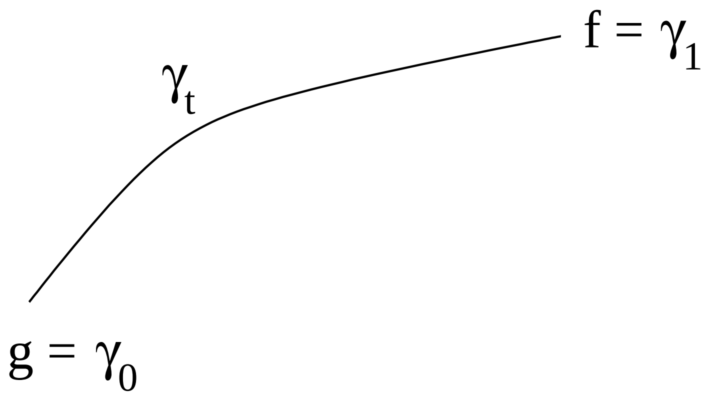

Figure 1: Path Connecting g and f

Abstract

Using the exponential and mixture connections in the space of distributions for sampling. Applications: Thermodynamic integration, The half Monty-Carlos Method for sampling from f by generating from g.

Paths in The Space of Distributions

Given two probability densities (or probability mass functions if the space of observations is discrete) f and g we define a path connecting them as a piece-wise smooth map from the interval $[0, 1]$ to the space of distributions for the data, $t \longrightarrow \gamma_t$ such that, $\gamma_0 = g$ and $\gamma_1 = f$. See the picture.

We assume that $f(x)$ is a complicated expression (with x possibly high dimensional) from which we want to sample. Due to the complexity of $f(x)$ we assume that there is no simple method for generating samples from f but we assume that there is available a method for generating samples from g.

It is intuitively clear how a path connecting the target distribution f with the simpler distribution g may help by providing a way to arrive to the target in little steps along the path. In the following sections we show two examples of the application of this simple but powerful idea.

There are many ways of building paths between densities but there are two special connections that have been proven useful, not only for generating random variables with the computer but also in the general theory of information geometry. I am of the opinion that the utility of the exponential and mixture connections is just the tip of the iceberg. I think that there is a whole unexplored geometric world lurking behind computer sampling.

The Exponential Connection

A path can be obtained by mixing the loglikelihoods of two extreme distributions,

$$\log(\gamma_t(x)) = t \log\ f(x) + (1-t) \log\ g(x) - \log\ Z_t$$

where Z_t is a normalization constant. We can also write,

$$\gamma_t(x) = \frac{f^t(x) g^{1-t}(x)}{Z_t}$$

with,

$$Z_t = \int f^t(x) g^{1-t}(x) dx$$

This path is known as the exponential connection between the densities g and f. The path is an exponential family with t as the natural parameter and the loglikelihood ratio $\log(f(x)/g(x))$ as sufficient statistic. This connection provides a notion of a straight line in the space of distributions that is diametrically opposed to the notion of straightness provided by the mixture connection introduced next.

The Mixture Connection

Another path connecting the target density f to the easier to sample from density g, is given by the mixture connection,

$$\gamma_t(x) = t f(x) + (1-t) g(x)$$

This path provides an alternative notion of a straight line in the space of distributions. Different ways of defining straightness give different ways of defining curvature and parallelism for example.

Thermodynamic Integration

Thermodynamic integration is a technique that makes use of the exponential connection for estimating the ratio of two normalization constants. The problem of computing the ratio of two normalization constants appears in the important problem of model selection or hypothesis testing. Suppose that we want to compare two competing models M_0, M_1 for the available data D. We need to compute the ratio of the two posterior probabilities for the models. By Bayes' theorem we get,

$$\log \frac{P(M_0|D)}{P(M_1|D)} = \log \frac{P(D|M_0)}{P(D|M_1)} + \log \frac{\pi(M_0)}{\pi(M_1)}$$

where, $\pi(M_i)$ are the prior probabilities for the two models. The above expression is a recurrent head ache for the faithful Bayesians for two reasons.

First, the standard improper non informative priors can't be used here. The ratio of infinities for the second term on the right is undefined when improper priors are used. Second, it is not enough to know the likelihood for the data up to a normalization constant since the first term on the right needs the ratio of both normalization constants in order to produce a definite numerical answer.

Let us then assume that the target density is only known up to a proportionality constant, i.e.,

$$f(x) = \frac{1}{Z_1} h(x)$$

with,

$$Z_1 = \int h(x) dx$$

unknown. Allow also the possibility of having g unnormalized, even though this is less common, i.e. define,

$$Z_0 = \int g(x) dx$$

which it is of course 1 when g is a proper density. The objective is to compute,

$$\begin{aligned}
\log \frac{Z_1}{Z_0} &= \log \frac{\int h}{\int g} \\
&= \log \frac{\int \frac{h}{g} g}{\int g} \\
&= \log \left\langle \frac{h(x)}{g(x)} \right\rangle_g \\
&\approx \log \left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{h(x_j)}{g(x_j)} \right\}
\end{aligned}$$

where the $x_1, x_2, \ldots$ are iid with density proportional to g. In a few dimensions when g is very close in shape to h the above approximation may work. In general, the simple importance sampling approximation above, fails to produce useful estimates. The use of the exponential connection can improve the estimates dramatically in the following way. Define,

$$Z_t = \int h^t(x) g^{1-t}(x) dx$$

which is the normalization constant for the mixture of the log likelihoods of h and g. Notice that the definition includes the extremes that we are after Z_0 and Z_1. From,

$$Z_{t+\epsilon} = \int \left(\frac{h(x)}{g(x)}\right)^{\epsilon} h^t(x) g^{1-t}(x) dx$$

we obtain,

$$\frac{Z_{t+\epsilon}}{Z_t} = \left\langle \left(\frac{h(x)}{g(x)}\right)^{\epsilon} \right\rangle_t$$

where by the last notation we mean the expected value of what is inside the brackets when x follows the exponential connection density with parameter t. For values of ϵ close to zero the sample means obtained by using the Metropolis algorithm produce very accurate estimates of the ratios for different values of t. Finally by climbing up the path we obtain,

$$\frac{Z_1}{Z_0} = \frac{Z_{1/n}}{Z_0} \frac{Z_{2/n}}{Z_{1/n}} \cdots \frac{Z_{(n-1)/n}}{Z_{(n-2)/n}} \frac{Z_1}{Z_{(n-1)/n}}$$

which is estimated as,

$$\begin{aligned}
\log \frac{Z_1}{Z_0} &= \sum_{i=0}^{n-1} \log \left\langle \left(\frac{h(x)}{g(x)}\right)^{1/n} \right\rangle_{i/n} \\
&\approx \sum_{i=0}^{n-1} \log \left(\frac{1}{N} \sum_{j=1}^{N} \left(\frac{h(x_{i,j})}{g(x_{i,j})}\right)^{1/n} \right)
\end{aligned}$$

where $x_{i,1}, x_{i,2}, \ldots, x_{i,N}$ are sample from the exponential connection with parameter i/n. Thus, in order to get the ratio we need to implement a sequence of MCMCs.

It is also possible to get a good estimate of the ratio by running a single MC but at the expense of having to generate from an equiprobable mixture of exponential connections. Here is how: Write,

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{Z_{t+\epsilon} - Z_t}{\epsilon Z_t} &= \lim_{\epsilon \to 0} \frac{\left\langle \left(\frac{h(x)}{g(x)}\right)^{\epsilon} \right\rangle_t - 1}{\epsilon} \\
\frac{d}{dt} \log Z_t &= \left\langle \frac{d}{d\epsilon} \exp(\epsilon \log \frac{h}{g}) \right\rangle_t = \left\langle \log \left(\frac{h(x)}{g(x)}\right) \right\rangle_t
\end{aligned}$$

from where we get,

$$\log Z_t = \int_0^t \left\langle \log \frac{h(x)}{g(x)} \right\rangle_{\lambda} d\lambda + \log Z_0$$

denoting the exponential connection with parameter λ by γ_λ and interchanging the order of integration we obtain,

$$\log \frac{Z_1}{Z_0} = \int \log \frac{h(x)}{g(x)} \left[\int_0^1 \gamma_\lambda(x) d\lambda \right] dx$$

and by calling,

$$G^*(x) = \int_0^1 \gamma_t(x) dt$$

the continuous equiprobable mixture of the elements of the path connecting f with g we can write,

$$\log \frac{Z_1}{Z_0} = \left\langle \log \frac{h(x)}{g(x)} \right\rangle_{G^*}$$

a very beautiful formula in its own very self and also useful. It says something like

> *the log of the ratio of averages of two positive functions*
> *equals the average of log of ratios*
> *with respect to the average exponential connection*
> *linking those two functions*

The Plain Vanilla Thermodynamic Integrator

```
{
sum ← 0
for j = 1, 2, ..., N

    {
    u ← unif(0, 1)
    x ← sample γ_u by Metropolis
    sum ← sum + log h(x)/g(x)
    }

  return sum/N }
```

In the above algorithm the metropolis steps can be dramatically improved by starting from a previously observed sample from γ_t with t as close as possible to the current u so that only a few iterations are necessary.

The Full Monty-Carlos Method

This is a new, as far as I know previously unknown method, that allows to sample from any f starting from any g. The simplest version of the method works as long as the ratio of normalization constants of f and g is known. If that ratio is unknown, it must first be estimated by one of the methods of the previous sections. A modification of the simpler version allows the freedom of using unnormalized densities.

The idea is very simple and it exploits a simple but super useful property of the mixture connection.

Theorem 1 *Let,*

$$\gamma_t(x) = tf(x) + (1-t)g(x)$$

be the mixture connection between two fully normalized pdfs (or probability mass functions) f and g. Then, for all values of x

$$\gamma_{t+\epsilon}(x) < \left(1 + \frac{\epsilon}{t}\right) \gamma_t(x)$$

for all t and ϵ so that the mixture parameters are always in $[0, 1]$.

Proof: just write,

$$\begin{aligned} \lambda = \frac{\gamma_{t+\epsilon}(x)}{\gamma_t(x)} &= \frac{\gamma_t(x) + \epsilon(f(x) - g(x))}{\gamma_t(x)} \\ &= 1 + \epsilon \frac{f(x) - g(x)}{\gamma_t(x)} \end{aligned}$$

and consider two cases.

Case I: $f(x) \leq g(x)$ In this case it follows from the last equation that, $\lambda \leq 1$.

Case II: $f(x) > g(x)$ For this case write the denominator in the last equation above as, $\gamma_t(x) = g(x) + t(f(x) - g(x))$ and divide the numerator and the denominator by $(f(x) - g(x))$ to obtain,

$$\lambda = 1 + \frac{\epsilon}{t + g(x)/(f(x) - g(x))} < 1 + \frac{\epsilon}{t}$$

Q.E.D.

The Monty-Carlos method makes use of this theorem in the following way. Start by generating a sample from $g = \gamma_0$ then move (horizontally in the picture) with a few metropolis iterations to get a sample from $\gamma_{1/n}$. Then, try to climb up towards f as much as possible by using the exact rejection constants provided by the previous theorem. When the sample gets rejected then fall back to the last acceptance that assures that the sample is from $\gamma_{k/n}$. This x is now used as an initial value for metropolis to try to go one step forward to $\gamma_{(k+1)/n}$. Continue in this way until arriving at f.

When n is small and g is close to f the method can be used for generating independent samples from f. For really complicated high dimensional f a high

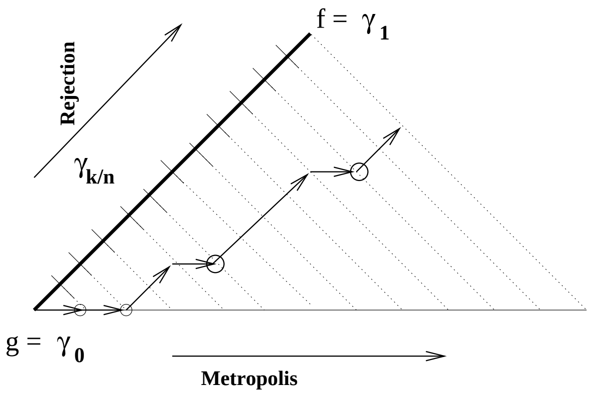

Figure 2: The Full Monty-Carlos Method

value for n is needed and there is no guarantee that g is close to f. In such cases, the method can be used for producing a single sample from f and after that continue with a version of metropolis but now knowing that metropolis has converged.

In order to provide a short description of the algorithm will make use of two functions:

$Met(x, t, Stop\_crit)$ Returns a sample from γ_t starting from x using the metropolis algorithm with a given stopping criterion (e.g. maximum number of iterations).

$Climb(x, k, n)$ Returns an integer k' with $k \leq k' \leq n$. The sample x from $\gamma_{k/n}$ is pushed-up as high as possible until it gets rejected at $\gamma_{k'/n}$.

The Full Monty-Carlos Algorithm

```
{
x ← sample from g(·)
k ← 1
while (k < n)

        {
        x ← Met(x, k/n, Stop_crit)
        k ← Climb(x, k, n)
        }

   Return x
}
```

Met$(x, t, Stop\_crit)$

```
{
until (Stop_crit)
        {
        y ← sample g(·)
        u ← unif(0,1)
        if u < min(1, γ_t(y)/γ_t(x))      then x ← y
        }
   Return x }
```

$$\text{if } u < \min\left(1, \frac{\gamma_t(y)}{\gamma_t(x)}\right) \quad \text{then } x \leftarrow y$$

Climb(x, k, n)

```
{
j ← 0
u ← unif(0,1)
while (k + j < n) AND
```

$$\left(u * \left[1 + \frac{1}{k+j-1}\right] * \gamma_{(k+j-1)/n}(x) < \gamma_{(k+j)/n}(x)\right)$$

```
        {
        u ← unif(0,1)
        j ← j + 1
        }
   Return k + j }
```

0.1 The Full-Monty Without Normalization Constants

If f and g are only known up to normalization constants the method works if we just change the **Climb** function to:

Climb(x, k, n)

```
{
j ← 0
u ← sample from exp(1)
while (k + j < n) AND
```

$$\left(u < \log\left(\frac{(k+j)\gamma_{(k+j-1)/n}(x)}{(k+j-1)\gamma_{(k+j)/n}(x)}\right)\right)$$

```
        {
        u ← sample from exp(1)
        j ← j + 1
        }
```

```
    Return k + j }
```

The fact that it works follows from the following (well known) theorem:

Lemma 1 *For a nonnegative function $h(x)$ the algorithm,*

```
{
Repeat
        {
        x ← sample from g()
        u ← sample from exp(1)
        }
   Until h(x) ≤ u
Return x }
```

produces a sample with density

$$f(x) = cg(x)\exp(-h(x))$$

proof:
For any borel set B we have,

$$\begin{aligned} P[X \in B] &= \frac{P[X \in B, U \geq h(X)]}{P[U \geq h(X)]} \\ &= \frac{\int_B \int_{y=h(x)}^{\infty} g(x)\exp(-y)dydx}{\int g(x)\exp(-h(x))dx} \\ &= \frac{\int_B g(x)\exp(-h(x))dx}{\int g(x)\exp(-h(x))dx} \\ &= \int_B f(x)dx \end{aligned}$$