Chapter 4

Markov chain Monte Carlo

The reality is that for most serious sampling problems one cannot generate independent samples from the target density π, nor can one generate samples from a reference density $\tilde{\pi}$ that is close enough to π to result in a reasonable importance sampling estimator. Under these conditions one needs to consider more general sequences of random variables. The logical first generalization is the replacement of sequences of independent random variables by Markov processes.

4.1 Markov processes

For our purposes, a Markov process is a random sequence $X^{(t)}$ with $t \in \mathbb{Z}$ or $t \in \mathbb{R}$, such that, for any $B \in \mathbb{R}^d$,

$$\mathbf{P}\left[X^{(t)} \in B \,|\, \mathcal{F}_s\right] = \mathbf{P}\left[X^{(t)} \in B \,|\, X^{(s)}\right] \quad \text{for all } t \geq s,$$

where $\mathcal{F}_t$ is an increasing sequence of σ-algebras (called a filtration) and $X^{(t)}$ is $\mathcal{F}_t$ measurable. When $t \in \mathbb{Z}$ we will refer to X as a Markov chain and when $t \in \mathbb{R}$ we will sometimes refer to it as a continuous time Markov process.

Markovianity can be expressed very simply as the requirement that, conditioned on the present value of $X^{(t)}$, its future and past values are independent. More precisely, for all t, conditioned on $\sigma\left(X^{(t)}\right)$, the σ-algebras

$\sigma\left(\{X^{(s)}\}_{s\leq t}\right)$ and $\sigma\left(\{X^{(s)}\}_{s\geq t}\right)$ are independent. To see this suppose that $A_- \in \sigma\left(\{X^{(s)}\}_{s\leq t}\right)$ and $A_+ \in \sigma\left(\{X^{(s)}\}_{s\geq t}\right)$ and note that by the tower property and the definition of conditional expectations,

$$\begin{aligned}\mathbf{P}\left[A_- \cap A_+ \,|\, X^{(t)}\right] &= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{A_-\cap A_+} \,|\, \{X^{(s)}\}_{s\leq t}\right] \,|\, X^{(t)}\right] \\ &= \mathbf{E}\left[\mathbf{1}_{A_-}\mathbf{P}\left[\mathbf{1}_{A_+} \,|\, \{X^{(s)}\}_{s\leq t}\right] \,|\, X^{(t)}\right].\end{aligned}$$

Markovianity of $X^{(t)}$ then implies that

$$\begin{aligned}\mathbf{P}\left[A_- \cap A_+ \,|\, X^{(t)}\right] &= \mathbf{E}\left[\mathbf{1}_{A_-}\mathbf{P}\left[\mathbf{1}_{A_+} \,|\, X^{(t)}\right] \,|\, X^{(t)}\right] \\ &= \mathbf{P}\left[\mathbf{1}_{A_+} \,|\, X^{(t)}\right]\mathbf{P}\left[\mathbf{1}_{A_-} \,|\, X^{(t)}\right].\end{aligned}$$

Incidentally, this observation implies that if $X^{(t)}$ is a Markov process and if for $t \in \mathbb{Z}$ we define the filtration $\mathcal{G}_t = \sigma\left(\{X^{(s)}\}_{s\geq -t}\right)$ then the process $Y^{(t)} = X^{(-t)}$ is a Markov process with respect to $\mathcal{G}_t$. Despite this, the law governing the evolution of $Y^{(t)}$ can be very different than the law governing the evolution of $X^{(t)}$. We'll return to this point again later.

Example 12. *The solutions of the ordinary differential equation*

$$\frac{d}{dt}y^{(t)} = b(t, y^{(t)})$$

are (continuous time) Markov processes since the distribution of $y^{(t)}$ (in this case a delta function) is completely determined by the value of $y^{(s)}$ for any $(s \leq t)$.

Example 13. *The solutions of the ordinary differential equation*

$$\frac{d^2}{dt^2}y^{(t)} = b(t, y^{(t)})$$

are not Markov processes. Defining $v = \frac{d}{dt}y$, this second order ODE can be rewritten as a system of first order ODE,

$$\frac{d}{dt}\begin{pmatrix} y^{(t)} \\ v^{(t)} \end{pmatrix} = \begin{pmatrix} v^{(t)} \\ b(t, y^{(t)}) \end{pmatrix}.$$

From this equation (and the uniqueness of solutions of ODE) it's clear that knowledge of $y^{(s)}$ at some initial time, s, is not enough to determine the

distribution of $y^{(t)}$ at future times $t > s$. One must also know $v^{(s)}$ at the initial time and, since $v^{(s)}$ is the derivative of y at time s, knowledge of v at time s is equivalent to knowledge of y at least at times close to but slightly less than s. The pair (y, v) is a Markov process.

As in this example, projection of a higher dimensional Markov process into a lower dimensional space usually results in a process that is no longer Markovian.

Example 14. *The simple random walk $X^{(k+1)} = X^{(k)} + \xi^{(k)}$ on the periodic lattice $\mathbb{Z}_L = \{0, 1, \ldots, L-1\}$ with independent $\xi^{(k)}$ distributed according to*

$$\mathbf{P}[\xi = 1] = \mathbf{P}[\xi = -1] = \mathbf{P}[\xi = 0] = \frac{1}{3}$$

is a Markov chain

It will be convenient to work with the distribution of a specific Markov chain directly rather than to work with the Markov chain itself and the original probability measure $\mathbf{P}$. For a specific Markov chain X for which we assume $X^{(0)}$ is drawn from a distribution μ, the distribution of X is a probability measure on the space of infinite sequences $x^{(0)}, x^{(1)}, \ldots$ which we will denote by $\mathbb{R}^\infty$. We can define that probability measure, which we will denote P_μ, by requiring that, for any subset $A \subset \mathbb{R}^\infty$ of the form

$$A = A_0 \times A_1 \times \cdots \times A_k \times \mathbb{R} \times \mathbb{R} \times \cdots$$

for subsets $A_i \subset \mathcal{B}$,

$$P_\mu[A] = \mathbf{P}\left[X^{(0)} \in A_0,\, X^{(1)} \in A_1, \ldots,\, X^{(k)} \in A_k\right].$$

When the initial distribution μ is a delta function at a single point x, we write P_x instead of P_{δ_x}. These definitions are straightforwardly generalized to define a probability measure, $P_{k,\mu}$ on $\mathbb{R}^{k,\infty}$ of infinite sequences whose first index is $t \in \mathbb{Z}$ (i.e. $x^{(k)}, x^{(k+1)}, \ldots$), and which is the probability distribution of a Markov chain for which we assume that $X^{(k)}$ is drawn from μ. We will denote expectations with respect to $P_{k,\mu}$ by $E_{k,\mu}$, i.e. a function $F : \mathbb{R}^{k,\infty} \to \mathbb{R}$ of the path of the Markov chain is now a random variable with expectation

$$E_{k,\mu}[F] = \int F(x^{(k)}, x^{(k+1)}, \ldots) P_{k,\mu}\left(dx^{(k)} \times dx^{(k+1)} \times \cdots\right).$$

In fact, for our purposes, we can completely forget about the original Markov chain and redefine the symbol $X^{(\ell)}$ to be a function taking values in $\mathbb{R}^{k,\infty}$ and returning values in $\mathbb{R}$ according to the formula

$$X^{(\ell)}(x^{(k)}, x^{(k+1)}, x^{(k+2)}, \dots) = x^{(\ell)},$$

i.e. the projection of the sequence onto the coordinate corresponding to index ℓ in the chain. With the appropriate definition of a σ-algebra on $\mathbb{R}^{k,\infty}$, the functions $X^{(\ell)}$ are again random variables. We will alternate between the two probability spaces, using whichever is more convenient to express any particular relation.

When it exists, the function

$$p(k+1, y \,|\, k, x) = \lim_{|dy| \to 0} \frac{P_{k,x}\left[X^{(k+1)} \in dy\right]}{|dy|}$$

is called the transition probability density for the chain. When the goal is to compute averages with respect to a given probably π, the process $X^{(k)}$ is often time-homogenous, i.e. the probability measure $P_{k,x}\left[X^{(k+1)} \in A\right]$ is independent of k, in which case we write the transition density (assuming it exists) as $p(y \,|\, x)$. All of the methods introduced in these notes can be analyzed as time-homogenous Markov processes and, unless otherwise specified we will assume that the chains we consider have that property.

It is also useful to define the transition operator $\mathcal{T}$ which operates on functions $f : \mathbb{R} \to \mathbb{R}$ from the right by

$$\mathcal{T}f(x) = E_x\left[f(X^{(1)})\right] = \int f(y) p(dy \,|\, x).$$

The action of $\mathcal{T}$ and on a distribution (or density) from the left, $\mu\mathcal{T}$ is defined by the requirement

$$\int f(x)[\mu\mathcal{T}](dx) = \int [\mathcal{T}f](x)\mu(dx) \tag{4.1}$$

for a sufficiently large set of test functions f. In other words, when μ is a density, $\mu\mathcal{T} = \mathcal{T}^*\mu$ where $\mathcal{T}^*$ is the adjoint of $\mathcal{T}$ in the inner product $\langle f, g\rangle = \int f(x)g(x)dx$. By choosing f in (4.1) to be the indicator function

of some set A (or perhaps a smooth approximation of that function), we see that

$$\mu\mathcal{T}(A) = P_\mu \left[X^{(1)} \in A\right] = \int_{\{x \in A\}} \int p(dx \,|\, y)\mu(dy).$$

Consistent with our notation so far, we will use the same symbol $\mu\mathcal{T}$ to denote the density (when it exists) of the distribution $\mu\mathcal{T}$, distinguishing the two only by the whether the argument is a set or a point. When μ is a density we will use the symbol $\mu\mathcal{T}$ to denote the distribution (or density) resulting from applying $\mathcal{T}$ to the distribution corresponding to μ. If we define the operator $\mathcal{T}^k$ by $\mathcal{T}^k f(x) = E_x \left[f(X^{(k)})\right]$ then, by the tower property of conditional expectations,

$$\begin{aligned}\mathcal{T}^j\mathcal{T}^k f(x) = E_x \left[E_{X^{(j)}} \left[f(X^{(k)})\right]\right] &= \mathbf{E}\left[\mathbf{E}\left[f(X^{(k+j)}) \,|\, X^{(j)}\right] \,|\, X^{(0)} = x\right] \\ &= \mathcal{T}^{k+j} f(x)\end{aligned}$$

so that in particular $\mathcal{T}^k = \mathcal{T}\mathcal{T}^{k-1} = \mathcal{T}^{k-1}\mathcal{T}$. In words, $\mathcal{T}^k$ is the kth power of the operator $\mathcal{T}$. The distribution of $X^{(k)}$ given that $X^{(0)}$ was drawn from μ is $\mu\mathcal{T}^k$.

Example 15. *Consider a Markov chain on a finite state space E. We might as well assume that the state space is $E = \{1, 2, \ldots, n\}$. Because the state space is finite, probability distributions and functions on this state space can be viewed as n-dimensional vectors. The action of $\mathcal{T}$ on a test function $f \in \mathbb{R}^n$ can then be viewed as multiplication on the left of f by the matrix T with entries*

$$T_{ij} = P_i \left[X^{(1)} = j\right].$$

Likewise, the action of $\mathcal{T}$ on a probability vector $\mu \in \mathbb{R}^n$ can be viewed as multiplication on the right of μ by T. The transition operator $\mathcal{T}$ can then be identified with T.

As before, given an objective function f, we will construct the estimator

$$\overline{f}_N = \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)})$$

of $\pi[f]$. The estimator $\overline{f}_N$ will no longer be unbiased but we still hope that

$$\overline{f}_N \to \pi[f]. \tag{4.2}$$

An alternative and closely related notion of ergodicity is the requirement that

$$\lim_{k\to\infty} \mu\mathcal{T}^k = \pi \tag{4.3}$$

for any initial distribution π, i.e. that $X^{(k)}$ converges to π in distribution. Convergence of $\overline{f}_N$ to $\pi[f]$ does not require (4.3). And convergence in distribution of $X^{(k)}$ to π implies no additional computational advantages in Monte Carlo simulation. Nonetheless, in practical situations it is difficult to construct a chain that satisfies (4.2) without also satisfying (4.3).

Example 16. *Consider sampling the uniform measure on $\mathbb{Z}_L$ by the Markov chain $X^{(k)}$ with*

$$P_i\left[X^{(1)} = (i+1) \bmod L\right] = 1,$$

i.e. $X^{(k)}$ moves to the right with every step. For convenience, let's assume that $X^{(0)} = L-1$. The estimator satisfies

$$\begin{aligned}\overline{f}_N = \frac{1}{N}\sum_{k=1}^{N} f(X^{(k)}) &= \frac{1}{N}\sum_{k=0}^{N-1} f\left(k \bmod L\right) \\ &= \frac{1}{N}\left(\left\lfloor\frac{N}{L}\right\rfloor\sum_{i=0}^{L-1} f(i) + \sum_{i=0}^{N \bmod L} f(i)\right) \\ &= \frac{1}{L}\sum_{i=0}^{L-1} f(i) + \mathcal{O}\left(\frac{L}{N}\right).\end{aligned}$$

The error for the estimator built from this Markov chain is $\mathcal{O}\left(1/N\right)$ and not our usual $\mathcal{O}\left(1/\sqrt{N}\right)$. From the Monte Carlo point of view the Markov chain $X^{(k)}$ is a fantastic choice.

On the other hand,

$$P_i\left[X^{(k)} = j\right] = \begin{cases} 1 & \text{if } j = (i+k) \bmod L \\ 0 & \text{otherwise} \end{cases}$$

is periodic in t and can never converge to anything, much less the uniform measure on $\mathbb{Z}_L$. So the chain does not satisfy (4.3).

Exercise 30. *Show that the eigenvalues of the transition matrix T for the process in the last example are the L roots of unity $e^{i2\pi\ell/L}$ for $\ell = 0, \ldots, L-$*

1. *What are the eigenvalues of the matrix T^k for any k? What are the eigenvalues of the matrix*

$$F = \frac{1}{L}\sum_{k=1}^{L} T^k?$$

As the previous example and exercise demonstrate, eigenvalues of the transition operator on the complex unit circle that are not equal to 1 prevent ergodicity in the sense of (4.3), but are not necessarily a problem for (4.2) (in fact they can improve convergence in (4.2)). As the next example shows, (4.2) will fail if the eigenvalue 1 has multiplicity more than one.

Example 17. *Consider sampling the uniform measure on $\mathbb{Z}_L$ for L even, by the Markov chain $X^{(k)}$ with*

$$P_i\left[X^{(1)} = (i+2) \bmod L\right] = 1,$$

i.e. $X^{(k)}$ moves to the right with every step. Notice that this Markov chain is not irreducible: if it starts on an even site it will never visit an odd site (the reverse statement is also true). To see the consequence of this, notice that the probability distribution π_e that is equal to $2/L$ on every even indexed site and 0 on every odd indexed site is invariant. The distribution π_o that is equal to $2/L$ on every odd indexed site and 0 on every even indexed site is also invariant. In fact, any probability vector obtained by

$$\alpha\pi_e + (1-\alpha)\pi_o$$

for $\alpha \in [0,1]$, is also an invariant probability vector.

In practice, when constructing a transition rule $p(y \mid x)$ for a time-homogenous Markov chain satisfying (4.2), the first criterion that must be kept in mind is invariance of the target distribution, i.e.

$$\pi\mathcal{T} = \pi. \tag{4.4}$$

Note that any transition operator $\mathcal{T}$ may have more than one invariant density. Some transition operators may have invariant measures that are not probability measures. In particular, it may happen that for some non-negative function p, $p\mathcal{T} = p$ but $\int p(x)dx = \infty$. As we will see in the next section, the assumption that $\mathcal{T}$ has a unique invariant measure and that that measure is a probability measure is already fairly powerful.

Example 18. *Consider the simple random walk on* $\mathbb{Z}$, $X^{(k+1)} = X^{(k)} + \xi^{(k)}$ *with independent* $\xi^{(k)}$ *distributed according to*

$$\mathbf{P}\left[\xi = 1\right] = \mathbf{P}\left[\xi = -1\right] = \mathbf{P}\left[\xi = 0\right] = \frac{1}{3}.$$

Note that this chain differs from the one in Example 14 in that its state space is not periodic. One can easily check that the distribution $\mu(i) = 1$ *is invariant for this chain, but cannot be normalized and written as a probability distribution.*

It is intuitively clear from either expression (4.2) or (4.3) that one must also ensure that the chain can visit all sets of non-zero π-probability and that it does so sufficiently often. The chain is called irreducible if, for every x and every set B of positive π-probability,

$$P_x\left[X^{(m)} \in B\right] > 0 \quad \text{for some} \quad m \in \mathbb{N}, \tag{4.5}$$

i.e. the chain can get from anywhere to anywhere.

In fact, to satisfy (4.2) or (4.3), it is not enough for the chain to preserve π and be irreducible (assuming the state space is infinite). One must show that, not only can the process reach every region in space, but it does so infinitely often. We will return to considering the convergence of Markov chain Monte Carlo methods later. For now we focus on conditions ensuring the invariance of π. That property and irreducibility are the most that can be rigorously guaranteed in most practical applications.

We will see that, given a density π, it is usually not hard to construct a chain satisfying (4.4) and (4.5). In most, but not all cases, in order to satisfy (4.4), one constructs a Markov Chain satisfying the so-called detailed balance condition,

$$P_\pi\left[X^{(1)} \in B_1 \text{ and } X^{(0)} \in B_0\right] = P_\pi\left[X^{(1)} \in B_0 \text{ and } X^{(0)} \in B_1\right] \tag{4.6}$$

for any pair of sets $B_0, B_1 \in \mathcal{B}$. Indeed, if condition (4.6) is satisfied then

$$\begin{aligned} P_\pi\left[X^{(1)} \in B\right] &= P_\pi\left[X^{(1)} \in B \text{ and } X^{(0)} \in B\right] \\ &\qquad + \mathbf{P}_\pi\left[X^{(1)} \in B \text{ and } X^{(0)} \notin B\right] \\ &= P_\pi\left[X^{(1)} \in B \text{ and } X^{(0)} \in B\right] \\ &\qquad + P_\pi\left[X^{(1)} \notin B \text{ and } X^{(0)} \in B\right] \\ &= P_\pi\left[X^{(0)} \in B\right] \end{aligned}$$

In terms of densities the detailed balance condition becomes,

$$p(y \,|\, x)\pi(x) = p(x \,|\, y)\pi(y) \tag{4.7}$$

and in terms of expectations of test functions, detailed balance becomes the requirement that

$$\int g(x)(\mathcal{T}f(x))\pi(dx) = \int f(x)(\mathcal{T}g(x))\pi(dx) \tag{4.8}$$

for continuous and bounded functions f and g.

Exercise 31. *By formally plugging* $f(x) = |dy|^{-1}\mathbf{1}_{\{dy\}}(x)$ *and* $g = |dx|^{-1}\mathbf{1}_{\{dx\}}(x)$ *derive* (4.7) *from the detailed balance condition in* (4.8).

Exercise 32. *Suppose that* $X^{(t)}$ *is a Markov process on a finite state space and that* $\mathcal{T}$ *is a symmetric matrix. Interpret these assumptions in terms of detailed balance.*

To better understand the full strength (and restrictiveness) of the detailed balance condition, let's consider what it says about transitions between sets in any partition of space, $\{B_i\}_{i=0}^{\infty}$ of space with $B_i \in \mathcal{B}$, $B_i \cap B_j = \{\}$ for $i \neq j$, and $\cup_{i=0}^{\infty} B_i = \mathbb{R}$. In terms of this partition, the detailed balance condition (4.6) becomes

$$P_\pi\left[X^{(1)} \in B_j \text{ and } X^{(0)} \in B_i\right] = P_\pi\left[X^{(1)} \in B_i \text{ and } X^{(0)} \in B_j\right].$$

Condition (4.4), on the other hand, implies that

$$P_\pi\left[X^{(1)} \in B_j\right] = P_\pi\left[X^{(0)} \in B_j\right]$$

which can be rewritten as

$$\begin{aligned} &P_\pi\left[X^{(1)} \in B_j \text{ and } X^{(0)} \in B_j\right] + P_\pi\left[X^{(1)} \in B_j \text{ and } X^{(0)} \notin B_j\right] \\ &\qquad = P_\pi\left[X^{(1)} \in B_j \text{ and } X^{(0)} \in B_j\right] + P_\pi\left[X^{(1)} \notin B_j \text{ and } X^{(0)} \in B_j\right]. \end{aligned}$$

Canceling the like term on both sides of the last equation and expanding the events $X^{(0)} \notin B_j$ and and $X^{(1)} \notin B_j$ in terms of the other sets in the

partition we obtain

$$\sum_{\substack{i\geq 1\\ i\neq j}} P_\pi\left[X^{(1)}\in B_i \text{ and } X^{(0)}\in B_j\right] = \sum_{\substack{i\geq 1\\ i\neq j}} P_\pi\left[X^{(1)}\in B_j \text{ and } X^{(0)}\in B_i\right] \quad (4.9)$$

This equation simply says that the total probably flux out of B_j and into all the other sets is equal to the total probability flux into B_j from all the other sets. Detailed balance on the other hand requires that the summands on both sides are equal, i.e. it requires that the flux between any pair of sets balances.

Example 19. *The deterministic chain in Example 16 does not satisfy detailed balance while the random chain in Example 14 does. So not only is detailed balance not required, it may lead to worse Monte Carlo estimators.*

To describe the next concept is is useful to extend our notation for the distribution of a Markov chain to chains initiated at $k=-\infty$, i.e. we define a probability measure $P_{-\infty,\pi}$ on the set of all bi-infinite sequences $\mathbb{R}^{\pm\infty}$ of the form $\ldots, x^{(-2)}, x^{(-1)}, x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ by setting

$$P_{-\infty,\pi} = \lim_{k\to\infty} P_{-k,\pi}$$

which will exist as long as π is an invariant measure for the Markov chain. Note that this new distribution inherits from the $P_{k,\pi}$ the property that, for any $k\in\mathbb{Z}$,

$$P_{-\infty,\pi}\left[X^{(k)}\in A\right] = \pi(A).$$

With this definition, notice that for any pair $\ell < k$, Markovianity of $X^{(k)}$ implies that

$$P_{-\infty,\pi}\left[X^{(\ell)}\in A_\ell,\ldots,X^{(k)}\in A_k\right] = P_{-\infty,\pi}\left[X^{(k)}\in A_k \,|\, X^{(k-1)}\in A_{k-1}\right] \times P_{-\infty,\pi}\left[X^{(\ell)}\in A_\ell,\ldots,X^{(k-1)}\in A_{k-1}\right].$$

Repeating this decomposition we find that

$$\begin{aligned} P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell, \ldots, X^{(k)} \in A_k\right] &= P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell\right] \\ &\quad \times \prod_{r=\ell+1}^{k} P_{-\infty,\pi}\left[X^{(r)} \in A_r \,|\, X^{(r-1)} \in A_{r-1}\right] \\ &= P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell\right] \\ &\quad \times \prod_{r=\ell+1}^{k} \frac{P_{-\infty,\pi}\left[X^{(r)} \in A_r,\, X^{(r-1)} \in A_{r-1}\right]}{P_{-\infty,\pi}\left[X^{(r-1)} \in A_{r-1}\right]}. \end{aligned}$$

If the process satisfies detailed balance then this expression can be rewritten as

$$\begin{aligned} P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell, \ldots, X^{(k)} \in A_k\right] &= P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell\right] \\ &\quad \times \prod_{r=\ell+1}^{k} \frac{P_{-\infty,\pi}\left[X^{(r-1)} \in A_r,\, X^{(r)} \in A_{r-1}\right]}{P_{-\infty,\pi}\left[X^{(r-1)} \in A_{r-1}\right]} \end{aligned}$$

which, after recalling that for any set A, since X preserves π, $P_{-\infty,\pi}\left[X^{(r)} \in A\right]$ is independent of r, can itself be rewritten as

$$\begin{aligned} P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell, \ldots, X^{(k)} \in A_k\right] &= P_{-\infty,\pi}\left[X^{(k)} \in A_k\right] \\ &\quad \times \prod_{r=\ell+1}^{k} \frac{P_{-\infty,\pi}\left[X^{(r-1)} \in A_r,\, X^{(r)} \in A_{r-1}\right]}{P_{-\infty,\pi}\left[X^{(r)} \in A_{r-1}\right]}. \end{aligned}$$

The last expression implies that

$$P_{-\infty,\pi}\left[X^{(\ell)} \in A_\ell, \ldots, X^{(k)} \in A_k\right] = P_{-\infty,\pi}\left[X^{(k)} \in A_\ell, \ldots, X^{(\ell)} \in A_k\right], \tag{4.10}$$

i.e. that if the Markov process $X^{(k)}$ satisfies detailed balance then, under $P_{-\infty,\pi}$, the process $Y^{(k)} = X^{(-k)}$ has exactly the same distribution as $X^{(k)}$. In other words, the distribution of $X^{(k)}$ is the same whether it is run forward or in reverse. A process $X^{(k)}$ with invariant measure π satisfying (4.10) is said to be reversible with respect to π. It is clear that if $X^{(k)}$ is reversible with respect to π then it satisfies detailed balance with respect to π. On the other hand, we have just shown that if $X^{(k)}$ satisfies detailed balance with respect to π then it is reversible with respect to π so that the two conditions are equivalent.

4.2 Generators

The generator $\mathcal{L}$ of a discrete time Markov process with transition operator $\mathcal{T}$, is

$$\mathcal{L} = \mathcal{T} - \mathcal{I}$$

where $\mathcal{I}$ is the identity operator. In terms of the generator, π is invariant if

$$0 = \int (\mathcal{L}f)(x)\pi(dx) = \int f(x)(\pi\mathcal{L})(dx)$$

for all test functions f. This implies that π is invariant when $\pi\mathcal{L} = 0$. The process $X^{(t)}$ is reversible if its generator satisfies

$$\int g(x)(\mathcal{L}f)(x)\pi(dx) = \int f(x)(\mathcal{L}g)(x)\pi(dx).$$

Notice that we can always write

$$f(X^{(k)}) = f(X^{(0)}) + \sum_{\ell=0}^{k-1} \mathcal{L}f(X^{(\ell)}) + M^{(k)}$$

where we have defined

$$M^{(k)} = \sum_{\ell=0}^{k-1} f(X^{(\ell+1)}) - \mathcal{T}f(X^{(\ell)}).$$

The process $M^{(k)}$ has the special property that, if $\mathcal{F}_k$ is the σ-algebra generated by $X^{(0)}, X^{(1)}, \ldots, X^{(k)}$, then, for any $\ell \leq k$

$$\mathbf{E}\left[M^{(k)} \,|\, \mathcal{F}_\ell\right] = M^{(\ell)}.$$

Processes with this property are called martingales, and play an important role in the theory of stochastic processes (see Chapter ??).

Because the generator maps constant functions to 0, it is not invertible in any space that includes non-zero constant functions. However, if we restrict the space of functions we consider to functions f with $\pi[f] = 0$ then there is a chance that $\mathcal{L}$ can be invertible. In fact, if we assume that for some $\alpha \in [0, 1)$, $\|\mathcal{T}(f - \pi[f])\| \leq \alpha\|f - \pi[f]\|$ (in a norm of your choosing) then

$$\left\| \sum_{k=0}^{N-1} \mathcal{T}^k (f - \pi[f]) \right\| \leq \sum_{k=0}^{N-1} \alpha^k \|f - \pi[f]\| \leq \frac{1}{1-\alpha}\|f - \pi[f]\|$$

and the infinite sum $\sum_{k=0}^{\infty} \mathcal{T}^k(f - \pi[f])$ is absolutely convergent as long as $\|f\| < \infty$. On the other hand, as one can verify,

$$-\mathcal{L}\left(\sum_{k=0}^{N-1} \mathcal{T}^k\right) = -\left(\sum_{k=0}^{N-1} \mathcal{T}^k\right)\mathcal{L} = I - \mathcal{T}^N$$

which, taking the large N limit, implies that $\mathcal{L}$ is invertible and

$$-\mathcal{L}^{-1} = \sum_{k=0}^{\infty} \mathcal{T}^k. \tag{4.11}$$

In fact, in this case, $\mathcal{L}^{-1}$ satisfies

$$\|\mathcal{L}^{-1}(f - \pi[f])\| \leq \frac{1}{1-\alpha}\|f - \pi[f]\|.$$

Note that the condition $\|\mathcal{T}(f - \pi[f])\| \leq \alpha\|f - \pi[f]\|$ for some $\alpha \in [0, 1)$ implies that

$$\|\mathcal{T}^k f - \pi[f]\| = \|\mathcal{T}^k(f - \pi[f])\| \leq \alpha^k\|f - \pi[f]\|.$$

Since this holds for all test functions, we can conclude that $X^{(k)}$ is ergodic.

We have just seen that assumptions about the ergodicity of $X^{(k)}$ can imply the existence of solutions to the equation $\mathcal{L}u = f - \pi[f]$. The properties of this solution, on the other hand, can also tell us about the ergodicity of $X^{(k)}$. For example, suppose that the solution, u, is bounded. In this case, the left hand side of the identity

$$\begin{aligned}\frac{u(X^{(N+1)}) - u(X^{(0)}) - \mathcal{L}u(X^{(0)})}{N} &= \frac{1}{N}\sum_{\ell=1}^{N} \mathcal{L}u(X^{(k)}) + \frac{1}{N}M^{(N+1)} \\ &= \overline{f}_N - \pi[f] + \frac{1}{N}M^{(N+1)}\end{aligned}$$

is $\mathcal{O}(N^{-1})$. Observe that

$$\begin{aligned}\frac{1}{N^2}\mathbf{E}\left[\left(M^{(N+1)}\right)^2\right] &= \frac{1}{N^2}\sum_{k=0}^{N} \mathbf{E}\left[\left(u(X^{(k+1)}) - \mathcal{T}u(X^{(k)})\right)^2\right] \\ &\quad + \frac{2}{N^2}\sum_{\ell<k\leq N} \mathbf{E}\left[\left(u(X^{(\ell+1)}) - \mathcal{T}u(X^{(\ell)})\right)\left(u(X^{(k+1)}) - \mathcal{T}u(X^{(k)})\right)\right].\end{aligned}$$

For each k letting $\mathcal{F}_k$ be the σ-algebra generated by $X^{(0)}, X^{(1)}, \ldots, X^{(k)}$ and noting that, for $\ell < k$,

$$\begin{aligned}
&\mathbf{E}\left[\left(u(X^{(\ell+1)}) - \mathcal{T}u(X^{(\ell)})\right)\left(u(X^{(k+1)}) - \mathcal{T}u(X^{(k)})\right)\right] \\
&\quad = \mathbf{E}\left[\left(u(X^{(\ell+1)}) - \mathcal{T}u(X^{(\ell)})\right)\mathbf{E}\left[u(X^{(k+1)}) - \mathcal{T}u(X^{(k)}) \,|\, \mathcal{F}_{\ell+1}\right]\right] \\
&\quad = 0
\end{aligned}$$

we see that the second sum in the expression for $\mathbf{E}\left[\left(M^{(N+1)}\right)^2\right]$ vanishes exactly. Using the fact that u is bounded, we find that

$$\frac{1}{N^2}\mathbf{E}\left[\left(M^{(N+1)}\right)^2\right] = \mathcal{O}\left(N^{-1}\right).$$

Thus we see that if u is bounded,

$$\mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2\right] = \mathcal{O}\left(N^{-1}\right).$$

Exercise 33. *Compute the generator for chain in Example 14 and show that it is reversible with respect to the uniform measure on the lattice.*

The generator is a particularly convenient tool when working with continuous time Markov processes. In that context we will use the notation

$$\mathcal{T}^t f(x) = E_x\left[f(X^{(t)})\right]$$

which resembles our notation for discrete time processes, but note that t is now a real number and not an integer. In the continuous time setting we will define the generator by

$$\mathcal{L}f(x) = \lim_{h\to 0}\frac{\mathcal{T}^h f(x) - f(x)}{h} = \frac{d}{dt}\mathcal{T}^t f(x)\Big|_{t=0}.$$

As for the transition operator, when μ is a distribution (or density), we define the distribution (or density) $\mu\mathcal{L}$ by the requirement that

$$\int f(x)[\mu\mathcal{L}](dx) = \int [\mathcal{L}f](x)\mu(dx)$$

so that in particular, when μ is a density, $\mu\mathcal{L} = \mathcal{L}^*\mu$ where $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$ in the $\langle f, g\rangle = \int f(x)g(x)dx$ inner product.

If we let $u(t,x) = \mathcal{T}^t f(x) = E_x\left[f(X^{(t)}\right]$ and observe that the tower property of conditional expectations implies that for any $h > 0$,

$$\mathcal{T}^t\mathcal{T}^h f(x) = \mathbf{E}_x\left[\mathbf{E}_{X^{(h)}}\left[f(X^{(t+h)})\right]\right] = \mathcal{T}^{t+h} f(x),$$

then the definition of $\mathcal{L}$ implies

$$\partial_t u(t,x) = \lim_{h\to 0} \frac{\mathcal{T}^{t+h} f(x) - \mathcal{T}^t f(x)}{h} = \lim_{h\to 0} \frac{\mathcal{T}^h u(t,x) - u(t,x)}{h} = \mathcal{L}u(t,x)$$

Solving this (infinite dimensional) linear ordinary differential equation (informally) justifies the expression

$$\mathcal{T}^t f(x) = e^{t\mathcal{L}} f(x).$$

In particular, we observe that the generator of the process governs the distribution of the process.

Example 20. *The generator of the d-dimensional ODE*

$$\frac{d}{dt} y^{(t)} = b(y^{(t)})$$

is defined by its action on test functions, f,

$$\mathcal{L}f(x) = \frac{d}{dt} f(y^{(t)})\Big|_{t=0} = \sum_{i=0}^{d-1} b_i(x)\, \frac{\partial}{\partial_{x_i}} f(x).$$

So

$$\mathcal{L} = \sum_{i=0}^{d-1} b_i(x)\, \frac{\partial}{\partial_{x_i}}.$$

This is the famous Liouvillian operator.

Exercise 34. *Compute the adjoint of the Liouvillian.*

In the continuous time context, if $\|\mathcal{T}^t(f - \pi[f])\| \leq \alpha^t \|f - \pi[f]\|$ for some $\alpha \in [0,1)$, then

$$-\mathcal{L}^{-1} = \int_0^\infty \mathcal{T}^t dt$$

is a bounded operator on functions with bounded norm and zero mean under π.

4.3 Convergence

In this section we briefly consider conditions under which (4.2) and stronger convergence results hold. We begin by recalling the famous Birkhoff Ergodic Theorem. That theorem addresses the ergodicity of sequences generated by measure preserving maps. On a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ a map $\theta : \Omega \to \Omega$ preserves the measure $\mathbf{P}$ if, for any event $A \in \mathcal{F}$,

$$\mathbf{P}\left[\{\omega : \theta(\omega) \in A\}\right] = \mathbf{P}\left[A\right].$$

Birkhoff's Ergodic Theorem tells us that, if X is any random variable satisfying $\mathbf{E}\left[|X|\right] < \infty$, then on a set of $\omega \in \Omega$ with probability 1,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} X(\theta^k(\omega)) = \mathbf{E}\left[X \,|\, \mathcal{I}\right](\omega) \tag{4.12}$$

where the σ-algebra $\mathcal{I}$ is generated by the invariant sets of θ, i.e. the sets $A \in \mathcal{F}$ satisfying $A = \{\omega : \theta(\omega) \in A\}$, with probability 1, i.e.

$$\mathbf{P}\left[\{\omega : \theta(\omega) \in A,\, \omega \notin A\}\right] = \mathbf{P}\left[\{\omega : \theta(\omega) \notin A,\, \omega \in A\}\right] = 0.$$

Exercise 35. *Show that $\mathcal{I}$ is a σ-algebra.*

When $\mathcal{I} = \{\{\}, \Omega\}$ so that the right hand side of (4.12) is actually a constant, we call $\mathbf{P}$ an ergodic measure for θ. Let $\mathcal{M}$ be the set of all measures preserved by θ. This set is convex, i.e. if $a \in [0, 1]$ and $P_0, P_1 \in \mathcal{M}$ then

$$P_a = (1 - a)P_0 + aP_1.$$

Any measure that *cannot* be written as $(1 - a)P_0 + aP_1$ for $P_0 \neq P_1$ and $0 < a < 1$ is called an extremal point of $\mathcal{M}$. It turns out that the ergodic measures for θ are exactly the extremal points of $\mathcal{M}$. So, for example, if there is a single measure preserved by θ then it must be ergodic.

We would like to use the last remark and Birkhoff's Ergodic Theorem to conclude that, if the Markov process $X^{(t)}$ has a unique invariant probability measure π, and if $\pi[|f|] < \infty$ then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)}) = \pi[f].$$

But Birkhoff's theorem can't be immediately applied to establish (4.2). To use the theorem we need to first define the shift map $\theta : \mathbb{R}^{\pm\infty} \to \mathbb{R}^{\pm\infty}$ specified by the relation

$$X^{(k)} \circ \theta = X^{(k+1)}$$

(recall that $X^{(k)}$ is the projection of a sequence in $\mathbb{R}^{\pm\infty}$ onto its index k component). The shift map preserves the measure $P_{-\infty,\pi}$ on $\mathbb{R}^{\pm\infty}$. With these definitions, Birkhoff's Ergodic Theorem now tells us that, if $F : \mathbb{R}^{\pm\infty} \to \mathbb{R}$ satisfies

$$E_{-\infty,\pi}\left[|F|\right] < \infty$$

then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} F(\theta^k(\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots)) \\ = E_{-\infty,\pi}\left[F \,|\, \mathcal{I}\right](\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots)$$

for all $(\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots) \in \mathbb{R}^{\pm\infty}$ in a set of $P_{-\infty,\pi}$-probability 1 where $\mathcal{I}$ is the σ-algebra of the $P_{-\infty,\pi}$-invariant sets of θ. It turns out a probability distribution π is an extremal point of the set

$$\mathcal{M}_{\mathcal{T}} = \{\pi : \pi\mathcal{T} = \pi\}$$

of all invariant measures for a given Markov process $X^{(k)}$ with transition operator $\mathcal{T}$ if and only if $P_{-\infty,\pi}$ is an ergodic distribution for the shift map θ. We can conclude therefore that if $X^{(k)}$ has a unique transition probability distribution, π, then

$$E_{-\infty,\pi}\left[F \,|\, \mathcal{I}\right](\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots) = E_{-\infty,\pi}\left[F\right].$$

Interpreting this last conclusion in terms of the original Markov process and applying it to the test function $F = f \circ X^{(0)}$ we find that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)}) = \pi[f] \tag{4.13}$$

for a subset of $\mathbb{R}^{\pm\infty}$ with P_π-probability 1 whenever π is the unique invariant measure for the chain. In other words, as long as the initial condition is drawn from the unique invariant measure π, then (4.2) holds with probably

1. But we don't expect our first point in the chain to be drawn exactly from π. Fortunately (4.13) easily yields a more general result since it implies

$$\int P_x \left[\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)}) = \pi[f] \right] \pi(dx) = 1.$$

Since the integrand in the last display does not exceed 1, it must be that, for a set of initial conditions x of π-probability 1,

$$P_x \left[\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)}) = \pi[f] \right] = 1.$$

This result is enough to tell us that, if π is the unique invariant measure for our Markov chain and if we choose N large enough, then our Markov chain Monte Carlo estimate will be accurate. But it has two important limitations. First, while it may be easy to check that a particular distribution π is left invariant by a Markov chain, it is generally much harder to verify that π is the unique invariant measure. And second, even if we knew that π was unique and therefore that, by the preceding discussion, (4.2) holds, this does not tell us how large we can expect the error in our estimate to be for finite N. For that we would like something like a Central Limit or Large Deviations Theorem.

Guarantees on uniqueness of the invariant measure and stronger convergence results are often established using so-called Lyapunov conditions. For example, we might require that, for functions $V : \mathbb{R} \to [0, \infty]$ and $f : \mathbb{R} \to [1, \infty)$ and a finite constant b,

$$\mathcal{L}V(x) \leq -f(x) + b\mathbf{1}_S(x) \tag{4.14}$$

where on the subset $S \subset \mathbb{R}$ the transition distribution satisfies the so-called minorization property

$$\inf_{x\in S} P_x \left[X^{(1)} \in A \right] \geq \alpha\, \nu(A) \tag{4.15}$$

for any subset $A \subset \mathcal{B}$ where $\alpha \in (0, 1)$ and ν is a probability distribution on $\mathbb{R}$. The function V in these relations is called a Lyapunov function.

Roughly, the purpose of the Lyapunov condition (4.14) is to ensure that the chain visits the set S sufficiently frequently. To see that this is the case, take $f \equiv 1$ in (4.14) and notice that, if

$$\tau_S = \inf\{k > 0 : X^{(k)} \in S\},$$

then (4.14) implies

$$\begin{aligned}
0 \leq E_x\left[V(X^{(\tau_S)})\right] &= V(x) + E_x\left[\sum_{k=0}^{\tau_S - 1} V(X^{(k+1)}) - V(X^{(k)})\right] \\
&= V(x) + E_x\left[\sum_{k=0}^{\tau_S - 1} E_x\left[V(X^{(k+1)}) - V(X^{(k)}) \,|\, \mathcal{F}_k\right]\right] \\
&= V(x) + E_x\left[\sum_{k=0}^{\tau_S - 1} E_{k,X^{(k)}}\left[V(X^{(k+1)})\right] - V(X^{(k)})\right] \\
&= V(x) + E_x\left[\sum_{k=0}^{\tau_S - 1} \mathcal{L}V(X^{(k)})\right] \\
&\leq V(x) - E_x\left[\tau_S\right] + bE_x\left[\sum_{k=0}^{\tau_S - 1} \mathbf{1}_S(X^{(k)})\right].
\end{aligned}$$

so that, for $x \notin S$,

$$E_x\left[\tau_S\right] \leq V(x).$$

Therefore, for general x we can conclude that

$$\begin{aligned}
E_x\left[\tau_S\right] &= 1 + E_x\left[E_{X^{(1)}}\left[\tau_S\right]\mathbf{1}_{S^c}(X^{(1)})\right] \\
&\leq 1 + E_x\left[V(X^{(1)})\right] \\
&= 1 + V(x) + \mathcal{L}V(x) \\
&\leq V(x) + b\mathbf{1}_S(x).
\end{aligned}$$

Thus, for example, the expected time of first return to S from an initial point in S is bounded as long as V is bounded on S.

To see that (4.15) is related to the uniqueness of π, assume for the moment that $X^{(k)}$ is initialized in S and remains in S for all k with probability 1. In this case condition (4.15) implies that for any two probability distributions η and μ,

$$\|\eta\mathcal{T} - \mu\mathcal{T}\|_{\text{TV}} \leq (1 - \alpha)\|\eta - \mu\|_{\text{TV}}. \tag{4.16}$$

To prove (4.16) we will use what is called a coupling argument the first step of which is to recall that the total variation distance between two measures η and μ can be written as

$$\|\eta - \mu\|_{\text{TV}} = \min_{\substack{X\sim\eta \\ Y\sim\mu}} \mathbf{P}\left[X \neq Y\right].$$

Note that the marginal distributions of X and Y are constrained in this minimization but the rest of their joint distribution is not. To bound $\|\eta\mathcal{T} - \mu\mathcal{T}\|_{\text{TV}}$ we need only find a pair of random variables $X^{(1)}$ and $Y^{(1)}$ distributed according to $\eta\mathcal{T}$ and $\mu\mathcal{T}$ respectively and evaluate $\mathbf{P}\left[X^{(1)} \neq Y^{(1)}\right]$. To that end, let $X^{(0)}$ and $Y^{(0)}$ be random variables respectively distributed according to η and μ and satisfying

$$\|\eta - \mu\|_{\text{TV}} = \mathbf{P}\left[X^{(0)} \neq Y^{(0)}\right].$$

We define the random variables $X^{(1)}$ and $Y^{(1)}$ according to the following rules. Let χ be an independent $Bernoulli(\alpha)$ random variable and let ξ be an independent random variable distributed according to ν. Given $X^{(0)}$ and χ define $X^{(1)}$ according to

$$\mathbf{P}\left[X^{(1)} \in A \,|\, X^{(0)},\, \chi\right] = \begin{cases} \nu(A) \text{ if } \chi = 1 \\ Q_{X^{(0)}}\left[X^{(1)} \in A\right] \text{ if } \chi = 0 \end{cases}$$

where

$$Q_x\left[X^{(1)} \in A\right] = \frac{P_x\left[X^{(1)} \in A\right] - \alpha\,\nu\left[A\right]}{1-\alpha}.$$

Notice that if we require that $X^{(0)} \in S$ then (4.15) implies that $Q_{X^{(0)}}\left[X^{(1)} \in A\right]$ is a probability distribution. The decomposition

$$P_x\left[X^{(1)} \in A\right] = \alpha\,\nu(A) + (1-\alpha)Q_x\left[X^{(1)} \in A\right]$$

along with the fact that $X^{(0)} \sim \eta$ together imply that $X^{(1)} \sim \eta\mathcal{T}$.

Exercise 36. *Show that* $X^{(1)} \sim \eta\mathcal{T}$.

If $Y^{(0)} = X^{(0)}$ or if $\chi = 1$ set $Y^{(1)} = X^{(1)}$, otherwise let $Y^{(1)}$ be an independent random variable drawn from $Q_{Y^{(0)}}\left[X^{(1)} \in A\right]$. With these choices, $Y^{(1)} \sim \nu\mathcal{T}$ and

$$\begin{aligned} \|\eta\mathcal{T} - \mu\mathcal{T}\|_{\text{TV}} &\leq \mathbf{P}\left[X^{(1)} \neq Y^{(1)}\right] \\ &\leq (1-\alpha)\mathbf{P}\left[X^{(0)} \neq Y^{(0)}\right] \\ &= (1-\alpha)\|\eta - \mu\|_{\text{TV}}. \end{aligned}$$

Exercise 37. *Show that* $Y^{(1)} \sim \mu\mathcal{T}$.

It is important to note that these calculations were very special to the total variation norm. In most cases even if $\mathcal{T}$ satisfies (4.16) in the total variation norm it will not satisfy similar expressions in other norms.

From expression (4.16) we see that if η and μ are invariant measures then we must have that $\|\eta - \mu\|_{\text{TV}} = 0$ (in fact, the expression also implies the existence of an invariant measure). If π is the invariant measure for $\mathcal{T}$ and η is any other probability distribution then (4.16) also implies that

$$\|\eta\mathcal{T}^k - \pi\|_{\text{TV}} = \|\eta\mathcal{T}^k - \pi\mathcal{T}^k\|_{\text{TV}} \leq (1-\alpha)^k \|\eta - \pi\|_{\text{TV}}$$

so that the Markov chain started from $X^{(0)} \sim \eta$ converges in distribution to a random variable drawn from π. Under conditions (4.14) and (4.15) when X does not remain in S for all time with probability 1, we still know that it does not remain outside of S for very long and one can show that the invariant measure is still unique and that $X^{(k)}$ still converges in distribution to that invariant measure.

When conditions (4.14) and (4.15) are satisfied, the Markov chain with transition operator $\mathcal{T}$ satisfies the Central Limit Theorem,

$$\lim_{N\to\infty} \sqrt{N}\,(\overline{g}_N - \pi[g]) = Z$$

for all functions g with $|g| \leq f$ where $Z \sim \mathcal{N}(0, \tau_g\,\sigma_g^2)$ where $\sigma_g^2 = \mathbf{var}_\pi(X^{(1)})$ and

$$\tau_g = 1 + 2\sum_{k=1}^{\infty} \mathbf{cor}_\pi\left(g(X^{(0)}), g(X^{(k)})\right)$$

as long as $\tau_g > 0$. Conditions (4.14) and (4.15) also imply that $\tau_g < \infty$.

The constant τ_g is called the integrated auto-correlation time (IAT). It captures the "cost" of the correlations between samples $g(X^{(k)})$ in terms of the rate of convergence of our estimator. Roughly speaking, it tells us that a single sample $g(X^{(k)})$ of this Markov chain has the statistical value of $1/\tau$ independent samples from π. To see that the appearance of τ_g in the asymptotic variance for $\sqrt{N}\,(\overline{g}_N - \pi[g])$ is reasonable, assume for a moment that

$X^{(0)} \sim \pi$ and consider the mean squared error,

$$\begin{aligned} N\mathbf{E}\left[(\overline{g}_N - \pi[g])^2\right] &= \frac{1}{N}\sum_{\ell,k=1}^{N} \mathbf{E}\left[\left(g(X^{(k)}) - \pi[g]\right)\left(f(X^{(\ell)}) - \pi[g]\right)\right] \\ &= \frac{1}{N}\sum_{k=1}^{N} \mathbf{E}\left[\left(g(X^{(k)}) - \pi[g]\right)^2\right] \\ &\qquad + \frac{2}{N}\sum_{\substack{1\le k<N \\ 1\le \ell\le N-k}} \mathbf{E}\left[\left(g(X^{(\ell)}) - \pi[g]\right)\left(g(X^{(\ell+k)}) - \pi[g]\right)\right] \end{aligned}$$

Because each $X^{(k)}$ is distributed according to π, the first term on the right hand side of the last display is σ_g^2. For the same reason, the second term is exactly

$$2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\mathbf{cov}_\pi\left(g(X^{(0)}), g(X^{(k)})\right)$$

which we can rewrite as

$$2\sigma_g^2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\mathbf{cor}_\pi\left(g(X^{(0)}), g(X^{(k)})\right).$$

In the large N limit we can expect that the number in the last display converges to $\sigma_g^2(\tau_g - 1)$ so that

$$\lim_{N\to\infty} N\mathbf{E}\left[(\overline{g}_N - \pi[g])^2\right] = \sigma_g^2 + \sigma_g^2(\tau_g - 1) = \sigma_g^2\tau_g.$$

The central limit theorem above suggests that τ_g is a natural measure of the quality of a Markov chain Monte Carlo scheme (note that σ_g does not depend on the particular chain used to sample from π). Unfortunately, it is notoriously difficult to accurately estimate τ_g. For the exercises in these notes I recommend that you use the IAT estimator included in the python `emcee` package.

For many reversible Markov chains, their generator $\mathcal{L}$ has a positive spectral gap, that is

$$\gamma = \inf_{\substack{\pi[f]=0 \\ \mathbf{var}_\pi f \neq 0}} \frac{-\int f(x)\mathcal{L}f(x)\pi(dx)}{\int f^2(x)\pi(dx)} > 0.$$

This implies a bound on the integrated autocorrelation time τ_g as follows. First notice that a positive spectral gap implies, that

$$\|\mathcal{T}(f - \pi[f])\|_\pi \leq (1 - \gamma)\|f - \pi[f]\|_\pi$$

where we have defined the norm $\|f\|_\pi = \left(\int f^2(x)\pi(dx)\right)^{\frac{1}{2}}$ and $\|f\|_\pi < \infty$. We have already seen that a bound of this type implies that we can define an inverse $\mathcal{L}^{-1}$ of $\mathcal{L}$ by

$$-\mathcal{L}^{-1} f = \sum_{k=0}^{\infty} \mathcal{T}^k f$$

as long as we only apply it to functions with $\pi[f] = 0$. From this formula it is also clear that

$$\|\mathcal{L}^{-1}(g - \pi[g])\|_\pi \leq \frac{1}{\gamma}\sigma_g.$$

Now notice that we can rewrite the formula for the asymptotic variance $\sigma_g^2\tau_g$ as

$$\begin{aligned}\sigma_g^2\tau_g &= 2\sum_{k=0}^{\infty}\int (g - \pi[g])\mathcal{T}^k[g - \pi[g]]\pi(dx) - \sigma_g^2 \\ &= -2\int (g - \pi[g])\mathcal{L}^{-1}[g - \pi[g]](x)\pi(dx) - \sigma_g^2 \qquad (4.17)\end{aligned}$$

By an application of the Cauchy-Schwartz inequality we find that

$$\sigma_g^2(\tau_g + 1) \leq 2\sigma_g\|\mathcal{L}^{-1}(g - \pi[g])\|_\pi \leq \sigma_g^2\frac{2}{\gamma},$$

or

$$\tau_g \leq \frac{2}{\gamma} - 1.$$

So a large spectral gap implies that averages with respect to any observable will converge quickly.

In fact, for many Markov chains, the spectral gap is an eigenvalue (the smallest non-zero eigenvalue) of $-\mathcal{L}$, i.e. there is a function $\psi(x)$ satisfying

$$-\mathcal{L}\psi = \gamma\psi, \quad \pi[\psi] = 0, \quad \mathbf{var}_\pi\psi = 1.$$

Note that the eigenfunction ψ is also an eigenfunction of $-\mathcal{L}^{-1}$ with eigenvalue γ^{-1}. From the formula for the asymptotic variance in (4.17), it is therefore clear that

$$\tau_\psi = \frac{2}{\gamma} - 1.$$

In other words, there is some function for which our upper bound on the integrated autocorrelation time is achieved by ψ. The spectral gap therefore often provides a worst case estimate of the convergence rate of a reversible Markov chain to its equilibrium distribution.

This close relationship between the spectral gap and the integrated auto correlation time is often useful for assessing the convergence of relatively simple Markov processes. For complex processes even accurate estimates of the spectral gap do not lead to practically useful bounds. This is because the functions whose averages converge the slowest in a complex system are often not the observables whose average we would like to compute. For example, in any molecular dynamics simulation there are many large scale rearrangements of the system that would not occur on any reasonable physical timescale. Though it is possible for a protein to tie itself in a knot, the probability that this happens within the lifetime of the protein is vanishingly small and the event is therefore of no biological interest.

4.4 Partial resampling

In this section we focus on transition rules for Markov chains that exactly preserve the value of some subset of the components (perhaps after a change of coordinates) of the chain at each step. We will assume that $X^{(t)} \sim \pi$ and that only a single component of $X^{(t)}$ is changed in any single step of the process. Each component could (and often will) have more than one dimension. Let j_t be the single component that is changed in moving from $X^{(t)}$ to $X^{(t+1)}$, i.e. that $X_j^{(t+1)} = X_j^{(t)}$ for $j \neq j_t$. We assume that the transformation of the j_t component preserves the conditional distribution of x_{j_t} given $x_{[j_t]} = (x_1, \ldots, x_{j_t-1}, x_{j_t+1}, \ldots, x_d)$. That is, we assume that

$$E_\pi \left[h(X_{j_t}^{(t+1)}) \,|\, X_{[j_t]}^{(t)} = x_{[j_t]} \right] = \int h(x_{j_t}) \pi(dx_{j_t} | x_{[j_t]}) \tag{4.18}$$

for any nice function h of x_{j_t}. If $p(y_{j_t}|x)$ is the probability density of the new j_tth component given the previous value of x then our assumption is that

$$\int p(y_{j_t}|x)\pi(dx_{j_t}|x_{[j_t]}) = \pi(y_{j_t}|x_{[j_t]}).$$

Our goal is to construct a chain that preserves π. In other words we would like to know if chains satisfying the requirements just described, preserve the complete π distribution. For any function f,

$$E_\pi\left[f(X^{(t+1)})\right] = E_\pi\left[f(X^{(t+1)}_{j_t}, X^{(t)}_{[j_t]})\right] = E_\pi\left[E_\pi\left[f(X^{(t+1)}_{j_t}, X^{(t)}_{[j_t]}) \,|\, X^{(t)}_{[j_t]}\right]\right]$$

Applying our assumption in (4.18) to $f(x_{j_t}, x_{[j_t]})$ for each fixed value of $x_{[j_t]}$, we find that

$$E_\pi\left[f(X^{(t+1)})\right] = \int f(x_{j_t}, x_{[j_t]})\pi(dx_{j_t}|x_{[x_{j_t}]})\pi(dx_{[j_t]}) = \int f(x)\pi(dx).$$

The argument above also applies if we first apply a coordinate transformation φ to $X^{(t)}$ to obtain a new variable $Y^{(t)} = \varphi(X^{(t)})$, then update one component of new variable $Y^{(t+1)} = (Y^{(t+1)}_{j_t}, Y^{(t)}_{[j_t]})$ and set $X^{(t+1)} = \varphi^{-1}(Y^{(t+1)})$. In this case the requirement is that the update of the j_t coordinate of Y should preserve the conditional distribution $\pi_\varphi(y_{j_t}|y_{[j_t]})$ where π_φ is the transformation of π under φ,

$$\pi_\varphi(y) = \frac{\pi(\varphi^{-1}(y))}{|D\varphi(\varphi^{-1}(y))|}$$

Thus transitions that involve linear combinations of coordinates, or even non-linear combinations of coordinates, can yield chains that preserve π as long as the low dimensional transitions preserve the appropriate conditional distribution.

Any scheme that fixes some components of the chain (in some coordinate frame) at each step and preserves the conditional density of π for the remaining components therefore leaves π invariant. This fact is referred to as the principle of partial resampling and is fundamental to many Markov chain Monte Carlo schemes. Of course any scheme that preserves the same coordinates of the state at each step cannot be ergodic.

Exercise 38. *Why not?*

Schemes relying on the partial resampling principle either use a deterministic schedule to choose component to be modified at step, j_t, or choose it randomly (and independent of the sample state) at each step.

4.5 Gibbs sampling

In the simplest realization of the partial resampling principle the transitions of the Markov chain are drawn directly and independently from a conditional distribution, i.e.

$$p(t+1, y_{j_t} \,|\, t, x) = \pi(y_{j_t})$$

Markov chains of this type are called Gibbs samplers.

Gibbs samplers are attractive in that, at each step, samples are drawn independently from a conditional distribution of π. However, it is not applicable unless the coordinates can be chosen so that the conditional distribution of each component is very simple (e.g. Gaussian). Even in the special cases in which this is possible, fixing the choice of coordinates can lead to slow convergence of the Markov chain due to poor conditioning of the target density in those coordinates (see Chapter **??**).

Example 21. *Consider a vector σ indexed on the periodic 2 dimensional lattice $\mathbb{Z}_L^2$ and with values in $\{-1, 1\}$. If we assign the density*

$$\pi(\sigma) = \frac{e^{\beta \sum_{\vec{i} \leftrightarrow \vec{j}} \sigma_{\vec{i}} \sigma_{\vec{j}}}}{\mathcal{Z}}$$

to the σ variables then this becomes the Ising model of statistical physics. Here $\vec{i}, \vec{j} \in \mathbb{Z}_L^2$ and here $\vec{i} \leftrightarrow \vec{j}$ indicates that $\vec{i}$ and $\vec{j}$ are neighboring sites on the lattice. The constant $\beta > 0$ is related to a physical temperature via $k_B T = \beta^{-1}$ where k_B is the Boltzmann constant. Then

$$\pi(\sigma_{\vec{i}_t} \,|\, \sigma_{[\vec{i}t]}) = w_+ \, \delta(\sigma_{\vec{i}_k} - 1) + w_- \, \delta(\sigma_{\vec{i}_k} + 1)$$

where

$$w_+ = \frac{e^{\beta \sum_{\vec{i} \leftrightarrow \vec{j}} \sigma_{\vec{i}}}}{e^{\beta \sum_{\vec{i} \leftrightarrow \vec{i}_t} \sigma_{\vec{i}}} + e^{-\beta \sum_{\vec{i} \leftrightarrow \vec{i}_t} \sigma_{\vec{i}}}}$$

and

$$w_- = \frac{e^{-\beta \sum_{\vec{i} \leftrightarrow \vec{i}_t} \sigma_{\vec{i}}}}{e^{\beta \sum_{\vec{i} \leftrightarrow \vec{i}_t} \sigma_{\vec{i}}} + e^{-\beta \sum_{\vec{i} \leftrightarrow \vec{i}_t} \sigma_{\vec{i}}}}$$

Exercise 39. *Write a routine that uses a Gibbs sampler to generate samples of the Ising model. Plot a histogram of the values of the magnetization*

$$f(\sigma) = \sum_{\vec{i} \in \mathbb{Z}_L^2} \sigma_{\vec{i}}.$$

Compute the integrated autocorrelation time for the magnetization. Do you find it better to select $\vec{i}_t$ randomly, or to sweep through the lattice deterministically? What happens to the integrated autocorrelation time when you change the temperature? What happens to the integrated autocorrelation time when you change the size of the lattice?

WARNING: integrated autocorrelation times are notoriously difficult to estimate. You should check that your estimate has converged by computing it on a few trajectories of increasing length.

4.6 The Metropolis–Hastings scheme

The building block of the vast majority of Markov chain Monte Carlo algorithms is called the Metropolis–Hastings scheme and proceeds from $X^{(t)}$ to $X^{(t+1)}$ as follows

1. `Generate a random variable` $Y^{(t+1)}$ `from some proposal distribution` $q(y \,|\, X^{(t)})$.
2. `With probability`

$$p_{acc}(X^{(t)}, Y^{(t+1)}) = \min\left\{1, \frac{\pi(Y^{(t+1)})\, q(X^{(t)} \,|\, Y^{(t+1)})}{\pi(X^{(t)})\, q(Y^{(t+1)} \,|\, X^{(t)})}\right\}$$

`set` $X^{(t+1)} = Y^{(t+1)}$. `Otherwise set` $X^{(t+1)} = X^{(t)}$.

The method has spawned a huge number of generalizations which focus mainly on the proposal choice (step 1). Notice that the method is in many ways similar to the rejection method. One key difference is that in this scheme we do not wait for an acceptance to adopt a new sample. Instead

when a proposal is rejected a copy of the last sample is adopted as the next sample.

The transition operator defined by the algorithm is given by

$$\mathcal{T}f = f(x)\, p_{rej}(x) + \int f(y)\, q(y \,|\, x)\, p_{acc}(x, y)\, dy$$

where

$$\begin{aligned} p_{rej}(x) &= \mathbf{P}_{k,x}\left[Y^{(k+1)} \text{ is rejected}\right] \\ &= \int (1 - p_{acc}(x, z))\, q(dz \,|\, x) \end{aligned}$$

is called the rejection probability.

Exercise 40. *Check that the transition operator is correct.*

The Metropolis–Hastings scheme generates a reversible chain with respect to the target density π. To see this notice that

$$\begin{aligned} q(y \,|\, x)\, p_{acc}(x, y)\pi(x) &= q(y \,|\, x) \min\left\{1, \frac{\pi(y)q(x \,|\, y)}{\pi(x)q(y \,|\, x)}\right\} \\ &= \min\left\{q(y \,|\, x)\pi(x), q(x \,|\, y)\pi(y)\right\} \\ &= q(x \,|\, y)\, p_{acc}(y, x)\pi(y) \end{aligned}$$

so that

$$\begin{aligned} \int g(x)\mathcal{T}f(x)\pi(dx) &= \int f(x)\, g(x)\, p_{rej}(x)\, \pi(dx) \\ &\qquad + \int g(x) \int f(y)\, q(y \,|\, x)\, p_{acc}(x, y)\, dy\, \pi(dx) \\ &= \int f(x)\, g(x)\, p_{rej}(x)\, \pi(dx) \\ &\qquad + \int f(y) \int g(x)\, q(x \,|\, y)\, p_{acc}(y, x)\, dx\, \pi(dy) \\ &= \int f(x)\mathcal{T}g(x)\pi(dx) \end{aligned}$$

Its intuitively clear that if you choose a proposal density $q(y \mid x)$ so that the average rate of rejection is too high, the chain $X^{(t)}$ cannot relax quickly (i.e. $\overline{f}_N$ will converge slowly). This means we must choose a $q(y \mid x)$ that is nearly reversible with respect to π (that it nearly or exactly preserves π is not enough) so that the acceptance probability is nearly 1. From our experience with importance sampling we might expect that this can be difficult in high dimensions. To see that this is indeed the case, consider the application of a Metropolis-Hastings scheme with

$$\pi(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

for some density p. We will also assume that

$$q(y_{1:d} \mid x_{1:d}) = \prod_{i=1}^{d} q_1(y_i \mid x_i).$$

In this setting, letting

$$w(x_i, y_i) = \frac{q_1(x_i \mid y_i)\pi(y_i)}{q_1(y_i \mid x_i)\pi(x_i)},$$

the average rejection probability is

$$\begin{aligned}\int p_{rej}(x)\pi(dx) &= 1 - \int \min\left\{1, e^{\sum_{i=1}^{d} \log w(x_i,y_i)}\right\} q(dy \mid x)\pi(dx) \\ &= 1 - \int e^{-d \max\left\{0, -\frac{1}{d}\sum_{i=1}^{d} \log w(x_i,y_i)\right\}} q(dy \mid x)\pi(dx)\end{aligned}$$

Under our assumptions, if (X, Y) is distributed according to $q(y \mid x)\pi(x)$ then the variables $w(X_i, Y_i)$ are independent. Their mean,

$$\bar{w} = \int - \left(\log \frac{q_1(x_i \mid y_i)\pi(y_i)}{q_1(y_i \mid x_i)\pi(x_i)}\right) q_1(dy_i \mid x_i)\pi(dx_i)$$

is the relative entropy of the density $q_1(x_i \mid y_i)\pi(y_i)$ with respect to $q_1(y_i \mid x_i)\pi(x_i)$ and is positive (by Jensen's inequality) unless the two densities are equal. We will assume that they aren't equal and that $\bar{w} > 0$. We will try to exploit the Large Deviations Principle for the sample average $-\frac{1}{d}\sum_{i=1}^{d} \log w(X_i, Y_i)$

to see that the average rejection probability is exponentially close to 1 as d grows. Recall that Cramer's Theorem for $-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)$ tells us that

$$\limsup_{d\to\infty}\frac{1}{d}\log\mathbf{P}\left[-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\in(a,b)\right]\leq\inf_{x\in(a,b)}I(x)$$

where the rate function I is defined via a Legendre transformation of the moment generating function of $-\log w(X_i,Y_i)$.

Now fix any constant $R>0$ and $K\in\mathbb{N}$, and write

$$\begin{aligned}&\mathbf{E}\left[e^{-d\max\left\{0,-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\right\}}\right]\\&=\sum_{k=0}^{K-1}\mathbf{E}\left[e^{-d\max\left\{0,-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\right\}};\,-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\in\left(\frac{k}{K},\frac{k+1}{K}\right]R\right]\\&\qquad+\mathbf{E}\left[e^{-d\max\left\{0,-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\right\}};\,-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)>R\right]\end{aligned}$$

This expression is bounded by

$$\begin{aligned}&\sum_{k=0}^{K-1}e^{-d\frac{kR}{K}}\,\mathbf{P}\left[-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)\in\left(\frac{k}{K},\frac{k+1}{K}\right]R\right]\\&\qquad+\mathbf{P}\left[-\frac{1}{d}\sum_{i=1}^{d}\log w(X_i,Y_i)>R\right].\end{aligned}$$

Combining this bound with Cramer's Theorem we find that

$$\begin{aligned}&\limsup_{d\to\infty}\frac{1}{d}\log\left(1-\int p_{rej}(x)\pi(dx)\right)\\&\qquad\leq\max\left\{\max_{k=0,1,\ldots,K}\left\{-\frac{kR}{K}-\inf_{x\in\left(\frac{k}{K},\frac{k+1}{K}\right]R}I(x)\right\},-\inf_{x>R}I(x)\right\}.\end{aligned}$$

I is continuous (in fact, its convex) and $I(x)\to\infty$ as $|x|\to\infty$, taking $K\to\infty$ and then $R\to\infty$, we obtain

$$\limsup_{d\to\infty}\frac{1}{d}\log\left(1-\int p_{rej}(x)\pi(dx)\right)\leq-\inf_{x>0}\left\{x+I(x)\right\},$$

so that, indeed, the average rejection probability is exponentially close to 1 as d grows.

Exercise 41. *For the setup in the last example show that the lower bound*

$$\liminf_{d\to\infty} \frac{1}{d} \log\left(1 - \int p_{rej}(x)\pi(dx)\right) \geq -\inf_{x>0}\{x + I(x)\}$$

also holds.

As for importance sampling, typical high dimensional sampling problems exhibit low dimensional structure which, if unknown, complicates sampling and leads to rejection rates that are even smaller than predicted by the independent component case. Fortunately, the flexibility in the structure of the Metropolis scheme allows for choices of proposal densities that, while not requiring detailed knowledge of the properties of π, can yield effective schemes in relatively high dimensions. In any case, the key to designing a successful Metropolis scheme for a complicated, high dimensional problem, is a good choice of q.

More precisely, one should use as much knowledge of the underlying problem (i.e. of π) as possible, to chose $q(y \,|\, x)$ so that the size of the typical displacement $Y^{(k+1)} - X^{(k)}$ is sufficiently large, for example you want to choose q so that

$$\mathbf{E}_{k,x}\left[\left(Y^{(k+1)} - x\right)^2\right]$$

is large, and so that the probability of a rejection p_{rej} is not too large. By examining p_{acc} we see that these two goals are in conflict. If we choose a q so that $Y^{(k+1)}$ is typically very close to $X^{(k)}$ and if π is a smooth density, then $\pi(Y^{(k+1)})$ will be very close to $\pi(X^{(k)})$ and (at least when q is symmetric), p_{acc} will be very close to 1.

A choice that requires very little knowledge of the underlying problem and has been successful in a wide range of problems of low to moderate dimension, is a proposal of the form,

$$Y_{i_k}^{(k+1)} = X_{i_k}^{(k)} + \xi^{(k+1)},$$

$$Y_{[i_k]}^{(k+1)} = X_{[i_k]}^{(k)}.$$

where $\xi^{(k+1)}$ is some one (or low) dimensional isotropic (the density of $\xi^{(k+1)}$ is a function only of distance from the origin) random variable and i_k is the indices of one (or a few) coordinates chosen either randomly or deterministically and varying from step to step (as in Gibbs sampling). Consequently, $q(y \mid x)$ is often symmetric in x and y and does not appear in p_{acc}. But even this choice, to be efficient, requires some special structure in π. To see this recall that at each step of the Metropolis scheme one must compute the ratio $\pi(Y^{(k+1)})/\pi(X^{(k)})$. On many problems evaluating $\pi(y)$ can be extremely expensive. This would seem to doom any method that has to re-evaluate π after perturbing just one or a few dimensions. Fortunately, for many problems it is much cheaper to evaluate the ratio $\pi(Y^{(k+1)})/\pi(X^{(k)})$ than to evaluate $\pi(X^{(k)})$ itself. The reason for this, as the next example demonstrates, is that π may be a product of terms, most of which do not depend on the variables being perturbed in any one step of the chain. Note that this property does not require independence of the components of X under π. It does, however, require some form of conditional independence.

Example 22. *Consider the Ising model and suppose that we are constructing a chain $X^{(k)}$ with values in the set of $L \times L$ matrices with entries in $\{-1, 1\}$, and that, at each step of the chain, $Y^{(k+1)}$ corresponds to flipping the sign of a single entry of $X^{(k)}$. Suppose that the index of the spin at which a sign flip is proposed at step k is $\vec{i}_k$, then*

$$\frac{\pi(Y^{(k+1)})}{\pi(X^{(k)})} = e^{-4\beta X_{\vec{i}_k}(k) \sum_{\vec{j} \leftrightarrow \vec{i}_k} X_{\vec{j}}(k)}.$$

Thus we need only sum the spins of the neighbors of $\vec{i}_k$, an operation much less costly than the $\mathcal{O}(L^2)$ operations required to evaluate π itself (ignoring the normalization constant which we don't need to know).

Exercise 42. *Write a Metropolis based scheme to sample the 2d-Ising model. Compare the magnetism integrated autocorrelation time for this scheme to the one you computed for the Gibbs sampler.*

There is a substantial body of work examining the behavior of the Metropolis scheme in high dimensions, and in particular, the optimal balance between proposal size and rejection rate. While mathematically rigorous statements can only be made in extremely restrictive settings, they do tend to agree with observations made by practitioners. In particular, the mathematical results

support the long held rule of thumb that one should choose a proposal size that results in a rejection rate of about %25. Of course a rule like this cannot apply to every possible setting. Nonetheless, the agreement between experience and theory is remarkable.

4.7 Importance weights for MCMC

Because the samples, $X^{(k)}$, generated by a typical MCMC scheme are only asymptotically (in the k large limit) distributed according to the target distribution π, for finite N the MCMC estimator $\overline{f}_N$ will have a bias. It is natural then to ask if the samples obtained from MCMC can be reweighted so that they can be used to compute averages with respect to π with no (or little) bias. Since the distribution of $X^{(k)}$ for finite k is not known in any explicit form, the answer to this question is not obvious. As we show in this section, with some assumptions on how the chain is generated, at least the final sample $X^{(N)}$ can be reweighted so that it can be used to compute unbiased (or nearly unbiased) averages against π even when N is finite.

Suppose that $X^{(0)}$ is drawn from a distribution $\tilde{\pi}$ and that our goal is to compute averages with respect to π. Instead of generating many steps of a Markov chain whose transitions leave π invariant, first introduce a sequence of distributions $\pi_0, \pi_1, \ldots, \pi_N$ with $\pi_0 = \tilde{\pi}$ and $\pi_N = \pi$, and assume that, for each $k \geq 1$, $\mathcal{T}_k$ is a transition operator that preserves π_{k-1}, i.e. $\pi_{k-1}\mathcal{T}_k = \pi_{k-1}$. In most applications, one chooses the π_k to "interpolate" between $\tilde{\pi}$ and π in the sense that π_k and π_{k+1} are close to one another. For example, a common and very simple choice is

$$\pi_k \propto \left(\frac{\pi}{\tilde{\pi}}\right)^{\frac{k}{N}} \tilde{\pi} \tag{4.19}$$

though there are many possibilities.

Let the in-homogenous chain $X^{(k)}$ be generated by this sequence of transition operators, i.e. for any test function f, $E_{k-1,x}\left[f(X^{(k)})\right] = \mathcal{T}_k f(x)$.

Beginning with $W^{(0)} = 1$, define importance weights recursively according to

$$W^{(k)} = W^{(k-1)}\, w_k(X^{(k)}) \quad \text{with} \quad w_k(x) = \frac{\pi_k(x)}{\pi_{k-1}(x)}$$

Notice that, for any test function f,

$$\pi_{k-1}\mathcal{T}_k[w_k f] = \pi_{k-1}[w_k f] = \pi_k[f]$$

As a consequence,

$$E_{\tilde{\pi}}\left[f(X^{(1)})W^{(1)}\right] = \pi_0\mathcal{T}_1[w_1 f] = \pi_1[f]$$

so that, after weighting by $W^{(1)}$, the sample $X^{(1)}$ is drawn from the density π_1 (without weighting it is drawn from $\pi_0\mathcal{T}_0$). For the purposes of an argument by induction, assume now that after weighting by $W^{(k-1)}$, the sample $X^{(k-1)}$ is drawn from π_{k-1}, i.e. that for any test function f,

$$E_{\tilde{\pi}}\left[f(X^{(k-1)})W^{(k-1)}\right] = \pi_{k-1}[f].$$

Then, letting $g(x) = \mathcal{T}_k[w_k f](x)$, we have that

$$\begin{aligned} E_{\tilde{\pi}}\left[f(X^{(k)})W^{(k)}\right] &= E_{\tilde{\pi}}\left[g(X^{(k-1)})W^{(k-1)}\right] \\ &= \pi_{k-1}\mathcal{T}_k[w_k f] \\ &= \pi_k[f], \end{aligned}$$

i.e. after weighting by $W^{(k)}$, the sample $X^{(k)}$ is drawn from π_k.

Of course in most cases, the function w_k will be known only up to a multiplicative constant. In this case we can generate M independent copies $X^{(k,j)}$ of the chain $X^{(k)}$ and use weights updated by the formula

$$W^{(k,j)} = \frac{W^{(k-1,j)}\, w_k(X^{(k-1,j)})}{\sum_{\ell=1}^{M} W^{(k-1,\ell)}\, w_k(X^{(k-1,\ell)})}$$

at the cost of some bias. We will call the resulting scheme Jarzynski's method.

In most applications, the trajectories of $X^{(k)}$ that result in non-negligible weights are somewhat rare resulting in an estimator with high variance. There is a natural strategy to attempt to remedy is problem. Notice that Jarzynki's method is the sequential implementation of importance sampling with multidimensional reference density

$$\tilde{\pi}(x_{0:N}) = \pi(x_0)\prod_{n=1}^{N} q_n(x_n \,|\, x_{n-1})$$

and multidimensional target density

$$\eta(x_{0:N}) = \pi(x_0) \prod_{n=1}^{N} \frac{\pi_n(x_n)}{\pi_{n-1}(x_n)} \, q_n(x_n \,|\, x_{n-1}).$$

Our considerations above imply both that η is a density and that

$$\eta(x_N) = \int \eta(x_{0:N}) \, dx_{0:N-1} = \pi(x_N),$$

i.e. the marginal distribution of the x_N variables under η is exactly the target density π. In light of this observation, we introduce a resampling of the weighted ensemble $\{X^{(k,j)}, W^{(k,j)}\}_{j=1}^{M}$ between each increment of k so that samples with small weight are removed and more effort is expended on samples with large weight exactly as one might normally apply in the context of sequential importance sampling and as described in Chapter ??.

Now consider Jarzynski's method (without resampling) when $\pi_k = \pi$ for $k = 1, 2, \ldots, N$. In this case the final N steps of the chain are all generated using a transition operator that preserves the target density π. One might hope then that in the limit of large N, since $X^{(N)}$ is asymptotically distributed according to π, the variance of the weights would vanish. Unfortunately this is not the case. In this case, for $k \geq 2$, $w_k = 1$, so $W^{(k)} = W^{(1)} = \pi(X^{(1)})/\tilde{\pi}(X^{(1)})$ so that increasing N does not change the weights at all.

If, on the other hand, we use the densities in (4.19) and we define the function

$$V(x) = N \log w_k = \log \left(\frac{\pi(x)}{\tilde{\pi}(x)} \right)$$

then, ignoring normalization of the weights,

$$W^{(N)} = \exp \left(\frac{1}{N} \sum_{k=1}^{N} V(X^{(k)}) \right).$$

If every step of $X^{(k)}$ preserved π, then you would expect that $\log W^{(N)}$ would converge to $\int V(x)\pi(dx)$ in the large N limit. Incidentally, the quantity $\int V(x)\pi(dx) = R(\tilde{\pi}\|\pi)$ is the relative entropy of $\tilde{\pi}$ with respect to π which is a commonly used measure of the difference between $\tilde{\pi}$ and π. Though the steps of $X^{(k)}$ do not preserve π, for large k they do preserve a distribution close to π, and we can still expect that the variance of the weights will decrease in the large N limit.

Exercise 43. *Use Jarzynski's method (without resampling) to generate weighted samples from the* 2*d-Ising model. Choose*

$$\pi_k = \pi^{\frac{k}{N}}$$

and try defining the transition operators $\mathcal{T}_k$ *for the Markov chain to be the ones you used in Exercises 39 and 42 with* π *in those exercises by* π_{k-1}*. Note that this means that* $X^{(0)}$ *is drawn from the distribution with independent spins (i.e.* $\beta = 0$*). Evaluate the performance of the estimator of the magnetization. How does the variance change when you increase* N*? What seems to be the optimal choice of* N *(considering variance and effort) for this problem? Try Jarzynski's method with resampling. Does resampling help? How would you compare Jarzynski's method to Gibbs or Metropolis sampling for this problem?*

Exercise 44. *At step* k *Jarzynski's method gives you an importance sampling scheme to sample from* π_k*. For large* k*, if* π_k *is close to* π_N *it seems a shame to make no use of the sample* $X^{(k)}$*. Can you think of a way to re-weight the (already weighted) sample* $X^{(k)}$ *so that it can contribute to estimates of averages with respect to* π_N*? Experiment with this on the* 2*d-Ising model. Does it help?*

4.8 bibliography