This article demonstrates my personal understanding of the **MCMC** techniques and its variations (Metropolis–Hastings, Gibbs sampling)

Some reference articles that inspire this brief summary are:

- UCSD CSE 250-A: Probabilistic Reasoning and Decision-Making
- [Personal Note from Wheaton College MATH 364: Math Modeling]
- MIT Note: Gibbs Sampling
- An Introduction to MCMC for Machine Learning

Bayesian Rule

Suppose we need to estimate some parameter θ given the observation data X. In order to calculate the probability distribution of θ given X, we will need implement Bayes rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

However, the problem of Bayes rule exists in the denominator. If we apply marginalization to $P(X)$, we get:

$$P(X) = \int_{\theta} P(x, \theta) d\theta$$

This means that to obtain the probability of the evidence node X, we need to integrating the quantity over **all possible parameter values**, and sometimes parameters can have a continuous range of values. **This makes the posterior impossible to compute with a closed-form solution**.

If we cannot solve the posterior **analytically**, how about we estimate $P(\theta|X)$ **quantitatively**?

Monte Carlo Simulation

The idea of Monte Carlo simulation presents a promising cure to the difficulty we experience from Bayes rules. The **theorem** states that if we draw an i.i.d (independently identically distributed) sample sets $x^{(i)}{}_{i=1}^{N}$ from a target density $p(x)$ defined on a high-dimensional space X, we are able to approximate the integrals $I(f)$ with tractable sum $I_N(f)$ that converges as follows

$$I_N(f) = \frac{1}{N} \sum_{i=1}^{N} f(x^(i)) \rightarrow I(f) = \int_X f(x)p(x)dx$$

where $n \to \infty$

But though this sampling is effective and promise convergence, for $p(x)$ that is not gaussian distributed, N can be very large til convergence. So the question remain: is there a sampling method that guarantees more effective convergence over $P(X)$?

Markov Chain (Recap)

Definitions:

- A **Markov Chain (MC)** is a stochastic model for a system that
  - may be in one of a finite number N of **states** $1, 2, \cdots, N$
  - may **transition** from one state to another
  - is observed at discrete **steps** (or **stages**)
- The **transition probability** p_{ij} is the probability that the system will transition from state i to state j at the next step. A Markov chain assumes that these transition probabilities are **constant** and **memoryless** (i.e. they do not depend on time or on the history of the system's past states). Because the system can only be in a finite number of states, the sum of transition probabilities from a state i to all other states (including i) must equal **1** ($\sum_{j=1}^{N} p_{ij} = 1$ for all i)
- The **state variable** $x_{k,n}$ represents the probability that the system is in state k at step n. The **state vector** $\vec{x_n}$ is the row vector $(x_{1,n}, x_{2,n}, x_{3,n}, \cdots, x_{N,n})$ that represents the probabilities that the system will be in the different states at step n. Because the system can only be in a finite number of states, the sum of the state variables at a given step must equal **1** ($\sum_{k=1}^{N} x_k = 1$)

Markov Chain as a Linear System

The evolution of a Markov chain can be represented as a linear discrete dynamical system:

$$[x_{1,n+1}, x_{2,n+1}, x_{1,n+1}, \cdots, x_{N,n+1}] = [x_{1,n}, x_{2,n}, x_{1,n}, \cdots, x_{N,n}] \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}$$

In a compact form, $x_{n+1} = x_n P$, where P is the transition matrix. With its closed-form solution:

$$x_k = x_0 P^k$$

Special Markov Chains

- A Markov chain is **ergodic** if the system can transition from any state to any other in a finite number of steps.
- A Markov chain is **regular** if all there exists some integer k such that all the elements of P^n are nonzero for $n \geq k$. (This implies that the system can transition from any state to any other in k steps, so a regular MC is always ergodic)
- An **absorbing state** i is a state such that $p_{ii} = 1$. A Markov chain is an absorbing Markov chain if it has one or more absorbing state, and the system is absorbed into one of them in the long run.

Fundamental Limit Theorem

Theorems

Let P be the transition matrix of a regular Markov chain. Then $\lim_{n\to\infty} P^n = W$, where all rows of W are the same vector $\vec{w}$, called the **steady state row vector**. This vector has the following properties:

- $\vec{w} = \lim_{n\to\infty} \vec{x_n}$ This is the long term state vector regardless of its initial state $\vec{x_0}$. So in the long run, the system "forgets" its initial state and settles onto the state vector $\vec{w}$
- $\vec{w}P = \vec{w}$ If the system's state vector is $\vec{w}$, it does not change at the next step.

The vector $\vec{w}$ can be found computationally (by calculating P^n for large n) or analytically (by solving $\vec{w}(P - I) = \vec{0}$ for $\vec{w}$)

Absorbing Markov Chains

If the absorbing states of an absorbing Markov chain are listed last, the its transition matrix can be divided into four block submatrices:

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

where

- Q is the transition submatrix between nonabsorbing states

- R is the transition submatrix from nonabsorbing states to absorbing states
- 0 is a matrix of zeroes
- I is an identity matrix

These matrices have the following properties:

- $\lim_{n\to\infty} Q^n = 0$. (Because the elements of each row of Q do not all add up to 1, Q^n eventually decays to zero as $n \to \infty$)
- The multistep transition matrix is $P^n = \begin{bmatrix} Q^n & (\sum_{k=1}^{n-1} Q^k)R \\ 0 & I \end{bmatrix}$ In the long term, the geometric sum equals $(I-Q)^{-1}$, so the matrix becomes:

$$\lim_{n\to\infty} P^n = \begin{bmatrix} 0 & (I-Q)^{-1}R \\ 0 & I \end{bmatrix}$$

* The matrix $(I-Q)^{-1}$ is called the **fundamental matrix of the markov chain**. The sum of elements in row i of this matrix is the expected time to absorption when starting at nonabsorbing state i. * The matrix $B = (I-Q)^{-1}R$ is the long term transition submatrix from nonabsorbing states to absorbing states. Each element b_{ij} represents the probability of long-term absorption into state j, when starting in nonabsorbing state i

Transition Matrix

Markov Chain provides us with an effective sampling method by **guarantee a more strategic acquisition of the samples** through the transition matrix in the Markov model. In this case, in order to have an ergodic MC model, we need to design a transition matrix T with these properties:

- **Irreducibility**: For any state of the Markov chain, there is a positive probability of visiting all other states.
- **Aperiodicity**: The chain should not get trapped in cycles.

A sufficient, but not necessary, condition to ensure that a particular $p(x)$ is the desired invariant distribution is the following reversibility (detailed balance)

$$p(x^{(i)})T(x^{i-1}|x^i) = p(x^{(i-1)})T(x^i|x^{i-1})$$

$$p(x^{(i)}) = \sum_{(x^{(i-1)})} p(x^{(i-1)})T(x^i|x^{i-1})$$

MCMC samplers are irreducible and aperiodic Markov chains that have the target distribution as the invariant distribution. One way to design these samplers is to ensure that detailed balance is satisfied. However, it is also important to design samplers that **converge quickly**.

MCMC (Metropolis-Hastings)

The general idea of Markov Chain Monte Carlo sampling method is to keep the sample roaming around the important portion of the distribution (normally a peak in the probabilistic space). However, since we cannot just simply discard the low probability, we use a proposal distribution matrix to simulate this uncertainty.

Proposal Distribution

The concept of proposal distribution matrix in MCMC is very similar to the transition matrix in Markov chain, and it is up to the design of proposal acceptance matrix that guarantees convergence and effective sampling results.

Metropolis-Hastings

Gibbs Sampling
