Monte Carlo Integration

Statistics · Module 9 — Computational Statistics

Node 9.2

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.2

Monte Carlo integration converts the analytical problem of evaluating an integral into the statistical problem of estimating a mean. It inherits all the tools of statistical inference – the law of large numbers for consistency, the central limit theorem for error quantification, and variance reduction techniques for efficiency. This node develops the method, its theoretical foundations, and the principal variance reduction strategies.

Formal Definition

Let $h : \mathcal{X} \to \mathbb{R}$ be a measurable function and P a probability distribution on $\mathcal{X}$. The target is

$$I = \mathbb{E}_P[h(X)] = \int h(x)\, dP(x).$$

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, the **Monte Carlo estimator** is

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

By the Strong Law of Large Numbers, $\hat{I}_n \overset{a.s.}{\to} I$ provided $\mathbb{E}_P[|h(X)|] < \infty$. By the Central Limit Theorem,

$$\sqrt{n}(\hat{I}_n - I) \overset{d}{\to} \mathcal{N}(0, \sigma_h^2), \quad \sigma_h^2 = \mathrm{Var}_P(h(X)),$$

so the Monte Carlo standard error is $\mathrm{SE}(\hat{I}_n) = \sigma_h / \sqrt{n}$.

Intuition

The convergence rate $n^{-1/2}$ is dimension-free. Deterministic quadrature in d dimensions converges at rate $n^{-r/d}$ for r-smooth integrands, suffering the curse of dimensionality. Monte Carlo does not: the rate is $n^{-1/2}$ whether $d = 1$ or $d = 10{,}000$. The constant σ_h depends on the integrand, but the rate does not depend on the dimension of the sample space. This makes Monte Carlo the default method for high-dimensional integration in Bayesian statistics, statistical physics, and finance.

Structural Role

Monte Carlo integration is the foundation for MCMC (Node 9.3), which extends the idea to non-i.i.d. samples from intractable distributions. It underlies the bootstrap (Node 9.4), which simulates the sampling distribution by resampling. Bayesian predictive inference (Node 6.5) uses Monte Carlo averages to compute posterior predictive quantities. The variance reduction techniques developed here reappear in importance-weighted MCMC and particle filtering.

Mathematical Development

Importance Sampling

When direct sampling from P is impossible or inefficient, let Q be a **proposal distribution** with $P \ll Q$. Define the importance weight $w(x) = p(x)/q(x)$. Then

$$I = \int h(x) \frac{p(x)}{q(x)} q(x)\, dx = \mathbb{E}_Q[h(X)w(X)].$$

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q$, the **importance sampling estimator** is

$$\hat{I}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i) w(X_i).$$

This is unbiased for I with variance

$$\text{Var}_Q(\hat{I}_n^{\text{IS}}) = \frac{1}{n} \text{Var}_Q(h(X)w(X)) = \frac{1}{n} \left[\int \frac{(h(x)p(x))^2}{q(x)}\, dx - I^2 \right].$$

Optimal proposal. The variance-minimizing proposal is $q^*(x) \propto |h(x)|p(x)$, which gives zero variance when $h \geq 0$. In practice, q^* is unknown (it requires the normalizing constant of $|h|p$), but it guides the design of good proposals: q should be large where $|h(x)|p(x)$ is large.

Self-Normalized Importance Sampling

When p is known only up to a normalizing constant, $p(x) = \tilde{p}(x)/Z$, define unnormalized weights $\tilde{w}(x) = \tilde{p}(x)/q(x)$. The **self-normalized estimator** is

$$\hat{I}_n^{\text{SN}} = \frac{\sum_{i=1}^n h(X_i)\tilde{w}(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)}.$$

This is biased but consistent, with bias $O(n^{-1})$. Its asymptotic variance depends on the **effective sample size**:

$$\text{ESS} = \frac{\left(\sum_{i=1}^n \tilde{w}_i\right)^2}{\sum_{i=1}^n \tilde{w}_i^2} \approx \frac{n}{1 + \text{Var}_Q(w)}.$$

When $Q = P$, all weights are equal and $\text{ESS} = n$. When Q is a poor match, a few weights dominate and $\text{ESS} \ll n$.

Control Variates

Let $g : \mathcal{X} \to \mathbb{R}$ with known expectation $\mu_g = \mathbb{E}_P[g(X)]$. The **control variate estimator** is

$$\hat{I}_n^{\mathrm{CV}} = \frac{1}{n} \sum_{i=1}^{n} \left[h(X_i) - c(g(X_i) - \mu_g) \right].$$

This is unbiased for any c. The variance is

$$\mathrm{Var}(\hat{I}_n^{\mathrm{CV}}) = \frac{1}{n} \left[\sigma_h^2 - 2c \,\mathrm{Cov}(h, g) + c^2 \sigma_g^2 \right].$$

The optimal coefficient is $c^* = \mathrm{Cov}(h, g)/\mathrm{Var}(g)$, yielding variance reduction factor $1 - \rho_{hg}^2$, where ρ_{hg} is the correlation between h and g. The closer $|\rho_{hg}|$ is to 1, the greater the reduction.

Antithetic Variates

For symmetric distributions, generate pairs (X_i, X_i') that are negatively correlated (e.g., $X_i' = -X_i$ for a distribution symmetric about 0). The estimator

$$\hat{I}_n^{\mathrm{AV}} = \frac{1}{n} \sum_{i=1}^{n/2} \frac{h(X_i) + h(X_i')}{2}$$

has variance reduced by $\mathrm{Cov}(h(X), h(X'))/\mathrm{Var}(h(X))$, which is negative when h is monotone.

Stratified Sampling

Partition $\mathcal{X} = \bigcup_{j=1}^{J} A_j$ with $P(A_j) = p_j$. Draw n_j samples from $P(\cdot \mid A_j)$ and estimate

$$\hat{I}_n^{\mathrm{strat}} = \sum_{j=1}^{J} p_j \hat{I}_{n_j}^{(j)}, \quad \hat{I}_{n_j}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} h(X_i^{(j)}).$$

With proportional allocation $n_j = np_j$, the variance is

$$\mathrm{Var}(\hat{I}_n^{\mathrm{strat}}) = \frac{1}{n} \sum_{j=1}^{J} p_j \sigma_j^2 \leq \frac{\sigma_h^2}{n},$$

where $\sigma_j^2 = \mathrm{Var}(h(X) \mid X \in A_j)$. The inequality is strict whenever $\mathbb{E}[h(X) \mid A_j]$ varies across strata, because stratification eliminates the between-stratum variance component.

Proof of variance reduction. By the law of total variance,

$$\mathrm{Var}(h(X)) = \mathbb{E}[\mathrm{Var}(h \mid \mathrm{stratum})] + \mathrm{Var}(\mathbb{E}[h \mid \mathrm{stratum}]) = \sum_j p_j \sigma_j^2 + \sum_j p_j (\mu_j - I)^2,$$

where $\mu_j = \mathbb{E}[h(X) \mid A_j]$. Simple Monte Carlo has variance $\mathrm{Var}(h)/n$; stratified sampling has variance $\sum p_j \sigma_j^2 / n$. The difference is $\sum p_j (\mu_j - I)^2 / n \geq 0$. □

Optimal allocation. Neyman allocation sets $n_j \propto p_j \sigma_j$, minimizing the stratified variance to $(\sum_j p_j \sigma_j)^2 / n$, which is further reduced when strata have unequal variability.

CLT for Monte Carlo Estimators

The central limit theorem provides the basis for confidence intervals. For the standard Monte Carlo estimator:

$$P\left(I \in \left[\hat{I}_n - z_{\alpha/2}\frac{\hat{\sigma}_h}{\sqrt{n}},\ \hat{I}_n + z_{\alpha/2}\frac{\hat{\sigma}_h}{\sqrt{n}}\right]\right) \to 1-\alpha,$$

where $\hat{\sigma}_h^2 = (n-1)^{-1}\sum(h(X_i) - \hat{I}_n)^2$. The confidence interval width is $O(n^{-1/2})$ regardless of dimension, confirming the dimension-free convergence rate. For importance sampling, the CLT applies to $h(X)w(X)$, but the effective variance may be much larger or smaller than σ_h^2 depending on the proposal quality.

Worked Examples

Example 1: Estimating a Posterior Mean

Let $\theta \sim \text{Gamma}(3,1)$ (prior) and $X \mid \theta \sim \text{Poisson}(\theta)$ with $x = 7$ observed. The posterior is $\theta \mid x \sim \text{Gamma}(10,2)$. The posterior mean is $\mathbb{E}[\theta \mid x] = 10/2 = 5$. To verify by Monte Carlo: draw $\theta_1, \ldots, \theta_n \sim \text{Gamma}(10,2)$ and compute $\hat{I}_n = n^{-1}\sum\theta_i$. With $n = 10{,}000$, the standard error is $\sqrt{10/4}/100 = 0.0158$, giving a 95% confidence interval of approximately $(4.969, 5.031)$.

Example 2: Importance Sampling for Rare-Event Probability

Estimate $P(X > 5)$ where $X \sim \mathcal{N}(0,1)$. Direct Monte Carlo requires enormous n since $P(X > 5) \approx 2.87 \times 10^{-7}$. Use importance sampling with proposal $Q = \mathcal{N}(5,1)$:

$$w(x) = \frac{\phi(x)}{\phi(x-5)} = \exp\left(-5x + \frac{25}{2}\right).$$

The estimator $\hat{p} = n^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i > 5)w(X_i)$ with $X_i \sim \mathcal{N}(5,1)$ concentrates samples in the region of interest. With $n = 1{,}000$, approximately half the draws exceed 5, and the weighted average gives an estimate with coefficient of variation below 0.1, compared to a coefficient of variation exceeding 50 for naive Monte Carlo at the same n.

Cross-Links

- **Linear Algebra:** Stratified sampling exploits the decomposition of total variance into within-stratum and between-stratum components, mirroring the ANOVA decomposition.
- **AI:** Stochastic gradient descent is Monte Carlo estimation of the population risk gradient; mini-batch size controls the variance-computation tradeoff. Policy gradient methods in reinforcement learning use control variates (baselines) to reduce gradient variance.
- **Economics:** Monte Carlo simulation is the standard tool for pricing path-dependent derivatives in finance; importance sampling is used to estimate the probability of extreme portfolio losses (Value-at-Risk).
- **Quantum:** Quantum Monte Carlo methods estimate ground-state properties by sampling from the path integral representation; the sign problem is a variance explosion analogous to importance weight degeneracy.

Structural Compression

Invariant: The Monte Carlo estimator $\hat{I}_n$ is unbiased with variance $\mathrm{Var}(h)/n$, converging at rate $n^{-1/2}$ regardless of the dimension of the sample space.

Mechanism: The SLLN guarantees almost-sure convergence of sample averages to population means; the CLT characterizes the $n^{-1/2}$ Gaussian fluctuations. Variance reduction techniques modify the integrand or sampling scheme to decrease the constant σ_h^2 without changing the rate.

Cross-System Echo: The dimension-free convergence rate of Monte Carlo is structurally identical to the statistical principle that sample means converge at rate $n^{-1/2}$ regardless of the population distribution. In every system where high-dimensional integration appears – posterior computation in statistics, partition function estimation in physics, expected utility computation in economics – Monte Carlo is the universal computational primitive precisely because it circumvents the curse of dimensionality.

Exercises

1. Prove that the importance sampling estimator $\hat{I}_n^{\mathrm{IS}}$ is unbiased and derive its variance in terms of $\mathbb{E}_Q[(h(X)w(X))^2]$.
2. For $h(x) = x^2$ and $P = \mathcal{N}(0,1)$, compute the optimal importance sampling proposal $q^*(x) \propto x^2\phi(x)$. Identify the distribution and show that the resulting estimator has zero variance.
3. Derive the optimal control variate coefficient $c^* = \mathrm{Cov}(h,g)/\mathrm{Var}(g)$ and show that the variance reduction factor is $1 - \rho_{hg}^2$.
4. Show that stratified sampling with proportional allocation always reduces variance compared to simple random sampling, with equality iff the stratum means are all equal.
5. For self-normalized importance sampling, derive the $O(n^{-1})$ bias using a Taylor expansion of the ratio estimator, and show that the bias vanishes as $n \to \infty$.
6. Estimate $\int_0^1 e^x\,dx$ using Monte Carlo with $n = 1{,}000$ and the control variate $g(x) = x$ (with $\mu_g = 1/2$). Compare the standard error with and without the control variate.
7. Argue, structurally, why the dimension-free convergence rate of Monte Carlo integration represents the same principle as the dimension-free convergence of sample means in the CLT: both are consequences of the fact that averaging independent random variables reduces variance at rate $1/n$, and this rate depends on the number of observations, not the complexity of the underlying space.

Statistics · Module 9 — Computational Statistics · Node 9.2
Concepts: monte-carlo-methods, law-of-large-numbers, central-limit-theorem
Prerequisites: 1.3, 1.4, 7.1
Enables: 9.3, 9.4

hbar.university — own.your.cognition