# Importance sampling

Many Monte Carlo calculations have as a goal computing an integral.

$$A = \int_{\mathbb{R}^d} u(x)\, dx \ . \tag{1}$$

For low dimensions ($d \leq 3$ or maybe even $d \leq 5$), such integrals are better calculated by direct methods such as the trapezoid rule. For "high" dimensions ($d > 5$), direct methods are often impractical, so we resort to methods that involve randomization, which are called Monte Carlo methods.

The direct Monte Carlo method uses a factorization of the integrand $u$ into a product where one of the factors is a probability density function

$$u(x) = V(x)f(x) \ , \quad f(x) \geq 0 \text{ for all } x \ , \quad \int_{\mathbb{R}^d} f(x)\, dx = 1 \ . \tag{2}$$

If we have a set of independent samples $X_j \sim f$, then the direct Monte Carlo estimate of $A$ is

$$\widehat{A} = \frac{1}{n} \sum_{j=1}^{n} V(X_j) \quad , \ X_j \sim f \text{ , i.i.d.} \tag{3}$$

The variance of this estimator is[1]

$$\text{var}_f\Big( \widehat{A} \Big) = \frac{1}{n} \text{var}_f( \, V(X) \, ) \ . \tag{4}$$

If we have more than one probability distribution, a subscript may determine the distribution. Thus, (4) refers to the variance when $X$ has $f$ as its PDF.

*Importance sampling* usually refers to choosing the factorization (2) in a systematic way. The factorization is not unique. If $g$ is "any" other[2] probability density function, then

$$V(x)f(x) = V(x)\frac{f(x)}{g(x)}g(x) = V(x)L(x)g(x) = W(x)g(x) \ .$$

The *likelihood ratio* is factor that accounts for the change of "measure" (probability density)[3] that replaces $f$ with $g$

$$L(x) = \frac{f(x)}{g(x)} \ . \tag{5}$$

[1] i.i.d. is for independent and identically destributed.

[2] "Restrictions apply."

[3] Fancy probability distributions are described using the mathematical formalism of measure theory. Changing the probability distribution is changing the measure. In measure theory, the likelihood ratio (5) is often called the *Radon Nikodym derivative*.

The new "observable" (quantity whose expected value is being taken) is

$$W(x) = V(x)L(x) \ .$$

The importance sampling alternative to (3) is

$$\widehat{A} = \frac{1}{n}\sum_{j=1}^{n} V(X_j)L(X_j) \quad , \ X_j \sim g \ , \text{ i.i.d.} \tag{6}$$

The variance of this estimator is

$$\text{var}_g\Big(\ \widehat{A}\ \Big) = \frac{1}{n}\text{var}_g(\ V(X)L(X)\ ) \ . \tag{7}$$

Many integration problems (1) come with a natural factorization (2). This happens, for example, when the problem is to find the expected value of some function $V$ in a given probability distribution. Other problems come directly in the form (1) and we are not given factorization. An example is the *evidence integral* used in Bayesian *model selection*. Keep in mind that a natural factorization (2) may not be the best one to use in the Monte Carlo estimation (3). The goal is to estimate $A$. There are more and less efficient ways to do that.

## Importance sampling to simplify the sampling problem

The Monte Carlo estimator (3) requires samples $X_j \sim f$. We have seen that some distributions are hard to simple.

**Example**. Assignment 6 involved the Gamma distribution

$$f(t) = \frac{1}{\Gamma(s)}\, t^{s-1}e^{-t} \ .$$

The rejection sampler in Assignment 6 is not very effective (has a low acceptance probability) when $s$ is large. However, when $s$ is large there is an approximate description of $f$ using the *Laplace approximation*. This starts by writing $f$ in the form of a Gibbs Boltzmann distribution

$$f(t) = \frac{1}{Z}e^{-\phi(t)} \quad , \quad \text{with } \phi(t) = t - (s-1)\log(t) \ . \tag{8}$$

The "most likely" $t$ value is the one that minimizes the potential:

$$t_* = \arg\ \min_t \phi(t) = s - 1 \ .$$

The two term Taylor approximation of $\phi$ near the most likely point is

$$\phi(t) \approx \phi_* + \frac{1}{2}\phi''(t_*)(t - t_*)^2 \ .$$

A calculation shows that this is

$$\phi(t) \approx \phi_* + \frac{1}{2(s-1)}(t-t_*)^2 \ .$$

This leads to an approximation of $f$ that takes the form

$$f(t) \approx \frac{1}{Z} e^{-\frac{(t-t_*)^2}{2(s-1)}} \ . \tag{9}$$

The normalization constant $Z$ is this formula is different from the one in (8). The approximation (9) suggests using the right side as an alternative sampling distribution. This is Gaussian with mean $t_* = s-1$ and variance $\sqrt{s-1}$. Altogether, the importance sampling (change of measure) strategy would use

$$g(t) = \frac{1}{(s-1)\sqrt{2\pi}} e^{-\frac{(t-t_*)^2}{2(s-1)}} \quad , \quad L(t) = \frac{(s-1)\sqrt{2\pi}}{\Gamma(s-1)} \frac{t^{s-1}e^{-t}}{e^{-\frac{(t-t_*)^2}{2(s-1)}}} \ .$$

These formulas are written out in detail to make the point that they are complicated but manageable. The normalization constant in the $g$ formula is not important because we sample $g$ using the normal sampler with a given mean and variance. The normalization constants in the $L$ formula do matter. This is a drawback of this kind of importance sampling.

## Variance reduction, rare events

Importance sampling (change of measure) can lower the variance of the estimator. The variance is high when the PDF $f$ puts most of its weight in regions where $u$ is small. In this situation, only rare samples $X$ are in a region that is important for the integrand, while most of them are wasted in regions where $u$ is small. In this context, *importance sampling* means finding an alternative PDF $g$ that puts more weight where $u$ is large. You succeed when the variance on the right of (7) is much smaller than on the right side of (4). Of course, lower variance means higher accuracy in the Monte Carlo estimator. Alternatively, it means achieving a given accuracy with fewer samples.

**Example.** The problem is to estimate the Gamma function for large $s$ by Monte Carlo integration.[4] That is, to fine

$$A = \int_0^\infty t^{s-1}e^{-t}\,dt \ .$$

Looking at this integral, you might think of taking the probability density to be the exponential $f(t) = e^{-t}$ and the observable to be the power part $V(t) = t^{s-1}$.

[4] This is a terrible way to estimate the Gamma function, partly because a one dimensional integral is done more accurately and cheaply by a deterministic quadrature algorithm such as the trapezoid rule. In the particular case of the Gamma function, special properties and alternative formulas allow it to be evaluated accurately in almost no computer time.

This is a high variance strategy because samples $T \sim e^{-t}$ are not far from zero while the integrand $u(t) = t^{s-1}e^{-t}$ has a strong maximum near $t = s - 1$. A better factorization of the integrand, we suspect, is

$$u(t) = t^{s-1}e^{-t} = \frac{t^{s-1}e^{-t}}{\mathcal{N}(s-1, \sqrt{s-t})(t)} \mathcal{N}(s-1, \sqrt{s-t})(t) \ .$$

In this formula we use $\mathcal{N}$ to represent the Gaussian probability density, as in

$$\mathcal{N}(\mu, \sigma^2)(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \ .$$

This importance sampling algorithm would be to generate $n$ independent samples $T_j \sim \mathcal{N}(s-1, \sqrt{s-t})$ and then use the approximation

$$\widehat{A} = \frac{1}{n} V(T_j) \ , \quad V(t) = \frac{t^{s-1}e^{-t}}{\mathcal{N}(s-1, \sqrt{s-t})(t)} \ .$$

This method is a little more complicated than the direct method using exponential samples and $V(t) = t^{s-1}$, but the variance is so much smaller (when $s$ is large) that you definitely should do it this way.

*Rare event simulation* is a particular case that comes up in many applications. The problem is to estimate the probability of some very rare event. To use importance sampling, you would look for an alternative probability distribution that makes the rare event more likely, simulate this, and then "discount" the *hits* by the likelihood ratio.