# 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), $\Theta$. 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 $\Theta$. 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 $\Theta$ 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 $\sigma^2$, 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 $\Theta$, we can easily generate samples of $X$. However, the goal here is usually to produce an estimate of the true value of $\Theta$ from obsevations of $X$. A reasonable approximation of the true value of $\Theta$ is given by the "posterior mean"

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

where $\pi(\theta|x)$ is the "posterior density" of $\Theta$ 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 $\Theta$ 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 $\pi$, 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 $\pi$ 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

$\delta$. Assume for the moment that $x \in \mathbb{R}$ and $\pi$ is supported on the interval $[0, 1]$. You probably remember from your calculus classes that if $f$ and $\pi$ 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 $\pi$ 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 $\alpha$ 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 $\delta$. 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 $\sigma$-algebras

First recall that a probability measure, $\mathbf{P}$, is a map from a collection of subsets $\mathcal{F}$ of some (possibly abstract) space $\Omega$ 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 $\sigma$-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,

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

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

and

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

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 $\Omega$ 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 $\Omega$ for which a particular statement is true. For example if $\Omega$ 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 $\sigma$-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 $\Omega$ 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 $\Omega$. Let*

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

*for each $\omega \in \Omega$. You can easily verify that $\mathcal{F}$ is a $\sigma$-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 $\sigma$-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 $\sigma$-algebra $\mathcal{B}$ on $\mathbb{R}^d$ is the smallest $\sigma$-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 $\omega$ 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 $\pi$ 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 $\pi$. 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 $\pi$ as a density or distribution depending on the context. When we write $\pi(x)$ we are implicitly assuming that the distribution $\pi$ 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 $\pi$ and

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

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

**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 $\sigma$-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 $\sigma$-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 $\sigma$-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 $\sigma$-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 $\{\}$, $\Omega$, $B$, and $B^{\mathrm{c}}$ 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 $\sigma$-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 $\mathbf{E}[X \mid Y]$ as a function of the possible values taken by $Y$ becomes

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

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

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

### 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($\lambda$) 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($\lambda$) 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 $\sigma^2/N$ which vanishes as $N \to 0$. Moreover, there is no reason to believe that $\sigma^2$ will depend on dimension. When the samples are pairwise independent random variables, the number of samples to achieve accuracy $\delta$ 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 $\omega$ in a subset of $\Omega$ 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 $\Lambda_N$ for $\lambda$ 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 $\lambda$ 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 $\lambda$, we expect is roughly of size $\mathcal{O}\left(N^{-3/2}\right)$ (this is true for all $\lambda$ 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 $\lambda$. 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 $\gamma$) 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 $\Lambda$ 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