Markov chains in a shoebox

Peter G. Doyle Gregory Leibon Jean Steiner

Version 4A5 dated 30 April 2010
GNU FDL*

Abstract

Fundamental theorems and formulas about ergodic Markov chains can be understood in a uniform way via the theory of bilinear forms on the hyperplane $H = \{x : \sum_i x_i = 0\}$ and generalized inverses of the matrices that define them.

1 Introduction

We are going to show how to understand fundamental facts and formulas about ergodic Markov chains in a uniform way via the theory of bilinear forms on the hyperplane $H = \{x : \sum_i x_i = 0\}$. We'll begin with the Kemeny-Snell theorem, which states that an ergodic chain is uniquely determined by its matrix of hitting times. In the course of this discussion we'll naturally encounter Kemeny's constant; Tetali's generalization of Foster's theorem; and the reversibility criterion of Coppersmith, Tetali, and Winkler. From there we'll consider general properties of bilinear forms on H; Green's functions and hitting times (including how to compute them); conformal variation of Markov chains; commute times and the commute time embedding; the relationship to the spectral embedding; the problem of 'reversifying' a non-reversible ergodic chain. We'll take some time out to explain the probabilistic intuition behind some of the key formulas. And at the end, in an appendix, we'll show how tensor notation (with upper and lower indices) can be used to advantage in Markov chain theory.

2 The Kemeny-Snell theorem

Let P be the n-by-n transition matrix of an ergodic Markov chain. Let π be its equilibrium vector:

$$\pi^{\mathsf{t}} P = \pi^{\mathsf{t}}; \quad \pi^{\mathsf{t}} \mathbf{1} = \mathbf{1}.$$

Here π^{t} denotes the incarnation of π as a row vector, and $\mathbf{1} = e_1 + \ldots + e_n$ denotes the vector of all 1's, with $e_1, \ldots, e_n$ being the standard basis vectors of $\mathbf{R}^n$. In general, we'll be treating all vectors as column vectors by default. This convention has its drawbacks: π, for example, would naturally be a row vector. The advantage of this convention is that it makes it much easier to tell which vectors in an expression are row vectors and which are column vectors.

Let M be the matrix of expected hitting times, so that M_{ij} is the expected time to reach j starting from i, with the convention that $M_{ii} = 0$.

Theorem 1 (Kemeny and Snell [**?**]) *From M we can recover P.*

Note that M has $n^2 - n$ degrees of freedom, which matches the degrees of freedom of the stochastic matrix P. If instead of M we took the matrix $\bar{M}$ agreeing with M off the diagonal, but with diagonal entries $\bar{M}_{ii} = 1/\pi_i$ telling the expected time to *return* to i starting from i, we could very quickly determine P. Likewise if we constrained the diagonal entries of P to be 0. Neither of these weaker results should be confused with the Kemeny-Snell theorem.

We will be giving the proof of this theorem presently. Beforehand, we will introduce important concepts and notation.

Define the *Kirchhoff matrix* K, with

$$K_{ij} = \pi_i (I - P)_{ij}.$$

We have

$$K\mathbf{1} = \mathbf{0}; \quad \mathbf{1}^{\mathsf{t}} K = \mathbf{0}^{\mathsf{t}}.$$

Let

$$H = \{x : \sum_i x_i = 0\}.$$

We call H the *obvious hyperplane*, ignoring the perhaps even better claim to this title of the parallel hyperplane $\{x : \sum_i x_i = 1\}$. K maps H to itself, and when restricted to H, the map is invertible: This is a consequence of our assumption that the chain is ergodic. So K has a generalized inverse

$$\Omega = (K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}})^{-1} - \mathbf{1}\mathbf{1}^{\mathbf{t}}.$$

K is, in turn, a generalized inverse of Ω:

$$K = (\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}})^{-1} - \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}}.$$

Proposition 2 *Ω is uniquely characterized by the properties*

$$K\Omega K = K; \quad \Omega\mathbf{1} = \mathbf{0}; \quad \mathbf{1}^{\mathbf{t}}\Omega = \mathbf{0}^{\mathbf{t}}.$$

Proof. Standard and straight-forward, but let's do it anyway, for practice with the notation.

First let's verify the normalization. $K+\frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}}$ maps $\mathbf{1}$ to $\frac{1}{n}\mathbf{1}$, so its inverse $\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}}$ maps $\mathbf{1}$ to $n\mathbf{1}$. Thus

$$\Omega\mathbf{1} = (\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}})\mathbf{1} - n\mathbf{1} = n\mathbf{1} - n\mathbf{1} = \mathbf{0}.$$

Similarly, $\mathbf{1}^{\mathbf{t}}\Omega = \mathbf{0}^{\mathbf{t}}$.

Now

$$K = K(\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}})(K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}}) = K\Omega K$$

because both $\mathbf{1}$ and $\mathbf{1}^{\mathbf{t}}$ kill both K and Ω.

As for the uniqueness, if Ω satisfies these properties then

$$(K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}})(\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}})(K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}}) = K\Omega K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}} = K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}}$$

and since the right side is invertible we have

$$(K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathbf{t}})(\Omega + \mathbf{1}\mathbf{1}^{\mathbf{t}}) = I$$

so

$$\Omega = (K + \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathrm{t}})^{-1} - \mathbf{1}\mathbf{1}^{\mathrm{t}}. \quad \blacksquare$$

Note. Ω also satisfies

$$\Omega K \Omega = \Omega,$$

but we prefer not to emphasize this because we will be considering other generalized inverses to K that do not have this property.

Proposition 3

$$KM = -I + \pi \mathbf{1}^{\mathrm{t}}.$$

Proof. A standard probabilistic argument about the effect of taking one step away from i on the expected time to reach j yields:

$$1 + \sum_k P_{ik} M_{kj} = M_{ij} \quad (i \neq j);$$

$$\sum_k P_{ik} M_{ki} = \frac{1}{\pi_i},$$

where in the second equation we use the fact that, starting from i, the expected time to *return* to i is $\frac{1}{\pi_i}$. Combining and rearranging these equations gives

$$[(I - P)M]_{ij} = -\frac{1}{\pi_i} I_{ij} + 1.$$

Multiplying by π_i gives

$$(KM)_{ij} = -I_{ij} + \pi_i = (-I + \pi \mathbf{1}^{\mathrm{t}})_{ij}. \quad \blacksquare$$

Corollary 4 (Kemeny's constant)

$$M\pi = c\mathbf{1}$$

for some constant c. π is the only probability vector with this property.

Proof.

$$KM\pi = (-I + \pi \mathbf{1}^{\mathrm{t}})\pi = -\pi + \pi = \mathbf{0}.$$

But the constant vectors are the only elements of the kernel of K (on either side). $\blacksquare$

Corollary 5 (Tetali's generalization of Foster's theorem)

$$\sum_{ij} \pi_i P_{ij} M_{ij} = n - 1.$$

Proof. Since the diagonal hitting times M_{ii} are 0, this is equivalent to

$$\mathrm{Tr}(KM) = n - 1,$$

which follows immediately from the proposition. ■

Note. According to Tetali [**?**], this generalization of Foster's theorem was implicit in the article [**?**], of Coppersmith et al., unbeknownst to the authors. Hopefully by the time we're done, this fact will seem so obvious that not even Coppersmith et al. could overlook it.

Proof of Theorem 1. Set

$$\Omega = -(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{t}})M(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{t}})$$

and observe that the name Ω is justified because this matrix satisfies the characterization of Ω above, namely

$$K\Omega K = K; \quad \Omega\mathbf{1} = \mathbf{0}; \quad \mathbf{1}^{\mathrm{t}}\Omega = \mathbf{0}^{\mathrm{t}}.$$

Ω determines K as above, so, starting from M, we get Ω, hence K. Once we have K we can get π from Proposition 3 above:

$$\pi\mathbf{1}^{\mathrm{t}} = KM + I.$$

Alternatively, we can get π directly from the fact that it is the unique probability vector for which $M\pi = c\mathbf{1}$ for some c. Either way, once we have K and π we get P. ■

3 The Green's function

The Kirchhoff matrix K has other generalized inverses, in addition to the standard one that we're calling Ω. An important one is the *Green's function*

$$G = (K + \pi\pi^{\mathrm{t}})^{-1} - \mathbf{1}\mathbf{1}^{\mathrm{t}},$$

which is uniquely characterized by

$$KGK = K; \quad G\pi = \mathbf{0}; \quad \pi^{\mathrm{t}}G = \mathbf{0}^{\mathrm{t}}.$$

G is just Ω renormalized to favor π instead of $\mathbf{1}$. From G it's easy to get Ω:

$$\Omega = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{t}})G(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{t}}).$$

And given π, we get back to G in a similar way:

$$G = (I - \pi\pi^{\mathsf{t}})\Omega(I - \pi\pi^{\mathsf{t}}).$$

In its dependence on both K and π, G is like M, and in fact it's easy to convert directly between them:

$$M_{ij} = G_{ij} - G_{jj};$$

$$G = -(I - \pi\pi^{\mathsf{t}})M(I - \pi\pi^{\mathsf{t}}).$$

Stop! It's high time to take a higher-level approach to this business, and explain where all these similar-looking formulas are coming from. What it comes down to is that Ω, G, and $-M$ all induce the same bilinear form on H, namely, the form inverse to that induced by the Kirchhoff matrix K. Where Ω depends only on K, G and M depend on K and π, and determine both. We can think of G and M as squirreling away the information about π where it won't affect the form induced on H. Once we understand such forms, and the various ways of normalizing the matrices that determine them, all this will be crystal clear.

4 Forms on the obvious hyperplane

An n-by-n matrix A determines a bilinear form L_A on $\mathbf{R}^n$:

$$L_A(x, y) = x^{\mathsf{t}}Ay.$$

Denote by $L_A|H$ the restriction of the bilinear form L_A to the obvious hyperplane

$$H = \{x \in \mathbf{R}^n : \sum_i x_i = 0\} = \{x \in \mathbf{R}^n : x^{\mathsf{t}}\mathbf{1} = 0\} = \{y \in \mathbf{R}^n : \mathbf{1}^{\mathsf{t}}y = 0\}$$

Proposition 6

$$L_A|H = 0 \iff \exists a, b : A_{ij} = a_i + b_j \iff \exists a, b : A = a\mathbf{1}^{\mathsf{t}} + \mathbf{1}b^{\mathsf{t}}.$$

So A and B determine the same bilinear form on H just if they differ by something that depends only on the column plus something that depends only on the row.

Proof. Straightforward, but let's do it anyway.

We just need to characterize matrices A for which $L_A|H = 0$. If $A = a\mathbf{1}^\mathsf{t} + \mathbf{1}b^\mathsf{t}$ and $x, y \in H$,

$$L_A(x,y) = x^\mathsf{t}Ay = x^\mathsf{t}a(\mathbf{1}^\mathsf{t}y) + (x^\mathsf{t}\mathbf{1})b^\mathsf{t}y = 0.$$

Conversely, if $L_A|H = 0$ then

$$(e_i - e_n)^\mathsf{t}A(e_j - e_n) = A_{ij} - A_{in} - A_{nj} + A_{nn} = 0$$

so

$$A_{ij} = A_{in} + A_{nj} - A_{nn}. \quad \blacksquare$$

There are various ways to normalize the matrix A of a bilinear form on H. For example, we may specify $\alpha, \beta \notin H$ to lie in the kernel of L_A on the left and right, respectively. This nails down the matrix. Assuming without loss of generality that

$$\alpha^\mathsf{t}\mathbf{1} = \mathbf{1}^\mathsf{t}\beta = 1,$$

set

$$\hat{A} = (I - \mathbf{1}\alpha^\mathsf{t})A(I - \beta\mathbf{1}^\mathsf{t}),$$

i.e.

$$\hat{A}_{ij} = A_{ij} - (A\beta)_i - (\alpha^\mathsf{t}A)_j + \alpha^\mathsf{t}A\beta,$$

to get

$$\alpha^\mathsf{t}\hat{A} = \alpha^\mathsf{t}(I - \mathbf{1}\alpha^\mathsf{t})A(I - \beta\mathbf{1}^\mathsf{t}) = (\alpha^\mathsf{t} - \alpha^\mathsf{t})A(I - \beta\mathbf{1}^\mathsf{t}) = \mathbf{0}^\mathsf{t},$$

and similarly

$$\hat{A}\beta = \mathbf{0}.$$

Taking $\alpha = \beta = \frac{1}{n}(1, \ldots, 1)$, $\hat{A}$ will have row and column sums 0. Taking $\alpha = \beta = e_n$, we get

$$\hat{A} = A_{ij} - A_{in} - A_{nj} + A_{nn}$$

with vanishing nth row and column.

Let's have a notation for our normalized representative matrix:

$$\mathrm{Gnd}(A, \alpha, \beta) = (I - \frac{1}{\alpha^\mathsf{t}\mathbf{1}}\mathbf{1}\alpha^\mathsf{t})A(I - \frac{1}{\mathbf{1}^\mathsf{t}\beta}\beta\mathbf{1}^\mathsf{t}),$$

where now we aren't assuming that α and β are probability vectors. This will allow us to write

$$A = \mathrm{Gnd}(A, \alpha, \beta)$$

to indicate that A is normalized so that

$$\alpha^{\mathsf{t}} A = \mathbf{0}^{\mathsf{t}}; \quad A\beta = \mathbf{0}.$$

Summing up the above discussion and notation, we have:

Proposition 7 *The following are equivalent:*

- $L_A|H = L_B|H$.
- $\forall \alpha, \beta \notin H$, $\mathrm{Gnd}(A, \alpha, \beta) = \mathrm{Gnd}(B, \alpha, \beta)$.
- $\exists \alpha, \beta \notin H$, $\mathrm{Gnd}(A, \alpha, \beta) = \mathrm{Gnd}(B, \alpha, \beta)$.
- $\mathrm{Gnd}(A, \mathbf{1}, \mathbf{1}) = \mathrm{Gnd}(B, \mathbf{1}, \mathbf{1})$.
- $\mathrm{Gnd}(A, e_n, e_n) = \mathrm{Gnd}(B, e_n, e_n)$.
- $\forall i, j$, $A_{ij} - A_{in} - A_{nj} + A_{nn} = B_{ij} - B_{in} - B_{nj} + B_{nn}$. ■

The point here is that the bilinear form $L_A|H$ on H determines a unique normalization of A for any α, β, so two matrices determine the same form on H just if they agree when normalized in the same way.

As an application of this, we have the following equivalent conditions for a form on H to be symmetric.

Proposition 8 *The following are equivalent:*

- $L_A|H$ *is symmetric.*
- $L_A|H = L_{A^{\mathsf{t}}}|H$.
- $\forall \alpha \notin H$, $\mathrm{Gnd}(A, \alpha, \alpha)$ *is symmetric.*
- $\exists \alpha \notin H$, $\mathrm{Gnd}(A, \alpha, \alpha)$ *is symmetric.*
- $\mathrm{Gnd}(A, \mathbf{1}, \mathbf{1})$ *is symmetric.*
- $\mathrm{Gnd}(A, e_n, e_n)$ *is symmetric.*
- $\forall i, j$, $A_{ij} - A_{in} - A_{nj} + A_{nn} = A_{ji} - A_{jn} - A_{ni} + A_{nn}$.
- $\forall i, j$, $A_{ij} + A_{jn} + A_{ni} = A_{in} + A_{nj} + A_{ji}$.

- $\forall i, j, k,\ A_{ij} + A_{jk} + A_{ki} = A_{ik} + A_{kj} + A_{ji}$.
- $\forall i_1, i_2, \ldots, i_m, i_{m+1} = i_1,\ A_{i_1 i_2} + \ldots + A_{i_m i_{m+1}} = A_{i_{m+1} i_m} + \ldots + A_{i_2 i_1}$. ∎

Here are conditions for symmetry phrased in terms of conditions for vanishing of the form $L_{\bar{A}}|H$ associated to an antisymmetric matrix $\bar{A}$. All except the last follow immediately from the previous proposition. We're phrasing them separately here because we want to emphasize that they can be proven immediately, from scratch.

Proposition 9 *Let $\bar{A}$ be anti-symmetric. The following are equivalent:*

- $L_{\bar{A}}|H = 0$.
- $\forall i, j,\ \bar{A}_{ij} + \bar{A}_{jn} + \bar{A}_{ni} = 0$.
- $\forall i, j, k,\ \bar{A}_{ij} + \bar{A}_{jk} + \bar{A}_{ki} = 0$.
- $\forall i_1, i_2, \ldots, i_m, i_{m+1} = i_1,\ \bar{A}_{i_1 i_2} + \ldots + \bar{A}_{i_m i_{m+1}} = 0$.
- $\exists a_1, \ldots, a_n,\ \bar{A}_{ij} = a_i - a_j$. ∎

Proof. As observed above, the equivalence of these conditions is immediate. What we're seeing here is a discrete version of the familiar facts from vector analysis that the curl of the gradient of a function vanishes, and if a vector field has vanishing curl, then it is the gradient of some function. ∎

By now we've beaten to death the following observation of Coppersmith, Tetali, and Winkler [**?**]:

Corollary 10 *A Markov chain is reversible just if the expected time to traverse any cycle of states is the same in either direction. It suffices to check only 3-cycles involving the fixed state n, or any other generating set for the space of cycles.*

Proof. This is an immediate consequence of the previous immediate proposition. ∎

The lesson here is that thinking about bilinear forms on H has straightened out our thinking about conditions for reversibility. In the end, we don't need them to prove this criterion. But that doesn't diminish their usefulness.

One more time

Here's another way to look at what has been going on here. A bilinear form on a vector space V corresponds to a linear form on the tensor product $V \otimes V$. We may identify the tensor product $\mathbf{R}^n \otimes \mathbf{R}^n$ with the space of n-by-n matrices. The tensor product $H \otimes H$ then corresponds to the subspace consisting of matrices with row and column sums 0. This space is spanned by matrices of the form $e_{ac} - e_{ad} - e_{bc} + e_{bd}$. The values of the linear form on these matrices is just what you get from the bilinear form by putting $e_a - e_b$ on the left and $e_c - e_d$ on the right. A bilinear form is symmetric just if it vanishes for everything in the $H \wedge H$, which is the space of antisymmetric matrices whose row sums (and hence also column sums) vanish. Our various criteria for symmetry amount to verifying the vanishing on one spanning set or another of this space.