Chapter 1

Probability distributions and Monte Carlo

1.1 Where do probability distributions come from?

Sampling problems arise in a huge array of applications in the physical, biological, and social sciences. For example estimating equilibrium or non-equilibrium properties of a molecular system, predicting the weather, and pricing stock options all require some form of sampling. The probability distributions encountered in these problems are extremely complex and usually very high dimensional (some times millions or billions of variables). As we will learn later, designing effective Monte Carlo methods for these problems is a major challenge.

Statistical inverse problems are a particularly ubiquitous variety of sampling problems that arise as follows. Suppose you believe that some observable quantity (e.g. the radial velocity of a star) is given by a known function of a number of parameters (e.g. the orbital parameters of a planet orbiting the star), Θ. Suppose that, in the absence of any observational information you believe that the parameters are distributed according to some density $p(\theta)$ which is referred to as a prior distribution and may be used to enforce, for

example, known physical constraints on the parameters Θ. Suppose further that this prior is supplemented by noisy measurements, X, of the observable. For example, you might model the relationship between observed values, X, and the values of Θ as

$$X = g(\Theta) + Z \tag{1.1}$$

where Z is the noise.

For definiteness, let's suppose that the noise is a Gaussian random variable with mean 0 and variance σ^2, i.e. $Z \sim \mathcal{N}(0, \sigma^2)$. We'll see later that it's not hard to generate samples from $\mathcal{N}(m, \sigma^2)$. So for a fixed value of Θ, we can easily generate samples of X. However, the goal here is usually to produce an estimate of the true value of Θ from obsevations of X. A reasonable approximation of the true value of Θ is given by the "posterior mean"

$$\hat{\theta} = \int \theta \, \pi(\theta|X) d\theta$$

where $\pi(\theta|x)$ is the "posterior density" of Θ implied by p and expression (1.1) with a fixed value of X, in this case

$$\pi(\theta|x) = \frac{e^{-\frac{(x-g(\theta))^2}{2\sigma^2}} p(\theta)}{\int e^{-\frac{(x-g(\xi))^2}{2\sigma^2}} p(\xi) d\xi}.$$

Note that for fixed X the distribution of Θ dictated by (1.1) can be far from Gaussian and impossible to sample directly.

Another common source of difficult sampling problems is statistical mechanics. Ludwig Boltzmann famously postulated that the positions $\hat{x}$ and momenta $\tilde{x}$ of the atoms in a molecular system of constant size (n), occupying a constant volume, and in contact with a heat bath (at constant temperature T), are distributed according to

$$\pi(\hat{x}, \tilde{x}) = \frac{e^{-\beta(V(\hat{x}) + \mathcal{K}(\tilde{x}))}}{\int e^{-\beta(V(\hat{x}) + \mathcal{K}(\tilde{x}))} d\hat{x} d\tilde{x}}$$

where V is a potential energy describing the interaction of the particles in the system,

$$\mathcal{K}(\tilde{x}) = \sum_{i=0}^{n-1} \frac{\tilde{x}_i^2}{m_i}$$

is the kinetic energy of the system, and $\beta = (k_B T)^{-1}$ is the inverse product of Boltzmann's constant and the temperature. The potential V is often a very rough function with many local minima. This and other typical features of the potential make computing averages with respect to $\pi(\hat{x}, \tilde{x})$ a very difficult undertaking. But the quantities that determine, for example, whether a new drug treatment might be effective, are defined as averages with respect to the Boltzmann distribution and computing integrals of the form

$$\int f(\hat{x}, \tilde{x})\, \pi(\hat{x}, \tilde{x}) d\hat{x}\, d\tilde{x}$$

cannot be avoided.

1.2 What is sampling and why Monte Carlo?

Suppose that $\pi \geq 0$, $\int \pi(x)dx = 1$, and you want to compute the integral (average)

$$\pi[f] = \int f(x)\, \pi(x)\, dx. \tag{1.2}$$

Of course for any complicated pair of functions f and π, evaluating the integral (1.2) by hand is out of the question. And because x may take infinite, even uncountably many, values we cannot even evaluate the integral exactly on a computer. The goal of numerical integration, random or otherwise, is then to select a finite set of points at which to evaluate f and π and to assemble an approximation to (1.2) from those values. We can express this goal more generally as the desire to construct an estimator of $\pi[f]$ of the form

$$\sum_{k=0}^{N-1} f(x^{(k)})\, w^{(k)}$$

where the $\{x^{(k)}\}_{k=0}^{N-1}$ are N points and the $w^{(k)}$ are, as yet unspecified, non-negative values referred to here as "weights" that depend on the particular scheme in question. Together the collection of points and weights $\{w^{(k)}, x^{(k)}\}$ are referred to as "samples," or as an "ensemble."

A useful way to asses the quality of an estimate of this type is to consider the number of samples, $N(\delta)$, required to achieve an estimate of $\pi[f]$ of accuracy

δ. Assume for the moment that $x \in \mathbb{R}$ and π is supported on the interval $[0, 1]$. You probably remember from your calculus classes that if f and π are smooth enough (continuously differentiable) then

$$\left| \frac{1}{N} \sum_{k=0}^{N-1} f(x^{(k)})\pi(x^{(k)}) - \int_0^1 f(x)\pi(x)dx \right| = \mathcal{O}\left(\frac{1}{N}\right) \tag{1.3}$$

where, for each $1 \leq k \leq N$, $x^{(k)}$ is any point in the interval $[(k-1)/N,\, k/N]$. The symbol $\mathcal{O}(z)$ will be used to denote terms that are bounded above by some unspecified constant multiple of the number z. Expression (1.3) tells us that, for this scheme, $N(\delta) = \mathcal{O}(1/\delta)$. You may also recall that if we choose $x^{(k)} = (k-0.5)/N$ (and if f and π are twice continuously differentiable) then the error with N points will be $\mathcal{O}(1/N^2)$ so that $N(\delta) = \mathcal{O}(1/\sqrt{\delta})$. More generally, for a deterministic numerical integration scheme of order $\alpha \geq 1$, one has $N(\delta) = \mathcal{O}(\delta^{-1/\alpha})$.

Now suppose that, instead of computing an integral in one dimension, we want to compute

$$\pi[f] = \int_0^1 \int_0^1 f(x, y)\, \pi(x, y)\, dxdy.$$

In this case an order α integration scheme will have $N(\delta) = \mathcal{O}(\delta^{-2/\alpha})$ because both the x and y variables need to be discretized with $\mathcal{O}(\delta^{-1/\alpha})$ samples to achieve an accuracy of δ. In higher dimensions the relationship becomes $N(\delta) = \mathcal{O}(\delta^{-d/\alpha})$ and you can see that numerical integration by these methods becomes hopeless very quickly as d increases. As a rule of thumb, deterministic methods can be used when $d \leq 4$ at most. Recent advances in so-called sparse gridding schemes can extend this bound by several dimensions but the basic exponential increase in cost with dimension remains.

As we will see in a moment when we introduce our first Monte Carlo estimator, we can expect the random numerical integration schemes that we study in this course to satisfy

$$\sqrt{\mathbf{E}\left[\left(\sum_{k=0}^{N-1} f(X^{(k)})W^{(k)} - \pi[f]\right)^2\right]} = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

where $X^{(k)}$ is a sequence of random variables, $W^{(k)}$ are a sequence of random weights (non-negative random variables), and f is any function for which $\pi[f^2]$ is finite. Though the convergence rate on the right hand side is slower than what we found for deterministic integration when d is small, the constant in the $\mathcal{O}$ term is typically only weakly dependent on the dimension of X. Moreover, when f is not smooth, this bound still holds while no deterministic scheme can be accurate. That said, in low dimensions and for typical objective functions f, the increase in accuracy resulting from an increased number of samples is much smaller for Monte Carlo than it is for deterministic methods. That brings us to the first rule of Monte Carlo:

Rule 1. *If you can use a deterministic scheme you probably should.*

1.3 A few concepts from probability

Before we introduce the basic Monte Carlo estimator we will review a few basic notions from probability. Some of these concepts are not needed to introduce most Monte Carlo methods. We introduce them now in the hopes that the reader will be familiar with them when they are needed to analyze and understand the methods that we present.

1.3.1 Probability measures and σ-algebras

First recall that a probability measure, $\mathbf{P}$, is a map from a collection of subsets $\mathcal{F}$ of some (possibly abstract) space Ω to $[0, 1]$ with the following properties:

1. $\mathbf{P}[\Omega] = 1$
2. If $\{A_i\}$ is a countable family of **disjoint** sets in $\mathcal{F}$ then

$$\mathbf{P}\left[\bigcup_{i=0} A_i\right] = \sum_{i=0} \mathbf{P}[A_i]$$

.

The collection $\mathcal{F}$ must be a σ-algebra, i.e. it must satisfy

1. $\Omega \in \mathcal{F}$.
2. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
3. if $\{A_i\}$ is a countable collection of elements of $\mathcal{F}$ then $\bigcup_{i=0} A_i \in \mathcal{F}$.

Note that our requirements in the definition of a probability measures have several immediate and important consequences. For example,

$$\mathbf{P}[\{\}] = 0,$$

$$\text{if} \quad B \subset A \quad \text{then} \quad \mathbf{P}[B] \leq \mathbf{P}[A],$$

and

$$\mathbf{P}\left[\bigcup_{i=0} A_i\right] \leq \sum_{i=0} \mathbf{P}[A_i]$$

for any (not necessarily disjoint) collection $\{A_i\} \subset \mathcal{F}$.

Exercise 1. *Establish the previous three properties of a probability measure from the definition.*

In everything we do below we will be assuming that there is some underlying triplet $(\Omega, \mathcal{F}, \mathbf{P})$ which we refer to as the probability space. One should think of Ω as the set of all possible outcomes of an experiment (which may involve many repetitions of smaller experiments). The collection $\mathcal{F}$ corresponds to collections of true/false statements that one can make about the outcome of an experiment in the sense that each set in $\mathcal{F}$ can be thought of as those outcomes in Ω for which a particular statement is true. For example if Ω consists of all possible versions of today's weather then one set in $\mathcal{F}$ might be all those outcomes for which the temperature in Hyde Park is above 77° (the statement being, "the temperature is above 77°"). In this context the restrictions in the definition of a σ-algebra become transparent: any sensible true or false statement about an experiment has a "complement" (e.g. "the temperature in Hyde Park is not above 77°") and we can string multiple

statements together by putting "or"s between them (which corresponds to taking unions) to get another true/false statement[1]

Example 1. *Let Ω denote the set of all length 100 sequences of H's and T's, and let $\mathcal{F} = \mathcal{P}(\Omega)$ the collection of all subsets of Ω. Let*

$$\mathbf{P}[\omega] = 2^{-100}$$

for each $\omega \in \Omega$. You can easily verify that $\mathcal{F}$ is a σ-algebra and that $\mathbf{P}$ is a probability measure. This is the probability space corresponding to an experiment involving 100 flips of an unbiased coin.

1.3.2 Random variables

A random variable is a measurable map $X : \Omega \to \mathbb{R}^d$. Just as the name implies, the concept of measurability refers to our degree of certainty in the value taken by X. More precisely, we say that X is measurable with respect to a σ-algebra $\mathcal{F}$ if, for each Borel subset $B \subset \mathbb{R}^d$, the set $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ is in $\mathcal{F}$. The Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}^d$ is the smallest σ-algebra on $\mathbb{R}^d$ containing all of the open sets in $\mathbb{R}^d$. It is generally safe to assume that any subset of $\mathbb{R}^d$ you encounter is in $\mathcal{B}$. So measurability of X with respect to $\mathcal{F}$ is simply the requirement that the statement "the value of X is in B" is contained in $\mathcal{F}$ for every B. If we knew the validity of all the statements in $\mathcal{F}$ then the value of X would be known with certainty.

Typically, the ω argument is dropped when writing random variables so that, for example, the probability $\mathbf{P}[\{\omega \in \Omega : X(\omega) \in A\}]$ is written $\mathbf{P}[X \in A]$.

Example 2. *Using the "100 coin flips" probability space described in the previous example we could define the random variable*

$$X(\omega) = \#\ H\text{'s in } \omega.$$

A particularly useful variety of random variables is the so called indicator functions

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

[1] Of course we can also string statements together by putting "and"s between them (which would correspond to taking intersections of sets) but this is taken care of by the other two requirements.

where $A \in \mathcal{F}$.

Exercise 2. *Show that when $A \in \mathcal{F}$, $\mathbf{1}_A$ is a random variable.*

In the previous example we can use indicators to express X as

$$X = \sum_{j=0}^{99} \mathbf{1}_{\{\omega:\, \omega_j = H\}}.$$

1.3.3 Expected values

When it exists, the expected value of a random variable X is the Lebesgue integral of the function X with respect to the probability measure $\mathbf{P}$, i.e.

$$\mathbf{E}[X] = \int X(\omega)\, \mathbf{P}[d\omega].$$

The Lebesgue integral is defined by first requiring that the Lebesgue integral of an indicator function is equal to the probability of the set on which the indicator takes the value 1. In other words,

$$\mathbf{P}[A] = \mathbf{E}[\mathbf{1}_A].$$

This relation is very useful in its own right. Starting from this requirement, integrals of more general functions are defined by approximating those functions by sums of indicator functions. The most important property of the expectation is that it is linear, i.e. if a and b are constant and X and Y are random variables, then

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

For our purposes we can assume that either

(i) X is a *discrete* random variable in which case X takes values in a discrete subset $\{x_i\} \subset \mathbb{R}^d$ and the expected value of $f(X)$ becomes

$$\mathbf{E}[f(X)] = \sum_{i=0}^{\infty} f(x_i)\, \pi_i$$

where $\pi_i = \mathbf{P}[X = x_i]$, or that

(ii) X is a *continuous* random variable, in which case

$$\mathbf{E}\left[f(X)\right] = \int f(x)\,\pi(x)dx$$

where π is the probability density function for the random variable X. That the above expression should hold for all continuous and bounded f can be taken as a definition of the function π. Alternatively we could define

$$\pi(x) = \lim_{|dx|\to 0} \frac{\mathbf{P}\left[X \in dx\right]}{|dx|}$$

where dx is a small volume element containing x in its interior and $|dx|$ is its volume.

Following the common practice in the Monte Carlo literature, we will variously refer the π as a density or distribution depending on the context. When we write $\pi(x)$ we are implicitly assuming that the distribution π has a density and we are using the symbol $\pi(x)$ to represent that density. Regardless of the context we will use the notation

$$\pi[f] = \int f(x)\pi(dx)$$

for the integral of a test function f against a probability distribution π and

$$\pi(A) = \int_A \pi(dx)$$

for the probability of a set $A \subset \mathbb{R}^d$ under π.

Exercise 3. *Show that if* $\mathbf{E}\left[X^2\right] < \infty$ *then* $c = \mathbf{E}\left[X\right]$ *is the value that minimizes the expression* $\mathbf{E}\left[(X-c)^2\right]$. *In other words the expectation of* X *is the best guess of its value in the sense that you expect deviations for the mean to be smaller than deviations from any other constant.*

Another expectation (or moment) that appears frequently in these notes is the covariance

$$\mathbf{cov}(X) = \mathbf{E}\left[\left(X - \mathbf{E}\left[X\right]\right)\left(X - \mathbf{E}\left[X\right]\right)^{\mathrm{T}}\right].$$

Each diagonal entry in this matrix, $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])^2\right]$, is the variance of the 1D random variable X_i. The off-diagonal entries

$$\mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right]$$

are the covariances of the components (often written $\mathbf{cov}(X_i, X_j)$). Notice that $\mathbf{cov}(X)$ is a $d \times d$ positive symmetric definite matrix. The correlation $\rho(X_0, X_1)$ between two one-dimensional random variables is

$$\rho(X_0, X_1) = \frac{\mathbf{cov}(X_0, X_1)}{\sqrt{\mathbf{var}(X_0)\mathbf{var}(X_1)}}.$$

1.3.4 Conditional expectations

Occasionally it is useful to compute the expected value of a random variable given some limited information. For example, one might like to know the expected temperature outside of your house given that it is a sunny day. As we have mentioned, a σ-algebra represents information in the form of true or false statements. When we ask for the expected value of a random variable given some information, that information is encoded in a σ-algebra $\mathcal{G} \subset \mathcal{F}$.

In the example just given, X is the current outside temperature and

$$\mathcal{G} = \{\{\}, \Omega, \{\text{it is sunny}\}, \{\text{it is not sunny}\}\}.$$

If I were to ask you to estimate the current temperature outside of your house without going outside and reading the temperature from a thermometer then you might intuitively answer that your best guess is $\mathbf{E}[X]$, the mean temperature. This is the best guess in the sense that it is the value m minimizing $\mathbf{E}[(X - m)^2]$. But if I then told you that you could look out the window you might revise your guess based on whether you observe the skies to be sunny or cloudy. In fact, before looking out the window you can decide what your guess will be if you see that the skies are sunny and what it will be if you see that the skies are cloudy. The final guess that you will make is unknown (random) until you finally look out the window. Naturally, when weighted by the probabilities of observing it to be a sunny day or a cloudy day, the average of your two guesses should be $\mathbf{E}[X]$. This random variable representing your two guesses is the conditional expectation of X given $\mathcal{G}$.

More precisely, the conditional expectation $\mathbf{E}[X \mid \mathcal{G}]$ is the $\mathcal{G}$ measurable random variable (unique up to a probability zero event) such that

$$\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}]\, \mathbf{1}_A] = \mathbf{E}[X \mathbf{1}_A] \tag{1.4}$$

for all $A \in \mathcal{G}$.

Exercise 4. *If a and b are constants and X and Y are random variables show that*

$$\mathbf{E}[aX + bY \mid \mathcal{G}] = a\mathbf{E}[X \mid \mathcal{G}] + b\mathbf{E}[Y \mid \mathcal{G}].$$

Hint: do this by checking whether the random variable on the right hand side of the last display satisfies the requirements in the definition of the conditional expectation.

Another consequence of this definition is that if Y is $\mathcal{G}$ measurable then

$$\mathbf{E}[XY \mid \mathcal{G}] = Y\, \mathbf{E}[X \mid \mathcal{G}]. \tag{1.5}$$

When $\mathbf{E}[X^2] < \infty$, we can also characterize the conditional expectation as the $\mathcal{G}$-measurable random variable Y minimizing $\mathbf{E}\left[(X - Y)^2\right]$. This confirms our intuition that the conditional expectation is the best (in the sense of mean squared error) estimator of a random variable given some information.

Exercise 5. *Use expression* (1.5) *to establish the assertion in the last sentence.*

Recall that measurability of $\mathbf{E}[X \mid \mathcal{G}]$ with respect to $\mathcal{G}$ means that if the validity of the statements in $\mathcal{G}$ is known then the value of $\mathbf{E}[X \mid \mathcal{G}]$ is certain. In the example this means that once I look out the window and see that it's sunny, my expectation of the temperature outside is a non-random quantity (though the temperature itself remains random).

Exercise 6. *What is $\mathbf{E}[X \mid \mathcal{G}]$ when $\mathcal{G}$ is the trivial σ-algebra $\mathcal{G} = \{\{\}, \Omega\}$?*

Exercise 7. *What is $\mathbf{E}[X \mid \mathcal{G}]$ when X is $\mathcal{G}$-measurable?*

We will use conditional expectations frequently in these notes and specifically we will need the "tower" property:

$$\mathbf{E}\left[\mathbf{E}\left[X \,|\, \mathcal{G}\right] \,|\, \mathcal{F}\right] = \mathbf{E}\left[\mathbf{E}\left[X \,|\, \mathcal{F}\right] \,|\, \mathcal{G}\right] = \mathbf{E}\left[X \,|\, \mathcal{G}\right]$$

anytime $\mathcal{F}$ and $\mathcal{G}$ are two σ-algebras with $\mathcal{G} \subset \mathcal{F}$.

Exercise 8. *Use the definition of conditional expectation to establish the tower property.*

When $\mathcal{G}$ consists only of the sets $\{\}$, Ω, B, and B^{c} for some $B \in \mathcal{F}$ with $0 < \mathbf{P}\left[B\right] < 1$, the conditional expectation takes a particularly simple form:

$$\mathbf{E}\left[X \,|\, \mathcal{G}\right] = \begin{cases} \frac{\mathbf{E}[X\,\mathbf{1}_B]}{\mathbf{P}[B]} & \text{if } \omega \in B \\ \frac{\mathbf{E}[X\,\mathbf{1}_{B^{\mathrm{c}}}]}{\mathbf{P}[B^{\mathrm{c}}]} & \text{otherwise} \end{cases}$$

as can be verified with equation (1.4).

Exercise 9. *Verify this above expression for* $\mathbf{E}\left[X \,|\, \mathcal{G}\right]$ *when* $\mathcal{G} = \left\{\{\}, \Omega, B, B^{c}\right\}$.

In this case we typically write $\mathbf{E}\left[X \,|\, B\right]$ for the value taken by $\mathbf{E}\left[X \,|\, \mathcal{G}\right]$ on the event B. By plugging the random variable $X = \mathbf{1}_A$ into the last display for some event A we obtain a definition for the probability of a set A given a set B satisfying the famous Bayes' formula

$$\mathbf{P}\left[A \,|\, B\right] = \frac{\mathbf{P}\left[A,\, B\right]}{\mathbf{P}\left[B\right]}.$$

When $\mathcal{G} = \sigma(Y)$ (the σ-algebra formed from all true false statements about the value of Y) we typically write $\mathbf{E}\left[X \,|\, Y\right]$ for the conditional expectation of X given $\mathcal{G}$, emphasizing the fact that the value of the conditional expectation is no longer random once the value of Y is revealed. In terms of a probability density $\pi(x, y)$ for two continuous random variables X, and Y, Bayes' formula becomes

$$\begin{aligned} \pi(x \,|\, y) &\equiv \lim_{\max\{|dx|,|dy|\}\to 0} \frac{\mathbf{P}\left[X \in dx \,|\, Y \in dy\right]}{|dx|} \\ &= \lim_{\max\{|dx|,|dy|\}\to 0} \frac{\mathbf{P}\left[X \in dx,\, Y \in dy\right]}{|dx||dy|} \frac{|dy|}{\mathbf{P}\left[Y \in dy\right]} \\ &= \frac{\pi(x, y)}{\pi(y)}. \end{aligned}$$

In this case, the conditional expectation $\mathbf{E}[X \,|\, Y]$ as a function of the possible values taken by Y becomes

$$\begin{aligned}\mathbf{E}[X \,|\, Y = y] &= \lim_{|dy|\to 0} \frac{\mathbf{E}\left[X \mathbf{1}_{\{Y \in dy\}}\right]}{\mathbf{P}[Y \in dy]} \\ &= \frac{\int x\, \pi(x,y) dx}{\pi(y)} \\ &= \int x\, \pi(x \,|\, y) dx\end{aligned}$$

A useful formula to remember when $Y = f(X)$ is

$$\pi(x \,|\, y) = \frac{\delta(y - f(x))\, \pi(x)}{\int \delta(y - f(x))\, \pi(x) dx}.$$

1.3.5 Independence

A collection of random variables $X_0, X_1, X_2 \ldots$ is independent if, for every finite collection of indices $i_0, i_1, \ldots, i_{K-1}$,

$$\mathbf{P}\left(X_{i_0} \in A_0, X_{i_1} \in A_1, \ldots, X_{i_{K-1}} \in A_{K-1}\right) = \prod_{k=0}^{K-1} \mathbf{P}\left(X_{i_0} \in A_0, X_{i_1} \in A_1, \ldots, X_{i_{K-1}} \in A_{K-1}\right)$$

for any collection $A_0, A_1, \ldots A_{K-1}$ of reasonable (say open) subsets of $\mathcal{B}$. Note that, in terms of the density of continuous random variables, independence becomes

$$\begin{aligned}\pi(x_{i_0}, x_{i_1}, \ldots, x_{i_{K-1}}) &= \lim_{\max_j\{|dx_{i_j}|\}\to 0} \frac{\mathbf{P}\left[X_{i_0} \in dx_{i_0}, X_{i_1} \in dx_{i_1}, \ldots, X_{i_{K-1}} \in dx_{i_{K-1}}\right]}{|dx_{i_0}||dx_{i_1}| \cdots |dx_{i_{K-1}}|} \\ &= \lim_{\max_j\{|dx_{i_j}|\}\to 0} \prod_{j=0}^{K-1} \frac{\mathbf{P}\left[X_{i_j} \in dx_{i_j}\right]}{|dx_{i_j}|} \\ &= \prod_{j=0}^{K-1} \pi(x_{i_j}).\end{aligned}$$

Independence implies that, for every finite collection of indices $i_0, i_1, \ldots, i_{K-1}$,

$$\mathbf{E}\left[\prod_{k=0}^{K-1} g_k(X_{i_k})\right] = \prod_{k=0}^{K-1} \mathbf{E}\left[g_k(X_{i_k})\right]$$

for any reasonable (so that the expression makes sense) collections of functions $g_0, g_1, \ldots, g_{K-1}$ from $\mathbb{R}^d$ to $\mathbb{R}$. One special case of this last relation occurs when we compute the covariance of two independent random variables X and Y in which case we obtain

$$\mathbf{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])]\,\mathbf{E}[(Y - \mathbf{E}[Y])] = 0.$$

Note that the reverse implication does not hold: $\mathbf{cov}(X, Y) = 0$ does not imply that X and Y are independent.

Exercise 10. *Can you think of two very simple random variables X and Y taking only values in $\{-1, 0, 1\}$ for which $\mathbf{cov}(X, Y) = 0$ but X and Y are not independent?*

We say that X_{i_1} and X_{i_2} are pairwise independent if the above statements hold with $K = 2$. Notice that pairwise independence of every pair in the collection $X_1, X_2, X_3 \ldots$ does not imply independence of the collection.

Exercise 11. *Suppose that X is a random variable with*

$$\mathbf{P}[X = 1] = \mathbf{P}[X = -1] = \frac{1}{2}$$

and that Y is independent of X and has the same distribution. Let

$$Z = XY.$$

Show that X, Y, Z is are pairwise independent but not independent.

1.4 Some simple probability distributions

1. *Bernoulli(p) distribution*: Suppose $0 < p < 1$. A Bernoulli random variable is a disctrete random variable with

   $$\pi(0) = 1 - p$$

   and

   $$\pi(1) = p.$$

2. *Multinomial(k, p) distribution*: Suppose $0 \leq p_i \leq 1$ for $i = 0, 1, \ldots, n-1$ and $\sum_{i=0}^{n-1} p_i = 1$. For non-negative integers, $m_0, m_1, \ldots, m_{n-1}$ with $\sum_{i=0}^{n-1} m_i = k$,

$$\pi(m_0, m_1, \ldots, m_{n-1}) = \frac{k!}{m_0!\, m_1! \ldots m_{n-1}!}\, p_0^{m_0} p_1^{m_1} \ldots p_{n-1}^{m_{n-1}}$$

3. *Poisson(λ) distribution*: Suppose $\lambda > 0$ is a real number. Then

$$\pi(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

is the probability distribution for a non-negative integer valued random variable.

4. *Uniform(a,b) distribution*: Suppose $a < b$ are real numbers. Then

$$\pi(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

is the probability density function for a continuous random variable taking values in (a, b).

5. *Exponential(λ) distribution*: Suppose that $\lambda > 0$ is a real number. Then

$$\pi(x) = \lambda e^{-\lambda x}$$

is the probability density function for a continous random variable taking values in $(0, \infty)$.

6. *Single variable Gaussian distribution* ($\mathcal{N}(m, \sigma^2)$): Suppose that m and $\sigma > 0$ are real numbers. Then

$$\pi(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-m)^2}{2\sigma^2}}$$

is the probability density function for a continuous random variable taking values in $\mathbb{R}$.

7. *Multivariate Gaussian distribution* ($\mathcal{N}(m, C)$): Suppose that $m \in \mathbb{R}^d$ and that C is a symmetric positive definite $d \times d$ matrix. Then

$$\pi(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|C|}}\, e^{-\frac{(x-m)^{\mathrm{T}} C^{-1} (x-m)}{2}}$$

is the probability density function for a continuous random variable taking values in $\mathbb{R}^d$.

1.5 Monte Carlo and notions of convergence

Our goal is usually to estimate the mean of some function $f(X)$ of a random variable X,

$$\pi[f] = \int f(x)\pi(dx).$$

The simplest Monte Carlo estimator is of the form

$$\overline{f}_N = \frac{1}{N}\sum_{k=0}^{N-1} f(X^{(k)})$$

where the $X^{(k)}$ are separate realizations of random variables whose distributions are close to that of X (so that, in particular, $\mathbf{E}\left[f(X^{(k)})\right] \approx \pi[f]$). In practice the most we can hope for is that the distribution of $X^{(k)}$ converges to that of X when k is very large. Construction of algorithms with that property will be a significant goal of this course.

For the moment we will assume that the distribution of each of the $X^{(k)}$ is exactly the distribution of X. Notice that in this case,

$$\mathbf{E}\left[\overline{f}_N\right] = \frac{1}{N}\sum_{k=0}^{N-1} \mathbf{E}\left[f(X^{(k)})\right]$$

so that, in particular, if for each k, $\mathbf{E}\left[f(X^{(k)})\right] = \pi[f]$, then our estimator $\overline{f}_N$ (which is a random variable) would be unbiased, i.e.

$$\mathbf{E}\left[\overline{f}_N\right] = \pi[f].$$

Our hope, of course is that

$$\overline{f}_N \to \pi[f] \qquad \text{as} \qquad N \to \infty.$$

Suppose that $\mathbf{var}(f(X)) = \sigma^2$. Then one can easily compute that

$$\begin{aligned}\mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2\right] &= \frac{1}{N^2}\mathbf{E}\left[\sum_{k,\ell=0}^{N}\left(X^{(k)} - \pi[f]\right)\left(X^{(\ell)} - \pi[f]\right)\right] \\ &= \frac{1}{N^2}\mathbf{E}\left[\sum_{k=0}^{N-1}\left(X^{(k)} - \pi[f]\right)^2\right] + \frac{2}{N^2}\sum_{k<\ell}\mathbf{cov}(X^{(k)}, X^{(\ell)}).\end{aligned}$$

Under our assumption that the distribution of each $X^{(k)}$ is identical to the distribution of X, the first term is a sum of identical values so that we obtain

$$\mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2\right] = \frac{\sigma^2}{N} + \frac{2}{N^2}\sum_{k<\ell}\mathbf{cov}(X^{(k)}, X^{(\ell)}).$$

Though we have specified the distribution of each $X^{(k)}$ individually, we have not specified the joint distribution of the entire collection $\{X^{(k)}\}_{k=0}^{N-1}$. If we make the assumption that the collection $\{X^{(k)}\}_{k=0}^{\infty}$ is pairwise independent then, as we have seen earlier in these notes, their covariances vanish and with them the second double summation term leaving only the first term σ^2/N which vanishes as $N \to 0$. Moreover, there is no reason to believe that σ^2 will depend on dimension. When the samples are pairwise independent random variables, the number of samples to achieve accuracy δ as measured by the root mean squared deviation

$$\mathbf{rmse}(\overline{f}_N) = \sqrt{\mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2\right]}$$

is $N(\delta) = \mathcal{O}(1/\delta^2)$. This is the source of the assertion that the cost of a Monte Carlo scheme is independent of dimension. However, the second double sum may or may not vanish as N increases because it involves $\mathcal{O}(N^2)$ summands. Moreover, The dependence of Monte Carlo schemes on dimension is largely determined by the this second term. Finding schemes that keep that term small is our primary goal in designing new algorithms.

There are a few other forms of convergence beyond convergence in **rmse** (or L^2) that we might be interested in. The most basic is *convergence in probability*

$$\lim_{N\to\infty} \mathbf{P}\left[|\overline{f}_N - \pi[f]| > \delta\right] = 0 \qquad \text{for all } \delta > 0.$$

This is called the *Weak Law of Large Numbers* (WLLN) and is implied by convergence in RMSE because

$$\begin{aligned}\mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2\right] &= \mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2 \mathbf{1}_{\{|\overline{f}_N - \pi[f]|>\delta\}}\right] \\ &\qquad + \mathbf{E}\left[\left(\overline{f}_N - \pi[f]\right)^2 \mathbf{1}_{\{|\overline{f}_N - \pi[f]|\le\delta\}}\right] \\ &> \delta^2\,\mathbf{P}\left[|\overline{f}_N - \pi[f]| > \delta\right].\end{aligned}$$

The WLLN is weaker than the *Strong Law of Large Numbers* (SLLN) which states that

$$\lim_{N\to\infty} \overline{f}_N(\omega) = \pi[f]$$

for all ω in a subset of Ω that has probability 1. We will not show it here, but our assumptions above that the $X^{(k)}$ be independent and identically distributed (i.i.d.) with $\mathbf{E}\left[f(X^{(k)})\right] = \pi[f]$ is enough to guarantee that the SLLN holds.

The final form of convergence that we need to define is *convergence in distribution*. For example, one might ask what happens to moments like $\mathbf{E}\left[g(\overline{f}_N)\right]$ for some continuous and bounded objective function g, as $N \to \infty$. It won't surprise the reader that the convergence results above all imply that the distribution of $\overline{f}_N$ converges to a delta function at $\pi[f]$, i.e. $\mathbf{E}\left[g(\overline{f}_N)\right]$ converges to $g(\pi[f])$. More interesting is the limiting distribution of the scaled error, $Z_N = \sqrt{N}\left(\overline{f}_N - \pi[f]\right)$. Note that $\mathbf{E}\left[Z_N\right] = 0$ and, when the $X^{(k)}$ are i.i.d., the calculations above imply that $\mathbf{E}\left[Z_N^2\right] = \sigma^2$. In other words, the factor of $\sqrt{N}$ is just enough to keep the scaled error from vanishing as N is increased. It is natural to ask if Z_N converges to some non trivial (i.e. non-constant) random variable. The appropriate notion of convergence for this question is convergence in distribution.

There are certain classes of functions that are rich enough to completely specify the limit of the distribution of a sequence of random variables. One example, when their expectations exist, is the family of functions of the form $g(x) = e^{\lambda x}$ for $\lambda \in \mathbb{R}$. This family is continuous but not bounded so there are many distributions for which the expectation of these functions is infinite. When the expectations of the functions in this family do exist they are given a special name:

$$\Lambda_N(\lambda) = \mathbf{E}\left[e^{\lambda Z_N}\right] \tag{1.6}$$

is called the moment generating function (to see why differentiate it at $\lambda = 0$). Convergence of Λ_N for λ in an open interval containing 0 to the moment generating function of some other random variable Z is enough to guarantee that Z_N converges to Z in distribution.

Let's return to the case in which $X^{(k)}$ are i.i.d random variables drawn from p and now assume that $\mathbf{E}\left[e^{\lambda X}\right]$ is finite for λ in an interval containing 0.

Notice that for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda Z_N}\right] = \left(\mathbf{E}\left[e^{\frac{\lambda}{\sqrt{N}}(X^{(1)}-\pi[f])}\right]\right)^N. \tag{1.7}$$

Applying the expansion

$$f(x) = f(0) + x\, f'(0) + \frac{x^2}{2} f''(0) + \int_0^x \frac{(x-y)^2}{2} f'''(y) dy$$

to $f(x) = e^x$ with $Y = \lambda(X^{(1)} - \pi[f])/\sqrt{N}$ we obtain

$$\mathbf{E}\left[e^{\frac{\lambda}{\sqrt{N}}\sum_{i=1}^N (X^{(k)}-\pi[f])}\right] = \left(1 + \frac{\lambda^2}{2N}\sigma^2 + \mathbf{E}\left[\int_0^Y \frac{(Y-y)^2}{2} e^y dy\right]\right)^N \tag{1.8}$$

Note that arriving at this expression we have carried out the expectation exactly for the terms involving f' and f''. The term involving the integral in the last display is bounded above by

$$\frac{1}{3}\mathbf{E}\left[Y^3 e^{|Y|}\right]$$

which, for small enough λ, we expect is roughly of size $\mathcal{O}\left(N^{-3/2}\right)$ (this is true for all λ if the $X^{(k)}$ are bounded random variables). Neglecting this smallest term we find that

$$\lim_{N\to\infty} \mathbf{E}\left[e^{\lambda Z_N}\right] = e^{\frac{\lambda^2}{2}}.$$

for those values of λ. But, as one can check, this is the moment generating function of a Gaussian random variable with mean 0 and standard deviation 1.

Exercise 12. *Find the moment generating function for an $\mathcal{N}(0,1)$ random variable.*

This is an example of the *Central Limit Theorem* (CLT). In more general situations (correlated $X^{(k)}$) one can occasionally prove versions of the LLN and CLT. The LLN is easier to establish in general, but it guarantees only that if you take N to be very large you will get a reasonable estimate. It does not tell you (as the CLT does) how fast you can expect the error to decrease as you increase N.

To better understand what the CLT does and does not say let's consider one final type of convergence result. The *Large Deviations Principle* LDP for $\overline{f}_N$ says (when it holds) roughly that for some constant $\gamma(\epsilon)$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left(\left|\overline{f}_N - \pi[f]\right| > \epsilon\right) = -\gamma(\epsilon),$$

i.e. that the probability of a "large deviation" of the estimate vanishes exponentially (with rate γ) as $N \to \infty$. Note that the LDP usually predicts a different "tail behavior" than one might infer from the CLT. If we naively interpret the CLT we might believe that

$$\begin{aligned}\mathbf{P}\left[\overline{f}_N - \pi[f] > \epsilon\right] &= \mathbf{P}\left[Z_N > \sqrt{N}\epsilon\right] \\ &\approx \frac{1}{\sqrt{2\pi\sigma^2}} \int_{z>\sqrt{N}\epsilon} e^{-\frac{z^2}{2\sigma^2}} dz.\end{aligned}$$

Making the change of variables $y = z + \sqrt{N}\,\epsilon$ gives

$$\mathbf{P}\left[\overline{f}_N - \pi[f] > \epsilon\right] \approx \frac{e^{-\frac{N\epsilon^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \int_{y>0} e^{-\frac{y^2}{2\sigma^2}} e^{-\frac{y\sqrt{N}\epsilon}{\sigma^2}} dy.$$

When N is large, the function $e^{-\frac{y\sqrt{N}\,\epsilon}{\sigma^2}}$ is concentrated near 0 and gives almost no weight to regions farther from 0 (than say $N^{-1/4}$). Within such a narrow strip the function $e^{-\frac{y^2}{2\sigma^2}}$ is effectively equal to 1. To a good approximation, we can ignore the function $e^{-\frac{y^2}{2\sigma^2}}$ when we compute the integral. Our CLT based reasoning thus leads us to the approximation

$$\mathbf{P}\left[\overline{f}_N - \pi[f] > \epsilon\right] \approx \frac{\sigma}{\epsilon\sqrt{2N\pi}} e^{-\frac{N\epsilon^2}{2\sigma^2}}$$

i.e.

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\overline{f}_N - \pi[f] > \epsilon\right] = -\frac{\epsilon^2}{2\sigma^2}.$$

Unfortunately (unless the $X^{(k)}$ were Gaussian) this will not be the correct rate of decay of the probability. In the *i.i.d.* case again, Cramér's Large Deviation Theorem tells us that the correct rate of decay is (with a few qualifications)

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\overline{f}_N - \pi[f] \in (a,b)\epsilon\right] = -\inf_{x-\pi[f]\in(a,b)} I(x)$$

where $I(x)$ is the Legendre transform of the logarithm of the moment generating function Λ for X, i.e.

$$I(x) = \sup_{\lambda} \{x\lambda - \log \Lambda(\lambda)\}.$$

Our error was in applying the CLT to estimate the probability that Z_N lies within a set, $\left\{z > \sqrt{N}\epsilon\right\}$, that is shrinking too quickly with N.

Example 3. *To see the failure of the CLT approach for computing the probabilities of large deviations, suppose that the $X^{(k)}$ are exponentially distributed with mean 1, i.e. they are drawn according to the density*

$$\pi(x) = e^{-x}.$$

The moment generating function corresponding to this density it

$$\Lambda(\lambda) = \int_0^\infty e^{(\lambda-1)x} dx = \begin{cases} \frac{1}{1-\lambda}, & \text{for } \lambda < 1 \\ \infty, & \text{otherwise} \end{cases}.$$

The supremum defining the Legendre transform is obtained when

$$\lambda = 1 - \frac{1}{x}$$

so that

$$I(x) = x - 1 - \log x.$$

Using this function we find that, for example,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\overline{x}_N - 1 > \epsilon\right] = -\epsilon + \log(1+\epsilon).$$

Note that by the (flawed) CLT based reasoning in the previous paragraph we would obtain

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\overline{x}_N - 1 > \epsilon\right] = -\frac{\epsilon^2}{2}.$$

As is often the case in scientific computing, one is typically more interested in the relative error

$$\mathbf{rel\_err}(\overline{f}_N) = \frac{\mathbf{rmse}(\overline{f}_N)}{\pi[f]}$$

of a Monte Carlo estimator $\overline{f}_N$ of $\pi[f]$. When $\pi[f]$ is very small, controlling the relative error can be very difficult and requires either a very large number of samples N, or a very carefully designed estimator. The next exercise shows that the standard estimator of a small probability will have very large relative error.

Exercise 13. *Write a subroutine that takes N as an argument and generates a sample of the estimator $\overline{x}_N$ from the previous example (Python has a routine to generate N samples from the distribution $exp(1)$). Then write a routine that calls your subroutine to generate many copies of $\overline{x}_N$ and produces a histogram of the values of $\sqrt{N}\left(\overline{x}_N - \pi[x]\right)$. Produce this histogram for several values of N and show that for large N, the histograms approach the Gaussian density. A quantile–quantile (QQ) plot is a plot of the quantiles (i.e. the inverse of the cumulative distribution function) of two 1-dimensional distributions against one another. If the resulting curve is $y = x$, the distributions are the same. This is most often used when at least one of the distributions is empirical (i.e. a collection of samples) and you want to know how close those samples are to some specific distribution. Produce QQ plots to accompany your histograms.*

Next write a routine that constructs an estimate Q_N of the probability

$$p_N = \mathbf{P}\left[\overline{x}_N - 1 > 0.1\right]$$

by generating many samples of $\overline{x}_N$ (remember that probabilities are expectations of indicators). To make your estimator less costly to evaluate, it may help to recall that the $exp(\lambda)$ distribution is the same as the $Gamma(1, \lambda)$ distribution and that sums of gamma random variables, all of which have the same second parameter, is again a gamma random variable. Try to demonstrate the rate of decay we found in the last example. Estimating this quantity will require a huge number of samples of $\overline{x}_N$ as N increases. Write down a formula for the standard deviation of Q_N in terms of p_N. and compare it to p_N. Which, the standard deviation of Q_N or p_N, decays faster (you can answer this either by numerical test or by mathematical argument)?

1.6 bibliography

Chapter 2

A few exact sampling techniques

The basic building block of all of the methods that we will discuss are routines like `random()` in the $C/C++$ programming language which, with each call, returns an integer in the range $[0, M]$ where M is a very large positive integer. The best of these routines generate a deterministic periodic sequence of integers that, for most practical purposes, is indistinguishable from a random independent sequence of integers chosen uniformly (each one as likely as any other). Random number generators are interesting in their own right but we will not discuss them here.

Assuming that your random number generator (we'll assume it's `random()`) is actually producing an independent sequence of uniformly chosen integers, one can easily construct a very good approximation of a sequence of independent $\mathcal{U}(0, 1)$ random variables by the transformation

$$U = \frac{\texttt{random}()}{M}$$

with appropriate modifications if outcomes of $U = 0$ or $U = 1$ are problematic. In the next few sections we'll introduce techniques that can be used to transform samples from one distribution (e.g. $\mathcal{U}(0, 1)$) that can be easily generated into samples from a more complicated distribution.

2.1 Inversion

Suppose that our goal is to generate a sample from the distribution of a random variable X for which we have the function

$$F(x) = \mathbf{P}\left[X \leq x\right].$$

The function F is called the probability distribution function for X. This function is increasing and right continuous with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Exercise 14. *Use the requirements of a probability measure from Chapter 1 to show that F is increasing, right continuous, and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.*

If F is differentiable then X has density $\pi = F'$.

Exercise 15. *Suppose that $X \in \mathbb{R}$ has density π and distribution function F. Use the definition of the Riemann integral to show that, for any continuous, bounded function f,*

$$\mathbf{E}\left[f(X)\right] = \int f(x)F'(x)dx,$$

i.e. $\pi = F'$.

In fact, this function completely characterizes the statistical behavior of a random variable. Let

$$F^{\dagger}(u) = \inf\left\{x : F(x) \geq u\right\}.$$

If F happens to be one-to-one then $F^{\dagger}$ is just the usual inverse of F. Now notice that

$$\left\{u :\ F^{\dagger}(u) \leq y\right\} = \left\{u : u \leq F(y)\right\}$$

since on the one hand if $F(y) \geq u$ then $F^{\dagger}(u) \leq y$ by the definition of $F^{\dagger}$ and on the other hand note that by right continuity of F and the definition of $F^{\dagger}$, $F(F^{\dagger}(u)) \geq u$, so that since F is increasing $F^{\dagger}(u) \leq y$ implies that $u \leq F(y)$.

Therefore, if $U \sim \mathcal{U}(0,1)$, the probability distribution function of $Y = F^\dagger(U)$ is

$$\begin{aligned}\mathbf{P}[Y \leq y] &= \mathbf{P}\left[F^\dagger(U) \leq y\right] \\ &= \mathbf{P}[U \leq F(y)] \\ &= F(y),\end{aligned}$$

i.e. we have succeeded in generating a sample with the distribution of X.

Example 4. *Suppose we want to generate a single index Y from the set $\{0, 1, \ldots, n-1\}$ so that $\mathbf{P}[Y = i] = p_i$ where $p_i \geq 0$ and $\sum_{0=1}^{n-1} p_i = 1$. The probability distribution function for Y is*

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ s(j) & \text{if } x \in [j, j+1) \text{ for some } j \in \{0, 1, \ldots, n-2\} \\ 1 & \text{if } x \geq n-1 \end{cases}$$

where $s(j) = \sum_{i=0}^{j} p_i$. We also find that for $u \in [0,1)$,

$$F^\dagger(u) = \begin{cases} 0 & \text{if } u \in [0, s(0)) \\ j & \text{if } u \in [s(j-1), s(j)) \text{ for some } j \in \{1, 2, \ldots, n-1\}. \end{cases}$$

That $Y = F^\dagger(U)$ for $U \sim \mathcal{U}(0,1)$ has the correct distribution is intuitively clear since the length of each interval $[s(j-1), s(j))$ is p_j.

Example 5. *Suppose our goal is to sample an exponential random variable with parameter λ, i.e.*

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

We have just shown that if $U \sim \mathcal{U}(0,1)$ then

$$Y = F^\dagger(U) = -\frac{1}{\lambda}\log(1-U)$$

is distributed according to $exp(\lambda)$.

Exercise 16. *Consider the distribution on $(0,1)$ with*

$$\pi(x) = \frac{1}{2\sqrt{x}}.$$

Write a code that uses inversion to generate samples from π using samples from $\mathcal{U}(0,1)$ and assemble a histogram of the output of your scheme. Produce a QQ plot to graphically compare your samples to π.

2.2 Change of variables

In the last section we showed that a particular change of variables could always be used (assuming you can evaluate F^{-1}) to generate samples with a desired distribution. Let's now think slightly more generally about what happens to a continuous distribution under a change of variables. Let's see what happens when we apply some smooth invertible function φ to the variables sampled from a density $\tilde{\pi}$. Its easiest to consider what happens to the moments of $\tilde{\pi}$:

$$\int f(\varphi(y))\tilde{\pi}(y)dy.$$

Make the change of variables $x = \varphi(y)$. In 1D we obtain

$$\int f(x)|\varphi'(\varphi^{-1}(x))|^{-1}\tilde{\pi}(\varphi^{-1}(x))dx.$$

The variable $X = \varphi(Y)$ then has density

$$\pi(x) = |\varphi'(\varphi^{-1}(x))|^{-1}\tilde{\pi}(\varphi^{-1}(x)).$$

In higher dimensions the formula becomes

$$\pi(x) = \frac{\tilde{\pi}(\varphi^{-1}(x))}{|\det\left(D\varphi(\varphi^{-1}(x)\right)|}$$

with

$$(D\varphi)_{ij} = \frac{\partial \varphi_i}{\partial y_j}.$$

Thinking of the above calculations in the opposite direction, suppose the goal is to generate samples from π but we are unable to do that directly. Suppose that for some smooth, invertible change of variables φ we are able to efficiently generate samples $Y^{(k)}$ from $\tilde{\pi}(y) = |\det\left(D\varphi(y)\right)|\,\pi(\varphi(y))$. Then by taking $X^{(k)} = \varphi^{-1}(Y^{(k)})$ we obtain samples from π.

Example 6. *Box-Muller Suppose that u_0 and u_1 are independent $\mathcal{U}(0,1)$ random variables. Define the function $\varphi : [0,1]^2 \to \mathbb{R}^2$ by*

$$\varphi_1(u_0,u_1) = \sqrt{-2\log u_0}\cos(2\pi u_1), \qquad \varphi_2(u_0,u_1) = \sqrt{-2\log u_0}\sin(2\pi u_1).$$

The Jacobian of this transformation is the determinant of the matrix

$$\begin{bmatrix} -\frac{1}{\sqrt{-2\log u_0}u_0}\cos(2\pi u_1) & -2\pi\sqrt{-2\log u_0}\sin(2\pi u_1) \\ -\frac{1}{\sqrt{-2\log u_0}u_0}\sin(2\pi u_1) & -2\pi\sqrt{-2\log u_0}\cos(2\pi u_1) \end{bmatrix}$$

which can easily be computed to obtain $|det(D\varphi)| = \frac{2\pi}{u_0}$. *In terms of the variables* $(y_0, y_1) = \varphi(u_0, u_1)$, u_0 *can be written*

$$u_0 = e^{-\frac{y_0^2+y_1^2}{2}}.$$

The density of $X = \varphi(U)$ *is therefore*

$$\pi(x) = \frac{1}{2\pi}e^{-\frac{x_0^2+x_1^2}{2}}$$

(the π *on the right hand side of the last display is the area of the unit disk and not the density* $\pi(x)$*).*

Exercise 17. *Write a routine that generates two independent Gaussian random variables. Use your code to generate many samples and produce a (2 dimensional) histogram of the results. Produce a QQ plot to graphically compare the distribution of the samples you generate to the standard normal distribution.*

Exercise 18. *Use a change of variables similar to the one you used for the Gaussian to generate a uniformly distributed sample on the unit disk given two independent samples from* $\mathcal{U}(0,1)$ *and produce a (2 dimensional) histogram to verify your code.*

2.3 Rejection

For the algorithm described in this subsection we again assume that our goal is to draw samples from the density π. Assume that we can draw samples from $\tilde{\pi}$ instead and that for some constant $K \geq 1$,

$$\pi \leq K\tilde{\pi}. \tag{2.1}$$

We generate a sample $X \sim \pi$ by generating pairs $(Y^{(k)}, U^{(k)})$ of independent random variables with $Y^{(k)} \sim \tilde{\pi}$ and $U^{(k)} \sim \mathcal{U}(0,1)$ until index τ when

$$U^{(\tau)} \leq \frac{\pi(Y^{(\tau)})}{K\tilde{\pi}(Y^{(\tau)})}$$

at which point we set $X = Y^{(\tau)}$.

The first question one must ask is whether this scheme returns a sample in finite time, i.e. is $\tau < \infty$? We will return to this question in a moment. For now, we assume that $\mathbf{P}[\tau < \infty] = 1$. Having made that assumption, breaking up the event $\{Y^{(\tau)} \in dx\}$ according to the value of τ and plugging in the definition of τ yields the expansion

$$\begin{aligned}\mathbf{P}\left[Y^{(\tau)} \in dx\right] = \mathbf{P}\left[Y^{(\tau)} \in dx,\, \tau < \infty\right] &= \sum_{k=1}^{\infty} \mathbf{P}\left[Y^{(k)} \in dx,\, \tau = k\right] \\ &= \sum_{i=1}^{\infty} \mathbf{P}\left[Y^{(k)} \in dx,\, U^{(k)} \leq \frac{\pi(Y^{(k)})}{K\tilde{\pi}(Y^{(k)})},\, U^{(\ell)} > \frac{\pi(Y^{(\ell)})}{K\tilde{\pi}(Y^{(\ell)})}\,\forall\, \ell < k\right]\end{aligned}$$

Using the fact that the different samples are independent we can factor the last expression to obtain

$$\begin{aligned}\mathbf{P}\left[Y^{(\tau)} \in dx\right] = \sum_{k=1}^{\infty} \mathbf{P}\left[Y^{(k)} \in dx,\, U^{(k)} \leq \frac{\pi(Y^{(k)})}{K\tilde{\pi}(Y^{(k)})}\right] \\ \times \mathbf{P}\left[U^{(1)} > \frac{\pi(Y^{(1)})}{K\tilde{\pi}(Y^{(1)})}\right]^{k-1}\end{aligned}$$

Finally, appealing to the relation

$$\begin{aligned}\mathbf{P}\left[Y^{(k)} \in dx,\, U^{(k)} \leq \frac{\pi(Y^{(k)})}{K\tilde{\pi}(Y^{(k)})}\right] = \mathbf{P}\left[U^{(k)} \leq \frac{\pi(Y^{(k)})}{K\tilde{\pi}(Y^{(k)})} \,\middle|\, Y^{(k)} \in dx\right] \\ \times \mathbf{P}\left[Y^{(k)} \in dx\right]\end{aligned}$$

and the uniform distribution of the $U^{(\ell)}$ variables we conclude that

$$\begin{aligned}\lim_{|dx|\to 0} \frac{\mathbf{P}\left[Y^{(\tau)} \in dx\right]}{|dx|} &= \sum_{k=1}^{\infty} \left(1 - \int \frac{\pi(y)}{K\tilde{\pi}(y)}\tilde{\pi}(y)dy\right)^{k-1} \frac{\pi(x)}{K\tilde{\pi}(x)}\tilde{\pi}(x) \\ &= \frac{\pi(x)}{K} \sum_{k=0}^{\infty} \left(1 - \frac{1}{K}\right)^{k} \\ &= \pi(x)\end{aligned}$$

Note that to use this algorithm you need to relate some distribution that you can easily sample (say a Gaussian) to the distribution that you'd actually like to sample, by a bound of the form (2.1). In all but some simple cases this is not possible. Nonetheless, rejection sampling can be a useful component of more complicated sampling schemes.

We return now to considering the cost of this algorithm. As mentioned above, one must first address the question of whether or not $\tau < \infty$. Fortunately, again decomposing the event $\tau < \infty$ by the possible values taken by τ and using the independence of the variables, we find that

$$\begin{aligned}\mathbf{P}(\tau < \infty) &= \sum_{k=1}^{\infty} \mathbf{P}(\tau = k) \\ &= \frac{1}{K} \sum_{k=1}^{\infty} \left(1 - \frac{1}{K}\right)^{k-1} \\ &= 1\end{aligned}$$

so the algorithm will at least exit properly with probability one (in fact we already assumed this with our previous calculation).

While returning a sample from π in finite time is of course a very basic requirement, it is hardly enough. We would also like to know how much computational effort will be expended to generate that sample. The cost of the scheme is characterized by the expectation of τ. We can compute the the mean time to exit as

$$\begin{aligned}\mathbf{E}[\tau] = \mathbf{E}\left[\sum_{k=1}^{\infty} \mathbf{1}_{\{\tau \leq k\}}\right] &= \sum_{k=1}^{\infty} \mathbf{E}\left[\mathbf{1}_{\{\tau \leq k\}}\right] = \sum_{k=1}^{\infty} \mathbf{P}(\tau \geq k) \\ &= \sum_{k=1}^{\infty} \left(1 - \frac{1}{K}\right)^{k-1} \\ &= K.\end{aligned}$$

Clearly, if the K chosen is not optimal, or if the best possible K so that (2.1) holds is very large, the rejection algorithm will be very costly.

Exercise 19. *Write a routine to generate a single sample from the uniform measure on the unit disk from two independent samples from $\mathcal{U}(0,1)$ using*

rejection. Make sure you clearly identify the target density π, the reference density $\tilde{\pi}$, and your choice of K (which should be justified). There's a natural choice of K for this problem that allows you to apply the sampling algorithm without having to know the area of the unit disk in advance... what is it? Verify your code by producing a histogram. Compare (numerically) the cost of this approach with the more direct approach in Exercise 18 by comparing, e.g. the expected number of $\mathcal{U}(0,1)$ variables required per sample from the unit disk and the expected wall clock time per sample from the unit disk.

2.4 bibliography

Chapter 3

Importance Sampling

Now we move on to schemes that do not produce exact (up to the floating point and periodicity issues mentioned in the last section) samples but that can be applied to far more complex sampling problems. The first family of algorithms of this kind that we will consider are called importance sampling methods. In their simplest form, these methods produce very simple unbiased estimators comprised of sums of independent random variables. More precisely, suppose your goal is to compute

$$\pi[f] = \int f(x)\pi(dx).$$

The simplest estimator is

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

where the $X^{(k)}$ are independent and all sampled from π. Recall that this estimator is unbiased and that we can compute its **rmse**. There are two possible drawbacks to this algorithm. The first is that it may be very costly or impossible to generate independent samples from π. The second difficultly is that for may problems the **rmse** may be unacceptably large so that a reasonable estimate requires very large N. Now suppose that $\tilde{\pi}$ is some other

distribution that we can sample. We can then try to construct the estimator

$$\widetilde{f}_N = \frac{1}{N}\sum_{k=1}^{N} f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$$

where the $Y^{(k)}$ are independent samples from $\tilde{\pi}$. It will often be convenient to write this estimator as

$$\widetilde{f}_N = \sum_{k=1}^{N} f\left(Y^{(k)}\right) W^{(k)}$$

where

$$W^{(k)} = \frac{1}{N}\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}.$$

We will assume that if $\mathrm{supp}(f)$ is the set of points x for which $f(x) \neq 0$, then

$$\mathrm{supp}(f\pi) \subset \mathrm{supp}(f\tilde{\pi}).$$

If the random variable the random variable $f(X)$ was integrable then $f(Y^{(k)})\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ is also integrable and

$$\mathbf{E}\left[f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right] = \int f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\tilde{\pi}(dy) = \int f(y)\pi(dy) = \mathbf{E}\left[f(X)\right]$$

and the estimator $\widetilde{f}_N$ is unbiased.

As we did for $\overline{f}_N$, if the random variables $f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ have finite variance we can easily compute that

$$\mathbf{rmse}(\widetilde{f}_N) = \frac{\sqrt{\mathbf{var}\left(f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right)}}{\sqrt{N}}.$$

Since, for any random variable X with finite variance we have

$$\mathbf{var}(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2$$

and the mean of $f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ is $\pi[f]$,

$$\begin{aligned}\mathbf{var}\left(f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right) &= \int\left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2\tilde{\pi}(dy) - (\pi[f])^2 \\ &= \int (f(x))^2\frac{\pi(x)}{\tilde{\pi}(x)}\pi(dx) - (\pi[f])^2\end{aligned}$$

Example 7. *Consider importance sampling when $f = 1$, π is $\mathcal{N}(0,1)$, and $\tilde{\pi}$ is $\mathcal{N}(0,\sigma^2)$. The error is*

$$\mathbf{rmse}\left(\widetilde{f}_N\right) = \frac{1}{\sqrt{N}}\sqrt{\frac{\sigma}{\sqrt{2\pi}}\int e^{-\left(1-\frac{1}{2\sigma^2}\right)x^2}dx - 1}.$$

As soon as σ^2 becomes less than 1/2, the error becomes infinite, illustrating the fact that the tails of the reference density should, in general, be heavier than the tails of the target density (or at least of $|f|$ times the target density).

3.1 Optimal importance sampling

The goal in selecting an importance sampling reference density is to choose the reference density that results in an estimator $\widetilde{f}_N$ that has lower variance than $\overline{f}_N$. From our last computation we can focus our efforts on choosing a $\tilde{\pi}$ for which

$$\int \left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2 \tilde{\pi}(dy)$$

is as low as possible. The optimal choice of $\tilde{\pi}$ can easily be identified. Indeed, Jensen's inequality implies that

$$\int \left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2 \tilde{\pi}(dy) \geq \left(\int |f(y)|\pi(dy)\right)^2$$

and this lower bound is achieved by

$$\tilde{\pi}(x) = \frac{|f(x)|\pi(x)}{\int |f(y)|\pi(dy)}.$$

Of course it is extremely unlikely that one can sample from (or even evaluate) this optimal density even if you could sample from π. You'll notice that the optimal density involves a computation (the normalization constant) very similar to the original problem. The is an example of one of the central tenants of Monte Carlo: the more you know about the answer the better the solution you can design. In most practical situations one applies intuition about the problem at hand to design a reasonable $\tilde{\pi}$. However there are

some interesting situations in which one can derive mathematically justifiable choices of reference density. In the chapter on rare event simulation I provide one such example.

Exercise 20. *Use samples from* $\mathcal{N}(m, \sigma^2)$ *to estimate* $\mathbf{P}[X > 2]$ *for* $X \sim \mathcal{N}(0, 1)$ *using importance sampling. By comparing the variances of the estimators for different* m *and* σ*, draw conclusions about the values of* m *and* σ *that yield the best estimators.*

3.2 Normalization constants and an alternative estimator

Examining the estimator $\widetilde{f}_N$ more closely, you'll notice that to use it we need to evaluate the ratio $\pi/\tilde{\pi}$. In many applications this is only possible up to an unknown multiplicative constant. In this section we'll describe an importance sampling strategy for estimating this constant and use it to build an alternative importance sampling estimator that only requires that we can evaluate $\pi/\tilde{\pi}$ up to the unknown constant.

Given a non-negative and integrable function p, the normalization constant (or partition function) is the value $\mathcal{Z}_p = \int p(x)dx$. The problem of estimating a normalization constant arrises in a wide range of applications. In statistical mechanics one is often interested in computing the normalization constant for a family of densities indexed by some parameter θ, i.e.

$$\mathcal{Z}_{p_\theta} = \int p_\theta(dx).$$

Here, for each The function

$$F(\theta) = -\log \mathcal{Z}_{p_\theta}$$

is called a free energy. The marginal distribution

$$\pi(x) = \int \pi(x, y)dy$$

of the x variables determined by some joint distribution $\pi(x, y)$ is another quantity of frequent interest in statistical mechanics where $-\log \pi(x)$ is again referred to a free energy.

We have already seen that in Bayesian statistics, given a prior distribution $\pi(\theta)$ on the parameters and a data likelihood $\pi(y \,|\, \theta)$, a common goal is to sample from the posterior distribution

$$\pi(\theta \,|\, y) = \frac{\pi(y \,|\, \theta)\pi(\theta)}{\pi(y)}$$

where θ is a collection of parameters and y is a realization of the data. The normalization constant in this expression, $\pi(y)$ (the marginal density of the data), is called the model evidence and is used to distinguish between putative statistical models.

We'll now describe a very simple estimator of the ratio $\mathcal{Z}_p/\mathcal{Z}_q$ of the normalization constants of two non-negative, integrable functions p and q. Suppose that you can sample from the density

$$\tilde{\pi} = \frac{q(x)}{\mathcal{Z}_q}$$

and let $\pi = p/\mathcal{Z}_p$. Our importance sampling estimator for averages with respect to π with reference density $\tilde{\pi}$ satisfies

$$\widetilde{1}_N \longrightarrow \pi[1] = 1.$$

Multiplying both sides of this expression by $\mathcal{Z}_p/\mathcal{Z}_q$ we obtain the estimator

$$\frac{1}{N}\sum_{k=1}^{N} \frac{p(Y^{(k)})}{q(Y^{(k)})} \longrightarrow \frac{\mathcal{Z}_p}{\mathcal{Z}_q}.$$

Exercise 21. *Write a routine that uses $\mathcal{N}(0,1)$ samples to estimate the normalization constant for the density proportional to $e^{-|x|^3}$.*

Based on the approximation of the ratio of normalization constants above, a reasonable modification to the standard importance sampling estimator to deal with unknown normalization constants is

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = \frac{1}{N}\sum_{k=1}^{N} \frac{f(Y^{(k)})\frac{p(Y^{(k)})}{q(Y^{(k)})}}{\frac{1}{N}\sum_{\ell=1}^{N}\frac{p(Y^{(\ell)})}{q(Y^{(\ell)})}} = \sum_{k=1}^{N} f\left(Y^{(k)}\right) W^{(k)}$$

where now

$$W^{(k)} = \frac{\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}}{\sum_{\ell=1}^{N} \frac{\pi(Y^{(\ell)})}{\tilde{\pi}(Y^{(\ell)})}}.$$

The WLLN implies that the numerator in the rightmost expression in the last display converges to $\pi[f]$ and the denominator converges to 1. Therefore $\widetilde{f}_N/\widetilde{1}_N$ converges to $\pi[f]$. However, inspecting the mean we see that in general

$$\mathbf{E}\left[\frac{\widetilde{f}_N}{\widetilde{1}_N}\right] \neq \frac{\mathbf{E}\left[\sum_{k=1}^{N} f\left(Y^{(k)}\right) \frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right]}{\mathbf{E}\left[\sum_{k=1}^{N} \frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right]} = \pi[f],$$

i.e. $\widetilde{f}_N/\widetilde{1}_N$ is a biased estimator.

We can estimate the size of this bias by writing

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = h\left(\widetilde{f}_N, \widetilde{1}_N\right)$$

where $h(x, y) = x/y$ and taylor expanding around the means $(\pi[f], 1)$ (in this context this technique is refered to as the delta method). Let

$$\gamma(t) = \begin{pmatrix} \pi[f] \\ 1 \end{pmatrix} (1-t) + \begin{pmatrix} \widetilde{f}_N \\ \widetilde{1}_N \end{pmatrix} t, \qquad t \in [0,1].$$

We have

$$\begin{aligned} h\left(\widetilde{f}_N, \widetilde{1}_N\right) &= h(\gamma(0)) + \frac{d}{dt}\left(h(\gamma(t))\right)\Big|_{t=0} + \int_0^1 (1-s)\frac{d^2}{ds^2}\left(h(\gamma(s))\right) ds \\ &= h(\pi[f],1) + (\widetilde{f}_N - \pi[f])\partial_x h(\pi[f],1) + (\widetilde{1}_N - \pi[f])\partial_y h(\pi[f],1) \\ &\quad + (\widetilde{f}_N - \pi[f])^2 \int_0^1 (1-s)\partial_x^2 h(\gamma(s))ds \\ &\quad + (\widetilde{1}_N - 1)^2 \int_0^1 (1-s)\partial_y^2 h(\gamma(s))ds \\ &\quad + (\widetilde{f}_N - \pi[f])(\widetilde{1}_N - 1) \int_0^1 2(1-s)\partial_{xy} h(\gamma(s))ds. \end{aligned}$$

Since $\partial_x h(\pi[f], 1) = 1$ and $\partial_y h(\pi[f], 1) = -\pi[f]$, we find that the expectations of the second and third terms in the last expression vanish. We'll simply

pretend that the three integrals appearing in the formula are bounded and recall that

$$\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)^2\right] = \mathcal{O}\left(\frac{1}{N}\right), \quad \mathbf{E}\left[\left(\widetilde{1}_N - 1\right)^2\right] = \mathcal{O}\left(\frac{1}{N}\right),$$

and, by the Cauchy-Schwartz inequality,

$$\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)\left(\widetilde{1}_N - 1\right)\right] \leq \sqrt{\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)^2\right]}\sqrt{\mathbf{E}\left[\left(\widetilde{1}_N - 1\right)^2\right]}.$$

These calculations yield the basic conclusion that the bias of $\widetilde{f}_N/\widetilde{1}_N$ is smaller than the standard deviation of $\widetilde{f}_N$ (which are of order $N^{-1/2}$) and should therefore not trouble us too much in many applications.

To be more confident that the bias is small we would need to know more about the likelihood of very small values of $\widetilde{1}_N$ (which will result in large values for the integral terms we have ignored). Cramer's theorem tells us that the probability that $\widetilde{1}_N < \delta$ for any $\delta < 1$ is exponentially small in N, but this is not enough since, for example, if the event $\widetilde{1}_N = 0$ occurs with positive probability then the bias is infinite. At the cost of an additional small bias, the estimator can be modified so that these issues are avoided.

Our primary motivation for introducing the estimator $\widetilde{f}_N/\widetilde{1}_N$ was that the densities π and $\tilde{\pi}$ might only be known up to a multiplicative constant making it impossible to assemble $\widetilde{f}$. Is there a reason to prefer the biased estimator $\widetilde{f}_N/\widetilde{1}_N$ when unknown normalization constants are not an issue? Let's consider the mean squared error

$$\mathbf{rmse}^2\left(\frac{\widetilde{f}_N}{\widetilde{1}_N}\right) = \mathbf{E}\left[\left(\frac{\widetilde{f}_N}{\widetilde{1}_N} - \pi[f]\right)^2\right].$$

Using the same expansion of h that we used above, we obtain

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

Since the first two terms in the last display are $\mathcal{O}\left(N^{-1/2}\right)$ we'll neglect the

$\mathcal{O}(N^{-1})$ terms to obtain

$$\begin{aligned}\mathbf{rmse}^2\left(\frac{\widetilde{f}_N}{\widetilde{1}_N}\right) &\approx \mathbf{E}\left[\left(\widetilde{f}_N - \pi[f] - \pi[f](\widetilde{1}_N - 1)\right)^2\right] \\ &= \mathbf{var}\left(\widetilde{f}_N\right) + \frac{(\pi[f])^2}{N}\,\mathbf{var}\left(\frac{\pi(Y)}{\tilde{\pi}(Y)}\right) \\ &\qquad - \frac{2\pi[f]}{N}\,\mathbf{cov}\left(f(Y)\,\frac{\pi(Y)}{\tilde{\pi}(Y)}, \frac{\pi(Y)}{\tilde{\pi}(Y)}\right)\end{aligned}$$

where Y is distributed according to $\tilde{\pi}$.

For certain choices of f the covariance in the last display will be small or negative (e.g. if $f = \tilde{\pi}/\pi$) and the mean squared error of $\widetilde{f}_N/\widetilde{1}_N$ will be larger than that for $\widetilde{f}_N$. However, in many cases this covariance will be large and $\widetilde{f}_N/\widetilde{1}_N$ will have smaller error. As a dramatic example, suppose that f is nearly constant. Then $\pi[f]$ is approximately equal to this constant and the **rmse** of $\widetilde{f}_N/\widetilde{1}_N$ nearly vanishes.

Exercise 22. *Repeat the last exercise for the importance sampling estimator $\widetilde{f}_N/\widetilde{1}_N$ instead of $\widetilde{f}_N$. Which of these two estimators do you prefer? Does the answer depend on m and σ?*

3.3 Importance sampling in high dimensions

As a general rule, the need for a good approximation of the optimal importance sampling estimator becomes more acute in high dimensions. As a general measure of the quality of a reference density within the context of importance sampling one can consider the variance of the importance sampling weights, i.e.

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \mathbf{var}\left(\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right).$$

We use the symbol χ^2 because $\mathbf{var}\left(\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right)$ is Pearson's χ^2-divergence between π and $\tilde{\pi}$. Much like the relative entropy, $\chi^2(\pi \,\|\, \tilde{\pi})$ is a very strong measure of the distance between π and $\tilde{\pi}$ (though it is not a distance) in the

sense that it bounds the total variation distance between π and $\tilde{\pi}$. Indeed, by Jensen's inequality,

$$\begin{aligned}\chi^2(\pi \,\|\, \tilde{\pi}) &= \int \left| \frac{\pi(y)}{\tilde{\pi}(y)} - 1 \right|^2 \tilde{\pi}(dy) \\ &\geq \left(\int \left| \frac{\pi(y)}{\tilde{\pi}(y)} - 1 \right| \tilde{\pi}(dy) \right)^2 \\ &= 4\|\pi - \tilde{\pi}\|_{\mathrm{TV}}^2\end{aligned}$$

Suppose that the goal is to construct an estimator of $\pi[f]$ with **rmse** equal to δ. In the last section we estimated the error of the estimator $\widetilde{f}_N/\widetilde{1}_N$ as

$$\mathbf{rmse}^2(\widetilde{f}_N/\widetilde{1}_N) \approx \mathbf{E}\left[\left(\widetilde{f}_N - \pi[f] - \pi[f](\widetilde{1}_N - 1) \right)^2 \right]$$

where Y is a random variable distributed according to $\tilde{\pi}$. The expression on the right hand side of the last display is equal to

$$\frac{1}{N}\mathbf{E}\left[(f(Y) - \pi[f])^2 \left(\frac{\pi(Y)}{\tilde{\pi}(Y)} \right)^2 \right].$$

If we make the (usually very severe) assumption that the variables $f(Y)$ and $\pi(Y)/\tilde{\pi}(Y)$ are independent then we find that

$$\mathbf{rmse}^2(\widetilde{f}_N/\widetilde{1}_N) \approx \frac{1}{N}\mathbf{var}\left(f(X)\right)\left(1 + \chi^2(\pi \,\|\, \tilde{\pi})\right)$$

where X is distributed according to π. On the other hand, for the standard estimator using M independent samples from π, we know that

$$\mathbf{rmse}^2\left(\overline{f}_M\right) = \frac{\mathbf{var}\left(f(X)\right)}{M}.$$

As a consequence, we can make a very rough estimate the number of samples required by the estimator $\overline{f}$ to achieve the same accuracy as $\widetilde{f}_N/\widetilde{1}_N$ as

$$M \approx \frac{N}{1 + \chi^2(\pi \,\|\, \tilde{\pi})}.$$

The term on the right hand side of the last display is referred to as the effective sample size, ess_N, of the importance sampling estimator $\widetilde{f}_N/\widetilde{1}_N$. It gives a rough estimate of the number of independent samples from π that would be of similar statistical quality to the N weighted samples generated in importance sampling. By this measure, when $\chi^2(\pi \,\|\, \tilde{\pi})$ is large we expect importance sampling to yield poor results.

As the following example illustrates, unless the reference distribution is chosen very carefully we expect resampling to perform poorly in high dimensions.

Example 8. *Consider importance sampling the d dimensional Gaussian target density $\pi = \mathcal{N}(m, C)$ using the reference density $\tilde{\pi} = \mathcal{N}(m, (1-\alpha)C)$ for some $\alpha \in (0,1)$. Under π or $\tilde{\pi}$, the density of the function $V(x) = (x-m)^T C^{-1}(x-m)$ is independent of the choice of m and C (it is a chi-squared random variable with d degrees of freedom). As can be easily computed, the variance of the importance weights is*

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \int \left(\frac{\pi(x)}{\tilde{\pi}(x)}\right) \pi(dx) - 1 = \frac{1}{(1-\alpha^2)^{\frac{d}{2}}} - 1.$$

For fixed α, $\chi^2(\pi \,\|\, \tilde{\pi})$ grows exponentially with d.

We can further illustrate the failure of importance sampling in high dimensions by considering the case in which π is the density of d i.i.d. random variables, i.e.

$$\pi(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i)$$

for some density p of a single variable. One might encounter a density of this kind when assimilating d observations of an experiment and assuming that the error in the various observations are independent. Let's suppose that you want to use a reference density of the same form,

$$\tilde{\pi}(y_1, y_2, \ldots, y_d) = \prod_{i=1}^{d} q(y_i).$$

Then the independence of the $Y_j^{(k)}$ yeilds

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \sqrt{\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right]^d} - 1.$$

When $\tilde{\pi} \neq \pi$ we will have that

$$\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right] > \mathbf{E}\left[\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right]^2 = 1$$

and $\chi^2(\pi \,\|\, \tilde{\pi})$ will increase exponentially with d.

The preceding calculation should give pause to anyone seeking to use importance sampling in very high dimensional systems. In fact, it suggests that, when using importance sampling at least, Monte Carlo suffers from exactly the same problem as the deterministic integration schemes that we discussed earlier. It is important to keep in mind, however, that the case of independent, identically distributed components is a very special one. In fact typical high dimensional problems almost always exhibit low dimensional structure. This means that high dimensional densities encountered in typical sampling applications tend to concentrate on a lower dimensional subspace. Given our demonstration above of the dangers of importance sampling in high dimensions, one might take comfort in the observation of lower dimensional structure in high dimensional sampling problems. But often one has little or no information about the lower dimensional structure of the distribution, in which case that structure actually makes the problem much more difficult than the independent identically distributed components setting. One is forced to use a reference density $\tilde{\pi}$ for which the variance of $\pi/\tilde{\pi}$ may be extremely high.

There are, however, important situations in which importance sampling can be used to great advantage in high dimensions. The key to success, of course, is choosing a reference density sufficiently close to the optimal importance sampling density. In the next example, the marginal distributions, $\pi(x_j)$, change with dimension, and by choosing a reference density that respects the correct scaling with dimension we can ensure $\chi^2(\pi \,\|\, \tilde{\pi})$ is bounded for all d.

Example 9. *As in the discussion above, assume that*

$$\pi(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i) \quad \textit{and} \quad \tilde{\pi}(y_1, y_2, \ldots, y_d) = \prod_{i=1}^{d} q(Y_i).$$

Suppose that $p = \mathcal{N}(0, d^{-1})$ *and that* $q = \mathcal{N}(d^{-1}, d^{-1})$. *Then*

$$\begin{aligned}\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right] &= \frac{\sqrt{d}}{\sqrt{2\pi}}\int e^{-dy^2+d(y-d^{-1})^2}e^{-\frac{d(y-d^{-1})^2}{2}}dy \\ &= \frac{\sqrt{d}e^{\frac{1}{d}}}{\sqrt{2\pi}}\int e^{-\frac{d(y+d^{-1})^2}{2}}dy \\ &= e^{\frac{1}{d}}\end{aligned}$$

so that $\chi^2(\pi \,\|\, \tilde{\pi})$ *is stable as* $d \to \infty$. *This example is related to importance sampling for diffusions which we'll return to in Part II.*

3.4 Sequential importance sampling and resampling

Occasionally, the structure of a problem allows one to break a high dimensional sampling problem into manageable pieces. We can always decompose a multidimensional density π as

$$\pi(x_{1:d}) = \pi(x_1)\prod_{n=2}^{d}\pi(x_n \,|\, x_{1:n-1}) \tag{3.1}$$

where we have introduced the more compact notation

$$x_{m:n} = (x_m, x_{m+1}, \ldots, x_n) \quad \text{for} \quad m \le n.$$

To see this just recall that

$$\pi(x_n \,|\, x_{1:n-1}) = \frac{\pi(x_{1:n})}{\pi(x_{1:n-1})}.$$

The decomposition in (3.1) suggests a sampling strategy for π: first sample X_1 from $\pi(x_1)$ and then, at step n given the components $X_{1:n-1}$ generated so far, generate X_n from $\pi(x_n \,|\, X_{1:n-1})$.

Example 10. *Consider generating a simple random walk of length* d *on a periodic lattice* $\mathbb{Z}_L^2 = \{0, 1, \ldots, L-1\} \times \{0, 1, \ldots, L-1\}$. *A walk on the lattice*

of length d (an element of SRW(d)) is just a chain of states $x_{1:d}$ with x_{s+1} a neighbor on the lattice of x_s (denoted here $x_{s+1} \leftrightarrow x_s$) for all $s < d$. The density for a simple random walk on the lattice is

$$\pi(x_{1:d}) = \frac{1}{\mathcal{Z}_d} \begin{cases} 1 & \text{if } x_{1:d} \text{ is a walk on the lattice} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{Z}_d = L^2 4^{d-1}$ is the normalization constant.

In this case, the marginal

$$\pi(x_{1:n}) = \sum_{x_{n+1:d}} \pi(x_{1:d})$$

is just the density for the simple random walk of length n. We can easily compute that, if $x_{1:n-1}$ is a walk on the lattice,

$$\pi(x_n \,|\, x_{1:n-1}) = \frac{\mathcal{Z}_n}{\mathcal{Z}_{n-1}} = \begin{cases} 1/4 & \text{if } x_{1:n} \in SRW(n) \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, we can sample the walk by choosing an initial point, X_1, uniformly and then, at step $n > 1$, picking a neighbor of X_{n-1} uniformly.

Of course it is unlikely that we will know enough about π (and it's marginal and conditional densities) to carry out this procedure. You can check that even in the simple random walk example, if I permute the indices in the decomposition (3.1) it is already harder to see how to use the decomposition to sample from π. Fortunately, this decomposition strategy can sometimes be salvaged with the help of importance sampling. In fact, though the sequential importance sampling and resampling strategy that we develop in this section was originally introduced to sequentially construct samples from high dimensional distributions, it is now used in applications ranging from the simulation of rare events to online data assimilation.

3.4.1 Sequential importance sampling

We will form a reference density $\tilde{\pi}$ for use in importance sampling of π by replacing the various terms in the decomposition (3.1) by approximations. In

more detail, given a density $\tilde{\pi}_1(x_1)$ and conditional densities $q_n(x_n \,|\, x_{1:n-1})$ define the sequence of reference densities

$$\tilde{\pi}_n(x_{1:n}) = \tilde{\pi}_1(x_1) \prod_{\ell=2}^{n} q_\ell(x_\ell \,|\, x_{1:\ell-1}) \tag{3.2}$$

and set $\tilde{\pi} = \tilde{\pi}_d$. We will assume that one can sample from $\tilde{\pi}_1$ and the conditional densities q_n and evaluate them up to a normalization constant. Notice that the sequence of densities $\tilde{\pi}_n$ is closed under marginalization in the sense that, for $m \leq n$,

$$\tilde{\pi}_n(x_{1:m}) = \tilde{\pi}_m(x_{1:m}).$$

The normalized importance sampling estimator for an average with respect to π using reference density $\tilde{\pi}$ is

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = \sum_{k=1}^{N} f\left(Y_{1:d}^{(k)}\right) W_d^{(k)}$$

where $Y_{1:d}^{(k)}$ are samples from $\tilde{\pi}$ and

$$W_d^{(k)} = \frac{\pi(Y_{1:d}^{(k)})/\tilde{\pi}(Y_{1:d}^{(k)})}{\sum_{\ell=1}^{N} \pi(Y_{1:d}^{(\ell)})/\tilde{\pi}(Y_{1:d}^{(\ell)})}.$$

Comparing to the decomposition of π in (3.1), it is clear that our importance sampling estimator will perform well when

$$\tilde{\pi}_1(x_1) \approx \pi(x_1) \quad \text{and} \quad q_n(x_n \,|\, x_{1:n-1}) \approx \pi(x_n \,|\, x_{1:n-1}).$$

Our immediate goal is to represent this estimator in terms of a recursion on dimension. The generation of the samples $Y_{1:n}^{(k)}$ from $\tilde{\pi}_n$ is naturally carried out by recursion: given a sample $Y_{1:n-1}^{(k)}$ from $\tilde{\pi}_{n-1}$ we can build a sample $Y_{1:n}^{(k)}$ sampled from $\tilde{\pi}_n$ by first sampling $Y_n^{(k)}$ from $q_n(x_n \,|\, Y_{1:n-1}^{(k)})$ and then setting $Y_{1:n}^{(k)} = (Y_{1:n-1}^{(k)}, Y_n^{(k)})$.

We now find a recursion for the weights $W_n^{(k)}$. To do this we need to introducing a sequence of densities

$$\pi_n(x_{1:n})$$

for $n = 1, 2, \ldots, d$ with $\pi_d = \pi$. Note that we are not assuming that the marginal under the target density of the first n variables, $\pi(x_{1:n})$, is equal to $\pi_n(x_{1:n})$. We do assume that one can evaluate the ratios π_n/π_{n-1} up to an unknown normalization constant, but not that you can sample directly from π_n. We will characterize a recursion for importance sampling estimators for each of the π_n using reference density $\tilde{\pi}_n$ and built off of the estimator for π_{n-1}. This recursion has no immediate utility as the the resulting estimator for π is exactly the usual normalized importance sampling estimator for averages with respect to π using reference density $\tilde{\pi}$. It will become very useful later in this section when we introduce the notion of resampling.

The normalized importance sampling estimator for an average with respect to π_n using the reference density $\tilde{\pi}_n$ would use weights

$$W_n^{(k)} = \frac{\pi_n(Y_{1:n}^{(k)})/\tilde{\pi}_n(Y_{1:n}^{(k)})}{\sum_{\ell=1}^N \pi_n(Y_{1:n}^{(\ell)})/\tilde{\pi}_n(Y_{1:n}^{(\ell)})}$$

where $Y_{1:n}^{(k)}$ is drawn from $\tilde{\pi}_n$. Now observe that

$$\frac{\pi_n(x_{1:n})}{\tilde{\pi}_n(x_{1:n})} = \frac{\pi_{n-1}(x_{1:n-1})}{\tilde{\pi}_{n-1}(x_{1:n-1})} w_n(x_{1:n})$$

where we have defined the function

$$w_n(x_{1:n}) = \frac{\pi_n(x_{1:n})}{\pi_{n-1}(x_{1:n-1})\, q_n(x_n \,|\, x_{1:n-1})}.$$

As a consequence we obtain the recursion

$$W_n^{(k)} = \frac{W_{n-1}^{(k)}\, w_n(Y_{1:n}^{(k)})}{\sum_{\ell=1}^N W_{n-1}^{(\ell)}\, w_n(Y_{1:n}^{(\ell)})}.$$

It is important to note that, under our assumptions, one can evaluate w_n up to a multiplicative constant that cancels in the normalization of the $W_n^{(k)}$.

Example 11. *A chain $x_{1:d}$ of states $x_s \in \mathbb{Z}_L^2$ with $x_{s+1} \leftrightarrow x_s$ for $s < d$ and $x_t \neq x_s$ for all $s, t \leq d$ is called a self avoiding walk of length d (an element of SAW(d)). Imagine sampling from the density $\pi(x_{1:d})$ defined by*

$$\pi(x_{1:d}) = \frac{1}{\mathcal{Z}_d} \begin{cases} 1 & \text{if } x_{1:d} \in SAW(d) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{Z}_d$ is the (now unknown) normalization constant. One could imagine using importance sampling directly to compute averages with respect to π using as a reference density, the uniform measure on chains satisfying $x_{s+1} \leftrightarrow x_s$ for all $s < d$ (we don't know the normalizing constants so we'd have to use $\widetilde{f}_N/\widetilde{1}_N$). But it is very unlikely that a chain from this reference density would satisfy $x_s \neq x_t$ for $s, t \leq d$ and most of our effort would be spent generating samples that would later be assigned weight 0.

We can use sequential importance sampling instead. First, define π_n for $n \leq d$ as the uniform measure on SAW(n). In contrast with the simple random walk problem in the last example, the distribution π_{n-1} is not quite the marginal density $\pi(x_{1:n}) = \sum_{x_{n+1:d}} \pi(x_{1:d})$ of the first n states in a chain of length d drawn from π. To see this, observe that for $n < d$, an element $x_{1:n} \in SAW(n)$ for which there are no neighbors of x_n that have not already been visited, has $\pi_n(x_{1:n}) > 0$, but $\pi(x_{1:n}) = 0$. But our sequential importance sampling framework does not require this.

The factor w_n needed to update the weights can be written as

$$w_n(x_{1:n}) = \frac{\pi_n(x_n \,|\, x_{1:n-1})\, \pi_n(x_{1:n-1})}{q_n(x_n \,|\, x_{1:n-1})\, \pi_{n-1}(x_{1:n-1})}$$

Moreover,

$$\begin{aligned} \pi_n(x_{1:n-1}) &= \sum_{x_n} \pi_n(x_{1:n}) \\ &= \begin{cases} \frac{m_n(x_{1:n-1})}{\mathcal{Z}_n} & x_{1:n-1} \in SAW(n-1) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where

$$m_n(x_{1:n-1}) = \left|\{x_n : x_{1:n} \in SAW(n)\}\right|.$$

This implies that, as long as $m_n(x_{1:n-1}) > 0$,

$$\pi_n(x_n \,|\, x_{1:n-1}) = \frac{\pi_n(x_{1:n})}{\pi_{n-1}(x_{1:n-1})} = \begin{cases} \frac{1}{m_n} & \text{if } x_{1:n} \in SAW(n) \\ 0 & \text{otherwise.} \end{cases}$$

When $m_n(x_{1:n-1}) > 0$, this conditional distribution is easy enough to sample from and makes a natural choice for q_n. Given a chain in $SAW(n-1)$, one

simply chooses x_n from among those neighbors of x_{n-1} that have not yet been reached by the chain. Chains for which this is not possible ($m_n = 0$) will receive 0 weight and can be discarded. Indeed, having made this choice for q_n, the weight factors become

$$w_n(x_{1:n}) = \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} = \frac{m_n(x_{1:n-1})\mathcal{Z}_{n-1}}{\mathcal{Z}_n}.$$

Finally, notice in addition that for this q_n, the ratio of successive normalization constants $\mathcal{Z}_n/\mathcal{Z}_{n-1}$ can be written

$$\begin{aligned}\frac{\mathcal{Z}_n}{\mathcal{Z}_{n-1}} &= \sum_{x_{1:n}} m_n(x_{1:n-1}) q_n(x_n \,|\, x_{1:n-1}) \pi_{n-1}(x_{1:n-1}) \\ &= \sum_{x_{1:n-1}} m_n(x_{1:n-1}) \pi_{n-1}(x_{1:n-1})\end{aligned}$$

3.4.2 Sequential importance sampling with resampling

We have already demonstrated the difficulties suffered by importance sampling in high dimensions. If used as just described this scheme too should be expected to fail for large d unless $\tilde{\pi}$ happens to be a very good approximation of π. The utility of this recursive form of importance sampling is only fully exploited when we combine it with resampling. That is, instead of carrying samples with very low weight we replace low weight samples with copies of high weight samples in a statistically consistent manner. More precisely, assuming that at step n we have a weighted ensemble of N_n samples, $\left\{W_n^{(k)}, X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$, approximately drawn from π_n, in the sense that

$$\sum_{k=1}^{N_n} W_n^{(k)} f\left(X_{1:n}^{(k)}\right) \approx \int f(x_{1:n}) \pi_n(dx_{1:n})$$

for any test function f, we

1. Resample the weighted ensemble $\left\{W_n^{(k)}, X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$ to obtain a uniformly weighted ensemble $\left\{1/N, Y_{1:n}^{(k)}\right\}_{k=1}^{N_{n+1}}$ approximately drawn from π_n in

the sense that

$$\frac{1}{N}\sum_{k=1}^{N_{n+1}} f\left(Y_{1:n}^{(k)}\right) \approx \int f(x_{1:n})\pi_n(dx_{1:n})$$

for any test function f.

2. For $k = 1, 2, \ldots, N_{n+1}$ generate $X_{n+1}^{(k)}$ from $q_{n+1}(x_{n+1} \,|\, Y_{1:n}^{(k)})$ and set

$$X_{1:n+1}^{(k)} = \left(Y_{1:n}^{(k)}, X_{n+1}^{(k)}\right).$$

3. Compute the weights

$$W_{n+1}^{(k)} = \frac{w_{n+1}(X_{1:n+1}^{(k)})}{\sum_{\ell=1}^{N_{n+1}} w_{n+1}(X_{1:n+1}^{(\ell)})}.$$

The number of samples after resampling, N_n, need not be deterministic but will, in general, be close to the user specified value N. The basic technique used to generate the unweighted ensemble is to make multiple copies of samples in the weighted ensemble with large weights and to discard samples in the weighted ensemble with small weights. Note the absence of the weights $W_n^{(k)}$ in the formula for the new weights $W_{n+1}^{(k)}$. The previous weights are already accounted for in the duplication or removal of samples from the $\left\{X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$ ensemble. This resampling is done at each step in the recursion with the goal being to devote our computational resources only to those samples with a reasonable chance of contributing to the final average at step d.

In order to explain the concept of resampling in more detail, it is useful to view the ensemble of samples at any iteration of the scheme as a weighted empirical measure, i.e. consider the random distribution

$$\Psi_n(x_{1:n}) = \sum_{k=1}^{N_n} W_n^{(k)}\, \delta\left(x_{1:n} - X_{1:n}^{(k)}\right)$$

corresponding to the ensemble of samples generated by the above steps after n iterations. Note that Ψ_n is not quite a probability distribution unless $\sum_{k=1}^{N_n} W_n^{(k)} = 1$. Knowledge of Ψ_n is equivalent to knowledge of the ensemble of the weighted samples $\{W_n^{(k)}, X_{1:n}^{(k)}\}$.

Step 1 above corresponds to, starting from Ψ_n, generating a new random distribution

$$\tilde{\Psi}_n(x_{1:n}) = \frac{1}{N}\sum_{k=1}^{N_n} N_n^{(k)}\,\delta\left(x_{1:n} - X_{1:n}^{(k)}\right)$$

where $N_{n+1} = \sum_{k=1}^{N_n} N_n^{(k)}$ and the $N_n^{(k)}$ are random, non-negative integers satisfying

$$\mathbf{E}\left[N_n^{(k)} \,|\, \{W_n^{(\ell)}\}_{\ell=1}^{N_n}\right] = NW_n^{(k)}. \tag{3.3}$$

Defining a new collection of N_{n+1} points $\{Y_{1:n}^{(\ell)}\}_{\ell=1}^{N_{n+1}}$, exactly $N_n^{(k)}$ elements of which are equal to $X_{1:n}^{(k)}$, this last distribution can be rewritten as

$$\tilde{\Psi}_n(x_{1:n}) = \frac{1}{N}\sum_{k=1}^{N_{n+1}} \delta\left(x_{1:n} - Y_{1:n}^{(k)}\right).$$

In Steps 2 and 3 above, the samples $Y_{1:n}^{(k)}$ are augmented with a sample $X_{n+1}^{(k)}$ from $q_{n+1}\left(x_{n+1} \,|\, Y_{1:n}^{(k)}\right)$ to obtain $X_{1:n+1}^{(k)} = \left(Y_{1:n}^{(k)}, X_{n+1}^{(k)}\right)$ which is then weighted by

$$W_{n+1}^{(k)} = \frac{w_{n+1}(X_{1:n+1}^{(k)})}{\sum_{\ell=1}^{N_{n+1}} w_{n+1}(X_{1:n+1}^{(\ell)})}$$

to obtain the new distribution

$$\Psi_{n+1}(x_{1:n+1}) = \sum_{k=1}^{N_{n+1}} W_{n+1}^{(k)}\,\delta\left(x_{1:n+1} - X_{1:n+1}^{(k)}\right)$$

One possible choice for the distribution of the $\{N_n^{(k)}\}$ that satisfies condition (3.3) is the $Multinomial(N, p)$ distribution with the vector p having entries $p_k = W_n^{(k)}$, in which case $N_n = N$ exactly at each step. Notice that, if the $N_n^{(k)}$ are selected from $Multinomial(N, \{W_n^{(k)}\})$, then, since the variance of $N_n^{(k)}$ is $NW_n^{(k)}\left(1 - W_n^{(k)}\right)$,

$$\begin{aligned}\mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)^2 \,\middle|\, \Psi_n\right] &= \frac{1}{N^2}\mathbf{E}\left[\left(NW_n^{(k)} - N_n^{(k)}\right)^2 \,\middle|\, \Psi_n\right] \\ &= \frac{1}{N}W_n^{(k)}\left(1 - W_n^{(k)}\right).\end{aligned}$$

Similarly, since the covariance of $N_n^{(k)}$ and $N_n^{(\ell)}$ for $i \neq j$ is $-NW_n^{(k)}W_n^{(\ell)}$,

$$\mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) \,\middle|\, \Psi_n\right] = -\frac{W_n^{(k)}W_n^{(\ell)}}{N}.$$

These expressions imply that the error from a single resampling step is

$$\begin{aligned}
&\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \,\middle|\, \Psi_n\right] \\
&\quad = \mathbf{E}\left[\left(\sum_{\ell=1}^{N}\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) f(X_{1:n}^{(\ell)})\right)^2 \,\middle|\, \Psi_n\right] \\
&\quad = \sum_{\ell=1}^{N}\left(f(X_{1:n}^{(\ell)})\right)^2 \mathbf{E}\left[\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right)^2 \,\middle|\, \Psi_n\right] \\
&\qquad\quad + 2\sum_{k<\ell\leq N} f(X_{1:n}^{(k)})f(X_{1:n}^{(\ell)}) \\
&\qquad\qquad\qquad \times \mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) \,\middle|\, \Psi_n\right] \\
&\quad = \frac{1}{N}\sum_{\ell=1}^{N}\left(f(X_{1:n}^{(\ell)})\right)^2 W_n^{(\ell)}\left(1 - W_n^{(\ell)}\right) \\
&\qquad\qquad\qquad - \frac{2}{N}\sum_{k<\ell\leq N} f(X_{1:n}^{(k)})f(X_{1:n}^{(\ell)})W_n^{(k)}W_n^{(\ell)} \\
&\quad = \frac{1}{N}\sum_{k=1}^{N}\left(f(X_{1:n}^{(k)}) - \sum_{\ell=1}^{N} f(X_{1:n}^{(\ell)})W_n^{(\ell)}\right)^2 W_n^{(k)}. \qquad (3.4)
\end{aligned}$$

When f is bounded, i.e. when $\|f\|_\infty < \infty$, this last expression is bounded by $\|f\|_\infty^2/N$.

For each n, and any function f which takes $x_{1:n}$ as its argument, define the new function

$$\mathcal{Q}_n f(x_{1:n-1}) = \int f(x_{1:n})w_n(x_{1:n})q_n(dx_n \,|\, x_{1:n-1})$$

which takes $x_{1:n-1}$ as its argument. Note that

$$\begin{aligned}\|Q_n f\|_\infty &\leq \|f\|_\infty \left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \int \pi_n(dx_n \,|\, x_{1:n-1}) \right\|_\infty \\ &= \|f\|_\infty \left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \right\|_\infty .\end{aligned}$$

We'll assume that there is some, possibly unknown, constant K so that

$$\left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \right\|_\infty \leq K$$

for all n so that $\|Q_n f\|_\infty \leq K \|f\|_\infty$.

The total error in the sequential importance sampling scheme after n steps can be decomposed as follows,

$$\begin{aligned}\int f(x_{1:n}) \left(\Psi_n - \pi_n\right)(dx_{1:n}) &= \int f(x_{1:n}) \Psi_n(dx_{1:n}) \\ &\quad - \int \mathcal{Q}_n f(x_{1:n-1}) \pi_{n-1}(dx_{1:n-1}) \\ &= \int f(x_{1:n}) \Psi_n(dx_{1:n}) - \int \mathcal{Q}_n f(x_{1:n-1}) \tilde{\Psi}_{n-1}(dx_{1:n-1}) \\ &\quad + \int \mathcal{Q}_n f(x_{1:n-1}) \left(\tilde{\Psi}_{n-1} - \Psi_{n-1}\right)(dx_{1:n-1}) \\ &\quad + \int \mathcal{Q}_n f(x_{1:n-1}) \left(\Psi_{n-1} - \pi_{n-1}\right)(dx_{1:n-1}).\end{aligned}$$

Labeling the terms on each of the three lines in this decomposition I_1, I_2, and I_3 respectively, note that the independence of the random variables generated at each step of the algorithm imply that

$$\mathbf{E}\left[I_1 I_2 \,|\, \Psi_{n-1}, \tilde{\Psi}_{n-1}\right] = 0, \quad \mathbf{E}\left[I_1 I_3 \,|\, \Psi_{n-1}, \tilde{\Psi}_{n-1}\right] = 0,$$

and

$$\mathbf{E}\left[I_2 I_3 \,|\, \Psi_{n-1}\right] = 0.$$

Therefore

$$\mathbf{E}\left[\left(\int f(x_{1:n}) \left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right] = \mathbf{E}\left[I_1^2\right] + \mathbf{E}\left[I_2^2\right] + \mathbf{E}\left[I_3^2\right].$$

The first term in this sum can be re-expressed as

$$\mathbf{E}\left[\left(\sum_{\ell=1}^{N} W_n^{(\ell)} f(X_{1:n}^{(\ell)}) - \frac{1}{N}\sum_{\ell=1}^{N} \mathcal{Q}_n f(Y_{1:n-1}^{(\ell)})\right)^2 \,\middle|\, \tilde{\Psi}_{n-1}\right]$$
$$= \frac{1}{N^2}\sum_{\ell=1}^{N} \mathbf{E}\left[\left(N W_n^{(\ell)} f(Y_{1:n-1}^{(\ell)}, X_n^{(\ell)}) - \mathcal{Q}_n f(Y_{1:n-1}^{(\ell)})\right)^2 \,\middle|\, \tilde{\Psi}_{n-1}\right]$$

which, when f is bounded at least and when the weights have finite variance, we can expect to be of size $\|f\|_\infty^2/N$. And we have already seen in (3.4) that if $\mathcal{Q}_n f$ is bounded (which it will be when f is bounded), and if the $N_n^{(\ell)}$ are sampled from a multinomial distribution, then $\mathbf{E}\left[I_2^2 \,|\, \Psi_{n-1}\right]$ is of size $K^2\|f\|_\infty^2/N$.

At this point (under a few assumptions) we have shown that the error at step n is only slightly ($\mathcal{O}(1/N)$) larger than the step $n-1$ error in estimating the average of $\mathcal{Q}_n f$ against π_{n-1}, i.e. we have shown that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathbf{E}\left[\left(\int \mathcal{Q}_n f(x_{1:n-1})\left(\Psi_{n-1} - \pi_{n-1}\right)(dx_{1:n-1})\right)^2\right]$$
$$+ \mathcal{O}\left(\|f\|_\infty^2 \frac{1+K^2}{N}\right).$$

Repeating the same steps $n-2$ more times we find that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathbf{E}\left[\left(\int \mathcal{Q}_n \cdots \mathcal{Q}_2 f(x_1)\left(\Psi_1 - \pi_1\right)(dx_1)\right)^2\right]$$
$$+ \mathcal{O}\left(\|f\|_\infty^2 \frac{1+K^2+\cdots+K^{2(n-2)}}{N}\right).$$

Since the samples $X_1^{(\ell)}$ were drawn independently from π_1 we know that

$$\mathbf{E}\left[\left(\int \mathcal{Q}_n \cdots \mathcal{Q}_2 f(x_1)\left(\Psi_1 - \pi_1\right)(x_1)dx_1\right)^2\right] = \frac{\mathbf{var}\left(\mathcal{Q}_n \cdots \mathcal{Q}_2 f\left(X_1^{(\ell)}\right)\right)}{N}$$
$$\leq \|f\|_\infty^2 \frac{K^{2(n-1)}}{N}$$

so that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathcal{O}\left(\|f\|_\infty^2 \frac{1 + K^2 + \cdots + K^{2(n-1)}}{N}\right).$$

In other words, we have shown that the sequential importance sampling with resampling scheme does converge to the correct answer as N increases. On the other hand, our estimates have been crude and do not reveal any advantage for sequential importance sampling with resampling over direct importance sampling. The growth of our bound with n is one symptom of our loose estimates. With more work, and a few more assumptions, we could show that the error in sequential importance sampling with resampling can often be bounded independently of n, something that would not typically be possible for direct importance sampling. Among other requirements, some form of contraction from the operators $\mathcal{Q}_n$ will be important in obtaining global-in-time error bounds.

Before ending this section we briefly consider alternatives to the multinomial distribution for sampling the $N_n^{(\ell)}$ in the sequential importance sampling with resampling procedure. Ultimately our goal in choosing a resampling scheme is to make the expectation

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \,\middle|\, \Psi_n\right]$$

as small as possible. However, we have seen in our bound on the total error that we must control terms of this form for functions f that we may not know in closed form (e.g. $\mathcal{Q}_n f$). It is reasonable then to instead attempt to minimize the conditional variances

$$\mathbf{var}\left[N_n^{(k)} \mid \Psi_n\right].$$

Given the requirement that the $N_n^{(k)}$ are integers and that

$$\mathbf{E}\left[N_n^{(k)} \mid \Psi_n\right] = NW_n^{(k)},$$

the conditional variance above is easily seen to be minimized when

$$N_n^{(k)} = \begin{cases} \left\lfloor NW_n^{(k)} \right\rfloor & \text{w. p. } \left\lceil NW_n^{(k)} \right\rceil - NW_n^{(k)} \\ \left\lceil NW_n^{(k)} \right\rceil & \text{w. p. } NW_n^{(k)} - \left\lfloor NW_n^{(k)} \right\rfloor. \end{cases} \tag{3.5}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. With condition (3.5) enforced, the conditional variances become

$$\mathbf{var}\left[N_n^{(k)} \mid \Psi_n\right] = \left(\left\lceil NW_n^{(k)} \right\rceil - NW_n^{(k)}\right)\left(NW_n^{(k)} - \left\lfloor NW_n^{(k)} \right\rfloor\right) \tag{3.6}$$

Exercise 23. *Verify that the conditional variance is minimized when* (3.5) *is satisfied and that the minimum value is given by* (3.6)*.*

The $Multinomial(N, \{W_n^{(\ell)}\})$ distribution does not satisfy (3.5) and we have seen that

$$\mathbf{var}\left[N_n^{(k)} \mid \Psi_n\right] = NW_n^{(k)}\left(1 - W_n^{(k)}\right)$$

which is much larger than the minimal value in (3.6). We will now consider a few rules that do satisfy (3.5). We begin by noticing that (3.5) constrains only the marginal distribution of the $N_n^{(k)}$ and does not affect their joint distribution. The simplest possible choice is to make the $N_n^{(k)}$ independent:

$$N_n^{(k)} = \left\lfloor NW_n^{(k)} \right\rfloor + \mathbf{1}_{\left\{U_n^{(k)} < NW_n^{(k)} - \lfloor NW_n^{(k)} \rfloor\right\}} \tag{3.7}$$

where the $U_n^{(k)}$ are independent random variables drawn from $\mathcal{U}(0,1)$. This scheme is sometimes referred to as Bernoulli resampling.

Exercise 24. *Check that for the $N_n^{(k)}$ generated according to* (3.7)*,* (3.5) *is satisfied.*

While (3.7) minimizes the conditional variances of the $N_n^{(k)}$, the total number of resampled points, $N_n = \sum_{\ell=1}^{N_{n-1}} N_{n-1}^{(\ell)}$, is not exactly equal to N (though its expectation is equal to N).

Exercise 25. *Follow the steps used to derive expression* (3.4) *to derive a bound for*

$$\mathbf{E}\left[\left(\int f(x_{1:n}\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \mid \Psi_n\right]$$

when the $N_n^{(k)}$ *are generated according to* (3.7).

Finally, a rule for generating the $N_n^{(k)}$ that fixes $N_n = N$ and which requires that we generate only one random variable to generate all of the $N_n^{(k)}$ at iteration n, proceeds as follows. First, generate a single independent random variate U_n from $\mathcal{U}(0,1)$. Then, for $k = 1, 2, \ldots, N_n$, set

$$N_n^{(k)} = \left|\left\{j \leq N : \sum_{\ell=1}^{k-1} W_n^{(\ell)} \leq U_n^{(j)} < \sum_{\ell=1}^{k} W_n^{(\ell)}\right\}\right| \tag{3.8}$$

where, for $j = 1, 2, \ldots, N$,

$$U_n^{(j)} = \frac{1}{N}(j - U_n)$$

and the notation $|A|$ for a discrete set of points A refers to the number of points in A.

Exercise 26. *Show that for* $N_n^{(\ell)}$ *defined by* (3.8), $\sum_{\ell=1}^{N} N_n^{(\ell)} = N$ *and* (3.5) *is satisfied.*

The rule (3.8) is often referred to as systematic resampling and is observed to perform very well in practice. Despite its success in applications, it is unfortunately not possible to show that it converges in general.

Exercise 27. *Find a sequence of weights* $\{w^{(\ell)}\}_{\ell=1}^N$ *with* $\sum_{\ell=1}^N w^{(\ell)} = 1$, *and points* $\{x^{(\ell)}\}_{\ell=1}^N$ *so that*

$$\mathbf{E}\left[\left(\frac{1}{N}\sum_{\ell=1}^{N} N^{(\ell)} f(x^{(\ell)}) - N w^{(\ell)} f(x^{(\ell)})\right)^2\right]$$

does not converge when the $N^{(\ell)}$ *are generated according to* (3.8) *with* $w^{(\ell)}$ *in place of* $W_n^{(\ell)}$. *Hint: try an even length alternating sequence of two values,* x_0 *and* x_1, *and assume that if* $x^{(k)} = x^{(\ell)}$ *then* $w^{(k)} = w^{(\ell)}$ *(as would occur if the* $w^{(k)}$ *were importance weights).*

Exercise 28. *Write a routine to use $\mathcal{N}(0,1)$ random variables to generate approximate $\mathcal{N}(0,\sigma^2)$ random variables via an application of each of the three resampling methods (multinomial, Bernoulli, and systematic) discussed in this section. Numerically estimate the variance of the $N^{(k)}$ from each method. What do you observe? Are your observations robust to changing σ^2? Note that this test corresponds to a single resampling step: first sample from $\mathcal{N}(0,1)$, then weight the samples by the appropriate normalized importance weights, then resample.*

Exercise 29. *Write a routine that uses sequential importance sampling with and without resampling to compute averages with respect to the uniform measure on SAW(d) for a sequence of increasing d. Use the same reference density described in Example 11 in both schemes. Produce plots of a single sample path for each value of d. In validating your algorithms you may find it useful to check, e.g. the expected number of times a lattice site is visited (this should be the same for every site). Can you think of other statistics that might help you validate/debug your codes? How can you compare the two methods? Can you think of a way to use these simulations to estimate the normalization constants $\mathcal{Z}_d$? Estimate how quickly $\mathcal{Z}_d$ grows with d.*

3.5 bibliography

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*

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

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 $A_i \subset \mathcal{B}$,

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

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

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

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 $\ell < k$,

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

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 $q(y \,|\, x)$ 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 $q(y \,|\, x)$ 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

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

In this setting, letting

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

the average rejection probability is

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

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

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

is the relative entropy of the density $q_1(x_i \,|\, y_i)\pi(y_i)$ with respect to $q_1(y_i \,|\, x_i)\pi(x_i)$ 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 2d-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 2d-Ising model. Does it help?*

4.8 bibliography

Chapter 5

Stochastic Thermostats

The Metropolis accept/reject step allowed for the use of very naive proposal densities. Errors in the proposal step are corrected at the acceptance step at the cost of a potentially very high rejection probability. In this section we describe a family of methods that use continuous time Markov processes that can be used to sample π without detailed knowledge of its structure, and without an accept/reject step. In practice, the continuous time Markov processes must be discretized which introduces error. This error can be eliminated by the addition of a Metropolis accept/reject step or it can be reduced by decreasing the discretization parameter. These methods (typically without any accept/reject step) are the basis of most MCMC simulations in very high dimensions.

5.1 Overdamped Langevin schemes

In this section we will derive the first of our new, accept/reject free schemes starting from a very simple "isotropic" Metropolis scheme. For any $h > 0$, let X_h be a Markov chain in $\mathbb{R}$ generated according to the Metropolis-Hastings rule described in the previous chapter with proposal density $q(y \mid x) = \mathcal{N}(x, 2h)$.

We will begin by computing the generator corresponding to X_h. It will be

convenient to scale the generator by a factor of h^{-1} so that

$$\mathcal{L}_h f(x) = \frac{\mathbf{E}_x\left[f(X_h^{(1)})\right] - f(x)}{h}.$$

This rescaling corresponds to associating each discrete step of the chain with a size h increment of a continuous time variable.

Taylor Expanding $f(X_h^{(1)})$ about x we find that

$$\mathcal{L}_h f(x) = \frac{f'(x)\mathbf{E}_x\left[\triangle_0^1 X_h\right] + \frac{1}{2}f''(x)\mathbf{E}_x\left[\left(\triangle_0^1 X_h\right)^2\right] + \mathcal{O}\left(\mathbf{E}_x\left[|\triangle_0^1 X_h|^3\right]\right)}{h}$$

Now notice that, by a change of variables,

$$\begin{aligned} E_x\left[\triangle_0^1 X_h\right] &= \int (y-x)p(y\,|\,x)dy = \int (y-x)q(y\,|\,x)p_{acc}(x,y)dy \\ &= \sqrt{2h}\int z\, p_{acc}(x, x+\sqrt{2h}\,z)\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz. \end{aligned}$$

Examining p_{acc} in more detail we find that

$$p_{acc}(x, x+\sqrt{2h}\,z) = 1 + \min\left\{0, \frac{\pi'(x)}{\pi(x)}\sqrt{2h}\,z + \mathcal{O}(h)\right\}.$$

When h is small, the sign of the expression in the minimum is determined by the first order (in $\sqrt{h}$) term. Since

$$\int z\,\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz = 0$$

we obtain

$$E_x\left[\triangle_0^1 X_h\right] = 2h\frac{\pi'(x)}{\pi(x)}\int_{\pi'(x)z<0} z^2\,\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz + \mathcal{O}\left(h^{3/2}\right) = h(\log\pi(x))' + \mathcal{O}\left(h^{3/2}\right).$$

Similarly

$$\begin{aligned} E_x\left[\left(\triangle_0^1 X_h\right)^2\right] &= 2h\int z^2 p_{acc}(x, x+\sqrt{2h}z)\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz \\ &= 2h + \mathcal{O}\left(h^{3/2}\right) \end{aligned}$$

and

$$E_x\left[|\triangle_0^1 X_h|^3\right] = \mathcal{O}\left(h^{3/2}\right)$$

so that

$$\mathcal{L}_h f(x) = \mathcal{L}_O f(x) + \mathcal{O}\left(\sqrt{h}\right)$$

where we have introduced the second order differential operator

$$\mathcal{L}_O f = f'(x)\left(\log \pi(x)\right)' + f''(x) = \frac{1}{\pi(x)}\left(\pi(x) f'(x)\right)' \tag{5.1}$$

Were we to repeat this derivation in higher dimensions for the Metropolis-Hastings scheme with proposal density $q(y\,|\,x) = \mathcal{N}\left(x, 2hS(x)\right)$ for some symmetric $d \times d$ positive definite matrix $S(x)$ we would again find that $\mathcal{L}_h = \mathcal{L}_O + \mathcal{O}\left(\sqrt{h}\right)$, where now

$$\begin{aligned}\mathcal{L}_O f &= \nabla f(x)\frac{1}{\pi(x)}\mathrm{div}\left(\pi(x)S(x)\right) + \mathrm{trace}\left(D^2 f(x)\, S(x)\right) \\ &= \frac{1}{\pi(x)}\mathrm{div}\left(\pi(x)\nabla f(x) S(x)\right). \end{aligned} \tag{5.2}$$

In these formulae, the matrix $D^2 f$ is the $d \times d$ matrix of second derivatives of f and for any n, the divergence of an $n \times d$ matrix valued function $M(x)$ is the $n \times 1$ vector with entries

$$\left(\mathrm{div} M(x)\right)_i = \sum_{j=1}^{d} \partial_j M_{ij}(x).$$

Exercise 45. *Verify the expression* (5.2) *for the limiting generator of the Metropolis-Hastings scheme with* $q(y\,|\,x) = \mathcal{N}\left(x, 2hS(x)\right)$. *Note that this transition density is not symmetric which leads to the additional first order term* $\nabla f(x)\, div S(x)$.

Let's examine this operator $\mathcal{L}_O$ in (5.2) in slightly more detail. First, recall that, when μ is a density, the action of $\mathcal{L}_O$ on μ is $\mu\mathcal{L}_O = \mathcal{L}_O^*\mu$ where $\mathcal{L}_O^*$ is the adjoint of $\mathcal{L}_O$ in the inner product $\langle f, g\rangle = \int f(x)g(x)dx$. For this particular operator the adjoint is computed by an integration by parts. We

find that

$$\begin{aligned}\int f(x)\mu\mathcal{L}_O(x)dx &= \int \mathcal{L}_O f(x)\mu(x)dx \\ &= \int \operatorname{div}\left(\pi(x)\nabla f(x)S(x)\right)\frac{\mu(x)}{\pi(x)}dx \\ &= -\int \pi(x)\nabla f(x)S(x)\nabla^{\mathrm{T}}\left(\frac{\mu(x)}{\pi(x)}\right)dx \\ &= \int f(x)\operatorname{div}\left(\pi(x)\nabla\left(\frac{\mu(x)}{\pi(x)}\right)S(x)\right)dx\end{aligned}$$

for all f, so that

$$\mu\mathcal{L}_O(x) = \operatorname{div}\left(\pi(x)\nabla\left(\frac{\mu(x)}{\pi(x)}\right)S(x)\right). \tag{5.3}$$

Plugging in π in place of μ we see that $\pi\mathcal{L}_O = 0$. In fact, $\mathcal{L}_O$ is reversible with respect to π. In particular,

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

for all f and g.

Exercise 46. *Verify the formula in the last display.*

We already knew that the Metropolis scheme preserves π. But having used that scheme to derive the operator $\mathcal{L}_O$, and having observed that $\pi\mathcal{L}_O = 0$, we can consider alternative schemes that may not exactly preserve π, but whose generator (after rescaling by h^{-1}), $\mathcal{L}_h$, also approximates $\mathcal{L}_O$. One might expect that the invariant measure (should it exist) for one of these alternative schemes, π_h, while not exactly equal to π, is close to π for small h. Indeed, if we assume that π_h is a probability measure satisfying $\pi_h\mathcal{L}_h = 0$, and if for some test function f, u is the solution of the PDE

$$\mathcal{L}_O u = f - \pi[f], \tag{5.4}$$

then

$$\pi_h[f] - \pi[f] = \pi_h\left(\mathcal{L}_O - \mathcal{L}_h\right)u \tag{5.5}$$

The last term will be small (in h), for example, if u is smooth and bounded with bounded derivatives. For the Metropolis scheme the final bound is $\mathcal{O}(\sqrt{h})$, which we know is too pessimistic (in that case, $\pi_h = \pi$), but for other chains that do not exactly preserve π, this expression can tell us that Monte Carlo estimates produced using the chain X_h will have a bias that is small. Later in this section we use a similar argument to bound the error in the Markov chain Monte Carlo estimator of $\pi[f]$ using a chain introduced below for which $\mathcal{L}_h f = \mathcal{L}_O f + \mathcal{O}(h)$.

To find other candidates for Markov chains that might have invariant probability measures approximating π, we return to the key features of the Metropolis scheme responsible for the convergence of $\mathcal{L}_h$ to $\mathcal{L}_O$. Can we identify those same properties in other chains? As we will see in a moment, the answer to this question is yes. The most compelling argument for looking for alternatives to the Metropolis–Hastings framework is the slow convergence of those schemes in high dimensions, an unfortunate characteristic that we have already explored. The so called Langevin schemes that we will derive in this Chapter can avoid the accept reject step (at the cost of a small in h bias) and are generally much more effective in very high dimensional settings than Metropolis–Hastings schemes.

Looking back at the convergence argument, the key properties that lead to the derivation of the limiting generator (5.1) are that

$$E_x\left[\triangle_0^1 X_h\right] = h\left(\log \pi(x)\right)' + o(h) \quad \text{and} \quad E_x\left[\left(\triangle_0^1 X_h\right)^2\right] = 2h + o(h)$$

or, in higher dimensions,

$$E_x\left[\triangle_0^1 X_h\right] = h\frac{1}{\pi(x)}\text{div}\left(\pi(x)S(x)\right) + o(h)$$

and

$$E_x\left[\left(\triangle_0^1 X_h\right)\left(\triangle_0^1 X_h\right)^{\text{T}}\right] = 2hS(x) + o(h).$$

More generally, our manipulations suggest that a discrete time Markov chain X_h satisfying

$$E_x\left[\triangle_0^1 X_h\right] = b(x)h + o(h) \quad \text{and} \quad E_x\left[\left(\triangle_0^1 X_h\right)^2\right] = \sigma(x)\sigma^{\text{T}}(x)h + o(h) \tag{5.6}$$

should have limiting generator

$$\mathcal{L}f(x) = \nabla f(x)\, b(x) + \frac{1}{2}\text{trace}\left(\sigma(x)\sigma^{\text{T}}(x)D^2 f(x)\right). \tag{5.7}$$

With some additional restrictions, this is in fact the case.

The simple recursion

$$X_h^{(k+1)} = X_h^{(k)} + h\, b(X_h^{(k)}) + \sqrt{h}\, \sigma(X_h^{(k)})\, \xi^{(k+1)} \tag{5.8}$$

where each $\xi^{(k)}$ is, for example, a vector of independent random variables with $\mathbf{P}\left[(\xi^{(k)})_i = 1\right] = \mathbf{P}\left[(\xi^{(k)})_i = -1\right] = 1/2$ or a vector of independent standard Gaussian random variables, has limiting generator in (5.7).

Exercise 47. *Show that for the process in* (5.8) *and any smooth function f with bounded derivatives,*

$$\mathcal{L}_h f - \mathcal{L} f = \mathcal{O}(h)$$

for $\mathcal{L}$ in (5.7)*.*

When we plug the choices $b = \operatorname{div}(\pi S)/\pi$ and $\sigma\sigma^{\mathrm{T}} = 2S$ obtained above into (5.8) we obtain a discretization of the overdamped Langevin sampler with updates

$$X_h^{(k+1)} = X_h^{(k)} + h\, S(X_h^{(k)}) \nabla^{\mathrm{T}} \log \pi(X_h^{(k)}) + h \operatorname{div} S(X_h^{(k)}) + \sqrt{2hS(X_h^{(k)})}\, \xi^{(k+1)} \tag{5.9}$$

Note that to implement (5.9), one needs to compute the matrix square root of S e.g. by Choleski factorization. The chain in (5.9) will not preserve π exactly and for h too large will not have an invariant probability measure at all (it will be transient). However, under restrictions on π and for h small the chain will have an invariant probability measure π_h and, combining (5.5) and the result Exercise 47, we expect that $\pi_h[f] = \pi[f] + \mathcal{O}(h)$.

Exercise 48. *Compute the invariant measure π_h of the process* (5.9) *in the case when $S = I$ and the target density is $\pi = \mathcal{N}(\mu, M)$. Confirm that $\pi_h[f] = \pi[f] + \mathcal{O}(h)$.*

The operator $\mathcal{L}_O$ in (5.2) is an example of a Kolmogorov operator and is the generator of a continuous time Markov process that is the limit of our Metropolis scheme, thus justifying our use of the term limiting generator. That limiting Markov process is called a diffusion process. Diffusion processes

always have generators of the form (5.7) where the vector valued function b and the matrix valued function σ can be more general than the functions that we have derived above by considering the limit of the Metropolis scheme.

At the moment we have no particular interest in diffusion processes. On the other hand, in this Chapter we are precisely interested in MCMC schemes with diffusion limits. And because, as we have already seen, there are many discrete processes with the same diffusion limit (i.e. same limiting functions b and σ), it will sometimes be useful to characterize our sampling schemes by their limiting drift and diffusion coefficients or, equivalently, by their limiting generators. Diffusion processes are used to model many physical processes and will receive more attention in Part II of these notes.

The approach represented by (5.9) and its many generalizations is the most common method by which very high dimensional averages and integrals are computed. To get a basic feeling for the qualitative properties of (5.9) (and for those of the Metropolis scheme that we have just learned is intimately related), assume that S is constant and replace (5.9) by the deterministic iteration

$$x_h^{(k+1)} = x_h^{(k)} + hS\nabla^{\mathrm{T}} \log \pi(x_h^{(k)}). \tag{5.10}$$

The generator of this process (after rescaling by h^{-1}) converges to $\mathcal{L}$ defined by $\mathcal{L}f = \nabla f S \nabla^{\mathrm{T}} \log \pi$. The operator $\mathcal{L}$ is just the Liouvillian corresponding to the ODE $\frac{d}{dt}y^{(t)} = S\nabla^{\mathrm{T}} \log \pi(y^{(t)})$, which implies that $x_h^{(k)}$ converges to $y^{(kh)}$ as h decreases (with kh held constant). That ODE is a gradient ascent for the function $\log \pi$. In other words

$$\frac{d}{dt} \log \pi \left(y^{(t)}\right) = \nabla(\log \pi(y^{(t)}))S\nabla^{\mathrm{T}}(\log \pi(y^{(t)})).$$

The expression on the right hand side is always positive because S is a symmetric positive definite matrix. In fact, if $\lambda_1 > 0$ is the smallest eigenvalue of S then

$$\frac{d}{dt} \log \pi \left(y^{(t)}\right) \geq \lambda_1 \|\nabla(\log \pi(y^{(t)}))\|_2^2$$

so that as y evolves it can only increase the value of π. Similarly, the Markov chain X generated by (5.8) tends to move toward higher π-probability regions. The added noise in (5.9) prevents $X_h^{(k)}$ from converging to a local maximum of π. Because of its tendency to reduce the "energy" $-\log \pi$, any term of the form $S\nabla^{\mathrm{T}} \log \pi$ for a symmetric positive semi-definite matrix S

is referred to a dissipative or damping term. In fact, the subscript O in $\mathcal{L}_O$ references the fact that the operator $\mathcal{L}_O$ can be derived as a highly damped (or "overdamped") limit of a more general family of operators that we will introduce later in this chapter. We will refer to any scheme with a limiting generator of the forms considered in this section as an overdamped scheme.

In Chapter **??** we will consider how the choice of S can be used to speed convergence of the MCMC estimator corresponding to (5.9) with $b = \operatorname{div}(\pi S)/\pi$ and $\sigma\sigma^{\mathrm{T}} = 2S$. Before moving on to introducing further improvements to this estimator we return to the question of how large an error we should expect when we estimate $\pi[f]$ by

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

recalling that $\mathcal{L}_h$ only approximately preserves π.

In fact, if we again make the (strong) assumptions that the PDE (5.4) has a smooth solution u, which, along with S and $\log\pi$, is bounded with bounded derivatives we can bound the error of $\overline{f}_N$ directly. To see this note that

$$\frac{u(X_h^{(k+1)}) - u(X_h^{(1)})}{h\,k} = \frac{1}{k}\sum_{\ell=1}^{k} \mathcal{L}_h u(X_h^{(\ell)}) + \frac{1}{h\,k} M_h^{(k)}$$

where $M_h^{(k)}$ is the martingale

$$M_h^{(k)} = \sum_{\ell=1}^{k} u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right].$$

Using the PDE solved by u, we can write

$$\frac{u(X_h^{(k+1)}) - u(X_h^{(1)})}{h\,k} = \frac{1}{k}\sum_{\ell=1}^{k} f(X_h^{(\ell)}) - \pi[f] + \frac{1}{k} R_h^{(k)} + \frac{1}{h\,k} M_h^{(k)} \qquad (5.11)$$

where

$$R_h^{(k)} = \sum_{\ell=1}^{k} (\mathcal{L}_h - \mathcal{L})\, u(X_h^{(\ell)}).$$

Note that, if u is bounded with bounded derivatives and if $X_h^{(k)}$ is generated by (5.9), then $R_h^{(k)}$ is of size $\mathcal{O}(kh)$ so that

$$\overline{f}_k - \pi[f] = \frac{1}{h\,k} M_h^{(k)} + \mathcal{O}((hk)^{-1}) + \mathcal{O}(h).$$

Now consider the expected square of the martingale $M_h^{(k)}$,

$$\mathbf{E}\left[(M_h^{(k)})^2\right] = \sum_{\ell=1}^{k} \mathbf{E}\left[\left(u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right]\right)^2\right]. \tag{5.12}$$

The cross terms on the right hand side have vanished because, if $\mathcal{F}_\ell$ is the sigma algebra generated by $X_h^{(0)}, X_h^{(1)}, \ldots, X_h^{(\ell)}$, then, for $r < \ell$,

$$\begin{aligned}
&\mathbf{E}\left[\left(u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right]\right)\left(u(X_h^{(r+1)}) - \mathbf{E}_{X_h^{(r)}}\left[u(X_h^{(r+1)})\right]\right)\right] \\
&\quad = \mathbf{E}\left[\mathbf{E}\left[u(X_h^{(\ell+1)}) - \mathbf{E}_{X^{(\ell)}}\left[u(X_h^{(\ell+1)})\right] \mid \mathcal{F}_\ell\right]\left(u(X_h^{(r+1)}) - \mathbf{E}_{X_h^{(r)}}\left[u(X_h^{(r+1)})\right]\right)\right] \\
&\quad = 0.
\end{aligned}$$

Returning to (5.12), taylor expansion of u reveals that (again assuming $X_h^{(k)}$ is generated by (5.9)) the expected square of $M_h^{(k)}$ is of size $\mathcal{O}(kh)$ so that

$$\mathbf{E}\left[(\overline{f}_k - \pi[f])^2\right] = \mathcal{O}\left(\frac{1}{h\,k}\right) + \mathcal{O}(h^2).$$

Note that this is exactly the qualitative behavior you would expect. As h is decreased, larger k is required to maintain a fixed accuracy because the chain is perturbed less at each step. On the other hand, when k increases with h fixed, an error of size $\mathcal{O}(h)$ persists.

5.2 Hamilton's ODEs

The basis of the family of MCMC schemes built from continuous time dynamics that we will describe is a simple, but very important, set of ordinary differential equations. For $x \in \mathbb{R}^d$, any sufficiently smooth real valued function $H(x)$ and skew-symmetric, smooth, $d \times d$ matrix valued function $J(x)$ (i.e. $J^{\mathrm{T}} = -J$), the ODE

$$\frac{d}{dt} y^{(t)} = -J(y^{(t)}) \nabla^{\mathrm{T}} H(y^{(t)}) + \mathrm{div} J(y^{(t)}) \tag{5.13}$$

is referred to as a Hamiltonian system of ODE. The function $H(x)$ is called the Hamiltonian. We will assume that

$$\liminf_{x\to\infty} \frac{H(x)}{\|x\|} > 0.$$

This condition guarantees that the function $\exp(-H(x))$ is integrable for any $\beta > 0$. In the context of statistical mechanics, the density

$$\pi_H(x) \propto \exp(-H(x)) \tag{5.14}$$

is called a Boltzmann density. In terms of the Boltzmann density we can rewrite the Hamiltonian system of ODE (5.13) as

$$\frac{d}{dt}y^{(t)} = \frac{1}{\pi_H(y^{(t)})}\mathrm{div}\left(\pi_H(y^{(t)})J(y^{(t)})\right) \tag{5.15}$$

Example 23. *In the special case that $x = (\hat{x}, \tilde{x})$ and*

$$H(x) = \frac{1}{2}\tilde{x}^T M^{-1}\tilde{x} + U(\hat{x})$$

where $\tilde{x}$ and $\hat{x}$ are the positions and velocities of a system of particles with mass matrix M and experiencing the potential field U and

$$J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},$$

the Hamiltonian system can be rewritten

$$M\frac{d^2}{dt^2}\hat{y}^{(t)} = -\nabla^T U(\hat{y}^{(t)}),$$

which is just Newton's equations of motion i.e. force equals mass times acceleration.

The system (5.13) has several important characteristics. To derive these characteristics, recall that the generator corresponding to the ODE (5.13) is

$$\mathcal{L}_H f = \nabla f \frac{1}{\pi_H}\mathrm{div}\,(\pi_H J)$$

The action of $\mathcal{L}_H$ on a density μ is found (by an integration by parts) to be

$$\mu\mathcal{L}_H = -\text{div}\left(\frac{\mu}{\pi_H}\,\text{div}^{\text{T}}(\pi_H J)\right)$$

Expanding this expression we find that

$$\mu\mathcal{L}_H = -\nabla\left(\frac{\mu}{\pi_H}\right)\text{div}\,(\pi_H J) - \frac{\mu}{\pi_H}\text{div}\,(\text{div}^{\text{T}}\,(\pi_H J))$$

The second term of the last display vanishes because J is antisymmetric, leaving

$$\mu\mathcal{L}_H = -\nabla\left(\frac{\mu}{\pi_H}\right)\text{div}\,(\pi_H J) \tag{5.16}$$

Exercise 49. *Check this.*

Plugging in $\mu = \pi_H$ we find that

$$\pi_H\mathcal{L}_H = 0.$$

In fact, with respect to π_H, the operator $\mathcal{L}_H$ satisfies the even stronger property

$$\int g(x)\mathcal{L}_H f(x)\pi_H(dx) = -\int f(x)\mathcal{L}_H g(x)\pi_H(dx) \tag{5.17}$$

for any test functions f and g. Condition (5.17) closely resembles the reversibility condition introduced in the Chapter (**??**) and will be referred to here as skew-reversibility with respect to π_H. Like reversibility, skew-reversibility with respect to π_H implies that $\pi_H\mathcal{L}_H = 0$ (though we had already verified this).

Exercise 50. *Verify equation* (5.17).

Expression (5.17) in turn implies that

$$\int f(x)g(y^{(t)}(x))\pi_H(dx) = \int g(x)f(y^{(-t)}(x))\pi_H(dx). \tag{5.18}$$

Exercise 51. *Establish the expression in the last display. Hint: fix* $s \in [0,t]$ *and let* $w(s) = \int f(y^{(s-t)}(x))g(y^{(s)}(x))\pi_H(dx)$ *and then show that the derivative of* w *is zero.*

We will use these facts to derive a number of sampling schemes based loosely on the ODE in (5.13).

Some final useful properties of the solution to (5.13) can be derived if we make additional assumptions on J and H. We will assume that the variable x can be decomposed into two vectors $x = (\hat{x}, \tilde{x})$ with $\hat{x} \in \mathbb{R}^{\hat{d}}$ and $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ and that first, $H(\hat{x}, \tilde{x})$ is an even function of $\tilde{x}$, i.e that

$$H(\hat{x}, \tilde{x}) = H(\hat{x}, -\tilde{x}), \tag{5.19}$$

and second that J has the particular form

$$J(x) = \begin{bmatrix} 0 & -\hat{J}(\hat{x}) \\ \hat{J}^{\mathrm{T}}(\hat{x}) & 0 \end{bmatrix} \tag{5.20}$$

where $\hat{J}$ is a $\hat{d} \times \tilde{d}$ matrix valued function of only the $\hat{x}$ variables. Under these assumptions (5.13) becomes,

$$\frac{d}{dt} \begin{pmatrix} \hat{y}^{(t)} \\ \tilde{y}^{(t)} \end{pmatrix} = \begin{pmatrix} \hat{J}(\hat{y}^{(t)}) \nabla_{\tilde{x}}^{\mathrm{T}} H(y^{(t)}) \\ -\hat{J}^{\mathrm{T}}(\hat{y}^{(t)}) \nabla_{\hat{x}}^{\mathrm{T}} H(y^{(t)}) + \mathrm{div} \hat{J}^{\mathrm{T}}(\hat{y}^{(t)}) \end{pmatrix}.$$

Exercise 52. *Check that for J of the form in* (5.20)*,*

$$\mathrm{div} J = \begin{pmatrix} 0 \\ \mathrm{div} \hat{J}^{T} \end{pmatrix}.$$

Under assumptions (5.19) and (5.20) the action of the operator $\mathcal{L}_H$ on functions becomes

$$\mathcal{L}_H f = \nabla_{\hat{x}} f \hat{J} \nabla_{\tilde{x}}^{\mathrm{T}} H - \nabla_{\tilde{x}} f \hat{J}^{\mathrm{T}} \nabla_{\hat{x}}^{\mathrm{T}} H + \nabla_{\tilde{x}} f \, \mathrm{div} \hat{J}^{\mathrm{T}}$$

and its action on probability densities becomes

$$\mu \mathcal{L}_H = -\nabla_{\hat{x}} \mu \, \hat{J} \nabla_{\tilde{x}}^{\mathrm{T}} H + \nabla_{\tilde{x}} \mu \, \hat{J}^{\mathrm{T}} \nabla_{\hat{x}}^{\mathrm{T}} H - (\mu \, \nabla_{\tilde{x}} H + \nabla_{\tilde{x}} \mu) \, \mathrm{div} \hat{J}^{\mathrm{T}}$$

Notice that if for some test function f we set $f_-(x) = f(\hat{x}, -\tilde{x})$ then

$$\mathcal{L}_H f_-(\hat{x}, -\tilde{x}) = -\mathcal{L}_H f(x). \tag{5.21}$$

A similar formula holds if we apply $\mathcal{L}_H$ to $\mu_-(x) = \mu(\hat{x}, -\tilde{x})$. Relation (5.21) has several remarkable and useful ramifications. For example, it implies that

if f is an even (resp. odd) function of $\tilde{x}$ then $\mathcal{L}_H f$ is an odd (resp. even) function of $\tilde{x}$. In particular, if f and g are both even functions of $\tilde{x}$ then

$$\int g(x)\mathcal{L}_H f(x)dx = 0 \tag{5.22}$$

Exercise 53. *Establish expression* (5.22).

Expression (5.21) also implies that the functions

$$y^{(-t)}(x) \qquad \text{and} \qquad \big(\hat{y}^{(t)}(\hat{x}, -\tilde{x}), -\tilde{y}^{(t)}(\hat{x}, -\tilde{x})\big)$$

both solve (5.13) with the sign of the right hand side reversed and with initial condition x. By the uniqueness of solutions to the ODE, we find therefore that the two functions are equal, i.e.

$$y^{(-t)}(x) = \big(\hat{y}^{(t)}(\hat{x}, -\tilde{x}), -\tilde{y}^{(t)}(\hat{x}, -\tilde{x})\big) \tag{5.23}$$

Exercise 54. *Assuming uniques of solutions to* (5.13)*, establish* (5.23).

Equation (5.23) is called time reversal symmetry and tells us that the inverse of the flow map at time t can also be written as a forward-in-time integration using an $\tilde{x}$ initial condition with reversed sign. Time reversal symmetry is not to be confused with the notion of reversibility that we have introduced early and indeed, we have already seen that solutions to Hamilton's ODE are skew-reversible. But time reversal symmetry will have important implications in the next section where it will be used to show that a Markov process incorporating Hamilton's ODE are indeed reversible.

Finally, time reversal symmetry combined with expression (5.18) implies that if f and g are even functions of $\tilde{x}$ then

$$\int g(x)f(y^{(t)}(x))\pi_H(dx) = \int f(x)g(y^{(t)}(x))\pi_H(dx). \tag{5.24}$$

Exercise 55. *Establish equation* (5.24).

Before closing this section we make a few additional observations in the case when J is a constant matrix. Notice that, in this case, $\mathcal{L}_H H = 0$ so that the

value of H is exactly preserved by the solution to (5.13). Moreover, when J is constant, $\mu\mathcal{L}_H = 0$ for an density μ of the form $\mu(x) = \rho(H(x))$ for some function ρ. In words, the flow map $y^{(t)}(x)$ preserves any density of this form (including the constant density). However, it is also clear that if the value of H is preserved, the solutions to (5.13) cannot be ergodic (they cannot be irreducible). We will see in the next section that with some modification, (5.13) can be used to build effective MCMC algorithms.

5.3 Hamiltonian based MCMC schemes

In the last section we learned that solutions to the ODE (5.13) preserve the Boltzmann density. Since we have wide latitude in defining H, it is natural to ask if, given a target density $\pi(x)$, the unusual properties of solutions to (5.13) can be put to use in efficiently generating samples from π. The answer is yes, but, in general, requires the addition of some source of randomness. Indeed, as we have already mentioned, we do not expect solutions to (5.13) to be irreducible (much less ergodic).

Designing a Markov process based on Hamilton's ODE that is irreducible generally requires adding some additional "conjugate" dimensions to the system. In other words, if the target density is $\pi(\hat{x})$ where $\hat{x} \in \mathbb{R}^{\hat{d}}$, we set $x = (\hat{x}, \tilde{x})$ where $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ and $d = \hat{d} + \tilde{d}$. If we choose H of the form

$$H(\hat{x}, \tilde{x}) = -\log \pi(\hat{x}) + K(\tilde{x}) \tag{5.25}$$

then the marginal of the $\hat{x}$ variables under the density

$$\pi_H(\hat{x}, \tilde{x}) \propto e^{-H(\hat{x},\tilde{x})}$$

will be exactly π, i.e.

$$\int \pi_H(x) d\tilde{x} = \pi(\hat{x}).$$

In fact, under π_H, the $\hat{x}$ and $\tilde{x}$ variables are independent. We will take advantage of this observation by designing (higher dimensional) Markov chains to sample from π_H instead of π. Samples from π can be recovered from the $\hat{x}$ components of the resulting trajectory.

Assuming that the density proportional to $\exp(-K(\tilde{x}))$ can be sampled easily (we chose K after all), then a simple Markov chain sampling π_H can be constructed as follows: Fix an $s > 0$, draw $\tilde{X}^{(0)}$ from $\exp(-K(\tilde{x}))/\tilde{\mathcal{Z}}$, and proceed from a sample $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(k)})$ to generate $X^{(k+1)}$ by,

Algorithm 1. *Hamilton's ODE's with randomized conjugate variables*

1. `Independently sample a variable` $\tilde{Y}^{(k)}$ `from the density proportional to` $e^{-K(\tilde{x})}$.
2. `Set` $X^{(k+1)} = y^{(s)}(\hat{X}^{(k)}, \tilde{Y}^{(k)})$ `where` $y^{(s)}(x)$ `is solution to` (5.13) `at time` s `with initial condition` x.

That the Markov chain $X^{(t)}$ preserves π_H follows from the fact that $y^{(s)}$ preserves π_H.

When we enforce that J has the structure in (5.20) and that K is an even function (i.e. $K(-\tilde{x}) = K(\tilde{x})$), we can prove that the Markov chain $\hat{X}$ generated by Algorithm 1 is reversible with respect to π. To see this note that the transition operator for the Markov chain $\hat{X}^{(t)}$ generated by Algorithm 1 is given by

$$\mathcal{T}_s f(\hat{x}) = \frac{\int f(\hat{y}^{(s)}(\hat{x}, \tilde{x}))e^{-K(\tilde{x})}d\tilde{x}}{\tilde{\mathcal{Z}}}.$$

Appealing to time reversal symmetry and the fact that K is an even function we can apply expression (5.24) to find that

$$\begin{aligned}\int g(\hat{x})\mathcal{T}_s f(\hat{x})\pi(d\hat{x}) &= \int g(\hat{x})f(\hat{y}^{(s)}(x))\pi_H(dx) \\ &= \int f(\hat{x})g(\hat{y}^{(s)}(x))\pi_H(dx) \\ &= \int f(\hat{y})\mathcal{T}_s g(\hat{y})\pi(d\hat{y}),\end{aligned}$$

i.e. that the $\hat{X}^{(k)}$ process is reversible with respect to π.

There is one major issue that we have so far avoided. We need a way to approximate solutions of (5.13). Unfortunately it is not be possible to design a scheme that exactly preserves π_H as the exact solutions do. However, when

J is constant we can find simple discrete time approximations to (5.13) whose solutions are both symplectic and have time reversal symmetry. Assuming that J is constant and has the structure in (5.20), one such discretization is the Velocity Verlet scheme:

$$\begin{aligned}
\tilde{y}'_h &= \tilde{y}_h^{(\ell)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{y}_h^{(\ell)}) \\
\hat{y}_h^{(\ell+1)} &= \hat{y}_h^{(\ell)} + h\hat{J}\nabla^{\mathrm{T}}K(\tilde{y}'_h) \\
\tilde{y}_h^{(\ell+1)} &= \tilde{y}'_h + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{y}_h^{(\ell+1)})
\end{aligned} \tag{5.26}$$

where h is a small time-discretization parameter. Beginning with $y_h^{(0)} = x$, after $n = \lfloor s/h \rfloor$ iterations $y_h^{(n)}$ will approximate $y^{(s)}(x)$ up to an $\mathcal{O}(h^2)$ error. Modifications to this approximation are required when $\hat{J}$ is allowed to depend on $\hat{x}$.

Exercise 56. *Show that this is a consistent integration scheme with truncation error of order 3, i.e.*

$$\frac{y_h^{(1)}(x) - x}{h} = -J\nabla^{T}H(x) + \mathcal{O}(h^2)$$

Exercise 57. *Show that the Velocity Verlet scheme is symplectic (show that the jacobian of the map $x \to y_h^{(1)}(x)$ is 1), and time reversible in the same sense as the Hamiltonian ODE.*

With the discretization of (5.13) in (5.26), the practical alternative to (1) generates a chain $X_h^{(k+1)}$ according to the rule:

Algorithm 2. *Velocity Verlet with randomized conjugate variables*

1. `Independently sample variable` $\tilde{Y}_h^{(k)}$ `from the density proportional to` $\exp(-K(\tilde{x}))$
2. `Set` $X_h^{(k+1)} = y_h^{(n)}\left(\hat{X}_h^{(k)}, \tilde{Y}_h^{(k)}\right)$ `where` $y_h^{(n)}(x)$ `is the solution of` (5.26) `after` n `steps with initial conditions` x `and` n `chosen by the user.`

If we choose n very large in Algorithm 2, we will expend substantial effort to generate a single update of the chain $X^{(k)}$ and the scheme will become inefficient. On the other hand, consider the opposite extreme in which we choose $n = 1$ in Algorithm 2. Assume that we also make the typical choice $K(\tilde{x}) = \|\tilde{x}\|_2^2/2$. Then,

$$\begin{aligned}\hat{X}_h^{(k+1)} &= \hat{X}_h^{(k)} + h\hat{J}\left(\tilde{Y}_h^{(k)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{X}_h^{(k)})\right)\\ &= \hat{X}_h^{(k)} + \frac{h^2}{2}\hat{J}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{X}_h^{(k)}) + h\hat{J}\,\tilde{Y}_h^{(k)}.\end{aligned}$$

Setting $\delta = h^2/2$ and noting that $Y_h^{(k)}$ has mean 0 and identity covariance, we see that the above iteration is exactly of the overdamped form in (5.9). We therefore expect, that when n is small, the performance of this scheme is similar to the corresponding overdamped scheme. On the other hand, when n is very large, we expect (2) to converge slowly because H is nearly conserved by (5.26). It is often the case however, that for intermediate choices of n, the scheme in (2) outperforms its overdamped analogue (even accounting for the additional cost of the multiple evaluations of $\nabla\log\pi$ in Step 2).

An alternative approach to deriving possibly ergodic schemes based on the Hamiltonian ODE is to add appropriate random terms at each integration step. Indeed, adding $\mathcal{L}_H$ to the limiting generator $\mathcal{L}_O$ in (5.2) with π replaced by π_H yields a new limiting generator $\mathcal{L}_U = \mathcal{L}_H + \mathcal{L}_O$ or

$$\mathcal{L}_U f(x) = \frac{1}{\pi_H}\mathrm{div}\left(\pi_H(x)\nabla f(x)(S+J)(x)\right) \tag{5.27}$$

that satisfies

$$\pi_H\mathcal{L}_U = \pi_H\mathcal{L}_H + \pi_H\mathcal{L}_O = 0.$$

The subscript U in $\mathcal{L}_U$ stands for "underdamped."

Exercise 58. *Verify the above formula for the action of $\mathcal{L}_U$ on a function f.*

In fact, when J of form in (5.20) and H is an even function of $\tilde{x}$, expression (5.21) and the reversibility of $\mathcal{L}_O$ with respect to π_H imply that if f and g are even functions of $\tilde{x}$ then

$$\int f(x)\mathcal{L}_U g(x)\pi_H(dx) = \int g(x)\mathcal{L}_U f(x)\pi_H(dx). \tag{5.28}$$

Therefore we expect that discrete Markov chains corresponding (in the sense of Section **??**) to the limiting generator $\mathcal{L}_U$ should approximately preserve π_H. For example, the scheme

$$X_h^{(k+1)} = X_h^{(k)} - h\,(J+S)\,(X_h^{(k)})\nabla^{\mathrm{T}} H(X_h^{(k)}) + h\,\mathrm{div}\,(J+S)\,(X_h^{(k)}) + \sqrt{2h\,S(X_h^{(k)})}\,\xi^{(k)} \quad (5.29)$$

for independent $\xi^{(k)}$ with $\mathbf{E}\left[\xi^{(k)}\right] = 0$ and $\mathbf{cov}\left[\xi^{(k)}\right] = I$ (and finite higher moments) has limiting generator $\mathcal{L}_U$. Here S is assumed to be a symmetric positive semi-definite matrix and J is assumed to be anti-symmetric.

In the most common setup, one uses an J of form in (5.20) and H of form in (5.25) with

$$K(\tilde{x}) = \frac{\|\tilde{x}\|_2^2}{2},$$

and

$$S = \begin{bmatrix} 0 & 0 \\ 0 & \gamma \tilde{I} \end{bmatrix}$$

where $\gamma > 0$ and $\tilde{I}$ is the $\tilde{d} \times \tilde{d}$ identity matrix. With these choices and when $\hat{J}$ is constant, a discrete process with favorable properties (compared to (5.29)) is

$$\begin{aligned} \tilde{X}'_h &= \tilde{X}_h^{(\ell)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}} \log \pi(\hat{X}_h^{(\ell)}) \\ \hat{X}'_h &= \hat{X}_h^{(\ell)} + \frac{h}{2}\hat{J}\tilde{X}'_h \\ \tilde{X}''_h &= e^{-\gamma h}\tilde{X}'_h + \sqrt{(1-e^{-2\gamma h})}\,\xi^{(\ell+1)} \\ \hat{X}_h^{(\ell+1)} &= \hat{X}'_h + \frac{h}{2}\hat{J}\tilde{X}''_h \\ \tilde{X}_h^{(\ell+1)} &= \tilde{X}''_h + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}} \log \pi(\hat{X}_h^{(\ell+1)}) \end{aligned} \quad (5.30)$$

where ξ is a sequence of independent $\tilde{d}$ dimensional Gaussian random vectors with mean zero and identity covariance (and finite higher moments).

Exercise 59. *Check that the discrete time dynamics in* (5.30) *corresponds to the limiting generator $\mathcal{L}_U$ as claimed. How large (in terms of h) is the difference between the rescaled discrete generator $\mathcal{L}_h f$ and $\mathcal{L}_U f$ for smooth bounded test functions f with bounded derivatives?*

Finally, observe that if the "friction" coefficient γ is set to 0 this scheme reduces to the Velocity Verlet scheme (5.26).

5.4 Hybrid-MC and Metroplized Langevin schemes

The methods that we have described in this chapter all introduce some systematic error in our estimate of $\pi[f]$. For these schemes we expect that $\overline{f}_N$ converges not to $\pi[f]$ but to another quantity that converges to $\pi[f]$ when h is small. Given the many physical approximations often already inherent in the specification of π (e.g. approximate models), this additional systematic error can frequently be safely ignored. However, there are settings in which very high accuracy estimates of $\pi[f]$ are needed and the systematic error can be larger than the sampling (finite N) error. In these cases it may be worth while to "Metropolize" one of the schemes introduced in this chapter. By adding a Metropolis-Hastings type accept/reject step we can guarantee that $\overline{f}_N \to \pi[f]$ exactly as N increases. The cost for this improvement in accuracy is slower convergence (e.g. increased integrated autocorrelation time). For high dimensional problems it is often better to reduce the step size parameter h than to introduce Metroplization.

We begin by describing how our overdamped Langevin scheme in (5.9) can be modified so that it exactly preserves a target density π. As long as the density of noise variables, $\xi^{(k)}$, that we choose is non-zero everywhere, and the matrix S is positive definite, a single step of (5.9) defines a transition density $q(y \,|\, x)$ for which the ratio $q(x \,|\, y)/q(y \,|\, x)$ is finite and which can be therefore be used within the general Metropolis Hastings framework. For example, when the $\xi^{(k)}$ are Gaussian random variables and the matrix S is constant, the transition density is

$$q(y \,|\, x) \propto \exp\left(-\frac{(y - x - hS\nabla \log \pi(x))^{\mathrm{T}} S^{-1} (y - x - hS\nabla \log \pi(x))^{\mathrm{T}}}{4h}\right)$$

as can be verified by writing down the density for the Gaussian random variable $\xi^{(k)}$ and then changing variables from $\xi^{(k)}$ to $X_h^{(k)}$. Note that the formula is more complicated when S depends on position because the change of variables from $\xi^{(k)}$ to $X_h^{(k)}$ is non-linear. In any case, with q in hand we can use the standard Metropolis-Hastings procedure:

Algorithm 3. *Metropolized overdamped Langevin*

1. Let $Y^{(k+1)}$ be the result of a single step of (5.9) starting from initial point $X^{(k)}$.

2. With probability

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

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

Recalling that (5.9) is approximately reversible with respect to π and that the rejection probability is a measure of the distance between the densities $q(y \,|\, x)\pi(x)$ and $q(x \,|\, y)\pi(y)$, we should expect the rejection probability for a Metropolis-Hastings scheme with proposal density $q(y \,|\, x)$ corresponding to (5.9) to be very small. To see that this is indeed the case, consider the one dimensional case with $S = 1$. Letting $V(x) = -\log \pi(x)$, and using the change of variables

$$y \to x - hV'(x) + \sqrt{2h}\,\xi,$$

we obtain

$$\begin{aligned}\int p_{rej}(x)\pi(dx) &= \int \left|1 - \frac{q(x \,|\, y)\pi(y)}{q(y \,|\, x)\pi(x)}\right| q(y \,|\, x)\pi(x)dydx \\ &= \int \Bigg|1 - e^{-V(x-hV'(x)+\sqrt{2h}\,\xi)+V(x)} \\ &\qquad \times e^{\frac{-(\xi-\sqrt{h/2}V'(x)-\sqrt{h/2}V'(x-hV'(x)+\sqrt{2h}\,\xi))^2+\xi^2}{2}}\Bigg| e^{-\frac{\xi^2}{2}}\pi(x)dydx.\end{aligned}$$

The change of variables has re-expressed the average rejection probability as an integral of a non-negative quantity over two independent random variables both of which are independent of h. First, notice that

$$V'(x - hV'(x) + \sqrt{2h}\,\xi) = V'(x) + \sqrt{2h}V''(x)\xi + \mathcal{O}\left(h\right).$$

Expanding the terms in the integrand we find that

$$\begin{aligned}&-(\xi - \sqrt{h/2}V'(x) - \sqrt{h/2}V'(x - hV'(x) + \sqrt{2h}\,\xi))^2 + \xi^2 \\ &\qquad = -(\xi - \sqrt{2h}V'(x) - hV''(x)\xi + \mathcal{O}(h^{3/2}))^2 + \xi^2 \\ &\qquad = 2\sqrt{2h}V'(x)\xi + 2hV''(x)\xi^2 - 2h\left(V'(x)\right)^2 + \mathcal{O}(h^{3/2})\end{aligned}$$

and

$$V(x) - V(x - hV'(x) + \sqrt{2h}\,\xi) = h(V'(x))^2 - \sqrt{2h}V'(x)\xi \\ - \frac{1}{2}V''(x)\left(\sqrt{2h}\xi - hV'(x)\right)^2 + \mathcal{O}(h^{3/2}) \\ = h(V'(x))^2 - \sqrt{2h}V'(x)\xi - hV''(x)\xi^2 + \mathcal{O}(h^{3/2}).$$

So we see that

$$\int p_{rej}(x)\pi(dx) = \int \left|1 - e^{\mathcal{O}(h^{3/2})}\right| e^{-\frac{\xi^2}{2}}\pi(x)dydx = \mathcal{O}(h^{3/2})$$

and we expect the rejection rate to be very small at least when h is small.

Now let

$$H(\hat{x}, \tilde{y}) = -\log\pi(\hat{x}) + K(\tilde{y})$$

where $K(-\tilde{x}) = K(\tilde{x})$ and let π_H be the density proportional to e^{-H}. Suppose that we wish to modify Algorithm 2 so that it still preserves π_H, even though the exact solution to the Hamiltonian ODE (5.15), $y^{(t)}$, is replaced by $y_h^{(k)}$, the Velocity Verlet discrete time approximation in (5.26). The difference between the trajectories of $y^{(t)}$ and $y_h^{(k)}$ over a finite time interval (so over $\mathcal{O}(h^{-1})$ steps) is $\mathcal{O}(h^2)$. This small error translates into a small error in the sampling scheme (2). The error can either be reduced to tolerable levels by decreasing h (and correspondingly increasing n) or it can be eliminated altogether by introducing a Metropolis accept-reject step: sample $\tilde{X}^{(0)}$ from $\exp(-K(\tilde{x}))/\tilde{\mathcal{Z}}$ and, given a sample $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(k)})$, generate $X^{(k+1)}$ by

Algorithm 4. *Hybrid Monte Carlo*

1. Independently sample variable $\tilde{Y}(k)$ from the density proportional to $\exp(-K(\tilde{x}))$
2. Set $Y^{(k+1)} = y_h^{(n)}\left(\hat{X}^{(k)}, \tilde{Y}^{(k)}\right)$ where $y_h^{(\ell)}(x)$ solves (5.26) with initial conditions x and n is chosen by the user.
3. With probability

$$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{\pi_H(Y^{(k+1)})}{\pi_H(\hat{X}^{(k)}, \tilde{Y}^{(k)})}\right\}$$

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

Just as in Algorithm 1, Step 1 exactly preserves π_H. However, unlike Algorithm 1, Step 2 does not exactly preserve π_H. The purpose of Step 3 is to correct the error introduced in Step 2. Examining this procedure, the first question the reader should ask is "why does the transition density $q(y \,|\, x)$ not appear in the acceptance probability?" The proposal density corresponding to Step 2 is

$$q(y \,|\, x) = \delta(y - y_h^{(t_n)}(x)) \tag{5.31}$$

which certainly does not satisfy $q(y \,|\, x) = q(x \,|\, y)$.

In fact, Algorithm 1 is an example of another kind of Metropolis framework. Suppose that φ is a smooth involution, i.e. $\varphi = \varphi^{-1}$ and, given a target density μ and a proposal density $q(y|x)$ define the transformed densities

$$\mu_\varphi(x) = |D\varphi(x)|\, \mu(\varphi(x))$$

and

$$q_\varphi(y \,|\, x) = |D\varphi(y)|\, q(\varphi(y)|\varphi(x))$$

where we have used the fact that, for a smooth involution,

$$|D\varphi(x)| = \frac{1}{|D\varphi(\varphi(x))|}.$$

Note that the mapping $\varphi(x) = (\hat{x}, -\tilde{x})$ is a smooth involution. In fact, when H is an even function of the $\tilde{x}$ variables as we have assumed, this involution preserves the density π_H.

Now consider the following slight generalization of the Metropolis-Hastings scheme:

Algorithm 5. *Metroplis-Hastings with Involution*

1. `Generate a random variable` $Y^{(k+1)}$ `from the proposal distribution` $q(y \,|\, X^{(k)})$.

2. `With probability`

$$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{q_\varphi(X^{(k)} \,|\, Y^{(k+1)})\, \mu_\varphi(Y^{(k+1)})}{q(Y^{(k+1)} \,|\, X^{(k)})\, \mu(X^{(k)})}\right\}$$

`set` $X^{(k+1)} = Y^{(k+1)}$. `Otherwise set` $X^{(k+1)} = \varphi(X^{(k)})$.

Instead of enforcing reversibility, Algorithm 5 results in a chain with transition operator $\mathcal{T}$ satisfying

$$\int f_\varphi(x)\mathcal{T}g_\varphi(x)\mu(dx) = \int g(x)\mathcal{T}f(x)\mu(dx) \tag{5.32}$$

for any test functions f and g and with $f_\varphi(x) = f(\varphi(x))$. Like reversibility, this condition implies that the resulting Markov chain preserves μ.

Exercise 60. *Show that for any density μ, if the transition density for a Markov chain satisfies* (5.32) *then the Markov chain preserves μ.*

In order to show that Algorithm 5 satisfies (5.32), first observe that

$$\begin{aligned}
&p_{acc}(\varphi(x),\varphi(y))\, q_\varphi(y \mid x)\, \mu_\varphi(x) \\
&\qquad = \min\left\{ q_\varphi(y|x)\mu_\varphi(x), \frac{q_\varphi(\varphi(x)|\varphi(y))\mu_\varphi(\varphi(y))q_\varphi(y \mid x)\, \mu_\varphi(x)}{q(\varphi(y) \mid \varphi(x))\mu(\varphi(x))} \right\} \\
&\qquad = \min\left\{ q_\varphi(y|x)\mu_\varphi(x), q(x|y)\mu(y) \right\} \\
&\qquad = p_{acc}(y,x)\, q(x \mid y)\, \mu(y)
\end{aligned}$$

where, to obtain the third equality we have used the fact that

$$\frac{q_\varphi(\varphi(x) \mid \varphi(y))\mu_\varphi(\varphi(y))}{q(\varphi(y)|\varphi(x))\mu(\varphi(x))} = \frac{q(x|y)\mu(y)}{q_\varphi(y|x)\mu_\varphi(x)}.$$

The transition operator resulting from Algorithm 5 is

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

where

$$p_{rej}(x) = 1 - \int q(y \mid x)p_{acc}(x,y)dy.$$

From these expressions we see that

$$\begin{aligned}\int f_\varphi(x)\mathcal{T}g_\varphi(x)\mu(dx) &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx) \\ &\qquad + \int f_\varphi(x)g_\varphi(y)p_{acc}(x,y)q(y|x)\mu(x)dydx \\ &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx) \\ &\quad + \int f(x)g(y)\, p_{acc}(\varphi(x),\varphi(y))\, q_\varphi(y|x)\mu_\varphi(x)dxdy \\ &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx) \\ &\qquad + \int g(x)f(y)p_{acc}(x,y)q(y\,|\,x)\mu(x)dxdy \\ &= \int g(x)\mathcal{T}f(x)\mu(dx).\end{aligned}$$

Returning to Algorithm 4 and the particular transition density $q(y\,|\,x)$ in (5.31), observe that, for any test function f of $x \in \mathbb{R}^{\hat{d}}$ and $y \in \mathbb{R}^{\tilde{d}}$, volume preservation and time reversibility of the flow map $x \to y_h^{(n)}(x)$ imply that

$$\begin{aligned}\int f(x,y)q(x\,|\,y)dxdy &= \int f(x, y_h^{(n)}(x))dx \\ &= \int f(y_h^{(-n)}(y), y)dy \\ &= \int f((\hat{y}_h^{(n)}(\hat{y},-\tilde{y}), -\tilde{y}_h^{(n)}(\hat{y},-\tilde{y}), y)dy \\ &= \int f(x,y)q(\hat{x},-\tilde{x}\,|\,\hat{y},-\tilde{y})dxdy.\end{aligned}$$

Since this is true for any f we can conclude that

$$q_\varphi(x\,|\,y) = q(y\,|\,x)$$

for the particular involution $\varphi(x) = (\hat{x}, -\tilde{x})$, which explains why we can omit q in the acceptance probability in Algorithm 4. Since this involution also preserves π_H, we find that the acceptance probability in Algorithm 4 is indeed of the form in Algorithm 5. Finally note that, since the $\tilde{x}$ variables are

randomized at each iteration of Algorithm 4, we could set $X^{(k+1)} = \varphi(X^{(k)})$ upon rejection rather than $X^{(k+1)} = X^{(k)}$ without changing the distribution of resulting chain.

Exercise 61. *Though the full Markov chain X generated by Algorithm 4 does not satisfy detailed balance, the process $\hat{X}$ is also a Markov chain and is reversible. Show this.*

Before moving on we point out that the arguments above did not actually require that the Hamiltonian appearing in the acceptance probability in Algorithm 4 be the same as the one used to define the evolution (5.26). In fact, as long as the mapping $x \to y_h^{(n)}(x)$ is volume preserving time reversible and H is even in $\tilde{x}$, $y_h^{(n)}(x)$ need have no relationship at all with H. Of course the reason we expect Hybrid Monte Carlo to be an effective method is that, for sufficiently small h, the solution of (5.26) should very nearly conserve the value of H and therefore result in a high acceptance probability. Certainly if the error introduced is not controllable (by reducing some user defined parameter) then one cannot justify omitting the accept/reject step. So introducing an evolution that does not nearly conserve H may be counter-productive.

Finally, we turn our attention to Metropolizing the underdamped Langevin scheme in (5.30). The transition density for a single step of (5.30) is

$$q(y \,|\, x) \propto \delta\left(\hat{y} - \hat{x} - \frac{h}{2}\hat{J}(\tilde{y} + \tilde{x}) + \frac{h^2}{4}\hat{J}\hat{J}^{\mathrm{T}}\left(\nabla^{\mathrm{T}} \log \pi(\hat{y}) - \nabla^{\mathrm{T}} \log \pi(\hat{x})\right)\right) \times r(x, y) \quad (5.33)$$

where we have introduced the function

$$r(x, y) = \exp\left(-\frac{\|\tilde{y} - e^{-\gamma h}\tilde{x} - \frac{1}{2}h\hat{J}^{\mathrm{T}}\left(e^{-\gamma h}\nabla^{\mathrm{T}} \log \pi(\hat{x}) + \nabla^{\mathrm{T}} \log \pi(\hat{y})\right)\|_2^2}{2(1 - e^{-2h\gamma})}\right).$$

Exercise 62. *Show that the transition density for a single step of* (5.30) *is the one given in* (5.33).

The transition operator corresponding to the proposal distribution $q(y \,|\, x)$ in (5.33) satisfies (5.32) with $\mu = \pi_H$ and $\varphi(x) = (\hat{x}, -\tilde{x})$ to within an $\mathcal{O}(h^2)$ error.

Exercise 63. *Show this.*

Plugging the proposal density in (5.33) into Algorithm 5, we find that the delta functions exactly cancel and we obtain

Algorithm 6. *Metropolized underdamped Langevin*

1. `Let` $Y^{(k+1)}$ `be the result of a single step of` (5.30) `starting from initial point` $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(t)})$.

2. `With probability`

$$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{\pi_H(Y^{(k+1)})\, r(\hat{Y}^{(k+1)}, -\tilde{Y}^{(k+1)}, \hat{X}^{(k)}, -\tilde{X}^{(k)})}{\pi_H(X^{(k)})\, r(X^{(k)}, Y^{(k+1)})}\right\}$$

`set` $X^{(k+1)} = Y^{(k+1)}$. `Otherwise set` $X^{(k+1)} = \left(\hat{X}^{(k)}, -\tilde{X}^{(k)}\right)$.

Exercise 64. *Consider a set of 2-dimensional vectors, $\vec{\sigma}_i \in \mathbb{R}^2$ indexed by the 1-dimensional periodic lattice $\mathbb{Z}_L$ and with $\|\vec{\sigma}_i\|_2 = 1$. The nearest neighbor XY model of statistical physics assigns to these vectors the density*

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

In terms of the angles $\theta_i \in [-\pi, \pi)$ of the vectors $\vec{\sigma}_i$, this density becomes

$$\pi(\theta) = \frac{e^{\beta \sum_{i \leftrightarrow j} \cos(\theta_i - \theta_j)}}{\mathcal{Z}}.$$

Write a routine to sample the XY model using both (5.8) *and a Metropolized version of* (5.8)*. Compare the Metropolized and un-Metropolized schemes for different values of h (but note that the total number of time-steps you use should scale like h^{-1}). Make your comparisons in terms of the integrated autocorrelation time of the variable*

$$\frac{M_1(\sigma)}{\|M(\sigma)\|_2}$$

which is the cosine of the angle of the magnetization vector,

$$M(\sigma) = \sum_{i=0}^{L-1} \vec{\sigma}_i \in \mathbb{R}^2.$$

What do you observe when L increases.

Exercise 65. *Write a routine to sample the XY model described in Exercise 64 using Algorithm 4 both with and without the Metropolis accept/reject step. You're free to choose the other parameters of the algorithm (K, J, n, and h) as you like (but be clear about your choices). Compare the results to those of Exercise 64 using integrated autocorrelation time of the cosine of the angle of magnetization as your measure of efficiency. Make sure you are accounting correctly for the cost to generate each step of the chain (e.g. as measured by the number of evaluations of $\nabla \log \pi$.*

Exercise 66. *Sample the XY model described in Exercise 64 using* (5.30) *with $\hat{J} = \hat{I}$. Compare to the Metropolized scheme in Algorithm 6 for different values of h and for different values of γ using integrated autocorrelation time of the cosine of the angle of magnetization as your measure of efficiency. Compare to your results from Exercises 64 and 65. Which of the schemes do you prefer for sampling the XY model?*

5.5 bibliography