Mixing of Markov Chains

Arpon Basu

Last updated July 4, 2024

Contents

§1. Convergence of Markov Chains: Classical Techniques

In this chapter, we will see various functional analytic ways of proving the convergence of Markov chains.
Let Ω be the state space of a Markov chain. We will assume Ω to be finite, and write $n := |\Omega|$. Suppose π is a probability distribution on Ω such that $\pi(x) > 0$ for all $x \in \Omega$. Let $P \in [0,1]^{\Omega \times \Omega}$ be the transition operator of our Markov chain, i.e. for any $x, y \in \Omega$, $P(x,y)$ gives us the probability that our Markov chain moves from x to y. Clearly, we must have $\sum_{y \in \Omega} P(x,y) = 1$ for all $x \in \Omega$, i.e. $P\mathbb{1} = \mathbb{1}$. [1]
Now, suppose P is reversible w.r.t. π, i.e. for any $x, y \in \Omega$, we have:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \tag{1.1}$$

The above condition is also known as the *detailed balance* condition. Note that reversibility implies *stationarity*, [2] i.e.

$$\pi^{\mathsf{T}} P = \pi^{\mathsf{T}}$$

Indeed, $(\pi^{\mathsf{T}}P)_x = \sum_{y \in \Omega} \pi(y)P(y,x) = \sum_{y \in \Omega} \pi(x)P(x,y) = \pi(x) \sum_{y \in \Omega} P(x,y) = \pi(x)$.
Since P is reversible w.r.t. π, P is *self-adjoint* w.r.t. the inner product defined by π. Indeed, for $u, v \in \mathbb{R}^{\Omega}$, define:

$$\langle u, v \rangle_{\pi} := \sum_{x \in \Omega} \pi(x)u(x)v(x)$$

Also, let $D_{\pi} := \operatorname{diag}(\pi)$, i.e. $D_{\pi} \in \mathbb{R}^{\Omega \times \Omega}$ is a diagonal matrix such that $D_{\pi}(x,x) := \pi(x)$. Then Eq. (1.1) implies that

$$D_{\pi}P = P^{\mathsf{T}}D_{\pi} \iff D_{\pi}^{1/2}PD_{\pi}^{-1/2} = D_{\pi}^{-1/2}P^{\mathsf{T}}D_{\pi}^{1/2}$$

Thus the matrix $Q := D_{\pi}^{1/2}PD_{\pi}^{-1/2}$ is symmetric. Thus, Q has real eigenvalues. Furthermore, since P is similar to Q, the spectra of P and Q are exactly identical, i.e. P and Q have the same eigenvalues with the same multiplicities. Thus, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of P, and let $v^{(1)}, \ldots, v^{(n)}$ be the corresponding eigenvectors. Since P is self-adjoint w.r.t. $\langle \cdot, \cdot \rangle_{\pi}$, we can choose $v^{(1)}, \ldots, v^{(n)}$ to be orthonormal w.r.t. the $\langle \cdot, \cdot \rangle_{\pi}$ inner product, i.e. $\langle v^{(i)}, v^{(j)} \rangle_{\pi} = \delta_{ij} \iff v^{(i)}D_{\pi}v^{(j)} = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Furthermore, observe that $u^{(i)} := D_{\pi}^{1/2}v^{(i)}$ are orthonormal eigenvectors of Q, with eigenvalues λ_i. Thus, by the spectral theorem,

$$Q = \sum_{i=1}^{n} \lambda_i u^{(i)}(u^{(i)})^{\mathsf{T}} \implies P = \sum_{i=1}^{n} \lambda_i v^{(i)}(v^{(i)})^{\mathsf{T}}D_{\pi}$$

Now, since $P\mathbb{1} = \mathbb{1}$, WLOG we assume $\lambda_1 = 1, v^{(1)} = \mathbb{1}$. [3] Thus

$$P = \mathbb{1}\mathbb{1}^{\mathsf{T}}D_{\pi} + \sum_{i=2}^{n} \lambda_i v^{(i)}(v^{(i)})^{\mathsf{T}}D_{\pi} \implies P^t = \mathbb{1}\mathbb{1}^{\mathsf{T}}D_{\pi} + \sum_{i=2}^{n} \lambda_i^t v^{(i)}(v^{(i)})^{\mathsf{T}}D_{\pi}$$

Note that P^t is the t-fold application of the Markov chain kernel. Thus, to analyze convergence, we need to 'test' it against some initial distribution μ on Ω. Thus, let μ be a distribution on Ω. Then

$$\mu^{\mathsf{T}}P^t - \pi^{\mathsf{T}} = (\mu^{\mathsf{T}}\mathbb{1}\mathbb{1}^{\mathsf{T}}D_{\pi} - \pi^{\mathsf{T}}) + \sum_{i=2}^{n} \lambda_i^t \mu^{\mathsf{T}}v^{(i)}(v^{(i)})^{\mathsf{T}}D_{\pi}$$

Note that $\mu^{\mathsf{T}}\mathbb{1} = 1$, and $\mathbb{1}^{\mathsf{T}}D_{\pi} = \pi^{\mathsf{T}}$, and thus

$$\mu^{\mathsf{T}}P^t - \pi^{\mathsf{T}} = \sum_{i=2}^{n} \lambda_i^t \left(\mu^{\mathsf{T}}v^{(i)}\right) \cdot \left((v^{(i)})^{\mathsf{T}}D_{\pi}\right)$$

[1] non-negative matrices M which satisfy $M\mathbb{1} = \mathbb{1}$ are known as *stochastic*
[2] We treat π as a vector in $\mathbb{R}^{\Omega}$
[3] note that $\mathbb{1}$ is a unit vector in the $\langle \cdot, \cdot \rangle_{\pi}$ inner product: Indeed, $\langle \mathbb{1}, \mathbb{1} \rangle_{\pi} = \sum_{x \in \Omega} \pi(x) = 1$

Write $q := D_\pi^{-1}\mu$. Then rewriting the above expression yields

$$\mu^\mathsf{T} P^t - \pi^\mathsf{T} = \underbrace{\left(\sum_{i=2}^{n} \lambda_i^t \langle q, v^{(i)}\rangle_\pi (v^{(i)})^\mathsf{T}\right)}_{=:y^\mathsf{T}} D_\pi$$

Thus

$$\|\mu^\mathsf{T} P^t - \pi^\mathsf{T}\|_1 = \|y^\mathsf{T} D_\pi\|_1 = \sum_{x\in\Omega} \pi(x)|y(x)| = \mathbb{E}_{x\sim\pi}|y(x)| \overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}_{x\sim\pi} y(x)^2}$$

But note that $\mathbb{E}_{x\sim\pi} y(x)^2 = \sum_{x\in\Omega} \pi(x) y(x)^2 = \langle y, y\rangle_\pi = \|y\|_\pi^2$, where $\|\cdot\|_\pi$ is the norm associated with the inner product $\langle\cdot,\cdot\rangle_\pi$. Thus

$$\|\mu^\mathsf{T} P^t - \pi^\mathsf{T}\|_1 \leq \|y\|_\pi$$

Now,

$$\|y\|_\pi^2 = \sum_{i=2}^{n} \lambda_i^{2t} \langle q, v^{(i)}\rangle_\pi^2 \leq \lambda^{2t} \sum_{i=2}^{n} \langle q, v^{(i)}\rangle_\pi^2$$

where $\lambda := \sup_{i\geq 2} |\lambda_i|$. Now, note that $q = \sum_{i=1}^{n} \langle q, v^{(i)}\rangle_\pi v^{(i)}$. Thus

$$\sum_{i=2}^{n} \langle q, v^{(i)}\rangle_\pi^2 \leq \sum_{i=1}^{n} \langle q, v^{(i)}\rangle_\pi^2 = \|q\|_\pi^2$$

Consequently,

$$\|y\|_\pi \leq \lambda^t \|q\|_\pi$$

Now, $q = D_\pi^{-1}\mu$. Thus $q(x) = \mu(x)/\pi(x)$. Consequently,

$$\|q\|_\pi^2 = \sum_{x\in\Omega} \frac{\mu(x)^2}{\pi(x)} \leq \frac{1}{\pi_{\min}} \sum_{x\in\Omega} \mu(x)^2 \leq \frac{1}{\pi_{\min}}$$

where $\pi_{\min} := \min_{x\in\Omega} \pi(x)$.
Putting everything together, we get:

$$\|\mu^\mathsf{T} P^t - \pi^\mathsf{T}\|_{\mathrm{TV}} = \frac{1}{2}\|\mu^\mathsf{T} P^t - \pi^\mathsf{T}\|_1 \leq \frac{\lambda^t}{2\sqrt{\pi_{\min}}}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the *total variation distance*. Recall that for any measurable space $(\Omega, \mathcal{B})$, given two finite measures μ_1, μ_2 on Ω, we define their total variation distance to be:

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} := \sup_{A\in\mathcal{B}} |\mu_1(A) - \mu_2(A)|$$

In case Ω is finite, $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \|\mu_1 - \mu_2\|_1/2$.
We finally define the notion of *mixing time*:

Definition 1.1. Fix $\varepsilon > 0$. Let Ω be the state space of our Markov chain, and let P be its transition kernel. Suppose P is reversible w.r.t. π. Let μ be the initial distribution of our Markov chain. Then we define:

$$t_{\mathrm{mix}}(P, \varepsilon; \mu) := \inf_{t>0} \left\{\|P^t[\mu] - \pi\|_{\mathrm{TV}} \leq \varepsilon\right\}$$

where $P^t[\mu] := (\mu^\mathsf{T} P^t)^\mathsf{T}$.

We also define the *spectral gap* of our Markov chain:

Definition 1.2. Let P be a Markov transition kernel. Suppose P is reversible w.r.t. π, where $\pi \in (0,1]^\Omega$ is a probability distribution. Let $\lambda := \sup_{i \geq 2} |\lambda_i|$, where $1, \lambda_2, \ldots,$ are the eigenvalues of P. Then the spectral gap of P, denoted γ, is defined to be $1 - \lambda$.

Remark. It is easy to see that $\lambda \leq 1$: Indeed, let λ_i be any eigenvalue of P. Then by the Gerschgorin circle theorem,

$$|\lambda_i - P(x,x)| \leq \left| \sum_{y \neq x} P(x,y) \right| = \sum_{y \neq x} P(x,y) = 1 - P(x,x) \implies |\lambda_i| - |P(x,x)| \leq |\lambda_i - P(x,x)| \leq 1 - P(x,x) \implies |\lambda_i| \leq 1$$

We summarize the above discussion below:

Theorem 1.1. Let Ω be a finite set, and let P be a transition matrix over Ω. Suppose P is reversible w.r.t. π, where π is a probability distribution over Ω. Let γ be the spectral gap of P. Let $\pi_{\min} := \min_{x \in \Omega} \pi(x)$, and suppose $\pi_{\min} > 0$. Then for any distribution μ over Ω, and any $\varepsilon \in (0, 1/2)$, we have:

$$t_{\text{mix}}(P, \varepsilon; \mu) \leq \left\lceil \frac{\log(2\varepsilon\sqrt{\pi_{\min}})}{\log(1-\gamma)} \right\rceil \leq \left\lceil \frac{1}{\gamma} \cdot \left(\log \frac{1}{2\varepsilon} + \frac{1}{2} \log \frac{1}{\pi_{\min}} \right) \right\rceil = \mathcal{O}\left(\frac{1}{\gamma} \cdot \left(\log \frac{1}{\varepsilon} + \log \frac{1}{\pi_{\min}} \right) \right)$$

Remark. The above theorem establishes the central role of spectral gaps in the convergence of Markov chains.

One Small Patch: Lazy Markov Chains

Note that we proved that the spectral gap of a Markov chain is non-negative. However, if the spectral gap of a Markov chain is 0, then the above mixing time bound is useless. Now, the spectral gap does turn out to be 0 in some important cases. For example, let G be a d-regular bipartite graph. Consider a random walk on this graph, where given any vertex, we move to one of its neighbors, uniformly. Thus, the transition matrix for this random walk is A/d, where A is the adjacency matrix of G. Now, since G is bipartite, A/d has -1 as an eigenvalue, and consequently, $\lambda = 1$, and $\gamma = 0$.
The reason the above phenomenon happened is because the transition matrix was *periodic*, i.e. if our initial distribution was completely localized in one component of the bipartite graph, then no matter how many times we apply the transition matrix, the random walk will not spread over completely to the graph.
To remedy this, we introduce the notion of *lazy random walks*:

Definition 1.3. Let P be a transition matrix. Then $P' := (I + P)/2$ is the *lazy* version of P.

Note that:

1. P' is also a legitimate transition matrix.
2. If P is reversible w.r.t. π, then so is P'.
3. $\lambda_i(P') = (1 + \lambda_i(P))/2$ for all $i \in [n]$. Since $\lambda_i(P) \geq -1$ for all i, we get that $\lambda_i(P') \geq 0$ for all i, i.e. all eigenvalues of P' are non-negative. In particular, $\lambda(P') = \sup_{i \geq 2} |\lambda_i(P')| = \sup_{i \geq 2} \lambda_i(P') = \lambda_2(P')$, i.e. $\lambda(P')$ is just the second-largest eigenvalue of P'.
4. If the multiplicity of 1 as an eigenvalue of P is 1, then the multiplicity of 1 as an eigenvalue of P' is also 1, and consequently, $\lambda(P') < 1$. Now, if P is *irreducible*, i.e. for every $x, y \in \Omega$, there exists $n \geq 1$ such that $P^n(x,y) > 0$, then the multiplicity of 1 as an eigenvalue of P is 1. Consequently, the *lazy version of an irreducible transition matrix has a non-zero spectral gap*.

5. If the spectral gap of P is γ, then the spectral gap of P' is $\geq \gamma/2$. Thus, making a chain lazy doesn't (up to constants) hurt the time bound given by Theorem 1.1.

1.1. Poincaré Inequality

Let us revisit the definition of variance: For some distribution π, and a function $f : \Omega \mapsto \mathbb{R}$, we define the variance of f to be:

$$\mathbf{Var}_\pi(f) := \mathbb{E}_{x\sim\pi}\left[f(x)^2\right] - \mathbb{E}_{x\sim\pi}\left[f(x)\right]^2 = \mathbb{E}_{x\sim\pi}\left[\Big(f(x) - \big(\mathbb{E}_{y\sim\pi} f(y)\big)\Big)^2\right]$$

Let us give a more 'probabilistic' interpretation of the above definition.

Proposition 1. Let $X, Y \sim \pi$ be two i.i.d random variables. Then

$$\mathbf{Var}(f(X)) = \frac{1}{2}\mathbb{E}\left[(f(X) - f(Y))^2\right]$$

Proof.

$$\mathbb{E}\left[(f(X) - f(Y))^2\right] = \mathbb{E}\left[f(X)^2\right] + \mathbb{E}\left[f(Y)^2\right] - 2\mathbb{E}\left[f(X)f(Y)\right] = 2\mathbb{E}\left[f(X)^2\right] - 2\mathbb{E}\left[f(X)\right]^2$$

where the last equality follows from the fact $\mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[f(Y)^2\right]$ (since X, Y are identically distributed), and $\mathbb{E}\left[f(X)f(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[f(Y)\right]$ (since X, Y are independent). ■

Motivated by this, we define a quadratic form over $\mathbb{R}^\Omega$ (which is the space of functions mapping Ω to $\mathbb{R}$):

Definition 1.4 (Dirichlet Forms)**.** Let Ω be a finite set, and let $f, g \in \mathbb{R}^\Omega$ be functions. Let P be a Markov transition kernel, and let π be a distribution over Ω. Then define

$$\mathcal{E}(f,g) := \frac{1}{2}\mathbb{E}_{\substack{X\sim\pi \\ Y\sim P(X,\cdot)}}\left[(f(X) - f(Y))(g(X) - g(Y))\right] = \frac{1}{2}\sum_{x,y\in\Omega}\pi(x)P(x,y)(f(x) - f(y))(g(x) - g(y))$$

Remark. Note that *X and Y are not independent*. Also note that $\mathcal{E}(f,g) = \mathcal{E}(g,f)$, i.e. $\mathcal{E}$ is a symmetric functional.

We now prove that $\mathcal{E}(\cdot,\cdot)$ is very much like variance, but with a crucial difference: In Proposition 1, X and Y were identically distributed *independent* random variables. In the following lemma, both X and Y are still distributed as π, but $Y \sim P(X,\cdot)$, and thus they are no longer independent.

Proposition 2. Let $X \sim \pi, Y \sim P(X,\cdot)$. Let $f \in \mathbb{R}^\Omega$. Then

$$\mathcal{E}(f,f) = \frac{1}{2}\mathbb{E}\left[(f(X) - f(Y))^2\right] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}\left[f(X)f(Y)\right]$$

Proof.

$$\mathbb{E}\left[(f(X) - f(Y))^2\right] = \sum_{x,y\in\Omega}\Pr(X = x, Y = y)(f(x) - f(y))^2 = \sum_{x,y\in\Omega}\Pr(Y = y|X = x)\Pr(X = x)(f(x) - f(y))^2$$

$$= \sum_{x,y\in\Omega}P(x,y)\pi(x)(f(x) - f(y))^2$$

The second equality follows by expanding $(f(X) - f(Y))^2$ and noting that $\mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[f(Y)^2\right]$ since X and Y are identically distributed. ■

We are finally ready to introduce the main theorem which justifies the definition of Dirichlet energy:

Theorem 1.2 (Poincaré Inequality)**.** Let Ω be a finite set, and let P be a transition matrix over Ω. Suppose P is reversible w.r.t. π, where $\pi \in (0,1]^\Omega$ is a probability distribution over Ω. Let $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$ be the eigenvalues of P. Then

$$\mathcal{E}(f,f) \geq (1-\lambda_2)\,\mathbf{Var}(f)$$

Furthermore, there exists $f_* \in \mathbb{R}^\Omega$ such that $\mathcal{E}(f_*,f_*) = (1-\lambda_2)\,\mathbf{Var}(f_*)$, i.e. the above inequality is tight.

Remark. There are a few remarks due:

1. $1-\lambda_2$ is also known as the *Poincaré constant* of the system (P,π).
2. One can think of $\mathcal{E}(\cdot,\cdot)$ as some sort of 'local variance' across transitions of our Markov chain. The above inequality then bounds the ratio between local and 'global' variance.
3. Note that Poincaré's constant is $1-\lambda_2$, while the spectral gap of P is $1-\lambda$, where $\lambda = \sup_{i\geq 2}|\lambda_i| = \max\{\lambda_2, |\lambda_n|\}$. Thus, the Poincaré constant and the spectral gap of a system are slightly different quantities. However, as we saw earlier, for lazy chains these are the same quantity.
4. The above theorem gives a non-linear algebraic way to recover the spectral gap of our transition kernel.
5. We modified the definition of variance to arrive at the definition for Dirichlet forms. We were then able to recover the notion of spectral gaps by studying the relationship between Dirichlet forms and variance, i.e. by studying Dirichlet forms, we were able to gain insight into the rate of convergence of our Markov chain. This raises the question: Can we study any other functional which also gives us insight into Markov chain convergence? The answer is yes, and we shall deal with it soon.

Proof. We first recall the so-called *Rayleigh's criterion* from linear algebra: Let S be a symmetric matrix with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots$. Let the eigenvector corresponding to μ_1 be v_1. Then

$$\mu_2 = \sup_{\substack{v\neq 0 \\ \langle v, v_1\rangle = 0}} \frac{v^\mathsf{T} A v}{\|v\|_2^2}$$

Furthermore, equality is attained in the above supremum for $v = v_2$, where v_2 is an eigenvector of A corresponding to the eigenvalue μ_2.
Now, recall the matrix $Q := D_\pi^{1/2} P D_\pi^{-1/2}$. Since Q is symmetric, we can apply Rayleigh's criterion to it. Furthermore, the eigenvalues of Q are the same as the eigenvalues of P, and the eigenvector of Q corresponding to $\lambda_1 = 1$ is $\sqrt{\pi}$, where $\sqrt{\pi} \in \mathbb{R}^\Omega$ is component-wise square root of π. Then by Rayleigh's criterion,

$$\lambda_2 = \sup_{\sum_{x\in\Omega}\sqrt{\pi(x)}v(x)=0} \frac{\sum_{x,y\in\Omega} v(x)Q(x,y)v(y)}{\sum_{x\in\Omega} v(x)^2}$$

Now, set $v(x) = \sqrt{\pi(x)}f(x)$. Also recall that $Q(x,y) = P(x,y)\sqrt{\pi(x)/\pi(y)}$. Then

$$\sup_{\sum_{x\in\Omega}\sqrt{\pi(x)}v(x)=0} \frac{\sum_{x,y\in\Omega} v(x)Q(x,y)v(y)}{\sum_{x\in\Omega} v(x)^2} = \sup_{\sum_{x\in\Omega}\pi(x)f(x)=0} \frac{\sum_{x,y\in\Omega} \pi(x)P(x,y)f(x)f(y)}{\sum_{x\in\Omega}\pi(x)f(x)^2}$$

Now, $\sum_{x\in\Omega}\pi(x)f(x)^2 = \|f\|_\pi^2$. Also, if $X\sim\pi, Y\sim P(X,\cdot)$, then

$$\sum_{x,y\in\Omega}\pi(x)P(x,y)f(x)f(y) = \mathbb{E}[f(X)f(Y)]$$

Thus

$$\lambda_2 = \sup_{\sum_{x\in\Omega}\pi(x)f(x)=0} \frac{\mathbb{E}[f(X)f(Y)]}{\|f\|_\pi^2}$$

Now, by Proposition 2, $\mathbb{E}[f(X)f(Y)] = \mathbb{E}[f(X)^2] - \mathcal{E}(f,f)$. But $\mathbb{E}[f(X)^2] = \sum_{x\in\Omega}\pi(x)f(x)^2 = \|f\|_\pi^2$. Thus

$$\lambda_2 = \sup_{\sum_{x\in\Omega}\pi(x)f(x)=0} \frac{\|f\|_\pi^2 - \mathcal{E}(f,f)}{\|f\|_\pi^2} \implies \inf_{\sum_{x\in\Omega}\pi(x)f(x)=0} \frac{\mathcal{E}(f,f)}{\|f\|_\pi^2} = 1-\lambda_2$$

At this point, we are almost done, except for the fact that we only proved the inequality for functions f for which $\sum_{x\in\Omega}\pi(x)f(x) = 0 \iff \mathbb{E}[f(X)] = 0$. Now, let $g \in \mathbb{R}^\Omega$ be an arbitrary function. Set $f(x) := g(x) - \mathbb{E}[g(X)]$. Then $\mathbb{E}[f(X)] = 0$, and we have $\mathcal{E}(f,f) \geq (1-\lambda_2)\,\mathbf{Var}(f)$. But note that

$$\mathbf{Var}(g) = \mathbb{E}_{x\sim\pi}\left[\left(g(x) - \mathbb{E}_{y\sim\pi}g(y)\right)^2\right] = \mathbb{E}_{x\sim\pi}\left[f(x)^2\right] = \mathbb{E}[f(X)^2] = \mathbb{E}[f(X)^2] - \mathbb{E}[f(X)]^2 = \mathbf{Var}(f)$$

Also, for any $x, y \in \Omega$, $g(x) - g(y) = f(x) - f(y)$. Thus, by the very definition of Dirichlet forms, we have $\mathcal{E}(g,g) = \mathcal{E}(f,f)$. Thus we have $\mathcal{E}(f,f) = \mathcal{E}(g,g) \geq (1-\lambda_2)\,\mathbf{Var}(f) = (1-\lambda_2)\,\mathbf{Var}(g)$, as desired.
Finally, equality is attained in Poincaré's inequality: Indeed, if we trace the steps of the derivation, one finds that equality is attained for $f_* := D_\pi^{-1/2}u^{(2)}$, where $u^{(2)}$ is an eigenvector of Q corresponding to the eigenvalue λ_2. ■

Once we have established Poincaré's inequality, we can in fact derive mixing time bounds again without any linear algebra. Furthermore, in this derivation, we observe a clear monovariant (apart from the obvious ℓ_1 norm).

1.1.1. Dirichlet Energy as Rate of Change of Variance

Let μ be a distribution over Ω. We want to show that μ converges to π in a reasonable amount of time. We showed that directly by considering $\|P^t[\mu] - \pi\|_1$. However, is there any other proxy by which we can show convergence to π? The answer is yes, and that 'proxy' is variance.
However, we shall now move into continuous time, where certain calculations become much easier.
Let P be the transition matrix of our DTMC (Discrete Time Markov Chain). Now, when we are at any state $r \in \Omega$, we *wait* for a time given by an exponential random variable with parameter 1, and then we jump to another state according to the probability distribution dictated by (the r^{th} row of) P. Thus, our time parameter now takes values in $\mathbb{R}_{\geq 0}$, since exponential random variables are real-valued. Moreover, note what the trajectory of our CTMC (Continuous Time Markov Chain) looks like: Suppose we start from some state $r_0 \in \Omega$. Then the trajectory is

$$X_t = \begin{cases} r_0, & t \in [0, t_1) \\ r_1, & t \in [t_1, t_1+t_2) \\ r_2, & t \in [t_1+t_2, t_1+t_2+t_3) \\ \dots \end{cases}$$

where $t_1, t_2, t_3, \dots$ are i.i.d $\text{Exp}(1)$ random variables, and $(r_0, r_1, r_2, \dots)$ is a particular trajectory of the DTMC governed by P, and $\{X_t\}_{t\geq 0}$ is the trajectory of our CTMC. Indeed, the jump-and-hold description of a CTMC makes apparent the correspondence between a DTMC and a CTMC: One can take a particular trajectory of a DTMC, and separate its points with i.i.d exponential random variables to get a corresponding trajectory for the CTMC.
The description given above is known as the jump-and-hold description of a CTMC. However, for computational purposes, it is often convenient to work with another, equivalent, description of a CTMC, which is known as the infinitesimal description of a CTMC. That goes as follows: Suppose we have a CTMC which is generated in a jump-and-hold manner from a stochastic matrix P. Then we associate a transition matrix $L := P - I$ [4] to our CTMC, which has the interpretation that, for any $i \neq j, i, j \in \Omega$,

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

[4] L is also known as the Laplacian of the system

or equivalently,

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

Also, note the following facts about L:

1. For $i = j$, $L_{ii} = -\sum_{j \neq i} L(i,j)$, thus making $\sum_{j \in \Omega} L(i,j) = 0$ for every $i \in \Omega$.
2. If P is reversible w.r.t. π, then so is L.

The reason the infinitesimal description is useful is that it can be shown that $\mathbb{P}(X_t = j | X_0 = i) =: \mathbb{P}_i(X_t = j) = (\exp(Lt))_{ij}$. The set of matrices $\{\exp(tL)\}_{t \geq 0}$ is also known as the *semigroup generated by* L.
We shall now justify all of the above assertions in the theorem below.

Theorem 1.3 (Infinitesimal Description of CTMCs)**.** Let P be a transition matrix, and let μ_0 be the initial distribution of our Markov chain. Let μ_t be the distribution at time t of the CTMC governed by P. Then

$$\mu_t^{\mathsf{T}} = \mu_0^{\mathsf{T}} \exp(tL) = \mu_0^{\mathsf{T}} \exp(t(P - I))$$

Proof. Note that

$$\mu_t^{\mathsf{T}} = \sum_{k=0}^{\infty} \mu_0^{\mathsf{T}} P^k \cdot \Pr(\text{There are } k \text{ 'jumps' in time } t)$$

The distribution of the number of jumps in time t is actually a Poisson distribution with parameter t! Thus

$$\mu_t^{\mathsf{T}} = \mu_0^{\mathsf{T}} \sum_{k=0}^{\infty} P^k \cdot \frac{e^{-t} t^k}{k!} = \mu_0^{\mathsf{T}} e^{-t} \sum_{k=0}^{\infty} \frac{(tP)^k}{k!} = \mu_0^{\mathsf{T}} e^{-t} \exp(tP) = \mu_0^{\mathsf{T}} \exp(t(P - I))$$

■

Remark. A few remarks are due:

1. Yet another viewpoint of CTMCs (which is also highlighted in the proof above), which is equivalent to the jump-and-hold and infinitesimal descriptions is the observation that if $\{X_k\}_{k \in \mathbb{N}}$ is a DTMC, then $Y_t \sim X_{\text{Poisson}(t)}$ is the CTMC corresponding to $L = P - I$.
2. $\mathbb{P}(X_h = j | X_0 = i) = (\delta_i^{\mathsf{T}} \exp(hL))(j)$, where δ_i is the Dirac measure at i, i.e. $\delta_i \in \mathbb{R}^{\Omega}$ is the unit vector with 1 at i and 0 everywhere else. But $(\delta_i^{\mathsf{T}} \exp(hL))(j) = (\exp(hL))_{ij}$. Now, $\exp(hL) = I + hL + (hL)^2/2! + \cdots$, and thus $\exp(hL)_{ij} = \delta_{ij} + hL(i,j) + o(h)$. Consequently, if $i \neq j$, then $\mathbb{P}(X_h = j | X_0 = i) = hL(i,j) + o(h)$, as desired.

Now, let $\{\mu_t(j)\}_{j \in \Omega}$ denote the probability distribution on Ω at time t. Then

$$\Pr(X_{t+h} = j) = \sum_{i \in \Omega} \Pr(X_{t+h} = j | X_t = i) \Pr(X_t = i)$$

$$= \sum_{i \neq j} (L(i,j)h + o(h))\mu_t(i) + \Pr(X_{t+h} = j | X_t = j) \Pr(X_t = j)$$

Consequently,

$$\lim_{h \searrow 0} \frac{\Pr(X_{t+h} = j) - \Pr(X_t = j)}{h} = \sum_{i \neq j} L(i,j)\mu_t(i) + \underbrace{\Pr(X_t = j)}_{= \mu_t(j)} \lim_{h \searrow 0} \frac{\Pr(X_{t+h} = j | X_t = j) - 1}{h}$$

$$= \sum_{i \neq j} L(i,j)\mu_t(i) - \mu_t(j) \lim_{h \searrow 0} \frac{\Pr(X_{t+h} \neq j | X_t = j)}{h} = \sum_{i \neq j} L(i,j)\mu_t(i) - \mu_t(j) \underbrace{\sum_{k \neq j} L(j,k)}_{= -L(j,j)}$$

$$= \sum_{i \in \Omega} L(i,j)\mu_t(i)$$

But note that

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

Consequently, we arrive at the continuous version of the *Chapman-Kolmogorov theorem*, namely, for any $j \in \Omega$,

$$\frac{\mathrm{d}\mu_t(j)}{\mathrm{d}t} = \sum_{i \in \Omega} L(i,j)\mu_t(i) \tag{1.2}$$

Before we return to Markov chain mixing, we prove a useful identity involving $\mathcal{E}(\cdot,\cdot)$ and the Laplacian of the system.

Proposition 3. Let $L = P - I$. Then:

$$\mathcal{E}(f,g) = -\sum_{x,y \in \Omega} \pi(x)L(x,y)f(x)g(y)$$

Proof. Note that

$$2\mathcal{E}(f,g) = \sum_{x,y \in \Omega} \pi(x)P(x,y)(f(x) - f(y))(g(x) - g(y))$$

$$= \sum_{x,y \in \Omega} \pi(x)(L(x,y) + \delta_{xy})(f(x)g(x) - f(x)g(y) - f(y)g(x) + f(y)g(y))$$

Now, note that the above expression is symmetric in x and y, since $\pi(x)L(x,y) = \pi(y)L(y,x)$. Consequently, the above expression equals

$$= 2\sum_{x,y \in \Omega} \pi(x)(L(x,y)+\delta_{xy})(f(x)g(x)-f(x)g(y)) = 2\sum_{x,y \in \Omega} \pi(x)(L(x,y)f(x)g(x)+\delta_{xy}f(x)g(x)-L(x,y)f(x)g(y)-\delta_{xy}f(x)g(y))$$

Note that $\delta_{xy}f(x)g(x) = \delta_{xy}f(x)g(y)$. Thus the above expression equals

$$2\sum_{x,y \in \Omega} \pi(x)(L(x,y)f(x)g(x) - L(x,y)f(x)g(y)) = 2\sum_{x,y \in \Omega} \pi(x)L(x,y)f(x)(g(x) - g(y))$$

Thus

$$\mathcal{E}(f,g) = \sum_{x,y \in \Omega} \pi(x)L(x,y)f(x)(g(x) - g(y))$$

Thus if we can show that $\sum_{x,y \in \Omega} \pi(x)L(x,y)f(x)g(x) = 0$, then we'll be done. But

$$\sum_{x,y \in \Omega} \pi(x)L(x,y)f(x)g(x) = \sum_{x} \pi(x)f(x)g(x)\left(\sum_{y \in \Omega} L(x,y)\right) = 0$$

Where the last equality follows since the rows of the Laplacian do sum to 0 (since the rows of P sum to 1, and the rows of I also sum to 1). ∎

We will finally use this machinery to link Dirichlet energy to the rate of change of variance.

Lemma 1.4 (Dirichlet Energy as Rate of Change of Variance)**.** Let Ω be a finite set, and let P be a transition matrix over Ω. Suppose P is reversible w.r.t. π, where $\pi \in (0,1]^\Omega$ is a probability distribution over Ω. Let $\mu \in [0,1]^\Omega$ be a probability distribution over Ω. Let μ_t be the distribution of the CTMC at time t, governed by P (or more precisely L) with $\mu_0 = \mu$. Let $g_t := \mu_t/\pi$, i.e. $g_t(x) := \mu_t(x)/\pi(x)$. Then

$$\frac{\mathrm{d}\mathbf{Var}_\pi(g_t)}{\mathrm{d}t} = -2\mathcal{E}(g_t, g_t)$$

Remark. Note that if $\mu = \pi$, then $g \equiv 1$, and thus $\mathbf{Var}(g) = 0$. Thus $\mathbf{Var}_\pi(g)$ actually acts as some sort of distance between probability distributions and π.

Proof. Note that

$$\mathbf{Var}_\pi(g_t) = \sum_{x\in\Omega} \pi(x) g_t(x)^2 - \left(\sum_{x\in\Omega} \pi(x) g_t(x)\right)^2 = \sum_{x\in\Omega} \frac{\mu_t(x)^2}{\pi(x)} - \left(\sum_{x\in\Omega} \mu_t(x)\right)^2 = \sum_{x\in\Omega} \frac{\mu_t(x)^2}{\pi(x)} - 1$$

Thus

$$\frac{\mathrm{d}\mathbf{Var}_\pi(g_t)}{\mathrm{d}t} = \sum_{x\in\Omega} \frac{2\mu_t(x)}{\pi(x)} \frac{\mathrm{d}\mu_t(x)}{\mathrm{d}t} \overset{\text{Eq. (1.2)}}{=} \sum_{x\in\Omega} \frac{2\mu_t(x)}{\pi(x)} \sum_{y\in\Omega} L(y,x)\mu_t(y) = 2\sum_{x,y\in\Omega} \frac{\mu_t(x)}{\pi(x)} \frac{\mu_t(y)}{\pi(y)} \pi(y) L(y,x)$$

$$= 2\sum_{x,y\in\Omega} g_t(x) g_t(y) \pi(y) L(y,x) = 2\sum_{x,y\in\Omega} g_t(x) g_t(y) \pi(x) L(x,y)$$

At this point, we're done by Proposition 3. ∎

Theorem 1.5 (Exponential Decay of Variance)**.** Let P be a transition matrix, π be a distribution, with P reversible w.r.t. π. Let α be the Poincaré constant of the system. Let $\mu \in [0,1]^\Omega$ be a probability distribution over Ω. Let μ_t be the distribution of the CTMC at time t, governed by P with $\mu_0 = \mu$. Let $g_t := \mu_t/\pi$. Then

$$\mathbf{Var}_\pi(g_t) \le e^{-2\alpha t}\, \mathbf{Var}_\pi(g_0)$$

Proof. Note that

$$\frac{\mathrm{d}\mathbf{Var}_\pi(g_t)}{\mathrm{d}t} \overset{\text{Lemma 1.4}}{=} -2\mathcal{E}(g_t, g_t) \overset{\text{Theorem 1.2}}{\le} -2\alpha\, \mathbf{Var}_\pi(g_t) \implies \frac{\mathrm{d}\mathbf{Var}_\pi(g_t)}{\mathbf{Var}_\pi(g_t)} \le -2\alpha\, \mathrm{d}t \implies \mathrm{d}\ln \mathbf{Var}_\pi(g_t) \le -2\alpha\, \mathrm{d}t$$

Integrating the above inequality yields the desired result. ∎

1.2. The Modified Log-Sobolev Inequality

We saw that Poincaré inequality, variance, and the spectral gap of a system are very tightly linked. As asked above, one might wonder if some other functional might also lead to mixing time bounds.
Since variance is a 'quadratic form', one might like to think of variance as some sort of 'energy'. We shall now explore what sort of bounds we get when we use 'entropy' as our functional.

Definition 1.5 (Relative Entropy)**.** Let μ, π be two probability distributions on Ω, where Ω is a finite set. Further, assume μ is *absolutely continuous* w.r.t π, i.e. $\pi(x) = 0 \implies \mu(x) = 0$ for all $x \in \Omega$. Then we define the *relative entropy* of μ w.r.t. π to be

$$D_{\mathrm{KL}}(\mu\|\pi) := \sum_{x\in\Omega} \mu(x) \ln \frac{\mu(x)}{\pi(x)}$$

Remark. A few remarks are due:

1. $D_{\mathrm{KL}}(\pi\|\pi) = 0$ for any distribution π.
2. Relative entropy is also known as *informational divergence*, or *Kullback-Leibler divergence*.
3. $D_{\mathrm{KL}}(\mu\|\pi)$ should be taken as the 'entropic' analog of $\mathbf{Var}_\pi(\mu/\pi)$.

The way Cauchy-Schwartz inequality is the 'canonical inequality' for quadratic forms, the same way *Pinsker's inequality* is the canonical inequality for $D_{\mathrm{KL}}(\cdot\|\cdot)$.

Theorem 1.6 (Pinsker's Inequality)**.** Let μ, π be two probability distributions on Ω, where Ω is a finite set. Further, assume μ is absolutely continuous w.r.t π. Then

$$\|\mu - \pi\|_{\mathrm{TV}} \leq \sqrt{2D_{\mathrm{KL}}(\mu\|\pi)}$$

We will now prove an equivalent of Lemma 1.4 for D_{KL}.

Lemma 1.7. Let Ω be a finite set, and let P be a transition matrix over Ω. Suppose P is reversible w.r.t. π, where $\pi \in (0,1]^\Omega$ is a probability distribution over Ω. Let $\mu \in [0,1]^\Omega$ be a probability distribution over Ω. Let μ_t be the distribution of the CTMC at time t, governed by P with $\mu_0 = \mu$. Let $g_t := \mu_t/\pi$, i.e. $g_t(x) := \mu_t(x)/\pi(x)$. Then

$$\frac{\mathrm{d}D(\mu_t\|\pi)}{\mathrm{d}t} = -\mathcal{E}(g_t, \ln g_t)$$

Proof. Note that

$$\frac{\mathrm{d}D(\mu_t\|\pi)}{\mathrm{d}t} = \sum_{x\in\Omega} \frac{\mathrm{d}\mu_t(x)}{\mathrm{d}t}\left(1 + \ln\frac{\mu_t(x)}{\pi(x)}\right) \overset{\text{Eq. (1.2)}}{=} \sum_{x\in\Omega}\sum_{y\in\Omega} L(y,x)\mu_t(y)\left(1+\ln g_t(x)\right) = \sum_{x,y\in\Omega} \pi(y)L(y,x)g_t(y)(1+\ln g_t(x))$$

$$= \sum_{x,y\in\Omega} \pi(y)L(y,x)g_t(y) + \sum_{x,y\in\Omega} \pi(y)L(y,x)g_t(y)\ln g_t(x)$$

The second term equals $-\mathcal{E}(\ln g, g) = -\mathcal{E}(g, \ln g)$ by Proposition 3, and the first term becomes zero since $\sum_{x\in\Omega} L(y,x)$ factors out and becomes 0. ∎

Consequently, if we could have an entropic analog of Poincaré's inequality, we would have the entropic analog of variance decay.
The inequality we prove is the so-called *Modified Log-Sobolev inequality*. Before that, we define the *entropy* of a function.

Definition 1.6 (Entropy)**.** Let Ω be a set, let π be a probability distribution on Ω, and let $f \in \mathbb{R}^\Omega_{\geq 0}$. Then we define the *entropy* of f to be:

$$\mathrm{Ent}_\pi(f) := \mathbb{E}_\pi\left[f\ln f\right] - \mathbb{E}_\pi\left[f\right]\cdot\ln\mathbb{E}_\pi\left[f\right]$$

We now define the so-called *Modified Log-Sobolev constant*:

Definition 1.7 (Modified Log-Sobolev Constant)**.** Let P be a transition matrix on Ω. Then the Modified Log-Sobolev constant of (P, π) is defined as:

$$\rho := \inf \left\{ \frac{\mathcal{E}(f, \ln f)}{\operatorname{Ent}_\pi(f)} : f \in \mathbb{R}^\Omega_{\geq 0}, \operatorname{Ent}_\pi(f) \neq 0 \right\}$$

Remark. A few remarks are due:

1. Let ρ be the modified Log-Sobolev constant of P, and let α be the Poincaré constant of P. Then $\rho \leq 2\alpha$. However, in many important situations, ρ and α are of the same order. In that case, the modified Log-Sobolev constant gives better mixing time bounds, as we shall see now.
2. The reason this is called the 'modified' Log-Sobolev constant is that the "original" Log-Sobolev constant was $\inf \mathcal{E}(\sqrt{f}, \sqrt{f}) / \operatorname{Ent}_\pi(f)$. However, it was later realized that the 'modified' Log-Sobolev constant was more general, and more useful for a wider variety of Markov chains, so that's why we only study this.

Once we have defined the Modified Log-Sobolev constant, we can try to derive mixing time bounds from it.

Theorem 1.8. Let Ω be a finite set, and let P be a transition matrix over Ω. Suppose P is reversible w.r.t. π, where π is a probability distribution over Ω. Let ρ be the modified Log-Sobolev constant of (P, π). Let $\pi_{\min} := \min_{x \in \Omega} \pi(x)$, and suppose $\pi_{\min} > 0$. Then for any distribution μ_0 over Ω, we have

$$\|\mu_t - \pi\|_{\mathrm{TV}}^2 \leq 2 \log \frac{1}{\pi_{\min}} e^{-\rho t}$$

Although the above result is for CTMCs, we can conclude that for small enough $\varepsilon > 0$, we have:

$$t_{\mathrm{mix}}(P, \varepsilon; \mu) = \mathcal{O}\left(\frac{1}{\rho} \cdot \left(\log \frac{1}{\varepsilon} + \log \log \frac{1}{\pi_{\min}} \right) \right)$$

where t_{mix} is the usual mixing time (for DTMCs).

Proof. By Lemma 1.7,

$$\frac{\mathrm{d} D(\mu_t \| \pi)}{\mathrm{d} t} = -\mathcal{E}(g_t, \ln g_t) \leq -\rho \operatorname{Ent}_\pi(g_t)$$

With some work one can verify that $\operatorname{Ent}_\pi(g_t) = D(\mu_t \| \pi)$. Thus, similarly as in the proof of Theorem 1.5, we obtain $D(\mu_t \| \pi) \leq e^{-\rho t} D(\mu_0 \| \pi)$. By Theorem 1.6, we then have $\|\mu_t - \pi\|_{\mathrm{TV}}^2 \leq 2 e^{-\rho t} D(\mu_0 \| \pi)$. Finally, once again with some thought one sees that $D(\mu_0 \| \pi) \leq \ln \frac{1}{\pi_{\min}}$ (with equality being achieved by $\mu_0 = \delta_x$, where $\pi(x) = \pi_{\min}$), as desired. ∎