Math 1b TA: Padraic Bartlett

Recitation 9: Probability Matrices and Real Symmetric Matrices

Week 9 *Caltech 2011*

1 Random Question

Show that

$$1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\ddots}}}}=\varphi, \text{ the golden ratio, which is } =\frac{1+\sqrt{5}}{2}.$$

2 Homework comments

- Section average: $72/80 \cong 90\%$. This was about $2-3\%$ above the class average; nice work!
- As the above comment suggests, people did pretty well! There isn't much to say, other than "please include explanations of all the things you do."

Today, our talk is pretty much split into two topics: **probability matrices** and **real symmetric matrices**. For each subject, we'll include examples of both the calculations you'll be expected to do on the HW, as well as proofs that you might be asked to understand or replicate.

3 Probability Matrices: Definitions and Examples

Definition. A $n \times n$ matrix P is called a **probability matrix** if and only if the following two properties are satisfied:

- $P \geq 0$; in other words, $p_{ij} \geq 0$ for every entry p_{ij} of P.
- The column sums of P are all 1; in other words, $\sum_{i=1}^{n} p_{ij} = 1$, for every j.

Why do we call these kinds of matrices **probabilitiy matrices**? Well, consider the following way of interpreting such a matrix P:

- Suppose that you have an object that is capable of being in n different states. For example, if the object you're modeling is a Caltech undergrad, you could roughly model them as having the following three distinct states:

$$\{\text{sleeping}, \text{ doing sets}, \text{ eating}\}$$

- Suppose furthermore that at the end of every hour (or more generally, at the end of every **step**, where your step size is some arbitrary unit), the undergrad you're studying has certain **probabilities** of switching from the state they're in into a different state. I.e. you could assume that if your undergrad is working on a set for a hour, there's a decent chance (30%) that they pass out because it's difficult, a better chance (60%) that they keep working on the set, and a small chance (10%) that they get hungry and stop for a sandwich.

- If you have this information for all of your states, you can create a probability matrix with this data! To do this, let p_{ij} contain the probability of switching from state j to state i. Then, we have a matrix all of whose columns sum to 1 (this is because your undergrad is always in *one* of these three states, and thus when they leave any state j the sum of their chances of landing in the other states must be 1.) In other words: a probability matrix!

As it turns out, this method of describing a finite-state system isn't just notationally useful: we in fact have a pair of remarkably useful theorems for probability matrices!

4 Probability Matrices: Two Useful Theorems

Theorem 1 *If we have a probability matrix P representing some finite system with n states $\{1, \dots n\}$, then the probability of starting in state j and ending in state i in precisely m steps is the (i, j)-th entry in P^m.*

Proof. Last recitation, we proved that for a $n \times n$ matrix P, the (i, j)-th entry of P^m can be written as the sum

$$\sum_{c_1, \dots c_{m-1}=1}^{n} p_{i,c_1} \cdot p_{c_1,c_2}, \cdot \ldots \cdot p_{c_{m-2},c_{m-1}} \cdot p_{c_{m-1},j}$$

But what *is* one of these terms $p_{i,c_1} \cdot p_{c_1,c_2}, \cdot \ldots \cdot p_{c_{m-2},c_{m-1}} \cdot p_{c_{m-1},j}$? Well, if we read it from right to left, it's just the product

(probability of going from j to c_{m-1}) $\cdot$ (probability of going from c_{m-1} to c_{m-2})$\cdot$ ...
$\cdot$(probability of going from c_1 to i).

If we just multiply everything out, we can see that this is just the probability that we take the *specific path* $j \to c_{m-1} \to \ldots \to i$ from j to i! And our sum has one term for every such path from i to j: therefore, because our sum is just the sum of all of these probabilities, it represents the chances of us taking **any** of these paths from i to j. Therefore, the (i, j)-th entry represents the probability of us going from j to i in m steps along any of these paths – which is what we claimed.

An example of this theorem follows below:

Example. Suppose we have a Caltech student capable of entering three possible states $\{\text{sleeping}_1, \text{sets}_2, \text{eating}_3\}$, and suppose it switches between these three states every hour according to the following probability matrix:

$$P = \begin{pmatrix} .4 & .3 & 0 \\ .5 & .7 & .7 \\ .1 & 0 & .3 \end{pmatrix}$$

(I.e. if our student is asleep at noon, it has a 40% chance of staying asleep at 1, a 50% chance of waking up and starting a set, and a 10% chance of deciding it's hungry for a kebab.)

If your student is asleep at 2am, what are the chances that it will be working on a set at 4am?

Solution. So: by the theorem above, we just need to look at the $(2, 1)$ cell in P^2 to figure this out. Calculating gives us that

$$P^2 = \begin{pmatrix} .31 & .33 & 21 \\ .62 & .64 & .7 \\ .07 & .03 & .09 \end{pmatrix}$$

and thus that our poor student has a 62% chance of working on their sets at 4am.

This theorem allows us to say what happens at specific points and times in the future. However: what if we're not interested in specific points and times in the future, but rather in the long-term behavior of the system as a whole? The following two definitions and pair of theorems tell us what to do in that situation:

Definition. A vector $\mathbf{v} \in \mathbb{R}^n$ is called a **probability vector** if and only if $\sum_{i=1}^{n} v_i = 1$.

Definition. For a probability matrix P, $\mathbf{v}$ is called a **stable vector** if and only if $\mathbf{v}$ is a probability vector that is also an eigenvector of P with corresponding eigenvalue 1.

Theorem 2 *Every probability matrix has at least one stable vector.*

Theorem 3 *If P is a probability matrix such that $P^m > 0$ for some m – i.e. $p_{ij} > 0$ for every entry in P^m, for some m – then P has exactly **one** stable vector, $\mathbf{x}$.*

*Furthermore, this stable vector is an **attractor** for P: in other words, if $\mathbf{v}$ is any probability vector, we have that $\lim_{n\to\infty} P^n \cdot \mathbf{v} = \mathbf{x}$. Basically, this means that if we run the system represented by P for long enough, the odds of us being in any one of our n states are given by the entries of $\mathbf{x}$.*

The proof of this theorem is in the lecture notes and is kinda tricky, so we've omitted it here because of time constraints. Let me know if you have questions about it, though!

Instead, we offer an example, to illustrate what's going on here:

Example. If we take our Caltech student from earlier, in the long run, what are they more likely to be doing – sets or sleeping?

Solution. So, as we noted before, our probability matrix is

$$P = \begin{pmatrix} .4 & .3 & 0 \\ .5 & .7 & .7 \\ .1 & 0 & .3 \end{pmatrix}.$$

In our earlier example, we showed that $P^2 > 0$; therefore, we know that there is exactly one stable vector, and this stable vector tells us precisely what our student is likely to be doing in the long run. So: let's find it!

Recall that a stable vector is simply an eigenvector for 1 that happens to also be a probability vector. So, to find our stable vector $\mathbf{x}$, we simply need to find E_1:

$$\begin{aligned} E_1 &= \text{nullspace}\,(A - I) \\ &= \text{nullspace} \begin{pmatrix} -.6 & .3 & 0 \\ .5 & -.3 & .7 \\ .1 & 0 & -.7 \end{pmatrix} \\ &= \text{ all } (x, y, z) \text{ such that } (A - I) \\ &= \text{nullspace} \begin{pmatrix} -.6 & .3 & 0 \\ .5 & -.3 & .7 \\ .1 & 0 & -.7 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = (0,0,0); \text{i.e.} \\ 0 &= 3y - 6x, \\ 0 &= 5x - 3y + 7z, \text{ and} \\ 0 &= x - 7z. \end{aligned}$$

Solving the above equations for x, y and z tells us that E_1 is made up out of vectors of the form $c \cdot (7, 14, 1)$. Thus, if we want a vector of this form to have its entries sum up to 1, we simply let $c = 1/22$ and get that our unique stable vector is

$$\left(\frac{7}{22}, \frac{14}{22}, \frac{1}{22} \right).$$

Consequently, we can say that in the long run, our student is twice as likely to be studying (14/22) as it is to be sleeping (7/22).

This is most of what we know about probability matrices! We now change gears here somewhat abruptly, to begin our discussion of real-valued symmetric matrices:

5 Real Symmetric Matrices: A Theorem and its Use

First, as a reminder:

Definition. A $n \times n$ matrix A is called symmetric iff $a_{ij} = a_{ji}$, for every i and j; equivalently, A is called symmetric iff $A = A^T$.

In lecture, we discussed the following remarkable theorem about real-valued symmetric matrices:

Theorem 4 *If A is a real-valued symmetric matrix, then*

- *A has only real-valued eigenvalues,*
- *A is diagonalizable, and furthermore*
- *A is diagonalizable by an* ***orthogonal matrix***[1]*: in other words, there is an orthogonal matrix E made of eigenvectors and diagonal matrix D made out of eigenvalues such that $A = EDE^T$.*

As the proof was discussed in class and is kinda ponderous, we omit it here in favor of discussing two things: (1) how to use the above theorem, and (2) an example of the kinds of proofs you might be asked to do on HW/the final involving symmetric matrices.

First: how *do* we use this theorem? Well, as it turns out, you can pretty much follow the exact same process we used for diagonalizing matrices in general. Suppose that we have a real symmetric matrix A; then, if we want to diagonalize it, we need to simply do the following:

1. First, find all of A's eigenvalues $\lambda_1 \ldots \lambda_k$.
2. Once you've done that, find each eigenspace E_{λ_i}.
3. For each eigenspace E_{λ_i}, find an **orthonormal basis** for E_{λ_i}. This is the **only difference** between this process and normal diagonalization – the bit about making sure that your basis is **orthogonal**. (To do this, simply use Gram-Schmidt on a normal basis for E_{λ_i}, and then normalize all of your vectors by dividing by their length – if you've forgotten how to do Gram-Schmidt, consult week 4's notes, or contact me!)
4. Take all of the vectors you got from these orthogonal bases, and use all of them as the columns in a matrix E; then, take their corresponding eigenvalues, and put them in the diagonal entries in some diagonal matrix D. Then, $A = EDE^T$! and you're done.

Again, for emphasis: the *only* difference between this and normal diagonalization is the part about finding **orthogonal** bases for the E_{λ_i}'s.

We include an example of this algorithm here:

Example. Diagonalize the matrix

$$A = \begin{pmatrix} 5 & 1 & 0 \\ 1 & 5 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

with orthogonal matrices: i.e. find an orthogonal matrix E and diagonal matrix D such that $A = EDE^T$.

[1] A is called an orthogonal matrix iff $A^{-1} = A^T$: i.e. iff $A^2 = I$.

Solution. We follow our blueprint above.

First, we find A's eigenvalues, by calculating its characteristic polynomial:

$$\begin{aligned}
\det(\lambda I - A) &= \det \begin{pmatrix} \lambda - 5 & -1 & 0 \\ -1 & \lambda - 5 & 0 \\ 0 & 0 & \lambda - 6 \end{pmatrix} \\
&= (\lambda - 5) \cdot \det \begin{pmatrix} \lambda - 5 & 0 \\ 0 & \lambda - 6 \end{pmatrix} - (-1) \cdot \det \begin{pmatrix} -1 & 0 \\ 0 & \lambda - 6 \end{pmatrix} \\
&= (\lambda - 5)^2(\lambda - 6) - (\lambda - 6) \\
&= ((\lambda - 5)^2 - 1)(\lambda - 6) \\
&= (\lambda^2 - 10\lambda + 25 - 1)(\lambda - 6) \\
&= (\lambda - 6)(\lambda - 4)(\lambda - 6)
\end{aligned}$$

Therefore, 6 and 4 are our eigenvalues. We now find their eigenspaces:

$$\begin{aligned}
E_4 &= \text{nullspace}(A - 4I) \\
&= \text{nullspace} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix},
\end{aligned}$$

which is spanned by the vector $(-1, 1, 0)$. Normalizing gives us the vector $(-1/\sqrt{2}, 1/\sqrt{2}, 0)$.

$$\begin{aligned}
E_6 &= \text{nullspace}(A - 6I) \\
&= \text{nullspace} \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix},
\end{aligned}$$

which consists of all vectors (x, y, z) such that $x = y$. One such vector is $(1, 1, 1)$; by using Gram-Schmidt or just being clever, another such vector that's orthogonal to $(1, 1, 1)$ is $(1, 1, -2)$. Normalizing gives us $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$ and $(1/\sqrt{6}, 1/\sqrt{}, -2/\sqrt{6})$ as our vectors.

Now that we've found all of our vectors, we use them as the columns of the matrix $E = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ 0 & 1/\sqrt{3} & -2/\sqrt{6} \end{pmatrix}$. Then, if we let $D = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}$ be the diagonal matrix corresponding to the appropriate eigenvalues of our matrix, by our theorem, we know that

$$A = EDE^T$$

as claimed.

6 Real Symmetric Matrices: A Sample Proof

Finally, we have an example that illustrates the kind of proofs we may ask you to do in your HW, or on the final, involving symmetric matrices:

Proposition 5 *A matrix A is symmetric if and only if $\langle \mathbf{x}, A\mathbf{y} \rangle = \langle A\mathbf{x}, \mathbf{y} \rangle$ for every pair of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

Proof. As this is an "if and only if" proof, we must prove both directions of our claim.

We start by assuming that A is symmetric. Then, if we look at $\langle A\mathbf{x}, \mathbf{y} \rangle$ for any pair of vectors $\mathbf{x}, \mathbf{y}$, we have (by definition!) that

$$\langle A\mathbf{x}, \mathbf{y} \rangle = (A\mathbf{x})^T \cdot \mathbf{y} = x^T \cdot A^T \cdot \mathbf{y} = \mathbf{x}^T \cdot A \cdot \mathbf{y} = \mathbf{x}^T \cdot (A\mathbf{y}) = \langle \mathbf{x}, A\mathbf{y} \rangle,$$

which is exactly what we claimed.

Conversely, assume that $\langle \mathbf{x}, A\mathbf{y} \rangle = \langle A\mathbf{x}, \mathbf{y} \rangle$ for every pair of vectors $\mathbf{x}, \mathbf{y}$. It's not entirely clear how to proceed from here: so, let's try just exploring and seeing what this property gives us. In particular, let's see what this property tells us about A when we let $\mathbf{x}, \mathbf{y}$ be the standard basis vectors of $\mathbb{R}^n$: i.e. $\mathbf{x} = \mathbf{e}_i =$ the vector that has a 1 in its i-th spot and zeroes elsewhere, and $\mathbf{y} = \mathbf{e}_j =$ the vector that has a 1 in its j-th spot and zeroes elsewhere. (In general, if you have a vector property and don't understand it, try to find out what it means when you plug in things like the standard basis vectors, the all-1's vector, and any simple examples you can think of. This will often work!)

So: with $\mathbf{x}, \mathbf{y}$ defined as above, we have that

$$\begin{aligned}
A \cdot \mathbf{x} &= \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \ddots & \dots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \cdot \mathbf{e}_i \\
&= \text{the } i\text{-th column of A} \\
\Rightarrow \langle A\mathbf{x}, \mathbf{y} \rangle &= (\text{the } i\text{-th column of A}) \cdot \mathbf{e}_j \\
&= a_{ji}, \text{ and} \\
A \cdot \mathbf{y} &= \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \ddots & \dots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \cdot \mathbf{e}_j \\
&= \text{the } k\text{-th column of A} \\
\Rightarrow \langle \mathbf{x}, A\mathbf{y} \rangle &= \mathbf{e}_i \cdot (\text{the } i\text{-th column of A}) \\
&= a_{ij}
\end{aligned}$$

But if these two inner products are equal, we've just shown that $a_{ij} = a_{ji}$, for every i and j! In other words, A must be symmetric, as claimed.