Fundamental Concepts: Transformation, Rejection, and Reweighting

Peadar Coyle

August 17, 2012

Abstract

1 Fundamental concepts: Sampling

A fundamental concept in using Monte Carlo methods is sampling.

1.1 Transformation methods

One of the simplest methods of generating random samples from a distribution with cumulative distribution function (c.d.f.) F(x) = $\mathbb{P}(X \leq x)$ is based on the inverse of the c.d.f. The c.d.f. is an increasing function, however it is not necessarily continuous. Thus we define the *generalised inverse* $F^-(u) = \inf\{x : F(x) \geq u\}$

Theorem 1.1 (Inversion Method) *Let* $U \simeq \bigcup[0,1]$ *and* F *be a c.d.f. Then* $F^-(U)$ *has the c.d.f.* F*.*

Proof *It is easy to see (by drawing a diagram say) that* $F^-(u) \leq x$ *is equivalent to* $u \leq F(x)$*. Thus for* $U \simeq \bigcup[0,1]$

$$\mathbb{P}(F^-(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

thus F *is the c.d.f. of* $X = F^-(U)$*.*

Let us consider a classical example.

Exponential Distribution . The exponential distribution with λ ¿ 0 has the c.d.f. $F_\lambda(x) = 1 - \exp(-\lambda x)$ for $x \geq 0$. Thus $F_\lambda^-(x) = F_\lambda^{-1}(u) = -log(1-u)/\lambda$. Thus we can generate random samples from $Expo(\lambda)$ by applying the transformation $-log(1-U)/\lambda$ to a uniform $\bigcup[0,1]$ random variable U. As U and 1-U, of course have the same distribution we can use $-log(U)/\lambda$ as well. ◁

The Inversion Method is a very efficient tool for generating random numbers. However very few distributions possess a c.d.f. whose (generalised) inverse can be evaluated efficiently. Take the example of the Gaussian distribution , whose c.d.f. is not even available in closed form. Nevertheless there are other examples of transformations. An example is the Box-Muller method for generating Gaussian random variables.

Box-Muller Method for Sampling from Gaussians When sampling from the normal distribution, one faces the problem that neither the c.d.f. $\phi(\cdot)$, nor its inverse has a closed-form expression. Thus we cannnot use the inversion method. It turns out however, that if we consider a pair $X_1, X_2 \overset{i.i.d.}{\sim} N(0,1)$, as a point (X_1, X_2) in the plane, then its polar coordinates (R, θ) are again independent and have distributions we can easily sample from: $\bigcup[0, 2\pi]$ and $R^2 \sim Expo(1/2)$. Then the joint density of (θ, r^2) is

$$f_{(\theta,r^2)}(\theta, r^2) = \frac{1}{2\pi} 1_{[0,2\pi]}(\theta) \cdot \frac{1}{2} \exp\left(-\frac{1}{2} r^2\right) \cdot 1_{[0,2\pi]}(\theta)$$

To obtain the probability density function of

$$X_1 = \sqrt{R^2} \cdot \cos(\theta) \qquad X_2 = \sqrt{R^2} \cdot \sin(\theta)$$

we need to use the transformaiton of densities formula.

$$f_{(X_1,X_2)}(x_1, x_2) = f_{(\theta,r^2)}(\theta(x_1, x_2), r^2(x_1, x_2)) \cdot \left| \begin{matrix} \frac{\partial x_1}{\partial \theta} & \frac{\partial x_1}{\partial r^2} \\ \frac{\partial x_2}{\partial \theta} & \frac{\partial x_2}{\partial r^2} \end{matrix} \right|^{-1} = \frac{1}{4\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)^2\right) \cdot 2$$

$$= \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_1^2\right)\right) \cdot \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_2^2\right)\right)$$

as

$$\left| \begin{matrix} \frac{\partial x_1}{\partial \theta} & \frac{\partial x_1}{\partial r^2} \\ \frac{\partial x_2}{\partial \theta} & \frac{\partial x_2}{\partial r^2} \end{matrix} \right| = \left| \begin{matrix} -r\sin(\theta) & \frac{\cos(\theta)}{2r} \\ r\cos(\theta) & \frac{\sin(\theta)}{2r} \end{matrix} \right| = \left| -\frac{r\sin(\theta)}{2r} \quad -\frac{r\cos(\theta)^2}{2r} \right| = \frac{1}{2}$$

Thus $X_1, X_2 \sim N(0,1)$. As their joint density factorises, X_1 and X_2 are independent, as required. Thus we only need to generate $\theta \sim \bigcup[0, 2\pi]$, and $R^2 \sim Expo(1/2)$. Using U_1,$U_2 \overset{i.i.d.}{\sim} \bigcup[0, 1]$ and example 1.1 we can generate R $= \sqrt{R^2}$ and θ by

$$R = \sqrt{-2\log(U_1)}, \qquad \theta = 2\pi U_2,$$

and thus

$$X_1 = \sqrt{-2\log(U_1)} \cdot \cos(2\pi U_2), \quad X_2 = \sqrt{-2\log(U_1} \cdot \sin(2\pi U_2)$$

are two independent realisations from a $N(0,1)$ distribution $\quad \triangleleft$

The idea of transformation methods like the Inversion Method was to generate random samples from a distribution other than the target distribution and to transform them such that they come from the desired target distribution. In many situatioons, we cannot find such a transformation in closed form. In these cases we have to find other ways of correcting for the fact that we sample from the 'wrong' distribution. The next two sections present two such ideas: rejection sampling and importance sampling.

1.2 Rejection sampling

The basic idea of rejection sampling is to sample from an *instrumental distribution* and reject samples that are 'unlikely' under the target distribution. Assume that we want to sample from a target distribution whose density f is known to us. The simple idea underlying rejection sampling (and other Monte Carlo algorithms) is the rather trivial identity

$$f(x) = \int_0^{f(x)} 1du = \int_0^1 1_{0<u<f(x)} du$$

Thus f(x) can be interpreted as the marginal density of a uniform distribution on the area under the density f(x)

$$\{(x,u) : 0 \leq u \leq f(x)\}$$

Consider for instance the beta distribution

Sampling from a Beta distribution The $\texttt{Beta}(\texttt{a},\texttt{b})$ distribution (a,b $\geq$ 0) has the density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad for\ 0 < x < 1,$$

where $\Gamma(z) = \int_0^{+\infty} t^{a-1}\exp(-t)dt$ is the Gamma function. For a,b $>$ 1 the $\texttt{Beta}(\texttt{a},\texttt{b})$ density is unimodal [1] with mode (a-1)/(a+b-2). It attains its maximum at 1680/729 $\approx$ 2.305 at x = 1/3. Using the above identity we can draw from $\texttt{Beta}(\texttt{a},\texttt{b})$ by drawing from a uniform distribution on the area under the density (the area under the curve - if the distribution is plotted). We sample from the area under the density, by sampling from the rectangle and keeping only the samples that fall in the area under the curve. Mathematically speaking, we sample independently $X \sim \bigcup[0,1]$ and $U \sim \bigcup[0,2.4]$. We keep the pair (X,U) if $X < f(X)$, otherwise we reject it. The conditional probability that a pair (X,U) is kept if X = x is

$$\mathbb{P}(U < f(X)|X = x) = \mathbb{P}(U < f(x)) = f(x)/2.4$$

As X and U are drawn independently we can rewrite our algorithm as: Draw X from $\bigcup[0,1]$ and accept X with probability f(X)/2.4 otherwise reject X.

◁

The method proposed in 1.2 is based on bounding the density of the Beta distribution by a box. Whilst this is a powerful idea, it cannot be directly applied to other distributions, as the density might be unbounded or have infinite support. However we might be able to bound the density of f(x) by M·g(x), where g(x) is a density we can easily sample from.

Algorithm 1.2 (Rejection sampling) . *Given two densities f,g with f(x) < M·g(x) for all x, we can generate a sample from f as follows:*

[1] Think of what happens in the discrete case, when we count the number of items in a distribution, the most often occuring is the mode, which is clearly the maximum of the distribution, in the continuous case

1. *Draw $X \sim g$*

2. *Accept X as a sample from f with probability*

$$\frac{f(X)}{M \cdot g(X)},$$

otherwise return to step 1.

Proof *We have*

$$\mathbb{P}(X \in \xi \textit{and is accepted}) = \int_{\xi} g(x) \frac{f(x)}{M \cdot g(x)} dx = \frac{\int_{\xi} f(x)dx}{M} \tag{1}$$

and thus[2]

$$\mathbb{P}(X \textit{ is accepted}) = \mathbb{P}(X \in S \textit{ and is accepted}) = \frac{1}{M} \tag{2}$$

yielding

$$\mathbb{P}(x \in \xi | X \textit{ is accepted}) = \frac{\mathbb{P}(X \in \xi \textit{ and is accepted})}{\mathbb{P}(X \textit{ is accepted}} = \frac{\int_{\xi} f(x)dx/M}{1/M} = \int_{\xi} f(x)dx \tag{3}$$

Thus the density of the values accepted by the algorithm is f(·)

1.3 Importance sampling

In rejection sampling we have compensated for the fact that we sampled from the instrumental distribution $g(x)$ instead of f(x) by rejecting some of the values proposed by g(x). Importance sampling is based on the idea of using weights to correct for the fact that we sample from the instrumental distribution $g(x)$ instead of the target distribution $f(x)$. Importance sampling is based on the identity

$$\mathbb{P}(X \in A) = \int_A f(x)dx = \int_A g(x) \frac{f(x)}{g(x)} dx = \int_A g(x)w(x)dx \tag{4}$$

for all $g(\cdot)$, such that $g(x) > 0$ for (almost) all x with f(x) >. We can generalise this identity by considering the expectation $\mathbb{E}_f(h(X))$ of a measurable function h:

$$\mathbb{E}_f(h(X)) = \int_S f(x)h(x)dx = \int_S g(x) \frac{f(x)}{g(x)} h(x)dx = \int_S g(x)w(x)h(x)dx = \mathbb{E}_g(w(X) \cdot h(X)), \tag{5}$$

if g(x) > 0 for (almost) all x with f(x).h(x) $\neq 0$. Assume we have have a sample $X_1, \cdots, X_n \sim g$. Then, provided $\mathbb{E}_g|w(X) \cdot h(X)|$ exists,

$$\frac{1}{n} \sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\overset{n \to \infty}{\to}} \mathbb{E}_g(w(X) \cdot h(X))$$

[2]We denote by S the set of all possible values X can take, i.e. $\int_S f(x)dx = 1$

(by the Law of Large Numbers) and thus by (5)

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\overset{n\to\infty}{\to}} \mathbb{E}_f(h(X))$$

In other words, we can estimate $\mu = \mathbb{E}_g(h(X))$ by

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)$$

Note that whilst $\mathbb{E}_g(w(X)) = \int_S \frac{f(x)}{g(x)} g(x)dx = \int_S f(x) = 1$, the weights $w_1(X), \cdots, w_n(X)$ do not necessarily sum up the n, so one might want to consider the *self-normalised* version

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{n} w(X_i} \sum_{i=1}^{n} w(X_i)h(X_i)$$

Algorithm 1.3 (Importance Sampling) *1. For i = 1,$\cdots$,n:*

- *Generate* $X_i \sim g$.
- *Set* $w(X_i) = \dfrac{f(X_i)}{g(X_i}$

2. Return either

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)$$

or

$$\hat{\mu} = \frac{1}{\sum_{i=1}^{n} w(X_i} \sum_{i=1}^{n} w(X_i)h(X_i)$$

There are theorems for bias and the variance of importance sampling.

Theorem 1.4 (Bias and Variance of Importance Sampling) *1.* $\mathbb{E}_g(\tilde{\mu}) = \mu$

2. $Var_g(\tilde{\mu}) = \frac{Var_g(w(X)\cdot h(X)}{n}$

3. $\mathbb{E}_g(\hat{\mu}) = \mu + \dfrac{\mu Var_g(w(X)) - Cov_g(w(X), w(X)\cdot h(X)}{n}$

Proof *We shall only prove (1) and (2), the other two items are proved in Liu, 2001, p.35* [3]

1. $\mathbb{E}_g\left(\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g(w(X_i)h(X_i)) = \mathbb{E}_f(h(X))$

2. $Var_g\left(\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var_g(w(X_i)h(X_i)) = \dfrac{Var_g(w(X)h(X))}{n}$

[3] Monte Carlo Methods in Scientific Computing

Note that the theorem implies that in contrast to $\tilde{\mu}$ the self-normalised estimator $\hat{\mu}$ is biased. The self-normalised estimator $\hat{\mu}$ might however have a lower variance. In addition, it has another advantage: we only need to know the density up to a multiplicative constant, as it is often the case in hierarchical Bayesian modelling. Assume f(x) = $C \cdot \pi(x)$, then

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)} = \frac{\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}h(X_i)}{\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}} = \frac{\sum_{i=1}^{n} \frac{C\cdot\pi(X_i)}{g(X_i)}h(X_i)}{\sum_{i=1}^{n} \frac{C\cdot\pi(X_i)}{g(X_i)}} = \frac{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)}h(X_i)}{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)}}$$

i.e. the self-normalised estimator $\hat{\mu}$ does not depend on the normalisation constant C. On the other hand, as we seen in the proof of the theorem 1.4 that it is a lot harder to analyse the theoretical propoerties of the self-normalised estimator $\hat{\mu}$. Although the above equations (4) and (5) hold for every g with $supp(g) \supset supp(f.h)$