1

Basic Bounds on Mixing Times

1.1 Preliminaries: Distances and mixing times

Let $(\Omega, \mathsf{P}, \pi)$ denote a transition probability matrix (or Markov kernel) of a finite Markov chain on a finite state space Ω with a unique invariant measure π. That is

$$\mathsf{P}(x,y) \geq 0, \quad \text{for all } x,y \in \Omega, \quad \text{and } \sum_{y\in\Omega} \mathsf{P}(x,y) = 1, \quad \text{for all } x \in \Omega.$$

$$\sum_{x\in\Omega} \pi(x)\mathsf{P}(x,y) = \pi(y), \quad \text{for all } y \in \Omega.$$

We assume throughout this paper that P is irreducible (i.e. Ω is strongly connected under P) and that π has full support (Ω). The minimal holding probability $\alpha \in [0,1]$ satisfies $\forall x \in \Omega : \mathsf{P}(x,x) \geq \alpha$, and if $\alpha \geq 1/2$ the chain is said to be lazy. If $A, B \subset \Omega$ the ergodic flow is $\mathsf{Q}(A,B) = \sum_{x\in A,\, y\in B} \pi(x)\mathsf{P}(x,y)$, while $A^c = \Omega \setminus A$ is the complement. For standard definitions and introduction to finite Markov chains, we refer the reader to [67] or [1].

It is a classical fact that if P is aperiodic then the measures $\mathsf{P}^n(x,\cdot)$ approach π as $n \to \infty$. Alternatively, let $k_n^x(y) = \mathsf{P}^n(x,y)/\pi(y)$ denote the density with respect to π at time $n \geq 0$, or simply $k_n(y)$ when the

start state or the start distribution is unimportant or clear from the context. Then the density $k_n^x(y)$ converges to 1 as $n \to \infty$. A proper quantitative statement may be stated using any one of several norms. In terms of L^p-distance

$$\|k_n - 1\|_{p,\pi}^p = \sum_{y\in\Omega} |k_n(y) - 1|^p \, \pi(y) \quad 1 \le p < +\infty \,.$$

When $p = 1$ and $p = 2$ these are closely related to the total variation distance and variance, respectively, such that if μ is a probability distribution on Ω, then

$$\begin{aligned} \|\mu - \pi\|_{TV} &= \frac{1}{2} \left\| \frac{\mu}{\pi} - 1 \right\|_{1,\pi} = \frac{1}{2} \sum_{y\in\Omega} |\mu(y) - \pi(y)| \\ \mathrm{Var}_\pi(\mu/\pi) &= \left\| \frac{\mu}{\pi} - 1 \right\|_{2,\pi}^2 = \sum_{y\in\Omega} \pi(y) \left(\frac{\mu(y)}{\pi(y)} - 1 \right)^2 \end{aligned}$$

Another important measure of closeness (but not a norm) is the informational divergence,

$$\mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) = \mathrm{Ent}_\pi(k_n^x) = \sum_{y\in\Omega} \mathsf{P}^n(x,y) \log \frac{\mathsf{P}^n(x,y)}{\pi(y)} \,,$$

where the entropy $\mathrm{Ent}_\pi(f) = \mathbb{E}_\pi f \log \frac{f}{\mathbb{E}_\pi f}$.

Each of these distances are convex, in the sense that if μ and ν are two distributions, and $s \in [0,1]$ then $dist((1-s)\mu + s\nu, \pi) \le (1-s)\, dist(\mu,\pi) + s\, dist(\nu,\pi)$. For instance, $\mathsf{D}(\mu\|\pi) = \mathrm{Ent}_\pi(\mu/\pi) = \mathbb{E}_\pi \frac{\mu}{\pi} \log \frac{\mu}{\pi}$ is convex in μ because $f \log f$ is convex. A convex distance $dist(\mu,\pi)$ satisfies the condition

$$\begin{aligned} dist(\sigma\mathsf{P}^n, \pi) &= dist\left(\sum_{x\in\Omega} \sigma(x) \mathsf{P}^n(x,\cdot), \pi \right) \\ &\le \sum_{x\in\Omega} \sigma(x) dist\left(\mathsf{P}^n(x,\cdot), \pi\right) \\ &\le \max_{x\in\Omega} dist(\mathsf{P}^n(x,\cdot), \pi) \,, \end{aligned} \tag{1.1}$$

and so distance is maximized when the initial distribution is concentrated at a point. To study the rate of convergence it then suffices to

study the rate when the initial distribution is a point mass δ_x (where δ_x is 1 at point $x \in \Omega$ and 0 elsewhere; likewise, let 1_A be one only on set $A \subset \Omega$).

Definition 1.1. The total variation, relative entropy and L^2 mixing times are defined as follows.

$$\begin{aligned} \tau(\epsilon) &= \min\{n : \forall x \in \Omega, \|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \le \epsilon\} \\ \tau_{\mathsf{D}}(\epsilon) &= \min\{n : \forall x \in \Omega, \mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) \le \epsilon\} \\ \tau_2(\epsilon) &= \min\{n : \forall x \in \Omega, \|k_n^x - 1\|_{2,\pi} \le \epsilon\} \end{aligned}$$

One may also consider the chi-square (χ^2) distance, which is just $\mathrm{Var}(k_n^x)$ and mixes in $\tau_{\chi^2}(\epsilon) = \tau_2(\sqrt{\epsilon})$. In the Appendix it is seen that $\tau_2(\epsilon)$ usually gives a good bound on L^∞ convergence, and so for most purposes nothing stronger than L^2 mixing need be considered.

An important concept in studying Markov chains is the notion of reversibility. The time-reversal P^* is defined by the identity $\pi(x)\mathsf{P}^*(x,y) = \pi(y)\mathsf{P}(y,x)$, $x, y \in \Omega$ and is the adjoint of P in the standard inner product for $L^2(\pi)$, that is $\langle f, \mathsf{P}g\rangle_\pi = \langle \mathsf{P}^* f, g\rangle_\pi$ where

$$\langle f, g\rangle_\pi = \sum_{x\in\Omega} \pi(x) f(x) g(x)$$

and a matrix M acts on a function $f : \Omega \to \mathsf{R}$ as

$$M\, f(x) = \sum_{y\in\Omega} M(x,y) f(y)\,.$$

A useful property of the reversal is that $k_n = \mathsf{P}^* k_{n-1}$, and inductively $k_n = (\mathsf{P}^*)^n k_0$. If $\mathsf{P}^* = \mathsf{P}$ then P is said to be time-reversible, or to satisfy the detailed balance condition. Given any Markov kernel P, two natural reversible chains are the additive reversibilization $\frac{\mathsf{P}+\mathsf{P}^*}{2}$, and multiplicative reversibilization $\mathsf{P}\mathsf{P}^*$.

A straightforward way to bound the L^2-distance is to differentiate the variance. In Lemma 1.4 it will be found that $\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t)$, where $\mathcal{E}(f,g)$ denotes a Dirichlet form, as defined below, and h_t the continuous time density defined in the following section. More generally, the Dirichlet form can be used in a characterization of

eigenvalues of a reversible chain (see Lemma 1.21), and to define the spectral gap and the logarithmic Sobolev type inequalities:

Definition 1.2. For $f, g : \Omega \to \mathsf{R}$, let $\mathcal{E}(f,g) = \mathcal{E}_{\mathsf{P}}(f,g)$ denote the Dirichlet form,

$$\mathcal{E}(f,g) = \langle f, (\mathsf{I}-\mathsf{P})g\rangle_\pi = \sum_{x,y} f(x)\,(g(x)-g(y))\,\mathsf{P}(x,y)\pi(x)\,.$$

If $f = g$ then

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y\in\Omega}(f(x)-f(y))^2\mathsf{P}(x,y)\pi(x)\,, \tag{1.2}$$

and

$$\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\mathsf{P}^*}(f,f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f,f)\,, \tag{1.3}$$

while if P is reversible then also $\mathcal{E}(f,g) = \mathcal{E}(g,f)$.

Finally, we recall some notation from complexity theory which will be used occasionally. Given positive functions $f, g : \mathsf{R}_+ \to \mathsf{R}_+$ we say that $f = O(g)$ if $f \le c\,g$ for some constant $c \ge 0$, while $f = \Omega(g)$ if $f \ge c\,g$ for a constant $c \ge 0$, and finally $f = \Theta(g)$ if $c_1\,g \le f \le c_2\,g$ for constants $c_1, c_2 \ge 0$. For instance, while attempting to analyze an algorithm requiring $\tau(n) = 3n^4 + n$ steps to terminate on input of size n, it might be found that $\tau(n) = O(n^5)$, or $\tau(n) = \Omega(n\log n)$, when in fact $\tau(n) = \Theta(n^4)$.

1.2 Continuous Time

Many mixing time results arise in a natural, clean fashion in the continuous time setting, and so we consider this case first. The arguments developed here will then point the way for our later consideration of discrete time results.

Let $\mathcal{L}$ denote the (discrete) Laplacian operator given by $\mathcal{L} = -(\mathsf{I}-\mathsf{P})$. Then for $t \ge 0$, $H_t = e^{t\mathcal{L}}$ represents the continuized chain [1] (or the heat kernel) corresponding to the discrete Markov kernel P. The continuized chain simply represents a Markov process $\{X_t\}_{t\ge 0}$ in Ω with initial distribution, μ_0 (say), and transition matrices

$$H_t = e^{-t(\mathsf{I}-\mathsf{P})} = \sum_{n=0}^{\infty}\frac{t^n\mathcal{L}^n}{n!} = e^{-t}\sum_{n=0}^{\infty}\frac{t^n\mathsf{P}^n}{n!}, \quad t \ge 0,$$

with the generator $\mathcal{L} = -(\mathsf{I} - \mathsf{P})$. Thus $H_t(x, y)$ denotes the probability that the rate one continuous Markov chain having started at x is at y at time t. Let $h_t^x(y) = H_t(x, y)/\pi(y)$, for each $y \in \Omega$, denote its density with respect to π at time $t \geq 0$, and $h_t(y)$ when the start state or the start distribution is unimportant or clear from the context. Also, let

$$H_t^* = e^{t\mathcal{L}^*} = \sum_{n=0}^{\infty} \frac{t^n (\mathcal{L}^*)^n}{n!}$$

be the semigroup associated to the dual $\mathcal{L}^* = -(\mathsf{I} - \mathsf{P}^*)$. The following is elementary and a useful technical fact.

Lemma 1.3. For any h_0 and all $t \geq 0$, $h_t = H_t^* h_0$. Consequently, for any $x \in \Omega$,

$$\frac{dh_t(x)}{dt} = \mathcal{L}^* h_t(x).$$

Using Lemma 1.3, the following lemma is easy to establish.

Lemma 1.4.

$$\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t) \tag{1.4}$$

$$\frac{d}{dt}\mathrm{Ent}(h_t) = -\mathcal{E}(h_t, \log h_t) \tag{1.5}$$

Proof. Indeed,

$$\begin{aligned} \frac{d}{dt}\mathrm{Var}(h_t) &= \int \frac{d}{dt} h_t^2 d\pi = 2\int h_t \mathcal{L}^* h_t d\pi \\ &= 2\int \mathcal{L}(h_t) h_t d\pi = -2\mathcal{E}(h_t, h_t). \end{aligned}$$

$$\begin{aligned} \frac{d}{dt}\mathrm{Ent}(h_t) &= \int \frac{d}{dt} h_t \log h_t d\pi = \int (\log h_t + 1)\mathcal{L}^* h_t d\pi \\ &= \int \mathcal{L}(\log h_t) h_t d\pi = -\mathcal{E}(h_t, \log h_t). \end{aligned}$$

□

The above motivates the following definitions of the spectral gap λ and the entropy constant ρ_0.

Definition 1.5. Let $\lambda > 0$ and $\rho_0 > 0$ be the optimal constants in the inequalities:

$$\lambda \mathrm{Var}_\pi f \leq \mathcal{E}(f,f), \quad \text{for all } f : \Omega \to \mathsf{R}.$$

$$\rho_0 \mathrm{Ent}_\pi f \leq \mathcal{E}(f, \log f), \quad \text{for all } f : \Omega \to \mathsf{R}_+. \tag{1.6}$$

When it is necessary to specify the Markov chain K being considered then use the notation λ_{K}.

Lemma 1.21 (Courant-Fischer theorem) shows that for a reversible Markov chain, the second largest eigenvalue λ_1 (of P) satisfies the simple relation $1 - \lambda_1 = \lambda$. However, reversibility is not needed for the following result.

Corollary 1.6. Let $\pi_* = \min_{x \in \Omega} \pi(x)$. Then, in continuous time,

$$\tau_2(\epsilon) \leq \frac{1}{\lambda}\left(\frac{1}{2}\log\frac{1-\pi_*}{\pi_*} + \log\frac{1}{\epsilon}\right) \tag{1.7}$$

$$\tau_{\mathsf{D}}(\epsilon) \leq \frac{1}{\rho_0}\left(\log\log\frac{1}{\pi_*} + \log\frac{1}{\epsilon}\right). \tag{1.8}$$

Proof. Simply solve the differential equations,

$$\frac{d}{dt}\mathrm{Var}(h_t^x) = -2\mathcal{E}(h_t^x, h_t^x) \leq -2\lambda\,\mathrm{Var}(h_t^x) \tag{1.9}$$

and

$$\frac{d}{dt}\mathrm{Ent}(h_t^x) = -\mathcal{E}(h_t^x, \log h_t^x) \leq -\rho_0\,\mathrm{Ent}(h_t^x), \tag{1.10}$$

and note that $\mathrm{Var}(h_0) \leq \frac{1-\pi_*}{\pi_*}$ and $\mathrm{Ent}(h_0) \leq \log\frac{1}{\pi_*}$ (e.g. by equation (1.1)). □

It is worth noting here that the above functional constants λ and ρ_0 indeed capture the rate of decay of variance and relative entropy, respectively, of H_t for $t > 0$:

Proposition 1.7. If $c > 0$ then

(a) $\mathrm{Var}_\pi(H_t f) \le e^{-ct}\mathrm{Var}_\pi f$, for all f and $t > 0$, if and only if $\lambda \ge c$.

(b) $\mathrm{Ent}_\pi(H_t f) \le e^{-ct}\mathrm{Ent}_\pi f$, for all $f > 0$ and $t > 0$, if and only if $\rho_0 \ge c$.

Proof. The "if" part of the proofs follows from (1.9) and (1.10). The only if is also rather elementary and we bother only with that of part (b): Starting with the hypothesis, we may say, for every $f > 0$, and for $t > 0$,

$$\frac{1}{t}\Big(\mathrm{Ent}_\pi(H_t f) - \mathrm{Ent}_\pi f\Big) \le \frac{1}{t}\Big(e^{-ct} - 1\Big)\mathrm{Ent}_\pi f.$$

Letting $t \downarrow 0$, we get $-\mathcal{E}(f, \log f) \le -c\mathrm{Ent}_\pi f$. □

While there have been several techniques (linear-algebraic and functional-analytic) to help bound the spectral gap, the analogous problem of getting good estimates on ρ_0 seems challenging. The following inequality relating the two Dirichlet forms introduced above also motivates the study of the classical logarithmic Sobolev inequality. In practice this is a much easier quantity to bound, and moreover it will later be shown to bound the stronger L^2 mixing time, and hence L^∞ as well.

Lemma 1.8. If $f \ge 0$ then

$$2\mathcal{E}(\sqrt{f}, \sqrt{f}) \le \mathcal{E}(f, \log f)$$

Proof. Observe that

$$a(\log a - \log b) = 2a \log \frac{\sqrt{a}}{\sqrt{b}} \ge 2a\left(1 - \frac{\sqrt{b}}{\sqrt{a}}\right) = 2\sqrt{a}\,(\sqrt{a} - \sqrt{b})$$

by the relation $\log c \ge 1 - c^{-1}$. Then

$$\begin{aligned}
\mathcal{E}(f, \log f) &= \sum_{x,y} f(x)(\log f(x) - \log f(y))\mathsf{P}(x,y)\pi(x) \\
&\ge 2\sum_{x,y} f^{1/2}(x)(f^{1/2}(x) - f^{1/2}(y))\mathsf{P}(x,y)\pi(x) \\
&= 2\mathcal{E}(\sqrt{f}, \sqrt{f})
\end{aligned}$$

□

Let $\rho_{\mathsf{P}} > 0$ denote the logarithmic Sobolev constant of P defined as follows.

Definition 1.9.

$$\rho = \rho_{\mathsf{P}} = \inf_{\mathrm{Ent} f^2 \neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Ent} f^2}.$$

Proposition 1.10. For every irreducible chain P,

$$2\rho \leq \rho_0 \leq 2\lambda\,.$$

Proof. The first inequality is immediate, using Lemma 1.8. The second follows from applying (1.6) to functions $f = 1 + \epsilon g$, for $g \in L^2(\pi)$ with $\mathbb{E}_\pi g = 0$. Assume $\epsilon \ll 1$, so that $f \geq 0$. Then using the Taylor approximation, $\log(1+\epsilon g) = \epsilon g - 1/2(\epsilon)^2 g^2 + o(\epsilon^2)$, we may write

$$\mathrm{Ent}_\pi(f) = \frac{1}{2}\epsilon^2 \pi(g^2) + o(\epsilon^2),$$

and

$$\mathcal{E}(f, \log f) = -\epsilon \mathbb{E}_\pi((\mathcal{L}g)\log(1+\epsilon g)) = \epsilon^2 \mathcal{E}(g,g) + o(\epsilon^2).$$

Thus starting from (1.6), and applying to f as above, we get

$$\epsilon^2 \mathcal{E}(g,g) \geq \frac{\rho_0}{2}\epsilon^2 \mathbb{E}_\pi g^2 + o(\epsilon^2).$$

Canceling ϵ^2 and letting $\epsilon \downarrow 0$, yields the second inequality of the proposition, since $\mathbb{E}_\pi g = 0$.

□

Remark 1.11. The relation $2\rho \leq 2\lambda$ found in the lemma can be strengthened somewhat to $\rho \leq \lambda/2$, by a direct application of the method used above. Under the additional assumption of reversibility, the inequality in Lemma 1.8 can be strengthened by a factor of 2 to match this, as explained in [25], in turn improving the above proposition to $4\rho \leq \rho_0 \leq 2\lambda$ for reversible chains.