MTH 511a - 2020: Lecture 11

Instructor: Dootika Vats

We have so far learned many (many!) ways of sampling from different distributions. As motivated in the first week of the course, these sampling methodologies are useful in Monte Carlo estimation problems.

Suppose π is a distribution with density π (this is an abuse of notation that is commonly made in sampling literature). We are interested in estimating the expectation of a function $h : \mathcal{X} \to \mathbb{R}$ with respect to π. That is, we want to estimate

$$\theta := \mathrm{E}_\pi[h(X)] = \int_{\mathcal{X}} h(x)\pi(x)\, dx\,,$$

we assume that θ is finite.

Note: there is no data here, there is just an integral!

Note: notation $E_\pi[X]$ means the expectation is with respect to π. From now on, it is very important to keep track of what the expectation is with respect to.

Suppose we can draw iid samples $X_1, \dots, X_N \overset{iid}{\sim} \pi(x)$ (this we can do using the many methods we have learned). Then define the estimator:

$$\hat{\theta} = \frac{1}{N}\sum_{t=1}^{N} h(X_t)\,.$$

By the law of large numbers we know that as $N \to \infty$

$$\hat{\theta} \xrightarrow{p} \theta\,.$$

In addition, we can find the variance of the estimator:

$$\mathrm{Var}_\pi(\hat{\theta}) = \mathrm{Var}_\pi\left(\frac{1}{N}\sum_{t=1}^{N} h(X_t)\right)$$

$$
\begin{aligned}
&= \frac{1}{N^2}\sum_{t=1}^{N} \mathrm{Var}_\pi(h(X_t)) \quad \text{because of independence} \\
&= \frac{\mathrm{Var}_\pi(h(X_1))}{N} \quad \text{because of identical}\,.
\end{aligned}
$$

Naturally, a CLT holds if $\mathrm{Var}_\pi(h(X_1)) < \infty$.

Q. But is there a way we can obtain a better estimator of θ?

A. Possibly by using importance sampling.

1 Importance Sampling

1.1 Basic/simple importance sampling

Let G be a distribution with density g defined on $\mathcal{X}$ so that,

$$
\begin{aligned}
\mathrm{E}_\pi[h(X)] &= \int_{\mathcal{X}} h(x)\pi(x)dx \\
&= \int_{\mathcal{X}} \frac{h(x)\pi(x)}{g(x)} g(x)\, dx \\
&= \mathrm{E}_g\left[\frac{h(Z)\pi(Z)}{g(Z)}\right], \qquad Z \sim G
\end{aligned}
$$

If $Z_1, \dots, Z_N$ are iid samples from G, then an estimator of θ is

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

The estimator $\hat{\theta}_g$ is the *importance sampling estimator*, the method is called *importance sampling* and G is the *importance distribution.*

Example 1 (Moments of Gamma distribution). Suppose we want to estimate the kth moment of a Gamma distribution.That is, let π be the density of a Gamma(α, β) distribution. Then

$$
\theta = \int_0^\infty x^k \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx\,.
$$

Suppose we set G to be also an Exponential(λ) distribution. Let $Z_1, \dots, Z_n \sim \mathrm{Exp}(\lambda)$

$$
\hat{\theta}_g = \frac{1}{N}\sum_{t=1}^{N} \left[\frac{h(Z_t)\pi(Z_t)}{g(Z_t)}\right]
$$

$$= \frac{1}{N} \sum_{t=1}^{N} \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{Z_t^k Z_t^{\alpha-1} e^{-\beta Z_t}}{\lambda e^{-\lambda Z_t}} \right]$$

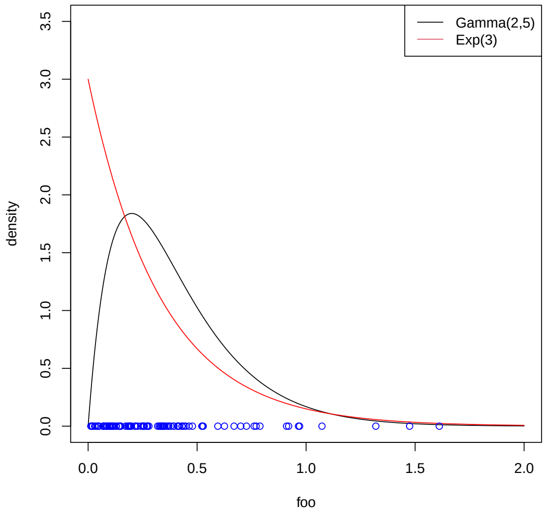
In the below simulation, I set $\alpha = 2, \beta = 5$ and set $\lambda = 3$.

```
set.seed(1)

alpha <- 2
beta <- 5

k <- 2 # second moment
(truth <- (alpha / beta^2) + (alpha/beta)^2) # true second moment
#[1] 0.24

lambda <- 3 #proposal

N <- 1e4
samp <- rexp(N, rate = lambda) # importance samples
func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp, rate
   = lambda)
mean(func) # truth is .24
#[1] 0.2385123
```

Our estimate is fairly close to the truth!

We can also visually compare the reference density π and the importance density g. Below, I also plot the first 100 draws from G.

```
foo <- seq(0, 2, length = 1e3)
plot(foo, dgamma(foo, shape = alpha, rate = beta),
  type = ’l’, col = "black", ylab = "density", ylim = c(0, 5))
lines(foo, dexp(foo, rate = lambda), col = "red")
points(x = samp[1:100],y = rep(0, 100), col = "blue")
legend("topright", legend = c("Gamma(2,10)", "Exp(5)"), col = c(1,2), lty =
   1)
```



Theorem 1 (Unbiasedness). *The importance sampling estimator $\hat{\theta}_g$ is unbiased for θ.*

Proof. To show an estimator is unbiased, we need to show that $\mathrm{E}_g[\hat{\theta}_g] = \theta$. Consider

$$\begin{aligned}
\mathrm{E}_g\left[\hat{\theta}_g\right] &= \mathrm{E}_g\left[\frac{1}{N}\sum_{t=1}^{N}\frac{h(Z_t)\pi(Z_t)}{g(Z_t)}\right] \\
&= \frac{1}{N}\sum_{t=1}^{N}\mathrm{E}_g\left[\frac{h(Z_t)\pi(Z_t)}{g(Z_t)}\right] \\
&= \frac{1}{N}\sum_{t=1}^{N}\mathrm{E}_g\left[\frac{h(Z_1)\pi(Z_1)}{g(Z_1)}\right] \\
&= \int_{\mathcal{X}}\frac{h(z)\pi(z)}{g(z)}g(z)\,dz \\
&= \int_{\mathcal{X}} h(z)\pi(z)dz \\
&= \theta\,.
\end{aligned}$$

□

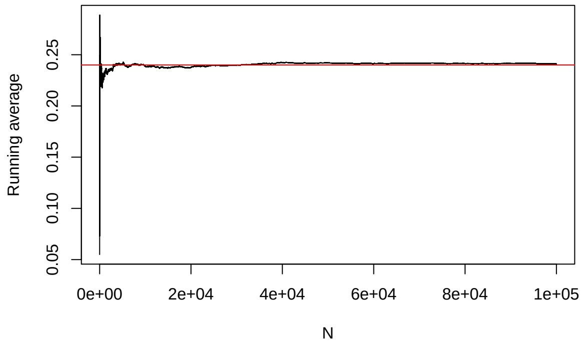
By the law of large numbers, as $N \to \infty$,

$$\hat{\theta}_g \xrightarrow{p} \mathrm{E}[\hat{\theta}_g] = \theta\,.$$

This means that as we get more and more samples from G, our estimator will get increasingly closer to the truth.

Example 2 (Gamma continued...). We can try to "verify" convergence by checking what happens as $N \to \infty$ in one simulation.

```
## Checking convergence
N <- 1e5 # very large N
samp <- rexp(N, rate = lambda) # importance samples
func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp, rate
    = lambda)

# Plotting the running average
plot(1:N, cumsum(func)/(1:N), type = 'l', xlab = "N", ylab = "Running
    average")
abline(h = truth, col = "red")
```



We will also try to "verify" via simulation that $\hat{\theta}_g$ obtained before is indeed unbiased. By definition of unbiasedness, $\text{E}_g[\hat{\theta}_g - \theta] = 0$. Thus, to mimic this in simulation, we will repeat the simulation *multiple times* (r times) so that we obtain

$$\hat{\theta}_g^1, \hat{\theta}_g^2, \dots, \hat{\theta}_g^r \,.$$

Then by definition, if r is large:

$$\text{Diff}_r = \frac{1}{r}\sum_{k=1}^{r}(\hat{\theta}_g^k - \theta) \approx \text{E}_g[\hat{\theta}_g - \theta] \,.$$

Thus, if $\text{Diff}_r \approx 0$, then we know that the procedure is likely unbiased.

```
## Checking if unbiased or not
N <- 1e4
r <- 1e3
ests <- numeric(length = r)
for(a in 1:r)
{
  samp <- rexp(N, rate = lambda) # importance samples
  func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp,
      rate = lambda)
  ests[a] <- mean(func)
}
mean(ests - truth) # very close to 0
[1] 2.700126e-05
```

However, we should never be happy with a point estimator! It is essential to quantify the variability in our estimator $\hat{\theta}_g$ in order to ascertain how "erratic" or "stable" the estimator. We also want to make confidence intervals around $\hat{\theta}_g$, but *does a central limit theorem hold?*. Note that, the variance of $\hat{\theta}_g$ is

$$\text{Var}_g(\hat{\theta}_g) = \text{Var}_g\left(\frac{1}{N}\sum_{t=1}^{N}\frac{h(Z_t)\pi(Z_t)}{g(Z_t)}\right) = \frac{1}{N}\text{Var}_g\left(\frac{h(Z_1)\pi(Z_1)}{g(Z_1)}\right) := \frac{\sigma_g^2}{N}\,.$$

A central limit theorem will hold if $\text{Var}_g\left(\dfrac{h(Z_1)\pi(Z_1)}{g(Z_1)}\right) < \infty$.

So the question is, when is this finite? The following provides a sufficient condition.

Theorem 2. *Suppose $Var_\pi(h(X)) < \infty$. If g is chosen such that*

$$\sup_{z\in\mathcal{X}}\frac{\pi(z)}{g(z)} \leq M < \infty$$

then

$$\sigma_g^2 < \infty\,.$$

Proof. First note that if the second moment is finite, then the variance is finite. So, consider the second moment of $\frac{h(Z)\pi(Z)}{g(Z)}$ where $Z \sim g$.

$$\begin{aligned}
\text{E}_g\left[\left(\frac{h(Z)\pi(Z)}{g(Z)}\right)^2\right] &= \int_{\mathcal{X}}\frac{h(z)^2\pi(z)^2}{g(z)^2}g(z)dz \\
&= \int_{\mathcal{X}} h(z)^2\,\frac{\pi(z)}{g(z)}\,\pi(z)dz \\
&\leq M\int_{\mathcal{X}} h(z)^2\pi(z)dz \\
&= M\,\text{E}_\pi(h(X)^2) < \infty \quad \text{by assumption}\,.
\end{aligned}$$

□

Thus, if an accept-reject on the same support is possible, the variance of the importance sampling method estimator is finite. Now, we have a central limit theorem that can hold. Recall

$$\sigma_g^2 = \text{Var}_g\left(\frac{h(Z)\pi(Z)}{g(Z)}\right)\,. \tag{1}$$

If $\sigma_g^2 < \infty$, then as $N \to \infty$,

$$\sqrt{n}(\hat{\theta}_g - \theta) \xrightarrow{d} N(0, \sigma_g^2)\,. \tag{2}$$

Example 3 (Gamma continued). We can and "verify" again via simulation if σ_g^2 is finite. Of course, we will not be able to verify this exactly, but we can guess, what's going on.

Recall from lecture 9 that for Gamma(α, β) with $\alpha > 1$ and exponential proposal for an accept-reject sampler will work only if $\lambda < \beta$. We chose $\lambda = 3$ and $\beta = 5$, thus our running example produces finite variance estimator of θ. We can "verify" by using our replications

$$\hat{\theta}_g^1, \hat{\theta}_g^2, \dots, \hat{\theta}_g^r$$

and noting that the sample variance of these r estimates should be σ_g^2/N. That is

$$\frac{1}{r-1}\sum_{k=1}^{r}(\hat{\theta}_g^k - \text{mean}(\hat{\theta}_g^{1:r})) \approx \frac{\sigma_g^2}{N}\,.$$

```
# looking at variance
var(ests) # This is var(theta_g) = sigma^2_g/N
# [1] 1.203824e-05

N*var(ests) # pretty small
# [1] 0.1203824
```

Now let's change things up, let's see what happens when we set $\lambda = 10$! For this setting, the accept-reject does not work, since

$$\sup_z \frac{\pi(z)}{g(z)} = \infty$$

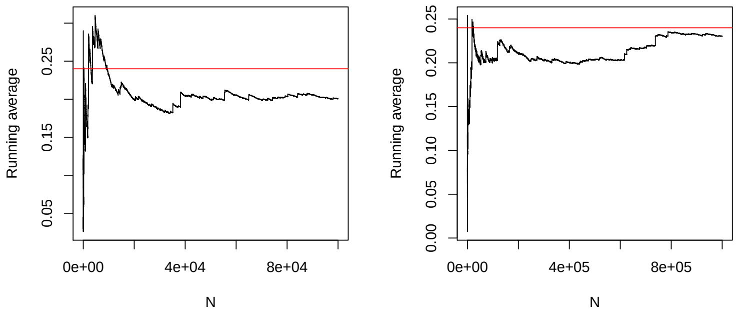
This does not imply that $\sigma_g^2 = \infty$ since the above theorem only provides a necessary condition. Nonetheless,

```
## When Accept-reject fails
# If lambda > beta, we know accept-reject fails, let's see what the
   variance is then
lambda <- 10
N <- 1e4
r <- 1e3
ests <- numeric(length = r)
for(a in 1:r)
{
  samp <- rexp(N, rate = lambda) # importance samples
  func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp,
     rate = lambda)
  ests[a] <- mean(func)
}
mean(ests - truth) # close to 0
# [1] -0.01287952

var(ests) # This is var(theta_g) = sigma^2_g/N
#[1] 0.02645068

N*var(ests) # Variance is much larger now!
# [1] 264.5068
```

We can see that the variance blows up! This means that our estimator cannot be trusted from one simulation to another!

Moreover, in this example the convergence plots can be analyzed as well. Note that, to convergence in the law of large numbers, we do not require a finite variance, so convergence will still occur. However, the sample size required to get close will be much large and for finite N, residual variability will remain.

```
## Checking convergence again
# Convergence is not affected, but it takes MUCH longer
# to get good convergence
par(mfrow = c(1,2))
N <- 1e5 # very large N
samp <- rexp(N, rate = lambda) # importance samples
func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp, rate
    = lambda)

# Plotting the running average
plot(1:N, cumsum(func)/(1:N), type = ’l’, xlab = "N", ylab = "Running
    average")
abline(h = truth, col = "red")


N <- 1e6 # very large N
samp <- rexp(N, rate = lambda) # importance samples
func <- samp^k * dgamma(samp, shape = alpha, rate = beta) / dexp(samp, rate
    = lambda)

# Plotting the running average
plot(1:N, cumsum(func)/(1:N), type = ’l’, xlab = "N", ylab = "Running
    average")
abline(h = truth, col = "red")
```



2 Questions to think about

1. Check what happens with $\beta = \lambda$ in this simulation.
2. Why would a CLT be useful here?
3. How would we check whether this importance sampler is better than IID Monte Carlo?