Dirichlet Forms on REM Chains

Arpon Basu

Indian Institute of Technology, Bombay

April 27, 2023

Introduction

The goals of this presentation are:

- We'll explore the relaxation times of finite space **R**eversible **E**rgodic **M**arkov (REM) chains, both in discrete time and continuous time.
- We introduce the notion of *Dirichlet forms* on Markov chains, which enable us to show that the speed with which a REM chain attains equilibrium is controlled by the second largest eigenvalue ('τ_2') of a particular symmetric matrix.

Notation and Convention

Throughout the report, we shall be working in the setting described below.

REM Chains

Our time homogenous Markov chain shall be represented as $\mathbf{X} := (X_t : t \in \mathbb{T})$, where the state space of $\mathbf{X}$ is denoted by $\mathcal{R}$, **which take to be finite.**

Further, our Markov chain is irreducible, ergodic, and reversible, and we denote the (unique) stationary distribution on $\mathbf{X}$ by π.

We shall denote by $\mathbb{E}_\rho [\cdot]$ the expectation of a random variable constituted by $\{X_t, t \in \mathbb{T}\}$, where the initial distribution, ie:- the distribution of X_0, is taken to be ρ. Further, if our Markov chain begins at some state i, then the corresponding expectation is denoted as $\mathbb{E}_i [\cdot]$. Note that if the distribution of X_0 is π, then our Markov chain is a stationary stochastic process.

For the discrete-time case, we represent the transition function of our Markov chain with $\mathbf{P} = (p_{ij})_{i,j \in \mathcal{R}}$.
The consequence of reversibility, then, is

Detailed Balance Equations

$$\pi_i p_{ij} = \pi_j p_{ji} \; \forall \; i, j \in \mathcal{R}$$

We also define

Hitting Time

$$T_i := \inf\{t \in \mathbb{Z}_{\geq 0} : X_t = i\}$$

Continuous Time Markov Chains

Let $\mathbf{P}$ be the transition matrix of our DTMC. For computational purposes, we work with the *infinitesimal* description of a CTMC. That goes as follows: We associate a transition matrix $\mathbf{Q} := \mathbf{P} - \mathbf{I}$ to our CTMC, which has the interpretation that, for any $i \neq j, i, j \in \mathcal{R}$,

$$q_{ij} = \lim_{h \searrow 0} \frac{\mathbb{P}(X_h = j | X_0 = i)}{h}$$

or equivalently,

$$\mathbb{P}(X_h = j | X_0 = i) = q_{ij} h + o(h)$$

Also, for $i = j$, we set $q_{ii} = -\sum_{j \neq i} q_{ij}$.

It can be shown that $\mathbb{P}(X_t = j | X_0 = i) =: \mathbb{P}_i(X_t = j) = q_{ij}^{(t)}$ is equal to $(\exp(\mathbf{Q}t))_{ij}$.

Continuous Time Chapman Kolmogorov Equations

Let $\{\rho_j(t)\}_{j\in\mathcal{R}}$ denote the probability distribution on $\mathcal{R}$ at time t. Then

$$\mathbb{P}(X_{t+h}=j)=\sum_{i\in\mathcal{R}}\mathbb{P}(X_{t+h}=j|X_t=i)\mathbb{P}(X_t=i)$$

$$=\sum_{i\neq j}(q_{ij}h+o(h))\rho_i(t)+\mathbb{P}(X_{t+h}=j|X_t=j)\mathbb{P}(X_t=j)$$

Consequently,

$$\lim_{h\searrow 0}\frac{\mathbb{P}(X_{t+h}=j)-\mathbb{P}(X_t=j)}{h}$$

$$=\sum_{i\neq j}q_{ij}\rho_i(t)-\rho_j(t)\lim_{h\searrow 0}\frac{\mathbb{P}(X_{t+h}\neq j|X_t=j)}{h}=\sum_{i\neq j}q_{ij}\rho_i(t)-\rho_j(t)\underbrace{\sum_{k\neq j}q_{jk}}_{=-q_{jj}}$$

$$=\sum_{i\in\mathcal{R}}q_{ij}\rho_i(t)$$

But note that

$$\lim_{h \searrow 0} \frac{\mathbb{P}(X_{t+h} = j) - \mathbb{P}(X_t = j)}{h} = \frac{d\rho_j(t)}{dt}$$

Consequently, we arrive at the continuous version of the Chapman-Kolmogorov theorem, namely, for any $j \in \mathcal{R}$,

$$\frac{d\rho_j(t)}{dt} = \sum_{i \in \mathcal{R}} q_{ij}\rho_i(t)$$

An Important Distinction b/w CTMCs and DTMCs

It is very important to mention that although many properties of DTMCs have natural analogs in the continuous-time case, one property where there is a notable difference is *periodicity*: Indeed, note that for CTMCs, for any $i \in \mathcal{R}$, the set $\{t : q_{ii}^{(t)} > 0\}$ is a subset of $\mathbb{R}_{\geq 0}$ in general, and consequently we can't take the greatest common divisor of this set in the usual sense. Thus, the concept of periodicity **is not defined** in the continuous time paradigm.

Spectral Representation of Reversible Markov Chains

We define the quantity

$$s_{ij} := p_{ij}\sqrt{\frac{\pi_i}{\pi_j}}$$

for all $i, j \in \mathcal{R}$. Furthermore, by the detailed balance equations, since we have $\pi_i p_{ij} = \pi_j p_{ji}$, we get that $s_{ij} = s_{ji}$, ie:- the matrix $\mathbf{S} := (s_{ij})_{i,j\in\mathcal{R}}$ is a real symmetric matrix. Consequently, we can diagonalize $\mathbf{S}$ to obtain

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^{\mathsf{T}}$$

where $\mathbf{U}$ is an orthonormal matrix, and Λ is the diagonal matrix containing the eigenvalues of $\mathbf{S}$ in decreasing order, ie:- $\Lambda_{ii} = \lambda_i$, and $\lambda_1 \geq \ldots \geq \lambda_n$.

Finally, a little linear algebra reveals that $\mathbf{S}$ and $\mathbf{P}$ have the same eigenvalues.

Spectral Representation

Since our Markov chain $\mathbf{X}$ is irreducible, the transition matrix $\mathbf{P}$ is a stochastic irreducible matrix. Consequently, by the Perron-Frobenius theorem, we have $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \geq -1$.
Further, note that

$$\mathbf{S}^t = \mathbf{U}\Lambda^t\mathbf{U}^\mathsf{T} \implies s_{ij}^{(t)} = \sum_{k=1}^{n} \lambda_k^t u_{ik} u_{jk}$$

Further, the Chapman-Kolmogorov equations yield $p_{ij}^{(t)} = s_{ij}^{(t)}\sqrt{\frac{\pi_j}{\pi_i}}$.
Thus, we can combine these two equations to yield the spectral representation formula, which goes as

Theorem (Spectral Representation formula)

We have

$$\mathbb{P}(X_t = j | X_0 = i) = \sqrt{\frac{\pi_j}{\pi_i}} \sum_{k=1}^{n} \lambda_k^t u_{ik} u_{jk}$$

Relaxation Time

This result can be easily *continuized*: If $\mathbf{Q}$ is the transition matrix of our CTMC $\mathbf{X}$, then $\mathbf{P} = \mathbf{Q} + \mathbf{I}$ is the transition matrix for some DTMC $\mathbf{Y}$. **Consequently, if the DTMC has an eigenvalue λ, then the corresponding CTMC has the eigenvalue $1 - \lambda$.**
Thus from this point onwards, we'll freely interchange between the spectral representations in the discrete and continuous time cases, such that $\lambda^{(c)} = 1 - \lambda^{(d)}$, where c, d denote continuous and discrete respectively.

Relaxation Time

$$\tau_2 := \begin{cases} 1/\lambda_2, \text{for continuous-time Markov chains} \\ 1/(1-\lambda_2), \text{for discrete-time Markov chains} \end{cases}$$

As we shall see throughout this report, τ_2 controls how fast the Markov chain converges towards its stationary distribution π.

Dirichlet Forms

Let **X** be an ergodic reversible Markov chain on a finite state space $\mathcal{R}$ with stationary distribution π. We define

$$\mathcal{E}(g,g) := \frac{1}{2}\sum_{i,j\in\mathcal{R}} \pi_i p_{ij}(g(j)-g(i))^2$$

for any function $g : \mathcal{R} \mapsto \mathbb{R}$, where we replace p_{ij} by q_{ij} in the continuous time analogue.

Some Basic Lemmata

Lemma

For the discrete-time case, we have

$$\mathcal{E}(g,g) = \frac{1}{2}\mathbb{E}_\pi\left[(g(X_1) - g(X_0))^2\right] = \mathbb{E}_\pi\left[g(X_0)(g(X_0) - g(X_1))\right]$$

Proof.

Note that

$$\mathbb{E}_\pi\left[(g(X_1) - g(X_0))^2\right] = \sum_{i,j\in\mathcal{R}} (g(j) - g(i))^2 \mathbb{P}(X_1 = j, X_0 = i)$$

$$= \sum_{i,j\in\mathcal{R}} (g(j) - g(i))^2 \mathbb{P}(X_1 = j | X_0 = i)\mathbb{P}(X_0 = i) = \sum_{i,j\in\mathcal{R}} (g(j) - g(i))^2 p_{ij}\pi_j$$

□

Proof.

Furthermore,

$$\frac{1}{2}\mathbb{E}_\pi\left[(g(X_1)-g(X_0))^2\right]=\frac{1}{2}\mathbb{E}_\pi\left[g(X_1)^2+g(X_0)^2-2g(X_0)g(X_1)\right]$$

Since we are taking expectation over a stationary distribution, we have $\mathbb{E}\left[g(X_1)^2\right]=\mathbb{E}\left[g(X_0)^2\right]$, yielding

$$\frac{1}{2}\mathbb{E}_\pi\left[g(X_1)^2+g(X_0)^2-2g(X_0)g(X_1)\right]=\mathbb{E}_\pi\left[g(X_0)^2-g(X_0)g(X_1)\right]$$

$\square$

Continuous Time Analogue

Lemma

For the continuous-time case, we have

$$\mathcal{E}(g,g) := -\sum_{i,j\in\mathcal{R}} \pi_i q_{ij} g(i)g(j)$$

Proof.

Note that for any $h > 0$

$$\mathbb{E}_\pi\left[(g(X_h)-g(X_0))^2\right] = \sum_{i,j\in\mathcal{R}} (g(j)-g(i))^2\mathbb{P}(X_h = j, X_0 = i)$$

$$= \sum_{i,j\in\mathcal{R}} (g(j)-g(i))^2\mathbb{P}(X_h = j | X_0 = i)\pi_i$$

□

Proof.

Thus

$$\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}_\pi\left[(g(X_h)-g(X_0))^2\right]=\sum_{i,j\in\mathcal{R}}\pi_i(g(j)-g(i))^2\lim_{h\searrow 0}\frac{\mathbb{P}(X_h=j|X_0=i)}{h}$$

$$=\sum_{i,j\in\mathcal{R}}\pi_i(g(j)-g(i))^2q_{ij}=2\mathcal{E}(g,g)$$

On the other hand,

$$\frac{1}{2}\lim_{h\searrow 0}\frac{\mathbb{E}_\pi\left[(g(X_h)-g(X_0))^2\right]}{h}=-\lim_{h\searrow 0}\frac{\mathbb{E}_\pi\left[g(X_0)(g(X_h)-g(X_0))\right]}{h}$$

as in the discrete-time case. But the above expression equals

$$-\sum_{i,j\in\mathcal{R}}g(i)(g(j)-g(i))\mathbb{P}(X_0=i)\lim_{h\searrow 0}\frac{\mathbb{P}(X_h=j|X_0=i)}{h}$$

Proof.

$$= -\sum_{i,j\in\mathcal{R}} g(i)(g(j)-g(i))\pi_i q_{ij} = -\sum_{i,j\in\mathcal{R}} g(i)g(j)\pi_i q_{ij} + \sum_{i,j\in\mathcal{R}} g(i)^2\pi_i q_{ij}$$

But

$$\sum_{i,j\in\mathcal{R}} g(i)^2\pi_i q_{ij} = \sum_{i\in\mathcal{R}}\left(g(i)^2\pi_i \sum_{j\in\mathcal{R}} q_{ij}\right) = 0$$

as desired. □

A Closer Look at Dirichlet Forms

When we want to quantify *how fast* a changing probability distribution (the canonical example of which is the continuous-time Markov chain) is tending to its stationary distribution, Dirichlet forms arise naturally. But before we can measure how fast our Markov chain is tending to equilibrium, we need to be able to define the distance between two probability distributions.

Statistical Distance between Probability Distributions

Let μ be a probability distribution on $\mathcal{R}$. Then we define

$$\|\mu - \pi\|_2^2 := \left(\sum_{i \in \mathcal{R}} \frac{\mu_i^2}{\pi_i} \right) - 1$$

Theorem

Let $\rho(t)$ be the distribution over $\mathcal{R}$ at time t. Then

$$\frac{d\|\rho(t)-\pi\|_2^2}{dt} = -2\mathcal{E}(f(t), f(t))$$

where $f_j(t) := \rho_j(t)/\pi_j, j \in \mathcal{R}$.

Proof.

Note that

$$\frac{d}{dt}\|\rho(t)-\pi\|_2^2 = \sum_{j\in\mathcal{R}} \frac{1}{\pi_j}\frac{d}{dt}\rho_j(t)^2 = \sum_{j\in\mathcal{R}} \frac{2}{\pi_j}\rho_j(t)\sum_{i\in\mathcal{R}}\rho_i(t)q_{ij}$$

$$= 2\sum_{i,j\in\mathcal{R}} \frac{\rho_j(t)}{\pi_j}\frac{\rho_i(t)}{\pi_i}\pi_i q_{ij} = -2\mathcal{E}(f(t), f(t))$$

□

Extremal Characterization of τ_2

We now define the extremal characterization of relaxation times. Although this might seem disconnected from the previous content, we shall tie everything together soon.

Theorem (Extremal characterization of τ)

$$\tau_2 = \sup_{\substack{g \neq 0 \\ \sum_{i \in \mathcal{R}} g(i)\pi_i = 0}} \frac{\|g\|_2^2}{\mathcal{E}(g,g)}$$

where $\|g\|_2^2$ *is defined to be* $\sum_{i \in \mathcal{R}} \pi_i g(i)^2 = \mathbb{E}_\pi\left[g(X_0)^2\right]$.

Proof.

Let **A** be any real symmetric matrix with eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$.

□

Proof.

Then from the Rayleigh characterization of eigenvalues, we have

$$\mu_2 = \sup_{\substack{x \neq 0 \\ \langle x, v \rangle = 0}} \frac{x^{\mathsf{T}} \mathbf{A} x}{\|x\|_2^2} = \sup_{\substack{x \neq 0 \\ \langle x, v \rangle = 0}} \frac{\sum x_i a_{ij} x_j}{\sum x_i^2}$$

where v is an eigenvector of $\mathbf{A}$ corresponding to μ_1.
We apply this to our spectral matrix $\mathbf{S} := (s_{ij})_{i,j \in \mathcal{R}}$. Note that the eigenvector corresponding to the highest eigenvalue μ_1 was v, where $v_i = \sqrt{\pi_i}$. Then setting $x_i = \sqrt{\pi_i} g(i)$ yields

$$\mu_2 = \sup_{\sum \sqrt{\pi_i} x_i = 0} \frac{\sum x_i s_{ij} x_j}{\sum x_i^2} = \sup_{\sum \pi_i g(i) = 0} \frac{\sum x_i s_{ij} x_j}{\sum \pi_i g(i)^2}$$

$$= \sup_{\sum \pi_i g(i) = 0} \frac{\sum \pi_i p_{ij} g(i) g(j)}{\|g\|_2^2}$$

$\square$

Proof.

$$= \sup_{\sum \pi_i g(i)=0} \frac{\mathbb{E}_\pi [g(X_0)g(X_1)]}{\|g\|_2^2} = \sup_{\sum \pi_i g(i)=0} \frac{\mathbb{E}_\pi \left[g(X_0)^2\right] - \mathcal{E}(g,g)}{\|g\|_2^2}$$

But $\mathbb{E}_\pi \left[g(X_0)^2\right] = \|g\|_2^2$, following which the desired result easily follows. □

An useful Corollary

Corollary

Suppose we collapse a subset $A \subseteq \mathcal{R}$ into a singleton $\{a\}$. Let τ_2^A be the relaxation time of this collapsed chain. Then $\tau_2^A \leq \tau_2$.

Proof.

For any function g on $(\mathcal{R} \setminus A) \cup \{a\}$, we can extend it to a function on $\mathcal{R}$ by setting $g(\alpha) = g(a)$ for every $\alpha \in A$. This doesn't change $\|g\|, \sum \pi_i g(i)$ and $\mathcal{E}(g, g)$, and thus $\tau_2 \geq \tau_2^A$ since for every candidate function in the supremum for τ_2^A, the same value is also attained in the corresponding supremum expression for τ_2. □

Rate of Decay of Markov Chains to Equilibrium

Theorem

Let $\rho(t)$ be the distribution of our continuous-time Markov chain (whose initial distribution is assumed to be arbitrary). Then

$$\|\rho(t) - \pi\|_2 \leq e^{-t/\tau_2}\|\rho(0) - \pi\|_2$$

Proof.

We have

$$\frac{d}{dt}\|\rho(t) - \pi\|_2^2 = -2\mathcal{E}(f(t), f(t))$$

where $f = \rho/\pi$.
But

$$-2\mathcal{E}(f(t), f(t)) = -2\mathcal{E}(f(t) - 1, f(t) - 1) \leq -2\frac{\|f(t) - 1\|_2^2}{\tau_2}$$

□

Proof.

Thus

$$\frac{d}{dt}\|\rho(t)-\pi\|_2^2 \leq -2\frac{\|f(t)-1\|_2^2}{\tau_2} = \frac{-2}{\tau_2}\|\rho(t)-\pi\|_2^2$$

Integration then yields our desired result. □

Some Examples

As promised, we have derived bounds for the rate of convergence of Markov chains to their equilibrium using extremal characterizations and Dirichlet forms.
We now see some examples.

On-off chains

On-off chains

The on-off chain is a Markov chain with two states, say α, β such that $p_{\alpha,\beta} = 1 - p_{\alpha,\alpha} = p \in (0,1)$, and $p_{\beta,\alpha} = 1 - p_{\beta,\beta} = q \in (0,1)$.
The transition matrix is given by $\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$.

- This chain is clearly irreducible, ergodic, and aperiodic. Its stationary distribution is given by π where $\pi_\alpha = \frac{q}{p+q} = 1 - \pi_\beta$.
- π satisfies the detailed balance equations, and consequently, the on-off chain is reversible.

Calculating τ_2 yields $\tau_2 = \frac{1}{p+q}$.

Qualitative Interpretation

Note that if both p, q are close to 0, then τ_2 is very large: Indeed, note that p, q are the probabilities that the walk goes from one state to another state. Consequently, p, q being small means that our on-off chain is "sticky", and it will take more time to mix, which tallies with a larger value of τ_2.

Random Walks on Graphs

A natural source of interesting finite space ergodic reversible chains is random walks on weighted graphs. However, we face a problem when we try to analyze random walks on graphs: Not all graph random walks are aperiodic. Indeed, consider a bipartite graph where our initial distribution is (concentrated on) some particular vertex. Then a random walk on this graph, where from every vertex of our graph we move to its neighbors, is **not ergodic** since we are on different partitions of the bipartite graph on consecutive time instants.
A workaround is to move to continuous time Markov chains on the same graph. As mentioned before, the concept of periodicity doesn't make any sense in the continuous-time régime. Moreover, the spectral properties of the graph carry over to the continuous-time regime without any hassle, and consequently, we can analyze the relaxation times of our graphs without worrying about periodicity.

The Cycle Graph

Consider the cycle graph C_n on n vertices, where at any given point, we move to its left/right neighbor with equal probability. This corresponds to a stochastic matrix **P** of the form

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \dots & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & \dots & 0 \\ 0 & \frac{1}{2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & 0 & 0 & \dots & 0 \end{bmatrix}$$

ie:- $p_{ij} = \frac{1}{2}$ if and only if $i = j \pm 1 \bmod n$. The chain is irreducible. Further, as discussed above, in the continuous time régime, we needn't worry about periodicity.
In this case, the stationary distribution π turns out to be identically equal to $\frac{1}{n}$ for every $i \in \{0, 1, \dots, n-1\}$, which is unsurprising given the regularity of the graph.

Moreover, this π satisfies the detailed balance equations too, so we have reversibility. Also, note that since π is the same for all points in its sample space, $\mathbf{S} = \mathbf{P}$.
Finally, we note that $\mathbf{P}$ is a *circulant* matrix, which has very special spectral properties. Since we don't have to completely work it out, we directly quote

$$\boxed{\tau_2 = \frac{1}{1 - \cos(\frac{2\pi}{n})} \sim \frac{n^2}{2\pi^2}}$$

At a qualitative level, it tells us that the Markov chain needs $\mathcal{O}(n^2)$ time to "mix thoroughly". One can't help but note the similarity with symmetric 1D random walks, where it takes $\mathcal{O}(n^2)$ time (in expectation) to cover a distance of $\mathcal{O}(n)$.

Some Other Special Graphs

We present here the values of τ_2 for random walks on some other special graphs [1].

Graphs	τ_2
n-path	$\sim (2/\pi^2)n^2$
Complete graph on n vertices	$(n-1)/n$
Star with n vertices	1
d-dimensional hypercube	$d/2$

[1] it is not difficult to verify that a random walk on a connected undirected graph is *always* irreducible and reversible

References

The references for this report are [1, Chapter 3], [1, Chapter 5]. A freely available online copy can be found here.

[1] Aldous and Fill. *Reversible Markov Chains and Random Walks on Graphs*. 2014.

Acknowledgements

I thank Prof. Ayan Bhattacharya for suggesting the topic of this presentation to me, and very patiently fielding and clearing all my queries regarding the same.

The End

Questions? Comments?