

Bayesian Theory and Computation

Lecture 3: Monte Carlo Methods



Cheng Zhang

School of Mathematical Sciences, Peking University

Feb 24, 2025

- ▶ We saw previously that in certain situations, the posterior distribution has a closed form (e.g., when the prior is conjugate), and the integrals are tractable.
- ▶ For many other problems, however, finding the posterior distribution and obtaining the expectation are far from trivial.
- ▶ Remember that even for the case of simple normal distribution with two parameters, the posterior didn't have a closed form unless we were willing to use noninformative priors or tie the variance of the mean to the variance of the data.
- ▶ In the following few lectures, we focus on problems where the posterior distribution is not analytically tractable.
- ▶ For this, we need to learn about Monte Carlo methods and Markov chain stochastic processes.

- ▶ Suppose we are interested in estimating $I(h) = \int_a^b h(x)dx$
- ▶ If we can draw iid samples, $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ uniformly from (a, b) , we can approximate the integral as

$$\hat{I}_n = (b - a) \frac{1}{n} \sum_{i=1}^n h(x^{(i)})$$

- ▶ Note that we can think about the integral as

$$(b - a) \int_a^b h(x) \cdot \frac{1}{b - a} dx$$

where $\frac{1}{b-a}$ is the density of $\text{Uniform}(a, b)$



- ▶ In general, we are interested in integrals of the form $\int_{\mathcal{X}} h(x)f(x)dx$, where $f(x)$ is a probability density function
- ▶ Analogous to the above argument, we can approximate this integral (or expectation) by drawing iid samples $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ from the density $f(x)$ and then

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

- ▶ Based on the law of large numbers, we know that

$$\lim_{n \rightarrow \infty} \hat{I}_n \xrightarrow{p} I$$

- ▶ And based on the central limit theorem

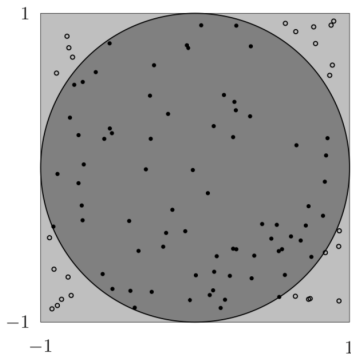
$$\sqrt{n}(\hat{I}_n - I) \rightarrow \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \text{Var}(h(X))$$



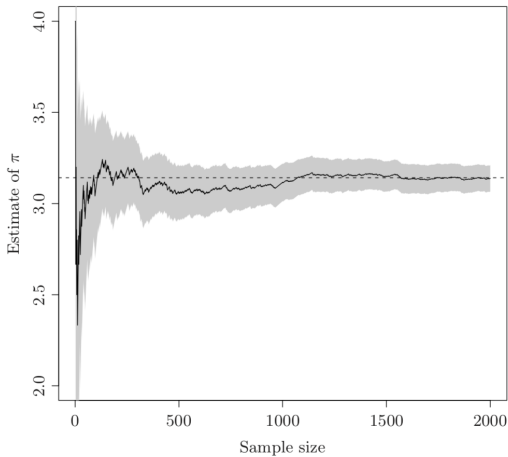
- ▶ Let $h(x) = \mathbf{1}_{B(0,1)}(x)$, then $\pi = 4 \int_{[-1,1]^2} h(x) \cdot \frac{1}{4} dx$
- ▶ Monte Carlo estimate of π

$$\hat{I}_n = \frac{4}{n} \sum_{i=1}^n \mathbf{1}_{B(0,1)}(x^{(i)})$$

$$x^{(i)} \sim \text{Uniform}([-1, 1]^2)$$



Monte Carlo estimate of π (with 90% confidence interval)



- ▶ Convergence rate for Monte Carlo: $\mathcal{O}(n^{-1/2})$

$$p\left(|\hat{I}_n - I| \leq \frac{\sigma}{\sqrt{n\delta}}\right) \geq 1 - \delta, \quad \forall \delta$$

often slower than quadrature methods ($\mathcal{O}(n^{-2})$ or better)

- ▶ However, the convergence rate of Monte Carlo does not depend on dimensionality
- ▶ On the other hand, quadrature methods are difficult to extend to multidimensional problems, because of the curse of dimensionality. The actual convergence rate becomes $\mathcal{O}(n^{-k/d})$, for any order k method in dimension d
- ▶ This makes Monte Carlo strategy very attractive for high dimensional problems

- ▶ Monte Carlo methods require sampling a set of points chosen randomly from a probability distribution
- ▶ For simple distribution $f(x)$ whose inverse cumulative distribution functions (CDF) exists, we can sampling x from f as follows

$$x = F^{-1}(u), \quad u \sim \text{Uniform}(0, 1)$$

where F^{-1} is the inverse CDF of f

- ▶ Proof.

$$p(a \leq x \leq b) = p(F(a) \leq u \leq F(b)) = F(b) - F(a)$$



- ▶ Exponential distribution: $f(x) = \theta \exp(-\theta x)$. The CDF is

$$F(a) = \int_0^a \theta \exp(-\theta x) = 1 - \exp(-\theta a)$$

therefore, $x = F^{-1}(u) = -\frac{1}{\theta} \log(1 - u) \sim f(x)$. Since $1 - u$ also follows the uniform distribution, we often use $x = -\frac{1}{\theta} \log(u)$ instead

- ▶ Normal distribution: $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. **Box-Muller Transform**

$$X = \sqrt{-2 \log U_1} \cos 2\pi U_2$$

$$Y = \sqrt{-2 \log U_1} \sin 2\pi U_2$$

where $U_1 \sim \text{Uniform}(0, 1)$, $U_2 \sim \text{Uniform}(0, 1)$



- ▶ Assume $Z = (X, Y)$ follows the standard bivariate normal distribution. Consider the following transform

$$X = R \cos \Theta, \quad Y = R \sin \Theta$$

- ▶ From symmetry, clearly Θ follows the uniform distribution on the interval $(0, 2\pi)$ and is independent of R
- ▶ What distribution does R follow? Let's take a look at its CDF

$$\begin{aligned} p(R \leq r) &= p(X^2 + Y^2 \leq r^2) \\ &= \frac{1}{2\pi} \int_0^r t \exp\left(-\frac{t^2}{2}\right) dt \int_0^{2\pi} d\theta = 1 - \exp\left(-\frac{r^2}{2}\right) \end{aligned}$$

Therefore, using the inverse CDF rule, $R = \sqrt{-2 \log U_1}$

- ▶ If it is difficult or computationally intensive to sample directly from $f(x)$ (as described above), we need to use other strategies
- ▶ Although it is difficult to sample from $f(x)$, suppose that we can evaluate the density at any given point up to a constant $f(x) = f^*(x)/Z$, where Z could be unknown (remember that this makes Bayesian inference convenient since we usually know the posterior distribution only up to a constant)
- ▶ Furthermore, assume that we can easily sample from another distribution with the density $g(x) = g^*(x)/Q$, where Q is also a constant

- ▶ Now we choose the constants c such that $cg^*(x)$ becomes the envelope (blanket) function for $f^*(x)$:

$$cg^*(x) \geq f^*(x), \quad \forall x$$

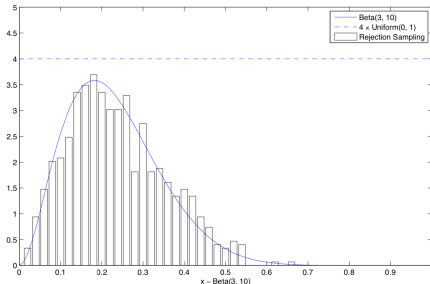
- ▶ Then, we can use a strategy known as *rejection sampling* in order to sample from $f(x)$ indirectly
- ▶ The rejection sampling method works as follows
 1. draw a sample x from $g(x)$
 2. generate $u \sim \text{Uniform}(0, 1)$
 3. if $u \leq \frac{f^*(x)}{cg^*(x)}$ we accept x as the new sample, otherwise, reject x (discard it)
 4. return to step 1

Rejection sampling generates samples from the target density, no approximation involved

$$\begin{aligned} p(X^R \leq y) &= p(X^g \leq y | U \leq \frac{f^*(X^g)}{cg^*(X^g)}) \\ &= p(X^g \leq y, U \leq \frac{f^*(X^g)}{cg^*(X^g)}) / p(U \leq \frac{f^*(X^g)}{cg^*(X^g)}) \\ &= \frac{\int_{-\infty}^y \int_0^{\frac{f^*(z)}{cg^*(z)}} dug(z) dz}{\int_{-\infty}^{\infty} \int_0^{\frac{f^*(z)}{cg^*(z)}} dug(z) dz} \\ &= \int_{-\infty}^y f(z) dz \end{aligned}$$



- ▶ Assume that it is difficult to sample from the Beta(3, 10) distribution (this is not the case of course)
- ▶ We use the Uniform(0, 1) distribution with $g(x) = 1, \forall x \in [0, 1]$, which has the envelop property: $4g(x) > f(x), \forall x \in [0, 1]$. The following graph shows the result after 3000 iterations

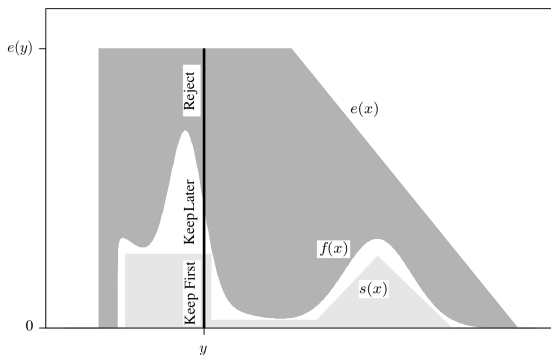


Rejection sampling becomes challenging as the dimension of x increases. A good rejection sampling algorithm must have three properties

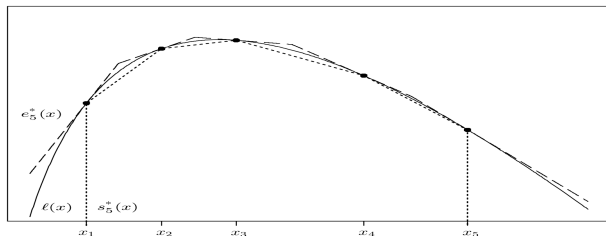
- ▶ It should be easy to construct envelopes that exceed the target everywhere
- ▶ The envelop distributions should be easy to sample
- ▶ It should have a low rejection rate

- ▶ When evaluating f^* is computationally expensive, we can improve the simulation speed of rejection sampling via *squeezed rejection sampling*
- ▶ Squeezed rejection sampling reduces the evaluation of f via a nonnegative squeezing function s that does not exceed f^* anywhere on the support of f : $s(x) \leq f^*(x), \forall x$
- ▶ The algorithm proceeds as follows:
 1. draw a sample x from $g(x)$
 2. generate $u \sim \text{Uniform}(0, 1)$
 3. if $u \leq \frac{s(x)}{cg^*(x)}$, we accept x as the new sample, return to step 1
 4. otherwise, determine whether $u \leq \frac{f^*(x)}{cg^*(x)}$. If this inequality holds, we accept x as the new sample, otherwise, we reject it.
 5. return to step 1



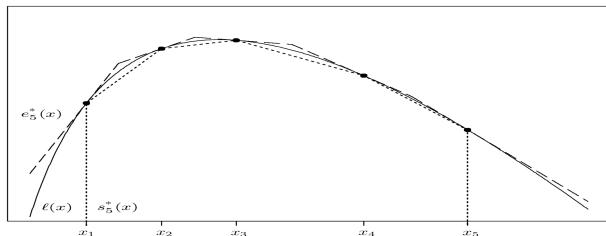


Remark: The proportion of iterations in which evaluation of f is avoided is $\int s(x)dx / \int e(x)dx$



- ▶ For a continuous, differentiable, log-concave density on a connected region of support, we can adapt the envelope construction (Gilks and Wild, 1992)
- ▶ Let $T = \{x_1, \dots, x_k\}$ be the set of k starting points.
- ▶ We first sample x^* from the piecewise linear upper envelope $e(x)$, formed by the tangents to the log-likelihood ℓ at each point in T_k .





- ▶ To sample from the upper envelop, we need to transform from log space by exponentiating and using properties of the exponential distribution
- ▶ We then either accept or reject x^* as in squeeze rejection sampling, with $s(x)$ being the piecewise linear lower bound formed from the chords between adjacent points in T
- ▶ Add x^* to T whenever the squeezing test fails.

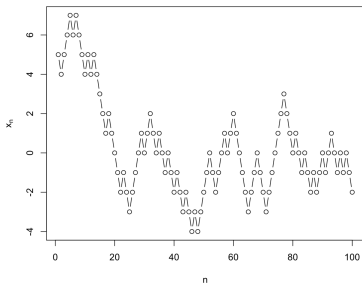


- ▶ For more complex distributions, we can use a Markov chain process to generate samples (which would not be independent anymore) and approximate the target distribution.
- ▶ This method is known as Markov chain Monte Carlo (MCMC) technique.
- ▶ However, we first need to discuss Markov chains and stochastic processes in general.

- ▶ Stochastic processes is a family of random variables, usually indexed by a set of numbers (time). A discrete time stochastic process is simply a sequence of random variables, X_0, X_1, \dots, X_n defined on the same probability space
- ▶ One of the simplest stochastic processes (and one of the most useful) is the simple random walk
- ▶ Consider a simple random walk on a graph $G = (\Omega, E)$. The stochastic process starts from an initial position $X_0 = x_0 \in \Omega$, and proceeds following a simple rule:

$$p(X_{n+1}|X_n = x_n) \sim \text{Discrete}(\mathcal{N}(x_n)), \forall n \geq 0$$

where $\mathcal{N}(x_n)$ denotes the neighborhood of x_n

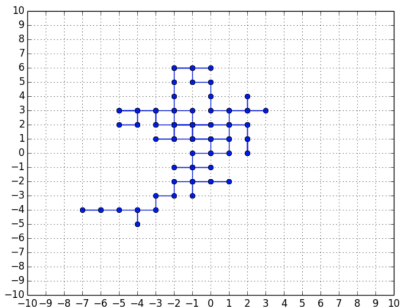
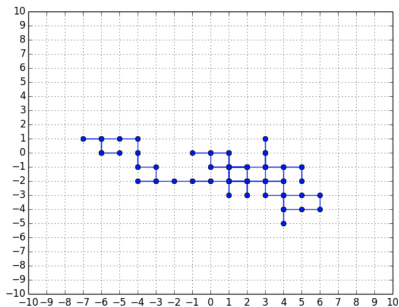


- ▶ Consider a sequence of iid random variables $\{Z_i\}$ such that $p(Z_i = 1) = p$, $p(Z_i = -1) = 1 - p$. A one dimension random work process can be defined as $X_0 = a$, $X_n = a + Z_1 + \cdots + Z_n$.
- ▶ The distribution of X_n

$$p(X_n = a + k) = \binom{n}{(n+k)/2} p^{(n+k)/2} (1-p)^{(n-k)/2}$$



Two random walks on a 20×20 grid graph



- ▶ The above simple random walk is a special case of another well-known stochastic process called *Markov chains*
- ▶ A Markov chain represents the stochastic movement of some particle in the state space over time. The particle initially starts from state i with probability $\pi_i^{(0)}$, and after that moves from the current state i at time t to the next state j with probability $p_{ij}(t)$
- ▶ A Markov chain has three main elements:
 1. A state space \mathcal{S}
 2. An initial distribution $\pi^{(0)}$ over \mathcal{S}
 3. Transition probabilities $p_{ij}(t)$ which are non-negative numbers representing the probability of going from state i to j , and $\sum_j p_{ij}(t) = 1$.
- ▶ When $p_{ij}(t)$ does not depend on time t , we say the Markov chain is time-homogenous



- ▶ Chain rule (in probability)

$$p(X_n = x_n, \dots, X_0 = x_0) = \prod_{i=1}^n p(X_i = x_i | X_{<i} = x_{<i})$$

- ▶ **Markov property**

$$p(X_{i+1} = x_{i+1} | X_i = x_i, \dots, X_0 = x_0) = p(X_{i+1} = x_{i+1} | X_i = x_i)$$

- ▶ Joint probability with Markov property

$$p(X_n = x_n, \dots, X_0 = x_0) = \prod_{i=1}^n p(X_i = x_i | X_{i-1} = x_{i-1})$$

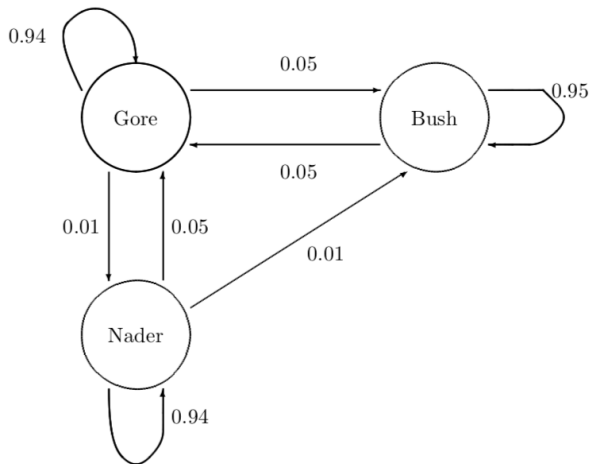
fully determined by the transition probabilities

- ▶ Consider the 2000 US presidential election with three candidates: Gore, Bush and Nader (just an illustrative example and does not reflect the reality of that election)
- ▶ We assume that the initial distribution of votes (i.e., probability of winning) was $\pi = (0.49, 0.45, 0.06)$ for Gore, Bush and Nader respectively
- ▶ Further, we assume the following transition probability matrix

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.94	0.05	0.01
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0.05	0.01	0.94



A probabilistic graph presentation of the Markov chain



- ▶ If we represent the transition probability a square matrix P such that $P_{ij} = p_{ij}$, we can obtain the distribution of states in step n , $\pi^{(n)}$, as follows

$$\pi^{(n)} = \pi^{(n-1)}P = \dots = \pi^{(0)}P^n$$

- ▶ For the above example, we have

$$\pi^{(0)} = (0.4900, 0.4500, 0.0600)$$

$$\pi^{(10)} = (0.4656, 0.4655, 0.0689)$$

$$\pi^{(100)} = (0.4545, 0.4697, 0.0758)$$

$$\pi^{(200)} = (0.4545, 0.4697, 0.0758)$$



- ▶ As we can see last, after several iterations, the above Markov chain converges to a distribution, $(0.4545, 0.4697, 0.0758)$
- ▶ In this example, the chain would have reached this distribution regardless of what initial distribution $\pi^{(0)}$ we chose. Therefore, $\pi = (0.4545, 0.4697, 0.0758)$ is the stationary distribution for the above Markov chain
- ▶ **Stationary distribution.** A distribution of Markov chain states is called to be stationary if it remains the same in the next time step, i.e.,

$$\pi = \pi P$$

- ▶ How can we find out whether such distribution exists?
- ▶ Even if such distribution exists, is it unique or not?
- ▶ Also, how do we know whether the chain would converge to this distribution?
- ▶ To find out the answer, we briefly discuss some properties of Markov chains

- ▶ Irreducible: A Markov chain is **irreducible** if the chain can move from any state to another state.
- ▶ Examples
 - ▶ The simple random walk is irreducible
 - ▶ The following chain, however, is reducible since Nader does not communicate with the other two states (Gore and Bush)

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.95	0.05	0
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0	0	1



- ▶ Period: the period of a state i is the greatest common divisor of the times at which it is possible to move from i to i .
- ▶ For example, all the states in the following Markov chain have period 3.

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

- ▶ Aperiodic: a Markov chain is said to be **aperiodic** if the period of each state is 1, otherwise the chain is periodic.

- ▶ **Recurrent** states: a state i is called recurrent if with probability 1, the chain would ever return to state i given that it started in state i .

	<i>Gore</i>	<i>Bush</i>	<i>Nader</i>
<i>Gore</i>	0.94	0.05	0.01
<i>Bush</i>	0.05	0.95	0
<i>Nader</i>	0.05	0.01	0.94

- ▶ **Positive recurrent:** a recurrent state j is called positive recurrent if the expected amount of time to return to state j given that the chain started in state j is finite
- ▶ For a positive recurrent Markov chain, the stationary distribution exists and is unique



- ▶ **Reversibility:** a Markov chain is said to be reversible with respect to a probability distribution π if $\pi_i p_{ij} = \pi_j p_{ji}$
- ▶ In fact, if a Markov chain is reversible with respect to π , then π is also a stationary distribution

$$\begin{aligned}\sum_i \pi_i p_{ij} &= \sum_i \pi_j p_{ji} \\ &= \pi_j \sum_i p_{ji} \\ &= \pi_j\end{aligned}$$

since $\sum_i p_{ji} = 1$ for all transition probability matrices

- ▶ This is also known as *detailed balance condition*.



- ▶ We can define a Markov chain on a general state space \mathcal{X} with initial distribution $\pi^{(0)}$ and transition probabilities $p(x, A)$ defined as the probability of jumping to the subset A from point $x \in \mathcal{X}$
- ▶ Similarly, with Markov property, we have the joint probability

$$p(X_0 \in A_0, \dots, X_n \in A_n) = \int_{A_0} \pi^{(0)}(dx_0) \dots \int_{A_n} p(x_{n-1}, dx_n)$$

- ▶ Example. Consider a Markov chain with the real line as its state space. The initial distribution is $\mathcal{N}(0, 1)$, and the transition probability is $p(x, \cdot) = \mathcal{N}(x, 1)$. This is just a **Brownian motion** (observed at discrete time)

- ▶ Unlike the discrete space, we now need to talk about the property of Markov chains with a continuous non-zero measure ϕ , on \mathcal{X} , and use sets A instead of points
- ▶ A chain is ϕ -irreducible if for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ and for all $x \in \mathcal{X}$, there exists a positive integer n such that

$$p^n(x, A) = p(X_n \in A | X_0 = x) > 0$$

- ▶ Similarly, we need to modify our definition of period

- ▶ A distribution π is a stationary distribution if

$$\pi(A) = \int_{\mathcal{X}} \pi(dx)p(x, A), \quad \forall A \subseteq \mathcal{X}$$

- ▶ As for the discrete case, a continuous space Markov chain is reversible with respect to π if

$$\pi(dx)p(x, dy) = \pi(dy)p(y, dx)$$

- ▶ Similarly, if the chain is reversible with respect to π , then π is a stationary distribution
- ▶ Example. Consider a Markov chain on the real line with initial distribution $\mathcal{N}(1, 1)$ and transition probability $p(x, \cdot) = \mathcal{N}(\frac{x}{2}, \frac{3}{4})$. It is easy to show that the chain converges to $\mathcal{N}(0, 1)$ ([Exercise](#))



- ▶ Ergodic: a Markov chain is ergodic if it is both irreducible and aperiodic, with stationary distribution π
- ▶ **Ergodic Theorem.** For an ergodic Markov chain on the state space \mathcal{X} having stationary distribution π , we have: (i) for all measurable $A \subseteq \mathcal{X}$ and π -a.e. $x \in \mathcal{X}$,

$$\lim_{t \rightarrow \infty} p^t(x, A) = \pi(A)$$

(ii) $\forall f$ with $\mathbb{E}_\pi |f(x)| < \infty$,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(X_t) = \int_{\mathcal{X}} f(x) \pi(x) dx, \quad \text{a.s.}$$

In particular, π is the **unique** stationary probability density function for the chain



- ▶ P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Academic, New York, 1984.
- ▶ W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.