

# Monte Carlo Methods

---

Juan F. Rubio-Ramírez

Emory University

- Solutions of many scientific problems involve intractable high-dimensional integrals.
- Standard deterministic numerical integration deteriorates rapidly with dimension.
- Monte Carlo methods are stochastic numerical methods to approximate high-dimensional integrals.
- Main application in this course: Bayesian statistics.

# Computing Integrals in 1D

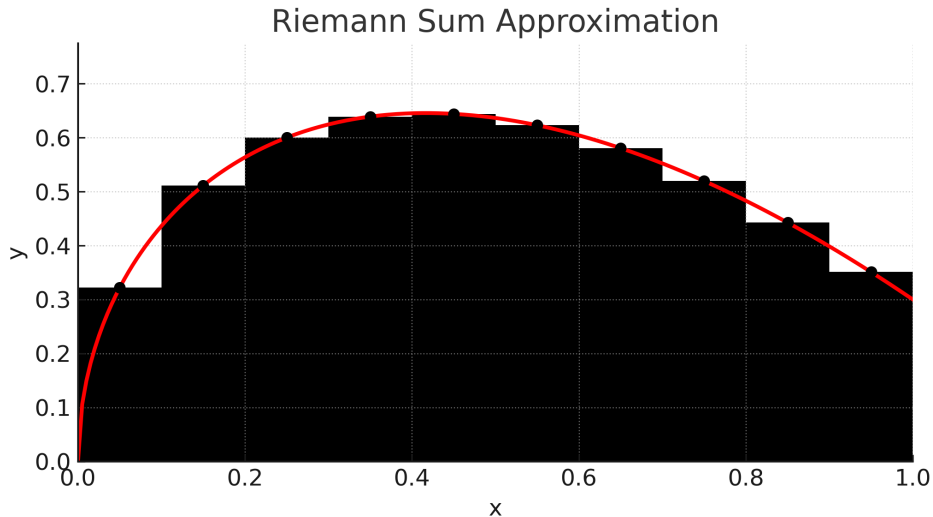
For  $f : X \rightarrow \mathbb{R}$ , let

$$I = \int_X f(x) dx.$$

When  $X = [0, 1]$ , approximate  $I$  via the midpoint Riemann sum

$$\hat{I}_n = \frac{1}{n} \sum_{i=0}^{n-1} f\left(\frac{i + \frac{1}{2}}{n}\right).$$

# Riemann Sums Figure



For a small interval  $[a, a + \varepsilon]$ ,

$$\int_a^{a+\varepsilon} f(x) dx \approx \varepsilon f(a).$$

We want to understand and bound the approximation error:

$$\text{Local error} = \left| \int_a^{a+\varepsilon} f(x) dx - \varepsilon f(a) \right|.$$

Then, we will use this local error to compute the global error.

## Step 1: Expressing $f(x)$ via its Derivative

By the Fundamental Theorem of Calculus,

$$f(x) = f(a) + \int_a^x f'(y) dy.$$

Substitute this expression into the integral:

$$\int_a^{a+\varepsilon} f(x) dx = \int_a^{a+\varepsilon} \left[ f(a) + \int_a^x f'(y) dy \right] dx.$$

## Step 2: Representing the Error as a Double Integral

Split the integral into two terms:

$$\int_a^{a+\varepsilon} f(x) dx = \underbrace{\int_a^{a+\varepsilon} f(a) dx}_{=\varepsilon f(a)} + \int_a^{a+\varepsilon} \int_a^x f'(y) dy dx.$$

Therefore,

$$\int_a^{a+\varepsilon} f(x) dx - \varepsilon f(a) = \int_a^{a+\varepsilon} \int_a^x f'(y) dy dx.$$

This expresses the integration error as the accumulated effect of the derivative  $f'(y)$  over the triangular region  $\{(x, y) : a \leq y \leq x \leq a + \varepsilon\}$  in the  $(x, y)$ -plane.

## Step 3: Bounding the Double Integral

Take absolute values and use the triangle inequality:

$$\left| \int_a^{a+\varepsilon} \int_a^x f'(y) \, dy \, dx \right| \leq \int_a^{a+\varepsilon} \int_a^x |f'(y)| \, dy \, dx.$$

Since  $|f'(y)| \leq \sup_{x \in [a, a+\varepsilon]} |f'(x)| =: M$ , we obtain

$$\leq M \int_a^{a+\varepsilon} \int_a^x 1 \, dy \, dx = M \int_a^{a+\varepsilon} (x - a) \, dx = M \frac{\varepsilon^2}{2}.$$

**Geometric intuition:** the inner integral  $\int_a^x 1 \, dy = x - a$  represents the width of a growing triangle, and integrating again gives the triangle's area  $\varepsilon^2/2$ .



## Step 4: From Local Error to Global Error

The previous bound applies to a single small interval  $[a, a + \varepsilon]$ :

$$\text{Local error} \leq M \frac{\varepsilon^2}{2}, \quad \text{where } M = \sup_{x \in [0,1]} |f'(x)|.$$

If we divide  $[0, 1]$  into  $n$  equal subintervals, then  $\varepsilon = 1/n$ .

$$\text{Local error} \leq \frac{M}{2n^2}.$$

There are  $n$  such intervals, so the total error satisfies

$$\text{Global error} \leq n \times \frac{M}{2n^2} = \frac{M}{2n} = O\left(\frac{1}{n}\right).$$

Thus, the midpoint (or naive) numerical integration rule converges at rate  $O(1/n)$ .

## Computing Integrals in 2D

Partition  $[0, 1]^2$  into an  $m \times m$  uniform grid with mesh  $h = 1/m$ . For cell midpoints

$$x_i = \frac{i + \frac{1}{2}}{m}, \quad y_j = \frac{j + \frac{1}{2}}{m}, \quad i, j = 0, \dots, m-1,$$

the 2D midpoint estimator with  $n = m^2$  evaluations is

$$\hat{I}_m = \frac{1}{m^2} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} f(x_i, y_j).$$

We show its error is  $O(1/m) = O(n^{-1/2})$ .

## Step 1: Local Error on One Cell (Analogy with 1D Case)

Consider one rectangular cell

$$R = [a, a + h] \times [b, b + h]$$

By the Fundamental Theorem of Calculus applied in both coordinates,

$$f(x, y) = f(a, b) + \int_a^x f_x(u, b) du + \int_b^y f_y(a, v) dv + \int_a^x \int_b^y f_{xy}(u, v) dv du.$$

$$\text{Local Error} = \left| \iint_R f(x, y) dx dy - h^2 f(a, b) \right|$$

We have that Local Error related to this:

$$\iint_R \left( \int_a^x f_x(u, b) du + \int_b^y f_y(a, v) dv + \int_a^x \int_b^y f_{xy}(u, v) dv du \right) dx dy.$$

## Step 2: Bounding the Local Error

Take absolute values and use the sup norm of the derivatives:

$$\begin{aligned}\text{Local Error} \leq & \|f_x\|_\infty \int_a^{a+h} \int_b^{b+h} (x-a) \, dx \, dy + \|f_y\|_\infty \int_a^{a+h} \int_b^{b+h} (y-b) \, dx \, dy \\ & + \|f_{xy}\|_\infty \int_a^{a+h} \int_b^{b+h} (x-a)(y-b) \, dx \, dy.\end{aligned}$$

Each term can be computed explicitly:

$$\int_a^{a+h} \int_b^{b+h} (x-a) \, dx \, dy = \frac{h^3}{2}, \quad \int_a^{a+h} \int_b^{b+h} (x-a)(y-b) \, dx \, dy = \frac{h^4}{4}.$$

Hence the local error is bounded by

$$\text{Local error on } R \leq M h^3, \quad M = C_1 \|f_x\|_\infty + C_2 \|f_y\|_\infty + C_3 \|f_{xy}\|_\infty.$$

### Step 3: From Local Error to Global Error

There are  $m^2$  cells, each with mesh  $h = 1/m$  and local error  $\leq Mh^3$ . Therefore the *global* error obeys

$$\text{Global error} \leq m^2 \times Mh^3 = m^2 \times M \left( \frac{1}{m} \right)^3 = \frac{M}{m}.$$

But  $n = m^2$ , thus

$$O\left(\frac{1}{m}\right) = O\left(n^{-1/2}\right).$$

## Extension to $d$ Dimensions

On  $[0, 1]^d$  with an  $m^d$  grid (mesh  $h = 1/m$ ,  $n = m^d$  points):

- The *local* midpoint error on one  $d$ -cube is  $O(h^{d+1})$  (one power of  $h$  from integrating each coordinate; an extra  $h$  from the derivative term).
- There are  $m^d$  cells.

Therefore

$$\text{Global error} \leq m^d \times O(h^{d+1}) = m^d \times O((1/m)^{d+1}) = O\left(\frac{1}{m}\right) = O(n^{-1/d}).$$

This is the *curse of dimensionality*: the rate degrades from  $O(n^{-1})$  (1D midpoint) to  $O(n^{-1/d})$  in  $d$  dimensions.

# Monte Carlo Integration

---

# Monte Carlo Integration

We are interested in computing

$$I = \int_{\mathcal{X}} \varphi(x) \pi(x) dx$$

where  $\pi$  is a pdf on  $\mathcal{X}$  and  $\varphi : \mathcal{X} \rightarrow \mathbb{R}$ .

**Monte Carlo method:** draw  $n$  i.i.d. samples  $X_1, \dots, X_n \sim \pi$  and compute

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

**Remark:** this corresponds to the empirical measure

$$\hat{\pi}_n(dx) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(dx).$$



**Law of Large Numbers.** If  $\mathbb{E}|\varphi(X)| < \infty$ , then  $\hat{I}_n \rightarrow I$  a.s.

**Central Limit Theorem.** If  $\sigma^2 = \text{Var}(\varphi(X)) = \int_{\mathcal{X}} [\varphi(x) - I]^2 \pi(x) dx < \infty$ , then

$$\text{Var}(\hat{I}_n) = \mathbb{E}[(\hat{I}_n - I)^2] = \frac{\sigma^2}{n} \text{ and, hence } \sqrt{n} \frac{\hat{I}_n - I}{\sigma} \Rightarrow \mathcal{N}(0, 1).$$

**Proposition.** If  $\sigma^2 = \text{Var}(\varphi(X)) < \infty$ , then

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (\varphi(X_i) - \hat{I}_n)^2$$

is an unbiased estimator of  $\sigma^2$ .

Let  $Y_i = \varphi(X_i)$ ,  $i = 1, \dots, n$ , be i.i.d. with

$$\mu = \mathbb{E}[Y], \quad \sigma^2 = \text{Var}(Y) < \infty, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

The sample variance is

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

**Goal:** show that  $\mathbb{E}[S_n^2] = \text{Var}(Y) = \sigma^2$ .

## Step 1. Algebraic Identity

Expand the deviations:

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_i (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2) \\ &= \sum_i Y_i^2 - 2\bar{Y} \sum_i Y_i + n\bar{Y}^2.\end{aligned}$$

Since  $\sum_i Y_i = n\bar{Y}$ ,

$$\boxed{\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2.}$$

This identity will let us compute  $\mathbb{E}[S_n^2]$  directly.

## Step 2. Take Expectations

By definition and linearity of expectation,

$$\begin{aligned}\mathbb{E}[S_n^2] &= \frac{1}{n-1} \mathbb{E} \left[ \sum_i (Y_i - \bar{Y})^2 \right] \\ &= \frac{1}{n-1} \left( \mathbb{E} \left[ \sum_i Y_i^2 \right] - n \mathbb{E}[\bar{Y}^2] \right).\end{aligned}$$

Because the  $Y_i$  are i.i.d.,

$$\mathbb{E} \left[ \sum_i Y_i^2 \right] = n \mathbb{E}[Y^2],$$

so

$$\mathbb{E}[S_n^2] = \frac{1}{n-1} (n \mathbb{E}[Y^2] - n \mathbb{E}[\bar{Y}^2]).$$

### Step 3. Evaluate $\mathbb{E}[\bar{Y}^2]$

Use  $\mathbb{E}[Z^2] = \text{Var}(Z) + (\mathbb{E}Z)^2$  with  $Z = \bar{Y}$ :

$$\mathbb{E}[\bar{Y}^2] = \text{Var}(\bar{Y}) + (\mathbb{E}[\bar{Y}])^2.$$

Since  $\mathbb{E}[\bar{Y}] = \mu$  and, for i.i.d. draws,

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n} \sum_i Y_i\right) = \frac{1}{n^2} \sum_i \text{Var}(Y_i) = \frac{\sigma^2}{n},$$

we obtain

$$\boxed{\mathbb{E}[\bar{Y}^2] = \frac{\sigma^2}{n} + \mu^2.}$$

## Step 4. Substitute and Simplify

Plug  $\mathbb{E}[Y^2] = \sigma^2 + \mu^2$  and  $\mathbb{E}[\bar{Y}^2] = \frac{\sigma^2}{n} + \mu^2$  into

$$\mathbb{E}[S_n^2] = \frac{1}{n-1} (n \mathbb{E}[Y^2] - n \mathbb{E}[\bar{Y}^2]) :$$

$$\begin{aligned} \mathbb{E}[S_n^2] &= \frac{1}{n-1} \left[ n(\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \right] \\ &= \frac{1}{n-1} ((n-1)\sigma^2) = \boxed{\sigma^2 = \text{Var}(Y) = \text{Var}(\varphi(X))}. \end{aligned}$$

Thus  $S_n^2$  is an *unbiased* estimator of the population variance.

**Chebyshev Inequality:**

$$\mathbb{P}\left(|\hat{I}_n - I| > c \frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{c^2}.$$

**Central Limit Theorem Approximation:**

$$\mathbb{P}\left(|\hat{I}_n - I| > c \frac{\sigma}{\sqrt{n}}\right) \approx 2(1 - \Phi(c)).$$

**Confidence Interval:**

$$\hat{I}_n \pm c_\alpha \frac{S_n}{\sqrt{n}}, \quad 2(1 - \Phi(c_\alpha)) = \alpha.$$

Rate of convergence:  $\mathcal{O}(n^{-1/2})$ .



## Chebyshev Inequality — Step by Step

We know that

$$\text{Var}(\hat{I}_n) = \mathbb{E}[(\hat{I}_n - I)^2] = \frac{\sigma^2}{n}.$$

Chebyshev's inequality states:

$$\mathbb{P}(|Z| > c) \leq \frac{\text{Var}(Z)}{c^2}.$$

Apply it to  $Z = \hat{I}_n - I$ :

$$\mathbb{P}(|\hat{I}_n - I| > t) \leq \frac{\text{Var}(\hat{I}_n)}{t^2} = \frac{\sigma^2/n}{t^2}.$$

Choose  $t = c\sigma/\sqrt{n}$ :

$$\mathbb{P}\left(|\hat{I}_n - I| > c \frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{c^2}.$$

Valid for all  $n$ , though conservative since it only uses the variance.

# Central Limit Theorem Approximation — From Inequality to Probability

From the Central Limit Theorem,

$$\frac{\sqrt{n}(\hat{I}_n - I)}{\sigma} \xrightarrow{D} \mathcal{N}(0, 1).$$

Hence, for large  $n$ ,

$$\mathbb{P}\left(|\hat{I}_n - I| > c \frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}(|Z| > c) = 2(1 - \Phi(c)).$$

- $\Phi(c)$ : CDF of the standard normal.
- Example values:  $c=1.96 \Rightarrow 95\%$  coverage;  $c=2.58 \Rightarrow 99\%$ .
- Much tighter than Chebyshev's bound.

# Building a Confidence Interval

From the Central Limit Theorem approximation and unbiased estimator of  $\sigma^2$ ,

$$\frac{\hat{I}_n - I}{S_n/\sqrt{n}} \approx Z \sim \mathcal{N}(0, 1).$$

Let  $c_\alpha$  satisfy  $2(1 - \Phi(c_\alpha)) = \alpha$ . Then

$$\mathbb{P}\left(-c_\alpha < \frac{\hat{I}_n - I}{S_n/\sqrt{n}} < c_\alpha\right) \approx 1 - \alpha.$$

Rearranging gives the  $(1 - \alpha)$  confidence interval for  $I$ :

$$I \in \left[\hat{I}_n - c_\alpha \frac{S_n}{\sqrt{n}}, \hat{I}_n + c_\alpha \frac{S_n}{\sqrt{n}}\right].$$

# Comparing Chebyshev and Central Limit Theorem Bounds

**Chebyshev (exact, loose):**

$$\mathbb{P}(|\hat{I}_n - I| > c\sigma/\sqrt{n}) \leq \frac{1}{c^2}.$$

$c$	Chebyshev	Central Limit Theorem (Normal)
1	1.00	0.317
2	0.25	0.0455
3	0.111	0.0027

**Central Limit Theorem (approx., tight):**

$$\mathbb{P}(|\hat{I}_n - I| > c\sigma/\sqrt{n}) \approx 2(1 - \Phi(c)).$$

⇒ Central Limit Theorem gives realistic probabilities; Chebyshev is universally valid but overly conservative.

Because  $\text{Var}(\hat{I}_n) = \sigma^2/n$ ,

$$\text{RMSE} = \sqrt{\text{Var}(\hat{I}_n)} = \frac{\sigma}{\sqrt{n}}.$$

Hence the typical estimation error decreases at rate

$$\boxed{\mathcal{O}(n^{-1/2})}.$$

## Implications:

- Doubling the precision (halving the error) requires  $4\times$  as many samples.
- This  $n^{-1/2}$  rate is fundamental—independent of the dimension of  $X$ .

Often the integral is  $I = \mathbb{E}_\pi[\varphi(X)]$  for a specific  $\varphi$  and target distribution  $\pi$ .

Monte Carlo approach relies on independent copies of  $X \sim \pi$ .

MC to approximate  $\mathbb{E}_\pi[\varphi(X)] \iff$  simulate from  $\pi$ .

Thus “Monte Carlo” sometimes refers broadly to simulation methods.

## Why Monte Carlo in our Case?

From previous chapter, we want to compute moments associated with: Posterior distribution:

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

The dimension of  $\theta$  can be large.

## A Bit of Historical Background and Intuition

Metropolis and Ulam (1949) and von Neumann (1951). Why the name “Monte Carlo”?

Two simple examples:

1. Probability of getting a total of six points when rolling two fair dice.
2. Throwing darts at a graph.



Assume we know how to generate draws from  $\pi(\theta \mid Y^T, i)$ . What does it mean to draw from it? Two basic questions:

1. Why do we want to do it?
2. How do we do it?

# How Do We Do It? Random Number Generators

Large literature. Two good surveys:

- Devroye (1986) *Non-Uniform Random Variate Generation*.
- Robert & Casella (2004) *Monte Carlo Statistical Methods*.

# Random Draws?

Natural sources of randomness are hard to use. A computer is deterministic! von Neumann (1951): “Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.”

- Pseudo-random number generators are highly non-linear iterative algorithms that “look random”.
- We focus on  $U(0, 1)$  draws.
- In general, other distributions arise from transforming uniforms.

Design iterative algorithms (Lehmer, 1951) that:

1. Are unpredictable for the uninitiated (relation to chaotic dynamical systems).
2. Pass standard statistical tests (K-S,  $\text{ARMA}(p, q)$ , etc.).

## Basic Component: Congruential Generators

Multiplicative congruential generator:

$$x_i = (ax_{i-1} + b) \bmod (M + 1) \text{ and } x_0 \text{ is the seed.}$$

Then:

$$u_i = x_i / (M + 1) \text{ is an } U(0, 1)$$

## Example: Generating Integers and Uniforms

Parameters:  $a = 5$ ,  $b = 3$ ,  $M = 16$ , seed  $x_0 = 7$ .

$$x_i = (ax_{i-1} + b) \bmod (M + 1) \quad \Rightarrow \quad x_i \in \{0, \dots, M\}.$$

$i$	Computation	$x_i$	$u_i = x_i / (M + 1)$
0	seed	7	0.412
1	$(57 + 3) \bmod 17 = 4$	4	0.235
2	$(54 + 3) \bmod 17 = 6$	6	0.353
3	$(56 + 3) \bmod 17 = 16$	16	0.941
4	$(516 + 3) \bmod 17 = 15$	15	0.882

Period/performance hinge on  $a, b, M$ . Bad choice example:  $a = 13$ ,  $c = 0$ ,  $M = 31$ ,  $x_0 = 1$  (historical bad examples: IBM RND, 1960s).



Traditional:  $a = 7^5 = 16807$ ,  $c = 0$ ,  $m = 2^{31} - 1$ . Period bounded by  $M$ . 32 vs 64 bit hardware matters. Beware IEEE floating-point standard. Alternatives exist.

Don't code your own RNG. MATLAB implements state-of-the-art (e.g., KISS by Marsaglia & Zaman, 1991). For Fortran/C++: see DIEHARD battery.

We often need non-uniform draws. Basic approach, move from uniforms via:

- Transformations (standard tricks).
- Inverse cdf method.

These underpin commercial software.

## Transformations Example: Normal via Box–Muller

Let  $U_1, U_2 \sim U(0, 1)$ . Then

$$x = \cos(2\pi U_1) \sqrt{-2 \log U_2}, \quad y = \sin(2\pi U_1) \sqrt{-2 \log U_2}$$

are i.i.d.  $N(0, 1)$  (points lie on a spiral in  $(x, y)$ ).

## Transformations Example: Multivariate Normal

If  $x \sim N(0, I)$  and  $\Sigma\Sigma^\top$  is the covariance, then

$$y = \mu + \Sigma x \sim N(\mu, \Sigma\Sigma^\top).$$

Use Cholesky for  $\Sigma$ .

# The Inverse Transform Method

Goal: Generate a random variable  $X$  with a known CDF using uniform draws.

- Let:

$$X = F^{-1}(U) \text{ and } U \sim U(0, 1)$$

- Then

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

- Hence  $X$  has distribution  $F$ .

## Example Inverse 1: Exponential Distribution

Let  $X \sim \text{Exp}(\lambda)$  with

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

Solve for  $x$  in terms of  $u = F(x)$ :

$$u = 1 - e^{-\lambda x} \quad \Rightarrow \quad x = -\frac{1}{\lambda} \ln(1 - u).$$

### Algorithm

1. Draw  $U \sim U(0, 1)$ .
2. Set  $X = -\frac{1}{\lambda} \ln U$ .

Then  $X$  follows  $\text{Exp}(\lambda)$ . Simple and exact.

## Example Inverse 2: Discrete Case (Bernoulli)

Let  $X \in \{0, 1\}$  with  $P(X = 1) = p$ . CDF:

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

**Algorithm:**

$$X = \begin{cases} 1, & \text{if } U < p, \\ 0, & \text{otherwise.} \end{cases}$$

**Comment:** works for any discrete distribution by comparing  $U$  with cumulative probabilities.



# Fundamental Theorem of Simulation

- Transformations and Inverse Method very limited set of distributions.
- We now present a general approach
- Suppose  $f(x)$  is a probability density on a measurable space  $\mathcal{X} \subseteq \mathbb{R}^d$ .
- Imagine the set under its graph:

$$A = \{(x, y) \in \mathcal{X} \times [0, \infty) : 0 \leq y \leq f(x)\}.$$

- If we draw points uniformly in  $A$ , the projection of these points onto the  $x$ -axis follows exactly the distribution with density  $f(x)$ .
- Intuitively, higher parts of the curve receive proportionally more projected points.

**Theorem.** Let  $f : \mathbb{R}^d \rightarrow [0, \infty)$  satisfy  $\int f(x) dx = 1$ . Define

$$A = \{(x, y) \in \mathbb{R}^d \times [0, \infty) : 0 \leq y \leq f(x)\}.$$

If  $(X, Y)$  is uniformly distributed on  $A$ , marginal distribution of  $X$  has density  $f(x)$ .

**Idea:** Uniform sampling under the density surface produces samples distributed according to that density.

- Since  $\int f(x) dx = 1$ , the total volume of  $A$  is one:

$$|A| = \int_{\mathbb{R}^d} \int_0^{f(x)} dy dx = 1.$$

- The joint density of  $(X, Y)$  is therefore

$$p_{X,Y}(x, y) = \mathbf{1}\{0 \leq y \leq f(x)\}.$$

- Integrating out  $y$ ,

$$p_X(x) = \int_0^{f(x)} 1 dy = f(x).$$

- Hence the marginal of  $X$  has density  $f(x)$ .  $\square$

# Acceptance Sampling: Motivation

We cannot easily draw points uniformly under  $f(x)$ .

- Introduce a simpler density  $g(x)$  such that we can draw from it.
- Find a constant  $a > 0$  such that:

$$f(x) \leq a g(x) \quad \forall x.$$

- The function  $a g(x)$  is called an **envelope** of  $f(x)$ .

**Idea:** sample uniformly under  $a g(x)$  and keep only points under  $f(x)$ .

# Acceptance Sampling: Algorithm

1. Draw  $X \sim g(x)$ .
2. Draw  $U \sim U(0, 1)$ .
3. Accept  $X$  if  $U \leq \frac{f(X)}{a g(X)}$ .

## Interpretation:

- The pair  $(X, Uag(X))$  is uniformly distributed under the curve  $ag(x)$ .
- Accepted draws correspond to points that lie under  $f(x)$ .
- Hence accepted  $X$  follow the target density  $f(x)$ .

Let  $f, g \geq 0$  and assume  $\text{supp}(f) \subseteq \text{supp}(g)$  (i.e.,  $g(x) = 0 \Rightarrow f(x) = 0$ ). Define

$$a = \sup_x \frac{f(x)}{g(x)}.$$

**Claim:**  $f(x) \leq a g(x)$  for all  $x$ .

## Is $a \geq 1$ Always?

Let  $f, g$  be densities with  $\text{supp}(f) \subseteq \text{supp}(g)$  and

$$a = \sup_x \frac{f(x)}{g(x)}.$$

**Claim.**  $a \geq 1$ , with  $a = 1$  iff  $f = g$  almost everywhere (w.r.t.  $g$ ). Hence  $a > 1$  whenever  $f \neq g$  on a set of positive  $g$ -measure.

Let  $R(x) = \frac{f(x)}{g(x)}$  where  $g(x) > 0$  (define  $R = 0$  when  $g = 0$ ; then  $f = 0$  there). Since  $f, g$  are densities,

$$\mathbb{E}_g[R(X)] = \int \frac{f(x)}{g(x)} g(x) dx = \int f(x) dx = 1.$$

If  $R(x) < 1$  for all  $x$  with  $g(x) > 0$ , then  $\mathbb{E}_g[R(X)] < 1$ , a contradiction. Thus  $\sup_x R(x) \geq 1$ , i.e.  $a \geq 1$ .

Moreover,  $a = 1$  iff  $R(x) \leq 1$  a.e. and  $\mathbb{E}_g[R] = 1$ , which forces  $R(x) = 1$  a.e., i.e.  $f(x) = g(x)$  a.e. □



- The acceptance rate is  $P(\text{accept}) = \frac{1}{a}$ .
- A tight envelope ( $a$  close to 1)  $\Rightarrow$  more efficient sampling.
- A loose envelope ( $a$  large)  $\Rightarrow$  many rejections.
- Ideally,  $g(x)$  resembles  $f(x)$  in shape and tails.

**Proof**  $P(\text{accept}) = \frac{1}{a}$

Let  $A = \{U \leq f(X)/(ag(X))\}$ . Then

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A] = \mathbb{E}\left[\mathbf{1}\left\{U \leq \frac{f(X)}{ag(X)}\right\}\right] = \mathbb{E}\left[\mathbb{E}\left(\mathbf{1}\left\{U \leq \frac{f(X)}{ag(X)}\right\} \mid X\right)\right].$$

Independence  $\Rightarrow$  given  $X = x$ ,  $U \sim U(0, 1)$  so

$$\mathbb{P}\left(U \leq \frac{f(X)}{ag(X)} \mid X\right) = \frac{f(X)}{ag(X)}.$$

Hence

$$\mathbb{P}(A) = \mathbb{E}\left[\frac{f(X)}{ag(X)}\right] = \int \frac{f(x)}{ag(x)} g(x) dx = \frac{1}{a} \int f(x) dx = \frac{1}{a}.$$

## Proof (One Line with the Supremum Property)

Define  $r(x) = \frac{f(x)}{g(x)}$  for  $g(x) > 0$  and set  $r(x) = 0$  when  $g(x) = 0$  (using  $g(x) = 0 \Rightarrow f(x) = 0$ ). By definition of the supremum,

$$r(x) \leq \sup_z r(z) = a \quad \text{for every } x.$$

Multiplying by  $g(x) \geq 0$  yields

$$f(x) = r(x) g(x) \leq a g(x) \quad \forall x.$$

□

- Suppose the target is a **truncated version** of an easy distribution  $g(x)$ :

$$f(x) = \frac{g(x) \mathbf{1}\{x \in A\}}{P_g(A)}, \quad A = \text{allowed region.}$$

- We can draw  $X \sim g$  easily, but we only want  $X \in A$ .
- Accept–Reject Sampling fits perfectly:
  1. Draw  $X \sim g$ .
  2. If  $X \in A$ , accept; otherwise, reject.

# Why It Works So Well

- Because  $a = \frac{1}{P_g(A)}$ , the acceptance probability is simply

$$P(\text{accept}) = P_g(X \in A).$$

- Efficiency depends only on how much mass of  $g$  lies inside  $A$ .
- If truncation is mild (e.g. 80–90% of the mass kept), then the acceptance rate is high and the algorithm is almost costless.
- Even for more severe truncation, the method is simple, exact, and needs no renormalization.

**Key idea:** For truncated densities, Accept–Reject  $\Rightarrow$  “Draw from the full  $g$  and keep what’s valid.” No extra math—just logical filtering.

## Truncated Distributions and the Accept–Reject Rule

Target density: truncated version of an easy  $g(x)$ ,

$$f(x) = \frac{g(x) \mathbf{1}\{x \in A\}}{P_g(A)}, \quad P_g(A) = \int_A g(x) dx.$$

Hence

$$\frac{f(x)}{g(x)} = \begin{cases} \frac{1}{P_g(A)}, & x \in A, \\ 0, & x \notin A. \end{cases}$$

To satisfy  $f(x) \leq a g(x)$  for all  $x$ , choose

$$a = \sup_x \frac{f(x)}{g(x)} = \frac{1}{P_g(A)}.$$

## Substitute into the Acceptance Condition

The generic rule:

$$U \leq \frac{f(X)}{a g(X)}.$$

Substitute the truncated expressions:

$$\frac{f(X)}{a g(X)} = \begin{cases} \frac{1/P_g(A)}{1/P_g(A)} = 1, & X \in A, \\ 0, & X \notin A. \end{cases}$$

Therefore

$$U \leq \begin{cases} 1, & X \in A, \\ 0, & X \notin A. \end{cases}$$

**Interpretation:**

- If  $X \in A$ :  $U \leq 1$  always true  $\Rightarrow$  accept.

## Conclusion: Why It Simplifies Perfectly

- For truncated densities, the acceptance test reduces to a simple membership check:

$$U \leq \frac{f(X)}{a g(X)} \iff X \in A.$$

- Inside  $A$ ,  $f$  and  $g$  have the same shape—just rescaled by  $1/P_g(A)$ .
- The acceptance rate is  $P_g(A)$ : the mass of  $g$  inside  $A$ .
- Hence the algorithm is extremely efficient:
  1. Draw  $X \sim g$ ,
  2. Accept if  $X \in A$ .

**Summary:** In the truncated case, Accept–Reject becomes “keep draws that lie in the truncation region.”



# Acceptance Pitfalls

1. Many rejections: minimize  $\alpha$ .
2. Need  $\pi/g$  bounded  $\Rightarrow g$  must have thicker tails.
3. Computing  $\alpha$  can be hard.

Can we do better? Yes—importance sampling.

Same setup. For any integrable  $h$ ,

$$\mathbb{E}_{\pi}[h(\theta)] = \int h(\theta) \frac{\pi(\theta)}{g(\theta)} g(\theta) d\theta.$$

With draws  $\{\theta_j\}_{j=1}^m$  from  $g$ ,

$$h_m^{IS} := \frac{1}{m} \sum_{j=1}^m h(\theta_j) \frac{\pi(\theta_j)}{g(\theta_j)} \rightarrow \mathbb{E}_\pi[h(\theta)].$$

## Importance Sampling III (CLT)

If  $\mathbb{E}_\pi\left[\frac{\pi(\theta)}{g(\theta)}\right]$  exists, then

$$m^{1/2}(h_m^{IS} - \mathbb{E}_\pi[h(\theta)]) \Rightarrow N(0, \sigma^2),$$

with

$$\sigma^2 \approx \frac{1}{m} \sum_{j=1}^m (h(\theta_j) - h_m^{IS})^2 \left( \frac{\pi(\theta_j)}{g(\theta_j)} \right)^2.$$

## Importance Sampling IV: Variance Intuition

We want the weight ratio  $\pi(\theta)/g(\theta)$  to be as flat as possible. Ideally  $g = \pi$ .

## Importance Sampling V: Picking $g$

Use a local (e.g., Taylor) approximation to  $\pi$  as  $g$ . Question: how to compute the Taylor approximation?

A simple sufficient condition:  $\pi(\theta)/g(\theta)$  bounded. Denote  $\omega(\theta) = \pi(\theta)/g(\theta)$ .

If only unnormalized densities  $\tilde{\pi}, \tilde{g}$  are available, then

$$\mathbb{E}_{\pi}[h(\theta)] = \frac{\int h(\theta) \frac{\tilde{\pi}(\theta)}{\tilde{g}(\theta)} \tilde{g}(\theta) d\theta}{\int \frac{\tilde{\pi}(\theta)}{\tilde{g}(\theta)} \tilde{g}(\theta) d\theta}.$$



$$h_m^{SN} = \frac{\sum_{j=1}^m h(\theta_j) \omega(\theta_j)}{\sum_{j=1}^m \omega(\theta_j)}, \quad \sigma^2 \approx \frac{m \sum_{j=1}^m (h(\theta_j) - h_m^{SN})^2 \omega(\theta_j)^2}{\left(\sum_{j=1}^m \omega(\theta_j)\right)^2}.$$

## Example I: Bad $g$ (Heavy Tails Target)

Suppose  $\pi$  is  $t_\nu$  but we sample from  $g = N(0, 1)$ . Estimate  $\mathbb{E}[X]$  of  $t_\nu$  using IS weights  $\omega(\theta) = t_\nu(\theta)/\phi(\theta)$ .

Draw  $\theta_j \sim N(0, 1)$ , then

$$\widehat{\text{mean}} = \frac{1}{m} \sum_{j=1}^m \theta_j \omega(\theta_j), \quad \widehat{\text{Var}}(\widehat{\text{mean}}) = \frac{1}{m} \sum_{j=1}^m (\theta_j - \widehat{\text{mean}})^2 \omega(\theta_j)^2.$$

## Illustration (Reported in Slides)

As  $\nu = 3, 4, 10, 100$ , the estimated variance of the mean falls, but can be extremely large for small  $\nu$  under normal  $g$ .

## Table (Example I)

$\nu$	3	4	10	100
Est. Mean	0.1026	0.0738	0.0198	0.0000
Est. Var(Est. Mean)	684.52	365.66	36.82	3.59

## Example II: Heavy-Tailed $g$ for Light-Tailed Target

Now  $\pi = N(0, 1)$ , but draw from  $g = t_\nu$ . Variance of the IS estimator is moderate across  $\nu$ .

## Table (Example II)

$t_\nu$	3	4	10	100
Est. Mean	-0.0104	-0.0075	0.0035	-0.0029
Est. Var(Est. Mean)	2.0404	2.1200	2.2477	2.7444

## Relative Numerical Efficiency (RNE)

If  $g = \pi$ , then

$$\sigma^2 \approx \frac{1}{m} \sum_{j=1}^m (h(\theta_j) - h_m^{IS})^2 \approx \text{Var}_{\pi}[h(\theta)].$$

Define

$$\text{RNE} = \frac{\text{Var}_{\pi}[h(\theta)]}{\sigma^2}.$$

RNE close to 1: good IS; near 0: poor IS.



**Target**  $t_\nu$ ,  $g = N(0, 1)$ :

$$\text{RNE} = \{0.0134, 0.0200, 0.0788, 0.2910\} \text{ for } \nu = \{3, 4, 10, 100\}.$$

**Target**  $N(0, 1)$ ,  $g = t_\nu$ :

$$\text{RNE} = \{0.4777, 0.4697, 0.4304, 0.3471\}.$$

Two researchers share the likelihood  $f(Y^T | \theta)$  but have priors  $\pi_1(\theta) \neq \pi_2(\theta)$ . If researcher 1 has draws  $\theta^{(j)} \sim \pi(\theta | Y^T, \pi_1)$ , then for any  $h$ ,

$$\int h(\theta) \pi(\theta | Y^T, \pi_2) d\theta \approx \frac{\sum_{j=1}^m h(\theta^{(j)}) \frac{\pi_2(\theta^{(j)})}{\pi_1(\theta^{(j)})}}{\sum_{j=1}^m \frac{\pi_2(\theta^{(j)})}{\pi_1(\theta^{(j)})}}.$$

$$\int h(\theta) \pi(\theta | Y^T, \pi_2) d\theta = \frac{\int h(\theta) f(Y^T | \theta) \pi_2(\theta) d\theta}{\int f(Y^T | \theta) \pi_2(\theta) d\theta} = \frac{\int h(\theta) \frac{\pi_2(\theta)}{\pi_1(\theta)} \pi(\theta | Y^T, \pi_1) d\theta}{\int \frac{\pi_2(\theta)}{\pi_1(\theta)} \pi(\theta | Y^T, \pi_1) d\theta}.$$

# Importance Sampling Summary

- Choose  $g$  close to  $\pi$  (match location, scale, tails).
- Use self-normalized IS if normalizing constants unknown.
- Diagnose with RNE; respecify  $g$  if RNE is low.

- Simulation approximates expectations under complex posteriors/marginals.
- RNG quality matters; use proven libraries.
- Acceptance sampling is simple but can be inefficient.
- Importance sampling is powerful; success hinges on a good proposal.