Lecture 5: Simulation of the Posterior Distribution

Thibault Randrianarisoa

UTSC

February 12, 2026

Outline

- Introduction to (Posterior) Simulation
- Inverse Transform Method
- Rejection & Importance Sampling
- MCMC
  - Markov Chains
  - Metropolis-Hastings
  - Gibbs sampling

Introduction to sampling

Simulation of the Posterior

In practice, the posterior distribution is often an extremely complicated object that requires computing difficult integrals (like the evidence/marginal density of observations).

- If we do not have direct access to a density with an explicit form, we can seek to simulate it (i.e., sample values according to this law).
- Except in special cases (conjugate families), it is difficult to explicitly determine quantities like the mean, median, or quantiles.

Example

For example, the posterior mean is written as an integral against the posterior law:

$$\int_{\Theta} \theta d\Pi(\theta \mid \mathbf{X})$$

We use simulation (Monte Carlo methods) to approximate such integrals.

Basic Ingredient: The Uniform Generator

In any random variable simulation method, we always assume we have a basic ingredient:

Assumption

We have access to a generator of the Uniform distribution on $[0, 1]$, capable of providing independent realizations.

- In reality, computers use *pseudo-random* numbers.
- For this course, we assume we can simulate true uniform random variables.

Method 1: Inverse Transform

The first major simulation method is the Inverse Transform method.
We wish to simulate a real random variable X with cumulative distribution function (CDF) F.

Definition

Let F^{-1} be the generalized inverse of F, defined as:

$$\forall u \in [0,1], \quad F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$$

with the conventions $\inf \mathbb{R} = -\infty$ and $\inf \varnothing = +\infty$.

Even if we don't have the equivalence $F^{-1}(u) = x \iff F(x) = u$, we always have:

$$F(F^{-1}(u)) \geqslant u \quad \text{and} \quad F^{-1}(u) \leqslant x \iff u \leqslant F(x)$$

Inverse Transform: The Result

Proposition

Let $U \sim \text{Unif}([0,1])$. For all $x \in \mathbb{R}$, we have:

$$\mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x)$$

In other words, $F^{-1}(U) \sim X$.

Implication: If we know how to calculate F^{-1}, we know how to simulate a random variable with c.d.f. F.

Limitations of Inverse Transform

This simple method is not always feasible in practice.

- It requires knowing how to invert the CDF, which is not always possible explicitly.
- Sometimes the CDF itself is not accessible other than as an integral of the density.
- **Example:** The Normal (Gaussian) distribution does not have an explicit closed-form inverse CDF.

Method 2: Rejection Sampling

We wish to simulate a random variable in $\mathbb{R}^d$ with density f, but f is too complicated to simulate directly.

Assumptions

1. We can simulate from another density g (the proposal).
2. There exists a constant $m \geqslant 1$ such that:

$$\forall y \in \mathbb{R}^d, \quad f(y) \leqslant m g(y)$$

3. We can calculate the ratio $r(y) = \frac{f(y)}{m g(y)}$ for any y where $g(y) > 0$.

Rejection Algorithm

Let $(U_i)_{i \geqslant 1}$ be i.i.d. Unif$([0,1])$ and $(Y_i)_{i \geqslant 1}$ be i.i.d. with density g, independent of (U_i).

Algorithm

Define the stopping time τ:

$$\tau = \inf\{i \in \mathbb{N}^* : r(Y_i) \geqslant U_i\}$$

The variable $X = Y_\tau$ follows the density f.

Procedure:

1. Sample $Y \sim g$ and $U \sim \text{Unif}[0,1]$.
2. If $U \leqslant \frac{f(Y)}{mg(Y)}$, accept $X = Y$.
3. Otherwise, reject and repeat.

Properties of Rejection Sampling

Proposition

The variable $X = Y_\tau$ has density f. Furthermore, τ follows a geometric distribution with parameter $1/m$ and is independent of X.

- **Efficiency:** The expectation of τ is m. This means, on average, we must wait m trials to obtain one simulation of X.
- **Optimization:** To limit the number of rejections, it is important to choose g close to f so that m is as close to 1 as possible.

Proof of Validity

Let A be a Borel set in $\mathbb{R}^d$. By independence of the trials:

$$\begin{aligned}\mathbf{P}(Y_\tau \in A, \tau = n) &= \mathbf{P}(r(Y_1) < U_1, \ldots, r(Y_{n-1}) < U_{n-1}, r(Y_n) \geqslant U_n, Y_n \in A) \\ &= \mathbf{P}(r(Y) < U)^{n-1}\mathbf{P}(r(Y) \geqslant U, Y \in A)\end{aligned}$$

Using the independence between Y and U:

$$\mathbf{P}(r(Y) \geqslant U, Y \in A) = \int_{\mathbb{R}^d}\int_0^1 \mathbb{1}_{\{r(y)\geqslant u\}}\mathbb{1}_{\{y\in A\}}g(y)dudy$$

$$= \int_A r(y)g(y)dy = \int_A \frac{f(y)}{mg(y)}g(y)dy = \frac{1}{m}\int_A f(y)dy$$

Similarly, $\mathbf{P}(r(Y) < U) = 1 - \frac{1}{m}$.

Proof of Validity and Independence

Summing the probability of acceptance over all possible trial counts n:

$$\mathbf{P}(Y_\tau \in A) = \sum_{n=1}^{\infty} \mathbf{P}(Y_\tau \in A, \tau = n) = \sum_{n=1}^{\infty} \left(1 - \frac{1}{m}\right)^{n-1} \frac{1}{m} \int_A f(y)dy$$

$$= \int_A f(y)dy$$

Independence: Since $\mathbf{P}(Y_\tau \in A, \tau = n) = \mathbf{P}(Y_\tau \in A)\mathbf{P}(\tau = n)$, the variables τ (number of trials) and Y_τ (the accepted sample) are independent.

Example: Uniform on a Subset

Let A be a subset of the cube $[0,1]^d$. It is easy to simulate uniform on $[0,1]^d$, but harder on A.

- **Target:** Uniform on A.
- **Proposal:** $Y_1, Y_2, \ldots$ independent uniform on $[0,1]^d$.
- **Algorithm:** Draw variables until the first time τ where $Y_\tau \in A$.

Here, the ratio is simply the indicator $\mathbb{1}_{\{y \in A\}}$. The acceptance probability is related to the volume $\lambda(A)$. If $\lambda(A)$ is very small, the method is computationally expensive (high rejection rate).

Monte-Carlo methods

Monte Carlo Methods for Integration

Suppose Q is a distribution with density q on a compact set in $\mathbb{R}^d$ (e.g., $[0,1]^d$). Let ϕ be a known measurable function. We wish to calculate:

$$\mathbf{I} = \int \phi(x) dQ(x) = \int \phi(x) q(x) dx$$

Deterministic Approach (Riemann Sums):

- Divide $[0,1]^d$ into N^d smaller sub-cubes. Approximate $\phi \times q$ by a constant on each.
- **The Curse of Dimensionality:** To achieve a precision ε, the number of points required typically scales as ε^{-d}.
- If $d \geqslant 3$, the computational cost explodes.

Monte Carlo methods introduce randomness to break this dependence on the dimension.

Instead of a fixed grid, we use random points. Let $X_1, \ldots, X_N$ be i.i.d. random variables with distribution Q.

Motivation

By the Law of Large Numbers (LLN):

$$\mathbf{I}_N = \frac{1}{N} \sum_{i=1}^{N} \phi(X_i) \xrightarrow{a.s.} \int \phi(x) dQ(x) = \mathbf{I}$$

Central Limit Theorem

If $\int \phi^2 dQ < \infty$, then as $N \to \infty$:

$$\sqrt{N}(\mathbf{I}_N - \mathbf{I}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(\phi(X)))$$

Advantages and Limitations

Advantage:

- The convergence rate is $N^{-1/2}$, independent of the dimension d.
- We do not need the explicit density q, only the ability to simulate from Q.

Limitations:

- Simulation Difficulty: We may not know how to simulate from Q directly.
- Rare Events: If $\mathbf{I} = \mathbb{P}(X > 3)$ for $X \sim \mathcal{N}(0, 1)$, the event is rare.
- We would need an extremely large N to observe enough samples in the region of interest (and not have a zero estimator).

Solution: Change the sampling distribution to target the important regions (Importance Sampling).

Importance Sampling: The Principle

We wish to approximate $\mathbf{I} = \int \phi dQ$ where Q has density q.

Let $\tilde{q}$ be another density on $\mathbb{R}^d$ (the *proposal*) such that:

1. We can simulate efficiently from $\tilde{q}$.
2. **Support Condition:** $\forall y \in \mathbb{R}^d, \quad \tilde{q}(y) = 0 \implies \phi(y)q(y) = 0.$

Importance Sampling Estimator

Let $Y_1, \ldots, Y_N$ be i.i.d. with density $\tilde{q}$. We define:

$$\mathbf{J}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi(Y_i)q(Y_i)}{\tilde{q}(Y_i)}$$

Convergence of Importance Sampling

Under the integrability condition on ϕ, the Law of Large Numbers gives:

$$\mathbf{J}_N \xrightarrow{a.s.} \mathbf{I}$$

Note: We rewrite the integral $\int \phi q = \int \frac{\phi q}{\tilde{q}} \tilde{q}$.

Central Limit Theorem: To obtain a CLT, we must verify the second-order moment condition under the proposal density $\tilde{q}$:

$$\mathbf{E}_{\tilde{q}}\left[\frac{\phi(Y)^2 q(Y)^2}{\tilde{q}(Y)^2}\right] = \int \frac{\phi(y)^2 q(y)^2}{\tilde{q}(y)} dy < \infty$$

- We no longer simulate according to q, but according to $\tilde{q}$, which we choose freely.
- This allows us to place more probability mass in regions where the integrand is large (e.g., rare events), thereby reducing variance.

Example: Rare Event Simulation

Consider estimating the probability of a rare event for a standard normal variable:

$$\mathbf{I} = \mathbb{P}(X > 3) = \int \mathbb{1}_{x>3} q(x) dx, \quad \text{where } X \sim \mathcal{N}(0,1).$$

Standard Monte Carlo Approach:

- We draw $X_1, \ldots, X_N \sim \mathcal{N}(0,1)$ and compute $\mathbf{I}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i>3}$.
- Since the event is rare ($\mathbf{I} \approx 0.00135$), we need N to be extremely large to get a non-zero estimate.
- The variance is $\sigma^2_{MC} = \mathbf{I}(1-\mathbf{I}) \approx 10^{-3}$.

Problem: Most samples fall near 0, providing no information about the tail $x > 3$.

Example: Importance Sampling Solution

Importance Sampling Strategy:

- Choose a proposal density $\tilde{q}$ that puts more mass in the region $\{x > 3\}$.
- Let's use $\tilde{q}$ as the density of $\mathcal{N}(3, 1)$.

We simulate $Y_1, \ldots, Y_N \sim \mathcal{N}(3, 1)$ and compute:

$$\mathbf{J}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}_{Y_i > 3} q(Y_i)}{\tilde{q}(Y_i)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i > 3} e^{3Y_i - 9/2}$$

Result:

- The variance σ_{IS}^2 can be calculated (or estimated) to be approximately 10^{-6}.
- Comparison: The IS variance is a factor of 1000 smaller than standard MC. We obtain the same precision with far fewer samples.

Optimal Choice of Proposal Density

A natural question is: *What is the best possible choice for $\tilde{q}$?*

Proposition

The optimal proposal density q^* (which minimizes the variance of the estimator) is proportional to the absolute value of the integrand:

$$q^*(x) = \frac{|\phi(x)|q(x)}{\int |\phi(y)|q(y)dy}$$

Proof Intuition: The variance term involves $\int \frac{\phi^2 q^2}{\tilde{q}}$. By Jensen's inequality (or Cauchy-Schwarz), this is minimized when $\tilde{q} \propto |\phi|q$.

Remark

If $\phi(x) \geqslant 0$, the optimal density is:

$$q^*(x) = \frac{\phi(x)q(x)}{\mathbf{I}}$$

In this case, the variance is exactly **zero**.

- If we could simulate from q^*, a *single* sample would give the exact answer:

$$\frac{\phi(Y)q(Y)}{q^*(Y)} = \mathbf{I}$$

- ⚠ **paradox:** To use q^*, we need to know the normalizing constant $\mathbf{I}$, which is exactly the integral we are trying to compute!
- **Practical Use:** This theoretical result guides us to choose $\tilde{q}$ that is shaped similarly to $|\phi|q$.

Application: Estimating the Posterior Mean

The Problem

Recall that the posterior distribution is given by Bayes' formula:

$$\pi(\theta \mid \mathbf{X}) = \frac{\pi(\theta) p_\theta(\mathbf{X})}{\int_\Theta \pi(\theta) p_\theta(\mathbf{X}) d\nu(\theta)}$$

- The numerator is usually easy to compute.
- The denominator (the marginal likelihood) involves a potentially difficult integral.
- **Goal:** We want to compute an expectation under the posterior, $\int \phi(\theta) d\pi(\theta \mid \mathbf{X})$ (e.g., the posterior mean if $\phi(\theta) = \theta$).

Direct simulation from $\Pi(\cdot \mid \mathbf{X})$ is hard. However, simulating from the prior $\Pi(\cdot)$ is often easy.

We can rewrite the posterior expectation as a ratio of two integrals against the prior distribution:

$$\mathbf{E}[\phi(\theta) \mid \mathbf{X}] = \int_{\Theta} \phi(\theta) d\pi(\theta \mid \mathbf{X}) = \frac{\int_{\Theta} \phi(\theta) p_{\theta}(\mathbf{X}) \pi(\theta) d\nu(\theta)}{\int_{\Theta} p_{\theta}(\mathbf{X}) \pi(\theta) d\nu(\theta)}$$

Monte Carlo Strategy

1. Generate i.i.d. samples $\theta_1, \ldots, \theta_m$ from the **prior** π.
2. Approximate the numerator by $\frac{1}{m} \sum_{j=1}^{m} \phi(\theta_j) p_{\theta_j}(\mathbf{X})$.
3. Approximate the denominator by $\frac{1}{m} \sum_{j=1}^{m} p_{\theta_j}(\mathbf{X})$.

This yields the following estimator for the posterior expectation:

$$\widehat{\phi}_n^{(m)} = \frac{\sum_{j=1}^m \phi(\theta_j) p_{\theta_j}(\mathbf{X})}{\sum_{j=1}^m p_{\theta_j}(\mathbf{X})}$$

Properties:

- **Consistency:** By the Law of Large Numbers, as $m \to \infty$:

$$\widehat{\phi}_n^{(m)} \xrightarrow{a.s.} \int_\Theta \phi(\theta) d\Pi(\theta \mid \mathbf{X})$$

- **Asymptotic Normality:** Using the Delta method in dimension 2, one can show that $\widehat{\phi}_n^{(m)}$ is asymptotically normal.

⚠ *Important:* In this convergence, the data size n is fixed. It is the number of simulations m that tends to $+\infty$.

Markov Chain Monte-Carlo

Definition

MCMC stands for Markov Chain Monte Carlo. The goal is to approximate a target distribution or an integral by constructing a Markov chain that explores the state space.

Homogeneous Markov Chain: A process $(X_t)_{t\in\mathbb{N}}$ on Ω (a subset of $\mathbb{R}^d$ or $\mathbb{N}^d$)where the transition depends only on the current state:

- If $X_t = x$, the next state X_{t+1} is chosen according to a fixed probability measure P_x.
- We assume densities $P(x, \cdot)$ exist in the continuous case:

$$P_x(A) = \mathbb{P}(X_{t+1} \in A \mid X_t = x) = \int_A P(x, y)dy$$

The function $P : \Omega \times \Omega \to [0, 1]$ is called the transition kernel. It satisfies $\int_\Omega P(x, y)dy = 1$ for all x.

Examples of Markov Chains

1. Random Walk on $\mathbb{R}$: Let $X_0 \sim \mathcal{N}(0,1)$ and $(\xi_i)_{i \geqslant 1}$ be i.i.d. $\mathcal{N}(0,1)$. Define:

$$X_{n+1} = X_n + \xi_{n+1}$$

This is a Markov chain with transition kernel $P(x,y) = \phi(y-x)$, where ϕ is the standard normal density. This is a *Gaussian random walk*.

2. Random Walk on a Finite Graph: Let $G = (V, E)$ be a finite graph. A walker moves from u to a neighbor v chosen uniformly at random.

$$P(u,v) = \begin{cases} \frac{1}{\deg(u)} & \text{if } \{u,v\} \in E \\ 0 & \text{otherwise} \end{cases}$$

In this discrete case, P is (or can be represented as) a stochastic matrix.

Evolution and Stationary Distribution

If X_0 has density ν_0, then the density of X_t, denoted ν_t, evolves according to:

$$\nu_t(y) = \int_\Omega \nu_{t-1}(x) P(x, y) d\mu(x)$$

In the discrete case with row vectors: $\nu_t = \nu_0 P^t$.

Definition: Stationary Distribution

A density π is stationary (or invariant) for the chain if:

$$\forall y \in \Omega, \quad \int_\Omega \pi(x) P(x, y) dx = \pi(y)$$

In the discrete finite case: $\pi P = \pi$.

⚠ **Remark:** A stationary distribution does not always exist (e.g., Simple Random Walk on $\mathbb{Z}$ "diffuses" to infinity and has no stationary probability distribution).

Irreducibility and Stationary Distribution

Below, we restrict ourselves to finite state spaces Ω for simplicity.

Definition

A transition kernel P on a finite set Ω is said to be irreducible if for all states $x, y \in \Omega$, there exists a time $t \in \mathbb{N}$ such that $P^t(x, y) > 0$ (i.e., it is possible to reach any state from any other state).

⚠ The time $t = t(x, y)$ above depends on x and y

Theorem

Let Ω be a finite set and P a transition kernel on Ω.

- P admits a stationary probability distribution π.
- If P is irreducible, this probability π is **unique** and charges all states ($\pi(x) > 0$ for all x).

Detailed Balance

A simple way to find a stationary probability is to look for one satisfying the detailed balance condition.

Proposition

Let P be a transition kernel on Ω. If π is a density on Ω satisfying:

$$\forall x, y \in \Omega, \quad \pi(x)P(x,y) = \pi(y)P(y,x)$$

(we say P is reversible with respect to π), then π is **stationary**.

Proof: Integrating with respect to x:

$$\int_\Omega \pi(x)P(x,y)d\mu(x) = \int_\Omega \pi(y)P(y,x)d\mu(x) = \pi(y)\underbrace{\int_\Omega P(y,x)d\mu(x)}_{=1} = \pi(y).$$

Ergodicity and Convergence

To guarantee convergence of the chain to π, irreducibility is not enough. We need a stronger property.

Definition

The kernel P is said to be ergodic if:

$$\exists t \in \mathbb{N}, \quad \forall x, y \in \Omega, \quad P^t(x, y) > 0$$

Theorem

If P is an ergodic kernel on a finite Ω, then the distance to stationarity tends to 0:

$$\max_{x \in \Omega} d_{\mathrm{TV}}(P^t(x, \cdot), \pi) \xrightarrow{t \to \infty} 0$$

This theorem is the basis of MCMC: ergodic chains converge to their stationary measure.

Theorem (Ergodic Law of Large Numbers)

Let P be an ergodic kernel on finite Ω and π its stationary probability. Let $f : \Omega \to \mathbb{R}$ be a π-integrable function. Then, for any initial probability measure $X \sim \nu_0$:

$$\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) \xrightarrow[t\to\infty]{a.s.} \mathbf{E}_\pi[f]$$

Consequence: To approximate $\mathbf{E}_\pi[f]$, we do not need to simulate exactly from π. It suffices to simulate a Markov chain (X_t) having π as its stationary distribution. For large t, the time average approximates the spatial average.

Metropolis-Hastings Algorithm

Goal: Simulate from a target density π (known up to a constant) or compute $\int f(x)\pi(x)d\mu(x)$.

Idea: Construct a chain (X_t) with stationary distribution π using a proposal kernel Q.

1. If current state is x, generate a candidate $y \sim Q(x, \cdot)$.
2. Accept the move to y with probability $\alpha(x, y)$; otherwise stay at x.

The MH Ratio

The acceptance probability is defined by $\alpha(x, y) = r(x, y) \wedge 1$, where:

$$r(x, y) = \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}$$

Validity of Metropolis-Hastings

The transition kernel P of the Metropolis-Hastings chain is:

$$P(x,y) = \begin{cases} Q(x,y)\alpha(x,y) & \text{if } y \neq x \\ 1 - \sum_{z \neq x} Q(x,z)\alpha(x,z) & \text{if } y = x \end{cases}$$

Theorem

The kernel P defined by the Metropolis-Hastings algorithm admits π as a stationary distribution.

Proof Strategy: Check the Detailed Balance condition. For $y \neq x$, verify that $\pi(x)Q(x,y)\alpha(x,y) = \pi(y)Q(y,x)\alpha(y,x)$.

Suppose we want to simulate a pair (X, Y) with values in $\mathcal{X} \times \mathcal{Y}$, where the joint distribution is difficult to sample from directly, but the conditional distributions are easy to simulate:

- $\mathcal{L}(X \mid Y = y)$
- $\mathcal{L}(Y \mid X = x)$

The Algorithm

Start with an arbitrary initial pair $(X_0, Y_0) = (x_0, y_0)$. At step $t \in \mathbb{N}^*$, given (x_t, y_t):

1. Generate X_{t+1} according to $\mathcal{L}(X \mid Y = y_t)$. Let x_{t+1} be the value obtained.
2. Generate Y_{t+1} according to $\mathcal{L}(Y \mid X = x_{t+1})$. Let y_{t+1} be the value obtained.

Properties of the Gibbs Sampler

Proposition

The sequence $(X_t, Y_t)_{t \geqslant 1}$ is a Markov chain for which the joint distribution $\mathcal{L}(X, Y)$ is a stationary distribution.

Intuition: The Gibbs sampler is a special case of Metropolis-Hastings where the proposal distributions are the conditional distributions, and the acceptance probability is always 1 (i.e., we always accept the move).

Note: This easily generalizes to dimensions $d > 2$ by updating each component one by one conditioned on all the others.

Example: Bivariate Simulation

Consider the density on $\mathbb{R}^2$:

$$h(x, y) = C \exp\left(-\frac{y^2}{2} - \frac{x^2(1 + y + y^2)}{2}\right)$$

Conditional Distributions:

- $\mathcal{L}(X \mid Y = y) = \mathcal{N}\left(0, \frac{1}{1+y+y^2}\right)$
- $\mathcal{L}(Y \mid X = x) = \mathcal{N}\left(-\frac{x^2}{2(1+x^2)}, \frac{1}{1+x^2}\right)$

Algorithm: Start at $(0, 0)$. At time t:

1. Draw $X_{t+1} \sim \mathcal{N}\left(0, \frac{1}{1+y_t+y_t^2}\right)$.
2. Draw $Y_{t+1} \sim \mathcal{N}\left(-\frac{X_{t+1}^2}{2(1+X_{t+1}^2)}, \frac{1}{1+X_{t+1}^2}\right)$.

Application to Bayesian Statistics

The Gibbs sampler is particularly useful for Hierarchical Models.

In such models, it is typically easy to simulate one variable knowing all the others (the *full conditional* distributions).

Example: Consider a model with parameters θ and hyperparameters α, and data $\mathbf{X}$. We want to sample from the posterior $\mathcal{L}(\theta, \alpha \mid \mathbf{X})$.

- If we can simulate from $\mathcal{L}(\theta \mid \alpha, \mathbf{X})$ and $\mathcal{L}(\alpha \mid \theta, \mathbf{X})$, then the Gibbs sampler allows us to simulate approximately from the joint posterior $\mathcal{L}((\theta, \alpha) \mid \mathbf{X})$ by alternating updates.