Contents

9

Importance sampling

In many applications we want to compute $\mu = \mathbb{E}(f(\boldsymbol{X}))$ where $f(\boldsymbol{x})$ is nearly zero outside a region A for which $\mathbb{P}(\boldsymbol{X} \in A)$ is small. The set A may have small volume, or it may be in the tail of the $\boldsymbol{X}$ distribution. A plain Monte Carlo sample from the distribution of $\boldsymbol{X}$ could fail to have even one point inside the region A. Problems of this type arise in high energy physics, Bayesian inference, rare event simulation for finance and insurance, and rendering in computer graphics among other areas.

It is clear intuitively that we must get some samples from the interesting or important region. We do this by sampling from a distribution that overweights the important region, hence the name ***importance sampling***. Having oversampled the important region, we have to adjust our estimate somehow to account for having sampled from this other distribution.

Importance sampling can bring enormous gains, making an otherwise infeasible problem amenable to Monte Carlo. It can also backfire, yielding an estimate with infinite variance when simple Monte Carlo would have had a finite variance. It is the hardest variance reduction method to use well.

Importance sampling is more than just a variance reduction method. It can be used to study one distribution while sampling from another. As a result we can use importance sampling as an alternative to acceptance-rejection sampling, as a method for sensitivity analysis and as the foundation for some methods of computing normalizing constants of probability densities. Importance sampling is also an important prerequisite for sequential Monte Carlo (Chapter 15). For these reasons we spend a whole chapter on it. For a mini-chapter on the basics of importance sampling, read §9.1 through §9.4.

9.1 Basic importance sampling

Suppose that our problem is to find $\mu = \mathbb{E}(f(\boldsymbol{X})) = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where p is a probability density function on $\mathcal{D} \subseteq \mathbb{R}^d$ and f is the integrand. We take $p(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \notin \mathcal{D}$. If q is a positive probability density function on $\mathbb{R}^d$, then

$$\mu = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}} \frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}\,q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}_q\biggl(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\biggr), \tag{9.1}$$

where $\mathbb{E}_q(\cdot)$ denotes expectation for $\boldsymbol{X} \sim q$. We also write $\mathrm{Var}_q(\cdot)$, $\mathrm{Cov}_q(\cdot,\cdot)$, and $\mathrm{Corr}_q(\cdot,\cdot)$ for variance, covariance and correlation when $\boldsymbol{X} \sim q$. Our original goal then is to find $\mathbb{E}_p(f(\boldsymbol{X}))$. By making a multiplicative adjustment to f we compensate for sampling from q instead of p. The adjustment factor $p(\boldsymbol{x})/q(\boldsymbol{x})$ is called the ***likelihood ratio***. The distribution q is the ***importance distribution*** and p is the ***nominal distribution***.

The importance distribution q does not have to be positive everywhere. It is enough to have $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. That is, for $\mathcal{Q} = \{\boldsymbol{x} \mid q(\boldsymbol{x}) > 0\}$ we have $\boldsymbol{x} \in \mathcal{Q}$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. So if $\boldsymbol{x} \in \mathcal{D} \cap \mathcal{Q}^c$ we know that $f(\boldsymbol{x}) = 0$, while if $\boldsymbol{x} \in \mathcal{Q} \cap \mathcal{D}^c$ we have $p(\boldsymbol{x}) = 0$. Now

$$\begin{aligned}
\mathbb{E}_q\biggl(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\biggr) &= \int_{\mathcal{Q}} \frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}\,q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{Q}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{\mathcal{Q}\cap\mathcal{D}^c} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{D}\cap\mathcal{Q}^c} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu.
\end{aligned} \tag{9.2}$$

It is natural to wonder what happens for $\boldsymbol{x}$ with $q(\boldsymbol{x}) = 0$ in the denominator. The answer is that there are no such points $\boldsymbol{x} \in \mathcal{Q}$ and we will never see one when sampling $\boldsymbol{X} \sim q$. Later we will see examples where $q(\boldsymbol{x})$ close to 0 causes extreme difficulties, but $q(\boldsymbol{x}) = 0$ is not a problem if $f(\boldsymbol{x})p(\boldsymbol{x}) = 0$ too.

When we want q to work for many different functions f_j then we need $q(\boldsymbol{x}) > 0$ at every $\boldsymbol{x}$ where any $f_j(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. Then a density q with $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$ will suffice, and will allow us to add new functions f_j to our list after we've drawn the sample.

The ***importance sampling estimate*** of $\mu = \mathbb{E}_p(f(\boldsymbol{X}))$ is

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)}, \quad \boldsymbol{X}_i \sim q. \tag{9.3}$$

To use (9.3) we must be able to compute fp/q. Assuming that we can compute f, this estimate requires that we can compute $p(\boldsymbol{x})/q(\boldsymbol{x})$ at any $\boldsymbol{x}$ we might sample. When p or q has an unknown normalization constant, then we will resort to a ratio estimate (see §9.2). For now, we assume that p/q is computable, and study the variance of $\hat{\mu}_q$. Exponential tilting (§9.6) is one way to choose q with computable p/q even when p is unnormalized.

Theorem 9.1. *Let $\hat{\mu}_q$ be given by (9.3) where $\mu = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. Then $\mathbb{E}_q(\hat{\mu}_q) = \mu$, and $\mathrm{Var}_q(\hat{\mu}_q) = \sigma_q^2/n$ where*

$$\begin{aligned}\sigma_q^2 &= \int_{\mathcal{D}} \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} - \mu^2 \\ &= \int_{\mathcal{D}} \frac{\big(f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x})\big)^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{9.4}$$

Proof. That $\mathbb{E}_q(\hat{\mu}_q) = \mu$ follows directly from (9.2). Using $\mathcal{Q} = \{\boldsymbol{x} \mid q(\boldsymbol{x}) > 0\}$, we find that

$$\begin{aligned}\mathrm{Var}_q(\hat{\mu}_q) &= \frac{1}{n}\Big[\int_{\mathcal{Q}} \Big(\frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}\Big)^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \mu^2\Big] \\ &= \frac{1}{n}\Big[\int_{\mathcal{D}} \Big(\frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}\Big)^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \mu^2\Big],\end{aligned}$$

because the contributions to the integral from $\boldsymbol{x}$ in $\mathcal{D} \cap \mathcal{Q}^c$ and $\mathcal{Q} \cap \mathcal{D}^c$ are zero. Simple rearrangements give the two forms in equation (9.4). □

To form a confidence interval for μ, we need to estimate σ_q^2. From the second expression in (9.4) we find that

$$\sigma_q^2 = \mathbb{E}_q((f(\boldsymbol{X})p(\boldsymbol{X}) - \mu q(\boldsymbol{X}))^2/q(\boldsymbol{X})^2).$$

Because $\boldsymbol{x}_i$ are sampled from q, the natural variance estimate is

$$\hat{\sigma}_q^2 = \frac{1}{n}\sum_{i=1}^n \Big(\frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} - \hat{\mu}_q\Big)^2 = \frac{1}{n}\sum_{i=1}^n \big(w_i f(\boldsymbol{x}_i) - \hat{\mu}_q\big)^2, \tag{9.5}$$

where $w_i = p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$. Then an approximate 99% confidence interval for μ is $\hat{\mu}_q \pm 2.58\hat{\sigma}_q/\sqrt{n}$.

Theorem 9.1 guides us in selecting a good importance sampling rule. The first expression in (9.4) is simpler to study. A better q is one that gives a smaller value of $\int_{\mathcal{D}}(fp)^2/q\,\mathrm{d}\boldsymbol{x}$. When we want upper bounds on σ_q^2 we can bound $\int_{\mathcal{D}}(fp)^2/q\,\mathrm{d}\boldsymbol{x}$.

The second integral expression in (9.4) illustrates how importance sampling can succeed or fail. The numerator in the integrand at the right of (9.4) is small when $f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x})$ is close to zero, that is, when $q(\boldsymbol{x})$ is nearly proportional to $f(\boldsymbol{x})p(\boldsymbol{x})$. From the denominator, we see that regions with small values of $q(\boldsymbol{x})$ greatly magnify whatever lack of proportionality appears in the numerator.

Suppose that $f(\boldsymbol{x}) \geqslant 0$ for all $\boldsymbol{x}$, and to rule out a trivial case, assume that $\mu > 0$ too. Then $q_{\mathrm{opt}}(\boldsymbol{x}) \equiv f(\boldsymbol{x})p(\boldsymbol{x})/\mu$ is a probability density, and it has $\sigma^2_{q_{\mathrm{opt}}} = 0$. It is optimal, but not really usable because $\hat{\mu}_{q_{\mathrm{opt}}}$ becomes an average of $f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i) = \mu$ meaning that we could compute μ directly from f, p, and q without any sampling. Likewise for $f(\boldsymbol{x}) \leqslant 0$ with $\mu < 0$ a zero variance is obtained for the density $q = -fp/\mu$.

Although zero-variance importance sampling densities are not usable, they provide insight into the design of a good importance sampling scheme. Suppose that $f(\boldsymbol{x}) \geqslant 0$. It may be good for q to have spikes in the same places that f does, or where p does, but it is better to have them where fp does. Moreover, the best q is proportional to fp not $\sqrt{fp}$ or f^2p or some other combination.

To choose a good importance sampling distribution requires some educated guessing and possibly numerical search. In many applications there is domain knowledge about where the spikes are. In a financial setting we may know which stock fluctuations will cause an option to go to its maximal value. For a queuing system it may be easy to know what combination of arrivals will cause the system to be overloaded.

In general, the density q^* that minimizes σ_q^2 is proportional to $|f(\boldsymbol{x})|p(\boldsymbol{x})$ (Kahn and Marshall, 1953), outside of trivial cases where $\int |f(\boldsymbol{x})|p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$. This optimal density does not have $\sigma_{q^*}^2 = 0$ on problems where fp can be positive for some $\boldsymbol{x}$ and negative for other $\boldsymbol{x}$. When fp takes both positive and negative values, there is still a zero variance method, but it requires sampling at two points. See §9.13.

To prove that $q^*(\boldsymbol{x}) = |f(\boldsymbol{x})|p(\boldsymbol{x})/\mathbb{E}_p(|f(\boldsymbol{X})|)$ is optimal, let q be any density that is positive when $fp \neq 0$. Then

$$\begin{aligned}\mu^2 + \sigma_{q^*}^2 &= \int \frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2}{q^*(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} = \int \frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2}{|f(\boldsymbol{x})|p(\boldsymbol{x})/\mathbb{E}_p(|f(\boldsymbol{X})|)}\,\mathrm{d}\boldsymbol{x}\\ &= \mathbb{E}_p(|f(\boldsymbol{X})|)^2 = \mathbb{E}_q(|f(\boldsymbol{X})|p(\boldsymbol{X})/q(\boldsymbol{X}))^2\\ &\leqslant \mathbb{E}_q(f(\boldsymbol{X})^2 p(\boldsymbol{X})^2/q(\boldsymbol{X})^2) = \mu^2 + \sigma_q^2,\end{aligned}$$

and so $\sigma_{q^*}^2 \leqslant \sigma_q^2$. This proof is a straightforward consequence of the Cauchy-Schwarz inequality. But it requires that we know the optimal density before applying Cauchy-Schwarz. For a principled way to find candidates for q^* in problems like importance sampling, we can turn to the calculus of variations, as outlined on page 38 of the chapter end notes.

We may use the likelihood ratio $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ as a means to understand which importance sampling densities are good or bad choices. The first term in σ_q^2 is $\int f(\boldsymbol{x})^2 p(\boldsymbol{x})^2/q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. We may write this term as $\mathbb{E}_p(f(\boldsymbol{X})^2 w(\boldsymbol{X})) = \mathbb{E}_q(f(\boldsymbol{X})^2 w(\boldsymbol{X})^2)$. The appearance of q in the denominator of w, means that light-tailed importance densities q are dangerous. If we are clever or lucky, then f might be small just where it needs to be to offset the small denominator. But we often need to use the same sample with multiple integrands f, and so as a rule q should have tails at least as heavy as p does.

When p is a Gaussian distribution, then a common tactic is to take q to be a student's t distribution. Such a q has heavier tails than p. It even has heavier tails than fp for integrands of the form $f(\boldsymbol{x}) = \exp(\boldsymbol{x}^\mathsf{T}\theta)$, because then fp is proportional to another Gaussian density.

The reverse practice, of using Gaussian importance distribution q for a student's t nominal distribution, can easily lead to $\sigma_q^2 = \infty$. Even when q is nearly proportional to fp we may still have $\sigma_q^2 = \infty$. The irony is that the infinite

variance could be largely attributable to the unimportant region of $\mathcal{D}$ where $|f|p$ is small (but q is extremely small).

Bounded weights are especially valuable. If $w(\boldsymbol{x}) \leqslant c$ then $\sigma_q^2 \leqslant c\sigma_p^2$. See Exercise 9.6.

Example 9.1 (Gaussian p and q)**.** We illustrate the effect of light-tailed q in a problem where we would not need importance sampling. Suppose that $f(x) = x \in \mathbb{R}$ and that $p(x) = \exp(-x^2/2)/\sqrt{2\pi}$. If $q(x) = \exp(-x^2/(2\sigma^2))/(\sigma\sqrt{2\pi})$ with $\sigma > 0$ then

$$\begin{aligned}\sigma_q^2 &= \int_{-\infty}^{\infty} x^2 \frac{\bigl(\exp(-x^2/2)/\sqrt{2\pi}\bigr)^2}{\exp(-x^2/(2\sigma^2))/\sqrt{2\pi}/\sigma}\,\mathrm{d}x\\ &= \sigma\int_{-\infty}^{\infty} x^2 \exp(-x^2(2-\sigma^{-2})/2)/\sqrt{2\pi}\,\mathrm{d}x\\ &= \begin{cases}\dfrac{\sigma}{(2-\sigma^{-2})^{3/2}}, & \sigma^2 > 1/2\\ \infty, & \text{else.}\end{cases}\end{aligned}$$

To estimate $\mathbb{E}_p(X)$ with finite variance in the normal family, the importance distribution q must have more than half the variance that p has. Interestingly, the optimal σ is not 1. See Exercise 9.7.

The likelihood ratio also reveals a dimension effect for importance sampling. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ and to keep things simple, that the components of $\boldsymbol{X}$ are independent under both p and q. Then $w(\boldsymbol{X}) = \prod_{j=1}^d p_j(X_j)/q_j(X_j)$. Though f plays a role, the value of σ_q^2 is also driven largely by $\mathbb{E}_q(w(\boldsymbol{X})^2) = \prod_{j=1}^d \mathbb{E}_q(w_j(X_j)^2)$ where $w_j = p_j/q_j$. Each $\mathbb{E}_q(w_j(X_j)) = 1$ and $\mathbb{E}_q(w(\boldsymbol{X})^2) = \prod_{j=1}^d(1+\mathrm{Var}_q(w_j(X_j)))$. This quantity will grow exponentially with d unless we arrange for the variances of $w_j(X_j)$ to eventually decrease sharply as j increases.

Example 9.2 (Standard Gaussian mean shift)**.** Let $\boldsymbol{X} \sim \mathcal{N}(0, I)$ under p and $\boldsymbol{X} \sim \mathcal{N}(\theta, I)$ under q, where $\theta \in \mathbb{R}^d$. Then $w(\boldsymbol{x}) = \exp(-\theta^{\mathsf{T}}\boldsymbol{x} + \theta^{\mathsf{T}}\theta/2)$. The moments of the likelihood ratio depend on the size $\|\theta\|$ of our shift, but not on the dimension d. In particular, $\mathbb{E}_q(w(\boldsymbol{X})^2) = \exp(\theta^{\mathsf{T}}\theta)$.

Example 9.2 shows that a large nominal distribution d does not of itself make importance sampling perform poorly. Taking $\theta = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ there has the same effect as $\theta = (1, 0, \ldots, 0)$ does. The latter is clearly a one dimensional perturbation, and so is the former, since it involves importance sampling for just one linear combination of the d variables. What we do see is that small changes to a large number of components of $\mathbb{E}(\boldsymbol{X})$ or moderately large changes to a few of them, do not necessarily produce an extreme likelihood ratio.

Example 9.3 (Standard Gaussian variance change)**.** Let $\boldsymbol{X} \sim \mathcal{N}(0, I)$ and suppose that q has $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 I)$ for $\sigma > 0$. Then

$$w(\boldsymbol{x}) = \sigma^d \exp(-\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}(1 - 1/(2\sigma^2))),$$

and $\mathbb{E}_p(w(\boldsymbol{X})) = (\sigma^2/\sqrt{2\sigma^2 - 1})^d$ for $\sigma^2 > 1/2$.

Example 9.3 shows that making any small change of variance will bring an extreme likelihood ratio in d dimensions when d is large. The allowable change to σ^2 decreases with d. If we take $\sigma^2 = 1 + A/d^{1/2}$ for $A > -1/2$ then a Taylor series expansion of $\log(\mathbb{E}_p(w(\boldsymbol{X})))$ shows that $(\sigma^2/\sqrt{2\sigma^2-1})^d$ remains bounded as d grows.

Although we can get a lot of insight from the likelihood ratio $w(\boldsymbol{x})$ the efficiency of importance sampling compared to alternatives still depends on f. We will see another example of that phenomenon in §9.3 where we discuss some measures of effective sample size for importance sampling.

9.2 Self-normalized importance sampling

Sometimes we can only compute an unnormalized version of p, $p_u(\boldsymbol{x}) = cp(\boldsymbol{x})$ where $c > 0$ is unknown. The same may be true of q. Suppose that we can compute $q_u(\boldsymbol{x}) = bq(\boldsymbol{x})$ where $b > 0$ might be unknown. If we are fortunate or clever enough to have $b = c$, then $p(\boldsymbol{x})/q(\boldsymbol{x}) = p_u(\boldsymbol{x})/q_u(\boldsymbol{x})$ and we can still use (9.3). Otherwise we may compute the ratio $w_u(\boldsymbol{x}) = p_u(\boldsymbol{x})/q_u(\boldsymbol{x}) = (c/b)p(\boldsymbol{x})/q(\boldsymbol{x})$ and consider the ***self-normalized importance sampling estimate***

$$\widetilde{\mu}_q = \frac{\sum_{i=1}^n f(\boldsymbol{X}_i) w_u(\boldsymbol{X}_i)}{\sum_{i=1}^n w_u(\boldsymbol{X}_i)} \tag{9.6}$$

where $\boldsymbol{X}_i \sim q$ are independent. The factor c/b cancels from numerator and denominator in (9.6), leading to the same estimate as if we had used the desired ratio $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ instead of the computable alternative $w_u(\boldsymbol{x})$. As a result we can write

$$\widetilde{\mu}_q = \frac{\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i) w(\boldsymbol{X}_i)}{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i)}. \tag{9.7}$$

Theorem 9.2. *Let p be a probability density function on $\mathbb{R}^d$ and let $f(\boldsymbol{x})$ be a function such that $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ exists. Suppose that $q(\boldsymbol{x})$ is a probability density function on $\mathbb{R}^d$ with $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim q$ be independent and let $\widetilde{\mu}_q$ be the self-normalized importance sampling estimate* (9.6)*. Then*

$$\mathbb{P}\Big(\lim_{n\to\infty} \widetilde{\mu}_q = \mu\Big) = 1.$$

Proof. We may use expression (9.7) for $\widetilde{\mu}_q$. The numerator in (9.7) is the average $\hat{\mu}_q$ of $f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i)$ under sampling from q. These are IID with mean μ by (9.2). The strong law of large numbers gives $\mathbb{P}\big(\lim_{n\to\infty}\hat{\mu}_q = \mu\big) = 1$. The denominator of (9.7) converges to 1, using the argument for the numerator on the function f with $f(\boldsymbol{x}) \equiv 1$. □

The self-normalized importance sampler $\widetilde{\mu}_q$ requires a stronger condition on q than the unbiased importance sampler $\hat{\mu}_q$ does. We now need $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$ even if $f(\boldsymbol{x})$ is zero with high probability.

To place a confidence interval around $\widetilde{\mu}_q$ we use the delta method and equation (2.30) from §2.7 for the variance of a ratio of means. Applying that formula to (9.7) yields an approximate variance for $\widetilde{\mu}_q$ of

$$\begin{aligned}\widetilde{\mathrm{Var}}(\widetilde{\mu}_q) &= \frac{1}{n}\frac{\mathbb{E}_q\big((f(\boldsymbol{X})w(\boldsymbol{X})-\mu w(\boldsymbol{X}))^2\big)}{\mathbb{E}_q(w(\boldsymbol{X}))^2} = \frac{\sigma^2_{q,\mathrm{sn}}}{n}, \quad \text{where}\\ \sigma^2_{q,\mathrm{sn}} &= \mathbb{E}_q\big(w(\boldsymbol{X})^2(f(\boldsymbol{X})-\mu)^2\big).\end{aligned} \tag{9.8}$$

For a computable variance estimate, we employ w_u instead of w, getting

$$\widehat{\widetilde{\mathrm{Var}}}(\widetilde{\mu}_q) = \frac{1}{n}\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^n w_u(\boldsymbol{x}_i)^2(f(\boldsymbol{x}_i)-\widetilde{\mu}_q)^2}{\Big(\dfrac{1}{n}\displaystyle\sum_{i=1}^n w_u(\boldsymbol{x}_i)\Big)^2} = \sum_{i=1}^n w_i^2(f(\boldsymbol{x}_i)-\widetilde{\mu}_q)^2 \tag{9.9}$$

where $w_i = w_u(\boldsymbol{x}_i)/\sum_{j=1}^n w_u(\boldsymbol{x}_j)$ is the i'th normalized weight. The delta method 99% confidence interval based for self-normalized importance sampling is

$$\widetilde{\mu}_q \pm 2.58\sqrt{\widehat{\widetilde{\mathrm{Var}}}(\widetilde{\mu}_q)/n}.$$

When the weight function $w(\boldsymbol{x})$ is computable, then we can choose to use either the unbiased estimate (9.3) or the ratio estimate (9.6). They each have their strengths. Their variance formulas are:

$$\begin{aligned} n\mathrm{Var}(\hat{\mu}_q) &= \int \frac{(f(\boldsymbol{x})p(\boldsymbol{x})-\mu q(\boldsymbol{x}))^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}, \quad \text{and}\\ n\widetilde{\mathrm{Var}}(\widetilde{\mu}_q) &= \int \frac{p(\boldsymbol{x})^2(f(\boldsymbol{x})-\mu)^2}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

An argument for $\widetilde{\mu}_q$ is that it is exact when $f(\boldsymbol{x}) = C$ is constant, while $\hat{\mu}_q$ is not. We are not usually focused on estimating the mean of a constant, but if we don't get that right, then our estimates of $\mathbb{P}(X \in B)$ and $\mathbb{P}(X \notin B)$ for a set B, won't sum to one, and $\widehat{\mathbb{E}}(f(\boldsymbol{X})+C)$ won't equal $\widehat{\mathbb{E}}(f(\boldsymbol{X}))+C$, both of which could lead to strange results.

The arguments in favor of $\hat{\mu}_q$ are that it is unbiased, has a simple variance estimate, and the variance can be made arbitrarily close to 0 by making ever better choices of q.

It is not possible for the self-normalized estimate $\widetilde{\mu}_q$ to approach 0 variance with ever better choices of q. The optimal density for self-normalized importance sampling has the form $q(\boldsymbol{x}) \propto |f(\boldsymbol{x})-\mu|p(\boldsymbol{x})$ (Hesterberg, 1988, Chapter 2). Using this formula we find (Exercise 9.15) that

$$\sigma^2_{q,\mathrm{sn}} \geqslant \mathbb{E}_p(|f(\boldsymbol{X})-\mu|)^2. \tag{9.10}$$

It is zero only for constant $f(\boldsymbol{x})$. Self-normalized importance sampling cannot reduce the variance by more than $\mathbb{E}_p((f(\boldsymbol{X})-\mu)^2)/\mathbb{E}_p(|f(\boldsymbol{X})-\mu|)^2$.

Example 9.4 (Event probability)**.** Let $f(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in A\}$, so $\mu = \mathbb{E}(f(\boldsymbol{X}))$ is the probability of event A. The optimal distribution for self-normalized importance sampling is $q(\boldsymbol{x}) \propto p(\boldsymbol{x})|\mathbb{1}\{\boldsymbol{x} \in A\} - \mu|$. Under this distribution

$$\int_A q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{\int_A p(\boldsymbol{x})|\mathbb{1}\{\boldsymbol{x} \in A\} - \mu|\,\mathrm{d}\boldsymbol{x}}{\int p(\boldsymbol{x})|\mathbb{1}\{\boldsymbol{x} \in A\} - \mu|\,\mathrm{d}\boldsymbol{x}} = \frac{\mu(1-\mu)}{\mu(1-\mu) + (1-\mu)\mu} = \frac{1}{2}.$$

Whether A is rare or not, the asymptotically optimal density for self-normalized importance sampling puts half of its probability inside A. By contrast, in ordinary importance sampling, the optimal distribution places all of its probability inside A.

When $w(\boldsymbol{x})$ is computable, then a reasonable answer to the question of whether to use $\hat{\mu}_q$ or $\widetilde{\mu}_q$ is 'neither'. In this case we can use the regression estimator of §8.9 with the control variate $w(\boldsymbol{X})$. This variate has known mean, $\mathbb{E}_q(w(\boldsymbol{X})) = 1$. The regression estimator has a bias but it cannot increase the asymptotic variance compared to $\hat{\mu}_q$. The regression estimator has approximate variance $(1-\rho^2)\sigma_q^2/n$ compared to σ_q^2/n for $\hat{\mu}_q$ of (9.3), where $\rho = \mathrm{Corr}_q(f(\boldsymbol{X})w(\boldsymbol{X}), w(\boldsymbol{X}))$. The regression estimator is also at least as efficient as the ratio estimator which in this case is $\widetilde{\mu}_q$. The regression estimator is exact if either $q(\boldsymbol{x}) \propto f(\boldsymbol{x})p(\boldsymbol{x})$ (for nonnegative f) or f is constant.

The regression estimator here takes the form

$$\hat{\mu}_{q,\mathrm{reg}} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i)w(\boldsymbol{X}_i) - \hat{\beta}(w(\boldsymbol{X}_i) - 1), \tag{9.11}$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^n (w(\boldsymbol{X}_i) - \bar{w})f(\boldsymbol{X}_i)w(\boldsymbol{X}_i)}{\sum_{i=1}^n (w(\boldsymbol{X}_i) - \bar{w})^2}$$

for $\bar{w} = (1/n)\sum_{i=1}^n w(\boldsymbol{X}_i)$. To get both $\hat{\mu}_{q,\mathrm{reg}}$ and its standard error, we can use least squares regression as in Algorithm 8.3. We regress $f(\boldsymbol{X}_i)w(\boldsymbol{X}_i)$ on the centered variable $w(\boldsymbol{X}_i) - 1$ and retain the estimated intercept term as $\hat{\mu}_{q,\mathrm{reg}}$. Most regression software will also compute a standard error $\mathrm{se}(\hat{\mu}_{q,\mathrm{reg}})$ for the intercept. Our 99% confidence interval is $\hat{\mu}_{q,\mathrm{reg}} \pm 2.58\,\mathrm{se}(\hat{\mu}_{q,\mathrm{reg}})$.

The regression estimator version of importance sampling is a very good choice when we are able to compute $w(\boldsymbol{x})$. It has one small drawback relative to $\hat{\mu}_q$. In extreme cases we could get $\hat{\mu}_{q,\mathrm{reg}} < 0$ even though $f(\boldsymbol{x}) > 0$ always holds. In such cases it may be preferable to use a reweighting approach as in §8.10. That will give $\hat{\mu}_{q,\mathrm{reg}} > 0$ when $\min_{1\leqslant i\leqslant n} f(\boldsymbol{x}_i) > 0$ and $\min_{1\leqslant i\leqslant n} p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i) < 1 < \max_i p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$. On the other hand, getting $\hat{\mu}_{q,\mathrm{reg}} < 0$ when $\mathbb{P}(f(\boldsymbol{X}) > 0) = 1$ suggests that the sampling method has not worked well and perhaps we should increase n or look for a better q instead of trying to make better use of the given sample values.

9.3 Importance sampling diagnostics

Importance sampling uses unequally weighted observations. The weights are $w_i = p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i) \geqslant 0$ for $i = 1, \ldots, n$. In extreme settings, one of the w_i may be vastly larger than all the others and then we have effectively only got one observation. We would like to have a diagnostic to tell when the weights are problematic. It is even possible that $w_1 = w_2 = \cdots = w_n = 0$. In that case, importance sampling has clearly failed and we don't need a diagnostic to tell us so. Hence, we may assume that $\sum_{i=1}^n w_i > 0$.

Consider a hypothetical linear combination

$$S_{\boldsymbol{w}} = \frac{\sum_{i=1}^n w_i Z_i}{\sum_{i=1}^n w_i} \tag{9.12}$$

where Z_i are independent random variables with a common mean and common variance $\sigma^2 > 0$ and $\boldsymbol{w} \in [0, \infty)^n$ is the vector of weights. The unweighted average of n_{e} independent random variables Z_i has variance σ^2/n_{e}. Setting $\mathrm{Var}(S_{\boldsymbol{w}}) = \sigma^2/n_{\mathrm{e}}$ and solving for n_{e} yields the ***effective sample size***

$$n_{\mathrm{e}} = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} = \frac{n\bar{w}^2}{\overline{w^2}} \tag{9.13}$$

where $\bar{w} = (1/n)\sum_{i=1}^n w_i$ and $\overline{w^2} = (1/n)\sum_{i=1}^n w_i^2$. If the weights are too imbalanced then the result is similar to averaging only $n_{\mathrm{e}} \ll n$ observations and might therefore be unreliable. The point at which n_{e} becomes alarmingly small is hard to specify, because it is application specific.

The effective sample size (9.13) is computed from the generated values. For simple enough distributions, we can obtain a population version of n_{e}, as

$$n_{\mathrm{e}}^* = \frac{n\,\mathbb{E}_q(w)^2}{\mathbb{E}_q(w^2)} = \frac{n}{\mathbb{E}_q(w^2)} = \frac{n}{\mathbb{E}_p(w)}. \tag{9.14}$$

If n_{e}^* is too small, then we know q will produce imbalanced weights.

Another effective sample size formula is

$$\frac{n}{1 + \mathrm{cv}(\boldsymbol{w})^2} \tag{9.15}$$

where $\mathrm{cv}(\boldsymbol{w}) = \left((n-1)^{-1}\sum_{i=1}^n (w_i - \bar{w})^2\right)^{1/2}/\bar{w}$ is the coefficient of variation of the weights. If we replace the $n-1$ by n in $\mathrm{cv}(\boldsymbol{w})$ and substitute into $n/(1+\mathrm{cv}(\boldsymbol{w})^2)$, the result is n_{e} of (9.13), so the two formulas are essentially the same.

Instead of using an effective sample size, we might just estimate the variance of $\hat{\mu}_q$ (or of $\widetilde{\mu}_q$) and use the variance as a diagnostic. When that variance is quite large we would conclude that the importance sampling had not worked well. Unfortunately the variance estimate is itself based on the same weights that the estimate has used. Badly skewed weights could give a badly estimated mean along with a bad variance estimate that masks the problem.

The variance estimate (9.9) for $\widetilde{\mu}_q$ uses a weighted average of $(f(\boldsymbol{x}_i)-\widetilde{\mu}_q)^2$ with weights equal to w_i^2. Similarly, for $\hat{\mu}_q$ we have $\hat{\sigma}_q^2 = (1/n)\sum_{i=1}^n w_i^2 f(\boldsymbol{x}_i)^2 - \hat{\mu}_q^2$ which again depends on weights w_i^2. We can compute an effective sample size for variance estimates by

$$n_{\mathrm{e},\sigma} = \frac{\bigl(\sum_{i=1}^n w_i^2\bigr)^2}{\sum_{i=1}^n w_i^4}. \tag{9.16}$$

If $n_{\mathrm{e},\sigma}$ is small, then we cannot trust the variance estimate. Estimating a variance well is typically a harder problem than estimating a mean well, and here we have $n_{\mathrm{e},\sigma} \leqslant n_{\mathrm{e},\mu}$ (Exercise 9.17).

The accuracy of the central limit theorem for a mean depends on the skewness of the random variables being summed. To gauge the reliability of the CLT for $\hat{\mu}_q$ and $\widetilde{\mu}_q$ we find that the skewness of $S_{\boldsymbol{w}}$ matches that of the average of

$$n_{\mathrm{e},\gamma} = \frac{\bigl(\sum_i w_i^2\bigr)^3}{\bigl(\sum_i w_i^3\bigr)^2}$$

equally weighted observations (Exercise 9.19).

Effective sample sizes are imperfect diagnostics. When they are too small then we have a sign that the importance sampling weights are problematic. When they are large we still cannot conclude that importance sampling has worked. It remains possible that some important region was missed by all of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

The diagnostics above used the weights w_i only through their relative values $w_i/\sum_j w_j$. In cases where w_i is computable, it is the observed value of a random variable with mean $\mathbb{E}_q(p(\boldsymbol{X})/q(\boldsymbol{X})) = 1$. If the sample mean of the weights is far from 1 then that is a sign that q was poorly chosen. That is, $\bar{w} = (1/n)\sum_{i=1}^n w_i$ is another diagnostic.

These weight based diagnostics are the same for every function f. They are convenient summaries, but the effectiveness of importance sampling depends also on f. If every $f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i) = 0$ it is likely that our importance sampling has failed and no further diagnostic is needed. Otherwise, for a diagnostic that is specific to f let

$$\widetilde{w}_i(f) = \frac{|f(\boldsymbol{X}_i)|p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i)}{\sum_{j=1}^n |f(\boldsymbol{X}_j)|p(\boldsymbol{X}_j)/q(\boldsymbol{X}_j)}.$$

After some algebra, the sample coefficient of variation of these weights is $\mathrm{cv}(\boldsymbol{w}, f) = \sqrt{n\sum_{i=1}^n \widetilde{w}_i^2 - 1}$. An effective sample size customized to f is

$$n_{\mathrm{e}}(f) = \frac{n}{1+\mathrm{cv}(\boldsymbol{w}, f)^2} = \frac{1}{\sum_{i=1}^n \widetilde{w}_i(f)^2}. \tag{9.17}$$

There is a further discussion of effective sample size on page 39 of the chapter end notes.

j	Task	Predecessors	Duration
1	Planning	None	4
2	Database Design	1	4
3	Module Layout	1	2
4	Database Capture	2	5
5	Database Interface	2	2
6	Input Module	3	3
7	Output Module	3	2
8	GUI Structure	3	3
9	I/O Interface Implementation	5,6,7	2
10	Final Testing	4,8,9	2

Table 9.1: A PERT problem from Chinneck (2009, Chapter 11). Used with the author's permission. Here we replace the durations by exponential random variables. This example was downloaded from `http://optlab-server.sce.carleton.ca/POAnimations2007/PERT.html`.

9.4 Example: PERT

Importance sampling can be very effective but it is also somewhat delicate. Here we look at a numerical example and develop an importance sampler for it. The example is based on Project Evaluation and Review Technique (PERT), a project planning tool. Consider the software project described in Table 9.1. It has 10 tasks (activities), indexed by $j = 1, \dots, 10$. The project is completed when all of the tasks are completed. A task can begin only after all of its predecessors have been completed. Figure 9.1 shows the predecessor-successor relations in this example, with a node for each activity.

The project starts at time 0. Task j starts at time S_j, takes time T_j and ends at time $E_j = S_j + T_j$. Any task j with no predecessors (here only task 1) starts at $S_j = 0$. The start time for a task with predecessors is the maximum of the ending times of its predecessors. For example, $S_4 = E_2$ and $S_9 = \max(E_5, E_6, E_7)$. The project as a whole ends at time E_{10}.

The duration θ_j for task j is listed in Table 9.1. If every task takes the time listed there, then the project takes $E_{10} = 15$ days. For this example, we change the PERT problem to make T_j independent exponentially distributed random variables with means given in the final column of the table. The completion time is then a random variable with a mean just over 18 days, and a long tail to the right, as can be easily seen from a simple Monte Carlo.

Now suppose that there will be a severe penalty should the project miss a deadline in 70 days time. In a direct simulation of the project $n = 10{,}000$ times, $E_{10} > 70$ happened only 2 times. Missing that deadline is a moderately rare event and importance sampling can help us get a better estimate of its probability.

Here $(T_1, \dots, T_{10})$ plays the role of $\boldsymbol{X}$. The distribution p is comprised of

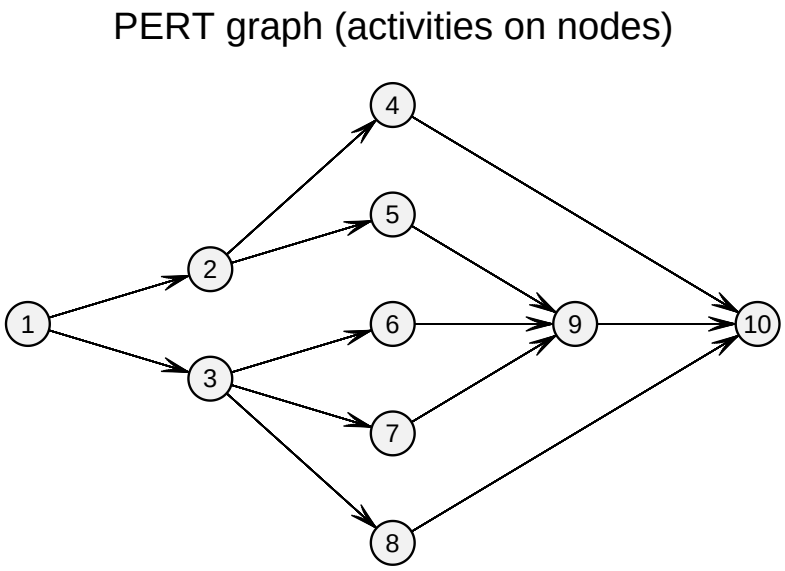

Figure 9.1: This figure shows the 10 activities for the PERT example in Table 9.1. The arrows point to tasks from their predecessors.

independent exponential random variables, $T_j \sim \theta_j \times \mathrm{Exp}(1)$. The function f is 1 for $E_{10}(T_1, \ldots, T_{10}) > 70$ and 0 otherwise. For simplicity, we take the distribution q to be that of independent random variables $T_j \sim \lambda_j \times \mathrm{Exp}(1)$. The importance sampling estimate of $\mu = \mathbb{P}(E_{10} > 70)$ is

$$\hat{\mu}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{E_{i,10} > 70\} \prod_{j=1}^{10} \frac{\exp(-T_{ij}/\theta_i)/\theta_i}{\exp(-T_{ij}/\lambda_i)/\lambda_i}.$$

We want to generate more $E_{10} > 70$ events, and we can do this by taking $\lambda_i > \theta_i$. We can make the mean of E_{10} close to 70 by taking $\lambda = 4\theta$. Importance sampling with $\lambda = 4\theta$ and $n = 10{,}000$ gave $\hat{\mu}_\lambda = 2.02 \times 10^{-5}$ with a standard error of 4.7×10^{-6}. It is a good sign that the standard error is reasonably small compared to the estimate. But checking the effective sample sizes, we find that

$$\begin{pmatrix} n_{\mathrm{e},\mu} & n_{\mathrm{e},\sigma} & n_{\mathrm{e},\gamma} \end{pmatrix} \doteq \begin{pmatrix} 4.90 & 1.15 & 1.21 \end{pmatrix}.$$

Simply scaling the means by a factor of 4 has made the weights too uneven. The largest weight was about 43% of the total weight. Inspecting the standard error did not expose the problem with these weights. Similarly, the average of $p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$ was 0.997 which is not alarmingly far from 1. This problem is simple enough that we could have computed n^*_{e} (Exercise 9.20).

Multiplying all the durations by 4 is too much of a distortion. We can get better results by using somewhat smaller multiples of θ. When exploring changes in λ it is advantageous to couple the simulations. One way to do this is

Task	1	2	3	4	5	6	7	8	9	10
θ	4	4	2	5	2	3	2	3	2	2
$\Delta^{(+1)}\, E_{10}$	1	1		1						1
$\Delta^{(\times 2)}\, E_{10}$	4	4		5	1	1			1	2

Table 9.2: For tasks $j = 1, \ldots, 10$ the mean durations are in the row θ. The row $\Delta^{(+1)}$ shows the increase in completion time E_{10} when we add one to T_j (getting $\theta_j + 1$) keeping all other durations at their means. Only the nonzero changes are shown. The row $\Delta^{(\times 2)}$ similarly shows the delay from doubling the time on task j to $2\theta_j$, keeping other times at their means.

to generate an $n \times 10$ matrix X of Exp(1) random variables and take $T_{ij} = X_{ij}\lambda_j$. We find this way that there is a limit to how well scale multiples of $\lambda = \kappa\theta$ can perform. Each component of θ that we increase makes the importance sampling weights more extreme. But the components do not all contribute equally to moving E_{10} out towards 70. We may do better by using a larger multiple on the more important durations.

Table 9.2 shows the results of some simple perturbations of $T_1, \ldots, T_{10}$. If all tasks but one are held fixed at their mean duration and the other is increased by 1 day, then it turns out that E_{10} either stays fixed or increases by 1. It increases by 1 for tasks 1, 2, 4, and 10. Those tasks are on the critical path. If they are delayed then the project is delayed. The other tasks can suffer small delays without delaying the project. Larger delays are different. A task that takes twice as long as expected could delay the project, even if it is not on the critical path. Table 9.2 also shows the effects of these doublings.

Suppose that we take $\lambda_j = \theta_j$ for j not in the critical path and $\lambda_j = \kappa\theta_j$ for j in the critical path $\{1, 2, 4, 10\}$. The estimates below

κ	3.0	3.5	4.0	4.5	5.0
$10^5 \times \hat{\mu}$	3.3	3.6	3.1	3.1	3.1
$10^6 \times \mathrm{se}(\hat{\mu})$	1.8	2.1	1.6	1.5	1.6
$n_{\mathrm{e},\mu}$	939	564	359	239	165
$n_{\mathrm{e},\sigma}$	165	88	52	33	23
$n_{\mathrm{e},\gamma}$	52	28	17	12	8

were obtained from $n = 10{,}000$ simulations using multiples λ_j of common random numbers X_{ij}.

The effective sample sizes move as we would predict. The estimates and standard errors do not move much. There does not seem to be a clearly superior choice among these values of κ. Using $\kappa = 4$ and $n = 200{,}000$ and a different random seed, the estimate $\hat{\mu}_q = 3.18 \times 10^{-5}$ was found. At this level of precision $\hat{\mu}_q \doteq \widetilde{\mu}_q \doteq \hat{\mu}_{q,\mathrm{reg}}$, the latter using $p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$ as a control variate. The standard error for $\hat{\mu}_q$ was 3.62×10^{-7}. The self-normalized importance sampling estimate $\widetilde{\mu}_q$ had a larger standard error of 5.22×10^{-7}. The standard error of $\hat{\mu}_{q,\mathrm{reg}}$ was barely smaller than the one for $\hat{\mu}_q$ but they agreed to the

precision given here. The effective sample sizes, rounded to integer values, were $\begin{pmatrix} n_{\mathrm{e},\mu} & n_{\mathrm{e},\sigma} & n_{\mathrm{e},\gamma} \end{pmatrix} \doteq \begin{pmatrix} 7472 & 992 & 1472 \end{pmatrix}$. The mean weight was 0.9992 and the coefficient of variation for $f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$, was 5.10 leading to an f-specific effective sample size (equation (9.17)) of $n_{\mathrm{e}}(f) \doteq 7405$.

We can estimate the variance reduction due to importance sampling as

$$\frac{\hat{\mu}_q(1-\hat{\mu}_q)/n}{\mathrm{se}(\hat{\mu}_q)^2} \doteq \frac{\hat{\mu}_q/n}{\mathrm{se}(\hat{\mu}_q)^2} \doteq 1209.$$

That is $n = 200{,}000$ samples with importance sampling are about as effective as roughly 242 million samples from the nominal distribution. Plain importance sampling was roughly twice as efficient as self-normalized importance sampling on this example: $\mathrm{se}(\widetilde{\mu}_q)^2/\mathrm{se}(\hat{\mu}_q)^2 \doteq 2.08$.

It is possible to improve still further. This importance sampler quadrupled all the task durations on the critical path. But quadrupling the 10'th time incurs the same likelihood ratio penalty as quadrupling the others without adding as much to the duration. We might do better using a larger multiple for task 5 and a smaller one for task 10. Of course, any importance sampler yielding positive variance can be improved and we must at some point stop improving the estimate and move on to the next problem.

It is not necessary to sample the 10'th duration T_{10}, from the nominal distribution or any other. Once E_4, E_8 and E_9 are available, either $S_{10} \geqslant 70$ in which case we are sure that $E_{10} > 70$, or $S_{10} < 70$ and we can compute $\mathbb{P}(E_{10} > 70 \,|\, T_1, \ldots, T_9) = \exp(-(70 - S_{10})/\theta_{10})$ using the tail probability for an exponential distribution. That is, we can use conditional Monte Carlo. Here two principles conflict. We should use conditional Monte Carlo on principle because using known integrals where possible reduces variance. A contrary principle is that we do better to stay with simple methods because human time is worth much more than computer time. For this example, we take the second principle and consider (but do not implement) some ways in which the analysis might be sharpened.

It is not just the 10'th time that can be replaced by conditional Monte Carlo. Given all the durations except T_j, there is a threshold $m \geqslant 0$ such that $E_{10} > 70$ when $T_j > m$. We could in principle condition on any one of the durations. Furthermore, it is clear that antithetic sampling would reduce variance because E_{10} is monotone in the durations $T_1, \ldots, T_{10}$. See Exercise 8.4 for the effects of antithetic sampling on spiky functions.

In this example importance sampling worked well but had to be customized to the problem. Finding the critical path and making it take longer worked well. That cannot be a general finding for all PERT problems. There may sometimes be a second path, just barely faster than the critical path, and having little or no overlap with the critical path. Increasing the mean duration of the critical path alone might leave a very high sampling variance due to contributions from the second path. Importance sampling must be used with some caution. Defensive importance sampling and other mixture based methods in §9.11 provide some protection.

9.5 Importance sampling versus acceptance-rejection

Importance sampling and acceptance-rejection sampling are quite similar ideas. Both of them distort a sample from one distribution in order to sample from another. It is natural then to compare the two methods. There are two main features on which to compare the methods: implementation difficulty, and efficiency. Importance sampling is usually easier to implement. Either one can be more efficient.

For implementation, let p be the target density and q the proposal density. We might only have unnormalized versions $\widetilde{p}$ and $\widetilde{q}$. Ordinary importance sampling requires that we can compute $p(\boldsymbol{x})/q(\boldsymbol{x})$. Self-normalized importance sampling has the weaker requirement that we can compute $\widetilde{p}(\boldsymbol{x})/\widetilde{q}(\boldsymbol{x})$. Acceptance-rejection sampling can be applied with either the ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$ or $\widetilde{p}(\boldsymbol{x})/\widetilde{q}(\boldsymbol{x})$. But either way we must know an upper bound c for this ratio in order to use acceptance-rejection. We don't need to know the bound for importance sampling. Furthermore, when $\sup_{\boldsymbol{x}} p(\boldsymbol{x})/q(\boldsymbol{x}) = \infty$ then acceptance-rejection is impossible while importance sampling is still available, though it might have a high variance.

As a result importance sampling is easier to implement than acceptance-rejection sampling, for straightforward estimation of μ as above. On the other hand, if the problem involves many different random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ each being generated in different places in our software, then using exact samples from each distribution may be easier to do than managing and combining importance weights for every variable in the system.

To compare the efficiencies we have to take account of costs. At one extreme, suppose that almost all of the cost is in computing f so that the costs of sampling from q and deciding whether to accept or reject that sample are negligible. Then we would use the same value of n in either importance sampling or acceptance-rejection sampling. As a result the efficiency comparison is the same as that between importance sampling and IID sampling. In this case, the relative efficiency of importance sampling from q to acceptance-rejection is σ_p^2/σ_q^2. For self-normalized importance sampling it is $\sigma_p^2/\sigma_{q,\mathrm{sn}}^2$.

At the other extreme, suppose that almost all of our costs are in sampling from q and computing the ratio q/p, while f is very inexpensive. Maybe $\boldsymbol{X}$ is a complicated stochastic process and f is simply whether it ends at a positive value. Then in the time it takes us to get n samples by importance sampling we would have accepted a random number of samples by acceptance-rejection. This number of samples will have the $\mathrm{Bin}(n, 1/c)$ distribution. For large n we might as well suppose it is n/c points where the acceptance-rejection rule keeps points when $p(\boldsymbol{X})/q(\boldsymbol{X}) \leqslant c$. In this extreme setting the efficiency of importance sampling is $c\sigma_p^2/\sigma_q^2$ and that of self-normalized importance sampling is $c\sigma_p^2/\sigma_{q,\mathrm{sn}}^2$.

We can get this same efficiency answer from a longer argument writing acceptance-rejection estimate as $\hat{\mu}_{AR} = \sum_{i=1}^n A_i Y_i / \sum_{i=1}^n A_i$, where A_i is 1 if $\boldsymbol{X}_i$ is accepted and 0 otherwise, and then using the ratio estimation variance formula (2.29) of §2.7.

Family	$p(\cdot\,;\theta)$	$w(\cdot)$	Θ
Normal	$\mathcal{N}(\theta,\Sigma)$	$\exp(\boldsymbol{x}^{\mathsf{T}}\Sigma^{-1}(\theta_0-\theta)+\frac{1}{2}\theta^{\mathsf{T}}\Sigma^{-1}\theta-\frac{1}{2}\theta_0^{\mathsf{T}}\Sigma^{-1}\theta_0)$	$\mathbb{R}^d$
Poisson	$\mathrm{Poi}(\theta)$	$\exp(\theta-\theta_0)(\theta_0/\theta)^x$	$(0,\infty)$
Binomial	$\mathrm{Bin}(m,\theta)$	$(\theta_0/\theta)^x((1-\theta_0)/(1-\theta))^{m-x}$	$(0,1)$
Gamma	$\mathrm{Gam}(\theta)$	$x^{\theta_0/\theta}\,\Gamma(\theta)/\Gamma(\theta_0)$	$(0,\infty)$

Table 9.3: Some example parametric families with their importance sampling ratios. The normal family shown shares a non-singular Σ.

We expect that our problem is somewhere between these extremes, and so the efficiency of importance sampling compared to acceptance-rejection is somewhere between σ_p^2/σ_q^2 and $c\sigma_p^2/\sigma_q^2$. The relative efficiency of self-normalized importance sampling to acceptance rejection is between $\sigma_p^2/\sigma_{q,\mathrm{sn}}^2$ and $c\sigma_p^2/\sigma_{q,\mathrm{sn}}^2$.

9.6 Exponential tilting

There are a small number of frequently used strategies for picking the importance distribution q. Here we describe exponential tilting. In §9.7 we consider matching the Hessian of q to that of p at the mode.

Often the nominal distribution p is a member of a parametric family, which we may write as $p(\boldsymbol{x}) = p(\boldsymbol{x};\theta_0)$. When we can sample from any member of this family it becomes convenient to define the importance sampling distribution by a change of parameter: $q_\theta(\boldsymbol{x}) = p(\boldsymbol{x};\theta)$. Many examples in this chapter are of this type. Table 9.3 gives the weight function $w(\boldsymbol{x})$ for several parametric families.

The example distributions in Table 9.3 are all exponential families. In an ***exponential family***,

$$p(\boldsymbol{x};\theta) = \exp(\eta(\theta)^{\mathsf{T}}T(\boldsymbol{x}) - A(\boldsymbol{x}) - C(\theta)), \quad \theta\in\Theta,$$

for functions $\eta(\cdot)$, $T(\cdot)$, $A(\cdot)$ and $C(\cdot)$. The normalizing constant (in the denominator of p) is $c(\theta) = \exp(C(\theta))$. Here p may be a probability density function or a probability mass function depending on the context. It takes a bit of rearrangement to put a distribution into the exponential family framework. In this form we find that

$$w(\boldsymbol{x}) = \frac{p(\boldsymbol{x};\theta_0)}{p(\boldsymbol{x};\theta)} = e^{(\eta(\theta_0)-\eta(\theta))^{\mathsf{T}}T(\boldsymbol{x})}\times c(\theta)/c(\theta_0).$$

The factor $\exp(-A(\boldsymbol{x}))$ cancels, resulting in a simpler formula that may also be much faster than computing the ratio explicitly.

Often $\eta(\theta)=\theta$ and $T(\boldsymbol{x})=\boldsymbol{x}$ and then $w(\boldsymbol{x}) = \exp((\theta_0-\theta)^{\mathsf{T}}\boldsymbol{x})c(\theta)/c(\theta_0)$. In this case, the importance sampling estimate takes the particularly simple

form

$$\hat{\mu}_\theta = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) e^{(\theta_0-\theta)^{\mathsf{T}}\boldsymbol{X}_i} \times \frac{c(\theta)}{c(\theta_0)}.$$

Importance sampling by changing the parameter in an exponential family is variously called ***exponential tilting*** or ***exponential twisting***. The normalizing constants for popular exponential families can usually be looked up. The exponential form is particularly convenient when $\boldsymbol{X}$ is a stochastic process. We will see an example in §9.9.

Exponential tilting can be used outside of exponential families. If $p(\boldsymbol{x})$ is a density function then $q_\eta(\boldsymbol{x}) \propto p(\boldsymbol{x})\exp(-\eta^{\mathsf{T}}T(\boldsymbol{x}))$ may still be a useful density (if it can be normalized). But q_η will no longer be in the same parametric family as p, perhaps requiring new sampling algorithms.

9.7 Modes and Hessians

In a Bayesian setting the target distribution is the posterior distribution of a parameter. In keeping with that context we will use the letter X to represent some data and the quantity that we importance sample will be represented by θ. A very general approach is to approximate the log posterior distribution by a quadratic Taylor approximation around its mode. That suggests a Gaussian distribution with covariance matrix equal to the inverse of the Hessian matrix of second derivatives. We use logistic regression as a concrete example but the concepts apply to many parametric statistical models.

In logistic regression, the binary random variable $Y_i \in \{0,1\}$ is related to a vector $\boldsymbol{x}_i \in \mathbb{R}^d$ by

$$\mathbb{P}(Y_i = 1) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)}{1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)} = \frac{1}{1+\exp(-\boldsymbol{x}_i^{\mathsf{T}}\theta)}.$$

Typically the first component of $\boldsymbol{x}_i$ is 1 which puts an intercept term in the model. The Y_i are independent and we let $\mathcal{Y} = (Y_1, \ldots, Y_n)$.

The Bayesian approach models θ as a random variable with a prior distribution $\pi(\theta)$. We often work with an un-normalized version $\pi_u(\theta) \propto \pi(\theta)$. Then the posterior distribution of θ given n independent observations is $\pi(\theta \,|\, \mathcal{Y}) = \pi_u(\theta \,|\, \mathcal{Y})/c$, for an un-normalized posterior distribution

$$\pi_u(\theta \,|\, \mathcal{Y}) = \pi_u(\theta)\prod_{i=1}^{n}\Big(\frac{\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)}{1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)}\Big)^{Y_i}\Big(\frac{1}{1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)}\Big)^{1-Y_i}.$$

We might not know the normalizing constant c even when we know the constant for $\pi(\theta)$. Sometimes an improper prior $\pi_u(\theta) = 1$ is used. That prior cannot be normalized and then whether $\pi_u(\theta \,|\, \mathcal{Y})$ can be normalized depends on the data.

The logarithm of the posterior distribution is

$$-\log(c) + \log(\pi_u(\theta)) + \sum_{i=1}^{n} Y_i \boldsymbol{x}_i^{\mathsf{T}}\theta - \sum_{i=1}^{n}\log(1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta)).$$

The sums over i grow with n while $\pi_u(\theta)$ does not, and so for n large enough the data dominate $\log(\pi(\theta))$. The first sum is linear in θ. Minus the second sum is a concave function of θ. We will assume that $\log(\pi_u(\theta\,|\,\mathcal{Y}))$ is strictly concave and takes its maximum at a finite value. With $\pi_u(\theta) = 1$, that could fail to happen, trivially if all $Y_i = 0$, or in a more complicated way if the $\boldsymbol{x}_i$ with $Y_i = 0$ can be linearly separated from those with $Y_i = 1$. Such separability failures may be rare when $d \ll n$ but they are not rare if d and n are of comparable size. We can force strict concavity through a choice of π_u that decays at least as fast as $\exp(-\|\theta\|^{1+\epsilon})$ for some $\epsilon > 0$.

Let θ_* be the maximizer of $\pi_u(\theta\,|\,\mathcal{Y})$. If $\pi_u(\theta)$ is not too complicated then we may be able to compute θ_* numerically. The first derivatives of $\log(\pi_u(\theta\,|\,\mathcal{Y}))$ with respect to θ vanish at this optimum. Then a Taylor approximation around this maximum a posteriori estimate is

$$\log(\pi_u(\theta\,|\,\mathcal{Y})) \doteq \log(\pi_u(\theta_*\,|\,\mathcal{Y}) - \frac{1}{2}(\theta-\theta_*)^{\mathsf{T}} H_*(\theta-\theta_*) \tag{9.18}$$

where

$$H_* = -\frac{\partial^2}{\partial\theta\partial\theta^{\mathsf{T}}}\log(\pi_u(\theta))|_{\theta=\theta_*}$$

is (minus) the matrix of second derivatives of $\log(\pi_u(\theta\,|\,\mathcal{Y}))$. We can usually approximate H_* by divided differences and, for convenient π_u, we can get H_* in closed form.

We interpret (9.18) as $\theta \mid \mathcal{Y} \mathrel{\dot\sim} \mathcal{N}(\theta_*, H_*^{-1})$. Then

$$\mathbb{E}(f(\theta)\,|\,\mathcal{Y}) = \frac{\int_{\mathbb{R}^d} f(\theta)\pi_u(\theta\,|\,\mathcal{Y})\,\mathrm{d}\theta}{\int_{\mathbb{R}^d}\pi_u(\theta\,|\,\mathcal{Y})\,\mathrm{d}\theta} \doteq \mathbb{E}\big(f(\theta)\,|\,\theta\sim\mathcal{N}(\theta_*, H_*^{-1})\big). \tag{9.19}$$

This approximation may be good enough for our purposes and depending on $f(\cdot)$ we may or may not have to resort to Monte Carlo sampling from $\mathcal{N}(\theta_*, H_*^{-1})$.

When we want to estimate posterior properties of θ without the Gaussian approximation then we can importance sample from a distribution similar to $\pi_u(\theta\,|\,\mathcal{Y})$. Even when we believe that $\mathcal{N}(\theta_*, H_*^{-1})$ is fairly close to $\pi_u(\theta\,|\,\mathcal{Y})$ it is unwise to importance sample from a Gaussian distribution. The Gaussian distribution has very light tails.

A popular choice is to apply self-normalized importance sampling drawing IID samples θ_i from the multivariate t distribution $\mathrm{t}(\theta_*, H_*^{-1}, \nu)$ of §5.2. Then we estimate $\mathbb{E}(f(\theta)\,|\,\mathcal{Y})$ by

$$\frac{\sum_{i=1}^n f(\theta_i)\pi_u(\theta_i\,|\,\mathcal{Y})\big(1+(\theta_i-\theta_*)^{\mathsf{T}}H_*(\theta_i-\theta_*)\big)^{-(\nu+d)/2}}{\sum_{i=1}^n \pi_u(\theta_i\,|\,\mathcal{Y})\big(1+(\theta_i-\theta_*)^{\mathsf{T}}H_*(\theta_i-\theta_*)\big)^{-(\nu+d)/2}}.$$

The degrees of freedom ν determine how heavy the tails of the t distribution are. They decay as $(1+\|\theta-\theta_*\|^2)^{-(\nu+d)/2} \approx \|\theta-\theta_*\|^{-\nu-d}$ for large $\|\theta\|$. For the logistic regression example above $\log(1+\exp(z))$ is asymptotic to z as $z\to\infty$ and to 0 as $z\to-\infty$. It follows that the data contribution to the

log posterior decays at the rate $\exp(-\|\theta\|)$. Any ν will lead to an importance sampling distribution with heavier tails than the posterior distribution.

There have been several refinements of this Hessian approach to account for skewness or multimodality of the posterior distribution. See page 38 of the chapter end notes.

The method cannot be applied when H_* is not invertible. For a simple one dimensional example of this problem, let the posterior distribution be proportional to $\exp(-(\theta-\theta_*)^4)$ for $\theta \in \mathbb{R}$. Then $H_* = 0$.

If we have a normalized posterior distribution $\pi(\theta \mid \mathcal{Y})$ then we might consider ordinary importance sampling. There the optimal importance distribution is proportional to $f(\theta)\pi(\theta \mid \mathcal{Y})$. For $f \geqslant 0$ we might then choose θ_* to maximize $f(\theta)\pi(\theta \mid \mathcal{Y})$ and take H_* from the Hessian matrix of $\log(f(\theta_*)) + \log(\pi(\theta_* \mid \mathcal{Y}))$. The result is very much like the Laplace approximation method in quadrature (§7.6).

9.8 General variables and stochastic processes

So far we have assumed that the nominal and sampling distributions of $\boldsymbol{X}$ are continuous and finite dimensional. Here we extend importance sampling to more general random variables and to processes.

For discrete random variables $\boldsymbol{X}$, the ratio $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$ is now a ratio of probability mass functions. The sampling distribution q has to satisfy $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x})f(\boldsymbol{x}) \neq 0$. The variance of $\hat{\mu}_q$ for discrete sampling is σ_q^2/n where

$$\sigma_q^2 = \sum_j \frac{(f(x_j)p(x_j) - \mu q(x_j))^2}{q(x_j)}$$

taking the sum over all j with $q(x_j) > 0$. For self-normalized importance sampling we require $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$.

We can also use importance sampling on mixed random variables having both continuous and discrete components. Suppose that with probability λ the vector $\boldsymbol{X}$ is drawn from a discrete mass function p_d while with probability $1-\lambda$ it is sampled from a probability density function p_c. That is $p = \lambda p_d + (1-\lambda)p_c$ for $0 < \lambda < 1$. Now we sample from a distribution $q = \eta q_d + (1-\eta)q_c$ where $0 < \eta < 1$ for a discrete mass function q_d and a probability density function p_c. We assume that $q_c(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p_c(\boldsymbol{x}) \neq 0$ and that $q_d(\boldsymbol{x}) > 0$ whenever $p_d(\boldsymbol{x})f(\boldsymbol{x}) \neq 0$. Then the importance sampling estimate of $\mu = \mathbb{E}_p(f(\boldsymbol{X}))$ is

$$\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)w(\boldsymbol{X}_i)$$

where $\boldsymbol{X}_i \sim q$ and

$$w(\boldsymbol{x}) = \begin{cases} \dfrac{\lambda p_d(\boldsymbol{x})}{\eta q_d(\boldsymbol{x})}, & q_d(\boldsymbol{x}) > 0 \\ \dfrac{(1-\lambda) p_c(\boldsymbol{x})}{(1-\eta) q_c(\boldsymbol{x})}, & q_d(\boldsymbol{x}) = 0. \end{cases}$$

Notice that $w(\boldsymbol{x})$ is determined only by the discrete parts at points $\boldsymbol{x}$ for which both $q_d(\boldsymbol{x}) > 0$ and $q_c(\boldsymbol{x}) > 0$. The discrete component trumps the continuous one.

Now suppose that $\boldsymbol{X} = (X_1, X_2, \dots)$ is a process of potentially infinite dimension with nominal distribution $p(\boldsymbol{x}) = \prod_{j \geqslant 1} p_j(x_j \,|\, x_1, \dots, x_{j-1})$. For $j = 1$ we interpret $p_j(x_j \,|\, x_1, \dots, x_{j-1})$ as simply $p_1(x_1)$.

If we importance sample with $q(\boldsymbol{x}) = \prod_{j \geqslant 1} q_j(x_j \,|\, x_1, \dots, x_{j-1})$ then the likelihood ratio is

$$w(\boldsymbol{x}) = \prod_{j \geqslant 1} \frac{p_j(x_j \,|\, x_1, \dots, x_{j-1})}{q_j(x_j \,|\, x_1, \dots, x_{j-1})}.$$

An infinite product leads to very badly distributed weights unless we arrange for the factors to rapidly approach 1.

In an application, we could never compute an infinite number of components of $\boldsymbol{x}$. Instead we apply importance sampling to problems where we can compute $f(\boldsymbol{X})$ from just an initial subsequence $X_1, X_2, \dots, X_M$ of the process. The time M at which the value of $f(\boldsymbol{X})$ becomes known is generally random. When $M = m$, then $f(\boldsymbol{X}) = f_m(X_1, \dots, X_m)$ for some function f_m. Assuming that the i'th sampled process generates $x_{i1}, \dots, x_{iM_i}$ where $M_i = m < \infty$ we will truncate the likelihood ratio at m factors using

$$w_m(\boldsymbol{x}) = \prod_{j=1}^{m} \frac{p_j(x_j \,|\, x_1, \dots, x_{j-1})}{q_j(x_j \,|\, x_1, \dots, x_{j-1})}.$$

We have assumed that $\mathbb{P}_q(M < \infty) = 1$. A simulation without this property might fail to terminate so we need it. For efficiency, we need the stronger condition that $\mathbb{E}_q(M) < \infty$. We also need $M = M(\boldsymbol{X})$ to be a ***stopping time***, which means that we can tell whether $M_i \leqslant m$ by looking at the values $x_{i1}, \dots, x_{im}$. For example, the hitting time $M_i = \min\{m \geqslant 1 \mid x_{im} \in A\}$ of a set A is a valid stopping time as is $\min(M_i, 1000)$, or more generally the minimum of several stopping times. But $M_i - 1$ is generally not a valid stopping time because we usually can't tell that the process is just about to hit the set A.

For a process, the importance sampling estimator is

$$\hat{\mu}_q = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) w_{M_i}(\boldsymbol{X}_i).$$

Next, we find $\mathbb{E}_q(\hat{\mu}_q)$. We will assume that $q_j(x_j \,|\, x_1, \dots, x_{j-1}) > 0$ whenever $p_j(x_j \,|\, x_1, \dots, x_{j-1}) > 0$.

We assume that $\mathbb{E}_q(f(\boldsymbol{X})w_M(\boldsymbol{X}))$ exists and $\mathbb{P}_q(M < \infty) = 1$. Then, recalling that $\sum_{m=1}^{\infty}$ denotes $\sum_{1 \leqslant m < \infty}$,

$$\begin{aligned}
\mathbb{E}_q\big(f(\boldsymbol{X})w_{M(\boldsymbol{X})}(\boldsymbol{X})\big) &= \sum_{m=1}^{\infty} \mathbb{E}_q\big(f_m(\boldsymbol{X})w_{M(\boldsymbol{X})}(\boldsymbol{X})\mathbb{1}_{M(\boldsymbol{X})=m}\big) \\
&= \sum_{m=1}^{\infty} \mathbb{E}_q\left(f_m(\boldsymbol{X})\prod_{j=1}^{m}\frac{p_j(X_j \mid X_1,\dots,X_{j-1})}{q_j(X_j \mid X_1,\dots,X_{j-1})}\mathbb{1}_{M(\boldsymbol{X})=m}\right) \\
&= \sum_{m=1}^{\infty} \mathbb{E}_p(f_m(\boldsymbol{X})\mathbb{1}_{M(\boldsymbol{X})=m}) \\
&= \mathbb{E}_p\big(f(\boldsymbol{X})\mathbb{1}_{M(\boldsymbol{X})<\infty}\big).
\end{aligned}$$

We needed $\mathbb{P}(M < \infty) = 1$ for the first equality. For the third equality, $\mathbb{1}_{M(\boldsymbol{X})=m}$ is a function of $X_1,\dots,X_m$ because M is a stopping time. Therefore the whole integrand there is an expectation over $X_1,\dots,X_m$ from q, which can be replaced by expectation from p, after canceling the likelihood ratio.

We have found that

$$\mathbb{E}_q(\hat{\mu}_q) = \mathbb{E}_p(f(\boldsymbol{X})\mathbb{1}_{M<\infty}).$$

If $\mathbb{P}_q(M < \infty) = \mathbb{P}_p(M < \infty) = 1$ then $\mathbb{E}_q(\hat{\mu}_q) = \mathbb{E}_p(f(\boldsymbol{X}))$ as we might naively have expected. But in general we can have $\mathbb{E}_q(\hat{\mu}_q) \neq \mathbb{E}_p(f(\boldsymbol{X}))$. The difference can be quite useful. For example, suppose that we want to find $\mathbb{P}_p(M < \infty)$. Then we can define $f(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$ and look among distributions q with $\mathbb{P}_q(M < \infty) = 1$ for one that gives a good estimate. Example 9.5 in §9.9 finds the probability that a Gaussian random walk with a negative drift ever exceeds some positive value.

9.9 Example: exit probabilities

A good example of importance sampling is in Siegmund's method for computing the probability that a random walk reaches the level b before reaching a where $a \leqslant 0 < b$. Suppose that X_j are IID random variables with distribution p having mean $\mu < 0$ and $\mathbb{P}(X > 0) > 0$. Let $S_0 = 0$, and for $m \geqslant 1$, let $S_m = \sum_{j=1}^{m} X_j$. Define $M = M_{a,b} = \min\{m \geqslant 1 \mid S_m \geqslant b \text{ or } S_m \leqslant a\}$. The process S_m is a random walk starting at 0, taking p-distributed jumps, and M represents the first time, after 0, that the process is outside the interval (a, b). The desired quantity, $\mu = \mathbb{P}(S_M \geqslant b)$, gives the probability that at the time of the first exit from (a, b), the walk is positive. The walk has a negative drift. If b is large enough, then $S_M \geqslant b$ is a rare event.

In one application, X_j represents a gambler's winnings, losses being negative winnings, at the j'th play of the game. The game continues until either the gambler's winnings reach b, or the losses reach $|a|$. Then μ is the probability that the game ends with the gambler having positive winnings.

A second application is the educational testing problem of §6.2. The sequential probability ratio test was driven by a random walk that drifted up for students who had mastered the material and down for those who needed remediation.

Suppose now that p is in an exponential family with parameter θ_0. Let $q(x) = p(x)\exp((\theta-\theta_0)x)c_{\theta_0}/c_\theta$. Then if we run our random walk with $X_i \sim q$ we find that

$$\begin{aligned}\hat{\mu}_q &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{S_{M_i} \geqslant b\} \prod_{j=1}^{M_i} \frac{p(X_{ij})}{q(X_{ij})} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{S_{M_i} \geqslant b\} \Big(\frac{c_\theta}{c_{\theta_0}}\Big)^{M_i} \exp((\theta_0-\theta)S_{M_i}).\end{aligned}$$

The estimator has a very desirable property that the likelihood ratio $w(\boldsymbol{X}) = \prod_{j=1}^{M} p(X_j)/q(X_j)$ only depends on $\boldsymbol{X}$ through the number M of steps taken and the end point S_M, but not the entire path from 0 to S_M.

Often there is a special value $\theta_* \neq \theta_0$ with $c_{\theta_*} = c_{\theta_0}$. If we choose $\theta = \theta_*$ then the estimator simplifies to

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{S_{M_i} \geqslant b\} \exp((\theta_0-\theta_*)S_{M_i})$$

which depends only on the end-point and not even on how many steps it took to get there. A celebrated result of Siegmund (1976) is that this choice θ_* is asymptotically optimal in the limit as $b \to \infty$. Let the variance using importance sampling with tilt θ be σ_θ^2/n. If $\theta \neq \theta_*$ then his Theorem 1 has

$$\lim_{b\to\infty} \frac{\mu^2+\sigma_{\theta_*}^2}{\mu^2+\sigma_\theta^2} = 0$$

at an exponential rate, for any a. This implies not only that θ_* is best but that in this limit no other tilting parameter θ can have σ_θ/μ remain bounded. That result holds for a walk on real values or on integer multiples of some $h > 0$, that is $\{0, \pm h, \pm 2h, \pm 3h, \dots\}$ with b an integer multiple of h.

We can illustrate this method with a normal random walk. Suppose that p is $\mathcal{N}(\theta_0, \sigma^2)$. We can assume that $\sigma = 1$ for otherwise we could divide a, b, θ_0, and σ by σ without changing μ. With $p = \mathcal{N}(\theta_0, 1)$ and $q = \mathcal{N}(\theta, 1)$ we find that $q(x) = p(x)\exp((\theta_0-\theta)x)c(\theta_0)/c(\theta)$ where the normalizing constant is $c(\theta) = \theta^2/2$. Choosing $\theta_* = -\theta_0 > 0$ we obtain $c(\theta_*) = c(\theta_0)$. This choice reverses the drift and now

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{S_{M_i} \geqslant b\} \exp(2\theta_0 S_{M_i}). \tag{9.20}$$

The advantage of reversing the drift parameter becomes apparent when $b > 0$ becomes very large while a remains fixed. For large b the nominal simulation will

seldom cross the upper boundary. The reversed simulation now drifts towards the upper boundary, the important region. The estimate $\hat{\mu}_q$ keeps score, by averaging $\exp(2\theta_0 S_{M_i})$. These are very small numbers, being no larger than $\exp(-2|\theta_0|b)$.

Example 9.5. A numerical example with $a = -5$, $b = 10$, $\theta_0 = -1$ and $n = 50{,}000$ yielded the following results: $\hat{\mu}_{\theta_*} = 6.6 \times 10^{-10}$ with $\sqrt{\widehat{\mathrm{Var}}(\hat{\mu}_{\theta_*})} = 2.57 \times 10^{-12}$. A nominal simulation with μ on the order of 10^{-9} would be extremely expensive. It is clear from (9.20) that $\mu \leqslant \exp(2\theta_0 b) = \exp(-20) \doteq 2.06 \times 10^{-9}$. Because S_M is larger than b we get $\mu < \exp(2\theta_0 b)$.

As $a \to -\infty$ we have $\mu \to \mathbb{P}(\max_{m \geqslant 1} S_m \geqslant b) = \mathbb{P}(M_{-\infty,b} < \infty)$, the probability that the random walk will ever exceed b. Because we reversed the drift, we can be sure that under the importance distribution the walk will always exceed b.

Example 9.6 (Insurance company failure)**.** A similar simulation can be used to estimate the probability that an insurance company is bankrupted. Between claims the amount of money the company receives is a per unit time, from premiums collected minus overhead expenses. Claim i arrives at time T_i and costs Y_i. The company's balance at claim times is a random walk $B_i = B_{i-1} + X_i$ for $X_i = a(T_i - T_{i-1}) - Y_i$. If they have priced risk well then $\mathbb{E}(X_i) > 0$, so the walk drifts up. Their ruin probability when starting with a balance $B_0 > 0$ is $\mathbb{P}(\min_{i \geqslant 1} B_i < 0)$. The random variables X_i are not necessarily from an exponential family, so the likelihood ratio $w(\boldsymbol{x})$ may be more complicated. But the general notion of reversing the drift is still applicable.

9.10 Control variates in importance sampling

Control variates can be usefully combined with importance sampling. A control variate is a vector $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_J(\boldsymbol{x}))^{\mathsf{T}}$ for which $\int h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta$ for a known value $\theta \in \mathbb{R}^J$. The integral is over the entire set of possible $\boldsymbol{x}$ values, such as all of $\mathbb{R}^d$. To use integrals over a subset we make $h(\boldsymbol{x}) = 0$ outside that subset. Some of the equations below simplify if $\theta = 0$. We can make $\theta = 0$ by using $h(\boldsymbol{x}) - \theta$ as the control variate.

A second form of control variate has a vector $g(\boldsymbol{x})$ where $\int g(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta$ is known. We will work mainly with the first form. When the density p is normalized the first form includes the second via $h(\boldsymbol{x}) = g(\boldsymbol{x})p(\boldsymbol{x})$.

Combining control variates with importance sampling from q leads to the estimate

$$\hat{\mu}_{q,\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i) - \beta^{\mathsf{T}} h(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} + \beta^{\mathsf{T}}\theta$$

for $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q$.

Theorem 9.3. *Let q be a probability density function with $q(\boldsymbol{x}) > 0$ whenever either $h(\boldsymbol{x}) \neq 0$ or $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. Then $\mathbb{E}_q(\hat{\mu}_{q,\beta}) = \mu$ for any $\beta \in \mathbb{R}^J$.*

Proof. The result follows from these two equalities: $\mathbb{E}_q(f(\boldsymbol{X})p(\boldsymbol{X})/q(\boldsymbol{X})) = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu$, and $\mathbb{E}_q(h_j(\boldsymbol{X})/q(\boldsymbol{X})) = \int h_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta_j$. □

The variance of $\hat{\mu}_{q,\beta}$ is $\sigma^2_{q,\beta}/n$ where

$$\sigma^2_{q,\beta} = \int \left(\frac{f(\boldsymbol{x})p(\boldsymbol{x}) - \beta^{\mathsf{T}} h(\boldsymbol{x})}{q(\boldsymbol{x})} - \mu + \beta^{\mathsf{T}}\theta \right)^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{9.21}$$

If we consider q fixed then we want the vector β^{opt} which minimizes (9.21). We can seldom get a closed form expression for β^{opt} but we can easily estimate it from the sample data.

The variance (9.21) is a mean squared error of a linear model relating fp/q to h/q under sampling $\boldsymbol{X} \sim q$. Our sample was drawn from this distribution q so we may estimate β^{opt} by least squares regression on the sample. We let $Y_i = f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i)$, $Z_{ij} = h_j(\boldsymbol{X}_i)/q(\boldsymbol{X}_i) - \theta$ and minimize $\sum_{i=1}^n (Y_i - \mu - \beta^{\mathsf{T}}\boldsymbol{Z}_i)^2$ over $\beta \in \mathbb{R}^J$ and $\mu \in \mathbb{R}$ obtaining $\hat{\beta}$ and the corresponding $\hat{\mu}_{q,\hat{\beta}}$. As usual with control variates, $\hat{\mu}_{q,\hat{\beta}}$ has a nonzero but asymptotically small bias.

For a 99% confidence interval we find

$$\hat{\sigma}^2_{q,\hat{\beta}} = \frac{1}{n-J-1}\sum_{i=1}^n (Y_i - \hat{\mu}_{q,\hat{\beta}} - \hat{\beta}^{\mathsf{T}}\boldsymbol{Z}_i)^2$$

and then take $\hat{\mu}_{q,\hat{\beta}} \pm 2.58\,\hat{\sigma}_{q,\hat{\beta}}/\sqrt{n}$ for a 99% confidence interval.

Theorem 9.4. *Suppose that there is a unique vector $\beta^{\mathrm{opt}} \in \mathbb{R}^J$ which minimizes $\sigma^2_{q,\beta}$ of (9.21). Suppose further that*

$$\mathbb{E}_q(h_j(\boldsymbol{X})^4/p(\boldsymbol{X})^4) < \infty \quad \textit{and} \quad \mathbb{E}_q(h_j(\boldsymbol{X})^2 f(\boldsymbol{X})^2/p(\boldsymbol{X})^4) < \infty$$

for $j = 1, \dots, J$. Then as $n \to \infty$,

$$\hat{\beta} = \beta^{\mathrm{opt}} + O_p(n^{-1/2}) \quad \textit{and} \quad \hat{\mu}_{q,\hat{\beta}} = \hat{\mu}_{q,\beta^{\mathrm{opt}}} + O_p(n^{-1}).$$

Proof. This is Theorem 1 of Owen and Zhou (2000). □

The conclusion of Theorem 9.4 is that we get essentially the same asymptotic error from an estimated coefficient as we would with the optimal one. The uniqueness condition in Theorem 9.4 will fail if two of the predictors h_j/q are collinear. In that case we may drop one or more redundant predictors to attain uniqueness.

Probability density functions provide a useful class of control variates h_j. We know that $\int h_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$. In §9.11 we will see a strong advantage to using a distribution q that is a mixture of densities with those same densities used as control variates.

9.11 Mixture importance sampling

In ***mixture importance sampling***, we sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the mixture distribution $q_\alpha = \sum_{j=1}^J \alpha_j q_j$ where $\alpha_j \geqslant 0$, $\sum_{j=1}^J \alpha_j = 1$ and the q_j are distributions. The closely related multiple importance sampling method of §9.12 takes $n_j \geqslant 1$ observations $\boldsymbol{X}_{ij} \sim q_j$ all independently, and then combines them to estimate μ.

Mixtures are convenient to sample from. Mixtures of unimodal densities provide a flexible approximation to multimodal target densities p that may arise in some Bayesian contexts. Similarly, since the ideal q is proportional to fp and not p, we might seek a mixture describing peaks in fp. For instance when a physical structure described by $\boldsymbol{X}$ has several different ways to fail, represented by the event $\boldsymbol{X} \in A_1 \cup A_2 \cup \cdots \cup A_J$ we may choose a q_j that shifts $\boldsymbol{X}$ towards A_j. In computer graphics, $f(\boldsymbol{x})$ may be the total contribution of photons reaching a point in an image. If there are multiple light sources and reflective objects in the scene, then mixture sampling is suitable with components q_j representing different light paths reaching the image.

Mixtures also allow us to avoid an importance density q with tails that are too light. If one mixture component is the nominal distribution p then the tails of q cannot be much lighter than those of p. Theorem 9.6 below show that mixture components used as control variates allow us to make several guesses as to which importance sampling distribution is best and get results comparable to those of the unknown best distribution among our choices.

In mixture importance sampling, we estimate $\mu = \mathbb{E}(f(\boldsymbol{X}))$ by the usual importance sampling equation (9.3), which here reduces to

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^n \frac{f(\boldsymbol{X}_i) p(\boldsymbol{X}_i)}{\sum_{j=1}^J \alpha_j q_j(\boldsymbol{X}_i)}. \tag{9.22}$$

Notice that equation (9.22) does not take account of which component q_j actually delivered $\boldsymbol{X}_i$. If we had sampled $\boldsymbol{X}_i$ from the mixture distribution by inversion, then we would not even know which q_j that was, but we could still use (9.22). We do have to compute $q_j(\boldsymbol{X}_i)$ for all $1 \leqslant j \leqslant J$, regardless of which proposal density generated $\boldsymbol{X}_i$.

An important special case of mixture IS is defensive importance sampling. In ***defensive importance sampling***, we take a distribution q thought to be a good importance sampler, mix it with the nominal distribution p, and then use

$$q_\alpha(\boldsymbol{x}) = \alpha_1 p(\boldsymbol{x}) + \alpha_2 q(\boldsymbol{x})$$

as the importance distribution, where $\alpha_j \geqslant 0$ and $\alpha_1 + \alpha_2 = 1$. For $\alpha_1 > 0$ we find that

$$\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} = \frac{p(\boldsymbol{x})}{\alpha_1 p(\boldsymbol{x}) + \alpha_2 q(\boldsymbol{x})} \leqslant \frac{1}{\alpha_1}$$

so the tails of q_α are not extremely light. A consequence is that

$$\mathrm{Var}(\hat{\mu}_{q_\alpha}) = \frac{1}{n}\Big(\int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q_\alpha(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} - \mu^2 \Big)$$

$$\leqslant \frac{1}{n}\Big(\int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{\alpha_1 p(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} - \mu^2\Big)$$
$$= \frac{1}{n\alpha_1}\big(\sigma_p^2 + \alpha_2\mu^2\big), \tag{9.23}$$

where σ_p^2/n is the variance of $\hat{\mu}$ under sampling from the nominal distribution. If $\mathrm{Var}_p(f(\boldsymbol{X})) < \infty$ then we are assured that $\sigma^2_{q_\alpha} < \infty$ too. For spiky f it is not unusual to find that $\mu \ll \sigma_p$ and so the defensive mixture could inflate the variance by only slightly more than $1/\alpha_1$. The bound (9.23) is quite conservative. It arises by putting $q(\boldsymbol{x}) = 0$. Self-normalized importance sampling with the defensive mixture component achieves an even better bound.

Theorem 9.5. *Let $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q_\alpha = \alpha_1 p + \alpha_2 q$ for $0 < \alpha_1 < 1$, $\alpha_1 + \alpha_2 = 1$, and $i = 1, \ldots, n$. Let $\widetilde{\mu}_{q_\alpha}$ be the self-normalized importance sampling estimate* (9.6) *and $\sigma^2_{q_\alpha,\mathrm{sn}}$ be the asymptotic variance of $\widetilde{\mu}_{q_\alpha}$. Then $\sigma^2_{q_\alpha,\mathrm{sn}} \leqslant \sigma_p^2/\alpha_1$.*

Proof. Hesterberg (1995). □

Example 9.7. Suppose that $f(x) = \varphi((x - 0.7)/0.05)/0.05$ (a Gaussian probability density function) and that $p(x)$ is the $\mathbf{U}(0, 1)$ distribution. If $q(x)$ is the Beta$(70, 30)$ distribution, then $q(x)$ is a good approximation to $f(x)p(x)$ in the important region. However q has very light tails and pf/q is unbounded on $(0, 1)$. The defensive mixture $q_{0.3} = 0.3q + 0.7p$ has a bounded likelihood ratio function. See Figure 9.2.

It is also worth comparing defensive importance sampling to what we might have attained by importance sampling from q. When q but not p is nearly proportional to fp, then q_α will not be nearly proportional to fp, possibly undoing what would have been a great variance reduction. Once again, the self-normalized version does better.

The self-normalized version of a general mixture is never much worse than the best individual mixture component would have been had we known which one that was. For self-normalized sampling we require $q_\alpha(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$. For $\widetilde{\mu}_{q_j}$ to be consistent we need $q_j(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$. In that case the ratio of asymptotic variances is

$$\frac{\sigma^2_{q_\alpha,\mathrm{sn}}}{\sigma^2_{q_j,\mathrm{sn}}} = \frac{\int (p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}))^2 (f(\boldsymbol{x}) - \mu)^2 q_\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int (p(\boldsymbol{x})/q_j(\boldsymbol{x}))^2 (f(\boldsymbol{x}) - \mu)^2 q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}$$
$$\leqslant \frac{\alpha_j^{-2}\int (p(\boldsymbol{x})/q_j(\boldsymbol{x}))^2 (f(\boldsymbol{x}) - \mu)^2 q_\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\alpha_j^{-1}\int (p(\boldsymbol{x})/q_j(\boldsymbol{x}))^2 (f(\boldsymbol{x}) - \mu)^2 q_\alpha(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}$$
$$= \frac{1}{\alpha_j}. \tag{9.24}$$

For $q_j = p$ we recover Theorem 9.5. For $q_j = q$ we find that the mixture would be at least as good as self-normalized importance sampling with $n\alpha_j$ observations from q. That does not mean it is as good as ordinary importance sampling would have been from $q_j \neq p$.

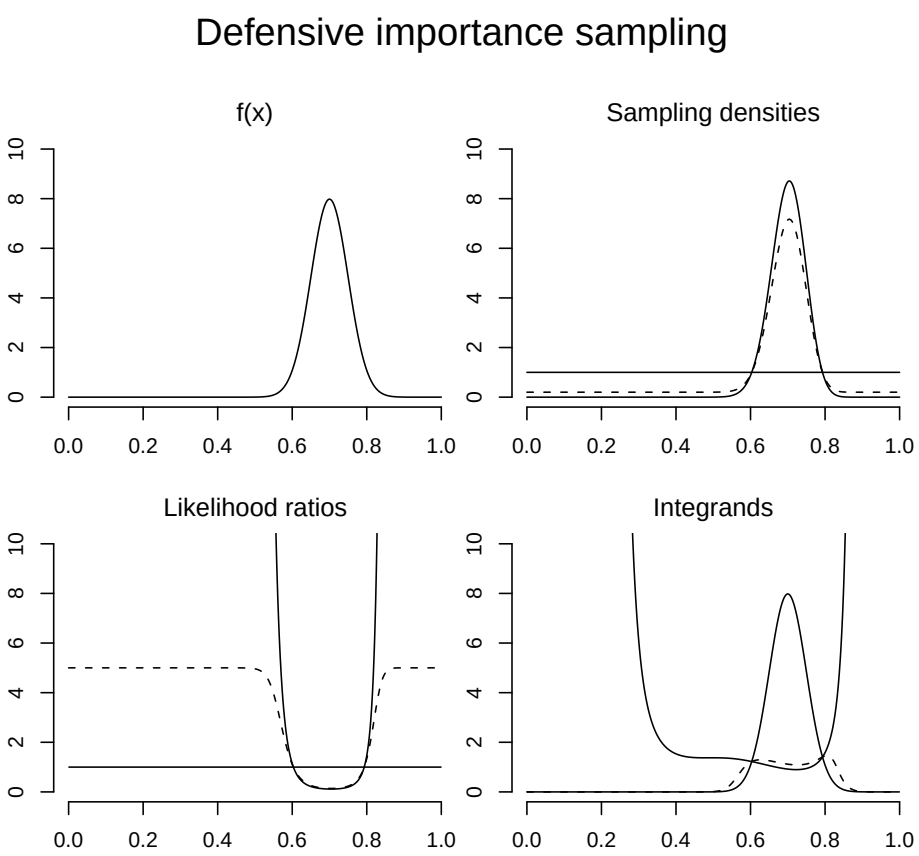

Figure 9.2: This figure shows defensive importance sampling from Example 9.7. The upper left panel shows f. The upper right panel shows the nominal density $p(x) = 1$ as well as the importance density $q(x) \propto x^{69}(1-x)^{29}$ as solid lines. The defensive mixture with $\alpha = (0.2, 0.8)$ is a dashed line. The bottom panels show likelihood ratios p/q, p/p, p/q_α and integrands fp/q, fp/p, fp/q_α, with the defensive case dashed. The defensive integrand is much flatter than the others.

Using the density function as a control variate provides at least as good a variance reduction as we get from self-normalized importance sampling or ordinary importance sampling or nominal sampling. This method also generalizes to mixtures with more than two components. We know that both $\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$ and $\int q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$ because they are densities. As long as these densities are normalized, we can use them as control variates as described in §9.10.

When we combine mixture sampling with control variates based on the com-

Algorithm 9.1 Mixture IS with component control variates by regression

given $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q_\alpha = \sum_{j=1}^J \alpha_j q_j$, $p(\boldsymbol{X}_i)$ and $f(\boldsymbol{X}_i)$, $\quad i = 1, \dots, n$
$Y_i \leftarrow f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)/q_\alpha(\boldsymbol{X}_i)$, $\quad i = 1, \dots, n$
$Z_{ij} \leftarrow q_j(\boldsymbol{X}_i)/q_\alpha(\boldsymbol{X}_i) - 1 \quad i = 1, \dots, n, \quad j = 1, \dots, J-1$ // dropped J'th
MLR $\leftarrow$ multiple linear regression of Y_i on Z_{ij}
$\hat{\mu}_{\text{reg}} \leftarrow$ estimated intercept from MLR
se $\leftarrow$ intercept standard error from MLR
deliver $\hat{\mu}$, se

This algorithm shows how to use linear regression software to implement mixture importance sampling with component control variates. One component has been dropped to cope with collinearity. For additional control variates h_m with $\int h_m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta_m$, include $Z_{i,J-1+m} = (h_m(\boldsymbol{X}_i) - \theta_m)/q_\alpha(\boldsymbol{X}_i)$ in the multiple regression.

ponent densities, the estimate of μ is

$$\hat{\mu}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^n \frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i) - \sum_{j=1}^J \beta_j q_j(\boldsymbol{X}_i)}{\sum_{j=1}^J \alpha_j q_j(\boldsymbol{X}_i)} + \sum_{j=1}^J \beta_j \tag{9.25}$$

where p is the nominal distribution. For defensive purposes we may take one of the q_j to be the same as p, or some other distribution known to yield a good estimator.

We may run into numerical difficulties using these control variates because $\sum_{j=1}^J q_j(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) = 1$ by definition of q_α. We can simply drop one of these redundant control variates. Algorithm 9.1 describes that process. It also shows how to include additional control variates $\int h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \theta$. If $\mathbb{E}_p(g(\boldsymbol{X})) = \theta$ then use $h(\boldsymbol{x}) = g(\boldsymbol{x})p(\boldsymbol{x})$ there and if $\mathbb{E}_q(g(\boldsymbol{X})) = \theta$ use $h(\boldsymbol{x}) = g(\boldsymbol{x})q(\boldsymbol{x})$.

We can compare the variance of mixture importance sampling to that of importance sampling with the individual mixture components q_j. Had we used q_j, the variance of $\hat{\mu}_{q_j}$ would be $\sigma^2_{q_j}/n$ where $\sigma^2_{q_j} = \mathrm{Var}_{q_j}(f(\boldsymbol{X})p(\boldsymbol{X})/q_j(\boldsymbol{X}))$.

Theorem 9.6. *Let $\hat{\mu}_{\alpha,\beta}$ be given by (9.25) for $\alpha_j > 0$, $\sum_{j=1}^J \alpha_j = 1$ and $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \sum_{j=1}^J \alpha_j q_j$. Let $\beta^{\text{opt}} \in \mathbb{R}^J$ be any minimizer over β of $\mathrm{Var}(\hat{\mu}_{\alpha,\beta})$. Then*

$$\mathrm{Var}(\hat{\mu}_{\alpha,\beta^{\text{opt}}}) \leqslant \min_{1\leqslant j\leqslant J} \frac{\sigma^2_{q_j}}{n\alpha_j}. \tag{9.26}$$

Proof. This is from Theorem 2 of Owen and Zhou (2000). □

We expect to get $n\alpha_j$ sample values from density q_j. The quantity $\sigma^2_{q_j}/(n\alpha_j)$ is the variance we would obtain from $n\alpha_j$ such values alone. We should not expect a better bound. Indeed, when $\sigma^2_{q_j} = \infty$ for all but one of the mixture components it is reassuring that those bad components don't make the estimate worse than what we would have had from the one good component.

If any one of the q_j is proportional to fp then Theorem 9.6 shows that we will get a zero variance estimator of μ.

It is common to have multiple functions f_m for $m = 1, \dots, M$ to average over $\boldsymbol{X}$. We may want to use the same $\boldsymbol{X}_i$ for all of those functions even though a good importance distribution for f_m might be poor for $f_{m'}$. Theorem 9.6 shows that we can include a mixture component q_j designed specifically for f_j without too much adverse impact on our estimate of $\mathbb{E}(f_m(\boldsymbol{X}))$ for $m \neq j$.

Theorem 9.6 does not compare $\hat{\mu}_{\alpha,\beta^{\mathrm{opt}}}$ to the best self-normalized importance sampler we could have used. When there is a single importance density then using it as a control variate has smaller asymptotic variance than using it in self-normalized importance sampling. Here we get a generalization of that result, as long as one of our component densities is the nominal one.

Theorem 9.7. *Let $\hat{\mu}_{\alpha,\beta}$ be given by (9.25) for $\alpha_j > 0$, $\sum_{j=1}^J \alpha_j = 1$ and $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \sum_{j=1}^J \alpha_j q_j$ for densities q_j one of which is p. Let $\beta^{\mathrm{opt}} \in \mathbb{R}^J$ be any minimizer over β of $\mathrm{Var}(\hat{\mu}_{\alpha,\beta})$. Then*

$$\mathrm{Var}(\hat{\mu}_{\alpha,\beta^{\mathrm{opt}}}) \leqslant \min_{1 \leqslant j \leqslant J} \frac{\sigma^2_{q_j,\mathrm{sn}}}{n\alpha_j}. \tag{9.27}$$

Proof. Without loss of generality, suppose that $q_1 = p$. Let $\beta = (\mu, 0, \dots, 0)^{\mathsf{T}}$. Then the optimal asymptotic variance is

$$\begin{aligned}
\sigma^2_{\alpha,\beta^{\mathrm{opt}}} \leqslant \sigma^2_{\alpha,\beta} &= \int \left(\frac{fp - \sum_{j=1}^J \beta_j q_j}{q_\alpha} - \mu + \sum_{j=1}^J \beta_j \right)^2 q_\alpha \, \mathrm{d}\boldsymbol{x} \\
&= \int \left(\frac{fp - \mu p}{q_\alpha} \right)^2 q_\alpha \, \mathrm{d}\boldsymbol{x} = \int \frac{p^2(f-\mu)^2}{q_\alpha} \, \mathrm{d}\boldsymbol{x} \\
&\leqslant \frac{\sigma^2_{q_j,\mathrm{sn}}}{\alpha_j}
\end{aligned}$$

for any $j = 1, \dots, J$. □

By inspecting the proof of Theorem 9.7 we see that it is not strictly necessary for one of the q_j to be p. It is enough for there to be constants $\eta_j \geqslant 0$ with $p(\boldsymbol{x}) = \sum_{j=1}^J \eta_j q_j(\boldsymbol{x})$. See Exercise 9.26.

9.12 Multiple importance sampling

In ***multiple importance sampling*** we take n_j observations from q_j for $j = 1, \dots, J$ arriving at a total of $n = \sum_{j=1}^J n_j$ observations. The sample values may be used as if they had been sampled from the mixture with fraction $\alpha_j = n_j/n$ coming from component q_j, and we will consider that case below, but there are many other possibilities. Multiple importance sampling also provides a zero variance importance sampler for use when $f(\boldsymbol{x})$ takes both positive and negative values. See §9.13 on positivisation.

A ***partition of unity*** is a collection of $J \geqslant 1$ weight functions $\omega_j(\boldsymbol{x}) \geqslant 0$ which satisfy $\sum_{j=1}^J \omega_j(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$. Suppose that $\boldsymbol{X}_{ij} \sim q_j$ for $i = 1, \dots, n_j$ and $j = 1, \dots, J$ and that ω_j are a partition of unity. The multiple importance sampling estimate is

$$\widehat{\mu}_\omega = \sum_{j=1}^J \frac{1}{n_j} \sum_{i=1}^{n_j} \omega_j(\boldsymbol{X}_{ij}) \frac{f(\boldsymbol{X}_{ij})p(\boldsymbol{X}_{ij})}{q_j(\boldsymbol{X}_{ij})}. \tag{9.28}$$

We saw an estimate of this form in §8.4 on stratified sampling. If we let $\omega_j(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in \mathcal{D}_j$ and 0 otherwise then (9.28) becomes independent importance sampling within strata. If we further take q_j to be p restricted to $\mathcal{D}_j$ then (9.28) becomes stratified sampling. Multiple importance sampling (9.28) may be thought of as a generalization of stratified sampling in which strata are allowed to overlap and sampling within strata need not be proportional to p.

Now assume that $q_j(\boldsymbol{x}) > 0$ whenever $\omega_j(\boldsymbol{x})p(\boldsymbol{x})f(\boldsymbol{x}) \neq 0$. Then multiple importance sampling is unbiased, because

$$\mathbb{E}(\widehat{\mu}_\omega) = \sum_{j=1}^J \mathbb{E}_{q_j}\left(\omega_j(\boldsymbol{X}) \frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q_j(\boldsymbol{X})}\right) = \sum_{j=1}^J \int \omega_j(\boldsymbol{x})f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu.$$

Among the proposals for functions $\omega_j(\boldsymbol{x})$, the most studied one is the ***balance heuristic*** with $\omega_j(\boldsymbol{x}) \propto n_j q_j(\boldsymbol{x})$, that is

$$\omega_j(\boldsymbol{x}) = \omega_j^{\mathrm{BH}}(\boldsymbol{x}) \equiv \frac{n_j q_j(\boldsymbol{x})}{\sum_{k=1}^J n_k q_k(\boldsymbol{x})}.$$

By construction $q_j(\boldsymbol{x}) > 0$ holds whenever $(\omega_j^{\mathrm{BH}} pf)(\boldsymbol{x}) \neq 0$. Let $n = \sum_{j=1}^J n_j$ and define $\alpha_j = n_j/n$. Then using the balance heuristic, $\widehat{\mu}_{\omega^{\mathrm{BH}}}$ simplifies to

$$\widehat{\mu}_\alpha = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{f(\boldsymbol{X}_{ij})p(\boldsymbol{X}_{ij})}{\sum_{j=1}^J \alpha_j q_j(\boldsymbol{X}_{ij})}. \tag{9.29}$$

In other words, multiple importance sampling, with weights from the balance heuristic reduces to the same estimator we would use in mixture importance sampling with mixture weights $\alpha_j = n_j/n$. Once again, the weight on a given sampled value $\boldsymbol{X}_{ij}$ does not depend on which mixture component it came from. The balance heuristic is nearly optimal in the following sense:

Theorem 9.8. *Let $n_j \geqslant 1$ be positive integers for $j = 1, \dots, J$. Let $\omega_1, \dots, \omega_J$ be a partition of unity and let ω^{BH} be the balance heuristic. Suppose that $q_j(\boldsymbol{x}) > 0$ whenever $\omega_j(\boldsymbol{x})p(\boldsymbol{x})f(\boldsymbol{x}) \neq 0$. Then*

$$\mathrm{Var}(\widehat{\mu}_{\omega^{\mathrm{BH}}}) \leqslant \mathrm{Var}(\widehat{\mu}_\omega) + \left(\frac{1}{\min_j n_j} - \frac{1}{\sum_j n_j}\right)\mu^2.$$

Proof. This is Theorem 1 of Veach and Guibas (1995). □

Other heuristics have been used in computer graphics. Among these are the ***power heuristic*** $w_j(\boldsymbol{x}) \propto (n_j q_j(\boldsymbol{x}))^\beta$ for $\beta \geqslant 1$ and the ***cutoff heuristic*** $w_j(\boldsymbol{x}) \propto \mathbb{1}\{n_j q_j(\boldsymbol{x}) \geqslant \alpha \max_\ell n_\ell q_\ell(\boldsymbol{x})\}$ for $\alpha > 0$. In each case the w_j are normalized to sum to 1 to get a partition of unity. The intuition is to put extra weight on the component with the largest $n_j q_j$ values that are more locally important (or locally proportional to fp). The ***maximum heuristic*** is the cutoff heuristic with $\alpha = 1$ or the power heuristic as $\beta \to \infty$. It puts all weight on the component with the largest value of $n_j q_j(\boldsymbol{x})$ or shares that weight equally when multiple components are tied for the maximum.

These heuristics diminish the problem that when fp is nearly proportional to one of q_j it might not be nearly proportional to $\sum_j n_j q_j$. There are versions of Theorem 9.8 for these three heuristics in Veach (1997, Theorem 9.3). The upper bounds are less favorable for them than the one for the balance heuristic. The upper bound in Theorem 9.8 contains a term like μ^2/n. The same holds for defensive importance sampling (9.23) if we don't use self-normalization. Using the densities as control variates removes the μ^2/n term.

We can also incorporate control variates into multiple importance sampling by the balance heuristic. Choosing the control variates to be mixture components is a derandomization of multiple IS with mixture controls. Because this derandomization reduces variance, Theorem 9.6 applies so that multiple importance sampling with the balance heuristic and β^{opt} has variance at most $\min_{1\leqslant j\leqslant N} \sigma^2_{q_j}/n_j$. Furthermore, Theorem 9.4 on consistency of $\hat\beta$ for β^{opt} also applies to multiple importance sampling with the balance heuristic. As a consequence, a very effective way to use multiple q_j is to sample a deterministic number n_j of observations from q_j and then run Algorithm 9.1 as if the $\boldsymbol{X}_i$ had been sampled IID from q_α.

9.13 Positivisation

It is a nuisance that a zero variance importance sampler is not available when f takes both positive and negative signs. There is a simple remedy based on multiple importance sampling.

We use a standard decomposition of f into positive and negative parts. Define

$$\begin{aligned} f_+(\boldsymbol{x}) &= \max(f(\boldsymbol{x}), 0), \quad \text{and} \\ f_-(\boldsymbol{x}) &= \max(-f(\boldsymbol{x}), 0). \end{aligned}$$

Then $f(\boldsymbol{x}) = f_+(\boldsymbol{x}) - f_-(\boldsymbol{x})$.

Now let q_+ be a density function which is positive whenever $pf_+ > 0$ and let q_- be a density function which is positive whenever $pf_- > 0$. We sample $\boldsymbol{X}_{i+} \sim q_+$ for $i = 1, \dots, n_+$ and $\boldsymbol{X}_{i-} \sim q_-$ for $i = 1, \dots, n_-$ (all independently) and define

$$\hat\mu^{\pm}_{\mathrm{mis}} = \frac{1}{n_+}\sum_{i=1}^{n_+} \frac{f_+(\boldsymbol{X}_{i+})p(\boldsymbol{X}_{i+})}{q_+(\boldsymbol{X}_{i+})} - \frac{1}{n_-}\sum_{i=1}^{n_-} \frac{f_-(\boldsymbol{X}_{i-})p(\boldsymbol{X}_{i-})}{q_-(\boldsymbol{X}_{i-})}.$$

While it is necessary to have $q_+(\boldsymbol{x}) > 0$ at any point where $f_+(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$, it is not an error if $q_+(\boldsymbol{x}) > 0$ at a point $\boldsymbol{x}$ with $f(\boldsymbol{x}) < 0$. All that happens is that a value $f_+(\boldsymbol{x}) = 0$ is averaged into the estimate $\hat{\mu}_+$. Exercise 9.27 verifies that $\hat{\mu}^{\pm}_{\text{mis}}$ really is multiple importance sampling as given by (9.28).

This two sample estimator is unbiased, because

$$\mathbb{E}(\hat{\mu}^{\pm}_{\text{mis}}) = \int f_+(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int f_-(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu,$$

using $f_+ - f_- = f$. Next

$$\mathrm{Var}(\hat{\mu}^{\pm}_{\text{mis}}) = \frac{1}{n_+}\int \frac{(pf_+ - \mu_+ q_+)^2}{q_+}\,\mathrm{d}\boldsymbol{x} + \frac{1}{n_-}\int \frac{(pf_- - \mu_- q_-)^2}{q_-}\,\mathrm{d}\boldsymbol{x}$$

where $\mu_\pm = \int f_\pm(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The variance can be estimated in a straightforward way by summing variances estimates for the two terms.

We get $\mathrm{Var}(\hat{\mu}^{\pm}_{\text{mis}}) = 0$ with $n = 2$ by taking $q_\pm \propto f_\pm p$ with $n_\pm = 1$. As with ordinary importance sampling, we don't expect to attain an estimate with variance 0, but we can use the idea to select good sample densities: q_+ should be roughly proportional to p times the positive part of f, erring on the side of having heavier tails instead of lighter ones, and similarly for q_-.

We can generalize the positivisation approach by writing

$$f(\boldsymbol{x}) = c + (f(\boldsymbol{x}) - c)_+ - (f(\boldsymbol{x}) - c)_-$$

for any $c \in \mathbb{R}$ that is convenient and estimating μ by

$$c + \frac{1}{n_+}\sum_{i=1}^{n_+} w_+(\boldsymbol{X}_{i+})(f(\boldsymbol{X}_{i+}) - c)_+ - \frac{1}{n_-}\sum_{i=1}^{n_-} w_-(\boldsymbol{X}_{i-})(f(\boldsymbol{X}_{i-}) - c)_-$$

for independent $\boldsymbol{X}_{i\pm} \sim q_\pm$ where $q_\pm(\boldsymbol{x}) \neq 0$ whenever $p(\boldsymbol{x})(f(\boldsymbol{x}) - c)_\pm \neq 0$ with $w_\pm(\boldsymbol{x}) = p(\boldsymbol{x})/q_\pm(\boldsymbol{x})$. A good choice for c would be one where $f(\boldsymbol{X}) = c$ has positive probability.

Still more generally, suppose that $g(\boldsymbol{x})$ is a function for which we know $\int g(\boldsymbol{X})p(\boldsymbol{X})\,\mathrm{d}\boldsymbol{x} = \theta$. Then we can estimate μ by

$$\theta + \frac{1}{n_+}\sum_{i=1}^{n_+} w_+(\boldsymbol{X}_{i+})(f(\boldsymbol{X}_{i+}) - g(\boldsymbol{X}_{i+}))_+ - \frac{1}{n_-}\sum_{i=1}^{n_-} w_-(\boldsymbol{X}_{i-})(f(\boldsymbol{X}_{i-}) - g(\boldsymbol{X}_{i-}))_-.$$

This approach can be valuable when $\mathbb{P}(f(\boldsymbol{X}) = g(\boldsymbol{X}))$ is close to one. For instance we might have a closed form expression for $\theta = \int g(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ but not for

$$f(\boldsymbol{x}) = \begin{cases} L, & g(\boldsymbol{x}) \leqslant L \\ U, & g(\boldsymbol{x}) \geqslant U \\ g(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

The problem context may give qualitative information about when $f > g$ or $f < g$ which we can use to choose $q_\pm$. It would be wise to incorporate a defensive mixture component from p into each of $q_\pm$.

A very simple form of positivisation is available when $f(\boldsymbol{x}) \geqslant B > -\infty$ for some $B < 0$. In that case we can replace f by $f + c$ for $c \geqslant -B$ to get a positive function. Then we subtract c from the importance sampled estimate of $\mathbb{E}_p(f(\boldsymbol{X}) + c)$. It may be hard to find a density q that is nearly proportional to $f + c$. This simple shift is used in some particle transport problems described in §10.5 where it helps bound some estimates away from 0.

9.14 What-if simulations

Although the primary motivation for importance sampling is to reduce variance for very skewed sampling problems, we can also use it to estimate $\mathbb{E}(f(\boldsymbol{X}))$ under multiple alternative distributions for $\boldsymbol{X}$. This is sometimes called ***what if*** simulation.

Suppose for instance that the probability density function p has a parametric form $p(\boldsymbol{x};\theta)$ for $\theta \in \Theta \subset \mathbb{R}^k$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim p(\cdot\,;\theta_0)$ where $\theta_0 \in \Theta$. Then for some other value $\theta \in \Theta$ we can estimate $\mu(\theta) = \int f(\boldsymbol{x})p(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x}$ by

$$\hat{\mu}_{\theta_0}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(\boldsymbol{x}_i;\theta)}{p(\boldsymbol{x}_i;\theta_0)} f(\boldsymbol{x}_i).$$

The estimation spares us from having to sample from lots of different distributions. Instead we simply reweight the simulated values from one of the simulations. The result is a bit like using common random numbers as in §8.6. Changing the likelihood ratio $p(\boldsymbol{x};\theta)/p(\boldsymbol{x};\theta_0)$ can be much faster than using common random numbers, if the latter has to recompute f for each value of θ.

The tails of $p(\cdot\,;\theta_0)$ should be at least as heavy as any of the $p(\cdot\,;\theta)$ that we're interested in. We don't have to sample values from one of the $p(\cdot\,;\theta)$ distributions. We can instead take $\boldsymbol{X}_i \sim q$ for a heavier tailed distribution q and then use the estimates

$$\hat{\mu}_q(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p(\boldsymbol{x}_i;\theta)}{q(\boldsymbol{x}_i)} f(\boldsymbol{x}_i).$$

Reweighting the sample is a good strategy for a local analysis in which θ is a small perturbation of θ_0. If $p(\cdot\,;\theta_0)$ is too different from $p(\cdot\,;\theta)$ then we can expect a bad result. For a concrete example, suppose that we sample from the $\mathcal{N}(\theta_0, I)$ distribution and plan to study the $\mathcal{N}(\theta, I)$ distribution by reweighting the samples. Using Example 9.2 we find that the population effective sample size of (9.14) is $n^*_{\mathrm{e}} = n\exp(-\|\theta-\theta_0\|^2)$. If we require $n^*_{\mathrm{e}} \geqslant 0.01n$, then we need $\|\theta-\theta_0\| \leqslant \log(10) \doteq 2.30$. More generally, if Σ is nonsingular, p is $\mathcal{N}(\theta_0, \Sigma)$ and q is $\mathcal{N}(\theta, \Sigma)$, then $n^*_{\mathrm{e}} \geqslant n/100$ holds for $\big((\theta-\theta_0)^{\mathsf{T}}\Sigma^{-1}(\theta-\theta_0)\big)^{1/2} \leqslant \log(10)$.

Example 9.8 (Expected maximum of d Poissons)**.** For $\boldsymbol{x} \in \mathbb{R}^d$, let $\max(\boldsymbol{x}) = \max_{1\leqslant j\leqslant d} x_j$. The $\boldsymbol{X}$ we consider has components $X_j \overset{\text{iid}}{\sim} \mathrm{Poi}(\nu)$ and we will

estimate $\mu(\nu) = \mathbb{E}(\max(\boldsymbol{X}))$. If we sample $X_{ij} \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ for $i = 1, \dots, n$ and $j = 1, \dots, d$ then we can estimate $\mu(\nu)$ by

$$\begin{aligned}\hat{\mu}(\nu) &= \frac{1}{n}\sum_{i=1}^{n} \max(\boldsymbol{X}_i) \times \prod_{j=1}^{d} \frac{e^{-\nu}\nu^{X_{ij}}/X_{ij}!}{e^{-\lambda}\lambda^{X_{ij}}/X_{ij}!} \\ &= \frac{1}{n}\sum_{i=1}^{n} \max(\boldsymbol{X}_i) e^{d(\lambda-\nu)} \Big(\frac{\nu}{\lambda}\Big)^{\sum_{j=1}^{d} X_{ij}}.\end{aligned}$$

In this case we can find n^*_{e}, the effective sample size (9.14) in the population. For $d = 1$ it is $n/e^{-2\lambda+\nu+\lambda^2/\nu}$ (Exercise 9.31) and for d independent components it is $n/(e^{-2\lambda+\nu+\lambda^2/\nu})^d$. Figure 9.3 illustrates this computation for $\lambda = 1$ and $d = 5$ using $n = 10{,}000$.

For ν near λ, the estimated effective sample sizes are close to $n^*_{\text{e}}(\nu)$. For ν much greater than λ we get very small values for $n^*_{\text{e}}(\nu)$ and unreliable sample values of n_{e}. The estimates $\hat{\mu}(\nu)$ are monotone increasing in ν. The 99% confidence intervals become very wide for $\nu > \lambda = 1$ even when the effective sample size is not small. The root of the problem is that the sampling density $\text{Poi}(\lambda)$ has lighter tails than $\text{Poi}(\nu)$. For large enough ν a sample from $\text{Poi}(\nu)$ will have a maximum value smaller than $\mathbb{E}(\max(\boldsymbol{X}); \nu)$ and no reweighting will fix that. In this example a sample from $\text{Poi}(\lambda)$ was effective for all of the ν below λ and for ν only modestly larger than λ.

We can even use reweighting to get an idea whether another importance sampling distribution might have been better than the one we used. Suppose that we want $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Now we have a parametric family of importance sampling candidates, $q(\boldsymbol{x}, \theta)$, for $\theta \in \Theta$ and p is not necessarily a member of this family. The variance of importance sampling for μ via $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q(\boldsymbol{x}; \theta)$ is

$$\sigma^2_\theta = \int \frac{(p(\boldsymbol{x})f(\boldsymbol{x}))^2}{q(\boldsymbol{x};\theta)}\,\mathrm{d}\boldsymbol{x} - \mu^2.$$

The second term does not depend on θ and so we prefer θ which makes the first term small. Rewriting the first term as

$$\text{MS}_\theta = \int \frac{(p(\boldsymbol{x})f(\boldsymbol{x}))^2}{q(\boldsymbol{x};\theta)q(\boldsymbol{x};\theta_0)} q(\boldsymbol{x};\theta_0)\,\mathrm{d}\boldsymbol{x},$$

we see that an unbiased estimate of it is

$$\widehat{\text{MS}}_{\theta(\theta_0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{(p(\boldsymbol{X}_i)f(\boldsymbol{X}_i))^2}{q(\boldsymbol{X}_i;\theta)q(\boldsymbol{X}_i;\theta_0)},$$

for $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q(\cdot, \theta_0)$. We might then minimize $\widehat{\text{MS}}_{\theta(\theta_0)}$ numerically over θ_0 to find a new value θ_1 to sample from. There are numerous adaptive importance sampling algorithms that make use of sample values from one distribution to choose a better importance sampler through some number of iterations. See §10.5.

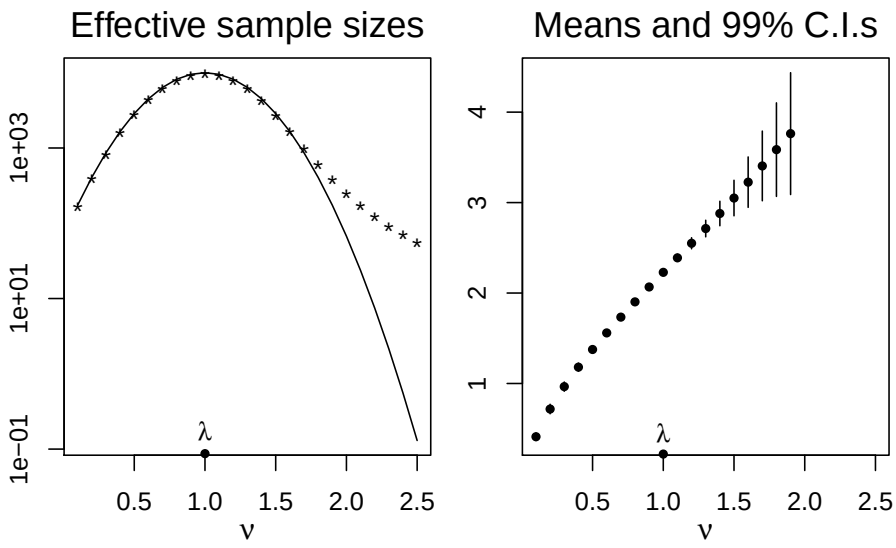

Figure 9.3: These figures describe the what-if simulation of Example 9.8. The left panel shows the population effective sample n_{e}^* size versus ν as a curve and the sample estimates as points. For ν with population $n_{\mathrm{e}}^* \geqslant n/100$ the right panel shows the estimates of $\mathbb{E}(\max(\boldsymbol{X}))$ for $X_1, \ldots, X_5 \overset{\text{iid}}{\sim} \mathrm{Poi}(\nu)$. The estimates were based on reweighting values from $\mathrm{Poi}(\lambda)$ where $\lambda = 1$.

Chapter end notes

Two very early descriptions of importance sampling are Kahn (1950a,b). The famous result that the optimal importance density $q(\boldsymbol{x})$ is proportional to $p(\boldsymbol{x})|f(\boldsymbol{x})|$ appears in Kahn and Marshall (1953). They describe a variational approach and also present the optimal importance sampler for the marginal distribution of X_1 when $\boldsymbol{X} = (X_1, X_2)$ and X_2 is sampled from its nominal conditional distribution given X_1.

Trotter and Tukey (1956) consider whether to use ordinary or self-normalized importance sampling. They lean towards the former and then suggest a hybrid of the methods. Kong (1992) connects effective sample size, via a delta method, to an approximate relative efficiency for self-normalized importance sampling. Hesterberg (1995) has a good discussion on the comparative advantages of ordinary versus self-normalized importance sampling, when both can be implemented. He finds that the ordinary method is superior for rare events but inferior in cases where fluctations in the likelihood ratio dominate.

Trotter and Tukey (1956) pointed out how observations from one distribution can be reweighted to estimate quantities from many other distributions, providing the basis for what-if simulations. Arsham et al. (1989) use that idea for a sensitivity analysis of some discrete event systems.

Using $\mathbb{E}_q(p(\boldsymbol{X})/q(\boldsymbol{X})) = 1$ to construct a hybrid of control variates and the regression estimator appears in Hesterberg (1987, 1995) and also in Arsham et al. (1989).

Hessian approach

Here we describe some refinements of the Hessian approach. If $\pi(\beta \mid \mathcal{Y})$ is skewed we can get a better fit by using a distribution that is not symmetric around β_*. One approach is to use a different variance to the left and right of β_*. In d dimensions there are potentially $2d$ different variances to select. The ***split-t*** distribution of Geweke (1989) is constructed as follows. First set

$$Z \sim \mathcal{N}(0, I_d), \quad Q \sim \chi^2_{(\nu)}, \quad \text{and}$$
$$V_j = \begin{cases} \sigma_{Lj} Z_j, & Z_j \leqslant 0 \\ \sigma_{Rj} Z_j, & Z_j > 0, \end{cases} \quad j = 1, \ldots, d,$$

with Z independent of Q. Then the split-t random variable is $\mu + C^{1/2}V/\sqrt{Q}$. The parameters $\sigma_{Lj} > 0$ and $\sigma_{Rj} > 0$ allow different variances to the left and right of the mode μ of V. Geweke (1989) gives the form of the density for split-t variables and some advice on how to select parameters. There is also a split-normal distribution.

Mixtures

Mixtures have long been used for importance sampling. In a Bayesian context, if the posterior distribution $\pi_u(\cdot)$ is multimodal then we might attempt to find all of the modes and approximate π_u by a mixture distribution with one component centered on each of those modes. Oh and Berger (1993) use a mixture of t distributions.

Defensive mixtures are from the dissertation of Hesterberg (1988), and are also in the article Hesterberg (1995).

Multiple importance sampling was proposed by Veach and Guibas (1995) for use with bidirectional path sampling in graphical rendering. Bidirectional path sampling was proposed independently by Lafortune and Willems (1993) and Veach and Guibas (1994). The combination of multiple control variates with importance sampling in §9.10 and §9.11 is taken from Owen and Zhou (2000).

The positivisation approach from §9.13 is based on Owen and Zhou (2000). They also considered a 'partition of identity' trick to smooth out the cusps from $(f(\boldsymbol{x}) - g(\boldsymbol{x}))_\pm$.

Calculus of variations

In §9.1 we proved that $q^*(\boldsymbol{x}) \propto |f(\boldsymbol{x})|p(\boldsymbol{x})$ is the optimal importance sampling density. The proof is unattractive in that we had to know the answer first.

We want to minimize $\int f(\boldsymbol{x})^2 p(\boldsymbol{x})^2/q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ over q subject to the constraint $\int q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$. This is a calculus problem, except that the unknown q is a function, not a vector. The solution may be found using the calculus of variations. A thorough description of the calculus of variations is given by Gelfand and Fomin (2000).

To minimize $\int f(\boldsymbol{x})^2 p(\boldsymbol{x})^2/q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ over q subject to $\int q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$, we form the Lagrangian

$$G(q) = \int f(\boldsymbol{x})^2 p(\boldsymbol{x})^2/q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \lambda\left(\int q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 1\right)$$

and working formally, set $\partial G/\partial q(\boldsymbol{x}) = 0$ along with $\partial G/\partial\lambda = 0$.

Setting the partial derivative with respect to $q(\boldsymbol{x})$ to zero, yields

$$-\frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2}{q(\boldsymbol{x})^2} + \lambda = 0$$

which is satisfied by the density $q^*(\boldsymbol{x}) = \sqrt{f(\boldsymbol{x})^2 p(\boldsymbol{x})^2/\lambda}$, that is, $q^* \propto |f(\boldsymbol{x})|p(\boldsymbol{x})$.

Just as in multivariable calculus, setting the first derivative to zero can give a minimum, maximum or saddle point, and it may not take proper account of constraints. To show that q^* is a local minimum we might turn to second order conditions (Gelfand and Fomin, 2000, Chapters 5 and 6). In many importance sampling problems we can make a direct proof that the candidate function is a global minimum, typically via the Cauchy-Schwartz inequality.

Effective sample size

The linear combination $S_{\boldsymbol{w}}$ of (9.12) used to motivate the formula for n_{e} raises some awkward conceptual issues. It is hard to consider the random variables $Z_i = f(\boldsymbol{X}_i)$ in it as independent given $w_i = p(\boldsymbol{X}_i)/q(\boldsymbol{X}_i)$. Both Z_i and w_i are deterministic functions of the same random $\boldsymbol{X}_i$. The derivation of equation (9.15) was based on a delta method argument (Kong, 1992). After some simplification the variance of the normalized importance sampler $\widetilde{\mu}_q$ can be close to $1 + \mathrm{cv}(\boldsymbol{w})^2$ times the variance of $\hat{\mu}_p$, the estimate we would use sampling directly from the nominal distribution p. Then instead of getting variance σ^2/n we get variance σ^2/n_{e}. By this approximation, self-normalized importance sampling cannot be better than direct sampling from the nominal distribution.

Self-normalized importance sampling can in fact be better than direct sampling. The difficulty arises from one step in the approximation. Kong (1992) points out the exact term whose omission gives rise to the difference. The effective sample size is a convenient way to interpret the amount of inequality in the weights but it does not translate cleanly into a comparison of variances.

The f-specific effective sample sizes are based on an idea from Evans and Swartz (1995). The statistic that they use is $\mathrm{cv}(\boldsymbol{w}, f)/\sqrt{n}$, which may be thought of as a coefficient of variation for $\hat{\mu}_q$.

The difficult choice with diagnostics is that we would like one summary effective sample size for all f, but the actual variance reduction in importance

sampling depends on f. For an example with good performance even though n_{e} is not large see Exercise 9.11.

Exercises

9.1. Suppose that p is the $\mathbf{U}(0,1)$ distribution, q is the $\mathbf{U}(0,1/2)$ distribution, and that $f(x) = x^2$. Show that $\mathbb{E}_q(f(X)p(X)/q(X))$ is not equal to $\mathbb{E}_p(f(X))$. Note: when $X \sim q$, we will never see $X > 1/2$.

9.2. Suppose that $q(x) = \exp(-x)$ for $x > 0$ and that $f(x) = |x|$ for all $x \in \mathbb{R}$, so that $q(x) = 0$ at some x where $f(x) \neq 0$.

a) Give a density $p(x)$ for which the expectation of $f(X)p(X)/q(X)$ for $X \sim q$ matches the expectation of $f(X)$ for $X \sim p$.

b) Give a density $p(x)$ for which the expectation of $f(X)p(X)/q(X)$ for $X \sim q$ does not match the expectation of $f(X)$ for $X \sim p$.

9.3. Suppose that p is the $\mathcal{N}(0,1)$ distribution, and that $f(x) = \exp(-(x-10)^2/2)$. Find the optimal importance sampling density q.

9.4. Suppose that p is the $\mathcal{N}(0,1)$ distribution, and that f is $\exp(kx)$ for $k \neq 0$. Find the optimal importance sampling density q.

9.5. Let p be the $\mathcal{N}(0, I)$ distribution in dimension $d \geqslant 1$.

a) Generalize Exercise 9.3 to the case where f is the density of $\mathcal{N}(\theta, I)$.

b) Generalize Exercise 9.4 to the case where $f(\boldsymbol{x}) = \exp(\boldsymbol{k}^{\mathsf{T}}\boldsymbol{x})$ for $\boldsymbol{k} \in \mathbb{R}^d$.

9.6. In the importance sampling notation of §9.1, suppose that $\sigma_p^2 < \infty$ and that $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x}) \leqslant c$ holds for all $\boldsymbol{x}$. Prove that $\sigma_q^2 \leqslant c\sigma_p^2 + (c-1)\mu^2$.

9.7. Here we revisit Example 9.1. If the nominal distribution p has $X \sim \mathcal{N}(0,1)$, and one uses an importance sampling distribution q which is $\mathcal{N}(0,\sigma^2)$, then what choice of σ^2 will minimize the variance of $(1/n)\sum_{i=1}^n X_i p(X_i)/q(X_i)$? By how much does the optimum σ^2 reduce the variance compared to using $\sigma^2 = 1$?

9.8. For self-normalized importance sampling suppose that $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x}) \leqslant c < \infty$ for all $\boldsymbol{x}$ with $p(\boldsymbol{x}) > 0$. Prove that the asymptotic variance satisfies $\sigma^2_{q,\mathrm{sn}} \leqslant c\sigma_p^2$.

9.9. Let $p(x) = \lambda e^{-\lambda x}$, for $x \geqslant 0$ and $q(x) = \eta e^{-\eta x}$ also for $x \geqslant 0$. The constant $\lambda > 0$ is given. For which values of η is the likelihood ratio $w(x)$ bounded?

9.10 (Optimal shift in importance sampling)**.** The variance of an ordinary importance sampling estimate changes if we add a constant to f. Here we find the optimal constant to add. For $c \in \mathbb{R}$ let $f_c(\boldsymbol{x}) = f_0(\boldsymbol{x}) + c$. Let the nominal density be $p(\boldsymbol{x})$ and the importance density be some other density $q(\boldsymbol{x})$. For this problem we assume that $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$. Let

$$\hat{\mu}_{q,c} = \frac{1}{n}\sum_{i=1}^n \frac{f_c(\boldsymbol{X}_i)p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} - c$$

for $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q$ be our estimate of $\mu = \int f_0(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Show that $\mathrm{Var}(\hat{\mu}_{q,c})$ is minimized by

$$c = -\frac{\int (w(\boldsymbol{x})f_0(\boldsymbol{x}) - \mu)(w(\boldsymbol{x}) - 1)\,\mathrm{d}\boldsymbol{x}}{\int (w(\boldsymbol{x}) - 1)^2\,\mathrm{d}\boldsymbol{x}}.$$

You may assume that $\mathrm{Var}(\hat{\mu}_{q,c}) < \infty$ for all $c \in \mathbb{R}$.

9.11. A bad effective sample size sometimes happens when importance sampling is actually working well. This exercise is based on an example due to Tim Hesterberg. Let p be $\mathcal{N}(0,1)$, $f(x) = \mathbb{1}_{x>\tau}$ for $\tau > 0$ and q be $\mathcal{N}(\tau, 1)$.

a) Show that $n^*_{\mathrm{e}} = \exp(-\tau^2)n$.

b) Show that $\dfrac{\sigma_p^2}{\sigma_q^2} = \dfrac{\Phi(\tau)\Phi(-\tau)}{e^{\tau^2}\Phi(-2\tau) - \Phi(-\tau)^2}$.

c) Evaluate these quantities for $\tau = 3$ and $\tau = 10$.

d) $\tau = 10$ is a pretty extreme case. What value do we get for σ_p/μ and σ_q/μ? As usual, $\mu = \mathbb{E}_p(f(X))$.

e) Simulate the cases $\tau = 3$ and $\tau = 10$ for $n = 10^6$. Report the Evans and Swartz effective sample sizes (9.17) that you get.

9.12. In the content uniformity problem of §8.6 the method first samples 10 items from $p = \mathcal{N}(\mu, \sigma^2)$. The test might pass based on those 10 items, or it may be necessary to sample 20 more items to make the decision. Suppose that we importance sample with a distribution q and use the ordinary importance sampling estimate $\hat{\mu}_q$ taken as an average of weighted acceptance outcomes, $(1/n)\sum_{i=1}^n w(\boldsymbol{X}_i)A(\boldsymbol{X}_i)$ for a weight function w and $A(\boldsymbol{x})$ which is 1 for accepted lots and 0 otherwise.

a) Will our estimate of the acceptance probability be unbiased if we always use $w(\boldsymbol{X}_i) = \prod_{j=1}^{30} p(X_{ij})/q(X_{ij})$, even in cases where the lot was accepted based on the first 10 items?

b) Will our estimate of the acceptance probability be unbiased if we use $w(\boldsymbol{X}_i) = \prod_{j=1}^{10} p(X_{ij})/q(X_{ij})$ when the lot is accepted after the first 10 items and use $w(\boldsymbol{X}_i) = \prod_{j=1}^{30} p(X_{ij})/q(X_{ij})$ otherwise?

9.13. Here we revisit the PERT problem of §9.4. This time we assume that the duration for task j is χ^2 with degrees of freedom equal to the value in the last column of Table 9.1. For example, task 1 (planning) takes $\chi^2_{(4)}$ and task 10 (final testing) takes $\chi^2_{(2)}$ time, and so on. Use importance sampling to estimate the probability that the completion time is larger than 50. Document the importance sampling distribution you used and how you arrived at it. It is reasonable to explore a few distributions with small samples before running a larger simulation to get the answer. Use hand tuning, and not automated adaptive importance samplers.

9.14. Let $A \subset \mathbb{R}^d$ and let $A^c = \{\boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{x} \notin A\}$. Let p be a probability density function on $\mathbb{R}^d$ with $0 < \int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < 1$. Let q_1 be the asymptotically optimal sampling density for estimating $\mu = \mathbb{E}_p(\mathbb{1}\{\boldsymbol{x} \in A\})$, using self-normalized importance sampling. Let q_2 be the corresponding optimal density for $\mathbb{E}_p(\mathbb{1}\{\boldsymbol{x} \notin A\})$. Show that q_1 and q_2 are the same distribution.

9.15. Equation (9.8) gives the delta method approximation to the variance of the ratio estimation version of the importance sampler. The optimal sampling distribution q for this estimate is proportional to $p(\boldsymbol{x})|f(\boldsymbol{x}) - \mu|$.

a) Prove that equation (9.10) is an equality if q is the optimal self-normalized importance sampling density.

b) Now suppose that $f(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in A$ and 0 otherwise, so $\mu = \mathbb{P}(\boldsymbol{X} \in A)$ for $\boldsymbol{X} \sim p$. What form does the variance from the previous part take? Compare it to $\mu(1-\mu)/n$ paying special attention to extreme cases for μ.

9.16 (Rare events and self-normalized importance sampling)**.** For each $\epsilon > 0$ let A_ϵ be an event with $\mathbb{P}_p(\boldsymbol{X} \in A_\epsilon)$. Sampling from the nominal distribution p leads to the estimate $\hat{\mu}_{p,\epsilon} = (1/n)\sum_{i=1}^n \mathbb{1}_{\boldsymbol{X}_i \in A_\epsilon} \sim \mathrm{Bin}(n, \epsilon)$. As a result,

$$\frac{\sqrt{\mathrm{Var}(\hat{\mu}_{p,\epsilon})}}{\epsilon} = \sqrt{\frac{1-\epsilon}{n\epsilon}} \to \infty, \quad \text{as } \epsilon \to 0. \tag{9.30}$$

Now suppose that for each $\epsilon > 0$, we are able to find and sample from the optimal self-normalized importance sampling density. Call it q_ϵ and the resulting estimator $\widetilde{\mu}_{q,\epsilon}$.

a) Find $\lim_{\epsilon \to 0^+} \sqrt{\widetilde{\mathrm{Var}}(\widetilde{\mu}_{q,\epsilon})}/\epsilon$.

b) If the limit in part **a** is finite then the method has ***bounded relative error***. Does the optimal self-normalized importance sampling estimator have bounded relative error?

c) Does the optimal ordinary importance sampling estimator have bounded relative error?

Equation (9.30) shows that sampling from the nominal distribution does ***not*** have bounded relative error.

9.17 (Effective sample sizes)**.** For $i = 1, \ldots, n$, let $w_i \geqslant 0$ with $\sum_{i=1}^n w_i$ strictly positive and $n > 1$.

a) Show that $n_{\mathrm{e},\sigma} \leqslant n_{\mathrm{e}}$ where $n_{\mathrm{e},\sigma}$ is given by (9.16).

b) Give an example to show that $n_{\mathrm{e},\sigma} = n_{\mathrm{e}}$ is possible here. [There are at least two quite different ways to make this happen.]

9.18. For $i = 1, \ldots, n$, let $w_i \geqslant 0$ with $\sum_{i=1}^n w_i$ strictly positive and $n \geqslant 1$. For $k \geqslant 1$ let $n_{\mathrm{e},k} = (\sum_i w_i^k)^2 / \sum_i w_i^{2k}$. What is $\lim_{k\to\infty} n_{\mathrm{e},k}$?

9.19. The skewness of a random variable X is $\gamma = \mathbb{E}((X-\mu)^3)/\sigma^3$ where μ and σ^2 are the mean and variance of X. Suppose that Z_i has (finite) mean, variance, and skewness μ, σ^2 and γ respectively. Let $w_1, \dots, w_n \geqslant 0$ be nonrandom with $\sum_{i=1}^n w_i > 0$. Show that the skewness of $Y = \sum_{i=1}^n w_i Z_i / \sum_{i=1}^n w_i$ is

$$\left(\sum_{i=1}^n w_i^3\right) \Big/ \left(\sum_{i=1}^n w_i^2\right)^{3/2}.$$

9.20. Sometimes we can compute a population version of effective sample size, without doing any sampling. Let $p(\boldsymbol{x}) = \prod_{j=1}^d \exp(-x_j/\theta_j)/\theta_j$ for $\boldsymbol{x} \in (0,1)^d$ and $\theta_j > 0$. Let $q(\boldsymbol{x}) = \prod_{j=1}^d \exp(-x_j/\lambda_j)/\lambda_j$ where $\lambda_j = \kappa_j\theta_j$ for $\kappa_j > 0$. Show that n_{e}^* given by (9.14) satisfies

$$n_{\mathrm{e}}^* = n \times \prod_{j=1}^d \frac{2\kappa_j - 1}{\kappa_j^2}$$

if $\min_j \kappa_j > 1/2$ and $n_{\mathrm{e}}^* = 0$ otherwise.

9.21. Derive population versions of $n_{\mathrm{e},\sigma}$ and $n_{\mathrm{e},\gamma}$ for the setting of Exercise 9.20.

9.22. For the PERT example of §9.4, use importance sampling to estimate $\mathbb{P}(E_{10} < 3)$. In view of Exercise 9.20 we should be cautious about importance sampling with small multiples of exponential means. Defensive importance sampling may help. Document the steps you took in finding an importance sampling strategy.

9.23. Let $p = \sum_{j=0}^1 \sum_{k=0}^1 \alpha_{jk} p_{jk}$ be a distribution on $\mathbb{R}^2$ with

$$\begin{aligned} p_{00} &= \delta_0 \times \delta_0 \\ p_{01} &= \delta_0 \times \mathcal{N}(0,1) \\ p_{10} &= \mathcal{N}(0,1) \times \delta_0, \quad \text{and} \\ p_{11} &= \mathcal{N}(0, I_2) \end{aligned}$$

where δ_0 is the degenerate distribution on $\mathbb{R}$ taking the value 0 with probability one. The coefficients satisfy $\alpha_{jk} > 0$ and $\sum_j \sum_k \alpha_{jk} = 1$. The notation $\boldsymbol{X} \sim f \times g$ means that $\boldsymbol{X} = (X_1, X_2)$ with $X_1 \sim f$ independently of $X_2 \sim g$.

Now let $q = \sum_{j=0}^1 \sum_{k=0}^1 \beta_{jk} q_{jk}$ where $q_{jk} = q_j \times q_k$ with $q_0 = \delta_0$ and q_1 a continuous distribution on $\mathbb{R}$ with a strictly positive density, $\beta_{jk} > 0$ and $\sum_j \sum_k \beta_{jk} = 1$.

Let $\boldsymbol{X}_i \sim q$ independently for $i = 1, \dots, n$ and suppose that $f(\boldsymbol{x})$ is a function on $\mathbb{R}^2$ with $\mathbb{E}_p(|f(\boldsymbol{X})|) < \infty$. Develop an estimate $\hat{\mu}_q$ of $\mu = \mathbb{E}_p(f(\boldsymbol{X}))$ using the observed values $\boldsymbol{x}_i$ of $\boldsymbol{X}_i \sim q$ and prove that $\mathbb{E}_q(\hat{\mu}_q) = \mu$.

9.24. Here we revisit the boundary crossing problem of Example 9.5 but with different parameters. Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(-1.5, 1)$. Estimate the probability that the random walk given by $X_1 + X_2 + \cdots + X_m$ exceeds $b = 20$ before it goes below $a = -7$. Include a 99% confidence interval.

9.25. Suppose that we use ordinary importance sampling with the defensive mixture $q_\alpha(\boldsymbol{x}) = \alpha_1 p(\boldsymbol{x}) + \alpha_2(\boldsymbol{x})$ where $0 < \alpha_1 < 1$ and $\alpha_1 + \alpha_2 = 1$. Prove that

$$\mathrm{Var}(\hat{\mu}_{q_\alpha}) \leqslant \frac{\sigma_q^2}{n\alpha_2} + \frac{1}{n}\frac{\alpha_1}{\alpha_2}\mu^2$$

where $\mu = \mathbb{E}_p(f(\boldsymbol{X}))$ and σ_q^2 is the asymptotic variance when importance sampling from q.

9.26. Prove Theorem 9.7 comparing mixture importance sampling to self-normalized importance sampling, under the weaker condition that there are constants $\eta_j \geqslant 0$ with $p(\boldsymbol{x}) = \sum_{j=1}^J \eta_j q_j(\boldsymbol{x})$.

9.27. Let $f(\boldsymbol{x})$ take both positive and negative values. Show that $\hat{\mu}_{\mathrm{mis}}$ of §9.13 really is multiple importance sampling (9.28) by exhibiting densities q_1 and q_2, weight functions w_1 and w_2, and sample sizes n_1 and n_2 for which (9.28) simplifies to $\hat{\mu}_{\mathrm{mis}}$.

9.28. The Horvitz-Thompson estimator is a mainstay of sampling methods for finite populations as described in Cochran (1977) and other references. Which of the heuristics in §9.12 is the closest match to the Horvitz-Thompson estimator?

9.29. The multiple importance sampling estimator of $\mu = \mathbb{E}_p(f(\boldsymbol{X}))$ is

$$\widetilde{\mu}_w = \sum_{j=1}^J \frac{1}{n_j}\sum_{i=1}^{n_j} \omega_j(\boldsymbol{X}_{ij}) \frac{f(\boldsymbol{X}_{ij})p(\boldsymbol{X}_{ij})}{q_j(\boldsymbol{X}_{ij})},$$

where $\omega_1, \ldots, \omega_J$ form a partition of unity, $n_j \geqslant 1$, p and q_j are probability density functions on $\mathbb{R}^d$ and $\boldsymbol{X}_{ij} \sim q_j$ independently. We assume that $q_j(\boldsymbol{x}) > 0$ whenever $(\omega_j p f)(\boldsymbol{x}) \neq 0$. Let $h(\boldsymbol{x})$ be such that $\int h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \eta \in \mathbb{R}^J$ is known.

Develop an unbiased control variate estimator $\widetilde{\mu}_{\omega,\beta}$ that combines multiple importance sampling with a control variate coefficient of β on h.

9.30. Suppose that $\boldsymbol{X} \sim p$ and for each $\boldsymbol{x} \in \mathbb{R}^d$ that $f(\boldsymbol{x})$ is a complex number. Show that there exists a multiple importance sampling scheme which will evaluate $\mathbb{E}(f(\boldsymbol{X}))$ exactly using 4 sample points. You may assume that $\mathbb{E}(f(\boldsymbol{X}))$ exists.

9.31. Here we find the population effective sample size (9.14) for the what-if simulation of Example 9.8. Let p be the Poi(ν) distribution and q be the Poi(λ) distribution. Show that $\mathbb{E}_p(w) = \exp(\lambda - 2\nu + \nu^2/\lambda)$.

9.32. We are about to do a what-if simulation for the distributions Gam(1), Gam(2), Gam(3) and Gam(4). Each of these distributions plays the role of p in its turn. One of them will be our importance sampling density q. Which one should we choose, and why?

Bibliography

Arsham, H., Feuerverger, A., McLeish, D. L., Kreimer, J., and Rubinstein, R. Y. (1989). Sensitivity analysis and the 'what if' problem in simulation analysis. *Mathematical and Computer Modelling*, 12(2):193–219.

Chinneck, J. W. (2009). Practical optimization: a gentle introduction. `http://www.sce.carleton.ca/faculty/chinneck/po.html`.

Cochran, W. G. (1977). *Sampling Techniques (3rd Ed)*. John Wiley & Sons, New York.

Evans, M. J. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10(3):254–272.

Gelfand, I. M. and Fomin, S. V. (2000). *Calculus of variations*. Dover, Mineola, NY.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Hesterberg, T. C. (1987). Importance sampling in multivariate problems. In *Proceedings of the Statistical Computing Section, American Statistical Association 1987 Meeting*, pages 412–417.

Hesterberg, T. C. (1988). *Advances in importance sampling*. PhD thesis, Stanford University.

Hesterberg, T. C. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–192.

Kahn, H. (1950a). Random sampling (Monte Carlo) techniques in neutron attenuation problems, I. *Nucleonics*, 6(5):27–37.

Kahn, H. (1950b). Random sampling (Monte Carlo) techniques in neutron attenuation problems, II. *Nucleonics*, 6(6):60–65.

Kahn, H. and Marshall, A. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278.

Kong, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, University of Chicago.

Lafortune, E. P. and Willems, Y. D. (1993). Bidirectional path tracing. In *Proceedings of CompuGraphics*, pages 95–104.

Oh, M.-S. and Berger, J. O. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association*, 88(422):450–456.

Owen, A. B. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, 4(4):673–684.

Trotter, H. F. and Tukey, J. W. (1956). Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo Methods*, pages 64–79, New York. Wiley.

Veach, E. (1997). *Robust Monte Carlo methods for light transport simulation.* PhD thesis, Stanford University.

Veach, E. and Guibas, L. (1994). Bidirectional estimators for light transport. In *5th Annual Eurographics Workshop on Rendering*, pages 147–162.

Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *SIGGRAPH '95 Conference Proceedings*, pages 419–428. Addison-Wesley.