Sampling From a Probability Distribution

Apurva Nakade

2025-02-01

Table of contents

In this module, we will discuss how to sample from a probability distribution. Sampling from a probability distribution is a fundamental problem in statistics and machine learning. It is used in various applications like Monte Carlo methods, Bayesian inference, and reinforcement learning.

We'll use this week to review many of the standard probability distributions and how to sample from them. We'll also discuss the concept of the cumulative distribution function (CDF) and how it can be used to sample from a probability distribution using the inverse transform sampling method.

From now on, we'll assume that we have a reliable way to generate random numbers from the uniform distribution $U(0,1)$.

To understand what it means to sample from a probability distribution, let's consider a simple example. Let X be a discrete random variable that takes values $1, 2, \dots, n$ with probabilities $p_1, p_2, \dots, p_n$. To sample from this distribution, we want to generate a random variable X such that $\mathbb{P}(X = i) = p_i$ for all $i = 1, 2, \dots, n$ i.e. we want to select a random integer i with probability p_i. If we generate enough samples $x_1, x_2, \dots, x_N$ from this distribution, then the fraction of samples that are equal to i will be approximately equal to p_i for large N.

Discrete Case

Let's consider a simple example where X is a discrete random variable that takes values $1, 2, 3$ with probabilities $0.2, 0.3, 0.5$ respectively. To sample from this distribution, we can use the following algorithm:

1. Generate a random number u from the uniform distribution $U(0, 1)$.
2. If $u \leq 0.2$, set $X = 1$.
3. If $0.2 < u \leq 0.5$, set $X = 2$.
4. If $0.5 < u \leq 1$, set $X = 3$.
5. Return X.

This algorithm can be easily extended to the case where X takes n values. This algorithm can be interpreted as a special case of the inverse transform sampling method described below.

Binomial Distribution

Let X be a random variable with binomial distribution with parameters n and p. The probability mass function of X is given by

$$\text{Binomial}(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

One way to sample from the binomial distribution is to treat it as a discrete distribution and use the above algorithm. However, we can use the Bernoulli distribution to sample from the binomial distribution.

If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with Bernoulli distribution with parameter p, then the random variable $X = Y_1 + Y_2 + \ldots + Y_n$ has binomial distribution with parameters n and p. This gives us a simple algorithm to sample from the binomial distribution:

1. Generate n random numbers $u_1, u_2, \ldots, u_n$ from the uniform distribution $U(0, 1)$.
2. Set $Y_i = 1$ if $u_i \leq p$ and $Y_i = 0$ otherwise for $i = 1, 2, \ldots, n$.
3. Compute $X = Y_1 + Y_2 + \ldots + Y_n$.
4. Return X.

Inverse Transform Sampling

Consider a continuous random variable X with probability density function $f(x)$. Let U be a random variable with uniform distribution $U(0,1)$.

Theorem 0.1. ***(Inverse Transform Sampling)****: Let $F(x)$ be the cumulative distribution function of X. If $F(x)$ is strictly increasing and continuous, then the random variable $Y = F^{-1}(U)$ has the same distribution as X.*

Proof. Let $F(x)$ be the cumulative distribution function of X. The cumulative distribution function of U is given by

$$\mathbb{P}(U \leq u) = u.$$

Since $F(x)$ is strictly increasing and continuous, it has an inverse $F^{-1}(u)$.

The cumulative distribution function of $Y = F^{-1}(U)$ is given by

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

Thus, Y has the same distribution as X. ■

The inverse transform sampling method can be used to sample from any probability distribution X for which we can compute the cumulative distribution function $F(x)$ and its inverse $F^{-1}(u)$. The algorithm to sample from a probability distribution using the inverse transform sampling method is as follows:

1. Generate a random number u from the uniform distribution $U(0,1)$.
2. Compute $x = F^{-1}(u)$.
3. Return x.

Even when $F(x)$ is not strictly increasing, we can still use the inverse transform sampling method by using the generalized inverse of $F(x)$. The generalized inverse of $F(x)$ is defined as

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

You can think of inf as min for simplicity.

□

Exponential Distribution

Let X be a random variable with exponential distribution with rate parameter $\lambda > 0$. The probability density function of X is given by

$$\text{Exp}(\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

The cumulative distribution function of X is given by

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

The inverse of the cumulative distribution function is given by

$$F^{-1}(u) = -\frac{1}{\lambda} \log(1 - u).$$

Hence, the inverse transform sampling method can be used to sample from the exponential distribution.

Weibull Distribution

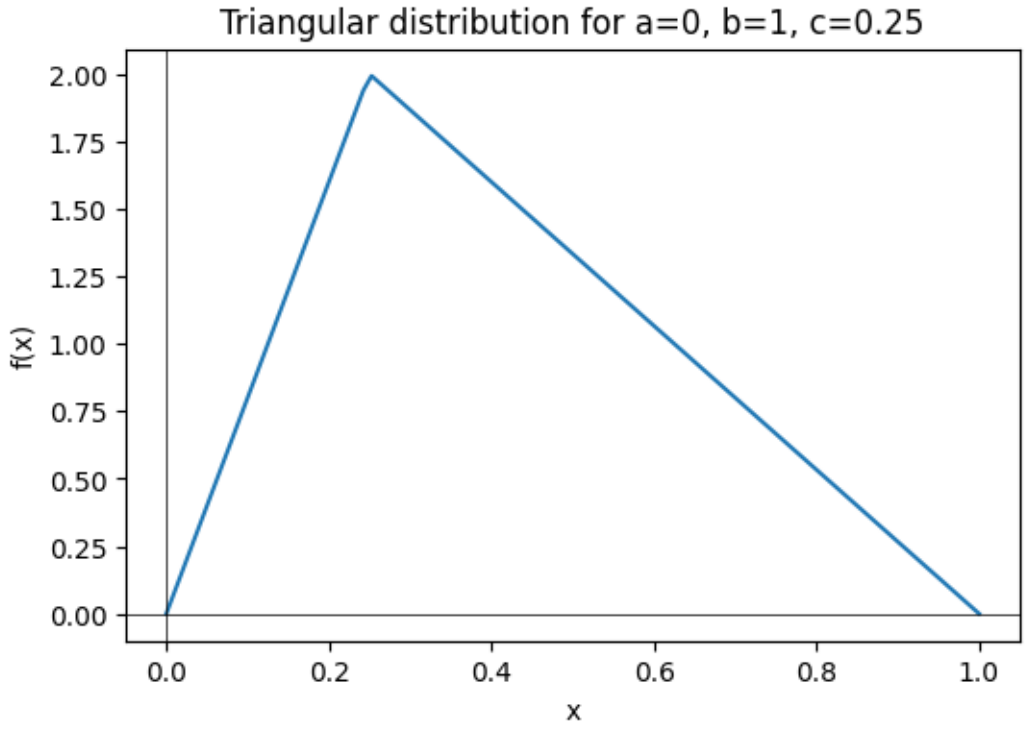
The Weibull distribution is a generalization of the exponential distribution. Let X be a random variable with Weibull distribution with shape parameter $k > 0$ and scale parameter $\lambda > 0$. The probability density function of X is given by

$$f(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0.$$

The inverse transform sampling method can be used to sample from the Weibull distribution.

Triangular Distribution

Let X be a random variable with triangular distribution supported over the interval $[a, b]$ with maximum value at $c \in [a, b]$. We can use the inverse transform sampling method to sample from the triangular distribution.



Normal Distribution

The standard normal distribution with mean 0 and variance 1 is given has the probability density function

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It is not possible to sample from the normal distribution using the inverse transform sampling method because the cumulative distribution function of the normal distribution does not have a closed-form inverse. However, there are other methods to sample from the normal distribution. One such method is the Box-Muller transform. The Box-Muller transform is based on the

following idea. Consider a 2D random variable (X, Y) with standard normal distribution. The joint probability density function of (X, Y) is given by

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \quad -\infty < x, y < \infty.$$

Let $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arctan(Y/X)$ be the polar coordinates of (X, Y). Then notice that θ is uniformly distributed in $[0, 2\pi]$. We can calculate the cdf of R as follows:

$$\begin{aligned}
\mathbb{P}(R \le r) &= \mathbb{P}(X^2 + Y^2 \le r^2) \\
&= \int_{x^2+y^2 \le r^2} f(x, y)\, dx\, dy \\
&= \int_{x^2+y^2 \le r^2} e^{-(x^2+y^2)/2}\, dx\, dy \\
&= \int_0^{2\pi} \int_0^r \frac{1}{2\pi} e^{-r^2/2} r\, dr\, d\theta \\
&= 1 - e^{-r^2/2}.
\end{aligned}$$

We can invert this to get the inverse cdf of R:

$$F_R^{-1}(u) = \sqrt{-2\log(1-u)}.$$

We can summarize the above discussion in the following theorem.

Theorem 0.2. ***(Box-Muller Transform)****: Let U_1, U_2 be independent random variables with uniform distribution $U(0, 1)$. Let $R = \sqrt{-2\log U_1}$ and $\Theta = 2\pi U_2$. Then, the random variables $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$ are independent and have standard normal distribution.*

Note that we are using U_1 instead of $1 - U_1$ in the formula for R. This is because $1 - U_1$ is also uniformly distributed in $[0, 1]$.

This gives us the following algorithm to sample from the normal distribution:

1. Generate two random numbers u_1, u_2 from the uniform distribution $U(0, 1)$.
2. Compute $R = \sqrt{-2\log u_1}$ and $\Theta = 2\pi u_2$.
3. Compute $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$.
4. Return X (or Y).

Poisson Distribution

The Poisson distribution with parameter $\lambda > 0$ is a discrete distribution that models the number of events occurring in a fixed interval of time or space, where λ is the average rate of events. In unit time T, the expected number of events is λT. The probability mass function of the Poisson distribution is given by

$$\text{Pois}(n) = \frac{e^{-\lambda}\lambda^n}{n!}, \quad n \in \mathbb{N}.$$

The pmf of the Poisson distribution measures the probability of observing n events in time T.

Relation between Poisson and Binomial Distribution

The Poisson distribution can be approximated by the binomial distribution when the number of trials n is large and the probability of success p is small. Let X be a random variable with binomial distribution with parameters n and p. As $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains constant, the pmf of X converges to the pmf of the Poisson distribution with parameter λ. This gives us a simple algorithm to sample from the Poisson distribution:

1. Set $X = 0$.
2. Choose n be a large integer (something like $n > 10\lambda$).
3. Set $p = \lambda/n$.
4. Generate a X according to the binomial distribution with parameters n and p.

This is a fast method to sample from the Binomial approximation to the Poisson distribution and is good when λ is small. For large λ, the method described below is more efficient as the number of trials n required for the binomial distribution to approximate the Poisson distribution becomes very large.

Relation between Poisson and Exponential Distribution

We exploit the relation between the Poisson distribution and the exponential distribution to sample from the Poisson distribution. When events occur at a constant rate λ, the time between events follows an exponential distribution with rate parameter λ. More precisely,

Theorem 0.3. ***(Inter-arrival Times)****:*

Let $X_1, X_2, ...$ be independent random variables with exponential distribution with rate parameter λ. Define

$$N = \max\{n : X_1 + X_2 + \cdots + X_n \leq 1\}.$$

Then, N has Poisson distribution with parameter λ.

The proof of this requires us to understand the relation between exponential and gamma distributions. We will skip the proof for now. The Poisson-Exponential connection gives us a simple algorithm to sample from the Poisson distribution:

1. Set $S = 0$ and $N = 0$.
2. While True:
    1. Generate a random number $x \sim \text{Exp}(\lambda)$.
    2. Set $S = S + x$.
    3. If $S > 1$, return N.
    4. Else, set $N = N + 1$.

Beta Distribution

The Beta distribution is a continuous distribution defined on the interval $[0, 1]$. Let X be a random variable with Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. The probability density function of X is given by

$$f(x) = cx^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1,$$

where $c = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$ is the normalizing constant. (Note that when α and β are not integers, we use Γ function to as a generalization of the factorial function.)

This is a simple function supported over the interval $[0, 1]$. The Beta distribution is used as a prior distribution in Bayesian statistics. When α and β are non-zero, the cdf of the Beta distribution does not have a closed-form expression. However, we can use the inverse transform sampling method to sample from the Beta distribution. Instead, we can use the Beta-Order Statistics connection to sample from the Beta distribution.

Definition 0.1. (Order Statistics): Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables. The k-th order statistic, denoted $X_{(k)}$, is the k-th smallest value among $X_1, X_2, \ldots, X_n$.

Theorem 0.4. ***(Beta-Order Statistics)****: Let $U_1, U_2, \ldots, U_n$ be independent random variables with uniform distribution $U(0, 1)$. Then the random variable $X = U_{(k)}$ has Beta distribution with parameters $\alpha = k$ and $\beta = n - k + 1$.*

Proof. We'll work out a partial proof of the theorem. Let $X = U_{(k)}$. The cumulative distribution function of X is given by

$$\begin{aligned} F(x) &= \mathbb{P}(U_{(k)} \leq x) \\ &= \mathbb{P}(\text{at least } k \text{ variables among } U_1, U_2, \ldots, U_n \text{ are less than } x) \\ &= \sum_{i=k}^{n} \binom{n}{i} x^i (1-x)^{n-i}. \end{aligned}$$

We differentiate both sides to get the probability density function of X:

$$\begin{aligned} f(x) &= \frac{d}{dx} F(x) \\ &= \sum_{i=k}^{n} \binom{n}{i} \frac{d}{dx} x^i (1-x)^{n-i} \\ &= \sum_{i=k}^{n} \binom{n}{i} i x^{i-1} (1-x)^{n-i} - \binom{n}{i} (n-i) x^i (1-x)^{n-i-1}. \end{aligned}$$

The rest of the proof involves checking that the higher terms in the alternating sum cancel out and only the first term with $i = k$ remains. ■

□

This theorem gives us a simple algorithm to sample from the Beta distribution:

1. Generate n random numbers $u_1, u_2, \ldots, u_n$ from the uniform distribution $U(0, 1)$.
2. Sort the numbers in increasing order $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$.
3. Return $u_{(k)}$.

Mixture Distributions

A mixture distribution is a probability distribution that is formed by taking a weighted sum of two or more probability distributions. Let X be a random variable that is a mixture of distributions $f_1(x), f_2(x), \ldots, f_n(x)$ with weights $w_1, w_2, \ldots, w_n$. The probability density function of X is given by

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \ldots + w_n f_n(x).$$

Mixture distributions are used to model complex distributions that cannot be modeled by a single distribution. We can sample from a mixture distribution by sampling from the component distributions and then taking a weighted sum of the samples.

$$X = Y_1 \text{ with probability } w_1, \quad X = Y_2 \text{ with probability } w_2, \quad \dots, \quad X = Y_n \text{ with probability } w_n.$$

To sample from a mixture distribution, we can use the following algorithm:

1. Sample from the discrete distribution $[w_1, w_2, \dots, w_n]$ to select a component distribution.
2. Sample from the selected component distribution.
3. Return the sample.

Note that this is not the same as constructing a linear combination of the component distributions. For example, if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then $w_1 X_1 + w_2 X_2$ is is a normal distribution with mean $w_1\mu_1 + w_2\mu_2$ and variance $w_1^2\sigma_1^2 + w_2^2\sigma_2^2$. This is not the same as a mixture of two normal distributions.