STATS 406 F15: Lab 07
Two forms of importance sampling

Post-lab reading assignment: Art Owen's lecture note on importance sampling.
`http://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf`
Part of this note is a user-friendly version of Professor Owen's note.

1 Recap: two forms of importance sampling

Problem of interest: (slightly different from last time)
We want to compute:

$$I := \mathbb{E}_p[f(\boldsymbol{X})] \tag{1}$$

where $\boldsymbol{X} \overset{\text{pdf}}{\sim} p(\boldsymbol{x})$. Recall that we did:

$$I = \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int \frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})} \cdot q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_q\left[\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] \tag{2}$$

NOTICE the transition from $\mathbb{E}_p$ in (1) to $\mathbb{E}_q$ in (2).

1. **Importance sampling (primary form):**

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} \tag{3}$$

   where $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q(\boldsymbol{x})$.

2. **A variant of importance sampling (self-normalized form):**

$$\tilde{I} = \frac{\sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)p_0(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)}}{\sum_{i=1}^{n} \frac{p_0(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)}} \tag{4}$$

   where $p_0(\boldsymbol{x}) \propto p(\boldsymbol{x})$ with a potentially unknown constant and $\boldsymbol{X}_i \overset{\text{iid}}{\sim} q(\boldsymbol{x})$.

Remarks:

1. (Last time) We showed the asymptotic unbiasedness of $\tilde{I}$ defined by (4), but it is almost always **biased** in finite samples.

2. The case where we must use (4): when the normalizing constant in $p(\boldsymbol{x})$ is unknown.

3. (This class) When we know $p(\boldsymbol{x})$ exactly, there comes the question of choosing between (3) and (4).

2 Known $p(\boldsymbol{x})$: primary form vs. self-normalized form

Recall that $\hat{I}$ is always unbiased, while $\tilde{I}$ is usually biased, so $\hat{I}$ is superior in this sense. On the other hand, $\tilde{I}$ is usually a variance-reduction version of $\hat{I}$. Next, we compare their variances.

2.1 Asymptotic variance of $\hat{I}$

Very simple:

$$n \cdot \mathrm{Var}[\hat{I}] = \mathrm{Var}\left(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right) = \mathbb{E}\left[\frac{f^2(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - I^2 \tag{5}$$

2.2 Asymptotic variance of $\tilde{I}$

Not so simple:

$$n \cdot \mathrm{Var}[\tilde{I}] = \mathbb{E}\left[(f(\boldsymbol{X}) - I)^2 \left(\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right)\right] \tag{6}$$

2.3 Comparison

Now, (5) - (6) gives that $\mathrm{Var}[\tilde{I}] \leq \mathrm{Var}[\hat{I}]$ if and only if

$$I\left(\mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] + 1\right) \leq 2\mathbb{E}\left[\frac{f(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] \tag{7}$$

Remarks:

1. In general, it is difficult to analytically calculate the terms in (7).

2. **Special case 1:** If

   $$f(\boldsymbol{x}) \equiv \text{Constant}$$

   Then by definition, I must equal to this constant, thus $\mathrm{Var}[\tilde{I}] = 0$. Of course $\mathrm{Var}[\tilde{I}] \leq \mathrm{Var}[\hat{I}]$, which can also be seen from (7).

3. **Special case 2:** Since

$$1 = \mathbb{E}\left[\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right]^2 \leq \mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right]$$

A sufficient condition for (7) to hold is

$$I \cdot \mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] \leq \mathbb{E}\left[\frac{f(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] \tag{8}$$

If we "revive" some shorted or omitted terms in (8), we can rewrite (8) as follows

$$\mathbb{E}\left[\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] \cdot \mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] \leq \mathbb{E}\left[\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right] \cdot \mathbb{E}\left[\frac{f(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] \tag{9}$$

The message in (9) is very clear: if $q(\boldsymbol{x})$ is chosen poorly such that $p(\boldsymbol{x})/q(\boldsymbol{x})$ is large where $f(\boldsymbol{x})$ is also large, then $(p(\boldsymbol{x})/q(\boldsymbol{x}))$ will enlarge such large values of $f(\boldsymbol{x})$ much more than $p(\boldsymbol{x})/q(\boldsymbol{x})$, and as a result (9) will hold.

2.4 In-lab exercise

For any vector $\boldsymbol{x} \in \mathbb{R}^4$, define

$$f(\boldsymbol{x}) := \mathbb{1}[\|x\|_2 \leq 1] \cdot a \cdot e^{-\frac{\|\boldsymbol{x}\|_2^2}{2\sigma^2}}$$

When you program this, first test with the configuration $a = 10$ and $\sigma = 1$. Use three methods(naive Monte Carlo and the two forms of importance sampling) to compute

$$I := \mathbb{E}[f(\boldsymbol{X})]$$

where $\boldsymbol{X}$ follows a uniform distribution on the 4-dimensional unit ball $\{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^4, \|\boldsymbol{x}\|_2 \leq 1\}$. For your chosen sample size n, compare the mean and the variance of different methods.

Hints:

- Let $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} N(0,1)$ random variables, then $Z_1^2 + \ldots + Z_k^2 \sim \mathcal{X}_k^2$ – a Chi-squared distribution with k degrees of freedom. In R, use

  ```
  pchisq(t, df);
  ```

  to access its CDF, where t is the threshold and df is the degrees of freedom.

- The volume of 4-dimensional ball with radius R is $\pi^2 R^4/2$.

```
# Solution: see Lab_7.R (will post later)
```

2.5 Proof of (6)

We need to introduce the delta method:

Delta method: Recall that by CLT, for X with finite second moment, letting $\mu := \mathbb{E}[X]$, we have

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

then for a smooth function $g(t)$, we have

$$\sqrt{n}\left(g(\bar{X}) - g(\mu)\right) \xrightarrow{d} N\left(0, [g'(\mu)]^2 \cdot \sigma^2\right) \tag{10}$$

Proof of the delta method. By Talor expansion:

$$g(\bar{X}) = g(\mu) + (\bar{X} - \mu)g'(\mu) + O\left((\bar{X} - \mu)^2\right)$$

That is,

$$\sqrt{n}\left(g(\bar{X}) - g(\mu)\right) = \sqrt{n}(\bar{X} - \mu)g'(\mu) + \sqrt{n} \cdot O\left((\bar{X} - \mu)^2\right)$$

by CLT, $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N\left(0, \sigma^2\right)$ so

$$\sqrt{n}(\bar{X} - \mu)g'(\mu) \xrightarrow{d} N\left(0, [g'(\mu)]^2 \cdot \sigma^2\right)$$

and $n(\bar{X} - \mu)^2 \xrightarrow{d} O(1)$, which yields

$$(\bar{X} - \mu)^2 \xrightarrow{d} O(n^{-1})$$

□

The multivariate version of the delta method is also true(the proof is similar and omitted). Specifically, we use the 2-dimensional case, let $\mu := \mathbb{E}[Y]$ and $\lambda := \mathbb{E}[Z]$. By CLT, we have:

$$\sqrt{n}\begin{pmatrix}\bar{Y} - \mu \\ \bar{Z} - \lambda\end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{pmatrix}\sigma_{yy} & \sigma_{yz} \\ \sigma_{yz} & \sigma_{zz}\end{pmatrix}\right)$$

Define $g(s, t) := s/t$ and apply the delta method:

$$\sqrt{n}\left(g(\bar{Y}, \bar{Z}) - g(\mu, \lambda)\right) \xrightarrow{d} N\left(0, \left(\frac{\partial g}{\partial s}(\mu, \lambda), \frac{\partial g}{\partial t}(\mu, \lambda)\right)\begin{pmatrix}\sigma_{yy} & \sigma_{yz} \\ \sigma_{yz} & \sigma_{zz}\end{pmatrix}\begin{pmatrix}\frac{\partial g}{\partial s}(\mu, \lambda) \\ \frac{\partial g}{\partial t}(\mu, \lambda)\end{pmatrix}\right)$$

Since

$$\frac{\partial g}{\partial s} = \frac{1}{t}$$
$$\frac{\partial g}{\partial t} = -\frac{s}{t^2}$$

we have

$$\begin{aligned}
&\left(\frac{\partial g}{\partial s}(\mu,\lambda),\frac{\partial g}{\partial t}(\mu,\lambda)\right)\begin{pmatrix}\sigma_{yy} & \sigma_{yz}\\ \sigma_{yz} & \sigma_{zz}\end{pmatrix}\begin{pmatrix}\frac{\partial g}{\partial s}(\mu,\lambda)\\ \frac{\partial g}{\partial t}(\mu,\lambda)\end{pmatrix}\\
=&\left(\frac{\partial g}{\partial s}(\mu,\lambda)\right)^2\sigma_{yy}+2\frac{\partial g}{\partial s}(\mu,\lambda)\frac{\partial g}{\partial t}(\mu,\lambda)\sigma_{yz}+\left(\frac{\partial g}{\partial t}(\mu,\lambda)\right)^2\sigma_{zz}\\
=&\frac{1}{\lambda^2}\sigma_{yy}-\frac{2\mu}{\lambda^3}\sigma_{yz}+\frac{\mu^2}{\lambda^4}\sigma_{zz} \qquad (11)
\end{aligned}$$

Replace Y by $\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}$ and Z by $\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}$, we see that $\mu = I$ and $\lambda = 1$. Also

$$\begin{aligned}
\sigma_{yy} &= \mathrm{Var}\left(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right) = \mathbb{E}\left[\frac{f^2(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - I^2\\
\sigma_{yz} &= \mathrm{Cov}\left(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})},\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right) = \mathbb{E}\left[\frac{f(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - I\\
\sigma_{zz} &= \mathrm{Var}\left(\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}\right) = \mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - 1^2
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\text{RHS of (11)} = \sigma_{yy} - 2I\cdot\sigma_{yz} + I^2\cdot\sigma_{zz}\\
=&\left(\mathbb{E}\left[\frac{f^2(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - I^2\right) - 2I\cdot\left(\mathbb{E}\left[\frac{f(\boldsymbol{X})p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - I\right) + I^2\cdot\left(\mathbb{E}\left[\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] - 1\right)\\
=&\mathbb{E}\left[\frac{\left(f^(\boldsymbol{X}) - 2I\cdot f(\boldsymbol{X}) + I^2\right)p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right] = \mathbb{E}\left[(f(\boldsymbol{X})-I)^2\left(\frac{p^2(\boldsymbol{X})}{q^2(\boldsymbol{X})}\right)\right]
\end{aligned}$$

The proof is complete.

3 Importance sampling diagnostics: effective sample size(ESS) and the rule of thumb

3.1 Motivation of introducing diagnostic tools for importance sampling

- In practice, researchers can often face very complicated or peculiar $f(\boldsymbol{x})p(\boldsymbol{x})$ such that it is too difficult to choose a $q(\boldsymbol{x})$ we can sample from that has a similar shape.
- As a result, we choose a $q(\boldsymbol{x})$ that is only capable to respect high value regions of $f(\boldsymbol{x})p(\boldsymbol{x})$ but cannot accurately mimic its shape.
- We need to (intuitively and very roughly) assess if such $q(\boldsymbol{x})$ is a decent choice.
- One way is to judge based on the sample variance, but it is not unit-free. In many cases, it is just difficult to know an observed value of the sample variance is "large" or "small"!

3.2 The concept of ESS

Weighted sample mean: For an iid sample $\{X_1, \ldots, X_n\}$ and a set of constant weights $w_1, \ldots, w_n$, consider the weighted sample mean:

$$\tilde{X} := \frac{w_1 X_1 + \ldots + w_n X_n}{w_1 + \ldots + w_n} \tag{12}$$

and compare it to the ordinary (unweighted) sample mean:

$$\hat{X} := \frac{1}{n}(X_1 + \ldots + X_n)$$

Now

$$\mathrm{Var}(\tilde{X}) = \frac{w_1^2 + \ldots + w_n^2}{(w_1 + \ldots + w_n)^2} \cdot \mathrm{Var}(X_i) \geq \frac{1}{n} \cdot \mathrm{Var}(X_i) = \mathrm{Var}(\hat{X}) \quad \text{(Why?)}$$

In the sense of variance reduction, we have a smaller sample size than the nominal sample size n. In other words, the variance suppression effect of $\tilde{X}$ is equally powerful to the ordinary sample mean of a sample at size

$$n_e := \frac{n\mathrm{Var}(\tilde{X})}{\mathrm{Var}(\hat{X})} \tag{13}$$

and it is easy to see that here

$$n_e = \frac{(w_1 + \ldots + w_n)^2}{w_1^2 + \ldots + w_n^2} \leq n \tag{14}$$

where equality only holds for the equal w_i case. Here, n_e is called the effective sample size(ESS). This n_e equals the one you saw in lecture. That is,

$$n_e = \frac{n}{1 + \mathrm{cv}(w)^2} \quad \text{(Verify this.)} \tag{15}$$

where the coefficient of variation is

$$\mathrm{cv}(w) := \left(\sum_{i=1}^{n} (w_i - \bar{w})^2 / n \right)^{1/2} \Big/ \bar{w}$$

Here, notice that I replaced the division of $n-1$ in most textbooks with n inside the square root to make the equality hold – this doesn't matter asymptotically.

3.3 The plug-in version of ESS in importance sampling

- **ESS-IS as a plug-in version of** (14)**:** Now let's derive the plug-in version of ESS for (4) by replacing X_i with $f(X_i)$ and w_i with $p_0(\boldsymbol{X}_i)/q(\boldsymbol{X}_i)$. We obtain the plug-in

version of (14) as follows:

$$n_e^{IS} := \frac{\left(\frac{p(\boldsymbol{X}_1)}{q(\boldsymbol{X}_1)} + \ldots + \frac{p(\boldsymbol{X}_n)}{q(\boldsymbol{X}_n)}\right)^2}{\left(\frac{p(\boldsymbol{X}_1)}{q(\boldsymbol{X}_1)}\right)^2 + \ldots + \left(\frac{p(\boldsymbol{X}_n)}{q(\boldsymbol{X}_n)}\right)^2} \tag{16}$$

The quantity n_e^{IS} defined by (16) is very straightforward to compute.

- If we want to derive a similar ESS-IS for (3), simply vary (12) by throwing away its demoninator, and derive an analogous version of (16). The derivation is exactly similar and omitted here.
- The ESS-IS (16) has a few essential differences from its original form:
  1. In ESS-IS (16), weights w_i's are now functions of X_i's and are no longer constants.
  2. ESS-IS only borrows the form of ESS. It is not even clear which two variances ESS-IS is comparing! On the other hand, however, a partial justification for ESS-IS states that under mild conditions, ESS-IS is a decent approximation for the ESS defined as (13), where the baseline variance is the variance of

$$\hat{I}_0 := \frac{1}{n}\{f(\boldsymbol{X}_1) + \ldots + f(\boldsymbol{X}_n)\}$$

     where $\boldsymbol{X} \overset{\text{pdf}}{\sim} p(\boldsymbol{x})$, compared to the variance of $\hat{I}$ defined in (3). The proof is highly technical. See Jun Liu's book (Section 2.5.3, page 35) for details.
  3. Another important fact that helps you realize that ESS-IS only borrows the form of ESS is that ESS-IS does not depend on $f(\boldsymbol{x})$. This is unimaginable for a rigorously defined ESS as that in (13).
- Rule of thumb: if n_e^{ISS} is not too small (now we can compare it to n), we think the choice of $q(\boldsymbol{x})$ is fine.