

Stat 451 Lecture Notes 06¹²

Monte Carlo Integration

Ryan Martin
UIC

www.math.uic.edu/~rgmartin

¹Based on Chapter 6 in Givens & Hoeting, Chapter 23 in Lange, and Chapters 3–4 in Robert & Casella

²Updated: March 18, 2016

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo

- Sampling and integration are important techniques for a number of statistical inference problems.
- But often the target distributions are too complicated and/or the integrands are too complex or high-dimensional for these problems to be solved using the basic methods.
- In this section we will discuss a couple of clever but fairly simple techniques for handling such difficulties, both based on a concept of “importance sampling.”
- Some more powerful techniques, namely *Markov Chain Monte Carlo* (MCMC), will be discussed in the next section.

- Let $f(x)$ be a pdf defined on a sample space $\mathbb{X}$.
- $f(x)$ may be a Bayesian posterior density that's known only up to a proportionality constant.
- Let $\varphi(x)$ be a function mapping $\mathbb{X}$ to $\mathbb{R}$.
- The goal is to estimate $E\{\varphi(X)\}$ when $X \sim f(x)$; i.e.,

$$\mu_f(\varphi) := E\{\varphi(X)\} = \int_{\mathbb{X}} \varphi(x)f(x) dx.$$

- The function $\varphi(x)$ can be almost anything — in general, we will only assume that $\mu_f(|\varphi|) < \infty$.
- $g(x)$ will denote a generic pdf on $\mathbb{X}$, different from $f(x)$.

- The general Monte Carlo method is based on the two most important results of probability theory — the law of large numbers (LLN) and the central limit theorem (CLT).

LLN. If $\mu_f(|\varphi|) < \infty$ and $X_1, X_2, \dots$ are iid f , then

$$\bar{\varphi}_n := \frac{1}{n} \sum_{i=1}^n \varphi(X_i) \rightarrow \mu_f(\varphi), \quad \text{with probability 1.}$$

CLT. If $\mu_f(\varphi^2) < \infty$ and $X_1, X_2, \dots$ are iid f , then

$$\sqrt{n}\{\bar{\varphi}_n - \mu_f(\varphi)\} \rightarrow \mathbf{N}(0, \sigma_f^2(\varphi)), \quad \text{in distribution.}$$

- Note that CLT requires a finite variance while LLN does not.

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo

- Assume that we know how to sample from $f(x)$.
- Let $X_1, \dots, X_n$ be an independent sample from $f(x)$.
- Then LLN states that $\bar{\varphi}_n = \frac{1}{n} \sum_{i=1}^n \varphi(X_i)$ should be a good estimate of $\mu_f(\varphi)$ provided that n is large enough.
- It's an unbiased estimate for all n .
- If $\mu_f(\varphi^2) < \infty$, then CLT allows us to construct a confidence interval for $\mu_f(\varphi)$ based on our sample.
- In particular, a $100(1 - \alpha)\%$ CI for $\mu_f(\varphi)$ is

$$\text{mean}\{Y_1, \dots, Y_n\} \pm z_{1-\alpha/2}^* \frac{\text{sd}\{Y_1, \dots, Y_n\}}{\sqrt{n}},$$

where $Y_i = \varphi(X_i)$, $i = 1, \dots, n$.

Example

- Suppose $X \sim \text{Unif}(-\pi, \pi)$.
- The goal is to estimate $E\{\varphi(X)\}$, where $\varphi(x) = \sin(x)$, which we know to be 0.
- Take an iid sample of size $n = 1000$, from the uniform distribution and evaluate $Y_i = \sin(X_i)$.
- Summary statistics for the Y -sample include:

$$\text{mean} = -0.0167 \quad \text{and} \quad \text{sd} = 0.699.$$

- Then a 99% confidence interval for $\mu_f(\varphi)$ is

$$-0.0167 \pm \underbrace{2.5758}_{\text{qnorm}(0.995)} \frac{0.699}{\sqrt{1000}} = [-0.074, 0.040].$$

Example: p-value of the likelihood ratio test

- Let $X_1, \dots, X_n$ iid $\text{Pois}(\theta)$.
- Goal is to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- One idea is the likelihood ratio test, but formula is messy:

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} = e^{n\bar{X} - n\theta_0} \left(\frac{n\bar{X}}{n\theta_0} \right)^{-n\bar{X}}.$$

- Need null distribution of the likelihood ratio statistic to compute, say, a p-value, but this is not available.³
- Straightforward to get a Monte Carlo p-value.
- Note that Λ depends only on the sufficient statistic $n\bar{X}$, which is distributed as $\text{Pois}(n\theta_0)$ under H_0 .

³Wilks's theorem gives us a large-sample approximation...

- Advantages of Monte Carlo:
 - Does not depend on the dimension of the random variables.
 - Basically works for all functions $\varphi(x)$.
 - A number of different things can be estimated with the same simulated X_i 's.
- Disadvantages of Monte Carlo:
 - Can be slow.
 - Need to be able to sample from $f(x)$.
 - Error bounds are not as tight as for numerical integration.

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling**
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo

- Importance sampling techniques are useful in a number of situations; in particular,
 - when the target distribution $f(x)$ is difficult to sample from,
 - or to reduce the variance of basic Monte Carlo estimates.
- Next slides gives a simple example of the latter.
- Importance sampling is the general idea of sampling from a different distribution but weighting the observations to make them look more like a sample from the target.
- Similar in spirit to SIR...

Motivating example

- Goal is to estimate the probability that a fair die lands on 1.
- A basic Monte Carlo estimate $\bar{X}$ based on n iid $\text{Ber}(\frac{1}{6})$ samples has mean $\frac{1}{6}$ and variance $\frac{5}{36n}$.
- If we change the die to have three 1's and three non-1's, then the event of observing a 1 is more likely.
- To account for this change, we should weight each 1 observed with this new die by $\frac{1}{3}$.
- So, if $Y_i = \frac{1}{3} \times \text{Ber}(\frac{1}{2})$, then

$$E(Y_i) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \quad \text{and} \quad V(Y_i) = \left(\frac{1}{3}\right)^2 \cdot \frac{1}{4} = \frac{1}{36}.$$

- So, $\bar{Y}$ has the same mean as $\bar{X}$, but *much smaller* variance: $\frac{1}{36n}$ compared to $\frac{5}{36n}$.

- As before, $f(x)$ is the target and $g(x)$ is a generic envelope.
- Define importance ratios $w^*(x) = f(x)/g(x)$.
- Then the key observation is that

$$\mu_f(\varphi) = \mu_g(\varphi w^*).$$

- This motivates the (modified) Monte Carlo approach:
 - 1 Sample $X_1, \dots, X_n$ iid from $g(x)$.
 - 2 Estimate $\mu_f(\varphi)$ with $n^{-1} \sum_{i=1}^n \varphi(X_i) w^*(X_i)$.
- If $f(x)$ is known only up to proportionality constant, then use

$$\mu_f(\varphi) \approx \sum_{i=1}^n \varphi(X_i) w(X_i) = \frac{\sum_{i=1}^n \varphi(X_i) w^*(X_i)}{\sum_{i=1}^n w^*(X_i)}.$$

Another motivating example

- Consider estimating $\int_0^1 \cos(\pi x/2) dx$, which equals $2/\pi$.
- Treat as an expectation with respect to $X \sim \text{Unif}(0, 1)$.
- Can show that

$$V\left\{\cos\left(\frac{\pi X}{2}\right)\right\} \approx 0.095.$$

- Consider importance sampling with $g(y) = \frac{3}{2}(1 - y^2)$.
- Can show that

$$V\left\{\frac{2 \cos(\frac{\pi Y}{2})}{3(1 - Y^2)}\right\} \approx 0.00099.$$

- So, the importance sampling estimator will have much smaller variance than the basic Monte Carlo estimator!
- This is the same idea as in the simple dice example.

Example: size and power estimation

- Suppose we wish to test $H_0 : \lambda = 2$ vs. $H_1 : \lambda > 2$ based on a sample of size $n = 10$ from a $\text{Pois}(\lambda)$ distribution.
- One would be tempted to use the usual z-test:

$$Z = \frac{\bar{X} - 2}{\sqrt{2/10}}.$$

- Under the normal distribution theory, the size $\alpha = 0.05$ test rejects H_0 if $Z \geq 1.645$.
- But the distribution is not normal, so the actual Type I error probability for this test likely isn't 0.05.
- Used two approaches: basic Monte Carlo and importance sampling, with $g = \text{Pois}(2.4653)$.
- Respective 95% confidence intervals:

$$(0.0448, 0.0532) \quad \text{and} \quad (0.0520, 0.0611).$$

- The choice of envelope $g(x)$ is important. In particular, the importance ratio $w^*(x) = f(x)/g(x)$ must be well-behaved, otherwise the variance of the estimate could be too large.
- A general strategy is to take $g(x)$ to be a heavy-tailed distribution, like Student-t or a mixture thereof.
- To get an idea of what makes a good proposal, let's consider a practically useless result:⁴

$$\text{"optimal" proposal} \propto |\varphi(x)|f(x).$$

- Take-away message: want proposal to look like f , but we are less concerned in places where φ is near/equal to zero.

⁴See Theorem 3.12 in Robert & Casella.

Example: small tail probabilities

- Goal: estimate $P(Z > 4.5)$, where $Z \sim N(0, 1)$.
- Naive strategy: sample $Z_1, \dots, Z_N$ iid $N(0, 1)$ and compute

$$N^{-1} \sum_{j=1}^N I_{Z_j > 4.5}.$$

- However, we won't get many Z_j 's that exceed 4.5, so naive Monte Carlo estimate is likely to be zero.
- Here, $\varphi(z) = I_{z > 4.5}$ so we don't need to worry about how f and g compare outside the interval $[4.5, \infty)$.
- Idea: do importance sampling with proposal g as a shifted exponential distribution, i.e., $g(z) \propto e^{-(z-4.5)}$.
- Comparison: for $N = 10000$ we have

	MC	IS	truth
[1,]	0	3.316521e-06	3.397673e-06

Is the chosen proposal good?

- It is possible to use the weights to judge the proposal.
- For f known exactly, not just up to proportionality, define the *effective sample size*

$$\hat{N}(f, g) = \frac{n}{1 + \text{var}\{w^*(X_1), \dots, w^*(X_n)\}}.$$

- $\hat{N}(f, g)$ is bounded by n and measures approximately how many iid samples the weighted importance samples are worth.
- $\hat{N}(f, g)$ close to n indicates that $g(x)$ is a good proposal, close to 0 means $g(x)$ is a poor proposal.

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization**
- 5 Root-finding and optimization via Monte Carlo

- In statistics (e.g., Stat 411), the Rao–Blackwell theorem provides a recipe for reducing the variance of an unbiased estimator by conditioning.
- The basis of this result are the two simple formulas:

$$\begin{aligned}E(Y) &= E\{E(Y \mid X)\} \\V(Y) &= E\{V(Y \mid X)\} + V\{E(Y \mid X)\} \\&\geq V\{E(Y \mid X)\}.\end{aligned}$$

- Key point: both Y and $g(X) = E(Y \mid X)$ are unbiased estimators of $E(Y)$, but the latter has smaller variance.
- In the Monte Carlo context, replacing a naive estimator with its conditional expectation is called *Rao–Blackwellization*.

Example: bivariate normal probabilities

- Consider computing $P(X > Y)$ where (X, Y) is a (standard) bivariate normal with correlation ρ .
- Naive: simulate (X_i, Y_i) and count instances where $X_i > Y_i$.
- However, the conditional distribution of X , given $Y = y$, is available, i.e., $X \mid (Y = y) \sim N(\rho y, 1 - \rho^2)$, so

$$h(y) := P(X > y \mid Y = y) = 1 - \Phi\left(\sqrt{\frac{1 - \rho}{1 + \rho}} y\right).$$

- R-B: simulate $Y_i \sim N(0, 1)$ and compute the mean of $h(Y_i)$.
- Comparison — $M = 10000$ samples with $\rho = 0.7$:

	naive	RB
[1,]	0.5012	0.4990414

- What about variances?

Example: hierarchical Bayes

- Consider the model $X_i \sim N(\theta_i, 1)$, independent, $i = 1, \dots, n$.
- Exchangeable/hierarchical prior:

$$\psi \sim \pi(\psi) \quad \text{and} \quad (\theta_1, \dots, \theta_n) \mid \psi \stackrel{\text{iid}}{\sim} N(0, \psi).$$

- Goal: compute the posterior mean $E(\theta_i \mid X)$, $i = 1, \dots, n$.
- Simple, if we could sample $(\theta_1, \dots, \theta_n)$ from the posterior.
- Rao–Blackwell approach based on the identity:

$$E(\theta_i \mid X_i, \psi) = \frac{\psi X_i}{\psi + 1}.$$

- Suggests that we can just take a sample $\psi^{(1)}, \dots, \psi^{(M)}$ from the posterior distribution of ψ , given X , and compute

$$\widehat{E(\theta_i \mid X)} = \frac{1}{M} \sum_{m=1}^M \frac{\psi^{(m)} X_i}{\psi^{(m)} + 1}, \quad i = 1, \dots, n.$$

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo
 - Stochastic approximation
 - Simulated annealing

- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo
 - Stochastic approximation
 - Simulated annealing

- We discussed a number of methods for root-finding, e.g.,
 - bisection
 - Newton's method.
- These methods are fast but require exact evaluation of the target function.
- Suppose the target function itself is not available, but we know how to compute a Monte Carlo estimate of it.
- How to find the root? — Newton's method won't work...

Stochastic approximation

- Suppose the goal is to find the root of a function f .
- However, $f(x)$ cannot be observed exactly — we can only measure $y = f(x) + \varepsilon$, where ε is a mean-zero random error.
- Stochastic approximation is a sort of stochastic version of Newton's method; idea is to construct a sequence of random variables that converges (probabilistically) to the root.
- Let $\{w_t\}$ be a vanishing sequence of positive numbers.
- Fix X_0 and define

$$X_{t+1} = X_t + w_{t+1}\{f(X_t) + \varepsilon_{t+1}\}, \quad t \geq 0.$$

Stochastic approximation (cont.)

- Intuition: Assume that f is monotone increasing...
- Can prove that, f has a unique root x^* and if w_t satisfies

$$\sum_{t=1}^{\infty} w_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} w_t^2 < \infty,$$

then $X_t \rightarrow x^*$ as $t \rightarrow \infty$ with probability 1.

- First studies by Robbins & Monro (1951).
- “Modern” theory uses a cute combination of probability theory (martingales) and stability theory of differential equations.⁵
- Has applications in statistical computing:
 - stochastic approximation–EM
 - stochastic approximation Monte Carlo
 - ...

⁵My first published paper has a nice (?) review of this stuff...

Stochastic approximation (cont.)

- How can the above formulation be applied in practice?
- Let $h(x, z)$ be a function of two variables, and suppose the goal is to find the root of

$$f(x) = E\{h(x, Z)\}, \quad \text{where } Z \sim P, \text{ with } P \text{ known.}$$

- If we can sample Z_t from P , then $h(X_t, Z_t)$ is an unbiased estimator of $f(X_t)$, given X_t .
- Therefore, run stochastic approximation as

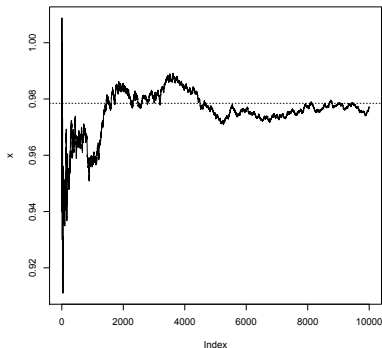
$$X_{t+1} = X_t + w_{t+1}h(X_t, Z_{t+1}), \quad Z_1, Z_2, \dots \stackrel{\text{iid}}{\sim} P.$$

Example: Student-t percentile

- Goal is to find the 100α th percentile of t_ν .
- Motivated by the scale-mixture form of Student-t, take

$$h(x, z) = \alpha - \Phi(x(z/\nu)^{-1/2}), \quad Z_t \sim \text{ChiSq}(\nu).$$

- Figure shows stochastic approximation output for $\nu = 3$, $\alpha = 0.8$, and $w_t = (1 + t)^{-0.75}$.



Example: exact confidence regions

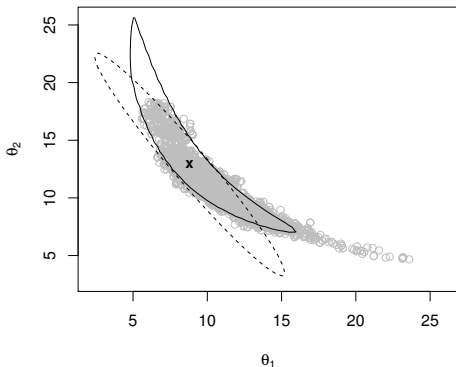
- First, consider testing $H_0 : \theta = \theta_0$.
- Let T_θ be a test statistic, and reject H_0 iff T_{θ_0} is too large.
- The p-value of the test is $p(\theta_0) = P_{\theta_0}(T_{\theta_0} > t_{\theta_0})$, where t_{θ_0} is the observed value of T_{θ_0} .
- An exact $100(1 - \alpha)\%$ confidence region for θ is

$$\{\theta : p(\theta) > \alpha\}.$$

- Can identify this region by solving the equation $p(\theta) = \alpha$.
- P-value is an expectation so falls in the category of problems that can be handled via stochastic approximation.
- Research problem: *how to do stochastic approximation efficiently here?*

Example: exact confidence regions (cont.)

- $n = 20$ real data points, modeled as $\text{Gamma}(\theta_1, \theta_2)$.
- Figure shows LRT p-value 10% contour.
- Compare with a Bayesian posterior sample and 90% confidence ellipse based on asymptotic normality of MLE.



- 1 Introduction
- 2 Basic Monte Carlo
- 3 Importance sampling
- 4 Rao–Blackwellization
- 5 Root-finding and optimization via Monte Carlo
 - Stochastic approximation
 - Simulated annealing

- We talked about a number of different methods for optimization, in particular, Newton's method.
- Required that the objective function have a derivative.
- But what if derivative isn't even defined?
- This is the case when the function domain is discrete.
- If the discrete space is small, then optimization is easy.
- What about if the discrete space is *huge*, so that it's not possible to enumerate all the function values?
- A Monte Carlo-based optimization method can be used here.

Simulated annealing

- Simulated annealing defines a random sequence of solutions x_t designed to target the global minimum of $f(x)$.
- The input variable x can be continuous, but let's focus on the discrete case.
- The idea is that at step $t + 1$, a point x_{new} is sampled from a distribution (possibly depending on t and x_t) and the new point x_{t+1} is either x_{new} or x_t , depending on the flip of a coin.
- The part about “flipping a coin” seems a bit strange but the purpose is to help the sequence avoid local minima.
- The key to the success of simulated annealing is a good choice of
 - *proposal distribution*, and
 - *cooling schedule*.

Simulated annealing algorithm

- 1 Specify a function $\tau(\cdot)$, a sequence of distributions $\Delta_t(\cdot)$, and a starting point x_0 . Set $t = 0$.
- 2 Sample $x_{\text{new}} \sim \Delta_t(x_t)$.
- 3 Calculate

$$\alpha = \min \left[1, \exp \left\{ \frac{f(x_t) - f(x_{\text{new}})}{\tau(t)} \right\} \right].$$

- 4 Flip a coin with prob α and set $x_{t+1} = x_{\text{new}}$ if the coin lands on Heads; otherwise, set $x_{t+1} = x_t$.
- 5 Set $t \leftarrow t + 1$ and return to Step 2.

For suitable $\Delta_t(\cdot)$ and $\tau(\cdot)$, x_t will tend (probabilistically) toward the global minimum of $f(x)$.

Example: variable selection in regression

- Consider a regression model with p predictor variables.
- There are a total of 2^p sub-models that one could consider and some are better than others – how to choose a good one?
- One objective criterion is to choose the model with the lowest Akaike Information Criterion (AIC)⁶.
- This is a problem of minimizing a function over a discrete set of indices — use simulated annealing.
- The `optim` routine in R: use `method=SANN` and the proposal is input as `gr` (the gradient).

⁶Basically the residual sum of squares plus a function increasing in the number of variables.

Example: variable selection in regression (cont.)

- For a given set of indices x_t , we sample x_{new} by choosing (essentially) at random to either add or delete an index. See the code for details.
- Just use the default cooling schedule in R.
- Code online uses a baseball data set from Givens & Hoeting.
- The goal is to identify a minimal collection of variables that explains the variability in player salaries.
- Run code to see how it goes...