Sampling methods

Why, when and how do we use them?

Ha Vu

06/14/2023

Contents

1 Why do we need sampling methods?

1.1 Easy-to-sample distribution

There are certain probabilities distributions from which it is very easy to sample. This is an answer provided by ChatGPT to the question **"What are some distributions that we know how to sample from?"** *Hint: just skim through, no need to understand.*

There are several common probability distributions for which we have known methods to sample from. Some of these distributions include:

- **Uniform distribution**: A continuous distribution where all values within a specified range are equally likely. Sampling from a uniform distribution can be done using algorithms like the inverse transform method or using random number generators.
- **Normal (Gaussian) distribution**: A continuous distribution characterized by a symmetric bell-shaped curve. Sampling from a normal distribution can be achieved using methods such as the Box-Muller transform or the Marsaglia polar method.
- **Exponential distribution**: A continuous distribution commonly used to model the time between events in a Poisson process. Sampling from an exponential distribution can be done using the inverse transform method or other techniques like the acceptance-rejection method.
- **Poisson distribution**: A discrete distribution that represents the number of events occurring within a fixed interval of time or space. Sampling from a Poisson distribution can be accomplished using algorithms such as the Knuth algorithm or the Poisson thinning method.

- **Binomial distribution**: A discrete distribution that models the number of successes in a fixed number of independent Bernoulli trials. Sampling from a binomial distribution can be done using methods like the inverse transform method or using built-in functions available in libraries like NumPy or SciPy.

- **Gamma distribution**: A continuous distribution often used to model positive-valued variables with a skewed distribution. Sampling from a gamma distribution can be achieved using techniques like the Marsaglia-Tsang method or the Cheng method.

These are just a few examples of distributions from which we know how to sample. There are many more distributions with known sampling methods, including the beta distribution, chi-squared distribution, geometric distribution, and more. The choice of distribution and sampling method depends on the specific problem and the characteristics of the data being modeled.

1.2 Why do we need sampling methods?

Off the top of my head, I can only sample from the Uniform (pick a number $[0, 1]$) and the Bernoulli distributions (toss a coin, or pick a number $[0, 1]$ and decide whether or not it is below/above a threshold p). In order to sample from other distributions (for example: exponential, normal, etc.) , we may need sampling methods. There are also other cases where we need to use sampling methods, simply because there's no official way to sample from the distributions that appear in our analyses.

- Sample from a distribution such that we know the inverse of the CDF $\longrightarrow$ **inverse CDF transform**.

- Estimate a certain function of variables from a distribution, i.e. $P(X^2 + Y^2 <= 1)$, given X and $Y \longrightarrow$ **Monte Carlo, importance sampling**.
- Estimate the posterior probability when the chose prior does not gives an analytical solution, i.e. if $X \sim Unif(0, 1)$, $Y|X \sim Bern(X)$, what is $P(Y|X)$? $\longrightarrow$ **Monte Carlo**
- Sample from a distribution $\pi(X)$ with no closed-form for the inverse of CDF, and for which we only know the value $\pi(X = x)$ for a value of x.$\longrightarrow$ **rejection sampling**
- sample from a multi-dim variable distribution $\longrightarrow$ **MCMC**

2 CDF transformation

2.1 Algorithm through example

Say we want to sample $X \sim f(X)$, such that we can derive a closed-form of the CDF, i.e. $F^{-1}(X)$ is a closed-form, with $F(x) = P(X <= x)$ where $X \sim f(X)$. Example: $X \sim Exponential(x, \lambda)$

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Then, the CDF is as follows: $F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

Can we find the inverse of F? Yes Say $y = F(x) = 1 - e^{-\lambda x}$, we have to swap x and y and solve for y, we have $x = 1 - e^{-\lambda y} \longrightarrow y = \frac{ln(1-x)}{-\lambda}$. So, $F^{-1}(y, \lambda) = \frac{ln(1-y)}{-\lambda}$.

Algorithm:

Algorithm 1: SampleExponential

Input : $\mathrm{U}_1, ..., U_n \overset{\text{iid}}{\sim} Uniform(0, 1)$.

Output: $\mathrm{X}_1, ..., X_n \overset{\text{iid}}{\sim} Exponential(\lambda)$.

1 **for** $i \leftarrow 1$ **to** n **do**

2 | Sample$U_i \sim \mathrm{Unif}(0, 1)$

3 | $X_i = \frac{ln(1-U_i)}{-\lambda}$

4 **end**

5 **return** $X_1, ..., X_n$

2.2 Proof: why it works?

$P(X <= x) = F(x) = u \Longrightarrow x = F^{-1}(u)$

3 Monte Carlo sampling

3.1 When is MC useful?

MC is also known as direct (naive) sampling. Useful when:

- **Bayesian inference**: calculate the posterior probability if we choose a prior such that the posterior doesn't have a nice distribution function. (**Note**: When do we have a nice form for the posterior distribution? $\longrightarrow$ Usually only when we have conjugate prior, which means the prior and posterior distrubtions belong to the same family (Ask Chat-GPT: "Examples of conjugate prior" and it will give: Beta-Binomial, Normal-Normal, Gamma-Exponential, Dirichlet-Multinomial).

  If $\theta \sim Unif(0, 1)$, and $P(X_i|\theta) = Bernouli(\theta)$, what is $P(\theta|X_1, ..., X_n)$?

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)}{P(\mathbf{X})}$$
$$= \frac{\prod_{i=1}^{n} P(X_i|\theta)P(\theta)}{\int_{\theta=0}^{1} \prod_{i=1}^{n} P(X_i|\theta)P(\theta)\, d\theta}$$

We can use MC and plug in multiple values of θ to estimate the integral in the denominator.

- **Calculate** $g(X)$: If $X \sim f(X)$, and we want to know $E(g(X))$ for some function g. It's the same as estimating the population mean by taking the sample mean.
- **When there is no analytical form for the distribution** P: $P(X^2 + Y^2 <= 1)$, given that X and $Y \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$.

4 Rejection sampling

Rejection sampling is most useful when we have a distribution $\pi(X)$ that is closed-form, i.e. we can easily query the value of $\pi(X = x)$ for each value of x BUT we cannot find a closed-form of the inverse of CDF of π (hence cannot use the CDF transform method to sample from π).

4.1 Algorithm

Suppose we know how to sample from a distribution $g(X)$ (For example, I only know how to sample using pen and pencils and coins from $Unif(a, b)$ and $Bernouli(\theta)$ distributions). We would like to generate samples from $g(X)$, and filter out samples that we think are not from $\pi(X)$. In particular, we would like to know $\pi(X)$ and a constant c such that

$\pi(X) <= c * g(X) \forall X \in \mathbf{R}$. Details as follows:

Algorithm 2: SampleRejection

Input : $\pi(X)$, $g(X)$

Output: $X_1, ..., X_n \overset{\text{iid}}{\sim} \pi(X)$.

1 **for** $i \leftarrow 1$ **to** n **do**

2 Sample $X_i \sim$ g(X)

3 Sample $U_i \sim$ Unif$(0, 1)$

4 If $U_i <= \frac{\pi(X)}{c*g(X)}$, then accept X_i, else resample X_i from line 2

5 **end**

6 **return** $X_1, ..., X_n$

4.2 Why does it work?

$$\text{Let } z = \begin{cases} 1 & \text{if } X \sim g(x) \text{ is accepted} \\ 0 & \text{otherwise} \end{cases}$$

Then $Z \sim Bernouli(\frac{\pi(X)}{c*g(X)})$

$$\begin{aligned} p(X = x \mid z = 1) &= \frac{p(X = x, z = 1)}{p(z = 1)} \\ &= \frac{p(z = 1 \mid X = x) \cdot g(x)}{\int p(z = 1 \mid X = x) \cdot g(x) dx} \\ &= \frac{\frac{\pi(x)}{c \cdot g(x)} \cdot g(x)}{\int \frac{\pi(x)}{c \cdot g(x)} \cdot g(x) dx} \\ &= \pi(x) \end{aligned}$$

The denominator transformation: $\int \frac{\pi(x)}{c} dx = \frac{1}{c}$

4.3 How to find the optimal $g(X)$?

See my hand-written note of a particular example, when $\pi(X) \propto \sigma(X) * I(x > c)$ for some constant c.

5 Importance sampling

Importance sampling is most helpful where we want to calculate $E[h(X)]$, where $X \sim \pi(X)$, from which we do not know, primitively, how to sample. Similarly to rejection sampling, we would like to sample $X \sim g(X)$ where g is a distribution that we DO know how to sample from.

Rejection and importance sampling serve different purposes. The purpose of rejection sampling is to obtain samples directly from the target distribution $\pi(X)$, while the purpose of importance sampling is to sample from a different distribution g in order to calculate the expectation of $h(X)$ where $X \sim \pi(X)$. **Importance sampling is more efficient than rejection sampling because it doesn't throw away points.**

5.1 Algorithm

1. Draw m points $x_1, \ldots, x_m$ from a "trial" distribution g.
2. Calculate the importance weight $w_i = \frac{\pi(x_i)}{g(x_i)}$
3. Estimate $\mathrm{E}_\pi[h(x)] = \frac{\sum_{i=1}^n h(x_i)w_i}{\sum_{i=1}^n w_i}$

5.2 Why does it work?

First, we can prove that $\frac{\sum_{i=1}^n h(x_i)w_i}{n}$ is a unbiased estimator of $\mathrm{E}_\pi[h(x)]$. See my hand-written note for this proof. The other estimator, $\frac{\sum_{i=1}^n h(x_i)w_i}{\sum_{i=1}^n w_i}$ is biased, but it reduces the variance of

the estimator, I still have not been able to prove that!

6 MCMC sampling

6.1 Markov Chain

$X_1, X_2, ..., X_n$ constitute a Markov Chain when $P\left(X_{n+1} \mid X_1 \dots X_n\right) = P\left(X_{n+1} \mid X_n\right)$

Transition probability is defined as: $T(a,b) = P\left(X_{n+1} = a \mid X_n = b\right)$

Our goal: Find a transition probability such that the resulting Markov chain still follows a target invariant distribution $\pi(X)$. **Goal of Markov Chain theory**: given $T(x,y)$, what is $\pi(x)$?. **Goal of MCMC**: given π, what is $T(x,y)$?

6.2 Reversibility Criterion

The Reversibility Criterion holds when:

$$P(X_i = a, X_{i+1} = b) = P(X_i = b, X_{i+1} = a)$$

$$P\left(X_{i+1} = b \mid X_i = a\right) \cdot P\left(X_i = a\right) = P\left(X_{i+1} = a \mid X_i = b\right) \cdot P\left(X_i = b\right)$$

$$T(b,a) \cdot \pi(a) = T(a,b) \cdot \pi(b)$$

Proof: Reversibility $\longrightarrow$ equilibrium Due to reversibility, the first equal will hold

$$\sum_a P\left(X_{i+1} = b \mid X_i = a\right) \cdot P\left(X_i = a\right) = \sum_a P\left(X_{i+1} = a \mid X_i = b\right) \cdot P\left(X_i = b\right)$$

$$\begin{aligned} \sum_a T(b,a) \cdot \pi(a) &= \sum_a T(a,b) \cdot \pi(b) \\ &= \pi(b) \cdot \sum_a T(a,b) \\ &= \pi(b) \end{aligned}$$

6.3 Markov Chain Monte Carlo

Given π as target distribution ($\forall x \in \mathbf{R}, \pi(x)$ is known), and *any* conditional distribution of our design $T(x_{n+1}, x_n)$, we would like to design a factor $\alpha(x_{n+1}, x_n)$ that will ensure reversibility. In particular, as an example, we design

$f(x_{n+1}|x_n) = \Phi(x_{n+1} - x_n) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_{n+1}-x_n)^2}{2}}$

To ensure reversibility, we will need $T(x_{n+1}, x_n) = f(x_{n+1}|x_n) \cdot \alpha(x_{n+1}, x_n)$, and $\alpha(x_{n+1}, x_n)$ is designed to satisfy the following:

$\pi(x_{n+1}) f(x_{n+1} \mid x_n) \alpha(x_{n+1}, x_n) = \pi(x_n) f(x_n \mid x_{n+1}) \alpha(x_n, x_{n+1})$

because both quantities are equal to $\pi(x_n)\pi(x_{n+1})$. If $\pi(x_{n+1}) f(x_{n+1} \mid x_n) > \pi(x_n) f(x_n \mid x_{n+1})$, then we would want $\alpha(x_{n+1}, x_n) = \frac{\pi(x_n) f(x_n|x_{n+1})}{\pi(x_{n+1}) f(x_{n+1}|x_n}$, and $\alpha(x_n, x_{n+1}) = 1$. More generally,

$$\alpha(x_{n+1}, x_n) = min\left(\frac{\pi(x_n) f(x_n \mid x_{n+1})}{\pi(x_{n+1}) f(x_{n+1} \mid x_n}, 1\right)$$

$$\alpha(x_n, x_{n+1}) = min\left(\frac{\pi(x_{n+1}) f(x_{n+1} \mid x_n)}{\pi(x_n) f(x_n \mid x_{n+1}}, 1\right)$$

6.4 Metropolis-Hastings Algorithm

In previous section, we learned about principals of designing the transition probability $T(x_{n+1}, x_n) = f(x_{n+1}|x_n) \cdot \alpha(x_{n+1}, x_n)$. Here, this algorithm demonstrates how to actually implement this principle into the sampling procedure: Given a random value of $x1$, we will generate $x2$ as follows:

1. Sample $\hat{x}2$ from $f(x2|x1)$, given that we already picked the form of f.

2. Calculate $\alpha(\hat{x2}, x1)$ (using formula from previous section), and accept $\hat{x2}$ as $x2$ as a Bernoulli process with parameter $\alpha(\hat{x2}, x1)$. If $x2$ was not successfully set as $\hat{x2}$, set $x2 = x1$, and continue the sampling process.

6.5 Gibbs Sampler

Gibbs sampling is used when we want to sample from a multi-dimensional distribution, that is usually hard to directly sample from. Without loss of generality, suppose $p(x, y)$ is the target distribution (2D variable). Gibbs sampling will try to sample x and y in a chain given that we can more easily derive the sampling procedure from $p(x|y)$ and $p(y|x)$. The procedure is as follows:

Algorithm 3: Sample Gibbs

Input :

Output: $(X_1, Y_1), ..., (X_n, Y_n) \sim P(X, Y)$.

Initialize (x_0, y_0) to some values
for $i \leftarrow 1$ **to** n **do**
    Sample $X_i \sim P(X \mid Y_{i-1})$
    Sample $Y_i \sim P(Y \mid X_i)$
end
return $(X_1, Y_1), ..., (X_n, Y_n)$

To generate the algorithm to higher dimension, Gibbs sampling also tries to sample one dimension at a time, i.e. sample $X_i \sim P(X_i \mid X^{-i})$, where X^{-i} denotes variables from $\mathbf{X}$ excluding the i-th dimension.

How is it different from Metropolis-Hastings?: Gibbs sampling assumes direct simulation from full conditionals. If full conditional simulation is not possible, then we could use Metropolis-Hastings. **Examples**:

- Sample from $P(x, y) \propto e^{-xy}\mathbf{1}(x, y \in (0, c))$ for some constant $c > 0$: See hand-written note.

- Sample from $\mathbf{X} \sim \text{Normal}(\mu, \Sigma)$: We really focus on the proof that the condition distribution on one variable from multivariate normal is also Normal.