Functions of matrices. Matrices with non-negative entries.

A. Eremenko

August 25, 2024

Functions of matrices. If p is a polynomial, and A a square matrix, then

$$p(A) = c_0 + c_1 A + \ldots + c_d A^d$$

is defined, and this is a matrix of the same size as A.

One can then take limits of polynomials, when limits exist, for example, we defined

$$e^A = \sum_{m=0}^{\infty} \frac{A^m}{m!}.$$

When A is diagonalizable, we have

$$A = B\Lambda B^{-1},$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and B is the matrix whose columns are eigenvectors which make a basis of the whole space.

Then we have a closed formula for a function of A

$$p(A) = Bp(\Lambda)B^{-1}, \tag{1}$$

which is proved first for powers of A, then for polynomials, and then for functions which are limits of polynomials. Notice that $p(\Lambda) = \mathrm{diag}(p(\lambda_1), \ldots, p(\lambda_n))$, so we have an explicit formula for $p(A)$.

About functions of matrices, there are two important theorems:

Spectral Mapping Theorem. *Eigenvalues of $p(A)$ are $p(\lambda)$, where λ are eigenvalues of A. If $\mathbf{v}$ is an eigenvector of A with eigenvalue λ then the same $\mathbf{v}$ is also an eigenvector of $p(A)$ with eigenvalue $p(\lambda)$.*

The second statement is evident: if $A\mathbf{v} = \lambda\mathbf{v}$ then $A^m\mathbf{v} = \lambda^m\mathbf{v}$, so for polynomials p we have $p(A)\mathbf{v} = p(\lambda)\mathbf{v}$ and then we can pass to the limit. The first part of the theorem includes the statement that $p(A)$ has no other eigenvalues except $p(\lambda)$, where λ is an eigenvalue of A. This is easy to prove for diagonalizable matrices. Non-diagonalizable (deficient) matrices will be considered later.

Hamilton-Cayley Theorem. *For a square matrix A and its characteristic polynomial p_A, we always have $p_A(A) = 0$.*

This theorem is true for *all* square matrices, and the proof for diagonalizable matrices is evident from (1): if p is the characteristic polynomial of A the RHS of (1) is 0.

Remark. 1. I did not discuss the exact notion of limit when "limits of polynomials" were mentioned. In all cases in this course polynomials converge uniformly on each bounded set of the complex plane. For example, the Taylor series of e^z converges in this way. A more careful analysis shows that one only needs uniform convergence in a neighborhood of the set of eigenvalues. This is easy to see for diagonalizable matrices to apply the formula

$$p(A) = Bp(\Lambda)B^{-1}$$

one only needs to know $p(\lambda_j)$, not the values of p at other points.

For example we can define sin of any square matrix by the series

$$\sin A = \sum_{m=0}^{\infty} (-1)^m A^{2m+1}/(2m+1)! = A - A^3/6 + A^5/120 - \ldots.$$

Exercise: show that $X(t) = \sin(At)$ solves the second order matrix differential equation

$$X'' = -A^2 X, \quad \text{and} \quad X(0) = 0, \quad X'(0) = I.$$

2. The set of all eigenvalues of a matrix (or of an operator) is called the *spectrum*. This explains the names of the theorems in this lecture. Why such a name? It has a long and very interesting story. I may tell some of it in one of the next lectures. Meanwhile you may think what this word "spectrum" means in physics.

Matrices with non-negative entries

As we have seen before, eigenvalues contain the most important information about a matrix, but on the other hand they may be difficult to find. So it is important to be able to tell something about them without finding them explicitly.

Today we consider matrices with non-negative entries. One motivation for their study is a model of many natural and social phenomena which is called a *Finite Markov chain.* Imagine a system consisting of many objects, each object can be in one of the n *states* at every moment of time. For example, the objects are American people, and the states are the states United states. Every person resides in some definite state in a given year (say, for tax purposes).

At every moment of time m the probability that an object is in state j is denoted by a_j, $1 \leq j \leq n$. (Alternatively one can say that the a_j is simply the number of objects in state j divided by the total number of objects.) So the distribution of our objects over states at the moment m of time is represented by a vector $\mathbf{a}(m)$ of dimension n whose coordinates are $a_j(m)$, $1 \leq j \leq n$.

In the next moment of time an object can move from state j to state i with probability $p_{i,j}$. (So that p_{jj} is the probability that it will not move).

The state of the whole big system at time m is described by a vector $\mathbf{a}(m)$ with coordinates $a_j(m)$. In the next moment of time the whole system will be in the new state

$$\mathbf{a}(m+1) = P\mathbf{a}(m), \quad \text{where} \quad P = (p_{ij}) \tag{2}$$

is an $n \times n$ matrix of *transition probabilities.*

Read Example on p. 257. In this example, $n = 2$, objects are people in the United States, and time is measured in years. The vector $\mathbf{a}(m) = (y_m, z_m)^T$ is the distribution of people between "Outside of California" and "Inside California". The transition matrix is

$$P = \left(\begin{array}{cc} 0.9 & 0.2 \\ 0.1 & 0.8 \end{array} \right).$$

Equation (2) gives

$$\mathbf{a}(m) = P^m\mathbf{a}(0).$$

We find eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0.7$ and corresponding eigenvectors $\mathbf{v}_1 = (2,1)^T$ and $\mathbf{v}_2 = (-1,1)^T$. They are linearly independent, so we can write

$$\mathbf{a}(0) = c_1\mathbf{v}_1 + c_2\mathbf{v}_2,$$

Then

$$\mathbf{a}(m) = c_1\lambda_1^n\mathbf{v}_1 + c_2\lambda_2^m\mathbf{v}_2.$$

Notice what happens "in the long run": since $|\lambda_2| < 1$, $\lambda_2^m \to 0$ as $m \to \infty$, while $\lambda_1^m = 1$ so

$$\mathbf{a}(m) \to c_1\mathbf{v}_1, \quad m \to \infty,$$

and this does not depend on the original distribution $\mathbf{a}_0$.

If the coordinates of $\mathbf{a}_m$ are interpreted as probabilities, they should add to 1, so $c_1 = 1/3$.

We obtained a limit distribution of population between the "rest of the US" and California; this distribution is $2:1$, it is given by the coordinates of the eigenvector $\mathbf{v}_1$ corresponding to the larger eigenvalue. This situation generalizes to any finite Markov chain with any number of states.

Everything depends on the following fundamental theorem about matrices with positive entries.

Perron's Theorem. *Let P be a square matrix with all entries $p_{ij} > 0$ (**strictly!**). Then the following statements hold:*

(i) There is a positive eigenvalue λ_1 which is greater than absolute value of any other eigenvalue: $\lambda_1 > |\lambda_j|$.

(ii) This eigenvalue λ_1 has one dimensional eigenspace spanned by an eigenvector $\mathbf{v}_1$ whose all coordinates are positive.

(iii) Eigenvectors corresponding to other eigenvalues $\lambda_j \neq \lambda_1$ cannot have all their coordinates of the same sign (some coordinates are positive some negative).

(iv) λ_1 is a simple root of the characteristic equation.

This eigenvalue λ_1 is called Perron's eigenvalue, and the corresponding eigenvector with positive coordinates is called a Perron eigenvector.

Returning to our finite Markov chain, we have a matrix P with non-negative entries (probabilities are non-negative!) Our matrix P has the additional property: all columns add to 1. Indeed, an object must move to *some* state (including the state where it was), so sum of all probabilities of all possible moves must be 1, that is

$$\sum_{i=1}^{n} p_{ij} = 1. \tag{3}$$

Matrices with non-negative entries and property (3) are called *stochastic.* Our vectors describing the states of the system also have non-negative entries which add to 1 (the system must be in *some* state, so probabilities of all states must add to 1).

Theorem. *Let P be a stochastic matrix with all entries strictly positive. Then the Perron eigenvalue $\lambda_1 = 1$. If $\mathbf{x}$ is any non-zero vector with non-negative coordinates then*

$$\lim_{m\to\infty} A^m \mathbf{x} = c\mathbf{v}_1,$$

where $\mathbf{v}_1$ is a Perron eigenvector.

Proof. Let $\mathbf{w} = (1, 1, \ldots, 1)$. Condition (3) can be conveniently written as

$$\mathbf{w}P = \mathbf{w}, \quad \text{or} \quad P^T\mathbf{w}^T = \mathbf{w}^T,$$

so $\mathbf{w}^T$ is an eigenvector of P^T with all coordinates positive and eigenvalue 1. Applying statement (iii) of Perron's theorem to P^T we conclude that Perron's eigenvalue of P^T equals 1. But P and P^T have the same characteristic polynomial, so all their eigenvalues are the same. So Perron's eigenvalue of P is also 1. So all other eigenvalues of P must have the absolute value less than 1. Suppose that P is diagonalizable. Then there is a basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and we can expand $\mathbf{x}$ in the basis

$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n.$$

applying P m times we obtain

$$P^m\mathbf{x} = c_1\mathbf{v}_1 + c_2\lambda_2^m\mathbf{v}_2 + \ldots + c_n\lambda_n^m\mathbf{v}_n.$$

since $|\lambda_j| < \lambda_1 = 1$ for all $j \geq 2$ we obtain the statement of the theorem.

The case of non-diagonalizable P will be dealt with later, when we study non-diagonalizable matrices.

Conditions of this theorem are not satisfied by all finite Markov chains: some entries of P can be zeros, and our theorem is not applicable in this case. Indeed, imagine a system whose states are divided into groups, and transition between groups never happens (the corresponding probabilities are zero). Then evidently it is not the case that every initial distribution will eventually evolve to a single equilibrium distribution.

The question is how to exclude this case. A stochastic matrix is called *ergodic* if *some power* P^N of P has strictly positive entries. In terms of the Markov chain this means that there is a positive probability of any transition $i \to j$ in N steps. It is exactly to these matrices that our theorem generalizes.

Control question. Determine (by hand, without a computer) whether this matrix is ergodic:

$$P = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/4 \\ 0 & 1/2 & 0 & 1/2 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/2 \end{pmatrix}.$$

If not ergodic, explain why. If ergodic, tell for which N will P^N have all entries positive.

Hint. Draw 6 dots on a sheet of paper, and label them with numbers 1 to 6. Think of the matrix P as representing a finite Markov chain. From each dot draw arrows representing possible transitions (that is transitions with positive probability). Thus an entry $p_{i,j} > 0$ is represented by the arrow from dot j to dot i. A brief inspection of the resulting picture allows you to answer the question: an entry p^N_{ij} of the matrix P^N is positive if a transition from j to i is possible in N steps.

Many other examples of applications of non-negative matrices are given in section 5.3, pp. 259–262, and in the problems to this section.

Let me mention one more application, different from Markov chains.

Consider a chess tournament with n players. How can we rate the players based on the results of the tournament? It is assumed, as this is usually done, that each plays one game with each other. The usual way to record the result of one game is +1 is you win, −1 is you loose, and 0 if it is a draw. But we want a non-negative matrix, so let us award 3 for winning, 2 for the draw and 1 for loosing a game. So the outcome can be represented as an

$n \times n$ matrix $A = (a_{i,j})$ where $a_{i,j}$ is the score of the player i in the game with players j. Each entry is $1, 2$ or 3.

The simplest way to rate the players is just to add their scores. We obtain a vector, let us call it $\mathbf{a}_1$ whose i-th entry is the total score of player i. Then one can write

$$\mathbf{a}_1 = A\mathbf{a}_0, \quad \text{where} \quad \mathbf{a}_0 = (1, 1, \ldots, 1)^T.$$

But this is not a very good rating: winning over a strong player should give you more credit that winning over a weak one. Since $\mathbf{a}_1$ nevertheless somehow reflects the strength of a player, we can introduce a refined rating:

$$\mathbf{a}_2 = A\mathbf{a}_1.$$

But why should one stop at this? We can define

$$\mathbf{a}_{m+1} = A\mathbf{a}_m, \quad m = 0, 1, 2, 3, \ldots.$$

So the "ultimate" rating will reasonably be some kind of limit of $\mathbf{a}_m$ as $m \to \infty$. Of course there is no reason to expect that the Perron eigenvalue will be 1 in this case, since our matrix is not stochastic (it just has non-negative entries). Since we are interested in a relative rating, multiplying a vector $\mathbf{a}_m$ by a constant does not change anything. So to obtain the ultimate rating we can consider the limit

$$\lim_{m \to \infty} \mathbf{a}_m / \lambda_1^m.$$

Perron's theorem implies that this limit exists, and is equal to Perron's eigenvector times a constant.

Thus the ultimate rating is done by Perron's eigenvector.

Interestingly, this is how Google rates web sites. In the Google case, the entries $a_{i,j}$ of matrix A are numbers of references on page i in pages j, and the obtained rating determines which pages appear first when you search something on the Google. For more detail on this you can look at the paper of T. Moh, "25 billion dollars eigenvector" posted on the course site. The proof of Perron's theorem is somewhat technical, and I do not give it here. You can read it in the post "Matrices with positive entries" or in the paper of Moh.