Monte Carlo Methods

Jesús Fernández-Villaverde
University of Pennsylvania

Why Monte Carlo?

- From previous chapter, we want to compute:

  1. Posterior distribution:

$$\pi\left(\theta|Y^T, i\right) = \frac{f(Y^T|\theta, i)\pi\left(\theta|i\right)}{\int_{\Theta_i} f(Y^T|\theta, i)\pi\left(\theta|i\right) d\theta}$$

  2. Marginal likelihood:

$$P\left(Y^T|i\right) = \int_{\Theta_i} f(Y^T|\theta, i)\pi\left(\theta|i\right) d\theta$$

- Difficult to asses analytically or even to approximate (Phillips, 1996).

- Resort to simulation.

A Bit of Historical Background and Intuition

- Metropolis and Ulam (1949) and Von Neuman (1951).

- Why the name "Monte Carlo"?

- Two silly examples:

  1. Probability of getting a total of six points when rolling two (fair) dices.

  2. Throwing darts at a graph.

Classical Monte Carlo Integration

- Assume we know how to generate draws from $\pi\left(\theta|Y^T, i\right)$.

- What does it mean to draw from $\pi\left(\theta|Y^T, i\right)$?

- Two Basic Questions:

    1. Why do we want to do it?

    2. How do we do it?

Why Do We Do It?

- Basic intuition: Glivenko-Cantelli's Theorem.

- Let $X_1, \ldots, X_n$ be iid as X with distribution function F. Let ω be the outcome and $F_n(x, \omega)$ be the empirical distribution function based on observations $X_1(\omega), \ldots, X_n(\omega)$. Then, as $n \to \infty$,

$$\sup_{-\infty < x < \infty} |F_n(x, \omega) - F(x)| \stackrel{a.s.}{\to} 0,$$

- It can be generalized to include dependence: A.W. Van der Vaart and J.A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, 1997.

Basic Result

- Let $h(\theta)$ be a function of interest: indicator, moment, etc...

- By the Law of Large Numbers:

$$E_{\pi\left(\cdot|Y^T,i\right)}\left[h(\theta)\right] = \int_{\Theta_i} h(\theta)\,\pi\left(\theta|Y^T,i\right)d\theta \simeq h_m = \frac{1}{m}\sum_{j=1}^{m} h\left(\theta_j\right)$$

- If $Var_{\pi\left(\cdot|Y^T,i\right)}\left[h(\theta)\right] < \infty$, by the Central Limit Theorem:

$$Var_{\pi\left(\cdot|Y^T,i\right)}\left[h_m\right] \simeq \frac{1}{m}\sum_{j=1}^{m}\left(h\left(\theta_j\right) - h_m\right)^2$$

How Do We Do It?

- Large literature.

- Two good surveys:

  1. Luc Devroye: *Non-Uniform Random Variate Generation*, Springer-Verlag, 1986. Available for free at

     `http://jeff.cs.mcgill.ca/~luc/rnbookindex.html`.

  2. Christian Robert and George Casella, *Monte Carlo Statistical Methods*, 2nd ed, Springer-Verlag, 2004.

Random Draws?

- Natural sources of randomness. Difficult to use.
- A computer...
- ...but a computer is a deterministic machine!
- Von Neumann (1951):

"Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin."

Was Von Neumann Right?

- Let's us do a simple experiment.

- Let's us start MATLAB, type `format long`, type `rand`.

- Did we get 0.95012928514718?

- This does not look terribly random.

- Why is this number appearing?

Basic Building Block

- MATLAB uses highly non-linear iterative algorithms that "look like" random.

- That is why sometimes we talk of pseudo-random number generators.

- We will concentrate on draws from a uniform distribution.

- Other (standard and nonstandard) distributions come from manipulations of the uniform.

Goal

Derive algorithms that, starting from some initial value and applying iterative methods, produce a sequence that: (Lehmer, 1951):

1. It is unpredictable for the uninitiated (relation with Chaotic dynamical systems).

2. It passes a battery of standard statistical tests of randomness (like Kolmogorov-Smirnov test, ARMA(p,q), etc).

Basic Component

- Multiplicative Congruential Generator:

$$x_i = (ax_{i-1} + b) \bmod (M + 1)$$

- x_i takes values on $\{0, 1, ..., M\}$.

- Transformation into a generator on $[0, 1]$ with:

$$x_i^* = \frac{ax_{i-1} + b}{M + 1}$$

- x_0 is called the *seed*.

Choices of Parameters

- Period and performance depends crucially on a, b, and M.
- Pick $a = 13$, $c = 0$, $M = 31$, and $x_0 = 1$.
- Let us run `badrandom.m`.
- Historical bad examples: `IBM RND` from the 1960's.

A Good Choice

- A traditional choice: $a = 7^5 = 16807$, $c = 0$, $m = 2^{31} - 1$.

- Period bounded by M. 32 bits versus 64 bits hardware.

- You may want to be aware that there is something called *IEEE standard for floating point arithmetic*.

- Problems and alternatives.

Real Life

- You do not want to code your own random number generator.
- Matlab implements the state of the art: `KISS` by Marsaglia and Zaman (1991).
- What about a compiler in Fortran or C++?
- `http://stat.fsu.edu/pub/diehard/`

Nonuniform Distributions

- In general we need something different from the uniform distribution.
- How do we move from a uniform draw to other distributions?
- Simple transformations for standard distributions.
- Foundations of commercial software.

Two Basic Approaches

1. Use transformations.

2. Use inverse method.

Why are those approaches attractive?

An Example of Transformations I: Normal Distributions

- Box and Muller (1958).

- Let U_1 and U_2 two uniform variables, then:

$$\begin{aligned} x &= \cos 2\pi U_1 \, (-2 \log U_2)^{0.5} \\ y &= \sin 2\pi U_1 \, (-2 \log U_2)^{0.5} \end{aligned}$$

  are independent normal variables.

- Problem: x and y fall in a spiral.

An Example of Transformations II: Multivariate Normal Distributions

- If $x \sim \mathcal{N}(0, I)$, then

$$y = \mu + \Sigma x$$

  is distributed as $\mathcal{N}(\mu, \Sigma'\Sigma)$.

- Σ can be the Cholesky decomposition of the matrix of variances-covariances.

An Example of Transformations III: Discrete Uniform

- We want to draw from $x \sim \mathcal{U}\{1, N\}$.

- Find $1/N$.

- Draw from $U \sim \mathcal{U}[0.1]$.

- If $u \in [0, 1/N] \Rightarrow x = 1$, if $u \in [1/N, 2/N] \Rightarrow x = 3$, and so on.

Inverse Method

- Conceptually general procedure for random variables on $\Re$.

- For a non-decreasionf function F on $\Re$, the *generalized inverse* of F, F^{-}, is the function

$$F^{-}(u) = \inf\{x : F(x) \geq u\}$$

- Lemma: If $U \sim \mathcal{U}[0.1]$, then the random variable $F^{-}(U)$ has the distribution F.

Proof:

For $\forall u \in [0.1]$ and $\forall x \in F^{-}([0.1])$, we satisfy:

$$F\left(F^{-}(u)\right) \geq u \text{ and } F^{-}(F(x)) \leq x$$

Therefore

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

and

$$P\left(F^{-}(U) \leq x\right) = P(U \leq F(x)) = F(x)$$

An Example of the Inverse Method

- Exponential distribution: $x \sim Exp(1)$.
- $F(x) = 1 - e^{-x}$.
- $x = -\log(1 - u)$.
- Thus, $X = -\log U$ is exponential if U is uniform.

Problems

- Algebraic tricks and the inverse method are difficult to generalize.
- Why? Complicated, multivariate distributions.
- Often, we only have a numerical procedure to evaluate the distribution we want to sample from.
- We need more general methods.

Acceptance Sampling

- $\theta \sim \pi\left(\theta|Y^T, i\right)$ with support C. $\pi\left(\theta|Y^T, i\right)$ is called the target density.

- $z \sim g(z)$ with support $C' \supseteq C$. g is called the source density.

- We require:

  1. We know how to draw from g.

  2. Condition:

$$\sup_{\theta \in C} \frac{\pi\left(\theta|Y^T, i\right)}{g(\theta)} = a < \infty$$

Acceptance Sampling Algorithm

Steps:

1. $u \sim U(0, 1)$.

2. $\theta^* \sim g$.

3. If $u > \frac{\pi\left(\theta^*|Y^T, i\right)}{a g(\theta^*)}$ return to step 1.

4. Set $\theta^m = \theta^*$.

Why Does Acceptance Sampling Work?

- Unconditional probability of moving from step 3 to 4:

$$\int_{C'} \frac{\pi\left(\theta|Y^T,i\right)}{ag(\theta)} g(\theta)\, d\theta = \int_{C'} \frac{\pi\left(\theta|Y^T,i\right)}{a} d\theta = \frac{1}{a}$$

- Unconditional probability of moving from step 3 to 4 when $\theta \in A$:

$$\int_{A} \frac{\pi\left(\theta|Y^T,i\right)}{ag(\theta)} g(\theta)\, d\theta = \int_{A} \frac{\pi\left(\theta|Y^T,i\right)}{a} d\theta = \frac{1}{a}\int_{A} \pi\left(\theta|Y^T,i\right) d\theta$$

- Dividing both expressions:

$$\frac{\frac{1}{a}\int_A \pi\left(\theta|Y^T,i\right) d\theta}{\frac{1}{a}} = \int_A \pi\left(\theta|Y^T,i\right) d\theta$$

An Example

- Target density:

$$\pi\left(\theta|Y^T, i\right) \propto \min\left[\exp\left(-\frac{\theta^2}{2}\right)\left(\sin\left(6\theta\right)^2 + 3\cos\left(\theta\right)^2\sin\left(4\theta\right)^2 + 1\right), 0\right]$$

- Source density:

$$g\left(\theta\right) \propto \frac{1}{\left(2\pi\right)^{0.5}}\exp\left(-\frac{\theta^2}{2}\right)$$

- Let's take a look: `acceptance.m`.

Problems of Acceptance Sampling

- Two issues:

  1. We disregard a lot of draws. We want to minimize a. How?

  2. We need to check π/g is bounded. Necessary condition: g has thicker tails than those of f.

  3. We need to evaluate bound a. Difficult to do.

- Can we do better? Yes, through importance sampling.

Importance Sampling I

- Similar framework than in acceptance sampling:

  1. $\theta \sim \pi\left(\theta|Y^T, i\right)$ with support C. $\pi\left(\theta|Y^T, i\right)$ is called the target density.

  2. $z \sim g(z)$ with support $C' \supseteq C$. g is called the source density.

- Note that we can write:

$$E_{\pi(\cdot|Y^T,i)}\left[h(\theta)\right] = \int_{\Theta_i} h(\theta)\,\pi\left(\theta|Y^T, i\right) d\theta = \int_{\Theta_i} h(\theta)\,\frac{\pi\left(\theta|Y^T, i\right)}{g(\theta)} g(\theta)\, d\theta$$

Importance Sampling II

- If $E_{\pi(\cdot|Y^T,i)}[h(\theta)]$ exists, a Law of Large Numbers holds:

$$\int_{\Theta_i} h(\theta) \frac{\pi\left(\theta|Y^T, i\right)}{g(\theta)} g(\theta)\, d\theta \simeq h_m^I = \frac{1}{m} \sum_{j=1}^{m} h\left(\theta_j\right) \frac{\pi\left(\theta_j|Y^T, i\right)}{g\left(\theta_j\right)}$$

- and

$$E_{\pi(\cdot|Y^T,i)}[h(\theta)] \simeq h_m^I$$

where $\left\{\theta_j\right\}_{j=1}^{m}$ are draws from $g(\theta)$ and $\frac{\pi\left(\theta_j|Y^T,i\right)}{g\left(\theta_j\right)}$ are the important sampling weights.

Importance Sampling III

- If $E_{\pi\left(\theta|Y^T,i\right)}\left[\frac{\pi\left(\theta|Y^T,i\right)}{g(\theta)}\right]$ exists, a Central Limit Theorem applies (see Geweke, 1989) and:

$$m^{1/2}\left(h_m^I - E_{\pi\left(\cdot|Y^T,i\right)}\left[h\left(\theta\right)\right]\right) \rightarrow \mathcal{N}\left(0, \sigma^2\right)$$

$$\sigma^2 \simeq \frac{1}{m}\sum_{j=1}^{m}\left(h\left(\theta_j\right) - h_m^I\right)^2\left(\frac{\pi\left(\theta_j|Y^T, i\right)}{g\left(\theta_j\right)}\right)^2$$

- Where, again, $\left\{\theta_j\right\}_{j=1}^{m}$ are draws from $g\left(\theta\right)$.

Importance Sampling IV

- Notice that:

$$\sigma^2 \simeq \frac{1}{m} \sum_{j=1}^{m} \left(h\left(\theta_j\right) - h_m^I \right)^2 \left(\frac{\pi\left(\theta_j | Y^T, i\right)}{g\left(\theta_j\right)} \right)^2$$

- Therefore, we want $\frac{\pi\left(\theta|Y^T,i\right)}{g(\theta)}$ to be almost flat.

Importance Sampling V

- Intuition: σ^2 is minimized when $\pi\left(\theta|Y^T, i\right) = g(\theta)$.,i.e. we are drawing from $\pi\left(\theta_j|Y^T, i\right)$.

- Hint: we can use as $g(\theta)$ the first terms of a Taylor approximation to $\pi\left(\theta|Y^T, i\right)$.

- How do we compute the Taylor approximation?

Conditions for the existence of $E_{\pi(\theta|Y^T,i)}\left[\frac{\pi\left(\theta|Y^T,i\right)}{g(\theta)}\right]$

- This has to be checked analytically.

- A simple condition: $\frac{\pi\left(\theta|Y^T,i\right)}{g(\theta)}$ has to be bounded.

- Some times, we label $\omega\left(\theta|Y^T,i\right)=\frac{\pi\left(\theta|Y^T,i\right)}{g(\theta)}$.

Normalizing Factor I

- Assume we do not know the normalizing constant for $\pi\left(\theta|Y^T,i\right)$ and $g(\theta)$.

- Let's call the unnormalized densities: $\widetilde{\pi}\left(\theta|Y^T,i\right)$ and $\widetilde{g}(\theta)$.

- Then:

$$E_{\pi\left(\cdot|Y^T,i\right)}\left[h(\theta)\right]=\frac{\int_{\Theta_i}h(\theta)\,\widetilde{\pi}\left(\theta|Y^T,i\right)d\theta}{\int_{\Theta_i}\widetilde{\pi}\left(\theta|Y^T,i\right)d\theta}=\frac{\int_{\Theta_i}h(\theta)\,\frac{\widetilde{\pi}\left(\theta|Y^T,i\right)}{\widetilde{g}(\theta)}\widetilde{g}(\theta)\,d\theta}{\int_{\Theta_i}\frac{\widetilde{\pi}\left(\theta|Y^T,i\right)}{\widetilde{g}(\theta)}\widetilde{g}(\theta)\,d\theta}$$

Normalizing Factor II

- Consequently:

$$h_m^I = \frac{\frac{1}{m}\sum_{j=1}^{m} h\left(\theta_j\right)\frac{\pi\left(\theta_j|Y^T,i\right)}{g\left(\theta_j\right)}}{\frac{1}{m}\sum_{j=1}^{m}\frac{\pi\left(\theta_j|Y^T,i\right)}{g\left(\theta_j\right)}} = \frac{\sum_{j=1}^{m} h\left(\theta_j\right)\omega\left(\theta_j|Y^T,i\right)}{\sum_{j=1}^{m}\omega\left(\theta_j|Y^T,i\right)}$$

- and:

$$\sigma^2 \simeq \frac{m\sum_{j=1}^{m}\left(h\left(\theta_j\right)-h_m^I\right)^2\left(\omega\left(\theta_j|Y^T,i\right)\right)^2}{\left(\sum_{j=1}^{m}\omega\left(\theta_j|Y^T,i\right)\right)^2}$$

The Importance of the Behavior of $\omega\left(\theta_j|Y^T,i\right)$: Example I

- Assume that we know $\pi\left(\theta_j|Y^T,i\right) = t_\nu$.

- But we do not know how to draw from it.

- Instead we draw from $\mathcal{N}(0,1)$.

- Why?

- Let's run `normalt.m`

- Evaluate the mean of t_v.

- Draw $\left\{\theta_j\right\}_{j=1}^{m}$ from $\mathcal{N}\,(0,1)$.

- Let $\frac{t_v(\theta_j)}{\phi(\theta_j)} = \omega\left(\theta_j\right)$.

- Evaluate

$$mean = \frac{\sum_{j=1}^{m} \theta_j \omega\left(\theta_j\right)}{m}$$

- Evaluate the variance of the estimated mean of t_v.

- Compute:

$$var\_est\_mean = \frac{\sum_{j=1}^{m} \left(\theta_j - mean\right)^2 \omega\left(\theta_j\right)^2}{m}$$

- Note: difference between:

  1. The variance of a function of interest.

  2. The variance of the computed mean of the function of interest.

Estimation of the Mean of t_v: `importancenormal.m`

v	3	4	10	100
Est. Mean	0.1026	0.0738	0.0198	0.0000
Est. of Var. of Est. Mean	684.5160	365.6558	36.8224	3.5881

The Importance of the Behavior of $\omega\left(\theta_j|Y^T, i\right)$: Example II

- Opposite case than before.

- Assume that we know $\pi\left(\theta_j|Y^T, i\right) = \mathcal{N}(0, 1)$.

- But we do not know how to draw from it.

- Instead we draw from t_v.

Estimation of the Mean of $\mathcal{N}(0,1)$: `importancet.m`

t_ν	3	4	10	100
Est. Mean	-0.0104	-0.0075	0.0035	-0.0029
Est. of Var. of Est. Mean	2.0404	2.1200	2.2477	2.7444

A Procedure to Check How Good is the Important Sampling Function

- This procedure is due to Geweke.

- It is called Relative Numerical Efficiency (*RNE*).

- First notice that if $g(\theta) = \pi\left(\theta|Y^T, i\right)$, we have that:

$$\begin{aligned}
\sigma^2 &\simeq \frac{1}{m}\sum_{j=1}^{m}\left(h\left(\theta_j\right) - h_m^I\right)^2 \left(\frac{\pi\left(\theta_j|Y^T, i\right)}{g\left(\theta_j\right)}\right)^2 = \\
&= \frac{1}{m}\sum_{j=1}^{m}\left(h\left(\theta_j\right) - h_m^I\right)^2 \simeq Var_{\pi\left(\cdot|Y^T,i\right)}\left[h\left(\theta\right)\right]
\end{aligned}$$

A Procedure of Checking how Good is the important Sampling Function II

- Therefore, for a given $g(\theta)$, the *RNE*:

$$RNE = \frac{Var_{\pi(\cdot|Y^T,i)}\left[h(\theta)\right]}{\sigma^2}$$

- If RNE closed to 1 the important sampling procedure is working properly.

- If RNE is very low, closed to 0, the procedure is not working as properly.

Estimation of the Mean of t_v

t_ν	3	4	10	100
RNE	0.0134	0.0200	0.0788	0.2910

Estimation of the Mean of $\mathcal{N}(0,1)$

t_ν	3	4	10	100
RNE	0.4777	0.4697	0.4304	0.3471

Important Sampling and Robustness of Priors

- Priors are researcher specific.

- Imagine researchers 1 and 2 are working with the same model, i.e. with the same likelihood function, $f(y^T|\theta, 1) = f(y^T|\theta, 2)$. (Now 1 and 2 do not imply different models but different researchers)

- But they have different priors $\pi(\theta|1) \neq \pi(\theta|2)$.

- Imagine that researcher 1 has draws from the her posterior distribution $\left\{\theta_j\right\}_{j=1}^N \sim \pi\left(\theta|Y^T, 1\right)$.

A Simple Manipulation

- If researcher 2 wants to compute

$$\int_{\Theta_i} h(\theta)\, \pi\left(\theta | Y^T, 2\right) d\theta$$

for any $\ell(\theta)$, he does not need to recompute everything.

- Note that

$$\int_{\Theta_i} h(\theta)\, \pi\left(\theta | Y^T, 2\right) d\theta = \int_{\Theta_i} h(\theta) \frac{\pi\left(\theta | Y^T, 2\right)}{\pi\left(\theta | Y^T, 1\right)} \pi\left(\theta | Y^T, 1\right) d\theta =$$

$$\frac{\int_{\Theta_i} h(\theta) \frac{f(y^T|\theta,2)\pi(\theta|2)}{f(y^T|\theta,1)\pi(\theta|1)} \pi\left(\theta | Y^T, 1\right) d\theta}{\int_{\Theta_i} \frac{f(y^T|\theta,2)\pi(\theta|1)}{f(y^T|\theta,1)\pi(\theta|1)} \pi\left(\theta | Y^T, 1\right) d\theta} = \frac{\int_{\Theta_i} h(\theta) \frac{\pi(\theta|2)}{\pi(\theta|1)} \pi\left(\theta | Y^T, 1\right) d\theta}{\int_{\Theta_i} \frac{\pi(\theta|2)}{\pi(\theta|1)} \pi\left(\theta | Y^T, 1\right) d\theta}$$

Importance Sampling

- Then:

$$\frac{\frac{1}{m}\sum_{j=1}^{m} h\left(\theta_j\right)\frac{\pi(\theta_j|2)}{\pi(\theta_j|1)}}{\frac{1}{m}\sum_{j=1}^{m}\frac{\pi(\theta_j|2)}{\pi(\theta_j|1)}} = \frac{\sum_{j=1}^{m} h\left(\theta_j\right)\frac{\pi(\theta_j|2)}{\pi(\theta_j|1)}}{\sum_{j=1}^{m}\frac{\pi(\theta_j|2)}{\pi(\theta_j|1)}} \rightarrow$$

$$\frac{\int_{\Theta_i} h\left(\theta\right)\frac{\pi(\theta|2)}{\pi(\theta|1)}\pi\left(\theta|Y^T,1\right)d\theta}{\int_{\Theta_i}\frac{\pi(\theta|2)}{\pi(\theta|1)}\pi\left(\theta|Y^T,1\right)d\theta} = \int_{\Theta_i} h\left(\theta\right)\pi\left(\theta|Y^T,2\right)d\theta$$

- Simple computation.

- Increased variance.