Chapter 3

Importance Sampling

Now we move on to schemes that do not produce exact (up to the floating point and periodicity issues mentioned in the last section) samples but that can be applied to far more complex sampling problems. The first family of algorithms of this kind that we will consider are called importance sampling methods. In their simplest form, these methods produce very simple unbiased estimators comprised of sums of independent random variables. More precisely, suppose your goal is to compute

$$\pi[f] = \int f(x)\pi(dx).$$

The simplest estimator is

$$\overline{f}_N = \frac{1}{N}\sum_{k=1}^{N} f\left(X^{(k)}\right)$$

where the $X^{(k)}$ are independent and all sampled from π. Recall that this estimator is unbiased and that we can compute its **rmse**. There are two possible drawbacks to this algorithm. The first is that it may be very costly or impossible to generate independent samples from π. The second difficultly is that for may problems the **rmse** may be unacceptably large so that a reasonable estimate requires very large N. Now suppose that $\tilde{\pi}$ is some other

distribution that we can sample. We can then try to construct the estimator

$$\widetilde{f}_N = \frac{1}{N}\sum_{k=1}^{N} f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$$

where the $Y^{(k)}$ are independent samples from $\tilde{\pi}$. It will often be convenient to write this estimator as

$$\widetilde{f}_N = \sum_{k=1}^{N} f\left(Y^{(k)}\right) W^{(k)}$$

where

$$W^{(k)} = \frac{1}{N}\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}.$$

We will assume that if $\mathrm{supp}(f)$ is the set of points x for which $f(x) \neq 0$, then

$$\mathrm{supp}(f\pi) \subset \mathrm{supp}(f\tilde{\pi}).$$

If the random variable the random variable $f(X)$ was integrable then $f(Y^{(k)})\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ is also integrable and

$$\mathbf{E}\left[f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right] = \int f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\tilde{\pi}(dy) = \int f(y)\pi(dy) = \mathbf{E}\left[f(X)\right]$$

and the estimator $\widetilde{f}_N$ is unbiased.

As we did for $\overline{f}_N$, if the random variables $f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ have finite variance we can easily compute that

$$\mathbf{rmse}(\widetilde{f}_N) = \frac{\sqrt{\mathbf{var}\left(f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right)}}{\sqrt{N}}.$$

Since, for any random variable X with finite variance we have

$$\mathbf{var}(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2$$

and the mean of $f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}$ is $\pi[f]$,

$$\begin{aligned}\mathbf{var}\left(f\left(Y^{(k)}\right)\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right) &= \int\left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2\tilde{\pi}(dy) - (\pi[f])^2 \\ &= \int (f(x))^2\frac{\pi(x)}{\tilde{\pi}(x)}\pi(dx) - (\pi[f])^2\end{aligned}$$

Example 7. *Consider importance sampling when $f = 1$, π is $\mathcal{N}(0,1)$, and $\tilde{\pi}$ is $\mathcal{N}(0,\sigma^2)$. The error is*

$$\mathbf{rmse}\left(\widetilde{f}_N\right) = \frac{1}{\sqrt{N}}\sqrt{\frac{\sigma}{\sqrt{2\pi}}\int e^{-\left(1-\frac{1}{2\sigma^2}\right)x^2}dx - 1}.$$

As soon as σ^2 becomes less than 1/2, the error becomes infinite, illustrating the fact that the tails of the reference density should, in general, be heavier than the tails of the target density (or at least of $|f|$ times the target density).

3.1 Optimal importance sampling

The goal in selecting an importance sampling reference density is to choose the reference density that results in an estimator $\widetilde{f}_N$ that has lower variance than $\overline{f}_N$. From our last computation we can focus our efforts on choosing a $\tilde{\pi}$ for which

$$\int \left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2 \tilde{\pi}(dy)$$

is as low as possible. The optimal choice of $\tilde{\pi}$ can easily be identified. Indeed, Jensen's inequality implies that

$$\int \left(f(y)\frac{\pi(y)}{\tilde{\pi}(y)}\right)^2 \tilde{\pi}(dy) \geq \left(\int |f(y)|\pi(dy)\right)^2$$

and this lower bound is achieved by

$$\tilde{\pi}(x) = \frac{|f(x)|\pi(x)}{\int |f(y)|\pi(dy)}.$$

Of course it is extremely unlikely that one can sample from (or even evaluate) this optimal density even if you could sample from π. You'll notice that the optimal density involves a computation (the normalization constant) very similar to the original problem. The is an example of one of the central tenants of Monte Carlo: the more you know about the answer the better the solution you can design. In most practical situations one applies intuition about the problem at hand to design a reasonable $\tilde{\pi}$. However there are

some interesting situations in which one can derive mathematically justifiable choices of reference density. In the chapter on rare event simulation I provide one such example.

Exercise 20. *Use samples from* $\mathcal{N}(m, \sigma^2)$ *to estimate* $\mathbf{P}[X > 2]$ *for* $X \sim \mathcal{N}(0, 1)$ *using importance sampling. By comparing the variances of the estimators for different* m *and* σ*, draw conclusions about the values of* m *and* σ *that yield the best estimators.*

3.2 Normalization constants and an alternative estimator

Examining the estimator $\widetilde{f}_N$ more closely, you'll notice that to use it we need to evaluate the ratio $\pi/\tilde{\pi}$. In many applications this is only possible up to an unknown multiplicative constant. In this section we'll describe an importance sampling strategy for estimating this constant and use it to build an alternative importance sampling estimator that only requires that we can evaluate $\pi/\tilde{\pi}$ up to the unknown constant.

Given a non-negative and integrable function p, the normalization constant (or partition function) is the value $\mathcal{Z}_p = \int p(x)dx$. The problem of estimating a normalization constant arrises in a wide range of applications. In statistical mechanics one is often interested in computing the normalization constant for a family of densities indexed by some parameter θ, i.e.

$$\mathcal{Z}_{p_\theta} = \int p_\theta(dx).$$

Here, for each The function

$$F(\theta) = -\log \mathcal{Z}_{p_\theta}$$

is called a free energy. The marginal distribution

$$\pi(x) = \int \pi(x, y)dy$$

of the x variables determined by some joint distribution $\pi(x, y)$ is another quantity of frequent interest in statistical mechanics where $-\log \pi(x)$ is again referred to a free energy.

We have already seen that in Bayesian statistics, given a prior distribution $\pi(\theta)$ on the parameters and a data likelihood $\pi(y \,|\, \theta)$, a common goal is to sample from the posterior distribution

$$\pi(\theta \,|\, y) = \frac{\pi(y \,|\, \theta)\pi(\theta)}{\pi(y)}$$

where θ is a collection of parameters and y is a realization of the data. The normalization constant in this expression, $\pi(y)$ (the marginal density of the data), is called the model evidence and is used to distinguish between putative statistical models.

We'll now describe a very simple estimator of the ratio $\mathcal{Z}_p/\mathcal{Z}_q$ of the normalization constants of two non-negative, integrable functions p and q. Suppose that you can sample from the density

$$\tilde{\pi} = \frac{q(x)}{\mathcal{Z}_q}$$

and let $\pi = p/\mathcal{Z}_p$. Our importance sampling estimator for averages with respect to π with reference density $\tilde{\pi}$ satisfies

$$\widetilde{1}_N \longrightarrow \pi[1] = 1.$$

Multiplying both sides of this expression by $\mathcal{Z}_p/\mathcal{Z}_q$ we obtain the estimator

$$\frac{1}{N}\sum_{k=1}^{N} \frac{p(Y^{(k)})}{q(Y^{(k)})} \longrightarrow \frac{\mathcal{Z}_p}{\mathcal{Z}_q}.$$

Exercise 21. *Write a routine that uses $\mathcal{N}(0,1)$ samples to estimate the normalization constant for the density proportional to $e^{-|x|^3}$.*

Based on the approximation of the ratio of normalization constants above, a reasonable modification to the standard importance sampling estimator to deal with unknown normalization constants is

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = \frac{1}{N}\sum_{k=1}^{N} \frac{f(Y^{(k)})\frac{p(Y^{(k)})}{q(Y^{(k)})}}{\frac{1}{N}\sum_{\ell=1}^{N}\frac{p(Y^{(\ell)})}{q(Y^{(\ell)})}} = \sum_{k=1}^{N} f\left(Y^{(k)}\right) W^{(k)}$$

where now

$$W^{(k)} = \frac{\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}}{\sum_{\ell=1}^{N} \frac{\pi(Y^{(\ell)})}{\tilde{\pi}(Y^{(\ell)})}}.$$

The WLLN implies that the numerator in the rightmost expression in the last display converges to $\pi[f]$ and the denominator converges to 1. Therefore $\widetilde{f}_N/\widetilde{1}_N$ converges to $\pi[f]$. However, inspecting the mean we see that in general

$$\mathbf{E}\left[\frac{\widetilde{f}_N}{\widetilde{1}_N}\right] \neq \frac{\mathbf{E}\left[\sum_{k=1}^{N} f\left(Y^{(k)}\right) \frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right]}{\mathbf{E}\left[\sum_{k=1}^{N} \frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right]} = \pi[f],$$

i.e. $\widetilde{f}_N/\widetilde{1}_N$ is a biased estimator.

We can estimate the size of this bias by writing

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = h\left(\widetilde{f}_N, \widetilde{1}_N\right)$$

where $h(x, y) = x/y$ and taylor expanding around the means $(\pi[f], 1)$ (in this context this technique is refered to as the delta method). Let

$$\gamma(t) = \begin{pmatrix} \pi[f] \\ 1 \end{pmatrix} (1-t) + \begin{pmatrix} \widetilde{f}_N \\ \widetilde{1}_N \end{pmatrix} t, \qquad t \in [0,1].$$

We have

$$\begin{aligned} h\left(\widetilde{f}_N, \widetilde{1}_N\right) &= h(\gamma(0)) + \frac{d}{dt}\left(h(\gamma(t))\right)\Big|_{t=0} + \int_0^1 (1-s)\frac{d^2}{ds^2}\left(h(\gamma(s))\right) ds \\ &= h(\pi[f],1) + (\widetilde{f}_N - \pi[f])\partial_x h(\pi[f],1) + (\widetilde{1}_N - \pi[f])\partial_y h(\pi[f],1) \\ &\quad + (\widetilde{f}_N - \pi[f])^2 \int_0^1 (1-s)\partial_x^2 h(\gamma(s))ds \\ &\quad + (\widetilde{1}_N - 1)^2 \int_0^1 (1-s)\partial_y^2 h(\gamma(s))ds \\ &\quad + (\widetilde{f}_N - \pi[f])(\widetilde{1}_N - 1) \int_0^1 2(1-s)\partial_{xy} h(\gamma(s))ds. \end{aligned}$$

Since $\partial_x h(\pi[f], 1) = 1$ and $\partial_y h(\pi[f], 1) = -\pi[f]$, we find that the expectations of the second and third terms in the last expression vanish. We'll simply

pretend that the three integrals appearing in the formula are bounded and recall that

$$\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)^2\right] = \mathcal{O}\left(\frac{1}{N}\right), \quad \mathbf{E}\left[\left(\widetilde{1}_N - 1\right)^2\right] = \mathcal{O}\left(\frac{1}{N}\right),$$

and, by the Cauchy-Schwartz inequality,

$$\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)\left(\widetilde{1}_N - 1\right)\right] \leq \sqrt{\mathbf{E}\left[\left(\widetilde{f}_N - \pi[f]\right)^2\right]}\sqrt{\mathbf{E}\left[\left(\widetilde{1}_N - 1\right)^2\right]}.$$

These calculations yield the basic conclusion that the bias of $\widetilde{f}_N/\widetilde{1}_N$ is smaller than the standard deviation of $\widetilde{f}_N$ (which are of order $N^{-1/2}$) and should therefore not trouble us too much in many applications.

To be more confident that the bias is small we would need to know more about the likelihood of very small values of $\widetilde{1}_N$ (which will result in large values for the integral terms we have ignored). Cramer's theorem tells us that the probability that $\widetilde{1}_N < \delta$ for any $\delta < 1$ is exponentially small in N, but this is not enough since, for example, if the event $\widetilde{1}_N = 0$ occurs with positive probability then the bias is infinite. At the cost of an additional small bias, the estimator can be modified so that these issues are avoided.

Our primary motivation for introducing the estimator $\widetilde{f}_N/\widetilde{1}_N$ was that the densities π and $\tilde{\pi}$ might only be known up to a multiplicative constant making it impossible to assemble $\widetilde{f}$. Is there a reason to prefer the biased estimator $\widetilde{f}_N/\widetilde{1}_N$ when unknown normalization constants are not an issue? Let's consider the mean squared error

$$\mathbf{rmse}^2\left(\frac{\widetilde{f}_N}{\widetilde{1}_N}\right) = \mathbf{E}\left[\left(\frac{\widetilde{f}_N}{\widetilde{1}_N} - \pi[f]\right)^2\right].$$

Using the same expansion of h that we used above, we obtain

$$\mathbf{rmse}^2\left(\frac{\widetilde{f}_N}{\widetilde{1}_N}\right) = \mathbf{E}\left[\left((\widetilde{f}_N - \pi[f]) - \pi[f](\widetilde{1}_N - 1) + \mathcal{O}\left(N^{-1}\right)\right)^2\right].$$

Since the first two terms in the last display are $\mathcal{O}\left(N^{-1/2}\right)$ we'll neglect the

$\mathcal{O}(N^{-1})$ terms to obtain

$$\begin{aligned}\mathbf{rmse}^2\left(\frac{\widetilde{f}_N}{\widetilde{1}_N}\right) &\approx \mathbf{E}\left[\left(\widetilde{f}_N - \pi[f] - \pi[f](\widetilde{1}_N - 1)\right)^2\right] \\ &= \mathbf{var}\left(\widetilde{f}_N\right) + \frac{(\pi[f])^2}{N}\,\mathbf{var}\left(\frac{\pi(Y)}{\tilde{\pi}(Y)}\right) \\ &\qquad - \frac{2\pi[f]}{N}\,\mathbf{cov}\left(f(Y)\,\frac{\pi(Y)}{\tilde{\pi}(Y)}, \frac{\pi(Y)}{\tilde{\pi}(Y)}\right)\end{aligned}$$

where Y is distributed according to $\tilde{\pi}$.

For certain choices of f the covariance in the last display will be small or negative (e.g. if $f = \tilde{\pi}/\pi$) and the mean squared error of $\widetilde{f}_N/\widetilde{1}_N$ will be larger than that for $\widetilde{f}_N$. However, in many cases this covariance will be large and $\widetilde{f}_N/\widetilde{1}_N$ will have smaller error. As a dramatic example, suppose that f is nearly constant. Then $\pi[f]$ is approximately equal to this constant and the **rmse** of $\widetilde{f}_N/\widetilde{1}_N$ nearly vanishes.

Exercise 22. *Repeat the last exercise for the importance sampling estimator $\widetilde{f}_N/\widetilde{1}_N$ instead of $\widetilde{f}_N$. Which of these two estimators do you prefer? Does the answer depend on m and σ?*

3.3 Importance sampling in high dimensions

As a general rule, the need for a good approximation of the optimal importance sampling estimator becomes more acute in high dimensions. As a general measure of the quality of a reference density within the context of importance sampling one can consider the variance of the importance sampling weights, i.e.

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \mathbf{var}\left(\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right).$$

We use the symbol χ^2 because $\mathbf{var}\left(\frac{\pi(Y^{(k)})}{\tilde{\pi}(Y^{(k)})}\right)$ is Pearson's χ^2-divergence between π and $\tilde{\pi}$. Much like the relative entropy, $\chi^2(\pi \,\|\, \tilde{\pi})$ is a very strong measure of the distance between π and $\tilde{\pi}$ (though it is not a distance) in the

sense that it bounds the total variation distance between π and $\tilde{\pi}$. Indeed, by Jensen's inequality,

$$\begin{aligned}\chi^2(\pi \,\|\, \tilde{\pi}) &= \int \left| \frac{\pi(y)}{\tilde{\pi}(y)} - 1 \right|^2 \tilde{\pi}(dy) \\ &\geq \left(\int \left| \frac{\pi(y)}{\tilde{\pi}(y)} - 1 \right| \tilde{\pi}(dy) \right)^2 \\ &= 4\|\pi - \tilde{\pi}\|_{\mathrm{TV}}^2\end{aligned}$$

Suppose that the goal is to construct an estimator of $\pi[f]$ with **rmse** equal to δ. In the last section we estimated the error of the estimator $\widetilde{f}_N/\widetilde{1}_N$ as

$$\mathbf{rmse}^2(\widetilde{f}_N/\widetilde{1}_N) \approx \mathbf{E}\left[\left(\widetilde{f}_N - \pi[f] - \pi[f](\widetilde{1}_N - 1) \right)^2 \right]$$

where Y is a random variable distributed according to $\tilde{\pi}$. The expression on the right hand side of the last display is equal to

$$\frac{1}{N}\mathbf{E}\left[(f(Y) - \pi[f])^2 \left(\frac{\pi(Y)}{\tilde{\pi}(Y)} \right)^2 \right].$$

If we make the (usually very severe) assumption that the variables $f(Y)$ and $\pi(Y)/\tilde{\pi}(Y)$ are independent then we find that

$$\mathbf{rmse}^2(\widetilde{f}_N/\widetilde{1}_N) \approx \frac{1}{N}\mathbf{var}\left(f(X)\right)\left(1 + \chi^2(\pi \,\|\, \tilde{\pi})\right)$$

where X is distributed according to π. On the other hand, for the standard estimator using M independent samples from π, we know that

$$\mathbf{rmse}^2\left(\overline{f}_M\right) = \frac{\mathbf{var}\left(f(X)\right)}{M}.$$

As a consequence, we can make a very rough estimate the number of samples required by the estimator $\overline{f}$ to achieve the same accuracy as $\widetilde{f}_N/\widetilde{1}_N$ as

$$M \approx \frac{N}{1 + \chi^2(\pi \,\|\, \tilde{\pi})}.$$

The term on the right hand side of the last display is referred to as the effective sample size, ess_N, of the importance sampling estimator $\widetilde{f}_N/\widetilde{1}_N$. It gives a rough estimate of the number of independent samples from π that would be of similar statistical quality to the N weighted samples generated in importance sampling. By this measure, when $\chi^2(\pi \,\|\, \tilde{\pi})$ is large we expect importance sampling to yield poor results.

As the following example illustrates, unless the reference distribution is chosen very carefully we expect resampling to perform poorly in high dimensions.

Example 8. *Consider importance sampling the d dimensional Gaussian target density $\pi = \mathcal{N}(m, C)$ using the reference density $\tilde{\pi} = \mathcal{N}(m, (1-\alpha)C)$ for some $\alpha \in (0,1)$. Under π or $\tilde{\pi}$, the density of the function $V(x) = (x-m)^T C^{-1}(x-m)$ is independent of the choice of m and C (it is a chi-squared random variable with d degrees of freedom). As can be easily computed, the variance of the importance weights is*

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \int \left(\frac{\pi(x)}{\tilde{\pi}(x)}\right) \pi(dx) - 1 = \frac{1}{(1-\alpha^2)^{\frac{d}{2}}} - 1.$$

For fixed α, $\chi^2(\pi \,\|\, \tilde{\pi})$ grows exponentially with d.

We can further illustrate the failure of importance sampling in high dimensions by considering the case in which π is the density of d i.i.d. random variables, i.e.

$$\pi(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i)$$

for some density p of a single variable. One might encounter a density of this kind when assimilating d observations of an experiment and assuming that the error in the various observations are independent. Let's suppose that you want to use a reference density of the same form,

$$\tilde{\pi}(y_1, y_2, \ldots, y_d) = \prod_{i=1}^{d} q(y_i).$$

Then the independence of the $Y_j^{(k)}$ yeilds

$$\chi^2(\pi \,\|\, \tilde{\pi}) = \sqrt{\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right]^d} - 1.$$

When $\tilde{\pi} \neq \pi$ we will have that

$$\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right] > \mathbf{E}\left[\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right]^2 = 1$$

and $\chi^2(\pi \,\|\, \tilde{\pi})$ will increase exponentially with d.

The preceding calculation should give pause to anyone seeking to use importance sampling in very high dimensional systems. In fact, it suggests that, when using importance sampling at least, Monte Carlo suffers from exactly the same problem as the deterministic integration schemes that we discussed earlier. It is important to keep in mind, however, that the case of independent, identically distributed components is a very special one. In fact typical high dimensional problems almost always exhibit low dimensional structure. This means that high dimensional densities encountered in typical sampling applications tend to concentrate on a lower dimensional subspace. Given our demonstration above of the dangers of importance sampling in high dimensions, one might take comfort in the observation of lower dimensional structure in high dimensional sampling problems. But often one has little or no information about the lower dimensional structure of the distribution, in which case that structure actually makes the problem much more difficult than the independent identically distributed components setting. One is forced to use a reference density $\tilde{\pi}$ for which the variance of $\pi/\tilde{\pi}$ may be extremely high.

There are, however, important situations in which importance sampling can be used to great advantage in high dimensions. The key to success, of course, is choosing a reference density sufficiently close to the optimal importance sampling density. In the next example, the marginal distributions, $\pi(x_j)$, change with dimension, and by choosing a reference density that respects the correct scaling with dimension we can ensure $\chi^2(\pi \,\|\, \tilde{\pi})$ is bounded for all d.

Example 9. *As in the discussion above, assume that*

$$\pi(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i) \quad \textit{and} \quad \tilde{\pi}(y_1, y_2, \ldots, y_d) = \prod_{i=1}^{d} q(Y_i).$$

Suppose that $p = \mathcal{N}(0, d^{-1})$ *and that* $q = \mathcal{N}(d^{-1}, d^{-1})$. *Then*

$$\begin{aligned}\mathbf{E}\left[\left(\frac{p(Y_1^{(1)})}{q(Y_1^{(1)})}\right)^2\right] &= \frac{\sqrt{d}}{\sqrt{2\pi}}\int e^{-dy^2+d(y-d^{-1})^2}e^{-\frac{d(y-d^{-1})^2}{2}}dy \\ &= \frac{\sqrt{d}e^{\frac{1}{d}}}{\sqrt{2\pi}}\int e^{-\frac{d(y+d^{-1})^2}{2}}dy \\ &= e^{\frac{1}{d}}\end{aligned}$$

so that $\chi^2(\pi \,\|\, \tilde{\pi})$ *is stable as* $d \to \infty$. *This example is related to importance sampling for diffusions which we'll return to in Part II.*

3.4 Sequential importance sampling and resampling

Occasionally, the structure of a problem allows one to break a high dimensional sampling problem into manageable pieces. We can always decompose a multidimensional density π as

$$\pi(x_{1:d}) = \pi(x_1)\prod_{n=2}^{d}\pi(x_n \,|\, x_{1:n-1}) \tag{3.1}$$

where we have introduced the more compact notation

$$x_{m:n} = (x_m, x_{m+1}, \ldots, x_n) \quad \text{for} \quad m \le n.$$

To see this just recall that

$$\pi(x_n \,|\, x_{1:n-1}) = \frac{\pi(x_{1:n})}{\pi(x_{1:n-1})}.$$

The decomposition in (3.1) suggests a sampling strategy for π: first sample X_1 from $\pi(x_1)$ and then, at step n given the components $X_{1:n-1}$ generated so far, generate X_n from $\pi(x_n \,|\, X_{1:n-1})$.

Example 10. *Consider generating a simple random walk of length* d *on a periodic lattice* $\mathbb{Z}_L^2 = \{0, 1, \ldots, L-1\} \times \{0, 1, \ldots, L-1\}$. *A walk on the lattice*

of length d (an element of SRW(d)) is just a chain of states $x_{1:d}$ with x_{s+1} a neighbor on the lattice of x_s (denoted here $x_{s+1} \leftrightarrow x_s$) for all $s < d$. The density for a simple random walk on the lattice is

$$\pi(x_{1:d}) = \frac{1}{\mathcal{Z}_d} \begin{cases} 1 & \text{if } x_{1:d} \text{ is a walk on the lattice} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{Z}_d = L^2 4^{d-1}$ is the normalization constant.

In this case, the marginal

$$\pi(x_{1:n}) = \sum_{x_{n+1:d}} \pi(x_{1:d})$$

is just the density for the simple random walk of length n. We can easily compute that, if $x_{1:n-1}$ is a walk on the lattice,

$$\pi(x_n \,|\, x_{1:n-1}) = \frac{\mathcal{Z}_n}{\mathcal{Z}_{n-1}} = \begin{cases} 1/4 & \text{if } x_{1:n} \in \text{SRW}(n) \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, we can sample the walk by choosing an initial point, X_1, uniformly and then, at step $n > 1$, picking a neighbor of X_{n-1} uniformly.

Of course it is unlikely that we will know enough about π (and it's marginal and conditional densities) to carry out this procedure. You can check that even in the simple random walk example, if I permute the indices in the decomposition (3.1) it is already harder to see how to use the decomposition to sample from π. Fortunately, this decomposition strategy can sometimes be salvaged with the help of importance sampling. In fact, though the sequential importance sampling and resampling strategy that we develop in this section was originally introduced to sequentially construct samples from high dimensional distributions, it is now used in applications ranging from the simulation of rare events to online data assimilation.

3.4.1 Sequential importance sampling

We will form a reference density $\tilde{\pi}$ for use in importance sampling of π by replacing the various terms in the decomposition (3.1) by approximations. In

more detail, given a density $\tilde{\pi}_1(x_1)$ and conditional densities $q_n(x_n \,|\, x_{1:n-1})$ define the sequence of reference densities

$$\tilde{\pi}_n(x_{1:n}) = \tilde{\pi}_1(x_1) \prod_{\ell=2}^{n} q_\ell(x_\ell \,|\, x_{1:\ell-1}) \tag{3.2}$$

and set $\tilde{\pi} = \tilde{\pi}_d$. We will assume that one can sample from $\tilde{\pi}_1$ and the conditional densities q_n and evaluate them up to a normalization constant. Notice that the sequence of densities $\tilde{\pi}_n$ is closed under marginalization in the sense that, for $m \leq n$,

$$\tilde{\pi}_n(x_{1:m}) = \tilde{\pi}_m(x_{1:m}).$$

The normalized importance sampling estimator for an average with respect to π using reference density $\tilde{\pi}$ is

$$\frac{\widetilde{f}_N}{\widetilde{1}_N} = \sum_{k=1}^{N} f\left(Y_{1:d}^{(k)}\right) W_d^{(k)}$$

where $Y_{1:d}^{(k)}$ are samples from $\tilde{\pi}$ and

$$W_d^{(k)} = \frac{\pi(Y_{1:d}^{(k)})/\tilde{\pi}(Y_{1:d}^{(k)})}{\sum_{\ell=1}^{N} \pi(Y_{1:d}^{(\ell)})/\tilde{\pi}(Y_{1:d}^{(\ell)})}.$$

Comparing to the decomposition of π in (3.1), it is clear that our importance sampling estimator will perform well when

$$\tilde{\pi}_1(x_1) \approx \pi(x_1) \quad \text{and} \quad q_n(x_n \,|\, x_{1:n-1}) \approx \pi(x_n \,|\, x_{1:n-1}).$$

Our immediate goal is to represent this estimator in terms of a recursion on dimension. The generation of the samples $Y_{1:n}^{(k)}$ from $\tilde{\pi}_n$ is naturally carried out by recursion: given a sample $Y_{1:n-1}^{(k)}$ from $\tilde{\pi}_{n-1}$ we can build a sample $Y_{1:n}^{(k)}$ sampled from $\tilde{\pi}_n$ by first sampling $Y_n^{(k)}$ from $q_n(x_n \,|\, Y_{1:n-1}^{(k)})$ and then setting $Y_{1:n}^{(k)} = (Y_{1:n-1}^{(k)}, Y_n^{(k)})$.

We now find a recursion for the weights $W_n^{(k)}$. To do this we need to introducing a sequence of densities

$$\pi_n(x_{1:n})$$

for $n = 1, 2, \ldots, d$ with $\pi_d = \pi$. Note that we are not assuming that the marginal under the target density of the first n variables, $\pi(x_{1:n})$, is equal to $\pi_n(x_{1:n})$. We do assume that one can evaluate the ratios π_n/π_{n-1} up to an unknown normalization constant, but not that you can sample directly from π_n. We will characterize a recursion for importance sampling estimators for each of the π_n using reference density $\tilde{\pi}_n$ and built off of the estimator for π_{n-1}. This recursion has no immediate utility as the the resulting estimator for π is exactly the usual normalized importance sampling estimator for averages with respect to π using reference density $\tilde{\pi}$. It will become very useful later in this section when we introduce the notion of resampling.

The normalized importance sampling estimator for an average with respect to π_n using the reference density $\tilde{\pi}_n$ would use weights

$$W_n^{(k)} = \frac{\pi_n(Y_{1:n}^{(k)})/\tilde{\pi}_n(Y_{1:n}^{(k)})}{\sum_{\ell=1}^N \pi_n(Y_{1:n}^{(\ell)})/\tilde{\pi}_n(Y_{1:n}^{(\ell)})}$$

where $Y_{1:n}^{(k)}$ is drawn from $\tilde{\pi}_n$. Now observe that

$$\frac{\pi_n(x_{1:n})}{\tilde{\pi}_n(x_{1:n})} = \frac{\pi_{n-1}(x_{1:n-1})}{\tilde{\pi}_{n-1}(x_{1:n-1})} w_n(x_{1:n})$$

where we have defined the function

$$w_n(x_{1:n}) = \frac{\pi_n(x_{1:n})}{\pi_{n-1}(x_{1:n-1})\, q_n(x_n \,|\, x_{1:n-1})}.$$

As a consequence we obtain the recursion

$$W_n^{(k)} = \frac{W_{n-1}^{(k)}\, w_n(Y_{1:n}^{(k)})}{\sum_{\ell=1}^N W_{n-1}^{(\ell)}\, w_n(Y_{1:n}^{(\ell)})}.$$

It is important to note that, under our assumptions, one can evaluate w_n up to a multiplicative constant that cancels in the normalization of the $W_n^{(k)}$.

Example 11. *A chain $x_{1:d}$ of states $x_s \in \mathbb{Z}_L^2$ with $x_{s+1} \leftrightarrow x_s$ for $s < d$ and $x_t \neq x_s$ for all $s, t \leq d$ is called a self avoiding walk of length d (an element of SAW(d)). Imagine sampling from the density $\pi(x_{1:d})$ defined by*

$$\pi(x_{1:d}) = \frac{1}{\mathcal{Z}_d} \begin{cases} 1 & \text{if } x_{1:d} \in SAW(d) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{Z}_d$ is the (now unknown) normalization constant. One could imagine using importance sampling directly to compute averages with respect to π using as a reference density, the uniform measure on chains satisfying $x_{s+1} \leftrightarrow x_s$ for all $s < d$ (we don't know the normalizing constants so we'd have to use $\widetilde{f}_N/\widetilde{1}_N$). But it is very unlikely that a chain from this reference density would satisfy $x_s \neq x_t$ for $s, t \leq d$ and most of our effort would be spent generating samples that would later be assigned weight 0.

We can use sequential importance sampling instead. First, define π_n for $n \leq d$ as the uniform measure on SAW(n). In contrast with the simple random walk problem in the last example, the distribution π_{n-1} is not quite the marginal density $\pi(x_{1:n}) = \sum_{x_{n+1:d}} \pi(x_{1:d})$ of the first n states in a chain of length d drawn from π. To see this, observe that for $n < d$, an element $x_{1:n} \in SAW(n)$ for which there are no neighbors of x_n that have not already been visited, has $\pi_n(x_{1:n}) > 0$, but $\pi(x_{1:n}) = 0$. But our sequential importance sampling framework does not require this.

The factor w_n needed to update the weights can be written as

$$w_n(x_{1:n}) = \frac{\pi_n(x_n \,|\, x_{1:n-1})\, \pi_n(x_{1:n-1})}{q_n(x_n \,|\, x_{1:n-1})\, \pi_{n-1}(x_{1:n-1})}$$

Moreover,

$$\begin{aligned} \pi_n(x_{1:n-1}) &= \sum_{x_n} \pi_n(x_{1:n}) \\ &= \begin{cases} \frac{m_n(x_{1:n-1})}{\mathcal{Z}_n} & x_{1:n-1} \in SAW(n-1) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where

$$m_n(x_{1:n-1}) = \left|\{x_n : x_{1:n} \in SAW(n)\}\right|.$$

This implies that, as long as $m_n(x_{1:n-1}) > 0$,

$$\pi_n(x_n \,|\, x_{1:n-1}) = \frac{\pi_n(x_{1:n})}{\pi_{n-1}(x_{1:n-1})} = \begin{cases} \frac{1}{m_n} & \text{if } x_{1:n} \in SAW(n) \\ 0 & \text{otherwise.} \end{cases}$$

When $m_n(x_{1:n-1}) > 0$, this conditional distribution is easy enough to sample from and makes a natural choice for q_n. Given a chain in $SAW(n-1)$, one

simply chooses x_n from among those neighbors of x_{n-1} that have not yet been reached by the chain. Chains for which this is not possible ($m_n = 0$) will receive 0 weight and can be discarded. Indeed, having made this choice for q_n, the weight factors become

$$w_n(x_{1:n}) = \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} = \frac{m_n(x_{1:n-1})\mathcal{Z}_{n-1}}{\mathcal{Z}_n}.$$

Finally, notice in addition that for this q_n, the ratio of successive normalization constants $\mathcal{Z}_n/\mathcal{Z}_{n-1}$ can be written

$$\begin{aligned}\frac{\mathcal{Z}_n}{\mathcal{Z}_{n-1}} &= \sum_{x_{1:n}} m_n(x_{1:n-1}) q_n(x_n \,|\, x_{1:n-1}) \pi_{n-1}(x_{1:n-1}) \\ &= \sum_{x_{1:n-1}} m_n(x_{1:n-1}) \pi_{n-1}(x_{1:n-1})\end{aligned}$$

3.4.2 Sequential importance sampling with resampling

We have already demonstrated the difficulties suffered by importance sampling in high dimensions. If used as just described this scheme too should be expected to fail for large d unless $\tilde{\pi}$ happens to be a very good approximation of π. The utility of this recursive form of importance sampling is only fully exploited when we combine it with resampling. That is, instead of carrying samples with very low weight we replace low weight samples with copies of high weight samples in a statistically consistent manner. More precisely, assuming that at step n we have a weighted ensemble of N_n samples, $\left\{W_n^{(k)}, X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$, approximately drawn from π_n, in the sense that

$$\sum_{k=1}^{N_n} W_n^{(k)} f\left(X_{1:n}^{(k)}\right) \approx \int f(x_{1:n}) \pi_n(dx_{1:n})$$

for any test function f, we

1. Resample the weighted ensemble $\left\{W_n^{(k)}, X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$ to obtain a uniformly weighted ensemble $\left\{1/N, Y_{1:n}^{(k)}\right\}_{k=1}^{N_{n+1}}$ approximately drawn from π_n in

the sense that

$$\frac{1}{N}\sum_{k=1}^{N_{n+1}} f\left(Y_{1:n}^{(k)}\right) \approx \int f(x_{1:n})\pi_n(dx_{1:n})$$

for any test function f.

2. For $k = 1, 2, \ldots, N_{n+1}$ generate $X_{n+1}^{(k)}$ from $q_{n+1}(x_{n+1} \,|\, Y_{1:n}^{(k)})$ and set

$$X_{1:n+1}^{(k)} = \left(Y_{1:n}^{(k)}, X_{n+1}^{(k)}\right).$$

3. Compute the weights

$$W_{n+1}^{(k)} = \frac{w_{n+1}(X_{1:n+1}^{(k)})}{\sum_{\ell=1}^{N_{n+1}} w_{n+1}(X_{1:n+1}^{(\ell)})}.$$

The number of samples after resampling, N_n, need not be deterministic but will, in general, be close to the user specified value N. The basic technique used to generate the unweighted ensemble is to make multiple copies of samples in the weighted ensemble with large weights and to discard samples in the weighted ensemble with small weights. Note the absence of the weights $W_n^{(k)}$ in the formula for the new weights $W_{n+1}^{(k)}$. The previous weights are already accounted for in the duplication or removal of samples from the $\left\{X_{1:n}^{(k)}\right\}_{k=1}^{N_n}$ ensemble. This resampling is done at each step in the recursion with the goal being to devote our computational resources only to those samples with a reasonable chance of contributing to the final average at step d.

In order to explain the concept of resampling in more detail, it is useful to view the ensemble of samples at any iteration of the scheme as a weighted empirical measure, i.e. consider the random distribution

$$\Psi_n(x_{1:n}) = \sum_{k=1}^{N_n} W_n^{(k)}\, \delta\left(x_{1:n} - X_{1:n}^{(k)}\right)$$

corresponding to the ensemble of samples generated by the above steps after n iterations. Note that Ψ_n is not quite a probability distribution unless $\sum_{k=1}^{N_n} W_n^{(k)} = 1$. Knowledge of Ψ_n is equivalent to knowledge of the ensemble of the weighted samples $\{W_n^{(k)}, X_{1:n}^{(k)}\}$.

Step 1 above corresponds to, starting from Ψ_n, generating a new random distribution

$$\tilde{\Psi}_n(x_{1:n}) = \frac{1}{N}\sum_{k=1}^{N_n} N_n^{(k)}\,\delta\left(x_{1:n} - X_{1:n}^{(k)}\right)$$

where $N_{n+1} = \sum_{k=1}^{N_n} N_n^{(k)}$ and the $N_n^{(k)}$ are random, non-negative integers satisfying

$$\mathbf{E}\left[N_n^{(k)} \,|\, \{W_n^{(\ell)}\}_{\ell=1}^{N_n}\right] = NW_n^{(k)}. \tag{3.3}$$

Defining a new collection of N_{n+1} points $\{Y_{1:n}^{(\ell)}\}_{\ell=1}^{N_{n+1}}$, exactly $N_n^{(k)}$ elements of which are equal to $X_{1:n}^{(k)}$, this last distribution can be rewritten as

$$\tilde{\Psi}_n(x_{1:n}) = \frac{1}{N}\sum_{k=1}^{N_{n+1}} \delta\left(x_{1:n} - Y_{1:n}^{(k)}\right).$$

In Steps 2 and 3 above, the samples $Y_{1:n}^{(k)}$ are augmented with a sample $X_{n+1}^{(k)}$ from $q_{n+1}\left(x_{n+1} \,|\, Y_{1:n}^{(k)}\right)$ to obtain $X_{1:n+1}^{(k)} = \left(Y_{1:n}^{(k)}, X_{n+1}^{(k)}\right)$ which is then weighted by

$$W_{n+1}^{(k)} = \frac{w_{n+1}(X_{1:n+1}^{(k)})}{\sum_{\ell=1}^{N_{n+1}} w_{n+1}(X_{1:n+1}^{(\ell)})}$$

to obtain the new distribution

$$\Psi_{n+1}(x_{1:n+1}) = \sum_{k=1}^{N_{n+1}} W_{n+1}^{(k)}\,\delta\left(x_{1:n+1} - X_{1:n+1}^{(k)}\right)$$

One possible choice for the distribution of the $\{N_n^{(k)}\}$ that satisfies condition (3.3) is the $Multinomial(N, p)$ distribution with the vector p having entries $p_k = W_n^{(k)}$, in which case $N_n = N$ exactly at each step. Notice that, if the $N_n^{(k)}$ are selected from $Multinomial(N, \{W_n^{(k)}\})$, then, since the variance of $N_n^{(k)}$ is $NW_n^{(k)}\left(1 - W_n^{(k)}\right)$,

$$\begin{aligned}\mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)^2 \,\middle|\, \Psi_n\right] &= \frac{1}{N^2}\mathbf{E}\left[\left(NW_n^{(k)} - N_n^{(k)}\right)^2 \,\middle|\, \Psi_n\right] \\ &= \frac{1}{N}W_n^{(k)}\left(1 - W_n^{(k)}\right).\end{aligned}$$

Similarly, since the covariance of $N_n^{(k)}$ and $N_n^{(\ell)}$ for $i \neq j$ is $-NW_n^{(k)}W_n^{(\ell)}$,

$$\mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) \,\middle|\, \Psi_n\right] = -\frac{W_n^{(k)}W_n^{(\ell)}}{N}.$$

These expressions imply that the error from a single resampling step is

$$\begin{aligned}
&\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \,\middle|\, \Psi_n\right] \\
&\quad = \mathbf{E}\left[\left(\sum_{\ell=1}^{N}\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) f(X_{1:n}^{(\ell)})\right)^2 \,\middle|\, \Psi_n\right] \\
&\quad = \sum_{\ell=1}^{N}\left(f(X_{1:n}^{(\ell)})\right)^2 \mathbf{E}\left[\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right)^2 \,\middle|\, \Psi_n\right] \\
&\qquad\quad + 2\sum_{k<\ell\leq N} f(X_{1:n}^{(k)})f(X_{1:n}^{(\ell)}) \\
&\qquad\qquad\qquad \times \mathbf{E}\left[\left(W_n^{(k)} - \frac{N_n^{(k)}}{N}\right)\left(W_n^{(\ell)} - \frac{N_n^{(\ell)}}{N}\right) \,\middle|\, \Psi_n\right] \\
&\quad = \frac{1}{N}\sum_{\ell=1}^{N}\left(f(X_{1:n}^{(\ell)})\right)^2 W_n^{(\ell)}\left(1 - W_n^{(\ell)}\right) \\
&\qquad\qquad\qquad - \frac{2}{N}\sum_{k<\ell\leq N} f(X_{1:n}^{(k)})f(X_{1:n}^{(\ell)})W_n^{(k)}W_n^{(\ell)} \\
&\quad = \frac{1}{N}\sum_{k=1}^{N}\left(f(X_{1:n}^{(k)}) - \sum_{\ell=1}^{N} f(X_{1:n}^{(\ell)})W_n^{(\ell)}\right)^2 W_n^{(k)}. \qquad (3.4)
\end{aligned}$$

When f is bounded, i.e. when $\|f\|_\infty < \infty$, this last expression is bounded by $\|f\|_\infty^2/N$.

For each n, and any function f which takes $x_{1:n}$ as its argument, define the new function

$$\mathcal{Q}_n f(x_{1:n-1}) = \int f(x_{1:n})w_n(x_{1:n})q_n(dx_n \,|\, x_{1:n-1})$$

which takes $x_{1:n-1}$ as its argument. Note that

$$\begin{aligned}\|Q_n f\|_\infty &\leq \|f\|_\infty \left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \int \pi_n(dx_n \mid x_{1:n-1}) \right\|_\infty \\ &= \|f\|_\infty \left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \right\|_\infty .\end{aligned}$$

We'll assume that there is some, possibly unknown, constant K so that

$$\left\| \frac{\pi_n(x_{1:n-1})}{\pi_{n-1}(x_{1:n-1})} \right\|_\infty \leq K$$

for all n so that $\|Q_n f\|_\infty \leq K \|f\|_\infty$.

The total error in the sequential importance sampling scheme after n steps can be decomposed as follows,

$$\begin{aligned}\int f(x_{1:n}) \left(\Psi_n - \pi_n\right)(dx_{1:n}) &= \int f(x_{1:n}) \Psi_n(dx_{1:n}) \\ &\quad - \int \mathcal{Q}_n f(x_{1:n-1}) \pi_{n-1}(dx_{1:n-1}) \\ &= \int f(x_{1:n}) \Psi_n(dx_{1:n}) - \int \mathcal{Q}_n f(x_{1:n-1}) \tilde{\Psi}_{n-1}(dx_{1:n-1}) \\ &\quad + \int \mathcal{Q}_n f(x_{1:n-1}) \left(\tilde{\Psi}_{n-1} - \Psi_{n-1}\right)(dx_{1:n-1}) \\ &\quad + \int \mathcal{Q}_n f(x_{1:n-1}) \left(\Psi_{n-1} - \pi_{n-1}\right)(dx_{1:n-1}).\end{aligned}$$

Labeling the terms on each of the three lines in this decomposition I_1, I_2, and I_3 respectively, note that the independence of the random variables generated at each step of the algorithm imply that

$$\mathbf{E}\left[I_1 I_2 \mid \Psi_{n-1}, \tilde{\Psi}_{n-1}\right] = 0, \quad \mathbf{E}\left[I_1 I_3 \mid \Psi_{n-1}, \tilde{\Psi}_{n-1}\right] = 0,$$

and

$$\mathbf{E}\left[I_2 I_3 \mid \Psi_{n-1}\right] = 0.$$

Therefore

$$\mathbf{E}\left[\left(\int f(x_{1:n}) \left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right] = \mathbf{E}\left[I_1^2\right] + \mathbf{E}\left[I_2^2\right] + \mathbf{E}\left[I_3^2\right].$$

The first term in this sum can be re-expressed as

$$\mathbf{E}\left[\left(\sum_{\ell=1}^{N} W_n^{(\ell)} f(X_{1:n}^{(\ell)}) - \frac{1}{N}\sum_{\ell=1}^{N} \mathcal{Q}_n f(Y_{1:n-1}^{(\ell)})\right)^2 \,\middle|\, \tilde{\Psi}_{n-1}\right]$$
$$= \frac{1}{N^2}\sum_{\ell=1}^{N} \mathbf{E}\left[\left(N W_n^{(\ell)} f(Y_{1:n-1}^{(\ell)}, X_n^{(\ell)}) - \mathcal{Q}_n f(Y_{1:n-1}^{(\ell)})\right)^2 \,\middle|\, \tilde{\Psi}_{n-1}\right]$$

which, when f is bounded at least and when the weights have finite variance, we can expect to be of size $\|f\|_\infty^2/N$. And we have already seen in (3.4) that if $\mathcal{Q}_n f$ is bounded (which it will be when f is bounded), and if the $N_n^{(\ell)}$ are sampled from a multinomial distribution, then $\mathbf{E}\left[I_2^2 \mid \Psi_{n-1}\right]$ is of size $K^2\|f\|_\infty^2/N$.

At this point (under a few assumptions) we have shown that the error at step n is only slightly ($\mathcal{O}(1/N)$) larger than the step $n-1$ error in estimating the average of $\mathcal{Q}_n f$ against π_{n-1}, i.e. we have shown that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathbf{E}\left[\left(\int \mathcal{Q}_n f(x_{1:n-1})\left(\Psi_{n-1} - \pi_{n-1}\right)(dx_{1:n-1})\right)^2\right]$$
$$+ \mathcal{O}\left(\|f\|_\infty^2 \frac{1+K^2}{N}\right).$$

Repeating the same steps $n-2$ more times we find that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathbf{E}\left[\left(\int \mathcal{Q}_n \cdots \mathcal{Q}_2 f(x_1)\left(\Psi_1 - \pi_1\right)(dx_1)\right)^2\right]$$
$$+ \mathcal{O}\left(\|f\|_\infty^2 \frac{1+K^2+\cdots+K^{2(n-2)}}{N}\right).$$

Since the samples $X_1^{(\ell)}$ were drawn independently from π_1 we know that

$$\mathbf{E}\left[\left(\int \mathcal{Q}_n \cdots \mathcal{Q}_2 f(x_1)\left(\Psi_1 - \pi_1\right)(x_1)dx_1\right)^2\right] = \frac{\mathbf{var}\left(\mathcal{Q}_n \cdots \mathcal{Q}_2 f\left(X_1^{(\ell)}\right)\right)}{N}$$
$$\leq \|f\|_\infty^2 \frac{K^{2(n-1)}}{N}$$

so that

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \pi_n\right)(dx_{1:n})\right)^2\right]$$
$$= \mathcal{O}\left(\|f\|_\infty^2 \frac{1 + K^2 + \cdots + K^{2(n-1)}}{N}\right).$$

In other words, we have shown that the sequential importance sampling with resampling scheme does converge to the correct answer as N increases. On the other hand, our estimates have been crude and do not reveal any advantage for sequential importance sampling with resampling over direct importance sampling. The growth of our bound with n is one symptom of our loose estimates. With more work, and a few more assumptions, we could show that the error in sequential importance sampling with resampling can often be bounded independently of n, something that would not typically be possible for direct importance sampling. Among other requirements, some form of contraction from the operators $\mathcal{Q}_n$ will be important in obtaining global-in-time error bounds.

Before ending this section we briefly consider alternatives to the multinomial distribution for sampling the $N_n^{(\ell)}$ in the sequential importance sampling with resampling procedure. Ultimately our goal in choosing a resampling scheme is to make the expectation

$$\mathbf{E}\left[\left(\int f(x_{1:n})\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \,\middle|\, \Psi_n\right]$$

as small as possible. However, we have seen in our bound on the total error that we must control terms of this form for functions f that we may not know in closed form (e.g. $\mathcal{Q}_n f$). It is reasonable then to instead attempt to minimize the conditional variances

$$\mathbf{var}\left[N_n^{(k)} \,|\, \Psi_n\right].$$

Given the requirement that the $N_n^{(k)}$ are integers and that

$$\mathbf{E}\left[N_n^{(k)} \mid \Psi_n\right] = NW_n^{(k)},$$

the conditional variance above is easily seen to be minimized when

$$N_n^{(k)} = \begin{cases} \left\lfloor NW_n^{(k)} \right\rfloor & \text{w. p. } \left\lceil NW_n^{(k)} \right\rceil - NW_n^{(k)} \\ \left\lceil NW_n^{(k)} \right\rceil & \text{w. p. } NW_n^{(k)} - \left\lfloor NW_n^{(k)} \right\rfloor. \end{cases} \tag{3.5}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. With condition (3.5) enforced, the conditional variances become

$$\mathbf{var}\left[N_n^{(k)} \mid \Psi_n\right] = \left(\left\lceil NW_n^{(k)} \right\rceil - NW_n^{(k)}\right)\left(NW_n^{(k)} - \left\lfloor NW_n^{(k)} \right\rfloor\right) \tag{3.6}$$

Exercise 23. *Verify that the conditional variance is minimized when* (3.5) *is satisfied and that the minimum value is given by* (3.6)*.*

The $Multinomial(N, \{W_n^{(\ell)}\})$ distribution does not satisfy (3.5) and we have seen that

$$\mathbf{var}\left[N_n^{(k)} \mid \Psi_n\right] = NW_n^{(k)}\left(1 - W_n^{(k)}\right)$$

which is much larger than the minimal value in (3.6). We will now consider a few rules that do satisfy (3.5). We begin by noticing that (3.5) constrains only the marginal distribution of the $N_n^{(k)}$ and does not affect their joint distribution. The simplest possible choice is to make the $N_n^{(k)}$ independent:

$$N_n^{(k)} = \left\lfloor NW_n^{(k)} \right\rfloor + \mathbf{1}_{\left\{U_n^{(k)} < NW_n^{(k)} - \lfloor NW_n^{(k)} \rfloor\right\}} \tag{3.7}$$

where the $U_n^{(k)}$ are independent random variables drawn from $\mathcal{U}(0,1)$. This scheme is sometimes referred to as Bernoulli resampling.

Exercise 24. *Check that for the $N_n^{(k)}$ generated according to* (3.7)*,* (3.5) *is satisfied.*

While (3.7) minimizes the conditional variances of the $N_n^{(k)}$, the total number of resampled points, $N_n = \sum_{\ell=1}^{N_{n-1}} N_{n-1}^{(\ell)}$, is not exactly equal to N (though its expectation is equal to N).

Exercise 25. *Follow the steps used to derive expression* (3.4) *to derive a bound for*

$$\mathbf{E}\left[\left(\int f(x_{1:n}\left(\Psi_n - \tilde{\Psi}_n\right)(dx_{1:n})\right)^2 \mid \Psi_n\right]$$

when the $N_n^{(k)}$ *are generated according to* (3.7).

Finally, a rule for generating the $N_n^{(k)}$ that fixes $N_n = N$ and which requires that we generate only one random variable to generate all of the $N_n^{(k)}$ at iteration n, proceeds as follows. First, generate a single independent random variate U_n from $\mathcal{U}(0,1)$. Then, for $k = 1, 2, \ldots, N_n$, set

$$N_n^{(k)} = \left|\left\{j \leq N : \sum_{\ell=1}^{k-1} W_n^{(\ell)} \leq U_n^{(j)} < \sum_{\ell=1}^{k} W_n^{(\ell)}\right\}\right| \tag{3.8}$$

where, for $j = 1, 2, \ldots, N$,

$$U_n^{(j)} = \frac{1}{N}(j - U_n)$$

and the notation $|A|$ for a discrete set of points A refers to the number of points in A.

Exercise 26. *Show that for* $N_n^{(\ell)}$ *defined by* (3.8), $\sum_{\ell=1}^{N} N_n^{(\ell)} = N$ *and* (3.5) *is satisfied.*

The rule (3.8) is often referred to as systematic resampling and is observed to perform very well in practice. Despite its success in applications, it is unfortunately not possible to show that it converges in general.

Exercise 27. *Find a sequence of weights* $\{w^{(\ell)}\}_{\ell=1}^N$ *with* $\sum_{\ell=1}^N w^{(\ell)} = 1$, *and points* $\{x^{(\ell)}\}_{\ell=1}^N$ *so that*

$$\mathbf{E}\left[\left(\frac{1}{N}\sum_{\ell=1}^{N} N^{(\ell)} f(x^{(\ell)}) - N w^{(\ell)} f(x^{(\ell)})\right)^2\right]$$

does not converge when the $N^{(\ell)}$ *are generated according to* (3.8) *with* $w^{(\ell)}$ *in place of* $W_n^{(\ell)}$. *Hint: try an even length alternating sequence of two values,* x_0 *and* x_1, *and assume that if* $x^{(k)} = x^{(\ell)}$ *then* $w^{(k)} = w^{(\ell)}$ *(as would occur if the* $w^{(k)}$ *were importance weights).*

Exercise 28. *Write a routine to use $\mathcal{N}(0,1)$ random variables to generate approximate $\mathcal{N}(0,\sigma^2)$ random variables via an application of each of the three resampling methods (multinomial, Bernoulli, and systematic) discussed in this section. Numerically estimate the variance of the $N^{(k)}$ from each method. What do you observe? Are your observations robust to changing σ^2? Note that this test corresponds to a single resampling step: first sample from $\mathcal{N}(0,1)$, then weight the samples by the appropriate normalized importance weights, then resample.*

Exercise 29. *Write a routine that uses sequential importance sampling with and without resampling to compute averages with respect to the uniform measure on SAW(d) for a sequence of increasing d. Use the same reference density described in Example 11 in both schemes. Produce plots of a single sample path for each value of d. In validating your algorithms you may find it useful to check, e.g. the expected number of times a lattice site is visited (this should be the same for every site). Can you think of other statistics that might help you validate/debug your codes? How can you compare the two methods? Can you think of a way to use these simulations to estimate the normalization constants $\mathcal{Z}_d$? Estimate how quickly $\mathcal{Z}_d$ grows with d.*

3.5 bibliography