Monte Carlo Integration and Importance Sampling

Adam Smith, Spring 2017

# 1 Monte Carlo Integration

Consider the problem of evaluating the expected value of a function of random variables $h(\theta)$ with respect to some density $f(\theta)$.

$$E^f[h(\theta)] = \int_\Theta h(\theta)f(\theta)d\theta \tag{1}$$

If $h(\theta) = \theta$ and $f(\theta) = \pi(\theta|y)$, for example, then $E^f[h(\theta)]$ is the posterior mean. Rather than trying to solve these potentially high dimensional or intractable integrals analytically, we will instead rely on computational techniques.

If $\theta^{(1)}, \dots, \theta^{(R)}$ are iid draws from $f(\theta)$, we can approximate (1) with the empirical average.

$$\bar{h}_R = \frac{1}{R}\sum_{r=1}^{R} h(\theta^{(r)}) \tag{2}$$

Here the Strong Law of Large Numbers guarantees that $\bar{h}_R$ is a good estimator, in that it will converge to the truth as $R$ gets large.

$$\bar{h}_R \xrightarrow{a.s.} E^f[h(\theta)] \quad \text{as} \quad R \to \infty \tag{3}$$

The process of approximating (1) with an empirical average is called Monte Carlo integration.

**Example 1.** Let $\theta \sim N(1, 5)$ and suppose we want to evaluate the third moment of a Normal distribution.

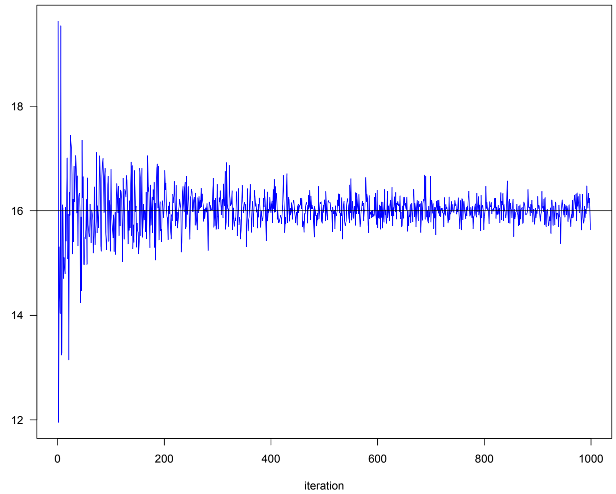
$$E^f[\theta^3] = \int_\Theta \theta^3 f(\theta)d\theta$$

This integral can be solved analytically. That is, by Stein's identity, we have

$$E^f[\theta^3] = 3\mu\sigma^2 + \mu^3 = 3(1)(5) + 1^3 = 16.$$

However, we can also use Monte Carlo integration to approximate this integral and assess its large-sample properties. To do this, we draw $\theta^{(r)} \sim N(1, 5)$ for $r = 1, \dots, R$ and various values of $R$ and then compute the empirical average in (2).

| $R$ | $\bar{h}_R$ |
|---|---|
| 100 | 11.5362 |
| 500 | 14.0838 |
| 1,000 | 14.8477 |
| 10,000 | 15.3902 |
| 100,000 | 15.9922 |




## 2 Importance Sampling

To implement Monte Carlo integration as stated above, we must be able to generate iid draws from $f(\theta)$. In many Bayesian problems, $f(\theta)$ is a posterior distribution which may not belong to a known class of distributions. In cases like these, generating iid draws will be hard if not impossible.

Importance sampling provides a solution to this problem. Suppose we are *unable* to generate iid draws from $f(\theta)$ but we can generate iid draws from a different distribution $g(\theta)$. First notice that we can rewrite (1) as an expectation with respect to $g(\theta)$.

$$E^f[h(\theta)] = \int_\Theta h(\theta)f(\theta)d\theta = \int_\Theta h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = E^g\left[h(\theta)\frac{f(\theta)}{g(\theta)}\right] \tag{4}$$

This new representation of $E^f[h(\theta)]$ will be valid as long as the support of $g$ is at least as big as the support of $f$. That is, we must choose $g$ so that the ratio $f(\theta)/g(\theta)$ is always finite.

Given $\theta^{(1)}, \ldots, \theta^{(R)}$ iid draws from $g(\theta)$, the importance sampling estimate of $E^f[h(\theta)]$ can be written as

$$\bar{h}_{IS,R} = \frac{1}{R}\sum_{r=1}^{R} h(\theta^{(r)})\frac{f(\theta^{(r)})}{g(\theta^{(r)})}. \tag{5}$$

**Example 2** (Binary Probit)**.** Consider the binary probit model specified as follows.

$$\begin{aligned} y &= I(z > 0) \\ z &= X\beta + \varepsilon,\ \ \varepsilon \sim N(0,1) \end{aligned} \tag{6}$$

Since $y$ is either zero or one, $y$ will have a Bernoulli distribution with probability

$$\begin{aligned} \Pr(y = 1) &= \Pr(z > 0) \\ &= \Pr(X\beta + \varepsilon > 0) \\ &= \Pr(\varepsilon > -X\beta) \\ &= 1 - \Pr(\varepsilon < -X\beta) \\ &= 1 - \Phi(-X\beta) \\ &= \Phi(X\beta) \ \ \text{by symmetry of the N(0,1)} \end{aligned} \tag{7}$$

where $\Phi(\cdot)$ is the standard normal CDF. We can formally write the likelihood function as

$$p(y|X,\beta) = \prod_{i=1}^{n} \Phi(X_i\beta)^{y_i}(1 - \Phi(X_i\beta))^{(1-y_i)}. \tag{8}$$

Finally, we will assume that $\beta$ has a $N(\bar{\beta}, A^{-1})$ prior. The posterior of $\beta$ has the form

$$\pi(\beta|y,X) = \frac{p(y|X,\beta)p(\beta)}{\int p(y|X,\beta)p(\beta)d\beta} = \frac{p(y|X,\beta)p(\beta)}{m(y)}. \tag{9}$$

Now suppose we want to find the posterior mean of $\beta$.

$$E^{\pi}[\beta] = \int \beta\pi(\beta|y,X)d\beta \tag{10}$$

If we could generate iid draws from $\pi(\beta|y,X)$, then we could simply use Monte Carlo integration to estimate $E^{\pi}[\beta]$. However, our Normal prior for $\beta$ is not conjugate to the Bernoulli likelihood, so iid sampling from $\pi(\beta|y,X)$ is not possible.

To use importance sampling, we must first pick a suitable distribution $g$. In this case, let

$$g(\beta) = MSt_{\nu}(\hat{\beta}_{MLE}, s \cdot (-H|_{\beta=\hat{\beta}_{MLE}})^{-1}) \tag{11}$$

which serves as a thick-tailed asymptotic approximation to $\pi(\beta|y,X)$. Here $\nu$ is a degrees of freedom parameter, $\hat{\beta}_{MLE}$ is the maximum likelihood estimate of $\beta$, $s$ is a scaling parameter, and $H$ is the Hessian matrix of the log-likelihood evaluated at $\hat{\beta}_{MLE}$. After sampling $\beta^{(1)}, \ldots, \beta^{(R)}$ from $g(\beta)$, we can use (5) to estimate the posterior mean.

$$\bar{h}_{IS,R} = \frac{1}{R}\sum_{r=1}^{R} \beta^{(r)} \frac{\pi(\beta^{(r)}|y)}{g(\beta^{(r)})}. \tag{12}$$

Now we have replaced the requirement of sampling from $\pi(\beta|y, X)$ with the requirement of evaluating $\pi(\beta|y, X)$ at various points. There is still one problem: given the lack of conjugacy, evaluating $\pi(\beta|y, X)$ is not possible as we only know how to evaluate the *unnormalized* posterior.

$$\bar{h}_{IS,R} = \frac{1}{R}\sum_{r=1}^{R} \beta^{(r)} \frac{p(y|X, \beta^{(r)})p(\beta^{(r)})}{m(y)g(\beta^{(r)})} = \frac{\frac{1}{R}\sum_{r=1}^{R} \beta^{(r)} \frac{p(y|X,\beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}}{m(y)} \tag{13}$$

That is, in the equation above, we can evaluate every term except for $m(y)$. Recall from (9) that the normalizing constant is also an integral of $\beta$.

$$m(y) = \int p(y|X, \beta)p(\beta)d\beta \tag{14}$$

Therefore, $m(y)$ can be rewritten as an expectation with respect to $g(\beta)$

$$m(y) = \int \frac{p(y|X, \beta)p(\beta)}{g(\beta)} g(\beta)d\beta = E^g\left[\frac{p(y|X, \beta)p(\beta)}{g(\beta)}\right] \tag{15}$$

which yields the importance sampling estimator

$$\bar{m}_{IS,R} = \frac{1}{R}\sum_{r=1}^{R} \frac{p(y|X, \beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}. \tag{16}$$

Combining (13) and (16) produces an importance sampling estimate of the posterior mean

$$\begin{aligned} E^{\pi}[\beta] &\approx \frac{\frac{1}{R}\sum_{r=1}^{R} \beta^{(r)} \frac{p(y|X,\beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}}{\frac{1}{R}\sum_{r=1}^{R} \frac{p(y|X,\beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}} \\ &= \frac{\sum_{r=1}^{R} \beta^{(r)} \frac{p(y|X,\beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}}{\sum_{r=1}^{R} \frac{p(y|X,\beta^{(r)})p(\beta^{(r)})}{g(\beta^{(r)})}} \\ &= \frac{\sum_{r=1}^{R} \beta^{(r)} w^{(r)}}{\sum_{r=1}^{R} w^{(r)}} \end{aligned} \tag{17}$$

where $w^{(r)} = p(y|X, \beta^{(r)})p(\beta^{(r)})/g(\beta^{(r)})$.

Finally, notice that the evaluation of $w^{(r)}$ does not require us to compute the normalizing constants of the prior $p(\beta)$ and importance density $g(\beta)$. To see why, suppose we only had access to the unnormalized densities $p^*$ and $g^*$.

$$\begin{aligned} p^*(\beta) &= c_1 p(\beta) \\ g^*(\beta) &= c_2 g(\beta) \end{aligned}$$

Then the weights based on these unnormalized densities can be written as

$$w^{*(r)} = \frac{p(y|X, \beta^{(r)}) p^*(\beta^{(r)})}{g^*(\beta^{(r)})} = w^{(r)} \times \frac{c_1}{c_2}.$$

But now the estimator based on the unnormalized weights reduces to the same estimator from (17).

$$\begin{aligned} E^{*\pi}[\beta] &\approx \frac{\sum_{r=1}^R \beta^{(r)} w^{*(r)}}{\sum_{r=1}^R w^{*(r)}} \\ &= \frac{\sum_{r=1}^R \beta^{(r)} w^{(r)} \frac{c_1}{c_2}}{\sum_{r=1}^R w^{(r)} \frac{c_1}{c_2}} \\ &= \frac{\frac{c_1}{c_2} \sum_{r=1}^R \beta^{(r)} w^{(r)}}{\frac{c_1}{c_2} \sum_{r=1}^R w^{(r)}} \\ &= \frac{\sum_{r=1}^R \beta^{(r)} w^{(r)}}{\sum_{r=1}^R w^{(r)}} \end{aligned}$$

# References

Robert, Christian P. and George Casella (2004), *Monte Carlo Statistical Methods.* Springer.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005), *Bayesian Statistics and Marketing.* New York: John Wiley and Sons.