Unit 1: Introduction, Rejection and Importance Sampling

Sampling methods can be used for a wide range of tasks in statistics, machine learning and data science. Typical applications include:

- Given a real-valued function $f \geq 0$, generate $x_1, \dots, x_n$ from the probability distribution with density function $Ce^{-f(x)}$, where C is the unknown normalizing constant.
- Find $\arg\min_x f(x)$.
- Approximate an (often high-dimensional) integral $\int H(x)\nu(\mathrm{d}x)$ for some real-valued function H and measure ν.
- Given samples $y_1, \dots, y_n$ from some unknown probability distribution π, generate a new sample x from π.

The first task will be the primary focus of this course, but we aim to cover sampling methods used for all the four tasks. We will prioritize methodology and computation over theory.

1.1 Monte Carlo Integration

Consider the integral $\theta \coloneqq \int_{\mathcal{X}} H(x)\pi(\mathrm{d}x)$ where π is a probability distribution defined on the space $\mathcal{X}$, and $H \colon \mathcal{X} \to \mathbb{R}$. We will always assume that θ is well-defined and finite. Given i.i.d. random variables $X_1, \dots, X_n$ drawn from π, we can approximate this integral by

$$\hat{\theta}_n = \frac{H(X_1) + \cdots + H(X_n)}{n}.$$

Assume $\int_{\mathcal{X}} H^2(x)\pi(\mathrm{d}x) < \infty$. Then, $H(X_1), \dots, H(X_n)$ are i.i.d. random variables with mean θ and finite variance. Hence, by the Law of Large Numbers, $\hat{\theta}_n$ converges to θ almost surely. By the Central Limit Theorem, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a normal random variable, so the error of the estimator $\hat{\theta}_n$ has order $O(n^{-1/2})$. Such a sampling scheme for integral approximation can be generalized by considering correlated random variables $X_1, X_2, \dots$ (e.g. in Markov chain Monte Carlo), and under some additional conditions CLT continues to hold.

Example 1.1. Consider this integral

$$\theta = \int_{\mathbb{R}} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x,$$

which is just the mean of a standard normal random variable and equals 0. Now let's generate a sequence of standard normal random variables as follows. Draw $X_1 \sim N(0,1)$, and for each $i \geq 2$, we let $X_i = \lambda X_{i-1} + Z_i$, where $\lambda \in (-1,1)$ is a constant, and $Z_1, Z_2, \dots$ are i.i.d. (and independent of $X_1, X_2, \dots$) with distribution $N(0, 1-\lambda^2)$. Clearly, every X_i follows $N(0,1)$. Define our estimator by $\hat{\theta}_n = (X_1 + \cdots X_n)/n$. Then, $\hat{\theta}_n$ is also normally distributed. For every fixed n, the variance of $\hat{\theta}_n$ decreases as λ decreases. In particular, when $\lambda < 0$, the estimator is more efficient than the average of i.i.d. samples from $N(0,1)$. For every fixed λ, as n goes to infinity, one can show that the variance of $\hat{\theta}_n$ is approximately $\frac{1+\lambda}{n(1-\lambda)}$.

Example 1.2. Let $V(p)$ denote the volume of the simplex

$$S_p = \{x = (x_1, \dots, x_p) \in \mathbb{R}^p \colon x_1 + \cdots + x_p \leq 1,\ x_i \geq 0 \text{ for each } i\}. \tag{1}$$

One can show that $V(p) = 1/p!$. Here is a very simple, though quite inefficient, method for numerically calculating $V(p)$. Observe that we can express $V(p)$ by

$$V(p) = \int_{[0,1]^p} \mathbb{1}_{S_p}(x)\mathrm{d}x = \mathbb{E}[\mathbb{1}_{S_p}(X)]$$

where $X \in \mathbb{R}^p$ follows a uniform distribution on $[0,1]^p$. Hence, by generating n random samples from $\mathrm{Unif}([0,1]^p)$, we can unbiasedly estimate $V(p)$; denote this estimator by $\hat{V}_n$, omitting the dependence on p. The standard deviation of $\hat{V}_n$ is $\{V(1-V)/n\}^{1/2}$.

1.2 Rejection Sampling

Consider Example 1.2 again. How to generate a random sample from the uniform distribution on S_p? The same idea applies. We draw X from $\mathrm{Unif}([0,1]^p)$ and discard it if $X \notin S_p$. This method is known as rejection sampling (other names include "accept-reject method", "acceptance-rejection sampling"); a general formulation is given in Theorem 1.1. When p is large, using rejection sampling to generate observations from $\mathrm{Unif}(S_p)$ can be a very bad idea, and in Exercise 1.1, we recall a simple method for direct sampling from $\mathrm{Unif}(S_p)$ [1].

Theorem 1.1 (Rejection Sampling). *Let f, g be two probability density functions with respect to a dominating measure μ on the space $\mathcal{X}$. Let M be a finite constant such that*

$$M \geq \sup_{x \in \mathcal{X}} \frac{f(x)}{g(x)}.$$

Let $Y_1, Y_2, \dots$ be i.i.d. with distribution g (for simplicity, we often refer to a density function as a distribution), and let $U_1, U_2, \dots$ be i.i.d. from $\mathrm{Unif}([0,1])$. *Define*

$$X = Y_\tau, \text{ where } \tau = \min\left\{ i \geq 1 \colon U_i \leq \frac{f(Y_i)}{Mg(Y_i)} \right\}.$$

Then X follows the distribution f.

Proof. Fix an arbitrary measurable $B \subset \mathbb{R}$. It suffices to show that $\mathbb{P}(X \in B) = \int_B f(x)\mu(\mathrm{d}x)$. Without loss of generality, consider the event $A = \{\tau = 1\}$. Then,

$$\mathbb{P}(X \in B \mid A) = \frac{\mathbb{P}(\{Y_1 \in B\} \cap A)}{\mathbb{P}(A)}.$$

For the numerator, we have

$$\begin{aligned}\mathbb{P}(\{Y_1 \in B\} \cap A) &= \int_{y \in B} \int_{u \in [0,1]} \mathbb{1}\left(u \leq \frac{f(y)}{Mg(y)} \right) \mathrm{d}u\, g(y)\mu(\mathrm{d}y) \\ &= \int_{y \in B} \frac{f(y)}{Mg(y)} g(y)\mu(\mathrm{d}y) = \frac{1}{M} \int_B f(y)\mu(\mathrm{d}y).\end{aligned}$$

Letting $B = \mathbb{R}$, we get $\mathbb{P}(A) = 1/M$. The claim then follows. □

Remark 1.1. The proof of Theorem 1.1 also reveals that the probability of acceptance is M^{-1}. This further implies that τ is a geometric random variable with success probability M^{-1}, and thus $\mathbb{E}[\tau] = M$. When M is large, this method is not very efficient. Another limitation of rejection sampling is that M may not exist even if f, g have the same support.

Exercise 1.1. Let $U_1, \dots, U_p$ be i.i.d. random variables drawn from the uniform distribution on $[0, 1]$. Denote the order statistics by $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_p$, and define $U_{(0)} = 0$. Show that $(X_1, \dots, X_p)$ follows the uniform distribution on S_n defined in (1), where $X_i = U_{(i)} - U_{(i-1)}$.

1.3 Importance Sampling

In rejection sampling, we generate samples using a reference distribution g. The same idea can be used to estimate the integral $\theta = \int_{\mathcal{X}} H(x)f(x)\mu(\mathrm{d}x)$. This technique is known as importance sampling, which essentially means a change of measure.

Theorem 1.2 (Importance Sampling). *Let f, g be two probability density functions, with respect to measure μ, and assume that $f > 0, g > 0$ everywhere. Given i.i.d. random variables $X_1, X_2, \dots$ with distribution g, we define*

$$\hat{\theta}_{g,n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i)w(X_i), \text{ where } w(x) = \frac{f(x)}{g(x)}. \tag{2}$$

Then, $\hat{\theta}_{g,n} \overset{a.s.}{\to} \theta = \int_{\mathcal{X}} H(x)f(x)\mu(\mathrm{d}x)$, provided that θ is well-defined and finite.

Proof. Observe that we can write

$$\theta = \int_{\mathcal{X}} I(x)g(x)\mu(\mathrm{d}x), \text{ where } I(x) = H(x)w(x).$$

Hence, $I(X_1), I(X_2), \dots$ are i.i.d. with finite expectation with respect to g. By the Strong Law of Large Numbers, their sample mean converges almost surely. □

Remark 1.2. We can replace $f, g > 0$ with weaker conditions. For example, the condition $\{x \colon g(x) = 0\} \subset \{x \colon f(x) = 0\}$ suffices, and the only change we need is to define $w = 0$ on the set $\{x \colon g(x) = 0\}$. This of course has no impact on implementation, since with probability one, we will not generate $X = x$ with $g(x) = 0$.

The variance of the estimator $\hat{\theta}_{g,n}$ can be calculated by

$$\mathrm{Var}(\hat{\theta}_{g,n}) = \frac{1}{n}\mathrm{Var}(I(X)) = \frac{1}{n}\left\{\int_{\mathcal{X}} \frac{H^2(x)f^2(x)}{g(x)}\mu(\mathrm{d}x) - \theta^2\right\}, \tag{3}$$

where $X \sim g$. It can be arbitrarily larger or smaller than the variance of the sample mean of n i.i.d. observations of $H(\tilde{X})$ with $\tilde{X} \sim f$; see Exercise 1.2.

In many applications, we can only evaluate f or g (or both) up to a normalizing constant, in which case the estimator defined in (2) cannot be used. This difficulty can be bypassed by using self-normalization.

Theorem 1.3 (Self-normalized Importance Sampling). *Consider the setting of Theorem 1.2. Define*

$$\tilde{\theta}_{g,n} = \frac{\sum_{i=1}^n H(X_i)w(X_i)}{\sum_{i=1}^n w(X_i)}.$$

Then, $\tilde{\theta}_{g,n} \overset{a.s.}{\to} \theta = \int_{\mathcal{X}} H(x)f(x)\mu(\mathrm{d}x)$, provided that θ is well-defined and finite.

Proof. Observe that we can write

$$\tilde{\theta}_{g,n} = \frac{\hat{\theta}_{g,n}}{n^{-1}\sum_{i=1}^n w(X_i)}.$$

Hence, by the continuous mapping theorem, it only remains to show that the denominator converges almost surely to 1. But this again follows from SLLN. □

Example 1.3. This example is from [2]. Consider

$$\theta = \int_{\mathbb{R}} xf(x)\mathrm{d}x, \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\theta)^2/2\sigma^2}.$$

So θ is just the mean of $X \sim N(\theta, \sigma^2)$. Let the reference distribution be $g_a(x) \propto f(x)^a$, parameterized by a (known as inverse temperature), and consider minimizing the variance given in (3) over $a > 0$. That is, we want to find

$$\operatorname*{arg\,min}_{a>0} \int_{\mathcal{X}} \frac{x^2 f^2(x)}{g_a(x)}\mathrm{d}x =: \operatorname*{arg\,min}_{a>0} J(a).$$

A straightforward calculation gives that for $a \in (0, 2)$,

$$J(a) = \frac{\theta^2 + (2-a)^{-1}\sigma^2}{\sqrt{a(2-a)}}.$$

Define $\gamma = \theta/\sigma$. The optimal value of a is given by $a^* = 1/2$ if $\gamma = 0$, and

$$a^* = \frac{3}{2} + \frac{1}{\gamma^2} - \frac{1}{2}\sqrt{\frac{4}{\gamma^4} + \frac{8}{\gamma^2} + 1}.$$

Note that a^* is always in $[1/2, 1)$, and it was shown in [2] that there always exists a^- (depending on γ) such that a reference distribution g_a with $a \in (a^-, 1)$ is more efficient than direct sampling from f. Intuitively, using some a slightly smaller than 1 should be advantageous, because (i) g_a still has a similar landscape to f so that samples are likely to be drawn around θ, and (ii) samples near θ have larger importance weights (since $a < 1$), which makes the importance sampling estimator more efficient.

Exercise 1.2. Consider the variance given in (3).

(a) Show that for fixed n, f, H, $\mathrm{Var}(\hat{\theta}_{g,n})$ is minimized when

$$g(x) = \frac{|H(x)|f(x)}{\int_{\mathcal{X}} |H(x)|f(x)\mu(\mathrm{d}x)},$$

provided that the denominator is greater than zero.

(b) Give an example where $\int_{\mathcal{X}} H^2(x)f(x)\mu(\mathrm{d}x) < \infty$ but $\mathrm{Var}(\hat{\theta}_{g,n}) = \infty$.

References

[1] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer Science & Business Media, 2013.

[2] Robert Gramacy, Richard Samworth, and Ruth King. Importance tempering. *Statistics and Computing*, 20(1):1–7, 2010.

[3] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.