Markov Chain Monte Carlo

Uri Shaham

January 20, 2025

1 Monte Carlo Methods

Monte Carlo simulation often refers to the estimation of deterministic quantities via sampling (e.g., estimating means using averages). For example, we can estimate the number π by sampling points in a 2d square with vertices at $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$. Defining an event A as 1 if the sampled point lies inside the unit circle and 0 otherwise, we have $\mathbb{E}[A] = \frac{\text{area of the circle}}{\text{area of the square}} = \frac{\pi}{4}$, so we can estimate π by the ratio of the number of points inside the circle to the total number of points.

Interestingly, Monte Carlo methods can sometimes be used to overcome the curse of dimensionality, as in the following example.

1.1 Numerical integration

Let $A = [0,1]^d$, and let $f : A \to \mathbb{R}$. We are interested in $\int_A f(x)dx$. We can define an ϵ-grid (i.e., a grid with with spacing ϵ, so the number of points is $n = \epsilon^{-d}$) with grid points x_i and then have

$$\int_A f(x)dx \approx \frac{1}{n}\sum_i f(x_i).$$

This doesn't scale well, since n is exponential in the dimension d. This is the curse of dimensionality (volume is exponential in the dimension).

Surprisingly, replacing the grid with uniform sampling can help. Let $\mathcal{P}$ be a uniform distribution on A with density p, i.e., $p(x) = 1$ if $x \in A$ and $p(x) = 0$ otherwise. Then

$$\int_A f(x)dx = \int f(x)p(x)dx = \mathbb{E}[f(x)].$$

Let's look at $\frac{1}{n}\sum_i f(x_i)$. Its expectation is

$$\mathbb{E}\left[\frac{1}{n}\sum_i f(x_i)\right] = \frac{1}{n}\sum_i \mathbb{E}[f(x_i)] = \mathbb{E}[f(x)] = \int_A f(x)dx,$$

i.e., it is an unbiased estimator of our desired integral. To find its rate of convergence, let's look at the

variance of the estimator:

$$\mathbb{E}\left(\frac{1}{n}\sum_i f(x_i) - \int_A f(x)dx\right)^2 = \mathrm{Var}\left(\frac{1}{n}\sum_i f(x_i)\right)$$
$$= \frac{1}{n^2}\sum_i \mathrm{Var} f(x_i)$$
$$= \frac{1}{n}\mathrm{Var}(f(x)), \quad (1)$$

so the rate of convergence is $\frac{1}{\sqrt{n}}$, regardless of the dimension!

2 Markov Chain Monte Carlo

Definition 2.1 (Markov chain)**.** *A Markov chain is a sequence of random variables $X_0, X_1, \ldots$, taking values from a (finite or infinite) state space $\mathcal{S} = \{1, 2, \ldots\}$, with the property that*

$$\Pr(X_n = i_n | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \Pr(X_n = i_n | X_{n-1} = i_{n-1}).$$

A Markov chain is specified by

- Initial distribution π_0 over S
- Transition rule. If $|\mathcal{S}| = N$, and the chain is time-homogeneous (i.e., the transition probabilities do not change over time), then this rule is a $N \times N$ matrix P, such that $P_{ij} = \Pr(X_n = j | x_{n-1} = i)$.

For any distribution π_0 of states, the distribution after one transition is given by $\pi_1^T = \pi_0^T P$.

Definition 2.2 (Stationary distribution)**.** *A stationary distribution is a vector π, such that $\pi^T P = \pi^T$.*

Definition 2.3 (Detailed balance)**.** *A Markov chain is called reversible if $\pi_0 = \pi$, and for all i, j, $\pi(i)P_{ij} = \pi(j)P_{ji}$. The last equation is called detailed balance.*

Detailed balance is a sufficient condition for the existence of stationary distribution (see homework).

3 The Metropolis-Hastings algorithm

MCMC methods are designed to obtain samples from a desired distribution when the distribution itself is only known up to a multiplicative factor. Let f be a positive function over a $\mathcal{S}$, which corresponds to a distribution given by $\pi(i) = \frac{f(i)}{\sum_i f(i)}$. To know π, we have to compute the denominator, which involves summation over possible very large or even infinite space, which is a problem. Metropolis Hastings lets us to sample from π, given the mere ability to compute f, and without requiring knowledge of $\sum_i f(i)$. This works by designing a Markov chain whose stationary distribution is π.

Given any proposal transition distribution $Q = \{Q(i|j)\}$ specifying the probabilities to propose state j given that the current state is i (and assume $Q(i|j)$ is positive for all i, j), we define the following transition matrix

$$P_{ij} = \begin{cases} Q(i|j) \min\left\{1, \frac{f(j)Q(i|j)}{f(i)Q(j|i)}\right\}, & i \neq j \\ 1 - \sum_{i \neq j} P_{ij}, & i = j. \end{cases} \quad (2)$$

Lemma 3.1. *The transition matrix defined by* (2) *satisfies $\pi^T P = \pi^T$.*

Proof. Let i, j be such that $i \neq j$. Then

$$\pi(i)P_{ij} = \frac{f(i)}{\sum_i f(i)} Q(j|i) \min\left\{1, \frac{f(j)Q(i|j)}{f(i)Q(j|i)}\right\} \propto \min\{f(i)Q(j|i), f(j)Q(i|j)\},$$

where $\propto$ means that this holds up to a multiplicative constant which does not depend on i, j. This is symmetric in i, j, hence $\pi(i)P_{ij} = \pi(j)P_{ji}$, i.e., detailed balance is satisfied for $i \neq j$, and trivially also for $i = j$. □

Note that since $\min\left\{1, \frac{f(j)Q(i|j)}{f(i)Q(j|i)}\right\}$ might be less than 1, it can be interpreted as a probabilistic decision to move from state i to state j, i.e., being at state i, state j is proposed and we move it with probability $\min\left\{1, \frac{f(j)Q(i|j)}{f(i)Q(j|i)}\right\}$, and with the remaining probability we stay at state i. The above is translated to the following sampling algorithm:

1. Initialize:
   (a) pick initial state i.
   (b) set $t = 0$.
2. Iterate:
   (a) sample a proposed state from $Q(j|i)$
   (b) Calculate the acceptance probability $A(i, j) = \min\left\{1, \frac{f(j)Q(i|j)}{f(i)Q(j|i)}\right\}$
   (c) With probability $A(i, j)$ accept j and set $x_t = j$. Otherwise $x_t = i$.
   (d) $t \leftarrow t + 1$.

Note that as we typically start from some arbitrary state distribution π_0, we should run the chain for some time to let the state distribution converge to π before we start collecting samples.

3.1 Application: numerical integration

Let $X \in \Omega$ be a random variable with density f, where Ω is a bounded region of $\mathbb{R}$, and let $s = s(X)$ be some statistic of X. Suppose we like to estimate $\mathbb{E}[s]$ on the tail $A \subset \Omega$ of f. This expectation is

$$\mathbb{E}[s|x \in A] = \int_\Omega f(x|x \in A)s(x)dx.$$

A straightforward Monte Carlo integration would draw samples from Ω corresponding to f, and estimate the integral by

$$\sum_{x \in A} \frac{1}{|\{x : x \in A\}|} s(x).$$

However, samples from A will be rare, by definition. MCMC can be utilized by using a proposal distribution that favors A.

3.2 Sampling from posterior

In Bayesian statistics, we estimate the posterior distribution of model parameters by

$$p(\theta|x) = \frac{p(\theta)\pi(x|\theta)}{p(x)} = \frac{p(\theta)\pi(x|\theta)}{\int_\theta p(\theta)p(x|\theta)}.$$

$p(\theta)$ is a prior distribution corresponding to our belief. $p(x|\theta)$ is typically given by our model. However, computing the denominator is often intractable because of the integration. MCMC lets us sample from $p(\theta|x)$ without knowing the denominator.

4 Gibbs Sampler

Gibbs sampler is a MCMC method for sampling high dimensional data, using conditional distributions. Specifically, let $x_t \in \mathbb{R}^d$ be a sample at time t. x_{t+1} is sampled from x_t by sampling the i'th entry from

$$p(\cdot|x_t[1], \ldots, x_t[i-1], x_t[i+1], \ldots, x[d]) := p(\cdot|x_t[-i]).$$

This is efficient, for example, in Restricted Boltzmann machines. To see why this works, note that

$$p(x[i]|x[-i]) = \frac{p(x)}{p(x[-i])},$$

i.e., if $x_t[-i]$ is sampled from the "$x_t[-i]$ - marginal", then sampling $x_{t+1}[i]$ from the conditional gives us a sample from the joint.

4.1 Connection between Gibbs sampler and MH

To see the connection of Gibbs sampling with MH, let's compute the MH acceptance probability, with $f(x) = p(x)$ and $Q(x_{t+1}|x_t) = p(x_{t+1}[i]|x_t[-i])$. Then

$$\begin{aligned}
A(x_t, x_{t+1}) &= \min\left\{1, \frac{p(x_{t+1})p(x_t[i]|x_{t+1}[-i])}{p(x_t)p(x_{t+1}[i]|x_t[-i])}\right\} \\
&= \min\left\{1, \frac{p(x_{t+1}[i]|x_t[-i])p(x_t[-i])p(x_t[i]|x_{t+1}[-i])}{p(x_t[i]|x_{t+1}[-i])p(x_{t+1}[-i])p(x_{t+1}[i]|x_t[-i])}\right\}. \\
&= \min\left\{1, \frac{p(x_{t+1}[-i])}{p(x_t[-i])}\right\}. \\
&= 1, \qquad (3)
\end{aligned}$$

where the last equality holds since in the transition from x_t to x_{t+1}, we only change the i'th entry. Thus Gibbs sampler can be viewed as a special case of MH, where the candidate is always accepted.