8 Additional Topics

8.1 Sampling from Probability Distributions

We will first talk about how to generate samples from a probability distribution. Why is this useful? Suppose we wish to simulate a radioactive decay process on a computer. A reasonable model for the times between subsequent decays of our radioactive isotope is the exponential distribution. Thus we need to be able to generate samples from an exponential distribution. For common distributions, such as the exponential distribution, there are built-in routines in Matlab or SciPy (python) to do this. It is useful, however, to know how this works since many times the distribution you want to sample does not have a built-in routine.

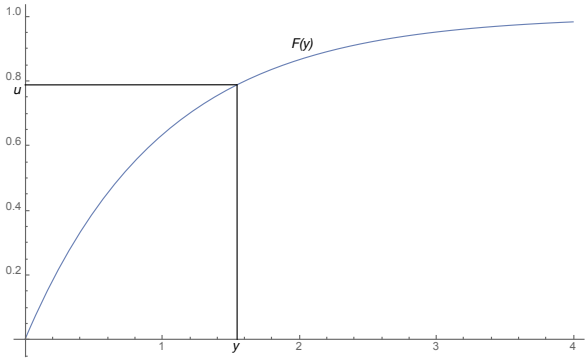
For simplicity, we will only discuss sampling from continuous random variables. The same ideas work for discrete random variable, with a few alterations. We will also assume that we have a method for generating samples from the Uniform[0, 1] distribution. This is a very interesting problem, and there are many ways to do it, most of which are not straightforward. If you find this interesting, the topic is discussed in courses on computational probability and in some computer science courses. We will discuss two methods of sampling from probability distributions: the inverse CDF method and rejection sampling.

8.1.1 Inverse CDF method

Suppose we have a population whose distribution is characterized by a CDF $F(y)$. Recall that for a random variable Y, the CDF is defined as $F(y) = \mathbb{P}(Y \leq y)$. We would like to generate samples from the population. The inverse CDF method works as follows:

1. Generate U, a sample from the Uniform[0, 1] distribution.
2. Let Y be the largest number x such that $F(x) \leq u$. If the CDF $F(y)$ is *strictly* increasing, then $F(y)$ is invertible so we can let $Y = F^{-1}(U)$.

It is perhaps easier to see this on a picture. This is an example of the inverse CDF method used on a population which has a exponential distribution with parameter $\lambda = 1$.



Why does this work. For simplicity, let's consider only the case where the CDF $F(y)$ is invertible. (Invertibility of the CDF is not required; in particular, it works for discrete random variables, whose CDF is not invertible). Let $U \sim \text{Uniform}[0, 1]$. Then we claim the distribution of $F^{-1}(U)$ is F. To see this, we look at the CDF of $F^{-1}(U)$.

$$\begin{aligned}\mathbb{P}(F^{-1}(U) \leq y) &= \mathbb{P}(U \leq F(y)) \\ &= F(y)\end{aligned}$$

where we used the fact that for a Uniform[0, 1] random variable U, the CDF is $\mathbb{P}(U \leq u) = u$ for $u \in [0, 1]$.

Now that we've seen the picture, let's use the inverse CDF method to sample from an exponentially distributed population.

Example. Suppose we have a population which has an exponential distribution with parameter λ. Let U be a Uniform[0, 1] random variable. Use the inverse CDF method to generate a sample from the population in terms of U.

First we need to find the CDF for the population. Integrating the density function from 0 to y:

$$\begin{aligned}F(y) &= \int_0^y \lambda e^{-\lambda t} dt \\ &= e^{-\lambda t}\Big|0^y \\ &= 1 - e^{-\lambda y}\end{aligned}$$

With appropriate bounds, the CDF is:

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

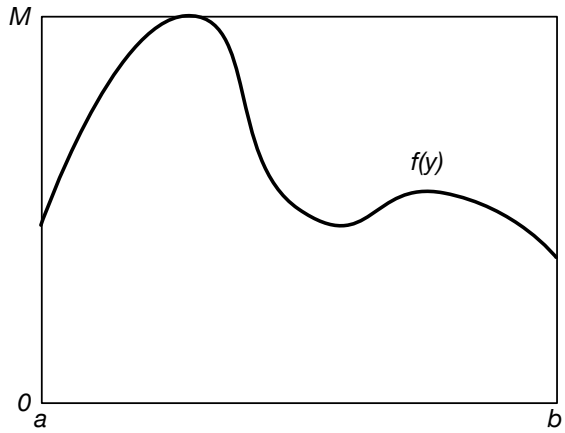
The CDF $F(y)$ is strictly increasing, it has an inverse. Since we want $Y = F^{-1}(U)$, we take $F(Y) = U$ and solve for Y.

$$\begin{aligned}1 - e^{-\lambda Y} &= U \\ e^{-\lambda Y} &= 1 - U \\ -\lambda Y &= \log(1 - U) \\ Y &= -\frac{1}{\lambda}\log(1 - U)\end{aligned}$$

For the exponential distribution this is very straightforward. In general this method works if we can easily compute the CDF. For continuous distributions, if the integral of the density has a nice closed form (as in the exponential case), the inverse CDF method usually works very well. For discrete distributions, this method also works well since to find the discrete CDF, all we have to do is add up the probabilities of the appropriate simple events. For cases where the CDF does not have a closed form (such as the normal distribution) or cases where the CDF is hard to invert, this method is not so good. For many of those cases, rejection sampling is the way to go.

8.1.2 Rejection Sampling

Rejection sampling is based on the "dartboard principle". Here's how it works. Imagine you have a continuous probability density function $f(x)$ you wish to sample from. Furthermore, imagine that the density function is nonzero only on the interval $[a, b]$. Put the density function on a rectangular dartboard. The bounds of the dartboard are from a to b in the x-direction and from 0 to M in the y-direction, where M is the maximum of $f(x)$ on $[a, b]$[1]. Here is an example of a possible dartboard.



Now throw darts uniformly at the dartboard until your dart lands under the density curve $f(x)$. In other words, reject all darts which do not land under the density curve (this is why we call this rejection sampling). Then your dart will be uniformly distributed in the region between the x-axis and the density curve, and the x-coordinate of your dart will be distributed according to the density $f(x)$. Intuitively, this works since there is more room on the board for (nonrejected) darts to land where the density curve is highest, i.e. where the probability density is greatest.

Let's show mathematically that this actually works. We have all the tools we need from the section on multivariate distributions! Let (X, Y) be the position of a nonrejected dart. Then the pair (X, Y) is uniformly distributed on the region between the x-axis and the density curve. Since we are assuming that the density $f(x)$ is zero outside $[a, b]$, the area of the region is $\int_a^b f(x)dx = 1$ since $f(x)$ is a probability density function. Since the joint density function of a uniform distribution is the reciprocal of the area of the region, for our joint density function we have joint density:

$$f(x, y) = \begin{cases} 1 & a \leq x \leq b, 0 \leq y \leq f(x) \\ 0 & \text{otherwise} \end{cases}$$

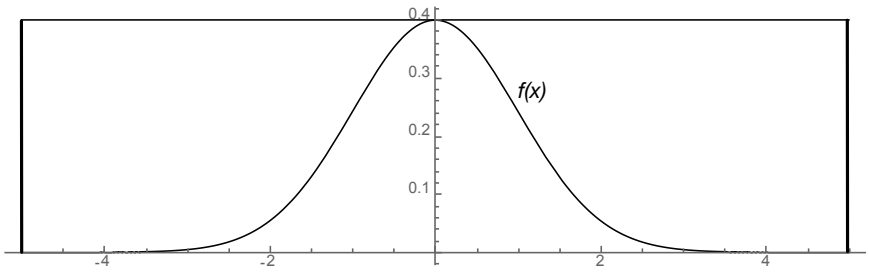
We claim the marginal density of X is the density $f(x)$. To see this, all we have to do is

[1]Recall that a continuous function has an absolute maximum on a closed interval $[a, b]$ and that density functions are nonnegative; thus this rectangular dartboard will have finite size.

integrate the joint density in y.

$$\begin{aligned} f_X(x) &= \int_0^{f(x)} 1 dy \\ &= y\Big|_0^{f(x)} \\ &= f(x) \end{aligned}$$

Thus the rejection sampling method produces works as advertised. There are a few disadvantages to the method, however. First, we need to have a bounded region for the density function. Many useful probability distributions, such as the normal distribution, are unbounded. To use rejection sampling for the standard normal distribution, for example, we have to impose artificial bounds on the density. For example, we could impose the bounds [-5, 5]. Since the probability is exceedingly low that a sample will be more than 5 standard deviations from the mean, this is not too unreasonable; it is important, however, to know that in doing this we are reducing the probability of extreme outliers to 0, which may not be what we want to do, especially if we are taking large numbers of samples. In addition, depending on the shape of the density function, we might have to throw many darts in order for one to not be rejected. This presents no problem theoretically, but might be a problem computationally. Take a look at the following picture of the standard normal density between -5 and 5.



From the picture, the area of the rectangular dartboard is approximately 4, and the area under the curve is approximately 1, since $f(x)$ is a probability density function. Thus we expect only 1/4 of our darts to be accepted, so we will have to throw on average 4 darts to get a single sample.

One way around this computation inefficiency is to note that the x-position of our darts does not have to be uniform. In fact, it makes sense to select the x-position of our darts according to a density function $g(x)$ which is similar to the one we are trying to simulate and is easy to take samples from. We can think of this as throwing darts at a non-square dartboard.

Once again, suppose we are trying to generate a sample from a probability density function $f(x)$, where $f(x)$ is zero outside a closed interval $[a, b]$. Suppose we can generate samples from another probability density function $g(x)$, either using the inverse CDF method or some other method. The function $g(x)$ will give us the shape of our dartboard. For this to work,

the function $f(x)$ must fit entirely on our dartboard. To make this happen, we scale $g(x)$ (if needed) by a constant factor M so that $f(x) \leq Mg(x)$ for all $x \in [a, b]$. We now throw darts uniformly at this dartboard.

At this point, the dartboard analogy breaks down a bit, and it is easier to just give the algorithm.

1. Choose Y uniformly from the interval $[0, 1]$.
2. Choose X according to probability density $g(x)$.
3. If $Y \leq f(X)/Mg(X)$, then accept the sample (X, Y). X is then the desired sample from the distribution $f(x)$.
4. Otherwise reject the sample and repeat from step 1.

Mathematically, why does this work. The proof uses Bayes' theorem. Let A be the event that the sample is accepted. Then by Bayes' theorem (fudging a little because density functions are not exactly probabilities but are close enough),

$$\mathbb{P}(X = x|A) = \frac{\mathbb{P}(A|X = x)P(X = x)}{\mathbb{P}(A)}$$

Since Y is a uniform random variable on [0, 1],

$$\begin{aligned}\mathbb{P}(A|X = x) &= \mathbb{P}\left(Y \leq \frac{f(x)}{Mg(x)}\right) \\ &= \frac{f(x)}{Mg(x)}\end{aligned}$$

where we used the density of the uniform random variable on [0, 1]. The probability $\mathbb{P}(X = x) = g(x)$, since X is chosen according to density function $g(x)$. As mentioned above, this is not quite accurate since densities are not really probabilities, but this is good enough for our purposes. By the Law of Total Probability for continuous random variables (essentially the same as the discrete case, except we replace summation with integration),

$$\begin{aligned}\mathbb{P}(A) &= \int_a^b \mathbb{P}(A|X = x)\mathbb{P}(X = x)dx \\ &= \int_a^b \frac{f(x)}{Mg(x)}g(x)dx \\ &= \frac{1}{M}\int_a^b f(x)dx \\ &= \frac{1}{M}\end{aligned}$$

where we used the fact that $f(x)$ is a density function, thus integrates to 1 over the interval $[a, b]$. Thus the probability of accepting a sample is $1/M$. Putting all of this together,

$$\begin{aligned}
\mathbb{P}(X = x|A) &= \frac{\mathbb{P}(A|X = x)P(X = x)}{\mathbb{P}(A)} \\
&= \frac{\frac{f(x)}{Mg(x)}g(x)}{\frac{1}{M}} \\
&= f(x)
\end{aligned}$$

Thus the rejection sampling method produces a sample which is distributed according to the desired distribution $f(x)$. The probability of accepting a sample is $1/M$. Thus if we treat the rejection sampling procedure as a sequence of Bernoulli trials with probability of success, then on average it should take M trials for a sample to be accepted. (We model the number of trials needed for the first success as a geometric random variable, and use the formula for the expected value of a geometric distribution.)