Chapter 6

Importance sampling

6.1 The basics

To movtivate our discussion consider the following situation. We want to use Monte Carlo to compute $\mu = E[X]$. There is an event E such that $P(E)$ is small but X is small outside of E. When we run the usual Monte Carlo algorithm the vast majority of our samples of X will be outside E. But outside of E, X is close to zero. Only rarely will we get a sample in E where X is not small.

Most of the time we think of our problem as trying to compute the mean of some random variable X. For importance sampling we need a little more structure. We assume that the random variable we want to compute the mean of is of the form $f(\vec{X})$ where $\vec{X}$ is a random vector. We will assume that the joint distribution of $\vec{X}$ is absolutely continous and let $p(\vec{x})$ be the density. (Everything we will do also works for the case where the random vector $\vec{X}$ is discrete.) So we focus on computing

$$E f(\vec{X}) = \int f(\vec{x}) p(\vec{x}) dx \tag{6.1}$$

Sometimes people restrict the region of integration to some subset D of R^d. (Owen does this.) We can (and will) instead just take $p(x) = 0$ outside of D and take the region of integration to be R^d.

The idea of importance sampling is to rewrite the mean as follows. Let $q(x)$ be another probability density on R^d such that $q(x) = 0$ implies $f(x)p(x) = 0$. Then

$$\mu = \int f(x) p(x) dx = \int \frac{f(x)p(x)}{q(x)} \, q(x) \, dx \tag{6.2}$$

We can write the last expression as

$$E_q \left[\frac{f(\vec{X})p(\vec{X})}{q(\vec{X})} \right] \tag{6.3}$$

where E_q is the expectation for a probability measure for which the distribution of $\vec{X}$ is $q(x)$ rather than $p(x)$. The density $p(x)$ is called the *nominal or target distribution* , $q(x)$ the *importance or proposal distribution* and $p(x)/q(x)$ the *likelihood ratio.* Note that we assumed that $f(x)p(x) = 0$ whenever $q(x) = 0$. Note that we do not have to have $p(x) = 0$ for all x where $q(x) = 0$.

The importance sampling algorithm is then as follows. Generate samples $\vec{X}_1, \cdots, \vec{X}_n$ according to the distribution $q(x)$. Then the estimator for μ is

$$\hat{\mu}_q = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\vec{X}_i)p(\vec{X}_i)}{q(\vec{X}_i)} \tag{6.4}$$

Of course this is doable only if $f(x)p(x)/q(x)$ is computable.

Theorem 1 *$\hat{\mu}_q$ is an unbaised estimator of μ, i.e., $E_q\hat{\mu}_q = \mu$. Its variance is σ_q^2/n where*

$$\sigma_q^2 = \int \frac{f(x)^2 p(x)^2}{q(x)} dx - \mu^2 = \int \frac{(f(x)p(x) - \mu q(x))^2}{q(x)} dx \tag{6.5}$$

Proof: Straightforward. **QED**

We can think of this importance sampling Monte Carlo algorithm as just ordinary Monte Carlo applied to $E_q[f(\vec{X})p(\vec{X})/q(\vec{X})]$. So a natural estimator for the variance is

$$\hat{\sigma}_q^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{f(\vec{X}_i)p(\vec{X}_i)}{q(\vec{X}_i)} - \hat{\mu}_q \right]^2 \tag{6.6}$$

What is the optimal choice of the importance distribution $q(x)$? Looking at the theorem we see that if we let $q(x) = f(x)p(x)/\mu$, then the variance will be zero. This is a legitimate probability density if $f(x) \geq 0$. Of course we cannot really do this since it would require knowing μ. But this gives us a strategy. We would like to find a density $q(x)$ which is close to being proportional to $f(x)p(x)$.

What if $f(x)$ is not positive? Then we will show that the variance is minimized by taking $q(x)$ to be proportional to $|f(x)|p(x)$.

Theorem 2 *Let $\overline{q}(x) = |f(x)|p(x)/c$ where c is the constant that makes this a probability density. Then for any probability density $q(x)$ we have $\sigma_{\overline{q}} \leq \sigma_q$*

Proof: Note that $c = \int |f(x)|q(x)dx$.

$$\begin{aligned}
\sigma_{\overline{q}} - \mu^2 &= \int \frac{f(x)^2 p(x)^2}{\overline{q}(x)} dx && (6.7)\\
&= c \int |f(x)|p(x)dx && (6.8)\\
&= \left(\int |f(x)|p(x)dx \right)^2 && (6.9)\\
&= \left(\int \frac{|f(x)|p(x)}{q(x)} q(x)dx \right)^2 && (6.10)\\
&\leq \int \frac{f(x)^2 p(x)^2}{q(x)^2} q(x)dx && (6.11)\\
&= \int \frac{f(x)^2 p(x)^2}{q(x)} dx && (6.12)\\
&= \sigma_q - \mu^2 && (6.13)
\end{aligned}$$

where we have used the Cauchy Schwarz inequality with respect to the probaility measure $q(x)dx$. (One factor is the function 1.) **QED**.

Since we do not know $\int f(x)p(x)dx$, we probably do not know $\int |f(x)|p(x)dx$ either. So the optimal sampling density given in the theorem is not realizable. But again, it gives us a strategy. We want a sampling density which is approximately proportional to $|f(x)|p(x)$.

Big warning: Even if the original $f(\vec{X})$ has finite variance, there is no guarantee that σ_q will be finite. Discuss heavy tails and light tails.

How the sampling distribution should be chosen depends very much on the particular problem. Nonetheless there are some general ideas which we illustrate with some trivial examples.

If the function $f(x)$ is unbounded then ordinary Monte Carlo may have a large variance, possibly even infinite. We may be able to use importance sampling to turn a problem with an unbounded random variable into a problem with a bounded random variable.

Example We want to compute the integral

$$I = \int_0^1 x^{-\alpha} e^{-x} \, dx \qquad (6.14)$$

where $0 < \alpha < 1$. So the integral is finite, but the integrand is unbounded. We take $f(x) = x^{-\alpha} e^{-x}$ and the nominal distribution is the uniform distribution on $[0, 1]$. Note that f will have infinite variance if $\alpha \leq -1/2$.

We take the sampling distribution to be

$$q(x) = \frac{1}{1-\alpha} x^{-\alpha} \tag{6.15}$$

on $[0, 1]$. This can be sampled using inversion. We have

$$f(x) \frac{p(x)}{q(x)} = e^{-x}(1-\alpha) \tag{6.16}$$

So we do a Monte Carlo simulation of $E_q[e^{-X}(1-\alpha)]$ where X has distribution q. Note that $e^{-X}(1-\alpha)$ is a bounded random variable.

The second general idea we illustrate involves rare-event simulation. This refers to the situation where you want to compute the probabily of an event when that probability is very small.

Example: Let Z have a standard normal distribution. We want to compute $P(Z \geq 4)$. We could do this by a Monte Carlo simulation. We generate a bunch of samples of Z and count how many satisfy $Z \geq 4$. The problem is that there won't be very many (probably zero). If $p = P(Z \geq 4)$, then the variance of $1_{Z \geq 4}$ is $p(1-p) \approx p$. So the error with n samples is of order $\sqrt{p/n}$. So this is small, but it will be small compared to p only if n is huge.

Our nominal distribution is

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \tag{6.17}$$

We take the sampling distribution to be

$$q(x) = \begin{cases} e^{-(x-4)}, & \text{if } x \geq 4, \\ 0, & \text{if } x < 4, \end{cases} \tag{6.18}$$

The sampling distribution is an exponential shifted to the right by 4. In other words, if Y has an exponential distribution with mean 1, then $Y + 4$ has the distribution q. The probability we want to compute is

$$\begin{aligned} p &= \int 1_{x \geq 4} p(x) \, dx && (6.19) \\ &= \int 1_{x \geq 4} \frac{p(x)}{q(x)} q(x) \, dx && (6.20) \end{aligned}$$

The likehood ratio is

$$w(x) = \frac{p(x)}{q(x)} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2 + x - 4) \tag{6.21}$$

On $[4, \infty)$ this function is decreasing. So its maximum is at 4 where its value is $\exp(-8)/\sqrt{2\pi}$ which is really small. The variance is no bigger than the second moment which is bounded by this number squared. This is $\exp(-16)/2\pi$. Compare this with the variance of ordinary MC which saw was of the order of p which is on the order of $\exp(-8)$. So the decrease in the variance is huge.

Example We return to the network example, following Kroese's review article. Let $U_1, U_2, \cdots, U_5$ be independent and uniform on $[0, 1]$. Let T_i be U_i multiplied by the approriate constant to give the desired distribution for the times T_i. We want to estimate the mean of $f(U_1, \cdots, U_5)$ where f is the minimum time. The nominal density is $p(u) = 1$ on $[0, 1]^5$. For our sampling density we take

$$g(u) = \prod_{i=1}^{5} \nu_i u_i^{\nu_i - 1} \tag{6.22}$$

where the ν_i are parameters. (This is a special case of the beta distribution.) Note that $\nu_i = 1$ gives the nominal distribution p. There is no obvious choice for the ν_i. Kroese finds that with $\nu = (1.3, 1.1, 1.1, 1.3, 1.1)$ the variance is reduced by roughly a factor of 2.

We have discussed importance sampling in the setting where we want to estimate $E[f(\vec{X})]$ and $\vec{X}$ is jointly absolutely continuous. Everything we have done works if $\vec{X}$ is a discrete RV. For this discussion I will drop the vector notation. So suppose we want to compute $\mu = E[f(X)]$ where X is discrete with probability mass function $p(x)$, i.e., $p(x) = P(X = x)$. If $q(x)$ is another discrete distribution such that $q(x) = 0$ implies $f(x)p(x) = 0$, then we have

$$\mu = E[f(X)] = \sum_x f(x)p(x) = \sum_x \frac{f(x)p(x)}{q(x)} q(x) = E_q[\frac{f(x)p(x)}{q(x)}] \tag{6.23}$$

where E_q means expectation with repect to $q(x)$.

Example - union counting problem (from Fishman)
We have a finite set which we will take to just be $\{1, 2, \cdots, r\}$ and will call Ω. We also have a collection $S_j, j = 1, \cdots, m$ of subsets of Ω. We know r, the cardinality of Ω and the cardinalities $|S_j|$ of all the given subsets. Througout this example we use $|\quad|$ to denote the cardinality of a set. We want to compute $l = |U|$ where

$$U = \cup_{j=1}^{m} S_j \tag{6.24}$$

We assume that r and l are huge so that we cannot do this explicitly by finding all the elements in the union. We can do this by a straightforward Monte Carlo if two conditions are met. First, we can sample from the uniform distribution on Ω. Second, given an $\omega \in \Omega$ we can determine if $\omega \in S_j$ in a reasonable amount of time. The MC algorithm is then to generate a large number, n, of samples ω_i from the uniform distribution on Ω and let X be the number that are in the union U. Our estimator is then rX/n. We are computing $E_p[f(\omega)]$, where

$$f(\omega) = r 1_{\omega \in U} \tag{6.25}$$

We are assuming r and n are both large, but suppose r/n is small. Then this will be an inefficient MC method.

For our importance sampling algorithm, define $s(\omega)$ to be the number of subsets S_j that contain ω, i.e.,

$$s(\omega) = |\{j : \omega \in S_j\}| \tag{6.26}$$

and let $s = \sum_\omega s(\omega)$. Note that $s = \sum_j |S_j|$. The importance distribution is taken to be

$$q(\omega) = \frac{s(\omega)}{s} \tag{6.27}$$

The likelihood ratio is just

$$\frac{p(\omega)}{q(\omega)} = \frac{s}{rs(\omega)} \tag{6.28}$$

Note that $q(\omega)$ is zero when $f(\omega)$ is zero. So $f(\omega)q(\omega)/p(\omega)$ is just $\frac{s}{s(\omega)}$. We then do a Monte Carlo to estimate

$$E_q[\frac{s}{s(\omega)}] \tag{6.29}$$

However, is it really feasible to sample from the q distribution? Since l is huge a direct attempt to sample from it may be impossible. We make two assumptions. We assume we know $|S_j|$ for all the subsets, and we assume that for each j, we can sample from the uniform distribution on S_j. Then we can sample from q as follows. First generate a random $J \in \{1, 2, \cdots, m\}$ with

$$P(J = j) = \frac{|S_j|}{\sum_{i=1}^m |S_i|} \tag{6.30}$$

Then sample ω from the uniform distribution on S_J. To see that this gives the desired density $q()$, first note that if ω is not in $\cup_i S_i$, then there is no chance of picking ω. If ω is in the union, then

$$P(\omega) = \sum_{j=1}^m P(\omega|J = j)P(J = j) = \sum_{j:\omega \in S_j} \frac{1}{|S_j|} \frac{|S_j|}{\sum_{i=1}^m |S_i|} = \frac{s(\omega)}{s} \tag{6.31}$$

Fishman does a pretty complete study of the variance for this importance sampling algorithm. Here we will just note the following. The variance will not depend on n. So if n, r are huge but r/n is small then the importance sampling algorithm will certainly do better than the simple Monte Carlo of just sampling uniformly from Ω.

Stop - Mon, 2/15

6.2 Self-normalized importance sampling

In many problems the density we want to sample from is only known up to an unknown constant, i.e., $p(x) = c_p p_0(x)$ where $p_0(x)$ is known, but c_p is not. Of course c_p is determined by the requirement that the integral of $p(x)$ be 1, but we may not be able to compute the integral. Suppose we are in this situation and we have another density $q(x)$ that we can sample from. It is also possible that $q(x)$ is only known up to a constant, i.e., $q(x) = c_q q_0(x)$ were $q_0(x)$ is known but c_q is not known.

The idea of self-normalizing is based on

$$\begin{aligned}
\int f(x)p(x)dx &= \int \frac{f(x)p(x)}{q(x)} q(x)dx \qquad (6.32)\\
&= \frac{\int \frac{f(x)p(x)}{q(x)} q(x)dx}{\int \frac{p(x)}{q(x)} q(x)dx} \qquad (6.33)\\
&= \frac{\int \frac{f(x)p_0(x)}{q_0(x)} q(x)dx}{\int \frac{p_0(x)}{q_0(x)} q(x)dx} \qquad (6.34)\\
&= \frac{\int f(x)w(x)q(x)dx}{\int w(x)q(x)dx} \qquad (6.35)\\
&= \frac{E_q[f(x)w(x)]}{E_q[w(x)]} \qquad (6.36)
\end{aligned}$$

where $w(x) = p_0(x)/q_0(x)$ is a known function.

The self-normalized importance sampling algorithm is as follows. We generate samples

$\vec{X}_1, \cdots, \vec{X}_n$ according to the distribution $q(x)$. Our estimator for $\mu = \int f(x)p(x)dx$ is

$$\hat{\mu} = \frac{\sum_{i=1}^n f(\vec{X}_i) w(\vec{X}_i)}{\sum_{i=1}^n w(\vec{X}_i)} \tag{6.37}$$

Theorem 3 hypotheses *The estimator $\hat{\mu}$ converges to μ with probability* 1.

Proof: Note that

$$\hat{\mu} = \frac{\frac{1}{n}\sum_{i=1}^n f(\vec{X}_i) w(\vec{X}_i)}{\frac{1}{n}\sum_{i=1}^n w(\vec{X}_i)} = \frac{\frac{1}{n}\sum_{i=1}^n f(\vec{X}_i) \frac{c_q}{c_p}\frac{p(\vec{X}_i)}{q(\vec{X}_i)}}{\frac{1}{n}\sum_{i=1}^n \frac{c_q}{c_p}\frac{p(\vec{X}_i)}{q(\vec{X}_i)}} = \frac{\frac{1}{n}\sum_{i=1}^n f(\vec{X}_i) \frac{p(\vec{X}_i)}{q(\vec{X}_i)}}{\frac{1}{n}\sum_{i=1}^n \frac{p(\vec{X}_i)}{q(\vec{X}_i)}} \tag{6.38}$$

Now apply the strong law of large number to the numerator and denominator separately. Remember that $\vec{X}$ is sampled from $q(x)$, so the numerator converges to $\int f(x)p(x)dx = \mu$. The denominator converges to $\int p(x)dx = 1$. **QED**

It should be noted that the expected value of $\hat{\mu}$ is not exactly μ. The estimator is slightly biased.

To find a confidence interval for self normalized importance sampling we need to compute the variance of $\hat{\mu}$. We already did this using the delta method. In $\hat{\mu}$ the numerator is the sample mean for fw and the denominator is the sample mean for w. Plugging this into our result from the delta method we find that an estimator for the variance of $\hat{\mu}$ is

$$\frac{\sum_{i=1}^n w(\vec{X}_i)^2 (f(\vec{X}_i) - \hat{\mu})^2}{(\sum_{i=1}^n w(\vec{X}_i))^2} \tag{6.39}$$

If we let $w_i = w(\vec{X}_i)/\sum_{j=1}^n w(\vec{X}_j)$, then this is just

$$\sum_{i=1}^n w_i^2 (f(\vec{X}_i) - \hat{\mu})^2 \tag{6.40}$$

In ordinary Monte Carlo all of our samples contribute with equal weight. In importance sampling we give them different weights. The total weight of the weights is $\sum_{i=1}^n w_i$. It is possible that most of this weight is concentrated in a just a few weights. If this happens we expect the important sampling Monte Carlo will have large error. We might hope that when this happens our estimate of the variance will be large and so this will alert us to the problem. However, our estimate of the variance σ_q uses the same set of weights, so it may not be accruate when this happens.

Another way to check if we are getting grossly imbalanced weights is to compute an effective sample size. Consider the following toy problem. Let $w_1, \cdots, w_n$ be constants (not random).

Let $Z_1, \cdots, Z_n$ be i.i.d. random variables with common variance σ^2. An estimator for the mean of the Z_i is

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i Z_i}{\sum_{i=1}^n w_i} \tag{6.41}$$

The variance of $\hat{\mu}$ is

$$var(\hat{\mu}) = \sigma^2 \frac{\sum_{i=1}^n w_i^2}{(\sum_{i=1}^n w_i)^2} \tag{6.42}$$

Now define the number of effective samples n_e to be the number of independent samples we would need to get the same variance if we did not use the weights. In this case the variance is σ^2/n_e. So

$$n_e = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2} \tag{6.43}$$

As an example, suppose that k of the w_i equal 1 and the rest are zero. The a trivial calculation shows $n_e = k$.

Note that this definition of the effective sample size only involves the weights. It does not take f into account. One can also define an effective sample size that depends on f. See Owen.

6.3 Variance minimization and exponential tilting

Rather than consider all possbile choices for the sampling distribution $q(x)$, one strategy is to restrict the set of $q(x)$ we consider to some family of distributions and minimize the variance σ_q over this family. So we assume we have a family of distributions $p(x, \theta)$ where θ parameterizes the family. Here x is multidimensional and so is θ, but the dimensions need not be the same. We let θ_0 be the parameter value that corresponds to our nominal distribution. So $p(x) = p(x, \theta_0)$. The weighting function is

$$w(x, \theta) = \frac{p(x, \theta_0)}{p(x, \theta)} \tag{6.44}$$

The importance sampling algorithm is based on

$$\mu = E_{\theta_0}[f(\vec{X})] = \int f(x) p(x, \theta_0) dx = \int f(x) w(x, \theta) p(x, \theta) dx = E_\theta[f(\vec{X}) w(\vec{X}, \theta)] \tag{6.45}$$

The variance for this is

$$\sigma^2(\theta) = \int f(x)^2 w(x, \theta)^2 p(x, \theta) dx - \mu^2 \tag{6.46}$$

We want to minimize this as a function of θ. One approach would be for each different value of θ we run a MC simulation where we sample from $p(x, \theta)$ and use these samples to estimate $\sigma^2(\theta)$. This is quite expensive since it involves a simulation for every value of θ we need to consider.

A faster approach to search for the best θ is the following. Rewrite the variance as

$$\sigma^2(\theta) = \int f(x)^2 w(x, \theta) p(x, \theta_0) dx - \mu^2 = E_{\theta_0}[f(\vec{X})^2 w(\vec{X}, \theta)] - \mu^2 \tag{6.47}$$

Now we run a single MC simulation where we sample from $p(x, \theta_0)$. Let $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_m$ be the samples. We then use the following to estimate σ_θ:

$$\hat{\sigma_0(\theta)}^2 = \frac{1}{m} \sum_{i=1}^{m} f(\vec{X}_i)^2 w(\vec{X}_i, \theta) - \mu^2 \tag{6.48}$$

The subscript 0 on the estimator is to remind us that we used a sample from $p(x, \theta_0)$ rather than $p(x, \theta)$ in the estimation. We then use our favorite numerical optimization method for minimizing a function of several variables to find the minimum of this as a function of θ. Let θ^* be the optimal value.

Now we return to the original problem of estimating μ. We generate samples of $\vec{X}$ according to the distribution $p(x, \theta^*)$. We then let

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(\vec{X}_i) w(\vec{X}_i, \theta^*) \tag{6.49}$$

The variance of this estimator is $\sigma^2_{\theta^*}/n$. Our estimator for the variance $\sigma^2_{\theta^*}$ is

$$\hat{\sigma^2(\theta^*)} = \frac{1}{m} \sum_{i=1}^{m} f(\vec{X}_i)^2 w(\vec{X}_i, \theta)^2 \tag{6.50}$$

The above algorithm can fail completely if the distribution $p(x, \theta_0)$ is too far from a good sampling distribution. We illustrate this with an example.

Example: We return to an earlier example. Z is a standard normal RV and we want to compute $P(Z > 4)$. We take for our family the normal distributions with variance 1 and mean θ. So

$$p(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x - \theta)^2) \tag{6.51}$$

and the nominal density $p(x)$ is $p(x, 0)$.

MORE MORE MORE MORE

We can fix the problem above as follows. Instead of doing our single MC run by sampling from $p(x,\theta_0)$, we sample from $p(x,\theta_r)$ that θ_r, our "reference" θ, is our best guess for a good choice of θ. Rewrite the variance as

$$\begin{aligned} \sigma^2(\theta) &= \int f(x)^2 w(x,\theta)^2 p(x,\theta)dx - \mu^2 & (6.52) \\ &= \int f(x)^2 \frac{p(x,\theta_0)^2}{p(x,\theta)p(x,\theta_r)} p(x,\theta_r)dx - \mu^2 & (6.53) \end{aligned}$$

We then generate samples $\vec{X}_1,\cdots,\vec{X}_n$ from $p(x,\theta_r)$. Our estimator for the variance $\sigma(\theta)^2$ is then

$$\hat{\sigma_r^2}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\vec{X}_i)^2 p(\vec{X}_i,\theta_0)^2}{p(\vec{X}_i,\theta)p(\vec{X}_i,\theta_r)} - \hat{\mu}^2 \tag{6.54}$$

For several well-known classes of distributions the ratio $p(x)/q(x)$ takes a simple form. An exponential family is a family such that

$$p(x;\theta) = \exp((\eta(\theta), T(x)) - A(x) - C(\theta)) \tag{6.55}$$

for functions $\eta(\theta), T(X), A(x), C(\theta)$. The following are examples. A multivariate normal with a fixed covariance matrix is an exponential distribution where the means are the parameters. The possion distribution is an exponential family where the parameter is the usual (one-dimensional) λ. If we fixed the number of trials, then the binominal distribution is an exponential family with parameter p. The gamma distribution is also an exponential family. In many cases the weight function just reduces to $\exp((\theta,x))$. Even if $p(x)$ does not come from an exponential family we can still look for a proposal density of the form

$$q(x) = \frac{1}{Z(\theta)}\exp((\theta,x))p(x) \tag{6.56}$$

where $Z(\theta)$ is just the normalizing constant. Importance sampling in this case is often called exponential tilting.

Example Comment on network example.

Stop - Wed, Feb 17

6.4 Processes

Now suppose that instead of a random vector we have a stochastic process $X_1, X_2, X_3, \cdots$. We will let X stand for $X_1, X_2, X_3, \cdots$. We want to estimate the mean of a function of the process $\mu = f(X)$. It doesn't make sense to try to give a probability density for the full infinite process. Instead we specify it through conditional densities: $p_1(x_1), p_2(x_2|x_1), p_3(x_3|x_1, x_2), \cdots, p_n(x_n|x_1, x_2, \cdots, x_{n-1}), \cdots$. Note that it is immediate from the definition of conditional density that

$$\begin{aligned} p(x_1, x_2, \cdots, x_n) &= p_n(x_n|x_1, x_2, \cdots, x_{n-1}) p_{n-1}(x_{n-1}|x_1, x_2, \cdots, x_{n-2}) && (6.57) \\ &\cdots\ p_3(x_3|x_1, x_2) p_2(x_2|x_1) p_1(x_1) && (6.58) \end{aligned}$$

We specify the proposal density in the same way:

$$\begin{aligned} q(x_1, x_2, \cdots, x_n) &= q_n(x_n|x_1, x_2, \cdots, x_{n-1}) q_{n-1}(x_{n-1}|x_1, x_2, \cdots, x_{n-2}) && (6.59) \\ &\cdots\ q_3(x_3|x_1, x_2) q_2(x_2|x_1) q_1(x_1) && (6.60) \end{aligned}$$

So the likehood function is

$$w(x) = \prod_{n \geq 1} \frac{p_n(x_n|x_1, x_2, \cdots, x_{n-1})}{q_n(x_n|x_1, x_2, \cdots, x_{n-1})} \tag{6.61}$$

An infinite product raises convergence questions. But in applications f typically either depends on a fixed, finite number of the X_i or f depends on a finite but random number of the X_i. So suppose that f only depends on $X_1, \cdots, X_M$ where M may be random. To be more precise we assume that there is a random variable M taking values in the non-negative integers such that if we are given that $M = m$, then $f(X_1, X_2, \cdots)$ only depends on $X_1, \cdots, X_m$. So we can write

$$f(X_1, X_2, \cdots) = \sum_{m=1}^{\infty} 1_{M=m} \, f_m(X_1, \cdots, X_m) \tag{6.62}$$

We also assume that M is a stopping time. This means that the event $M = m$ only depends on $X_1, \cdots, X_m$. Now we define

$$w(x) = \sum_{m=1}^{\infty} 1_{M=m}(x_1, \cdots, x_m) \prod_{n=1}^{m} \frac{p_n(x_n|x_1, x_2, \cdots, x_{n-1})}{q_n(x_n|x_1, x_2, \cdots, x_{n-1})} \tag{6.63}$$

Example - random walk exit: This follows an example in Owens. Let ξ_i be an i.i.d. sequence of random variables. Let $X_0 = 0$ and

$$X_n = \sum_{i=1}^{n} \xi_i \tag{6.64}$$

In probability this is called a random walk. It starts at 0. Now fix an interval (a,b) with $0 \in (a,b)$. We run the run until it exits this interval and then ask whether it exited to the right or the left. So we let

$$M = \inf\{n : X_n \geq b \quad or \quad X_n < a\} \tag{6.65}$$

So the stopping condition is $X_M \geq b$ or $X_M \leq a$. Then we want to compute $\mu = P(X_M \geq b)$. We are particularily interested in the case where $E\xi_i < 0$. So the walk drifts to the left on average and the probability μ will be small if b is relatively large.

We take the walk to have steps with a normal distribution with variance 1 and mean -1. So the walk drifts to the left. We take $(a,b) = (-5,10)$. We run the walk until is exits this interval and want to compute the probability it exits to the right. This is a very small probability. So the ξ_i are independent normal random variables with variance 1 and mean -1.

The conditional densities that determine the nominal distribution are given by

$$p(x_n|x_1,\cdots,x_{n-1}) = p(x_n|x_{n-1}) = f_{\xi_n}(x_n - x_{n-1}) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x_n - x_{n-1} - \theta_0)^2) \tag{6.66}$$

In our example we take $\theta_0 = -1$. **MORE Explain how we sample this** A Monte Carlo simulation with no importance sampling with 10^6 samples produced no samples that exited to the right. So it gives the useless estimate $\hat{p} = 0$.

For the sampling distribution we take a random walk whose step distribution is normal with variance 1 and mean θ. So

$$q(x_n|x_1,\cdots,x_{n-1}) = q(x_n|x_{n-1}) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x_n - x_{n-1} - \theta)^2) \tag{6.67}$$

The weight factors are then

$$w_n(x_1,\cdots,x_n) = \exp((\theta_0 - \theta)(x_n - x_{n-1}) - \frac{1}{2}\theta_0^2 + \frac{1}{2}\theta^2) \tag{6.68}$$

With no idea of how to choose θ, we try $\theta = 0$ and find with 10^6 samples

$$p = 6.74 \times 10^{-10} \pm 0.33 \times 10^{-10} \tag{6.69}$$

The confidence intervals is rather large, so we do a longer run with 10^7 samples and find

$$p = 6.53 \times 10^{-10} \pm 0.098 \times 10^{-10} \tag{6.70}$$

The choice of $\theta = 0$ is far from optimal. More on this in a homework problem.