MARKOV CHAIN MONTE CARLO: FROM THEORY TO METHOD

Deqian Kong[*]
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
deqiankong@ucla.edu

Shi Feng[*]
Department of Physics
The Ohio State University
Columbus, OH 43210
feng.934@osu.edu

1 Introduction

The essence of Markov Chain Monte Carlo (MCMC) method is to solve a problem by mapping it onto an iterative sampling problem of statistics, whereby the sampling procedure is governed by an engineered kernel so that the iteration converges to the result of the original problem. This is particularly useful when dealing with problems which have exponentially large searching space which makes them hard to solve by enumeration, or where there is a computational wall that is hard to penetrate by existing algorithms, like exactly diagonalizing a huge Hamiltonian.

Although MCMC is widely used to simulate statistical ensembles (e.g. thermal average), the target of the original problem need not have statistical essence (e.g. Finding the ground state wavefunction of a Hamiltonian). In the latter case, iterations in MCMC serves as a statistical detour around the computational wall that stands between the destination and the starting point, and may be perceived as a useful intermediary redundancy which is ultimately to be removed by convergence at the fixed point.

Feng: add more examples in physics

What does a physicist mean by sampling*?*

Statisticians and physicists use the word sampling not in exactly the same convention. In the view of a physicist (dummy physicist like Shi), sampling simply means imposing an observable operator (a Hermitian matrix) $\hat{O}$ on a system's Hilbert space for multiple times, which involves an average over either quantum or thermal ensemble, or both. In other words, given an observable $\hat{O}$, some information is inevitably coarse-grained by the sampling thus not detectable even though the statistical mechanics of the underlying microscopics are well-modeled. This perception is a top-down picture, whereby details of the system cannot be thoroughly pinned down as we stand at the top side, and the sampled data are perceived as shadows of true physical laws distorted and coarse-grained by probes. Such deficiency of physical sampling will always be with us, and every physicist has to learn to live with it. Nonetheless, as statisticians will tell us in MCMC, we sometimes can utilize this fact as our strengths to give predictions of physical properties with decent accuracy.

What does a statistician mean by sampling*?*

A statistician will argue, as the resolution of our probe will never be enough to pin down every detail of the underlying mechanics, why not just coarse-grain the theory at the first place, thus calculations can be rendered easier. Hence, in statistics, sampling means the selection of a subset of all theoretical elements, or rather, distributions, to resemble the essence of the theory such that we can give a good enough prediction with lower cost. The simplest example will be the uniform random sampling.

Suppose we want to evaluate the integral:

$$J = \int f(x)p(x)dx \tag{1.1}$$

[*] This is a pedagogical paper on MCMC

where $p(x)$ is a probability density distribution, $f(x)$ is some physical property that is dependent on microscopic states x. Naively all we need to do is evaluate the integral by brute force and get the number output. But instead of doing such a verbose calculation, a simpler way is to obtain independent and *evenly distributed* samples $\{x_1, x_2, \dots, x_N\}$ from $p(x)$, and calculate

$$J = \sum_{i}^{N} f(x_i)p(x_i)/N \tag{1.2}$$

But the problem with this method is that the sampling resolution has to be extremely sharp when the density of states is huge somewhere. Hence if the theoretical distribution $p(x)$ is spiky at a few x, we have to make significantly more sample points in order to tackle those peaks, even though the rest of $p(x)$ are flat that doesn't cost much. There is another preferred way that addresses this problem, whereby samples are picked up in such a way that they resembles the key information of the continuous distribution $p(x)$, thus J can be evaluated by

$$J = \sum_{i}^{N} f(x_i)/N \tag{1.3}$$

which should give a decent approximation of the original theory. In case of the canonical ensemble with Boltzmann distribution, the probability density $p(x) \propto \exp\{-\beta E\}$ has most of its weight close to $E = 0$, and a thin, long tail at higher E. Hence to do the aforesaid sampling, the subset of points that we need to pick up from $p(x)$ need to concentrate more at low energy and become sparse at high energy, so the essential information is captured. In MCMC, it is equivalent to

$$J = \sum_{t=0}^{N} f(x_t)/N \tag{1.4}$$

where each x_t is a configuration sampled at time t generated by some MCMC kernel, and it is expected to converge to the true J when N is large enough. This is exactly what the pioneers of MCMC in physics community did in [**?**] where authors used such nonuniform sampling to calculate thermal averages in canonical ensembles.

2 Markov Chain Basics

In this section we introduce the generic routine of MCMC, that is, to find a iteration kernel K that leads to the convergence to desired result, in which the final fixed point must satisfy the detailed balance condition (thus a global balance).

The current state in a Markov chain only depends on the most recent previous states, i.e,

$$P(X_t | X_{t-1}.X_{t-2}, \dots, X_0) = P(X_t | X_{t-1})$$

Definition 2.1 (Markov Chain). $MC = (\Omega, \hat{v}_0, \hat{K})$, where Ω is the state space, $\hat{v}_0 : \Omega \to \mathbb{R}$ is the initial probability distribution function over the states, $\hat{K} : \Omega \times \Omega \to \mathbb{R}$ is the transition probability function. Where the hats on $\hat{K}$ and $\hat{v}_0$ are used to emphasize they are essentially maps instead of numbers.

Remark: In many places the Markov Chain are defined without hats as $MC = (\Omega, v_0, K)$. This may lead to the confusion in expression like $v_0 K$ as it is usually written. Because maps do not multiply, but only interact via composition i.e. $v_0 \circ K$. Yet most of times $v_0 K$ are used to describe the probability distribution, which is a tuple of real numbers instead of a composite map. Therefore we use the hats to distinguish probability distribution from probability distribution function. In the forthcoming texts, we will denote the evaluated probability function by lower case letters. For example

$$v_0 = \hat{v}_0 \cdot \Omega, \quad K = \hat{K} \cdot (\Omega \times \Omega)$$

where v_0 and K can be perceived as real-valued vector and matrix. Elements in vector v_0 are real-valued probability of all configurations $\omega_i \in \Omega$, and elements in matrix K are conditional probabilities that connect pairs of configurations.

Example 1 In the simplest Ising model on an $a \times b = N$ sites with $S_i = \pm 1$, the state space Ω is the collection of all configurations i.e. $\Omega = \bigotimes^N \mathbb{Z}_2$, with the total number of states $\#\Omega = 2^N$. Each element $\omega \in \Omega$ is a N dimensional tuple whose elements have value ± 1. The initial probability distribution is denoted by the evaluated probability function v_0, a real-valued tuple, where the subscript 0 says that the probability distribution is at zeroth iteration. Of course the initial probability distribution v_0 is arbitrary thus not the desired distribution which we are trying to sample. Our goal is then to find a way to evolve v_0 to $v_1, \dots, v_n$, hoping such a series would ultimately converge to the true probability distribution, e.g. the Boltzmann weights $e^{-\beta E_\omega}$ for the Ising model in a canonical ensemble. This evolution of probability distribution is described by the aforementioned K.

At time n, the Markov Chain state will follow a probability for finite states, and the state converge to an invariant probability,

$$v_n = v_0 K^n \text{ and } \lim_{n\to\infty} v_0 K^n = \pi$$

Our objective is to design a Markov chain kernel K, such that π is the unique, invariant probability of K (a fixed point). Suppose we are given Ω and a target probability $\pi = (\pi_1, \cdots, \pi_N)_{(1\times N)}$, our goal is to design v_0 and K so that $\pi K = \pi$, which is a necessary condition.

Here we check the conditions for topology of transition matrix:

- stochastic matrix: $\sum_{j=1}^{N} K_{ij} = 1, \forall i \in \Omega, K_{ij} \geq 0$ or $K\mathbf{1}_{N\times 1} = \mathbf{1}$
- global balance: $\pi_{1\times N} K = \pi$, $\sum_{i=1}^{N} \pi_i K_{ij} = \pi_j, \forall j \in \Omega$
- detailed balance(*reversible*): $\pi_i K_{ij} = \pi_j K_{ji}, \forall i, j \in \Omega$

  We should know that the detailed balance is a sufficient but not necessary condition for global balance and we should know detailed balance indicates stationarity and in particular global balance,

  $$\pi K = \sum_{i=1}^{n} \pi_i [K_{i1}, \cdots, K_{iN}] = \sum_{i=1}^{n} [\pi_1 K_{1i}, \cdots, \pi_N K_{Ni}] = \pi$$

  $$\sum_{i=1}^{N} \pi_i K_{ij} = \pi_j \sum_{i=1}^{N} K_{ji} = \pi_j$$

- irreducibility: A Markov Chain is irreducible if its transition matrix K has only one communication class. $i \to j$, denotes a state j is accessible from i, if there exists a finite step M, such that

  $$K_{ij}^{M} = \sum_{i_1, i_2, \cdots, i_{M-1}} K_{ii_1} K_{i_1 i_2} \cdots K_{i_{M-2} i_{M-1}} K_{i_{M-1} j} > 0$$

  $i \leftrightarrow j$ generates a partition of the state space into disjoint equivalence(communication) classes given by $\Omega = \cup_{i=1}^{C} \Omega_i$ and one communication class means all the states are accessible from each other
- Aperiodicity: to define this, we need to define a periodic MC first. An irreducible MC with transition matrix K has period d if there is a unique patition of graph G into cyclic classes:

  $$C_1, C_2 \cdots, C_d, \sum_{j\in C_k} K_{ij} = 1, \forall i \in C_{k-1}$$

  In an periodic Markov Chain there is no connection between states within each individual cyclic class, and an irreducible Markov chain with transition matrix K is aperiodic if it's largest period is $d = 1$.
- stationarity distribution: A Markov chain with transition kernel K has stationary distribution π if

  $$\pi K = \pi$$

  There may be many stationary distributions w.r.t K. Even if there is a stationary distribution, a Markov chain may not always converge to it.

Theorem 2.1 (Perron-Frobenius*). *For any primitive (irreducible and aperiodic) $N \times N$ stochastic matrix K, with eigenvalues*

$$\lambda_1 > |\lambda_2| > \cdots > |\lambda_r|$$

and multiplicities as $m_1, m_2, \cdots, m_r$, with $\boldsymbol{u}_1 = \pi, \boldsymbol{v}_1 = \mathbf{1}$ has the biggest eigenvalue $\lambda = 1$ with $m_1 = 1$.

Proofs can be found here in both numerical($tr(K) = \sum_{i=1}^{n} K_{ii} = \sum_{i=1}^{r} m_i \lambda_i$) and geometric perspective(by defining sphere $x_1^2 + x_1^2 + \cdots + x_n^2 = 1$). More detailed one can be found here.

For a square matrix, we have eigen decomposition as $K = Q\Lambda Q^{-1}$ and as $n \to \infty$,

$$K^n = Q\Lambda Q^{-1} = Q\Lambda^n Q^{-1} \to \lambda_1 \boldsymbol{v}_1 \boldsymbol{u}_1 = \mathbf{1}\pi$$

2.1 A worked example of balance equations

The global balance equations are a set of equations that characterize the (dynamical) equilibrium distribution (or any stationary distribution, for example, a composite chemical system in a dynamical equilibrium).

Suppose at time t we have some initial pdf configuration $\pi^{(0)} = (p_1, p_2)$ which may be perceived as a discrete pdf of two distinct chemical compounds, and the transition matrix K is defined:

$$K = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}.$$

that is, at time $t+1$ the new distribution π' is:

$$\pi^{(1)} = \pi^{(0)} K = \begin{pmatrix} p_1 & p_2 \end{pmatrix} \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} = (p_1 k_{11} + p_2 k_{21},\ p_1 k_{12} + p_2 k_{22}) \tag{2.1}$$

with the row-major index notation we can write:

$$\pi_i^{(1)} = \sum_j \pi_j^{(0)} k_{ji}, \quad \text{with } i, j \in \{1, 2\} \tag{2.2}$$

If the system reaches an equilibrium state at time T, then the state at next moment $T+1$ should remain the same (to be precise, they are element-wise the same), that is:

$$\pi_i^{(T+1)} = \sum_j \pi_j^{(T)} k_{jk} = \pi_i^{(T)} \tag{2.3}$$

For short, we say at equilibrium we have a fixed point characterized by

$$\boxed{\pi_i = \sum_j \pi_j k_{ji}} \tag{2.4}$$

In the context of probability theory, let Ω be the total state space, and let $\mathbf{a}, \mathbf{b} \in \Omega$ be two different states. Suppose at time t the probability for the system to stay in state $\mathbf{a}$ is $p(\mathbf{a})$, and the probability of state $\mathbf{a}$ to transit into state $\mathbf{b}$ at next moment $t+1$ is $p(\mathbf{a}_t \to \mathbf{b}_{t+1})$. Therefore, the probability for the system to stay in state $\mathbf{b}$ at $t+1$ and in $\mathbf{a}$ at the previous moment is:

$$p(\mathbf{b}_{t+1}, \mathbf{a}_t) = p(\mathbf{a}_t) p(\mathbf{a}_t \to \mathbf{b}_{t+1}) \tag{2.5}$$

note that we have implicitly used that $p(\mathbf{a} \to \mathbf{b})$ is essentially a conditional distribution i.e. $p_t(\mathbf{a} \to \mathbf{b}) \equiv p_t(\mathbf{b}_{t+1} | \mathbf{a}_t)$. Now note that not only can $\mathbf{a}$ transit to $\mathbf{b}$, but there can be many other states which have non-zero probability to transit into $\mathbf{b}$ in the next moment. Therefore, the probability of the system to stay in state $\mathbf{b}$ is

$$p(\mathbf{b}_{t+1}) = \sum_n p(\mathbf{b}_{t+1}, \mathbf{n}_t) = \sum_n p(\mathbf{n}_t) p(\mathbf{n}_t \to \mathbf{b}_{t+1}) \tag{2.6}$$

Now we can identify $p(\mathbf{b}_{t+1})$ as $v_b^{(t+1)}$: the probability that the system be at $b-$th state at the $t+1$, and $p(\mathbf{a}_t \to \mathbf{b}_{t+1})$ as k_{ab}: the probability that the system transits from a-th state at t to b-th state at $t+1$:

$$\begin{aligned} p(\mathbf{b}_{t+1}) \quad &\Longleftrightarrow \quad v_b^{(t+1)} \\ p(\mathbf{a}_t \to \mathbf{b}_{t+1}) \quad &\Longleftrightarrow \quad k_{ab} \end{aligned}$$

Therefore the global interation can be written as:

$$v_b^{(t+1)} = \sum_j v_a^{(t)} k_{ab} \tag{2.7}$$

or in the matrix form:

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} \mathbf{K} \tag{2.8}$$

if the a fixed poin is reached at T then:

$$\mathbf{v}^{(T)} = \mathbf{v}^{(T+1)} = \mathbf{v}^{(T)} \mathbf{K} \tag{2.9}$$

3 Statistical Error Analysis and Binning

Note: MCMC samples are correlated. Discuss the independent case, then discuss the correlated case. Error bar can be crucial in Baysian mothods. Call central limit theorem.

Algorithm 1 The Metropolis-Hastings Algorithm

```
Input: Target probability distribution $\pi(x)$, current state $x^{(t)} \in \Omega$, and the proposal probability distribution $Q(x, y)$.
Output: New state $x^{(t+1)} \in \Omega$.
1. Propose a new state y by sampling from $Q(x^{(t)}, y)$.
2. Compute the acceptance probability:
    $\alpha(x, y) = \min\left(1, \frac{Q(y, x)}{Q(x, y)} \cdot \frac{\pi(y)}{\pi(x)}\right)$
3. With the probability $\alpha(x, y)$, we have $x^{(t+1)} = y$, otherwise $x^{(t+1)} = x^{(t)}$
```

4 The Metropolis-Hastings Algorithm

As the above notations, for $x, y \in \Omega$, we have

$$K(x, y) = \begin{cases} Q(x, y)\alpha(x, y) = Q(x, y) \min\left(1, \frac{Q(y,x)}{Q(x,y)} \cdot \frac{\pi(y)}{\pi(x)}\right) & \textit{if}\, y \neq x \\ 1 - \sum_{y \neq x} Q(x, y)\alpha(x, y) & \textit{if}\, y = x \end{cases}$$

Next we want to prove this definition of transition matrix K satisfies the detailed balance.

$$\pi(x)K(x, y) = \pi(x)Q(x, y) \min\left(1, \frac{Q(y, x)}{Q(x, y)} \cdot \frac{\pi(y)}{\pi(x)}\right) = \min\left(\pi(x)Q(x, y), \pi(y)Q(y, x)\right)$$

$$\pi(y)K(y, x) = \pi(y)Q(y, x) \min\left(1, \frac{Q(x, y)}{Q(y, x)} \cdot \frac{\pi(x)}{\pi(y)}\right) = \min\left(\pi(y)Q(y, x), \pi(x)Q(x, y)\right)$$

In many cases, the target distribution is written as a Gibbs distribution or Boltzmann distribution,

$$\pi(x) = \frac{1}{Z}e^{-E(x)}, \text{ or } \pi(x) = \frac{1}{Z}e^{-E(x)/T}$$

For simplicity let us used Gibbs for an example. While the normalizing constant is hard to compute, suppose the proposal probability is symmetric, i.e. $Q(x, y) = Q(y, x)$, then the acceptance probability becomes

$$\alpha(x, y) = \min(1, \frac{\pi(x)}{\pi(y)}) = \min(1, e^{-(E(x)-E(y))}) = \min(1, e^{-\triangle E})$$

In this way, if $\triangle E < 0$, i.e. state y has lower energy, $\alpha(x, y) = 1$, and $x^{(t+1)} = y$; if $\triangle E > 0$, i.e. state x has lower energy, $\alpha(x, y) = e^{-\triangle E}$. Note that $\triangle E$ is often computed locally as the two states x and y share most of their elements.

There exist other designs for the acceptance rate that guarantee the detailed balance equation, such as

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

Or more generally,

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

where $s(x, y)$ is a symmetric function.

With this method, we can maximize a function for optimization by slowly changing the stationary distribution $\pi(x)$ with additional temperature T. The temperature starts from a high T_0 and decreased to 0 as $n \to \infty$.

$$\pi(x, T_n) = \frac{1}{Z(T_n)}e^{-E(x)/T_n}$$

which is known as simulated annealing method.

4.1 Application of M-H method in Ising model

The first implementation of M-H algorithm is carried out in

[**?**], whose underlying theory was given much later in [**?**].

See example C++ code in https://github.com/fengshi96/MCMC.

5 Gibbs Sampler

Gibbs sampler was created for obtaining samples from distributions that are difficult to sample. Here we use the vector form, and $E(\mathbf{x})$ denotes the energy function,

$$\pi(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}, \mathbf{x} = (x_1, x_2, \cdots, x_d) \in \Omega$$

The Gibbs sampler was introduced as a stochastic version of the relaxation algorithm. In this way, we first introduce the relaxation algorithm, which has no no guarantees for finding the global optimum and in fact, it often gets stuck in local optima.

```
Algorithm 2 Relaxation Algorithm
Input:Energy function $E[\mathbf{x}]$, current state $\mathbf{x}^{(t)} = \{x_1, x_2, \cdots, x_d\} \in \Omega$ and each $x_i$ can have $L$ possible values as $\{y_1, y_2, \cdots, y_L\}$.
Output: New state $\mathbf{x}^{(t+1)} \in \Omega$.
1. Select an index variable $i \in \{1, 2, \cdots, d\}$ at random.
2. Compute
        $u = \arg\min_{y_l} \left( E(x_i = y_l | \mathbf{x}_{-i}) \right) \textit{for } l = 1, 2, \cdots, L$
3. Let
        $\mathbf{x}_{-i}^{(t+1)} = \mathbf{x}_{-i}^{(t)}$ and $x_i^{(t+1)} = u$
```

In formal, the goal of Gibbs sampler is to sample a joint probability,

$$\mathbf{x} = (x_1, x_2, \cdots, x_d) \sim \pi(x_1, x_2, \cdots, x_d)$$

by sampling in each dimension according to the conditional probability,

$$x_i \sim \pi(x_i | \mathbf{x}_{-i}) = \frac{1}{Z} \exp(-E(x_i | \mathbf{x}_{-i}))$$

```
Algorithm 3 Gibbs Sampler
Input:Target Probability function $\pi[\mathbf{x}]$, current state $\mathbf{x}^{(t)} = \{x_1, x_2, \cdots, x_d\} \in \Omega$ and each $x_i$ can have $L$ possible values as $\{y_1, y_2, \cdots, y_L\}$.
Output: New state $\mathbf{x}^{(t+1)} \in \Omega$.
1. Select an index variable $i \in \{1, 2, \cdots, d\}$ at random.
2. Compute conditional probability vector $\mathbf{u} = (u_1, u_2, \cdots, u_L)$with
        $u_l = \pi(x_i = y_l | \mathbf{x}_{-i})$
3. Sample $j \sim \mathbf{u}$ and set $\mathbf{x}_{-i}^{(t+1)} = \mathbf{x}_{-i}^{(t)}$ and $x_i^{(t+1)} = y_j$.
```

Here I want to point out the difference explicitly. When updating one element of the current sample, relaxation algorithm finds the current local minimum and Gibbs sampler samples from its distribution. We should know that only find the local minimum will limit the stochasticity.

A *sweep* of the Gibbs sampler is a sequential visit to all of the sites (variables) once. Although the transition matrix $K_i i$ for one Gibbs step may not be irreducible and aperiodic, it is easy to show that the total transition matrix $K = K_1 K_2 \cdots K_d$ does have these features after one sweep.

If we have $\mathbf{x}^{(t)} \sim \pi(\mathbf{x})$, by above procedure,

$$\begin{aligned}
\mathbf{x}^{(t)} &= (x_1, \cdots, x_i, x_{i+1}, \cdots, x_d) \sim \pi(\mathbf{x}) \\
\mathbf{x}^{(t+1)} &\sim \pi(y_j|\mathbf{x}_{-i})\pi(\mathbf{x}_{-i}) \\
&\sim \pi(y_j|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d)\pi(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d) \\
\mathbf{x}^{(t+1)} &= \pi(x_1, \cdots, x_{i-1}, y_j, x_{i+1}, \cdots, x_d) \sim \pi(\mathbf{x})
\end{aligned}$$

6 Cluster Sampling

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an adjacency graph. Each vertex $v_i \in \mathcal{V}$ has a state variable x_i with a finite number of labels, i.e. $x_i \in \{1, 2, \cdots, L\}$. If $\mathbf{X} = (x_1, x_2, \cdots, x_{|\mathcal{V}|})$ denotes the labeling of the graph, the Ising ($L = 2$) or Potts ($L \geq 3$) model is a Markov Random Field,

$$\pi(\mathbf{x}) = \frac{1}{Z} \exp\left(- \sum_{\langle s,t \rangle \in \mathcal{E}} \beta_{st} \mathbb{1}[x_s \neq x_t] \right)$$

The Swendsen-Wang (SW) algorithm introduces a set of random variables on the edges indicating if they are connected or not,

$$\mathbf{U} = \left\{ \mu_e : \mu_e \in \{0, 1\}, \forall e \in \mathcal{E} \right\}$$

The edge is connected is $\mu_e = 1$. The binary variable μ_e follows a Bernoulli distribution conditional on the labels of the vertices e connects, x_s , x_t,

$$\mu_e|(x_s, x_t) \sim Bernoulli\left((1 - e^{-\beta_{st}}) \mathbb{1}[x_s = x_t] \right), \forall e \in \mathcal{E}$$

$\mu_e = 1$ with probability $1 - e^{-\beta_{st}}$ if $x_s = x_t$ and $\mu_e = 0$ with probability 1 if $x_s \neq x_t$.

The SW algorithm iterates the clustering and flipping step. In clustering step, given the current labeling, we calculate the adjacency of the graph and form a set of connected components with the same label. In flipping step, we randomly select a connected component and assign an arbitrary color to all the lattice inside the connected component. In this step, one may choose to perform the random color flipping for some or all of the connected components in $\mathbf{CP}$ independently, as they are decoupled. By doing so, all possible labelings of the graph are connected in one step, just like one sweep of the Gibbs sampler.

Algorithm 4 Swendsen-Wang Algorithm

Initialize: The adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a set of random variables for each edge as $\mathbf{U} = \{\mu_e : \mu_e \in \{0, 1\}, \forall e \in \mathcal{E}\}$ denoting their connectivity and a set of random variables denoting the label of the lattice as $\mathbf{X} = \{x_i : x_i \in \{1, 2, \cdots, L\}, \forall i \in \mathcal{V}\}$.
Input: Current connectivity of edges $\mathbf{U}^{(t)}$ and labeling of lattice $\mathbf{X}^{(t)}$ at time t
Output: The connectivity of edges $\mathbf{U}^{(t+1)}$ and labeling of lattice $\mathbf{X}^{(t+1)}$ at time $t + 1$
1. the clustering step: sample the edges according to

$$\mu_e^{(t+1)}|(x_s^{(t)}, x_t^{(t)}) \sim Bernoulli((1 - e^{-\beta_{st}}) \mathbb{1}[x_s^{(t)} = x_t^{(t)}]), \forall e \in \mathcal{E}$$

In practice, we first let $\mu_e^{(t+1)} = 0$ if $x_s^{(t)} \neq x_t^{(t)}$ for each $e = \langle s, t \rangle$ and then the remaining $\mu_e^{(t+1)} = 0$ with the probability $e^{-\beta_{st}}$. Hence, the left edges form K connected components as $\mathbf{CP}(\mathbf{U}, \mathbf{X}) = \{cp_i : i = 1, 2, \cdots, K,$ with $\cup_{i=1}^{K} cp_i = \mathcal{V}\}$. Each connected component is a set of lattice with the same label.
2. the flipping step: randomly assign each connected component with a new label.
Select one connected component $V_o \in \mathbf{CP}$ at random and assign a common label l to all lattice in V_o. The new label l follows a discrete uniform distribution,

$$x_s^{(t+1)} = l, \forall s \in V_o, l \sim Uniform\{1, 2, \cdots, L\}$$

Next, we want to show that the SW algorithm can be interpreted as a Metroplis-Hastings step with acceptance rate 1.

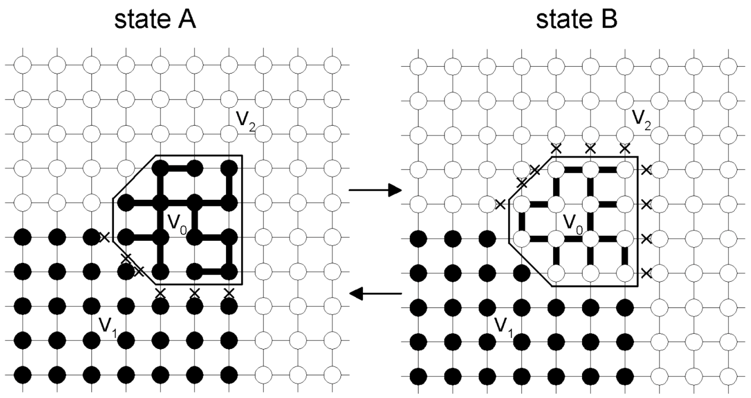

Figure 1: At each step, the SW algorithm flips a patch of vertices.

As shown in Figure **??**, suppose the current state is A where V_0 is connected to V_1. The edges between V_0 and V_1 are turned off in the clustering step with the probability $e^{-\beta}$, from which we form a cut C_{01} between V_0 and V_1 as $C_{01} = \{e = \langle s, t\rangle, s \in V_0, t \in V_1\}$(as crosses in figure). Similarly, if the Markov chain is currently at state B, in order to achieve A, we also form a cut $C_{02} = \{e = \langle s, t\rangle, s \in V_0, t \in V_2\}$. From the setting of Metroplis-Hastings, we need to compute the proposal probability $Q(A \to B)$ and $Q(B \to A)$, which is difficult but their ratio can be shown as

$$\frac{Q(A \to B)}{Q(B \to A)} = \frac{e^{-\beta|C_{01}|}}{e^{-\beta|C_{02}|}} = e^{-\beta(|C_{01}|-|C_{02}|)} \tag{6.1}$$

where $|\cdot|$ denotes the cardinality of a set. Remarkably, the ratio of the probability distribution is also decided by the size of the cuts because the probability distribution counts the number of connected edges.

$$\frac{\pi(A)}{\pi(B)} = \frac{e^{-\beta|C_{01}|}}{e^{-\beta|C_{02}|}} = e^{-\beta(|C_{01}|-|C_{02}|)} \tag{6.2}$$

Hence, the acceptance rate is given by

$$\alpha(A \to B) = \min(1, \frac{Q(A \to B)\pi(B)}{Q(B \to A)\pi(A)}) = 1 \tag{6.3}$$

At low temperature, $\beta \propto 1/T$ and thus the SW flips large patches with acceptance rate 1. Therefore, it can mix quickly even at critical temperatures.

Proof. (*of Equation* **??**) Let $\mathbf{U}_A|(\mathbf{X} = A)$ and $\mathbf{U}_B|(\mathbf{X} = B)$ be realizations of $\mathbf{U}$ at state A and state B. In the clustering step, we form two sets of connected components as $\mathbf{CP}(\mathbf{U}_A|\mathbf{X} = A)$ and $\mathbf{CP}(\mathbf{U}_B|\mathbf{X} = B)$.

For $\mathbf{U}_A|(\mathbf{X} = A)$, following the Bernoulli probabilities, we divide the $\mathbf{U}_A$ into on and off edges as $\mathbf{U}_A = \mathbf{U}_{A,\text{on}} \cup \mathbf{U}_{A,\text{off}}$, where $\mathbf{U}_{A,\text{on}} = \{\mu_e \in \mathbf{U}_A : \mu_e = 1\}$ and $\mathbf{U}_{A,\text{off}} = \{\mu_e \in \mathbf{U}_A : \mu_e = 0\}$.

However, we are only interested in those $\mathbf{U}_A$ which are able to yield V_0. We collect all such $\mathbf{U}_A$ including V_0 given A is a set, $\Omega(V_0|A) = \{\mathbf{U}_A : V_0 \in \mathbf{CP}(\mathbf{U}_A|\mathbf{X} = A)\}$. To be concrete, in order to get V_0, all edges between V_0 and V_1 must be cut off. We denote the remaining off edges as $^-\mathbf{U}_{A,\text{off}}$ in a sense that $^-\mathbf{U}_{A,\text{off}} \cup C_{01} = \mathbf{U}_{A,\text{off}}$ for all $\mathbf{U}_A \in \Omega(V_0|A)$.

Similarly, we have $^-\mathbf{U}_{B,\text{off}}$ as $^-\mathbf{U}_{B,\text{off}} \cup C_{02} = \mathbf{U}_{B,\text{off}}$ for all $\mathbf{U}_A \in \Omega(V_0|B)$.

A key observation in this formulation is that there is a one-to-one mapping between $\Omega(V_0|A)$ and $\Omega(V_0|B)$ because we have a one-to-one mapping between $\mathbf{U}_A$ and $\mathbf{U}_B$ by setting $\mathbf{U}_{B,\text{on}} = \mathbf{U}_{A,\text{on}}$ and $\mathbf{U}_{B,\text{off}} =^- \mathbf{U}_{A,\text{off}} \cup C_{0,2}$.

That is, $\mathbf{U}_A$ and $\mathbf{U}_B$ only differ in the cuts and all these random variables inside the cuts are set as off. In other words, their connected components are the same $\mathbf{CP}(\mathbf{U}_A|\mathbf{X} = A) = \mathbf{CP}(\mathbf{U}_B|\mathbf{X} = B)$. Similarly, any $\mathbf{U}_B \in \Omega(V_0|B)$ has a one-to-one mapping to $\mathbf{U}_A \in \Omega(V_0|A)$.

Now suppose we choose $V_0 \in \mathbf{CP}(\mathbf{U}_A|\mathbf{X} = A)$ randomly, its probability is

$$\mathbf{P}(V_0|A) = \sum_{\mathbf{U}_A \in \Omega(V_0|A)} \frac{1}{|\mathbf{CP}(\mathbf{U}_A|\mathbf{X} = A)|} \prod_{e \in \mathbf{U}_{A,\text{on}}} (1 - e^{-\beta_e}) \prod_{e \in ^-\mathbf{U}_{A,\text{off}}} e^{-\beta_e} \prod_{e \in C_{01}} e^{-\beta_e}$$

Similarly, the probability of choose V_0 in state B is

$$\mathbf{P}(V_0|B) = \sum_{\mathbf{U}_B \in \Omega(V_0|B)} \frac{1}{|\mathbf{CP}(\mathbf{U}_B|\mathbf{X} = B)|} \prod_{e \in \mathbf{U}_{B,\text{on}}} (1 - e^{-\beta_e}) \prod_{e \in -\mathbf{U}_{B,\text{off}}} e^{-\beta_e} \prod_{e \in C_{02}} e^{-\beta_e}$$

In this way, we have

$$\frac{Q(A \to B)}{Q(B \to A)} = \frac{\mathbf{P}(V_0|A)}{\mathbf{P}(V_0|B)} = \frac{e^{-\beta|C_{01}|}}{e^{-\beta|C_{02}|}} = e^{-\beta(|C_{01}|-|C_{02}|)}$$

□

7 Hamilton Monte Carlo

Kong: Just copy the original book for further understanding.

7.1 Hamilton Mechanics

Hamiltonian Monte Carlo (HMC) is a powerful framework for sampling from high-dimensional continuous distributions. Langevin Monte Carlo (LMC) is a special case of HMC that is widely used in Deep Learning applications. Given an n-dimensional continuous density $P(X)$, the only requirement for implementing HMC is the differentiability of the energy $U(X) = -\log P(X)$. Like other MCMC methods (e.g. slice sampling, Swendsen-Wang cuts), HMC introduces auxiliary variables to facilitate movement in the original space. In HMC, the original variables represent *position*, and the auxiliary variables represent *momentum*. Each position dimension has a single corresponding momentum variable, so the joint space of the original and auxiliary variables has dimension $2n$, twice the size of the original space. Once the momentum variables are introduced, Hamilton's Equations are used to simulate the time evolution of a physical system with potential energy U . The properties of Hamilton's Equations ensure that movement in the joint space preserves the distribution of P in the original space.

Hamiltonian Mechanics was originally developed as an alternative but equivalent formulation of Lagrangian Mechanics, and both are equivalent to Newtonian Mechanics. In Hamiltonian Mechanics, the states of a physical system are represented by a pair of n- dimensional variables q and p. The variable q represents *position* in the system, and p represents *momentum*. A joint state (q, p) provides a complete description of the physical system at a single instant in time.

The evolution of a state (q, p) over time is governed by a scalar-valued function $H(q, p)$ representing the energy of the system, and a pair of partial differential equations known as Hamilton's Equations:

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}$$
$$\frac{dp}{dt} = -\frac{\partial H}{\partial q}$$

$H(q, p)$ is often referred to as the Hamiltonian of the system, and it remains constant as (q, p) evolves over time.

References

[Metropolis et al.(1953)Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The Journal of Chemical Physics **21**, 1087 (1953), `https://doi.org/10.1063/1.1699114`, URL `https://doi.org/10.1063/1.1699114`.

[Hastings(1970)] W. K. Hastings, Biometrika **57**, 97 (1970), ISSN 00063444, URL `http://www.jstor.org/stable/2334940`.

[Chung(1960)] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, Die Grundlehren der Mathematischen Wissenschaften 104 (Springer-Verlag Berlin Heidelberg, 1960), 1st ed., ISBN 978-3-642-49408-6,978-3-642-49686-8.

Supplemental

Formal Definition

More details about this section can be found in [**?**].

The basic building blocks of a probability theory can be formalized in the following:

1. An abstract set Ω, termed *probability space* or *sample space*, whose elements ω are called the *elementary event* or *sample point*.
2. A Borel field $\mathfrak{F}$ of subsets of Ω, termed *measurable sets* or *events*, in which Ω is also a member.
3. An additive probability measure P defined on $\mathfrak{F}$

These together make the *probability triple* $(\Omega, \mathfrak{F}, P)$ TODO: what's its physics analogy.

Why Borel field?

Microcanonical ensembles

MCE focus on systems that are mechanically and adiabatically isolated from its environment ($\Delta E = W = Q = 0$). The general coordinates $\vec{x}$ is fixed, so that there is no work done i.e. $W = 0$; Internal energy E is also fixed since $Q = 0 \;\Rightarrow\; \Delta E = Q + W = 0$. This is the *macrostate* given $E, \vec{x}$, denoted by $M \equiv (E, \vec{x})$. *The corresponding set of mixed microstates form the microcanonical ensemble*.

A microstate in the phase space is labeled by μ i.e. phase space coordinate $\mu \equiv (x_1, p_1, \ldots, x_N, p_N)$, whose time evolution is governed by $\mathcal{H}(\mu)$. In MCE, the Hamiltonian conserves the total energy of a given system, so all valid microstates are confined to the surface $\mathcal{H}(\mu) = E$. The central postulate of Statmech states that the equilibrium probability distribution is given by:

$$P_{(E,\vec{x})} = \frac{1}{\Omega(E,\vec{x})} \cdot \begin{cases} 1, & \text{for} \mathcal{H}(\mu) = E \\ 0, & \text{otherwise} \end{cases} \tag{S1}$$

The zeroth law

Consider two microcanonical systems (each with a large dof), their state in phase space are μ_1 and μ_2 respectively. We allow them to exchange energy but not work. Remember these are systems modeled by MCE, so their state μ_i is determined by internal energy E_i and $\vec{x}$ only.

The combined system has energy:

$$E = E_1 + E_2.$$

For this big system, (at any moment), its position in phase space is spanned by $\mu = \mu_1 \otimes \mu_2$. Therefore the Hamiltonian is described by:

$$H(\mu_1 \otimes \mu_2) = H_1(\mu_1) + H_2(\mu_2) \tag{S2}$$

and

$$P_E(\mu_1 \otimes \mu_2) = \frac{1}{\Omega(E)} \cdot \begin{cases} 1 & \text{for } H_1(\mu_1) + H_2(\mu_2) = E \\ 0 & \text{otherwise} \end{cases} \tag{S3}$$

Note: Here the Big system itself is viewed as a MCE! Now we count how many states $\mu = \mu_1 \otimes \mu_2$ are possible. For each pair of $\{E_1 \pm dE_1/2, E_2 \pm dE_2/2\}$, there are $\Omega_1(E_1) \times \Omega_2(E_2)$ states. Therefore the total allowed states for the Big system is:

$$\Omega(E) = \int dE_1 \Omega_1(E_1) \Omega_2(E - E_1) \tag{S4}$$

we can write Ω as $\Omega = \exp\{\log(\Omega)\} = \exp\{S/k_B\}$, so:

$$\Omega(E) = \int dE_1 \exp\left\{\frac{S_1(E_1) + S_2(E - E_1)}{k_B}\right\} \tag{S5}$$

According to (3), all states are equal weighted, therefore the energy that produces largest $\Omega(E)$ is the equilibrium energy we are looking for. Since the integrand is exponentially large, we expect the mean contribution is from the peak defined at E_1^*, so that $S_1 + S_2$ is maximized, thus the total entropy is maximized:

$$S(E) = k_B \log \Omega(E) \simeq S_1(E_1^*) + S_2(E_2^*) \text{ is maximized} \tag{S6}$$

Now we find E_1 that maximize $S_1 + S_2$:

$$\begin{aligned} \frac{\partial}{\partial E_1}\Big(S_1(E_1) + S_2(E - E_1)\Big) &= \frac{\partial S_1}{\partial E_1} + \frac{\partial S_2(E - E_1)}{\partial E_1} \\ &= \frac{\partial S_1}{\partial E_1} - \frac{\partial S_2}{\partial E_2} = 0 \end{aligned} \tag{S7}$$

therefore:

$$\frac{\partial S_1}{\partial E_1} = \frac{\partial S_2}{\partial E_2} \tag{S8}$$

which must be satisfied when the joint system reach the equilibrium! It is consistent with zeroth law, that systems in equilibrium has equal temperautre:

$$\frac{\partial S}{\partial E}\Big|_{\mathbf{x}} = \frac{1}{T} \tag{S9}$$

note they are evaluated at their own fixed **x**.

Canonical ensembles

The story of sampling

This part, I will recap the history of sampling from the posterior distribution, which is a unique chance to grasp the main idea of sampling in statistics. Thanks S. Feng.

In history, with the power of Central Limit Theorem, we can use a single point estimate for a parameter and its standard error. (Kong: CLT talks about asymptotic normality of a distribution, but why we call it a single point estimation and does it has anything to do with estimate of sample mean?) However, in the view of Bayesian analysis, we seek to summarize the entire posterior distribution. The key difference lies in that here Bayesian tends to use entire posterior distribution rather than the mode of likelihood function and standard errors. In the same way, if we are able to summarize the entire posterior distribution for a parameter, there is no need to rely on asymptotic arguments about the normality of the distribution: It can be directly assessed.