MTH 511a - 2020: Lecture 12

Instructor: Dootika Vats

1 Importance Sampling

1.1 Basic/simple importance sampling

1.1.1 Intuition

Recall from the last lecture that for a distribution π, and a function h, interest is in estimating

$$\theta = \int_{\mathcal{X}} h(x)\pi(x)dx\,.$$

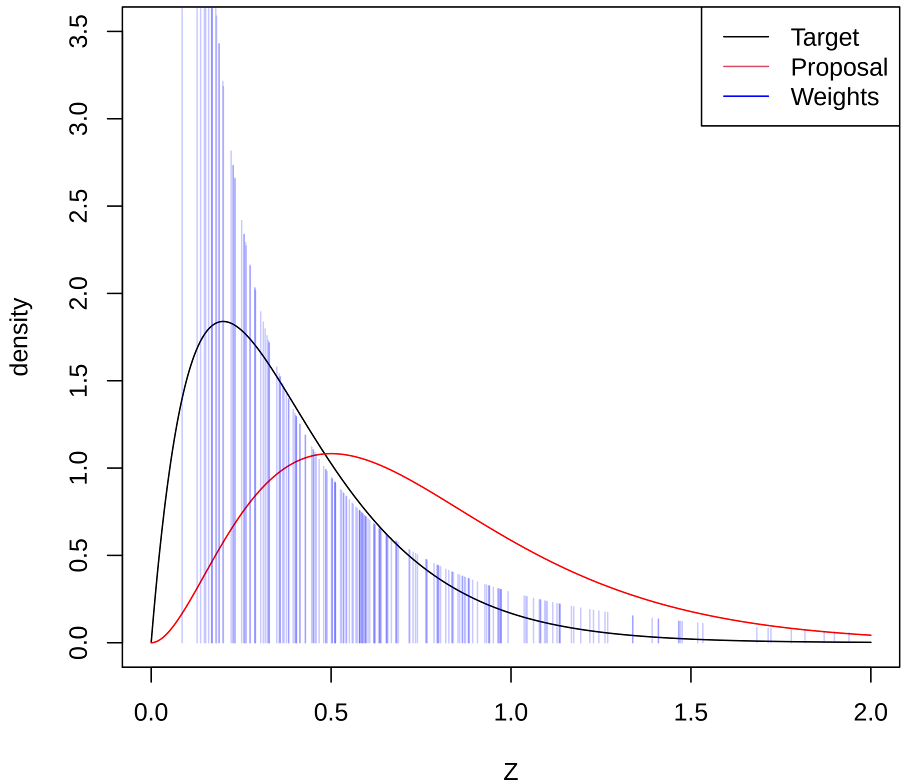
The importance sampling estimator obtains $Z_1, \dots, Z_N \sim G$, where G is a proposal distribution, then

$$\hat{\theta}_g = \frac{1}{N}\sum_{t=1}^{N} \frac{h(Z_t)\pi(Z_t)}{g(Z_t)}\,.$$

Let

$$w(Z_t) = \frac{\pi(Z_t)}{g(Z_t)}$$

when w are the weights and $\hat{\theta}_g$ is a weighted average of of $h(Z_t)$. Intuitively, this means that depending on how likely a sampled value is for π and g, a weight is assigned to that value. In the plot below, when values on the extreme left are proposed, those values are in an area of high probability for π, but unlikely to be proposed under G, thus they are assigned a large weight. Similarly, values that are likely under G but relatively less likely under π are assigned smaller weights.



1.1.2 Optimal proposals

How do we choose the importance distribution g? Note that, one reason to use importance sampling would be to obtain smaller variance estimators than the original. So, if we can choose g such that σ_g^2 is minimized that would be ideal. Let's see this term:

$$\sigma_g^2 = \mathrm{Var}_g\left(\frac{h(Z)\pi(Z)}{g(Z)}\right) = \mathrm{E}_g\left[\frac{h(Z)^2\pi(Z)^2}{g(Z)^2}\right] - \theta^2 = \underbrace{\int_{\mathcal{X}} \frac{h(z)^2\pi(z)^2}{g(z)}dz}_{A} - \theta^2$$

For the above to be small, term A should be close to θ^2.

Theorem 1. *The density g^* that minimizes σ_g^2 is*

$$g^*(z) = \frac{|h(z)|\pi(z)}{E_\pi\left[|h(x)|\right]}$$

as long as $\int |h(x)|\pi(x)dx \neq 0$.

Proof. The second moment

$$
\begin{aligned}
&\theta^2 + \sigma^2_{g^*} \\
&= \mathrm{E}_{g^*}\left[\left(\frac{h(Z)\pi(Z)}{g^*(Z)}\right)^2\right] \\
&= \int_{\mathcal{X}} \frac{h(z)^2\pi(z)^2}{g^*(z)^2} g^*(z)dz \\
&= \int_{\mathcal{X}} \frac{h(z)^2\pi(z)^2}{|h(z)|\pi(z)} \cdot \mathrm{E}_{\pi}\left[|h(x)|\right] dz \\
&= \mathrm{E}_{\pi}\left[|h(x)|\right] \int_{\mathcal{X}} |h(z)|\pi(z)dz \\
&= \left[\int_{\mathcal{X}} |h(z)|\pi(z)dz\right]^2 \\
&= \left[\int_{\mathcal{X}} \frac{|h(z)|\pi(z)}{g(z)} g(z)dz\right]^2 \quad \text{for any other } g \\
&= \left(\mathrm{E}_g\left[\frac{|h(z)|\pi(z)}{g(z)}\right]\right)^2 \\
&\leq \mathrm{E}_g\left[\frac{h(z)^2\pi(z)^2}{g^2(z)}\right] \quad \text{By Jensen's inequality: for a convex function } \phi,\ \phi(E[x]) \leq E(\phi(x)) \\
&= \theta^2 + \sigma^2_g
\end{aligned}
$$

$\Rightarrow\ \sigma^2_{g^*} \leq \sigma^2_g\,.$

Since this is true for all g, this implies that g^* produces the smallest $\sigma^2_{g^*}$ □

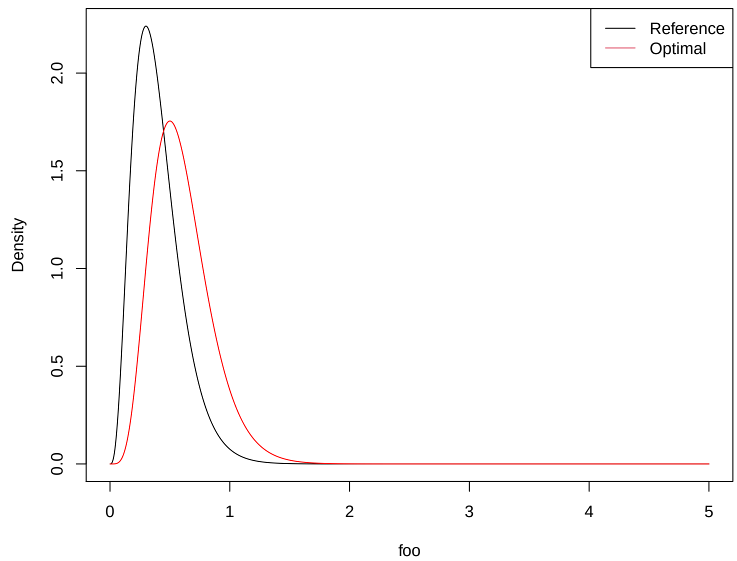
Example 1 (Gamma distribution). Consider estimating moments of a Gamma(α, β) distribution. We actually know the optimal importance distribution here! For estimating the kth moment

$$
\begin{aligned}
g^*(z) &\propto |h(z)|\pi(z) \\
&= |x^k| x^{\alpha-1} \exp\{-\beta x\} \\
&= x^{\alpha+k-1} \exp\{-\beta x\}\ .
\end{aligned}
$$

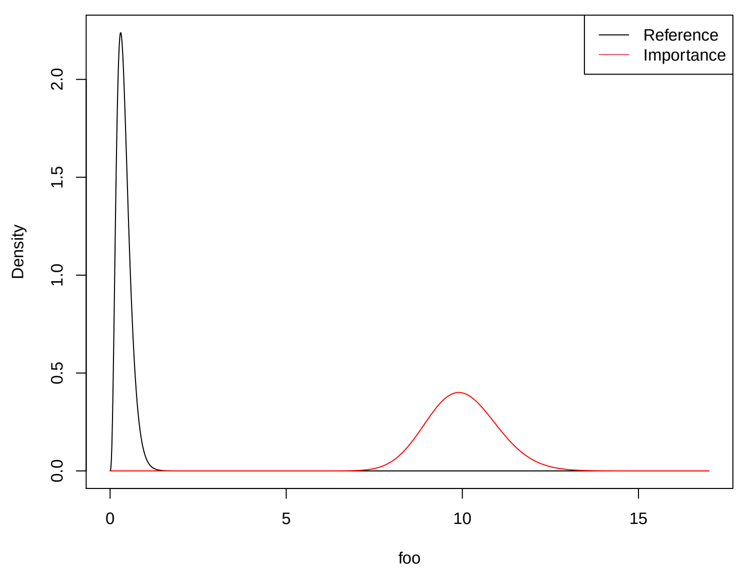
So the optimum importance distribution is Gamma$(\alpha+k, \beta)$. The variance in this case of the estimator will be quite close to 0.

```
###########################################
### Optimal importance sampling from Gamma
###########################################
set.seed(1)

# Function does importance sampling to estimate second moment of a gamma
   distribution
```

```r
imp_gamma <- function(N = 1e3, alpha = 4, beta = 10, moment = 2, imp.alpha
    = alpha + moment)
{
  fn.value <- numeric(length = N)

  draw <- rgamma(N, shape = imp.alpha, rate = beta) # draw imporance samples
  fn.value <- draw^moment * dgamma(draw, shape = alpha, rate = beta) /
      dgamma(draw, shape = imp.alpha, rate = beta)

  return(fn.value) #return all values
}

N <- 1e4
# Estimate 2nd moment from Gamma(4, 10) using Gamma(4, 10)
# this is IID Monte Carlo
imp_samp <- imp_gamma(N = N, imp.alpha = 4)
mean(imp_samp)
# [1] 0.2002069
var(imp_samp)
# [1] 0.04421469

# Estimate 2nd moment from Gamma(4, 10) using Gamma(6, 10)
# this is the optimal proposal
imp_samp <- imp_gamma(N = N)

mean(imp_samp)
# [1] 0.2
var(imp_samp)
# [1] 9.620212e-33

# why is the estimate good
foo <- seq(0.001, 5, length = 1e3)
plot(foo, dgamma(foo, shape = 4, rate = 10), type= 'l', ylab = "Density")
lines(foo, dgamma(foo, shape = 6, rate = 10), col = "red")
legend("topright", col = 1:2, lty = 1, legend = c("Reference", "Optimal"))
```




```
# Choosing a horrible proposal
# Estimate 2nd moment from Gamma(4, 10) using Gamma(100, 10)
imp_samp <- imp_gamma(N = N, imp.alpha = 100)
mean(imp_samp)  ## estimate is horrible
# [1] 1.107169e-22
var(imp_samp)
# [1] 5.679169e-41
```



Example 2 (Mean of standard normal). Let $h(x) = x$ and let $\pi(x)$ be the density of a standard normal distribution. So we are interested in estimating the mean of the standard normal distribution.

We can use classical Monte Carlo and obtain samples from π, but we will see that using importance sampling, we can get a better estimator of the mean!

Consider an importance distribution of $N(0, \sigma^2)$ for some $\sigma^2 > 0$. The variance of the importance estimator is

$$\begin{aligned}
\sigma_g^2 + \theta^2 = \sigma_g^2 & \\
&= \int_{-\infty}^{\infty} \frac{h(x)^2 \pi(x)^2}{g(x)} dx \\
&= \int_{-\infty}^{\infty} x^2 \frac{\sigma}{\sqrt{2\pi}} \exp\left\{ \frac{x^2}{2\sigma^2} - x^2 \right\} dx \\
&= \sigma \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{x^2}{2} \left(2 - \frac{1}{\sigma^2} \right) \right\} dx \\
&= \frac{\sigma}{\sqrt{2 - \sigma^{-2}}} \int_{-\infty}^{\infty} x^2 \, N(0, (2 - \sigma^{-2})^{-1}) \quad \text{if } \sigma^2 > 1/2 \\
&= \frac{\sigma}{(2 - \sigma^{-2})^{3/2}} \quad \text{if } \sigma^2 > 1/2
\end{aligned}$$

else variance is infinite. Also, minimizing the variance:

$$\arg \min_{\sigma > \sqrt{1/2}} \frac{\sigma}{(2 - \sigma^{-2})^{3/2}} = \sqrt{2}\,.$$

Thus the optimal proposal has standard deviation $\sigma = \sqrt{2}$, not 1! Also, at $\sigma^2 = 2$, the variance is .7698 which is less than 1.

1.1.3 Questions to think about

- Does this mean that $N(0, 2)$ is the optimal proposal for estimating the mean of a standard normal?
- What is the optimal proposal within the class of *Beta* proposals for estimating the mean of a Beta distribution?