that

$$\#\text{ paths of length } k = \sum_i \lambda_i(A)^k,$$

where $\lambda_i(A)$ are the eigenvalues of A; and asymptotically, the number of cycles of length k for very large k scales like $\lambda_1(A)^k$, where $\lambda_1(A)$ is the largest eigenvalue of the matrix A.

While the statement above deals only with eigenvalues and not with eigen-vectors, we can actually say more if we include the eigenvector; namely, if the graph A is irreducible (i.e. there is a path from every state to every other state), then the largest eigenvalue $\lambda_1(A)$ is a real, simple eigenvalue, and asymptotically the number of paths from any node i to node j scales like the (i, j) entry of the rank one matrix

$$\lambda_1^k vw^T$$

where v and w are the column and row eigenvectors of A corresponding to the eigenvalue λ_1, scaled so that $w^T v = 1$.