Sampling Methods

Chapter 9 (Ch10 from textbook)

Prof. Reza Azadeh

University of Massachusetts Lowell

Generating Pseudo Random Numbers

- Computer-based vs. Hardware-based
- An early computer-based PRNG, suggested by John von Neumann in 1946, is known as the middle-square method.
    1. take any number
    2. square it
    3. get the middle digits as the “random number”
    4. use that number as the seed for the next iteration.

Example: $1111 \rightarrow 01234321 \rightarrow 2343 \rightarrow \ldots$

Motivation

- For most probabilistic models of practical interest, exact inference is intractable, and so we have to resort to some form of approximation.
- There exists a set of inference algorithms that rely on deterministic approximation (e.g., variational Bayes and expectation propagation).
- In this chapter, we study another category of inference algorithms that rely on numerical sampling, known as **Monte Carlo** techniques.

Goals

- The Monte Carlo methods consider the overall problem of inferring the posterior distribution.
- For most situations the posterior distribution is required primarily for the purpose of evaluating expectations, for example for making predictions.
- In this chapter, we address the problem of finding the expectation of some function $f(\mathbf{z})$ where $\mathbf{z}$ can be discrete or continuous, or a combination of the two.

Expectation (1)

Recall that for a discrete variable $\mathbf{z}$, we can write the expectation of $f(\mathbf{z})$ as

$$\mathbb{E}[f] = \sum_{\mathbf{z}} f(\mathbf{z})p(\mathbf{z})$$

and for a continuous variable $\mathbf{z}$, we can write the expectation of $f(\mathbf{z})$ as

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Expectation (2)

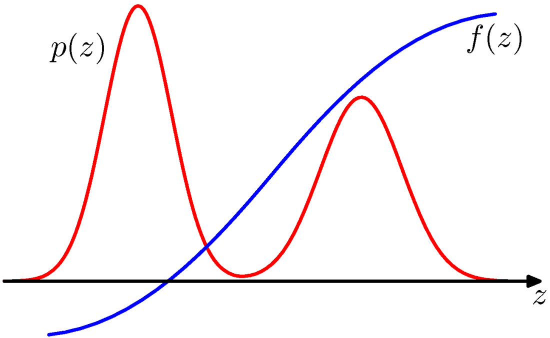

Figure: a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$

We suppose that such expectations are too complex to be evaluated exactly using analytical techniques.

How sampling methods approach this problem?

The general idea behind sampling methods is to obtain a set of samples $\mathbf{z}^{(l)}$ where $l = 1, \ldots, L$, drawn independently from the distribution $p(\mathbf{z})$.

This allows the expectation to be approximated by a finite sum

$$\hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}).$$

Estimator's Mean

As long as the samples are drawn from the distribution $p(\mathbf{z})$, then

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

and so the estimator $\hat{f}$ has the correct mean.

Estimator's Variance

The variance of the estimator is given by

$$\mathbb{V}[\hat{f}] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

is the variance of the function $f(\mathbf{z})$ under the distribution $p(\mathbf{z})$.

Estimator's Accuracy and Subsequent Issues

- The accuracy of the estimator does not depend on the dimensionality of $\mathbf{z}$, and high accuracy may be achieved with a relatively small number of samples.
- In practice, ten or twenty independent samples may suffice to estimate an expectation to sufficient accuracy.
- **Problem 1:** the samples might not be independent, and so the effective sample size might be much smaller than the apparent sample size.
- **problem 2:** if $f(\mathbf{z})$ is small in regions where $p(\mathbf{z})$ is large, and vice versa, then the expectation may be dominated by regions of small probability, implying that relatively large sample sizes will be required to achieve sufficient accuracy.

Basic Sampling Algorithms

- We consider some simple strategies for generating random samples from a given distribution.
- We assume that we have access to an algorithm that generates pseudo-random numbers distributed uniformly over $(0, 1)$.

In Python, you can use `numpy.random.rand()`, `numpy.random.random_sample()`, or `rng.random()` where `rng = np.random.default_rng()`.

Standard Distributions (1)

We first consider how to generate random numbers from simple nonuniform distributions, assuming that we already have access to a source of uniformly distributed random numbers.

Suppose that z is uniformly distributed over the interval $(0, 1)$, and that we transform the values of z using some function $f(.)$ so that $y = f(z)$. The distribution of y will be governed by

$$p(y) = p(z)|\frac{dz}{dy}|$$

where, in this case, $p(z) = 1$.

Goal: Choose function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$.

Standard Distributions (2)

The indefinite integral of $p(y)$ can be written as

$$z = h(y) \equiv \int_{-\infty}^{y} p(\hat{y}) d\hat{y}$$

Thus, $y = h^{-1}(z)$, and so we have to transform the uniformly distributed random numbers using a function which is the inverse of the indefinite integral of the desired distribution.

Standard Distributions (3)

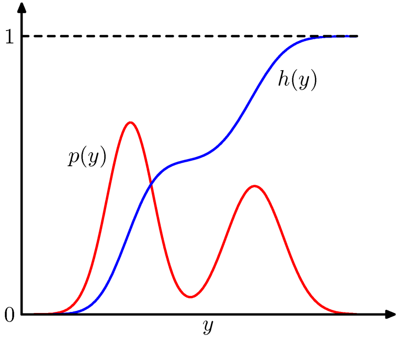

Figure: Geometrical interpretation of the transformation method for generating nonuniformly distributed random numbers. $h(y)$ is the indefinite integral of the desired distribution $p(y)$. If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$, then y will be distributed according to $p(y)$

Box-Muller Method (1)

This method generates samples from a Gaussian distribution.

1. First suppose we generate pairs of uniformly distributed random numbers $z_1, z_2 \in (-1, 1)$, which we can do by transforming a variable distributed uniformly over $(0, 1)$ using $z \to 2z - 1$.
2. Next we discard each pair unless it satisfies $z_1^2 + z_2^2 \leq 1$. This leads to a uniform distribution of points inside the unit circle with $p(z_1, z_2) = 1/\pi$.

Box-Muller Method (2)

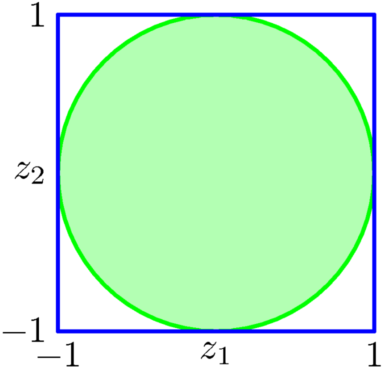

Figure: The Box-Muller method for generating Gaussian distributed random numbers starts by generating samples from a uniform distribution inside the unit circle.

Box-Muller Method (3)

3. Then, for each pair z_1, z_2, we evaluate the quantities

$$y_1, y_2 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}, z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$

   where $r^2 = z_1^2 + z_2^2$.

4. Then the joint distribution of y_1 and y_2 is given by:

$$\begin{aligned} p(y_1, y_2) &= p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| \\ &= \left[\frac{1}{\sqrt{2\pi}}\exp(-y_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-y_2^2/2)\right] \end{aligned}$$

   So y_1 and y_2 are independent and each has a Gaussian distribution with zero mean and unit variance.

Visualization of Box-Muller Method

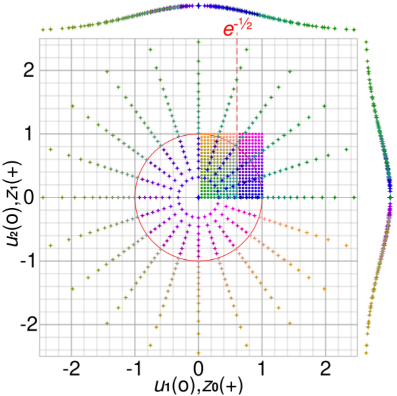

Figure: Colored points in the unit square (z_1, z_2), drawn as circles, are mapped to a 2D Gaussian (y_1, y_2), drawn as crosses. The plots at the margins are the probability distribution functions of y_1 and y_2. Source: Wikipedia

$\mathcal{N}(0,1) \to \mathcal{N}(\mu,\sigma^2)$

If y has a Gaussian distribution with zero mean and unit variance, then $\sigma y + \mu$ will have a Gaussian distribution with mean μ and variance σ^2.

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

To generate vector-valued variables having a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we can make use of the *Cholesky decomposition*, which takes the form $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$.

Then if $\mathbf{z}$ is a vector valued random variable whose components are independent and Gaussian distributed with zero mean and unit variance, then $\mathbf{y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$ will have mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Issues

The transformation technique depends on its success on the ability to calculate and then invert the indefinite integral of the required distribution. Such operations will only be feasible for a limited number of simple distributions, and so we must turn to alternative approaches in search of a more general strategy.

Rejection Sampling (1)

- Problem: Suppose we wish to sample from a distribution $p(\mathbf{z})$ that is not one the simple, standard distributions considered so far, and that sampling directly from $p(\mathbf{z})$ is difficult.
- Assumption: Suppose we can easily evaluate $p(\mathbf{z})$ for any given value of $\mathbf{z}$, up to some normalizing constant Z so that

$$p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

where $\tilde{p}(z)$ can readily be evaluated, but Z_p is unknown.

Rejection Sampling (2)

- We need a simpler distribution $q(z)$, sometimes called a **proposal distribution**, from which we can readily draw samples.
- We then introduce a constant k whose value is chosen such that $kq(z) \geq \tilde{p}(z)$ for all values of z. The function $kq(z)$ is called the comparison function.

Rejection Sampling (3)

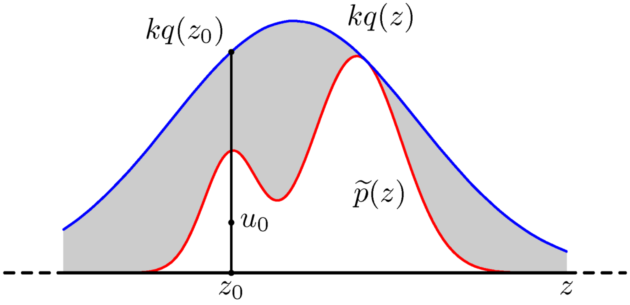

Figure: Samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\tilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\tilde{p}(z)$.

Rejection Sampling (4)

Each step of rejection sampling involves generating two random numbers.

1. Generate a number z_0 from the distribution $q(z)$.
2. Generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$. This pair of random numbers has uniform distribution under the curve of the function $kq(z)$.
3. If $u_0 > \tilde{p}(z_0)$ then the sample is rejected (lies in the grey area), otherwise u_0 is retained.
4. the remaining pairs then have uniform distribution under the curve of $\tilde{p}(z)$, and hence the corresponding z values are distributed according to $p(z)$, as desired.

Rejection Sampling - Example

Consider the task of sampling from a Gamma distribution

$$\mathrm{Gam}(z|a,b) = \frac{b^a z^{a-1} \exp(-bz)}{\Gamma(a)}$$

which for $a > 1$ has a bell-shaped form. A suitable proposal distribution is the Cauchy distribution because it too is bell-shaped and because we can use the transformation method, to sample from it.
The constant k must be selected as small as possible while still satisfying the requirement $kq(z) \geq \tilde{p}(z)$.

Rejection Sampling - Example

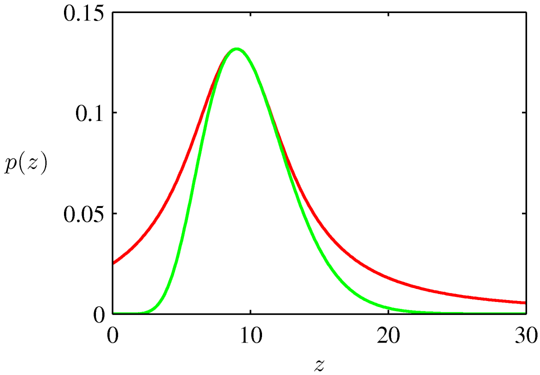

Figure: Gamma distribution (green), scaled Cauchy distribution (red). Samples from the Gamma distribution can be obtained by sampling from Cauchy and then applying the rejection sampling criteria.

Issues

- In many instances where we might wish to apply rejection sampling, it proves difficult to determine a suitable analytic form for the envelope distribution $q(z)$.
- An alternative approach is to construct the envelope function on the fly based on measured values of the distribution $p(z)$.
- This idea results in the *Adaptive Rejection Sampling* algorithm.

Importance Sampling (1)

- As mentioned before, one of the main reason for being able to sample from complicated probability distributions is to be able to evaluate expectations.
- Importance Sampling provides a method for approximating expectations directly but does not itself provide a mechanism for drawing samples from distribution $p(z)$.

Importance Sampling (2)

- Approximation of the expectation depends on being able to draw samples from the distribution $p(z)$.

$$\mathbb{E}[f] = \int p(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}.$$

- Assumption: It is impractical to sample from $p(\mathbf{z})$ and we only can evaluate $p(\mathbf{z})$ easily for any value of $\mathbf{z}$.

Importance Sampling (3)

- One simplistic strategy is to discretize $\mathbf{z}$-space into a uniform grid and to evaluate the integrand as a sum of the form

$$\mathbb{E}[f] \approx \sum_{l=1}^{L} p(\mathbf{z}^{(l)}) f(\mathbf{z}^{(l)}).$$

- An obvious problem is that the number of terms in the summation grows exponentially with the dimensionality of $\mathbf{z}$. Secondly, in high-dimensional problems, only a very small proportion of the samples will make significant contributions to the sum. And we prefer to choose samples to fall in regions where $p(\mathbf{z})$ is large.

Importance Sampling (4)

- Similar to rejection sampling, importance sampling is based on the use of a proposal distribution $q(\mathbf{z})$ from which it is easy to draw samples.
- The expectation then can be expressed in the form of a finite sum over samples $\{\mathbf{z}^{(l)}\}$ drawn from $q(\mathbf{z})$.

$$
\begin{aligned}
\mathbb{E}[f] &= \int p(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \\
&= \int \frac{p(\mathbf{z})}{q(\mathbf{z})} f(\mathbf{z}) q(\mathbf{z}) d\mathbf{z} \\
&\approx \frac{1}{L} \sum_{l=1}^{L} \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})} f(\mathbf{z}^{(l)})
\end{aligned}
$$

Importance Sampling (5)

- The quantities $r_l = \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}$ are known as *importance weights*, and they correct the bias introduced by sampling from the wrong distribution.
- Unlike rejection sampling, all of the generated samples are retained.

Importance Sampling - Example

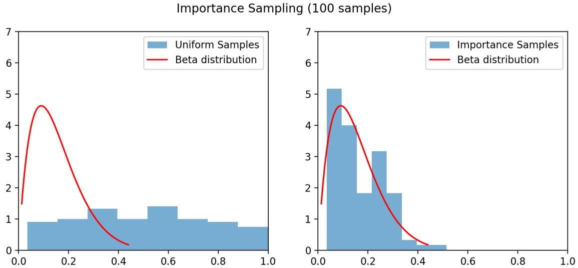

Figure: Sampling from a Beta distribution (red) with a uniform distribution as proposal. Number of samples: 100

Importance Sampling - Example

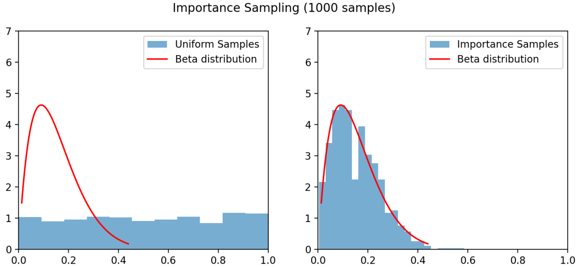

Figure: Sampling from a Beta distribution (red) with a uniform distribution as proposal. Number of samples: 10,000

Importance Sampling - Example

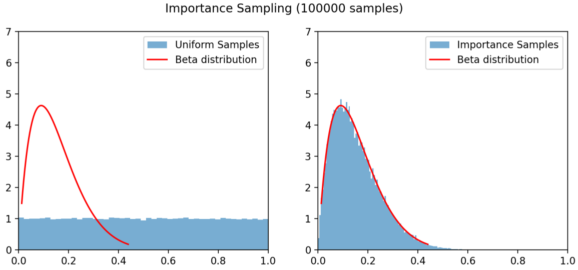

Figure: Sampling from a Beta distribution (red) with a uniform distribution as proposal. Number of samples: 100,000

Importance Sampling - Example

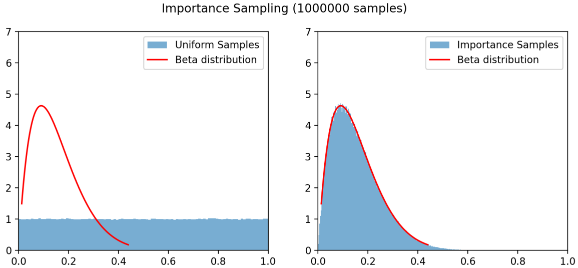

Figure: Sampling from a Beta distribution (red) with a uniform distribution as proposal. Number of samples: 1,000,000