4.3 IMPORTANCE SAMPLING

While the estimators constructed using samples exactly coming from p has desirable properties as we have seen above, in the majority of cases, we need to employ more complex sampling strategies. A few cases where we need this are summarised below.

- A typical problem arises when computing tail probabilities (also called *rare events*). We may have access to samples directly from $p(x)$, however, sampling from the tail of $p(x)$ might be extremely difficult. For example, consider the Gaussian random variable X with mean 0 and variance 1. The probability of X being larger than 4 is very small, i.e., $\mathbb{P}(X > 4) \approx 0.00003$. Sampling from the tail of this density directly would be very inefficient without further tricks.
- Another typical scenario where we may want to compute expectations with respect to $p(x)$ when we do not have direct samples from it. The standard example for this is the Bayesian setting. Given a prior $p(x)$ and a likelihood $p(y|x)$, we may want to compute the expectations w.r.t. the posterior density $p(x|y)$, i.e., $\mathbb{E}_{p(x|y)}[\varphi(X)]$. In this case, we do not have access to samples from $p(x|y)$ so we need to employ other strategies.

A strategy we will pursue in this section is specific to *Monte Carlo integration*. In other words, we will next describe a strategy where we can compute integrals and expectations w.r.t. a probability density without having access to samples from it. This is slightly different than directly aiming at *sampling* from the density (which can also be used to estimate integrals). While we will look at sampling methods in the following chapters, it is important to note that importance sampling is primarily an integration technique.

4.3.1 BASIC IMPORTANCE SAMPLING

Consider the basic task of estimating the integral

$$\bar{\varphi} = \mathbb{E}_p[\varphi(X)] = \int \varphi(x)p(x)\mathrm{d}x.$$

In this section, as opposed to previous sections, we assume that we cannot sample from p directly (or exactly, e.g., using rejection sampling[3]). However, we assume (in this section) we can evaluate the density $p(x)$ pointwise. We can still estimate this expectation (and compute integrals more generally), using samples from an instrumental, *proposal* distribution q. In other words, we can sample from a *proposal* and we can repurpose these samples to estimate expectations w.r.t. $p(x)$. This resembles the rejection sampling where we have also used a proposal to accept-reject samples. However, in this case, we will employ a different strategy of *weighting* samples and will not throw any of the samples away. The weights we will compute will *weight samples* so that the integral estimate gets closer to the true integral. In order to see how to do this, we compute

$$\begin{aligned}
\bar{\varphi} &= \int \varphi(x)p(x)\mathrm{d}x, \\
&= \int \varphi(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x, \qquad \text{“identity trick”} \qquad (4.11) \\
&= \int \varphi(x)w(x)q(x)\mathrm{d}x, \qquad (4.12)
\end{aligned}$$

[3] Recall that rejection sampling draws i.i.d samples from the density, not *approximate*.

where $w(x) = p(x)/q(x)$ (which is called the *weight function*). We know from Section 4.1 that we can estimate the integral in (4.12) using samples from q. Let $X_i \sim q$ be i.i.d samples from q for $i = 1, \ldots, N$. We can then estimate the integral in (4.12), hence the expectation $\bar{\varphi}$ using

$$\begin{aligned} \bar{\varphi} &= \int \varphi(x) w(x) q(x) \mathrm{d}x \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \mathsf{w}_i \varphi(X_i) = \hat{\varphi}_{\mathrm{IS}}^N, \end{aligned} \tag{4.13}$$

where $\mathsf{w}_i = w(X_i) = p(X_i)/q(X_i)$ are called the *weights*. The weights will play a crucial role throughout this section. The key idea of importance sampling is that, instead of throwing away the samples by rejection, we could reweight them according to their importance. This is why this strategy is called *importance sampling* (IS).

The importance sampling algorithm for this case then can be described relatively straightforwardly. Given $p(x)$ (which we can evaluate), we choose a proposal $q(x)$. Then, we sample $X_i \sim q(x)$ for $i = 1, \ldots, N$ and compute the IS estimator as

$$\hat{\varphi}_{\mathrm{IS}}^N = \frac{1}{N} \sum_{i=1}^{N} \mathsf{w}_i \varphi(X_i),$$

where $\mathsf{w}_i = \frac{p(X_i)}{q(X_i)}$ for $i = 1, \ldots, N$ are the importance weights. We summarise the method in Algorithm 7. In what follows, we will discuss some details of the method.

Algorithm 7 Pseudocode for basic importance sampling

1: Input: The number of samples N
2: **for** $i = 1, \ldots, N$ **do**
3: Sample $X_i \sim q(x)$
4: Compute weights $\mathsf{w}_i = \frac{p(x)}{q(x)}$
5: **end for**
6: Report the estimator

$$\hat{\varphi}_{\mathrm{IS}}^N = \frac{1}{N} \sum_{i=1}^{N} \mathsf{w}_i \varphi(X_i).$$

Remark 4.4. Unlike rejection sampling, in importance sampling, the proposal does not have to dominate the target density. Instead, the crucial requirement for the IS is that the support of the proposal should be the same as the support of the density. More precisely, we need $q(x) > 0$ whenever $p(x) > 0$. This is far less restrictive than the requirement of rejection sampling. Of course, the choice of proposal can still effect the performance of the IS. We will discuss this in more detail.

From Fig. 4.7, one can see an example plot of the target density $p(x)$, the proposal $q(x)$ and the associated weight function $w(x)$. See the caption for more details and intuition.

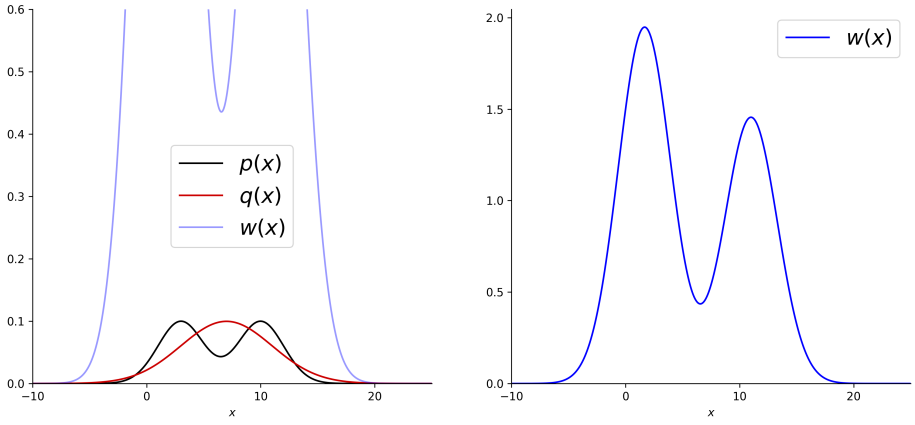

Figure 4.7: An example of target density $p(x)$, the proposal $q(x)$ and the associated weight function $w(x)$. One can see that if $q(x) < p(x)$ (which means fewer samples would be drawn from $q(x)$ in this region), then $w(x) > 1$ to account for this effect. The opposite is also true, since if $q(x) > p(x)$, this means that we would draw more samples than necessary, which should be downweighted, hence $w(x) < 1$ in these regions.

Example 4.6. Consider the problem of estimating $\mathbb{P}(X > 4)$ for $X \sim \mathcal{N}(0, 1)$. While we can exactly sample from this density, given that

$$\mathbb{P}(X > 4) = 3.16 \times 10^{-5},$$

it will be the case that very few of the samples from exact distribution will fall into this tail (Note that, while we know the exact value in this case, we will not know this in general – this is just a demonstrative example). In fact, a standard run with $N = 10000$ gives exactly zero samples that satisfy $X_i > 4$, hence provides the estimate as zero! It is obvious that this is not a great way to estimate the probability and we can use importance sampling for this. Consider a proposal $q(x) = \mathcal{N}(6, 1)$. This will draw a lot of samples from the region $X > 4$ and we can reweight this samples w.r.t. the target density using the IS estimator in (4.13). A standard run in this case with $N = 10000$ results in

$$\hat{\varphi}^N_{\text{IS}} = 3.18 \times 10^{-5},$$

which is obviously a much closer number to the truth.

One can next prove that the estimator $\hat{\varphi}^N_{\text{IS}}$ is unbiased.

Proposition 4.3. *The estimator $\hat{\varphi}^N_{IS}$ is unbiased, i.e.,*

$$\mathbb{E}_{q(x)}[\hat{\varphi}^N_{IS}] = \bar{\varphi}.$$

Proof. We simply write

$$\begin{aligned}
\mathbb{E}_q[\hat{\varphi}_{\text{IS}}^N] &= \mathbb{E}_{q(x)}\left[\frac{1}{N}\sum_{i=1}^N \mathsf{w}_i\varphi(X_i)\right] \\
&= \mathbb{E}_q\left[\frac{1}{N}\sum_{i=1}^N \frac{p(X_i)}{q(X_i)}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_q\left[\frac{p(X_i)}{q(X_i)}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^N \int \frac{p(x)}{q(x)}\varphi(x)q(x)\mathrm{d}x \qquad \text{since } X_i \sim q(x) \\
&= \int \varphi(x)p(x)\mathrm{d}x, \\
&= \bar{\varphi},
\end{aligned}$$

which completes the proof. □

An important quantity in IS is the *variance* of the estimator $\hat{\varphi}_{\text{IS}}^N$. The variance of the estimator is a measure of how much the estimator fluctuates around its expected value. The variance of the IS estimator (4.13) is given by the following proposition.

Proposition 4.4. *The variance of the estimator $\hat{\varphi}_{IS}^N$ is given by*

$$\mathsf{var}_q[\hat{\varphi}_{IS}^N] = \frac{1}{N}\left(\mathbb{E}_q[w^2(X)\varphi^2(X)] - \bar{\varphi}^2\right).$$

Proof. Next we write out the estimator $\hat{\varphi}_{\text{IS}}^N$ in (4.13)

$$\begin{aligned}
\mathsf{var}_q[\hat{\varphi}_{\text{IS}}^N] &= \mathsf{var}_q\left[\frac{1}{N}\sum_{i=1}^N \mathsf{w}_i\varphi(X_i)\right] \\
&= \frac{1}{N^2}\mathsf{var}_q\left[\sum_{i=1}^N w(X_i)\varphi(X_i)\right] \\
&= \frac{1}{N}\mathsf{var}_q\left[w(X)\varphi(X)\right] \qquad \text{where } X \sim q(x) \\
&= \frac{1}{N}\left(\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] - \mathbb{E}_q\left[w(X)\varphi(X)\right]^2\right) \\
&= \frac{1}{N}\left(\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] - \bar{\varphi}^2\right),
\end{aligned}$$

which concludes the proof. We have used the fact that the variance of the sum of independent random variables is the sum of the variances. □

One can see that this easily leads to the bound for the standard devation $\mathsf{std}_q[\hat{\varphi}_{\text{IS}}^N] \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. Also, we still have the result for the relative absolute error as

$$|\hat{\varphi}_{\text{IS}}^N - \bar{\varphi}| \leq \frac{V}{\sqrt{N}},$$

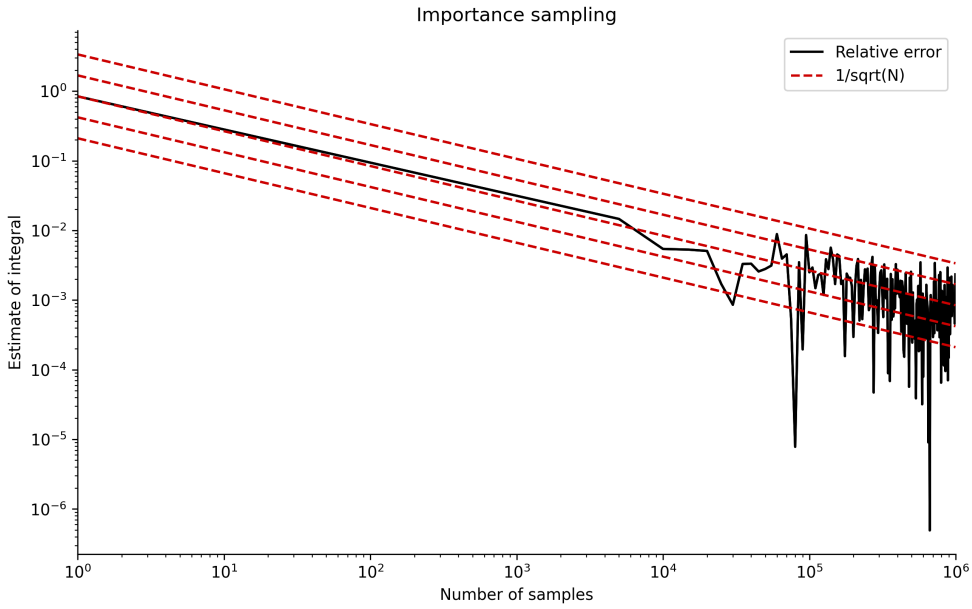

Figure 4.8: The importance sampling estimator $\hat{\varphi}_{\mathrm{IS}}^{N}$ is plotted against the number of samples N for the example in in Fig . 4.7, for $\varphi(x) = x^2$. This demonstrates that the random error in the IS case also satisfies $\mathcal{O}(1/\sqrt{N})$ convergence rate.

where V is an almost surely finite random variable. As in the perfect MC case, we will not prove this result as it is beyond our scope, but curious reader can refer to Corollary 2.2 in Akyildiz (2019) (which also holds for the self normalised case which will be introduced below). A demonstration of this rate for importance sampling can be seen from Fig. 4.8.

We can see that the variance of the IS estimator is finite if

$$\mathbb{E}_q[w^2(X)\varphi^2(X)] < \infty.$$

This implies that

$$\int w^2(x)\varphi^2(x)q(x)\mathrm{d}x = \int \frac{p(x)}{q(x)}\varphi^2(x)p(x)\mathrm{d}x < \infty.$$

In other words, for our importance sampling estimate to be well-defined, the ratio

$$\frac{p^2(x)}{q(x)}\varphi^2(x)$$

has to be integrable. We will see next an example where this condition is not satisfied.

Example 4.7 (Infinite variance IS, Example 3.8 from Robert and Casella (2010))**.** Consider the target

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2},$$

which is the Cauchy density. Let us choose the proposal

$$q(x) = \mathcal{N}(x; 0, 1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

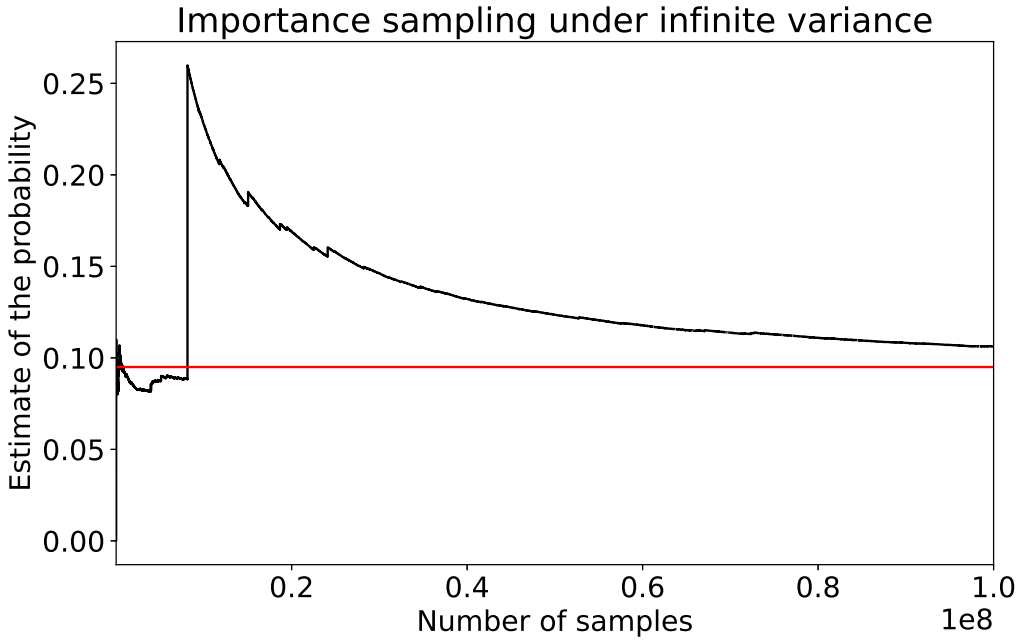

Figure 4.9: Estimating $\mathbb{P}(2 < X < 6)$ where X is Cauchy with $q(x) = \mathcal{N}(0, 1)$. The true value is plotted in red and the estimator value in black.

The ratio $\frac{p(x)}{q(x)} \propto \exp(x^2/2)/(1 + x^2)$ is explosive. This can result in problematic situations even if φ ensures that the variance is finite. For example, consider the problem of estimating $\mathbb{P}(2 < X < 6)$. One example run for this case can be seen from Fig. 4.9. One can see that the estimator in this case is unstable and cannot be reliably used.

Remark 4.5 (Optimal proposal)**.** We can try to inspect the variance expression to figure out which proposals can give us variance reduction. From Prop. 4.4, it follows that we have

$$\mathsf{var}_q[\hat{\varphi}^N_{\mathrm{IS}}] = \frac{1}{N}\mathsf{var}_q\left[w(X)\varphi(X)\right].$$

This means that minimising the variance of the IS estimator is the same as minmising the variance of the function $w(x)\varphi(x)$. Moreover, looking at the expression,

$$\mathsf{var}_q[\hat{\varphi}^N_{\mathrm{IS}}] = \frac{1}{N}\left(\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] - \bar{\varphi}^2\right),$$

we can see that since $\bar{\varphi}^2 > 0$ (which is independent of the proposal), we should choose a proposal that minimises $\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right]$. We can lower bound this quantity using Jensen's inequality:

$$\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] \geq \mathbb{E}_q\left[w(X)|\varphi(X)|\right]^2,$$

where we used the fact that $(\cdot)^2$ is a convex function (For a convex function f, Jensen's inequality states that $\mathbb{E}_q[f(X)] \geq f(\mathbb{E}_q[X])$.). Using $w(x) = p(x)/q(x)$, we arrive at the

following lower bound:

$$\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] \geq \mathbb{E}_p\left[|\varphi(X)|\right]^2. \tag{4.14}$$

Now let us expand the term $\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right]$ out and write

$$\begin{aligned}\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] &= \mathbb{E}_q\left[\frac{p^2(X)}{q^2(X)}\varphi^2(X)\right] \\ &= \int \frac{p^2(x)}{q^2(x)}\varphi^2(x)q(x)\mathrm{d}x \\ &= \int p(x)\frac{p(x)}{q(x)}\varphi^2(x)\mathrm{d}x, \\ &= \mathbb{E}_p\left[w(X)\varphi^2(X)\right]. \end{aligned} \tag{4.15}$$

The last equation, eq. (4.15), suggests that we can choose a proposal such that we attain the lower bound of this function (4.14) (which means that it would be the minimiser). In particular, if we choose a proposal $q(x)$ such that

$$w(x) = \frac{p(x)}{q(x)} = \frac{\mathbb{E}_p[|\varphi(X)|]}{|\varphi(x)|}$$

is satisfied, then (4.15) would be equal to the lower bound (4.14). This implies that

$$q_\star(x) = p(x)\frac{|\varphi(x)|}{\mathbb{E}_p[|\varphi(X)|]}, \tag{4.16}$$

would minimise the variance of the importance sampling estimator.

Choosing $q_\star$ as the proposal, one can see that the variance of the IS estimator satisfies

$$\begin{aligned}\mathsf{var}_{q_\star}[\hat{\varphi}^N_{\mathrm{IS}}] &= \frac{1}{N}\mathbb{E}_p\left[|\varphi(X)|\right]^2 - \frac{1}{N}\bar{\varphi}^2 \\ &\leq \frac{1}{N}\mathbb{E}_p\left[\varphi^2(X)\right] - \frac{1}{N}\bar{\varphi}^2 \\ &= \mathsf{var}_p[\hat{\varphi}^N_{\mathrm{MC}}],\end{aligned}$$

therefore we obtain

$$\mathsf{var}_{q_\star}[\hat{\varphi}^N_{\mathrm{IS}}] \leq \mathsf{var}_p[\hat{\varphi}^N_{\mathrm{MC}}],$$

i.e., a variance reduction. In fact, one can show that, if $\varphi(x) \geq 0$ for all $x \in \mathbb{R}$, then the variance of the IS estimator with optimal proposal $q_\star$ is equal to zero.

We note that this optimal construction of the proposal (4.16) is not possible to implement in practice. It requires the knowledge of the very quantity we want to estimate, namely, $\mathbb{E}_p[|\varphi(X)|]$! But in general, we can choose proposals that minimise the variance of the IS estimator where possible. This idea has been used in the literature to construct proposals that minimise the variance of the estimator, see, e.g., Akyildiz and Míguez (2021) and references therein. Within the context of this course, we will construct some simple

examples for this purpose later.

4.3.2 SELF-NORMALISED IMPORTANCE SAMPLING

As mentioned several times in past chapters, in many scenarios, we have access to the *unnormalised* density, i.e., given p, we can evaluate it up to a normalising constant. As usual, we denote this density $\bar{p}(x)$ and recall that it is related to p by

$$p(x) = \frac{\bar{p}(x)}{Z}$$

where $Z = \int \bar{p}(x)\mathrm{d}x$. In the context of Bayesian inference, we usually have an unnormalised posterior density $\bar{p}(x|y) \propto p(y|x)p(x)$. In the previous section, we have built an importance sampling estimator for the case where we have access to the normalised density. In this section, we will generalise the idea and assume we only have access to the unnormalised density.

Consider, again, the problem of estimating expectations of a given density p. For the case where we can only evaluate $\bar{p}(x)$, one way to estimate this expectation is to sample from a proposal distribution q and rewrite the integral as

$$\begin{aligned}\bar{\varphi} &= \int \varphi(x)p(x)\mathrm{d}x, \\ &= \frac{\int \varphi(x)\frac{\bar{p}(x)}{q(x)}q(x)\mathrm{d}x}{\int \frac{\bar{p}(x)}{q(x)}q(x)\mathrm{d}x},\end{aligned} \tag{4.17}$$

where we use the fact that $p(x) = \bar{p}(x)/Z$. This gives us two separate integration problems, one to estimate the numerator and one to estimate the denominator. We will estimate both quantities using samples from $q(x)$[4].

Let us now introduce the unnormalised weight function $W(x)$

$$W(x) = \frac{\bar{p}(x)}{q(x)},$$

which is analogous to the normalised weight function $w(x)$ in the previous section. Using $X_i \sim q(x)$ and building the Monte Carlo estimator of the numerator and denominator, we arrive at the following estimator of the (4.17):

$$\begin{aligned}\hat{\varphi}^N_{\text{SNIS}} &= \frac{\frac{1}{N}\sum_{i=1}^N \varphi(X_i)W(X_i)}{\frac{1}{N}\sum_{i=1}^N W(X_i)}, \\ &= \frac{\sum_{i=1}^N \varphi(X_i)W(X_i)}{\sum_{i=1}^N W(X_i)} \\ &= \sum_{i=1}^N \bar{\mathsf{w}}_i \varphi(X_i),\end{aligned} \tag{4.18}$$

where

$$\bar{\mathsf{w}}_i = \frac{W(X_i)}{\sum_{i=1}^N W(X_i)},$$

are the *normalised* weights. This estimator (4.18) is called the self-normalised importance sampling (SNIS) estimator.

[4] We do not have to, see, e.g., Lamberti et al. (2018).

Algorithm 8 Pseudocode for self-normalised importance sampling

1: Input: The number of samples N
2: **for** $i = 1, \dots, N$ **do**
3: Sample $X_i \sim q(x)$
4: Compute weights $\bar{\mathsf{w}}_i = \frac{\bar{p}(x)}{q(x)}$
5: **end for**
6: Report the estimator

$$\hat{\varphi}^N_{\text{SNIS}} = \sum_{i=1}^{N} \bar{\mathsf{w}}_i \varphi(X_i).$$

Remark 4.6 (Bias and variance)**.** As opposed to the normalised case, the estimator $\hat{\varphi}^N_{\text{SNIS}}$ is *biased*. The reason of this bias can be seen by recalling the integral (4.17). By sampling from $q(x)$, we can construct unbiased estimates of the numerator and denominator. However, the ratio of these two quantities is biased in general. However, it can be shown that the bias of the SNIS estimator decreases with a rate $\mathcal{O}(1/N)$ (Agapiou et al., 2017).

Since the SNIS estimator is biased, we can not use the same variance formula as in the previous section. It makes sense to consider the $\mathsf{MSE}(\hat{\varphi}^N_{\text{SNIS}})$ instead. However, this quantity is challenging to control in general – without bounded test functions. With bounded test functions, it is possible to show that the $\mathsf{MSE}(\hat{\varphi}^N_{\text{SNIS}})$ is controlled with a rate $\mathcal{O}(1/N)$ (Agapiou et al., 2017; Akyildiz and Míguez, 2021). We will not go into the details of this result here.

We can now describe the estimation procedure using SNIS. Given an unnormalised density $\hat{p}(x)$, we first sample N samples from a proposal, $X_1, \dots, X_N \sim q(x)$, and then compute normalised weights

$$\bar{\mathsf{w}}_i = \frac{W(X_i)}{\sum_{i=1}^{N} W(X_i)}.$$

where $W(x) = \frac{\bar{p}(x)}{q(x)}$. Finally, we compute the estimator

$$\hat{\varphi}^N_{\text{SNIS}} = \sum_{i=1}^{N} \bar{\mathsf{w}}_i \varphi(X_i).$$

In the following, we describe the algorithm, which is given in Algorithm 8.

4.4 IMPLEMENTATION, ALGORITHMS, DIAGNOSTICS

When implementing the IS or SNIS, there are several numerical considerations that need to be taken into account. Especially for SNIS, where the weight normalisation takes place, several numerical problems can arise for complex distributions that would prevent us from implementing them successfully.

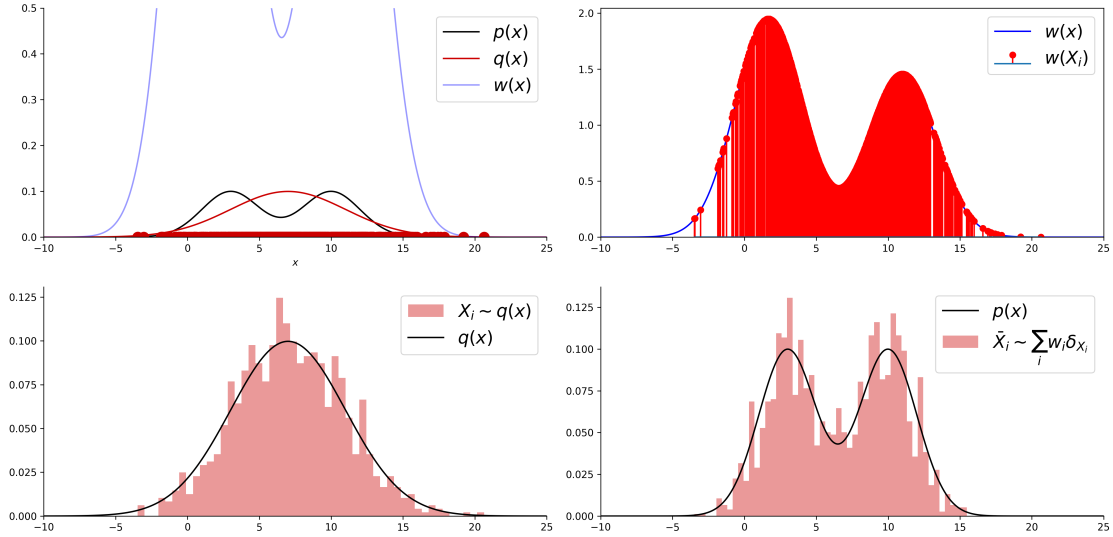

Figure 4.10: Top left shows the target density, the proposal, and the weight function. Top right shows the samples with their respective weights. Bottom left shows that these samples are indeed approximately distribution w.r.t. $q(x)$ (just with attached weights). Bottom right shows that we can resample these existing samples to obtain a new set of samples $\bar{X}_i$ that are distributed (approximately)according to $p(x)$.

4.4.1 COMPUTING WEIGHTS

In the case of SNIS, we have stated that the weights are computed as

$$\bar{\mathsf{w}}_i = \frac{W(X_i)}{\sum_{i=1}^N W(X_i)},$$

where $W(x) = \frac{\bar{p}(x)}{q(x)}$. However, this formula can be numerically ill-behaved for complex distributions. We have the same problem as in Example 3.12 where we also needed to compute the ratio $p(x)/q(x)$ (for the acceptance probability). To mitigate this, the weighting step is implemented as follows. We first compute log *unnormalised* weights:

$$\log \mathsf{W}_i = \log \bar{p}(X_i) - \log q(X_i).$$

However, directly applying exponentiation and normalisation can also lead to numerical problems. Note that, we will normalise these weights (e.g. multiplying them with a constant does not change the result), we can use this to our advantage. A common numerical trick is to subtract the maximum log weight from all weights:

$$\log \tilde{\mathsf{W}}_i = \log \bar{p}(X_i) - \log q(X_i) - \max_{i=1,\dots,N} \log \mathsf{W}_i.$$

This ensures that the maximum weight is 0 and all other weights are negative. We can now exponentiate the weights and normalise them:

$$\bar{\mathsf{w}}_i = \frac{\exp(\log \tilde{\mathsf{W}}_i)}{\sum_{i=1}^N \exp(\log \tilde{\mathsf{W}}_i)}.$$

Note that, this does not change the computation, just done for numerical stability.

4.4.2 SAMPLING IMPORTANCE RESAMPLING

We can also use the SNIS estimator as a sampler (Robert and Casella, 2004). Recall that, the SNIS estimator provides us an estimator of the distribution $p(x)$ as

$$p(x)\mathrm{d}x \approx \tilde{p}^N(x)\mathrm{d}x = \sum_{i=1}^{N} \bar{\mathsf{w}}_i \delta_{X_i}(x)\mathrm{d}x.$$

This p^N can be seen as a weighted distribution. By drawing samples from this distribution, we may also approximately sample from $p(x)$ (recall that IS based ideas here are just introduced for integration so far). We can then draw[5]

$$k \sim \text{Discrete}(\bar{\mathsf{w}}_1, \ldots, \bar{\mathsf{w}}_N),$$

and set new samples

$$\bar{X}_i = X_k.$$

This amounts to *resampling* the existing samples w.r.t. their weights. A demonstration of this idea can be seen from Figure 4.10.

4.4.3 DIAGNOSTICS FOR IMPORTANCE SAMPLING

It is important to have a good intuition and diagnostic tools to understand the performance of the IS and SNIS estimators. We first start with the effective sample size (ESS).

Definition 4.1 (Effective Sample Size)**.** *To measure the sample efficiency, one measure that is used in the literature is the effective sample size (ESS) which is given by*

$$\mathsf{ESS}_N = \frac{1}{\sum_{i=1}^{N} \bar{\mathsf{w}}_i^2},$$

for the SNIS estimator.

In order to see the meaning of the ESS, consider the case where $\bar{\mathsf{w}}_i = 1/N$ where we have an equally weighted sample. This means all samples are equally considered and in this case we have $\mathsf{ESS}_N = N$. On the other hand, if we have a sample X_i where $\bar{\mathsf{w}}_i = 1$ and, hence, $\bar{\mathsf{w}}_j = 0$ for every $j \neq i$, we obtain $\mathsf{ESS}_N = 1$. This means, we *effectively* have one sample which is the goal of the estimator. ESS is used to measure importance samplers and importance sampling-based estimators in the literature (Elvira et al., 2018). Note that the ESS_N takes values between 1 and N, i.e., $1 \leq \mathsf{ESS}_N \leq N$.

4.4.4 MIXTURE IMPORTANCE SAMPLING

Sometimes the target density $p(x)$ can be multimodal, therefore it is beneficial to use mixture densities as proposals (Owen, 2013). We have seen in previous chapters how to sample from a mixture. Let us define a proposal

$$q_\alpha(x) = \sum_{k=1}^{K} \alpha_k q_k(x),$$

[5]Note that here the weights $\bar{w}_i$ are normalised. Even in the basic IS case, we need to normalise weights (**just for resampling**) as they do not naturally sum up to one.

where $\alpha_k \geq 0$ and $\sum_{k=1}^K \alpha_k = 1$. In this version of the method, we just sample from the mixture proposal $X_i \sim q_\alpha(x)$ and then, given an unnormalised target $\bar{p}$, compute the importance weights as

$$\bar{\mathsf{w}}_i = \frac{W(X_i)}{\sum_{i=1}^N W(X_i)},$$

where

$$W(X_i) = \frac{\bar{p}(X_i)}{\sum_{k=1}^K \alpha_k q_k(X_i)}.$$

The computational concerns may arise in this situation too, as the denominator as a sum of densities and its log can be tricky to compute. In these cases, we can use the log-sum-exp trick to compute the log of the denominator.

4.5 EXAMPLES

In this section, we solve various examples regarding the Monte Carlo and importance sampling estimators.

Example 4.8 (Bayesian inference using importance sampling)**.** Self normalised IS is a natural choice for Bayesian inference. Assume that we have a prior $p(x)$ and a likelihood $p(y|x)$. The posterior is given by

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\mathrm{d}x}.$$

Let $\bar{p}(x|y) := p(y|x)p(x)$ as usual and design an importance sampler to estimate expectations of the form

$$\mathbb{E}_{p(x|y)}[\varphi(x)] = \int \varphi(x)p(x|y)\mathrm{d}x.$$

Assume that we choose $q(x)$ and decided to perform SNIS. We first sample $X_1, \ldots, X_N \sim q(x)$ and construct

$$\mathsf{W}_i = \frac{\bar{p}(X_i|y)}{q(X_i)} = \frac{p(y|X_i)p(X_i)}{q(X_i)}.$$

We can now normalise these weights and obtain

$$\bar{\mathsf{w}}_i = \frac{\mathsf{W}_i}{\sum_{i=1}^N \mathsf{W}_i},$$

which will give us the Monte Carlo estimator:

$$\mathbb{E}_{p(x|y)}[\varphi(x)] \approx \sum_{i=1}^N \bar{\mathsf{w}}_i \varphi(X_i).$$

It is useful to recall that this estimator is biased, since it is a SNIS estimator. However, as a byproduct of this estimator, we can also obtain an **unbiased** estimate of the marginal likelihood $p(y)$. Note that, this is already provided by the SNIS estimator

$$p(y) \approx \frac{1}{N}\sum_{i=1}^{N} \mathsf{W}_i.$$

In order to see this, let us compute

$$\begin{aligned}
\mathbb{E}_q\left[\sum_{i=1}^{N} \mathsf{W}_i\right] &= \sum_{i=1}^{N} \mathbb{E}_q\left[\frac{p(y|X_i)p(X_i)}{q(X_i)}\right] \\
&= N\mathbb{E}_q\left[\frac{p(y|X_i)p(X)}{q(X)}\right] \\
&= N\int \frac{p(y|x)p(x)}{q(x)} q(x)\mathrm{d}x \\
&= Np(y).
\end{aligned}$$

As we have seen before, a number of interesting problems require computing normalising constants, including model selection and prediction. SNIS estimators are very useful in the sense that they provide an unbiased estimate of it.

Example 4.9 (Marginal likelihood using importance sampling)**.** We have seen that we can get unbiased estimates of the marginal likelihood in the previous example. We will now focus on a sole integration problem and see how we can use importance sampling to compute the marginal likelihood. Note that, we have

$$p(y) = \int p(y|x)p(x)\mathrm{d}x,$$

for some prior $p(x)$ and likelihood $p(y|x)$. Note, as we mentioned before, in this case $p(x)$ can be seen as the distribution to sample from and $\varphi(x) = p(y|x)$ to obtain the standard problem of integration $\int \varphi(x)p(x)\mathrm{d}x$. A naive way to approximate this quantity (as we have seen before) is to sample i.i.d from $p(x)$ and approximate the integral, i.e., $X_1, \ldots, X_N \sim p(x)$ and write

$$p^N_{\mathrm{MC}}(y) = \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i) = \frac{1}{N}\sum_{i=1}^{N} p(y|X_i).$$

We can now look at the variance of this estimator

$$\mathsf{var}_p\left[p^N_{\mathrm{MC}}(y)\right] = \frac{1}{N}\mathsf{var}_{p(x)}[p(y|x)].$$

This quantity may depend on the prior-likelihood selection and can be large.

Let us take Remark 4.5 seriously and search for the optimal proposal $q_\star$. From (4.16), we can see that

$$q_\star(x) = p(x)\frac{|\varphi(x)|}{\mathbb{E}_p(x)[|\varphi(x)|]}.$$

In this case, however, we have $\varphi(x) = p(y|x)$ (and $|\varphi(x)| = \varphi(x)$ since the likelihood is positive everywhere). We can now write

$$q_\star(x) = p(x)\frac{p(y|x)}{\mathbb{E}_p[p(y|x)]} = p(x)\frac{p(y|x)}{p(y)}.$$

In other words, the optimal proposal is the posterior itself! Now we can compute the IS estimator variance where we plug $q_\star = p(x|y)$. Note to explore variance, we write

$$\mathsf{var}_{q_\star}[p^N_{\mathrm{IS}}(y)] = \frac{1}{N}\left(\mathbb{E}_{q_\star}\left[\left(\frac{p(x)}{q_\star(x)}\right)^2 p(y|x)^2\right] - p(y)^2\right).$$

We compute the first term in brackets,

$$\begin{aligned}
\mathbb{E}_{q_\star}\left[\left(\frac{p(x)}{q_\star(x)}\right)^2 p(y|x)^2\right] &= \int \frac{p^2(x)}{q_\star^2(x)} p(y|x)^2 q_\star(x)\mathrm{d}x \\
&= \int \frac{p^2(x)}{q_\star(x)} p(y|x)^2 \mathrm{d}x \\
&= \int \frac{p^2(x)}{p(x|y)} p(y|x)^2 \mathrm{d}x \\
&= \int \frac{p^2(x)p(y)}{p(y|x)p(x)} p(y|x)^2 \mathrm{d}x \\
&= p(y)\int p(x)p(y|x)\mathrm{d}x \\
&= p(y)^2.
\end{aligned}$$

Plugging this back into the above variance expression $\mathsf{var}_{q_\star}[p^N_{\mathrm{IS}}(y)]$, we obtain

$$\mathsf{var}_{q_\star}[p^N_{\mathrm{IS}}(y)] = \frac{1}{N}\left(p(y)^2 - p(y)^2\right) = 0.$$

It can be seen that we can achieve zero variance, but as we mentioned before, this required us to know the posterior density.

Example 4.10 (Minimum variance IS)**.** We are given an exponential distribution

$$p_\lambda(x) = \lambda \exp(-\lambda x).$$

and want to compute $\mathbb{P}(X > K)$. For example, $\lambda = 2$ and $K = 6$, we can analytically compute $\mathbb{P}(X > 6) = 6.144 \times 10^{-6}$. Therefore, we could not use the standard MC estimator to compute this probability. In order to mitigate the problem, we would like to use another exponential proposal $q_\mu(x)$ which may have higher probability concentration around 6. We would like to design our proposal using the minimum variance criterion (see

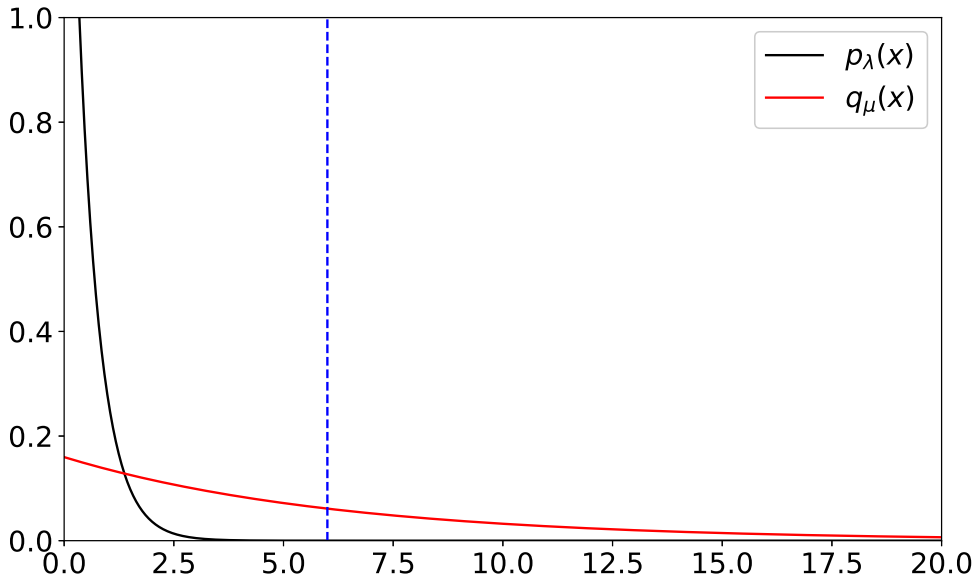

Figure 4.11: The density of the exponential distribution p_λ, the proposal q_μ and $K = 6$

Remark 4.5). Accordingly, we would like to find μ such that

$$\mu_\star = \operatorname*{argmin}_\mu \mathbb{E}_q \left[w^2(X)\varphi^2(X) \right].$$

In this case, note that we have $\varphi(x) = \mathbf{1}_{\{x>K\}}(x)$. In order to do this, we write next

$$\begin{aligned}
\mathbb{E}_{q_\mu}\left[\left(w^2(X)\varphi^2(X)\right)\right] &= \int \frac{p_\lambda(x)^2}{q_\mu(x)^2} q_\mu(x)\varphi^2(x)\mathrm{d}x, \\
&= \int_K^\infty \frac{p_\lambda(x)}{q_\mu(x)} p_\lambda(x)\mathrm{d}x, \\
&= \int_K^\infty \frac{\lambda^2 e^{-2\lambda x}}{\mu e^{-\mu x}}\mathrm{d}x, \\
&= \frac{\lambda^2}{\mu}\int_K^\infty e^{-(2\lambda-\mu)x}\mathrm{d}x.
\end{aligned}$$

Note at this stage that in order for this integral to be finite, we need to have $2\lambda - \mu > 0$. Therefore, we limit for $\mu \in (0, 2\lambda)$. In order to compute this, we can multiply and divide by $(2\lambda - \mu)$ and obtain

$$\mathbb{E}_{q_\mu}\left[\left(w^2(X)\varphi^2(X)\right)\right] = \frac{\lambda^2}{\mu(2\lambda-\mu)}\int_K^\infty (2\lambda-\mu)e^{-(2\lambda-\mu)x}\mathrm{d}x,$$

and using the CDF of the exponential distribution, we obtain

$$g(\mu) = \mathbb{E}_{q_\mu}\left[\left(w^2(X)\varphi^2(X)\right)\right] = \frac{\lambda^2}{\mu(2\lambda-\mu)}\left[1 - 1 + e^{-(2\lambda-\mu)K}\right]. \tag{4.19}$$

Now we optimise $g(\mu)$ w.r.t. μ. As usual, we compute first $\log$ (and drop the terms unrelated to μ as they will not matter in optimisation)

$$\log g(\mu) =^c -\log \mu - \log(2\lambda - \mu) + \mu K.$$

Computing

$$\frac{\mathrm{d}}{\mathrm{d}\mu} \log g(\mu) = -\frac{1}{\mu} + \frac{1}{2\lambda - \mu} + K,$$

Setting this to zero, we obtain

$$\begin{aligned} &-\frac{1}{\mu} + \frac{1}{2\lambda - \mu} + K = 0, \\ \Rightarrow &-(2\lambda - \mu) + \mu + K\mu(2\lambda - \mu) = 0, \\ \Rightarrow &K\mu^2 - 2K\mu\lambda + 2\lambda - 2\mu = 0, \\ \Rightarrow &K\mu^2 - 2(K\lambda + 1)\mu + 2\lambda = 0. \end{aligned}$$

This is a quadratic equation, therefore we will have two solutions:

$$\begin{aligned} \mu &= \frac{2(K\lambda + 1) \pm \sqrt{(2K\lambda + 2)^2 - 8K\lambda}}{2K}, \\ &= \frac{2(K\lambda + 1) \pm \sqrt{4K^2\lambda^2 + 4}}{2K}. \end{aligned}$$

If we inspect this solution, if we choose μ to be the sum of the two terms, we will then have $\mu > 2\lambda$ which is a violation of a condition condition we imposed for the integral to be finite. Therefore, we arrive at

$$\mu_\star = \frac{2(K\lambda + 1) - \sqrt{4K^2\lambda^2 + 4}}{2K}.$$

After this tedious computation, we can now verify the reduction in variance and estimation quality. Let us now set $K = 6$ and $\lambda = 2$. See Fig. 4.11 for plot of p_λ, $K = 6$ and $q_{\mu_\star}$ (the optimal exponential proposal). We can see that the proposal puts much higher mass to the right of K. A standard run for $N = 10^5$ samples gives us **zero** samples in the region of $X > 6$, therefore the standard MC estimate is zero! Compared to $\hat{\varphi}^N = 0$, using $q_{\mu_\star}$ as a proposal, we obtain $\hat{\varphi}^N_{\mathrm{IS}} = 6.08 \times 10^{-6}$ which is a much better estimate.

Let us compare the theoretical variances of two estimators. The standard variance of $\hat{\varphi}^N$ is

$$\mathsf{var}_p(\hat{\varphi}^N) = \frac{1}{N}\mathsf{var}_p(\varphi(X)),$$

where

$$\begin{aligned} \mathsf{var}_p(\varphi(X)) &= \int \varphi(x)^2 p_\lambda(x)\mathrm{d}x - \left(\int \varphi(x) p_\lambda(x)\mathrm{d}x\right)^2, \\ &= \int_K^\infty p_\lambda(x)\mathrm{d}x - \left(\int_K^\infty p_\lambda(x)\mathrm{d}x\right)^2. \end{aligned}$$

Using CDFs, we can compute this quantity hence can obtain the estimate of the variance for $\hat{\varphi}^N$.

Now set $\mu = \mu_\star$. The variance of $\hat{\varphi}^N_{\text{IS}}$ is given by (see Prop. 4.4)

$$\mathsf{var}_q[\hat{\varphi}^N_{\text{IS}}] = \frac{1}{N}\left(\mathbb{E}_q[w^2(X)\varphi^2(X)] - \bar{\varphi}^2\right),$$

We have already computed the term $\mathbb{E}_q[w^2(X)\varphi^2(X)]$ in Eq. (4.19). The second term is the true integral, which we also summarised how to compute above, i.e., $\bar{\varphi} = \int_K^\infty p_\lambda(x)\mathrm{d}x$ which can be computed using the exponential CDF. In this particular case, we compute

$$\mathsf{var}_q[\hat{\varphi}^N_{\text{IS}}] = \frac{1}{N}\left(g(\mu_\star) - \bar{\varphi}^2\right).$$

The theoretical variance of the naive MC estimator is given by 6.14×10^{-7} for $N = 10$ samples vs. the IS estimator variance is 6.04×10^{-11} for the same amount of samples. This is a huge improvement in the variance of the estimation.