STAT 7650 - Computational Statistics Lecture Slides

Monte Carlo Integration

Elvan Ceyhan

Updated: April, 2025

AU

- Based on parts of: Chapter 6 in Givens & Hoeting (Computational Statistics), Chapter 23 of Lange (Numerical Analysis for Statisticians), and Chapters 3-4 in Robert & Casella (Monte Carlo Statistical Methods).

Outline

Motivation

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

Notation

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

$$\mathbf{E}_f[h(X)] := \int_{\mathcal{X}} h(x) f(x)\, dx.$$

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

LLN and CLT

- 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 $\mathbf{E}_f|h(X)| < \infty$ and $X_1, X_2, \ldots$ are iid from f, then

  $$\overline{h_n(X)} := \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \mathbf{E}_f[h(X)], \text{ with probability 1.}$$

  CLT: If $\mathbf{E}_f[h^2(X)] < \infty$ and $X_1, X_2, \ldots$ are iid from f, then

  $$\sqrt{n}(\overline{h_n(X)} - \mathbf{E}_f[h(X)]) \to N(0, \sigma_f^2(h(X))), \text{ in distribution.}$$

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

Outline

Details on Estimation and Confidence Intervals

- Assume that we know how to sample from $f(x)$. Let $X_1, \ldots, X_n$ be a random sample from $f(x)$ and $Y_i = h(X_i)$ for $i = 1, 2, \ldots, n$.
- The LLN states that $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ should be a good estimate of $\mathbf{E}_f[h(X)]$ provided that n is large enough. It's an unbiased estimate for all n.
- If $\mathbf{E}_f[h^2(X)] < \infty$, then the CLT allows us to construct a confidence interval for $\mathbf{E}_f[h(X)]$ based on our sample.
- In particular, a $100(1-\alpha)\%$ Confidence Interval (CI) for $\mathbf{E}_f[h(X)]$ is

$$\bar{Y} \pm z_{\frac{1-\alpha}{2}} \times s_{\bar{Y}},$$

where $s_{\bar{Y}} := \frac{S_Y}{\sqrt{n}}$ is the sample standard error of $\bar{Y}$ with S_Y is sample standard deviation of $\{Y_1, \ldots, Y_n\}$.

Example: Estimating $\mathbf{E}[h(X)]$ with X from Uniform Distribution

Suppose $X \sim \text{Unif}(-\pi, \pi)$, so $f(x) = \frac{1}{2\pi}$ for $-\pi \leq x \leq \pi$.

The goal is to estimate $\mathbf{E}_f[h(X)]$, where $h(x) = \sin(x)$, which we know to be 0.

- Take an iid sample of size $n = 1000$ from the $\text{Unif}(-\pi, \pi)$ distribution and evaluate $Y_i = \sin(X_i)$.
- Summary statistics for the Y-sample include:
  - Mean $= \bar{Y} = -0.0167$
  - Standard deviation (SD) $= S_Y = 0.699$
- Then a 99% CI for $\mathbf{E}_f[h(X)]$ is calculated as:

$$-0.0167 \pm 2.576 \times \frac{0.699}{\sqrt{1000}} = [-0.074, 0.040]$$

Example: p-value of the Likelihood Ratio Test

- Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poi}(\lambda)$.
- Goal is to test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$.
- One idea is the likelihood ratio test, but the formula is messy:

$$\Lambda = \frac{L(\lambda_0)}{L(\widehat{\lambda})} = \exp\left(\sum X_i - n\lambda_0\right)\left(\frac{n\lambda_0}{\sum X_i}\right)^{\sum X_i}$$

  where $\widehat{\lambda} = \bar{X}$ (is the MLE).
- Need null distribution of the likelihood ratio statistic to compute, say, a p-value, but this is not directly available.[1]
- Straightforward to get a Monte Carlo p-value.
- Note that Λ depends only on the sufficient statistic $\sum X_i = n\bar{X}$, which is distributed as $\text{Poi}(n\lambda_0)$ under H_0.

[1] Wilks's theorem gives us a large-sample approximation...

Null Distribution of the Test Statistic

- The LRT statistic depends on the sufficient statistic $T = \sum X_i = n\bar{X}$.
- Under H_0, we have:

$$T \sim \text{Poi}(n\lambda_0).$$

- No closed-form for the distribution of Λ.
- For large n, Wilks's theorem implies:

$$-2 \log \Lambda \xrightarrow{d} \chi^2_1.$$

- For small or moderate n, use Monte Carlo simulation.

Monte Carlo p-value for LRT

- Simulate B datasets from $\text{Poi}(\lambda_0)$:

$$X_1^{(b)}, \ldots, X_n^{(b)} \sim \text{Poi}(\lambda_0), \quad b = 1, \ldots, B.$$

- Compute $\Lambda^{(b)}$ (or $-2 \log \Lambda^{(b)}$) for each dataset.
- Let $T_{\text{obs}} = -2 \log \Lambda_{\text{obs}}$ from original data.
- Monte Carlo p-value:

$$\widehat{p} = \frac{1}{B} \sum_{b=1}^{B} I\left(-2 \log \Lambda^{(b)} \geq T_{\text{obs}}\right).$$

Remarks on Monte Carlo Methods

Advantages of Monte Carlo:

- Does not depend on the dimension of the random variables.
- Basically works for all functions $h(x)$.
- 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.

Outline

Motivation for Importance Sampling

- Importance sampling techniques are useful in a number of situations; in particular:
  - When the target distribution $f(x)$ is difficult to sample from.
  - To reduce the variance of basic Monte Carlo estimates.
- The next slides give 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.
- This approach is similar in spirit to Sampling Importance Resampling (SIR)...

Motivating Example: Estimating Probabilities

- The goal is to estimate the probability that a fair die lands on 1. A basic Monte Carlo estimate $\widehat{X}$ based on n iid Ber $\left(\frac{1}{6}\right)$ samples has mean $\frac{1}{6}$ and variance $\frac{5}{36n}$.
- Changing the die to have three 1's and three non-1's increases the probability of observing a 1. To account for this, weight each 1 observed with the new die by $\frac{1}{3}$.
- So, if $Y_i = \frac{1}{3} \times \text{Ber}\left(\frac{1}{2}\right)$, then:
  - Expected value $\mathbf{E}(Y_i) = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
  - Variance $\text{Var}(Y_i) = \left(\frac{1}{3}\right)^2 \times \frac{1}{4} = \frac{1}{36}$
- Therefore, the new MC estimate, $\widehat{Y}$ has the same mean as $\widehat{X}$, but with a significantly smaller variance: $\frac{1}{36n}$ compared to $\frac{5}{36n}$.

Details on Importance Sampling

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

$$\mathbf{E}_f[h(X)] = \mathbf{E}_g[h(X)w^*(X)].$$

- This motivates the (modified) Monte Carlo approach:
  1. Sample $X_1, \ldots, X_n$ iid from $g(x)$.
  2. Estimate $\mathbf{E}_f[h(X)]$ with $\frac{1}{n}\sum_{i=1}^{n} h(X_i)w^*(X_i)$.
- If $f(x)$ is known only up to a proportionality constant, then use

$$\mathbf{E}_f[h(X)] \approx \sum_{i=1}^{n} h(X_i)w(X_i) = \frac{\sum_{i=1}^{n} h(X_i)w^*(X_i)}{\sum_{i=1}^{n} w^*(X_i)}$$

Importance Sampling to Estimate an Integral

- **Goal:** Estimate the integral

$$\int_0^1 \cos\left(\frac{\pi x}{2}\right) dx = \frac{2}{\pi} \approx 0.637.$$

- Interpret the integral as an expectation (with respect to Uniform distribution):

$$\mathbf{E}_f\left[\cos\left(\frac{\pi X}{2}\right)\right], \quad X \sim \text{Unif}(0,1)$$

since
$\int_0^1 \cos\left(\frac{\pi x}{2}\right) dx = \int_0^1 \cos\left(\frac{\pi x}{2}\right) \times f_X(x) dx = \int_0^1 \cos\left(\frac{\pi x}{2}\right) \times 1 dx.$

- Monte Carlo (MC) estimator:

$$\widehat{\mu}_{\text{MC}} = \frac{1}{n}\sum_{i=1}^{n} \cos\left(\frac{\pi X_i}{2}\right).$$

- It can be shown that $\text{Var}_f\left[\cos\left(\frac{\pi X}{2}\right)\right] \approx 0.095.$

Importance Sampling with Non-Uniform Proposal

- Use importance sampling with proposal density:

$$g(y) = \frac{3}{2}(1 - y^2), \quad y \in [0, 1].$$

- Importance sampling estimator:

$$\widehat{\mu}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(Y_i)}{g(Y_i)} h(Y_i), \quad Y_i \sim g,$$

  where $f(x) = 1$ on $[0, 1]$, and $h(x) = \cos\left(\frac{\pi x}{2}\right)$.

- So, the estimator simplifies to:

$$\widehat{\mu}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \cos\left(\frac{\pi Y_i}{2}\right)}{3(1 - Y_i^2)}.$$

Variance Reduction via Importance Sampling

- It can be shown that:

$$\mathrm{Var}_g \left[\frac{2\cos\left(\frac{\pi Y}{2}\right)}{3(1-Y^2)} \right] \approx 0.00099.$$

- This is significantly smaller than the variance under uniform sampling:

$$\frac{0.095}{0.00099} \approx 96.$$

- $\Rightarrow$ Importance sampling achieves over $95\times$ variance reduction.
- **Intuition:** $g(y)$ concentrates sampling where $\cos\left(\frac{\pi y}{2}\right)$ is large, improving efficiency.

Key Takeaways

- Importance sampling can significantly reduce variance when the proposal distribution aligns well with the shape of the integrand.
- Here, $g(y) = \frac{3}{2}(1 - y^2)$ closely matches the shape of $\cos\left(\frac{\pi y}{2}\right)$ on $[0, 1]$.
- Same principle as seen in the die example: reweighting improves precision.
- Variance reduction $\Rightarrow$ faster convergence for same n, or smaller n for same accuracy.

Another Motivating Example for IS (Cont'd)

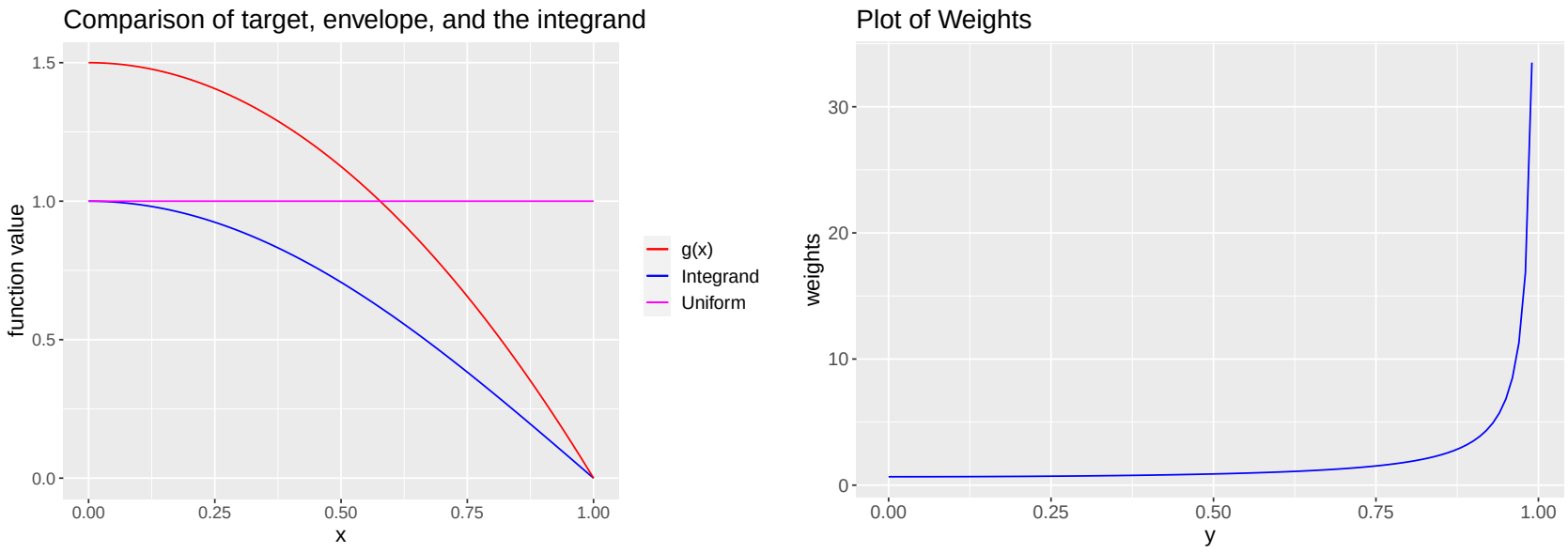

Figure 1: Left: The integrand function $\cos\left(\frac{\pi x}{2}\right)$, the envelope $g(x) = \frac{3}{2}(1 - x^2)$, and target density $f(x) = 1$. Right: This plot visualizes the weights $w(y) = f(y)/g(y) = 1/g(y)$ for $y \in [0, 0.99]$. The weights indicate the adjustment factor applied to each sampled value to estimate the integral using importance sampling.

Example: Size Estimation for a One-Sided Test

- Suppose $X_1, \ldots, X_{10} \overset{iid}{\sim} \text{Poi}(\lambda)$.
- We test:

$$H_0 : \lambda = 2 \quad \text{vs.} \quad H_1 : \lambda > 2.$$

- Use test statistic:

$$Z = \frac{\bar{X} - 2}{\sqrt{2/10}} = \frac{\bar{X} - 2}{\sqrt{0.2}}.$$

- Large-sample theory: Reject H_0 if $Z \geq 1.645$ for $\alpha = 0.05$.

Concern: Is the True Size Really 0.05?

- Poisson is discrete and skewed for small λ, and $n = 10$ is small.
- $\Rightarrow$ Normal approximation may not yield correct Type I error.
- Need to estimate the true size:

$$\alpha_{\text{true}} = P(Z \geq 1.645 | \lambda = 2).$$

- Two simulation approaches used:
  - Basic Monte Carlo sampling from Poi(2).
  - Importance sampling with proposal $g = \text{Poi}(2.4653)$.

Monte Carlo and Importance Sampling Estimators

Basic Monte Carlo:

- Generate B samples of size 10 from Poi(2).
- Compute Z and estimate $P(Z \geq 1.645)$ empirically.

Importance Sampling:

- Generate samples from $g = \text{Poi}(2.4653)$.
- Reweight using likelihood ratio:

$$w(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})} = \prod_{i=1}^{10} \frac{f(x_i)}{g(x_i)}.$$

- Estimate:

$$P_f(Z \geq 1.645) \approx \frac{1}{B} \sum_{b=1}^{B} w^{(b)} \cdot I(Z^{(b)} \geq 1.645).$$

Results: Estimated Type I Error Rates

- **Basic Monte Carlo:**

$$\widehat{\alpha}_{\text{MC}} = 0.0508, \quad 95\% \text{ CI}: \ (0.0465, 0.0551)$$

- **Importance Sampling:**

$$\widehat{\alpha}_{\text{IS}} = 0.0533, \quad 95\% \text{ CI}: \ (0.0520, 0.0611)$$

- Both methods suggest the actual size is slightly greater than 0.05.
- IS is more efficient: tighter confidence interval with fewer samples.

Key Takeaways

- Z-test assumes asymptotic normality of $\bar{X}$.
- For small n and non-normal distributions, the actual test size may differ from the nominal level.
- Simulation is a powerful tool to assess size and power empirically.
- Importance sampling is especially useful for tail probabilities.

Remarks on the Choice of Envelope $g(x)$

- The choice of envelope $g(x)$ is crucial. 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 ensure that $w^*(x)$ remains manageable.
- To get an idea of what makes a good proposal distribution, consider a practically useless result:

$$\text{“optimal” proposal} \propto |h(x)|f(x).$$

- **Take-away message:** We want the proposal distribution to resemble f, but we are less concerned in places where h is near or equal to zero.
- See Theorem 3.12 in Robert & Casella for more details.

Example: Estimating a Small Tail Probability

- **Goal:** Estimate the tail probability

$$p = P(Z > 4.5), \quad \text{where } Z \sim N(0, 1).$$

- Exact value:

$$p = 3.397673 \times 10^{-6}.$$

- This is a rare event: it occurs about 3–4 times in a million standard normal samples.

Naive Monte Carlo Estimation

- Sample $Z_1, \ldots, Z_M \overset{iid}{\sim} N(0,1)$.
- Estimate:

$$\widehat{p}_{\text{MC}} = \frac{1}{M} \sum_{j=1}^{M} I_{\{Z_j > 4.5\}}.$$

- With $M = 10{,}000$, we expect only 0.034 samples to exceed 4.5.
- So $\widehat{p}_{\text{MC}}$ is typically 0 — not a useful estimate.

Fixing the Problem with Importance Sampling

- Define $h(z) = I_{\{z>4.5\}}$.
- Then:

$$p = \mathbf{E}_{\varphi}[h(Z)] = \mathbf{E}_{g}\left[h(Z)\frac{\varphi(Z)}{g(Z)}\right],$$

  where φ is the standard normal density, and g is a proposal distribution.
- Since $h(z) = 0$ for $z \leq 4.5$, we only need to sample from g supported on $(4.5, \infty)$.

Choosing a Good Proposal Distribution

- Use a shifted exponential distribution as the proposal:

$$g(z) \propto e^{-(z-4.5)}, \quad z > 4.5.$$

- This concentrates all sampling effort in the region where $Z > 4.5$.
- Then the IS estimator becomes:

$$\widehat{p}_{\mathsf{IS}} = \frac{1}{M} \sum_{j=1}^{M} \frac{\varphi(Z_j)}{g(Z_j)}, \quad \text{where } Z_j \sim g \text{ and } Z_j > 4.5.$$

Results: Naive MC vs. Importance Sampling

For $M = 10{,}000$ samples:

Method	**Estimate**	**True Value**
Naive MC	0	3.397673×10^{-6}
IS	3.316521×10^{-6}	3.397673×10^{-6}

- Naive MC fails to detect rare events with small sample sizes.
- Importance Sampling gives accurate estimates by targeting the region of interest.

Key Takeaways

- Naive Monte Carlo is ineffective for estimating rare-event probabilities.
- Importance sampling improves efficiency by focusing sampling in the rare-event region.
- Careful choice of proposal distribution g is crucial for IS to work well.
- This technique is widely used in finance, reliability, and risk analysis.

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 (ESS)

$$N_{eff}(f, g) = \frac{n}{1 + s^2_{w*}}$$

  where s^2_{w*} is the sample variance of $\{w^*(X_1), \ldots, w^*(X_n)\}$.
- $N_{eff}(f, g)$ is bounded by n and measures approximately how many iid samples the weighted importance samples are worth.
- $N_{eff}(f, g)$ close to n indicates that $g(x)$ is a good proposal,
- Close to 0 means $g(x)$ is a poor proposal.

Outline

The Rao-Blackwell Theorem

- In statistics (e.g. in STAT 7610), the Rao-Blackwell Theorem provides a method for reducing the variance of an unbiased estimator by conditioning.
- The theorem is based on two simple formulas:

$$\mathbf{E}(Y) = \mathbf{E}[\mathbf{E}(Y|X)]$$

$$\text{Var}(Y) = \mathbf{E}[\text{Var}(Y|X)] + \text{Var}[\mathbf{E}(Y|X)] \geq \text{Var}[\mathbf{E}(Y|X)]$$

- **Key point:** Both Y and $h(X) = \mathbf{E}(Y|X)$ are unbiased estimators of $\mathbf{E}(Y) = \mu_Y$, but the latter has a smaller variance.
- In the Monte Carlo context, replacing a naive estimator with its conditional expectation is referred to as *Rao-Blackwellization*.

Example: Bivariate Normal Probabilities

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

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

Example: Bivariate Normal Probabilities

- **Rao-Blackwellization:** Simulate $X_i \sim N(0, 1)$ and compute the mean of $h(X_i)$.
- **Comparison:** $M = 10,000$ **samples with** $\rho = 0.7$**:**

Method	**Estimate**
Naive	0.5012
Rao-Blackwell	0.4990414

- **What about variances?**

Example: Hierarchical Bayesian Model

Consider the model $X_i \sim N(\mu_i, 1)$, independent, $i = 1, \ldots, n$.

- **Exchangeable/Hierarchical Prior:**

$$\mu_i|\tau \overset{iid}{\sim} N(0, \tau) \text{ for } i = 1, 2, \ldots, n \quad \text{and} \quad \tau \sim \pi(\tau).$$

- **Goal:** Compute the posterior mean $\mathbf{E}(\mu_i|\mathbf{X} = \mathbf{x})$, $i = 1, \ldots, n$ for $\mathbf{x} = (x_1, \ldots, x_n)$. Simple, if we could sample $(\mu_1, \ldots, \mu_n)$ from the posterior.
- **Rao-Blackwell Approach** based on the identity:

$$\mathbf{E}(\mu_i|x_i, \tau) = \frac{\tau x_i}{\tau + 1}.$$

- Suggests that we can just take a sample $\tau_1, \ldots, \tau_M$ from the posterior distribution of τ, given $\mathbf{X} = \mathbf{x}$, and compute

$$\widehat{\mathbf{E}(\mu_i|\mathbf{x})} = \frac{1}{M} \sum_{m=1}^{M} \frac{\tau_m x_i}{\tau_m + 1}, \; i = 1, \ldots, n.$$

Hierarchical Bayesian Model (Setup Details)

- Consider independent observations $X_i \sim N(\mu_i, 1)$, for $i = 1, \ldots, n$.
- That is, the data model is: $X_i | \mu_i \sim N(\mu_i, 1)$.
- **Hierarchical Prior:**
  - First Level: $\mu_i | \tau \overset{iid}{\sim} N(0, \tau)$ for $i = 1, 2, \ldots, n$.
  - Second Level: Prior $\pi(\tau)$ for hyperparameter τ; i.e., $\tau \sim \pi(\tau)$.
- **Goal:** Compute posterior mean $\mathbf{E}(\mu_i | \mathbf{X} = \mathbf{x})$, requires (computationally difficult) integration over τ and μ_i.

Hierarchical Bayesian Model — RB Approach and Derivation

- **Rao-Blackwell Theorem:** Use to reduce variance of estimates by conditioning.
- **Derivation of $\mathbf{E}(\mu_i|x_i,\tau)$:**
  - Posterior of μ_i given $X_i = x_i$ and τ: By Bayes' Theorem, combining the likelihood and the prior yields a posterior distribution for μ_i that is also normal.
  - That is, the likelihood function is $L(\mu_i|x_i) \propto \exp\left(-\frac{1}{2}(x_i-\mu_i)^2\right)$, and the prior for μ_i is $p(\mu_i|\tau) \propto \exp\left(-\frac{\mu_i^2}{2\tau}\right)$.
  - Multiplying the likelihood by the prior (omitting normalization constants) and completing the square in μ_i, we find that the posterior for μ_i is $\mu_i|x_i,\tau \sim N\left(\frac{\tau x_i}{\tau+1}, \frac{\tau}{\tau+1}\right)$.

Hierarchical Bayesian Model — RB Approach and Derivation

- **Expected Value $\mathbf{E}(\mu_i|x_i,\tau)$:** From the posterior distribution of μ_i, the expected value (mean) is directly the mean of the posterior normal distribution: $\dfrac{\tau x_i}{\tau+1}$.
- **Final Estimation with Rao-Blackwell:** To estimate $\mathbf{E}(\mu_i|\mathbf{X}=\mathbf{x})$, we first obtain samples $\tau_1,\ldots,\tau_M$ from the posterior distribution of τ given $\mathbf{x}$. Then, we compute the estimate of $\mathbf{E}(\mu_i|\mathbf{x})$ by averaging over these samples:

$$\widehat{\mathbf{E}(\mu_i|\mathbf{x})} = \frac{1}{M}\sum_{m=1}^{M}\frac{\tau_m x_i}{\tau_m+1}.$$

- This approach leverages the Rao-Blackwell Theorem to efficiently compute the posterior mean of μ_i by conditioning on τ and using the derived expression for $\mathbf{E}(\mu_i|x_i,\tau)$.

How to Find the Posterior Distribution of τ Given X?

Objective: Determine $\pi(\tau|\mathbf{X} = \mathbf{x})$.

1. **Joint Distribution of X and $\boldsymbol{\mu}$ Given τ:**

$$p(\mathbf{x}, \boldsymbol{\mu}|\tau) = \prod_{i=1}^{n} (N(x_i|\mu_i, 1) \times N(\mu_i|0, \tau))$$

2. **Marginal Likelihood of X Given τ:**

$$p(\mathbf{x}|\tau) = \int p(\mathbf{x}, \boldsymbol{\mu}|\tau) d\boldsymbol{\mu}.$$

   Generally not solvable in closed form.

3. **Posterior of τ:** $\pi(\tau|\mathbf{x}) \propto p(\mathbf{x}|\tau)\pi(\tau)$. Computation of $\pi(\tau|\mathbf{x})$ may require numerical integration or sampling (i.e. simulation methods like MCMC) due to the lack of a closed-form solution for the marginal likelihood.

Outline

Outline

Root-Finding with Monte Carlo Estimates

- We discussed a number of methods for root-finding (in S2), including:
  - Bisection
  - Newton's Method
- These methods are fast but require exact evaluation of the target function.
- **Challenge:** Suppose the target function itself is not directly available, but instead, we can compute a Monte Carlo estimate of it.
- **Question:** How to find the root? Newton's method and other traditional root-finding techniques may not be directly applicable due to the stochastic nature of the estimates.

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 or observe $y = f(x) + \varepsilon$, where ε is a mean-zero random error.
- **Stochastic approximation** is a sort of a stochastic version of Newton's Method; the 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

$$x_{t+1} = x_t + w_{t+1} \times (f(x_t) + \varepsilon_{t+1}), \quad t = 0, 1, 2, \ldots$$

Stochastic Approximation (Continued)

- **Intuition:** Assume that f is monotone increasing and 0 is in the target set (or codomain) of f...
- It can be proven 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 $x_t \to x^*$ as $t \to \infty$ with probability 1.
- First studied by Robbins & Monro (1951).
- **Modern theory** employs a blend of probability theory (martingales) and stability theory of differential equations.
- **Applications in Statistical Computing:**
  - Stochastic Approximation-EM (SA-EM)
  - Stochastic Approximation Monte Carlo (SAMC)
  - ...

Stochastic Approximation - Convergence Criteria

Key Conditions for Convergence

For a sequence of approximations x_t to converge to the true root x^* of f, the weights w_t used in the approximation must satisfy two conditions:

- The sum of the weights over all iterations tends to infinity: $\sum_{t=1}^{\infty} w_t = \infty$. This ensures that every point in the space is given infinite attention over time, allowing for the exploration of the function's behavior thoroughly.
- The sum of the squares of the weights is finite: $\sum_{t=1}^{\infty} w_t^2 < \infty$. This condition guarantees that the sequence of updates does not oscillate indefinitely, allowing the approximation process to stabilize and converge.

Stochastic Approximation (Continued)

- **Practical Application:** How can the stochastic approximation be applied in practice?
- Let $h(x, z)$ be a function of two variables, and suppose the goal is to find the root of

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

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

$$x_{t+1} = x_t + w_{t+1} \times h(x_t, Z_{t+1}); \quad Z_1, Z_2, \ldots \overset{iid}{\sim} F_Z.$$

Quantile Estimation for Student-t Distribution

- **Goal:** Estimate the 100α-th percentile of T_ν, i.e., find x_α such that
$$P(T_\nu \leq x_\alpha) = \alpha.$$
- **Challenge:** No closed-form expression for percentiles of t-distributions (except for special ν).

Scale-Mixture Representation of T_ν

- A Student-t variable can be written as
$$T_\nu = \frac{X}{\sqrt{Z/\nu}} = X\sqrt{\frac{\nu}{Z}},$$
where $X \sim N(0,1)$ and $Z \sim \chi^2_\nu$, independent.
- This leads to a representation of $P(T_\nu \leq x)$ via expectations over Z.

Quantile Condition as an Expectation

- Observe that $P(T_\nu \leq x) = \mathbf{E}_Z \left[\Phi\left(x\sqrt{\frac{Z}{\nu}} \right) \right].$
- So the quantile condition becomes:

$$P(T_\nu \leq x_\alpha) = \alpha \quad \Rightarrow \quad \mathbf{E}_Z \left[\Phi\left(x_\alpha \sqrt{\frac{Z}{\nu}} \right) \right] = \alpha.$$

Formulation as a Root-Finding Problem

- Define

$$h(x, z) = \alpha - \Phi\left(x\sqrt{\frac{z}{\nu}} \right).$$

  Then

$$f(x) = \mathbf{E}[h(x, Z)] = \alpha - \mathbf{E} \left[\Phi\left(x\sqrt{\frac{Z}{\nu}} \right) \right].$$

- The target is to find x^* such that $f(x^*) = 0$.

Stochastic Approximation Algorithm

- Let $Z_t \overset{iid}{\sim} \chi^2_\nu$ and use Robbins-Monro update:

$$x_{t+1} = x_t + w_{t+1} \cdot h(x_t, Z_{t+1}).$$

- For example, choose

$$w_t = (1+t)^{-0.75}, \quad t \geq 1.$$

- Then $x_t \to x_\alpha$ almost surely under standard assumptions.

Convergence Example

- Example parameters: $\nu = 3$, $\alpha = 0.8$, $x_0 = 1$, $w_t = (1+t)^{-0.75}$.

Example: Student-t Percentile

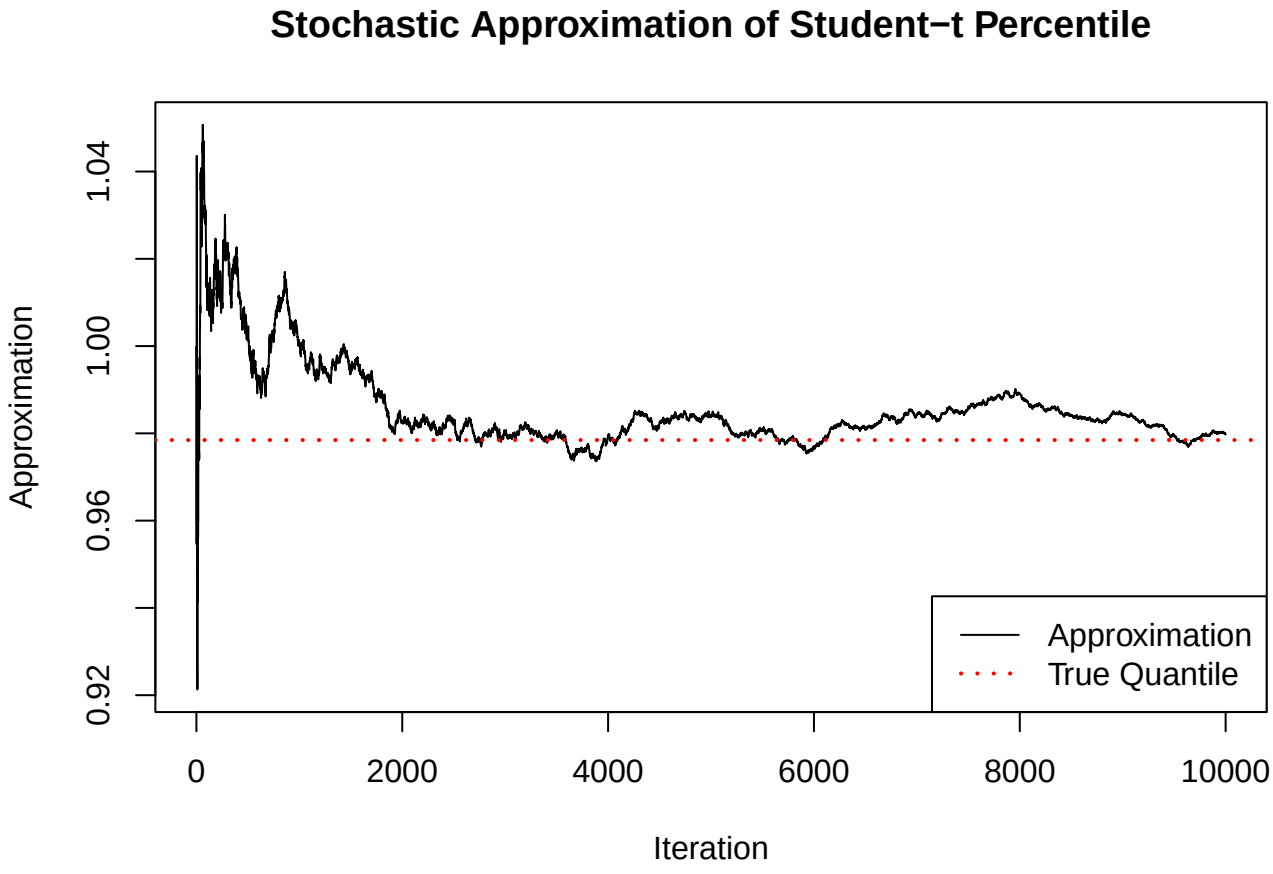

Figure 2: The convergence of x_t to the 100α-th percentile over iterations, indexed from 0 to 10,000, with x_t values ranging from 0.92 to 1.06.

Key Takeaways

- This method generalizes quantile estimation to settings where the CDF is intractable.
- Leverages a known scale-mixture representation and simulation-based approximation.
- Stochastic approximation offers a flexible and convergent algorithm.

Exact Confidence Regions via Hypothesis Test Inversion

- Consider testing $H_0 : \theta = \theta_0$ with a test statistic T_θ.
  - Reject H_0 if T_{θ_0} is too large.
  - Define p-value:

$$p(\theta_0) = P_{\theta_0}(T_{\theta_0} > t_{\text{obs}}).$$

  - Exact $(1 - \alpha) \times 100\%$ confidence region:

$$\mathcal{C}_\alpha = \{\theta : p(\theta) > \alpha\}.$$

- To construct $\mathcal{C}_\alpha$, solve $p(\theta) = \alpha$.

When is Stochastic Approximation Needed?

- In many problems, $p(\theta)$ cannot be computed analytically.
- Instead:

$$p(\theta) = \mathbf{E}_\theta \left[h(\theta, Z) \right], \quad .$$

  where $h(\theta, Z) = I_{\{T_\theta(Z) > t_{\text{obs}}\}}$ and $Z \sim P_\theta$.
- Goal: Solve

$$f(\theta) = p(\theta) - \alpha = 0.$$

- Use Robbins-Monro stochastic approximation:

$$\theta_{t+1} = \theta_t - w_{t+1} \cdot \left(h(\theta_t, Z_{t+1}) - \alpha \right),$$

  where $Z_{t+1} \sim P_{\theta_t}$.

Gamma Model Example: Likelihood Ratio-Based Confidence Region

Model Setup:

- Let $X_1, \ldots, X_{20} \overset{iid}{\sim} \text{Gamma}(\theta_1, \theta_2)$.
- Test $H_0 : (\theta_1, \theta_2) = (7, 13)$ using the Likelihood Ratio Test (LRT).
- Define the test statistic $T_\theta = -2 \log \left(\frac{L(\theta)}{L(\hat{\theta})} \right)$, where $\hat{\theta}$ is the MLE.

Gamma Model Example: Likelihood Ratio-Based Confidence Region

Computational Procedure:

- For each candidate $\theta = (\theta_1, \theta_2)$:
  - Generate B samples $Z^{(b)} \sim P_\theta$, compute $T_\theta^{(b)} = T_\theta(Z^{(b)})$.
  - Estimate the p-value via Monte Carlo:

$$\widehat{p}(\theta) = \frac{1}{B} \sum_{b=1}^{B} I_{\{T_\theta^{(b)} > t_{\text{obs}}\}}.$$

  - Solve $\widehat{p}(\theta) = \alpha$ using stochastic approximation to trace the boundary $\partial \mathcal{C}_\alpha$.

Visualization and Comparison

- **Plot contents:**
  - LRT p-value contour where $p(\theta) = 0.1$ (i.e., boundary of the 90% confidence region),
  - Bayesian posterior samples from prior over (θ_1, θ_2),
  - Asymptotic 90% confidence ellipse based on MLE normal approximation.
- **Interpretation:**
  - Posterior cloud aligns with, but not identical to, the exact region.
  - Asymptotic ellipse may underestimate or misalign in small samples.

Example: Exact Confidence Regions (Cont.)

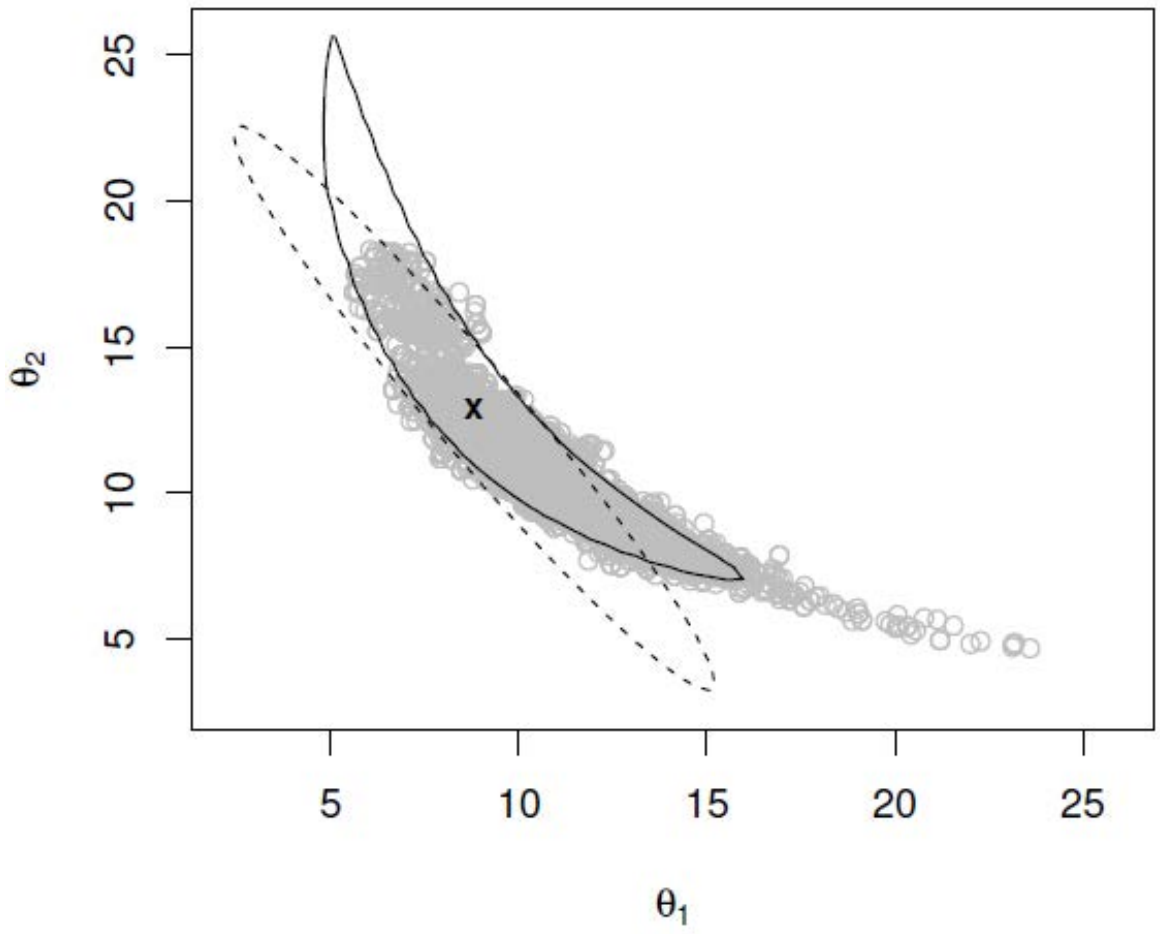

Figure 3: The plot showing the LRT p-value contour along with the Bayesian posterior sample and the confidence ellipse. The axes represent θ_1 and θ_2, ranging from 5 to 25.

Takeaways

- Inverting hypothesis tests yields exact confidence regions.
- When $p(\theta)$ is intractable, stochastic approximation enables efficient root-finding.
- The method is broadly applicable in simulation-based inference and non-standard models.
- Useful for comparing exact, Bayesian, and asymptotic confidence regions.

Outline

Optimization without Derivatives

We discussed various methods for optimization, particularly, Newton's Method, which requires the objective function to have a derivative.

- **Challenge:** What if the derivative isn't even defined? This scenario arises when the function domain is discrete.
- **Simple Case:** If the discrete space is small, then optimization is straightforward.
- **Complex Case:** What about when the discrete space is large, making it impractical to enumerate all function values?
- *Solution:* A Monte Carlo-based optimization method can be employed in these cases.

Simulated Annealing: Towards the Global Minimum

- Simulated annealing strategically explores the solution space to find the global minimum of a function $f(x)$, inspired by the metallurgical process of annealing.
- Though applicable to continuous variables, our discussion centers on discrete variables for clarity.
- **Decision Process at Each Step ($t+1$):**
  - Generate a candidate x_{new} in a neighborhood of x_t from a distribution (possibly depending on t and x_t)
  - Accept x_{new} with a probability, based on the flip of a (biased) coin, that favors lower $f(x)$ but allows for occasional increases to escape local minima.
- **Keys to Success:**
  1. A well-chosen proposal distribution.
  2. An effective cooling schedule.

"Cooling" in Simulated Annealing

- **Origin of the Term:** "Cooling" is inspired by the metallurgical process of annealing, where controlled cooling reduces defects in materials, enhancing their properties.
- **Analogy in Optimization:** In simulated annealing, "temperature" controls the acceptance of solutions, facilitating exploration of the solution space.
  - *High Temperature:* Encourages exploration, allowing acceptance of suboptimal solutions to avoid local minima.
  - *Cooling Down:* Gradually focuses the search on promising regions by being more selective in accepting solutions.
- **Cooling Schedule:** A strategy for decreasing temperature over time, balancing between diversification (exploration) and intensification (exploitation) to target the global minimum.

Simulated Annealing Algorithm

1. Specify a function $\beta(t)$, a sequence of distributions $\pi_t(x)$, and a starting point x_0. Set $t = 0$.
2. Sample $x_{\text{new}} \sim \pi_t(x_t)$.
3. Calculate
$$\alpha = \min\left\{1, \exp\left(\frac{f(x_t) - f(x_{\text{new}})}{\beta(t)}\right)\right\}.$$
4. Flip a coin with probability α of showing Heads and set $x_{t+1} = x_{\text{new}}$ if the coin lands on Heads; otherwise, set $x_{t+1} = x_t$.
5. Set $t = t + 1$ and return to Step 2.
6. Repeat until convergence.

For suitable $\pi_t(x)$ and $\beta(t)$, x_t will tend (probabilistically) toward the global minimum of $f(x)$.

About the Cooling Function in the SA Algorithm

- In the above algorithm, the cooling function is represented by $\beta(t)$, which is crucial for the algorithm's performance.
- It controls the "temperature" parameter in the simulated annealing algorithm, dictating how the acceptance probability for new solutions changes over time.
- As t (often representing time or iteration number) increases, $\beta(t)$ typically increases in simulated annealing contexts, effectively lowering the"temperature."
- The gradual cooling ensures a transition from a broad, explorative search across the solution space to a more focused, exploitative search around promising areas, aiming to converge to the global minimum of the objective function $f(x)$.
- The cooling function must balance between sufficient exploration at high temperatures and efficient exploitation as the temperature decreases.

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?
- **Objective Criterion:** One method is to choose the model with the lowest Akaike Information Criterion (AIC).[2]
- This becomes a problem of minimizing a function over a discrete set of indices — perfect for **simulated annealing**.
- **Implementation in R:**
  - Use the `optim` routine with `method="SANN"`.
  - The proposal function is input as `gr` (the gradient).
  - The "`SANN`" method in `optim` has an internal cooling schedule that is not directly adjustable via the function's parameters.

[2] AIC balances the residual sum of squares with a penalty for the number of variables.

Variable Selection in Regression (Cont.)

- For a given set of indices x_t, we sample x_{new} by choosing at random to either add or delete an index.
- **Procedure:**
  - Sample changes to the model indices essentially at random.
  - Use the default cooling schedule in `R` for simulated annealing.
- Code available on Canvas uses a baseball data set from Givens & Hoeting to identify a minimum collection of variables explaining the variability in baseball player salaries.
- Run the code to observe the variable selection process via simulated annealing.