Monte Carlo Statistical Methods

John C. Chao

Econ 721 Lecture Notes

October 18, 2022

Inversion Method

- It is well-known that there exist very good pseudo-random number generators for making $i.i.d.$ draws from a uniform distribution. See, for example, Knuth (1997). How do we make draws from a non-uniform distribution, however? In principle, we can use the uniform pseudo-random number generator to make draws from any distribution for which the cumulative distribution function (cdf) is known by making use of the so-called inversion method, as given by the lemma below.
- **Lemma:** Let

$$U \sim \text{Uniform}\left[0,1\right]$$

and let $F\left(\cdot\right)$ be a one-dimensional cdf. Then,

$$X \sim F^{-}\left(U\right)$$

has the distribution F. Here, we define

$$F^{-}\left(u\right) = \inf\left\{x : u \leq F\left(x\right)\right\}.$$

Inversion Method

- **Proof:** We will prove the lemma only for the case where $F(\cdot)$ is continuous and strictly increasing, i.e., X is a continuously distributed random variable. In this case, note that

$$\begin{aligned} \Pr(X \leq x) &= \Pr\left(F^{-1}(U) \leq x\right) \\ &= \Pr(U \leq F(x)) \\ &= F(x) \end{aligned}$$

  where the last inequality follows from the fact that U is uniformly distributed on the interval $[0, 1]$.
- **Remark:** In practice, however, for many distributions (including the normal distribution), $F(\cdot)$ is not known in closed form, so it is difficult to implement the inversion method directly.

Rejection (or Acceptance/Rejection) Method

- This method, due to von Neumann (1951) can be applied to make draws from any finite dimensional probability distribution with a density that is specified up to a normalization constant.
- Consider the following situation.

1. Suppose that we wish to make draws from $\pi(x)$, which is a density function or a probability mass function, but we do not know how to do that directly.
2. Suppose that, given x, it is at least easy to compute (or evaluate) the function
$$l(x) = a\pi(x)$$
for some possibly positive constant a.
3. Finally, suppose that it is easy to draw from a sampling distribution $g(x)$, and there exists a "covering constant" M such that the following envelop property holds
$$l(x) \leq Mg(x) \text{ for all } x.$$

Rejection (or Acceptance/Rejection) Method

- In this setting, we can apply the following algorithm:

(i) Draw x from $g(\cdot)$ and compute the ratio

$$p(x) = \frac{l(x)}{Mg(x)} \quad (\leq 1 \text{ given the envelope property})$$

(ii) Draw u from Uniform$[0, 1]$ and

$$\text{Accept } x \text{ as one of the draws if } u \leq p(x) = \frac{l(x)}{Mg(x)}$$

$$\text{Reject } x \text{ and go back to step (i) if } u > p(x) = \frac{l(x)}{Mg(x)}$$

(iii) Iterate between steps (i) and (ii) until we have collected the needed sample.

Rejection (or Acceptance/Rejection) Method

- We will show that the accepted sample generated by the algorithm above follows the target distribution $\pi(x)$.
- To proceed, let $\mathbb{I}_X$ denote an indicator function such that

$$\mathbb{I}_X = \begin{cases} 1 & \text{if } X \text{ drawn from } g(\cdot) \text{ is accepted} \\ 0 & \text{if } X \text{ drawn from } g(\cdot) \text{ is rejected} \end{cases}$$

- Note that

$$\begin{aligned} \Pr\{\mathbb{I}_X = 1\} &= \int \Pr\{\mathbb{I}_X = 1 | X = x\} g(x)\, dx \\ &= \int \frac{l(x)}{Mg(x)} g(x)\, dx = \int \frac{l(x)}{M} dx \\ &= \frac{a}{M} \int \frac{l(x)}{a} dx \\ &= \frac{a}{M} \int \pi(x)\, dx = \frac{a}{M}. \end{aligned}$$

Rejection (or Acceptance/Rejection) Method

- It follows that

$$\begin{aligned} p\left(x|\mathbb{I}_X = 1\right) &= \frac{\Pr\left\{\mathbb{I}_X = 1|X = x\right\} g\left(x\right)}{\Pr\left\{\mathbb{I}_X = 1\right\}} \text{ (by Bayes rule)} \\ &= a\left(\frac{1}{Mg\left(x\right)}\frac{l\left(x\right)}{a}\right) g\left(x\right) / \frac{a}{M} \\ &= \pi\left(x\right). \end{aligned}$$

- **Remark (i):** Note that the above calculations imply that we must have $a \leq M$.

Rejection (or Acceptance/Rejection) Method

- **Remark (ii):** Note also that our decision to accept or reject is based essentially on a draw from a Bernoulli distribution with probability of success (or acceptance) given by

$$p = p(x) = \frac{l(x)}{Mg(x)}$$

conditional on $X = x$. On the other hand, the unconditional probability of acceptance obtained by averaging with respect to the distribution $g(x)$ is given by

$$\overline{p} = E_g[p(x)] = \frac{a}{M} = \Pr\{\mathbb{I}_X = 1\}.$$

- **Remark (iii):** Hence, if we think this algorithm as being more efficient if we get high acceptance on average, then a good trial distribution $g(x)$ is one which gives a small M (since there is in general nothing we can do about the size of a, which might even be unknown).

Rejection (or Acceptance/Rejection) Method

- **Remark (iv):** We can also calculate the expected number of draws before we would obtain the first acceptance. To do this, note first that the probability of getting the first success in the y^{th} trial when the probability of success for each trial is

$$\overline{p} = \Pr\left\{\mathbb{I}_X = 1\right\} = \frac{a}{M}$$

is given by the probability distribution of a geometric random variable Y, i.e.,

$$\Pr\left\{Y = y\right\} = \overline{p}\left(1 - \overline{p}\right)^{y-1} \text{ for } y = 1, 2, ...$$

The expected value of this random variable can be calculated as

$$E\left[Y\right] = \frac{1}{\overline{p}} = \frac{M}{a}.$$

It follows that, if $a \approx 1$, then M is approximately the expected number of draws before we obtain the first acceptance.

Monte Carlo Integration

- Suppose that we want to compute the expectation (or integral)

$$\mu = \mathbb{E}_\pi \left[h\left(X \right) \right] = \int_{\mathcal{X}} h\left(x \right) \pi\left(x \right) dx.$$

It would seem that a natural way to proceed is to draw an $i.i.d.$ sample $\left(X^{(1)}, ..., X^{(m)} \right)$ from π and approximate $\mathbb{E}_\pi \left[h\left(X \right) \right]$ using the empirical average

$$\widetilde{\mu} = \frac{1}{m} \sum_{i=1}^{m} h\left(x^{(i)} \right)$$

- Assume the moment condition

$$E_\pi \left[\left| h\left(X \right) \right| \right] = \int_{\mathcal{X}} \left| h\left(x \right) \right| \pi\left(x \right) dx < \infty;$$

then, by the strong law of large numbers (SLLN)

$$\frac{1}{m} \sum_{i=1}^{m} h\left(X^{(i)} \right) \overset{a.s.}{\longrightarrow} \mathbb{E}_\pi \left[h\left(X \right) \right] \text{ as } m \longrightarrow \infty.$$

Monte Carlo Integration

- Suppose a stronger (2nd) moment condition holds, i.e.,

$$E_\pi\left\{\left[h\left(X\right)\right]^2\right\} < \infty;$$

then, by the Lindeberg-Lévy central limit theorem (CLT), we further obtain

$$\sqrt{m}\left(\frac{\widetilde{\mu} - \mathbb{E}_\pi\left[h\left(X\right)\right]}{\widehat{\sigma}_{h,m}}\right) \xrightarrow{d} N\left(0,1\right) \text{ as } m \to \infty,$$

where

$$\widehat{\sigma}^2_{h,m} = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - \widetilde{\mu}\right]^2$$

Monte Carlo Integration

- Note that the central limit theorem also provides us with a rate of convergence, so that, under the second moment condition, we have

$$\widetilde{\mu} - \mathbb{E}_{\pi}\left[h\left(X\right)\right] = O_p\left(\frac{1}{\sqrt{m}}\right)$$

- Interestingly, the rate of convergence depends only on m, the number of draws, and not on $\dim\left(x\right)$. This is why people are interested in Monte Carlo methods as a way of computing high-dimensional integral. Monte Carlo methods allow one to, in some sense, circumvent the so-called curse of dimensionality.

Importance Sampling

Some Motivation

- A problem with the Monte Carlo strategy discussed so far is that it could be very inefficient from a computational standpoint. In particular, if the support of $h(x)$ and $\pi(x)$ are substantially different, by drawing from the distribution $\pi(x)$, we could be wasting a lot of effort in regions of the integral where the product $h(x)\pi(x)$ (i.e., the integrand) is very close to zero.

Importance Sampling

- A second potential drawback of the above strategy is that we might want to evaluate not one but a family of related integrals, i.e.,

$$\int_{\mathcal{X}} h_1(x)\,\pi_1(x)\,dx,$$
$$\int_{\mathcal{X}} h_2(x)\,\pi_2(x)\,dx,$$
$$\vdots$$
$$\int_{\mathcal{X}} h_J(x)\,\pi_J(x)\,dx,$$

  If we estimate each integral

$$\int_{\mathcal{X}} h_i(x)\,\pi_i(x)\,dx$$

  by drawing from a separate distribution $\pi_i(x)$, this could again be inefficient and could lead to a lot of draws. Hence, we may want to have a method where we can draw one sample and possibly use it for several related problems.

Importance Sampling

- The method of importance sampling is an evaluation of the integral

$$\mu = \mathbb{E}_\pi \left[h\left(X \right) \right] = \int_{\mathcal{X}} h\left(x \right) \pi\left(x \right) dx.$$

based on generating a sample $\left(X^{(1)}, ..., X^{(m)} \right)$ from a given distribution g and approximating μ with

$$\begin{aligned} \overline{\mu} &= \frac{1}{m} \sum_{i=1}^{m} \frac{h\left(X^{(i)} \right) \pi\left(X^{(i)} \right)}{g\left(X^{(i)} \right)} \\ &= \frac{1}{m} \sum_{i=1}^{m} w\left(X^{(i)} \right) h\left(X^{(i)} \right) \end{aligned}$$

where

$$w\left(X^{(i)} \right) = \frac{\pi\left(X^{(i)} \right)}{g\left(X^{(i)} \right)}$$

are the importance weights.

Importance Sampling

- The intuition behind the importance sampling method is based on (alternative) representation of the integral of interest

$$\mathbb{E}_{\pi}\left[h\left(X\right)\right]=\int_{\mathcal{X}}\frac{h\left(x\right)\pi\left(x\right)}{g\left(x\right)}g\left(x\right)dx$$

  which is called the importance sampling fundamental identity.
- For this method to work, we need a support condition

$$\text{supp}\left(\pi\right)\subset\text{supp}\left(g\right)$$

Importance Sampling

- **Unbiasedness:** Suppose we draw an $i.i.d.$ sample $\left(X^{(1)}, ..., X^{(m)}\right)$ from g. Note that

$$\begin{aligned}
\mathbb{E}_g\left[\overline{\mu}\right] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_g\left[\frac{h\left(X^{(i)}\right)\pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)}\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{X}_g}\frac{h\left(x\right)\pi\left(x\right)}{g\left(x\right)}g\left(x\right)dx \\
&= \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{X}_g}h\left(x\right)\pi\left(x\right)dx \\
&= \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{X}_\pi}h\left(x\right)\pi\left(x\right)dx \ \ \left(\text{since supp}\left(\pi\right)\subset\text{supp}\left(g\right)\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_\pi\left[h\left(X\right)\right] \\
&= \mu.
\end{aligned}$$

Importance Sampling

- **Unbiasedness (con't):** Hence, $\overline{\mu}$ is an unbiased (Monte Carlo) estimator of μ when the expectation is taken with respect to g.
- Suppose, in addition, that

$$\begin{aligned}
& \mathbb{E}_g\left[\left|\frac{h\left(X^{(i)}\right)\pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)}\right|\right] \\
= & \int_{\mathcal{X}_g}\left|\frac{h(x)\,\pi(x)}{g(x)}\right| g(x)\,dx \\
= & \int_{\mathcal{X}_g}|h(x)|\,\pi(x)\,dx \\
= & \int_{\mathcal{X}_\pi}|h(x)|\,\pi(x)\,dx \;\; (\text{since } \operatorname{supp}(\pi)\subset\operatorname{supp}(g)) \\
< & \infty.
\end{aligned}$$

Importance Sampling

- Then, by the SLLN, we have

$$\overline{\mu} = \frac{1}{m}\sum_{i=1}^{m} \frac{h\left(X^{(i)}\right)\pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)} \xrightarrow{a.s.} \mu \text{ as } m \to \infty$$

- **Remark (i):** Given the support condition, the moment condition needed for the strong consistency of $\overline{\mu}$, i.e., the condition

$$E_{\pi}\left[\left|h\left(X\right)\right|\right] = \int_{\mathcal{X}} \left|h\left(x\right)\right| \pi\left(x\right) dx < \infty$$

is the same as moment condition which ensures the strong consistency of $\widetilde{\mu}$, even though we are now drawing from the distribution g instead of the distribution π.

Importance Sampling

- **Remark (ii):** Note that the importance sampling fundamental identity is a very general representation which expresses the fact that a given integral is not intrinsically associated with a given distribution. Importance sampling is therefore of considerable interest since it is a methodology that puts very little restriction on the choice of the instrumental distribution g, which can be conveniently chosen to be one that is easy to simulate from. Moreover, it may be possible to use the same sample (generated from g) repeatedly, not only for different functions h but also for different densities π, a feature that is particularly attractive for robustness and Bayesian sensitivity analysis.

Importance Sampling

- **Finite Variance Importance Sampler:** Although, in principle, to implement the importance sampling algorithm, we can sample from any distribution g that satisfies the support condition

$$\text{supp}\left(\pi\right) \subset \text{supp}\left(g\right),$$

in practice some choices might be better than others. More specifically, to get a CLT-type result, i.e.,

$$\sqrt{m}\left(\frac{\overline{\mu} - \mathbb{E}_{\pi}\left[h\left(X\right)\right]}{\overline{\sigma}_{h,m}}\right) \xrightarrow{d} N\left(0,1\right) \text{ as } m \longrightarrow \infty,$$

where

$$\begin{aligned}
\overline{\mu} &= \frac{1}{m}\sum_{i=1}^{m}\frac{h\left(X^{(i)}\right)\pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)}, \\
\overline{\sigma}_{h,m}^{2} &= \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - \overline{\mu}\right]^{2}
\end{aligned}$$

we also need a second moment condition.

Importance Sampling

- In this case, the variance is finite if

$$\begin{aligned}
\mathbb{E}_g\left[\frac{h^2(X)\pi^2(X)}{g^2(X)}\right] &= \int_{\mathcal{X}_g}\frac{h^2(x)\pi^2(x)}{g^2(x)}g(x)\,dx \\
&= \int_{\mathcal{X}_g}\frac{h^2(x)\pi(x)}{g(x)}\pi(x)\,dx \\
&= \int_{\mathcal{X}_\pi}\frac{h^2(x)\pi(x)}{g(x)}\pi(x)\,dx \\
&\quad (\text{since } \operatorname{supp}(\pi)\subset\operatorname{supp}(g)) \\
&= \mathbb{E}_\pi\left[\frac{h^2(X)\pi(X)}{g(X)}\right] \\
&< \infty.
\end{aligned}$$

Importance Sampling

- It follows that instrumental distributions with tails lighter than those of π (that is, those with unbounded ratios π/g) are not appropriate for importance sampling. In fact, in these cases, the variances of the corresponding estimators

$$\begin{aligned} \overline{\mu} &= \frac{1}{m}\sum_{i=1}^{m}\frac{h\left(X^{(i)}\right)\pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)} \\ &= \frac{1}{m}\sum_{i=1}^{m} w\left(X^{(i)}\right) h\left(X^{(i)}\right) \end{aligned}$$

  will be infinite for many functions h.

Importance Sampling

- More generally, if the ratio

$$w = \frac{\pi}{g}$$

is unbounded, the weights $w\left(x^{(i)}\right) = \pi\left(x^{(i)}\right)/g\left(x^{(i)}\right)$ may vary widely and possibly give too much importance to a few $x^{(i)}$ values. This, in turn, could lead to the value of the estimate $\overline{\mu}$ changing abruptly from one iteration to the next, even after many iterations. Conversely, specifying an instrumental distribution g with thicker tails than π could lead to more stable results.

- Geweke (1989) gives two types of sufficient conditions to ensure the finiteness of $\mathbb{E}_\pi\left[h^2\pi/g\right]$

(i) $w(x) = \pi(x)/g(x) < M < \infty$ for all $x \in \mathcal{X}$ and $\mathbb{E}_\pi\left[h^2\right] < \infty$

(ii) $\mathcal{X}$ is compact, $\pi(x) < C < \infty$ and $g(x) > \epsilon > 0$ for all $x \in \mathcal{X}$.

Importance Sampling

- A natural question to ask at this point is that, amongst all distributions g leading to finite variances for the estimator

$$\overline{\mu} = \frac{1}{m} \sum_{i=1}^{m} \frac{h\left(X^{(i)}\right) \pi\left(X^{(i)}\right)}{g\left(X^{(i)}\right)}$$

  what is the form of the optimal distribution given a particular function h and a fixed distribution π? The answer is provided by the following theorem from Rubinstein (1981).

- **Theorem:** The choice of g that minimizes the variance of the estimator $\overline{\mu}$ is

$$g^{*}(x) = \frac{|h(x)| \pi(x)}{\int_{\mathcal{X}_{\pi}} |h(x)| \pi(x) \, dx}$$

Importance Sampling

- **Proof of Theorem:** Note first that

$$
\begin{aligned}
& var_g\left[\frac{h(X)\pi(X)}{g(X)}\right] \\
= \; & \mathbb{E}_g\left[\frac{h^2(X)\pi^2(X)}{g^2(X)}\right] - \left(\mathbb{E}_g\left[\frac{h(X)\pi(X)}{g(X)}\right]\right)^2 \\
= \; & \mathbb{E}_g\left[\frac{h^2(X)\pi^2(X)}{g^2(X)}\right] - \left(\int_{\mathcal{X}_g}\frac{h(x)\pi(x)}{g(x)}g(x)\,dx\right)^2 \\
= \; & \mathbb{E}_g\left[\frac{h^2(X)\pi^2(X)}{g^2(X)}\right] - \left(\int_{\mathcal{X}_\pi}h(x)\pi(x)\,dx\right)^2 \\
= \; & \mathbb{E}_g\left[\frac{h^2(X)\pi^2(X)}{g^2(X)}\right] - \left(\mathbb{E}_\pi\left[h(X)\right]\right)^2
\end{aligned}
$$

so that the second term does not depend on g. Hence, to minimize the variance, we only need to choose g to minimize the first term.

Importance Sampling

- **Proof of Theorem (con't):** From Jensen's inequality, it follows that

$$\begin{aligned}
\mathbb{E}_g\left[\frac{h^2(X)\,\pi^2(X)}{g^2(X)}\right] &\geq \left(\mathbb{E}_g\left[\left|\frac{h(X)\,\pi(X)}{g(X)}\right|\right]\right)^2 \\
&= \left(\mathbb{E}_g\left[\frac{|h(X)|\,\pi(X)}{g(X)}\right]\right)^2 \\
&= \left(\int_{\mathcal{X}_\pi} |h(x)|\,\pi(x)\,dx\right)^2
\end{aligned}$$

which provides a lower bound that is independent of g.

Importance Sampling

- **Proof of Theorem (con't):** Now, take $g = g^*$, and we have

$$\begin{aligned} & \mathbb{E}_{g^*}\left[\frac{h^2(X)\pi^2(X)}{(g^*(X))^2}\right] \\ = & \mathbb{E}_{g^*}\left[\frac{h^2(X)\pi^2(X)}{h^2(X)\pi^2(X)}\left(\int_{\mathcal{X}_\pi}|h(x)|\pi(x)dx\right)^2\right] \\ = & \left(\int_{\mathcal{X}_\pi}|h(x)|\pi(x)dx\right)^2 \end{aligned}$$

so that setting $g = g^*$ satisfies the lower bound and, thus, is optimal. □