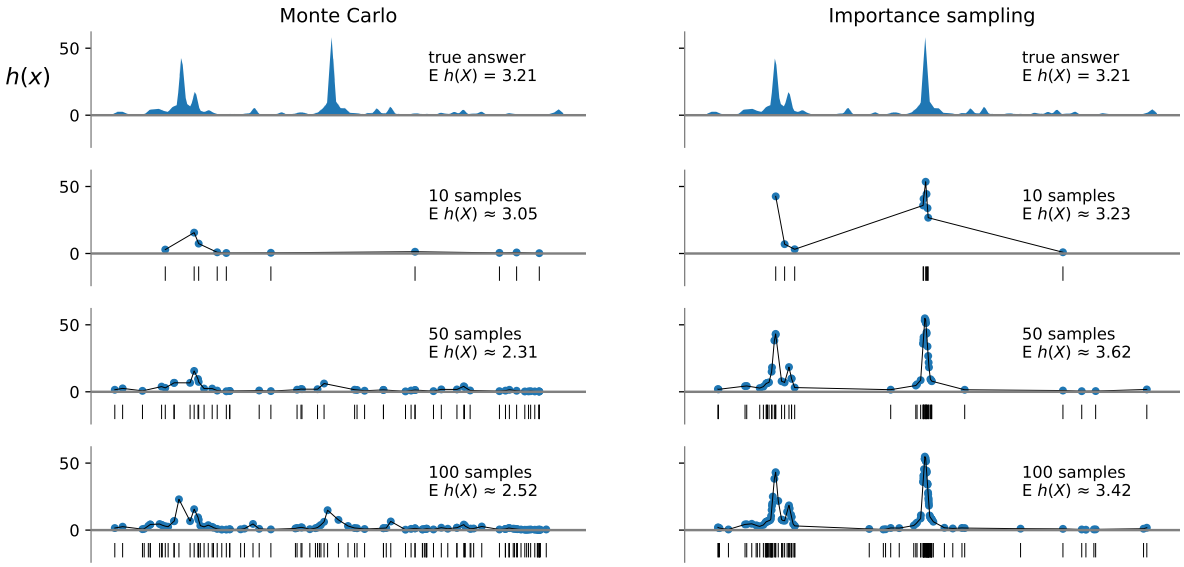
5.3. Importance sampling *

Monte Carlo says that, if X is a random variable and $h(\cdot)$ is some function of interest, then we can approximate $\mathbb{E}\,h(X)$ by

$$\mathbb{E}\,h(X) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i), \quad x_i \text{ sampled from } X.$$

Sometimes Monte Carlo doesn't work too well. In this plot below, $h(x)$ is a very spiky function and $X \sim \text{Uniform}[0,1]$ (which means that $\mathbb{E}\,h(X)$ is the area under the curve in the top plot). It takes very many samples before the sampled function even 'sees' the spikes, resulting in a Monte Carlo approximation that underestimates the true answer.



(What we actually see is more subtle than 'always underestimates'. For most of the samples $(x_1, \ldots, x_n)$ that we might draw we get an underestimate, but there's a small probability of drawing a sample that has lots of x_i at the spikes, leading to a wild overestimate. When we're using Monte Carlo, and we've drawn a sample, we don't know which of these two cases we've hit. In other words, our approximation procedure produces noisy estimates.)

Importance sampling is a modification to Monte Carlo in which we sample the x_i from a different distribution, call it $\tilde{X}$. The idea of $\tilde{X}$ is to make it more likely to get samples in the regions that matter most. In the example above, we had better make sure to get samples in the regions where $h(\cdot)$ has spikes, if we want an accurate picture and an accurate estimate of $\mathbb{E}\,h(X)$. Of course we have to do something to correct for this biased sampling, because otherwise we'd overestimate. The appropriate correction is:

Importance sampling approximation. We can approximate $\mathbb{E}\,h(X)$ by picking some other distribution $\tilde{X}$, called the *sampling distribution*, and then

$$\mathbb{E}\,h(X) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)\frac{\Pr_X(x_i)}{\Pr_{\tilde{X}}(x_i)}, \quad x_i \text{ sampled from } \tilde{X}.$$

The only restriction on the sampling distribution is that we require

$$\Pr_{\tilde{X}}(x) > 0 \qquad \text{wherever} \qquad h(x)\Pr_X(x) \neq 0.$$

HOW TO CHOOSE THE SAMPLING DISTRIBUTION

We have a whole design space of sampling distributions to choose from. A reasonable goal is to pick it so as to minimize the noisiness of the approximation, for example to pick it to minimize

$$\operatorname{Var} g(\tilde{X}) \qquad \text{where} \qquad g(x) = h(x)\frac{\Pr_X(x)}{\Pr_{\tilde{X}}(x)}.$$

If $h(x) \geq 0$, as it is in many of the uses we'll make of importance sampling, the answer is staring us in the face: pick the sampling distribution so that

$$\Pr_{\tilde{X}}(x) = \text{const} \times h(x)\Pr_X(x)$$

where the constant is whatever is needed to make $\Pr_{\tilde{X}}$ be a valid distribution. If we choose this then $g(\tilde{X})$ is constant, so the variance is zero, and we only need a single sample! There is of course a catch. We'd need to work out the constant, which is $1/\int_x h(x)\Pr_X(x)\,dx$, which is the exactly the integral that we're trying to approximate in the first place.

Nonetheless, this answer is morally right. If $h(x) \geq 0$ for all x, we should choose the sampling distribution to be roughly proportional to $h(x)\Pr_X(x)$. In the example illustrated above, I picked a two-component Gaussian mixture density roughly fitted to the shape of $h(x)\Pr_X(x)$. The nice thing about Gaussian sampling distributions is that it's very easy to write down $\Pr_{\tilde{X}}(x)$. We'll use this in section 5.4, where we train a neural network to find good parameters for a Gaussian sampling distribution.

density for a Gaussian mixture model: page 15

WHY IMPORTANCE SAMPLING WORKS

To justify the importance sampling approximation, let's first define

$$g(x) = h(x)\frac{\Pr_X(x)}{\Pr_{\tilde{X}}(x)}$$

and then use Monte Carlo integration to estimate $\mathbb{E}\,g(\tilde{X})$:

$$\mathbb{E}\,g(\tilde{X}) \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i), \quad x_i \text{ sampled from } \tilde{X}.$$

This gives the right hand side of the importance sampling approximation. To get the left hand side,

$$\begin{aligned}
\mathbb{E}\,g(\tilde{X}) &= \mathbb{E}_{z\sim\tilde{X}}\left\{h(z)\frac{\Pr_X(z)}{\Pr_{\tilde{X}}(z)}\right\} && \text{substituting in the definition of } g \\
&= \int_z h(z)\frac{\Pr_X(z)}{\Pr_{\tilde{X}}(z)}\Pr_{\tilde{X}}(z)\,dz && \text{writing out } \mathbb{E} \text{ as an integral (or sum)} \\
&= \int_z h(z)\Pr_X(z)\,dz && \text{since the } \Pr_{\tilde{X}}(z) \text{ terms cancel} \\
&= \mathbb{E}\,h(X) && \text{since } \mathbb{E}\,h(X) \text{ is just an integral (or sum).}
\end{aligned}$$

(This argument was cavalier about points where $\Pr_{\tilde{X}}(x) = 0$. It's easy to fix, but not illuminating.)