Markov Chain, Monte Carlo, and Markov Chain Monte Carlo

Danny Chen[1,*]

[1] *Department of Mathematics, Applied Mathematics, and Statistics, Case Western Reserve University, Cleveland, Ohio 44106, USA*

I. A MODEL THAT IS NEVER TRUE, BUT SOMETIMES GREAT

In this section, we will introduce the concept of Markov processes — dynamical systems that has only the memory of its most recent past. More specifically, we focus on the case where we have a dynamical system in discrete time taking up discrete (and finite) state spaces.

A. Probability Theory

Though the concept of probability and chance dates back a long time, it is generally considered that modern probability is developed by Pierre de Fermat – yes, the Fermat who stated the well-known theorem in the margin without a proof – and Blaise Pascal in 1654. It was used to study games with randomness in it, or to put it plainly, they want to win money [1]. However, this field was never formally developed and there are a lot of paradoxes. Then, Russian mathematician Andrey Kolmogorov came along and formalizes the notion of probability in his paper *General Theory of Measure and Probability Theory* in 1928[2] [3]. For the purpose of this project, we will not necessarily dive into the measure-theoretic background and stay within the realm where measure theory is not necessary — namely, we will work with discrete (perhaps large, but discrete nonetheless) state-spaces.[4] Let's begin the definition of a probability space.

Definition 1. *A probability space is a triple* $(\Omega, \mathcal{H}, \Pr)$ *where*

1. Ω *is a set of possible outcomes*
2. $\mathcal{H}$ *is a* σ*-algebra of all possible events*
3. $\Pr : \mathcal{H} \to \mathbb{R}^+$ *is a probability measure, that is,*
   - $\Pr(\emptyset) = 0$ *and* $\Pr(\Omega) = 1$
   - $\Pr(\cup_n E_n) = \sum_n \Pr(E_n)$ *for disjoint set* (E_n)*'s.*

For our purposes, a random variable X will be a variable that takes values in Ω at random with probability specified by the measure Pr. Then, a stochastic process $\{X_n : n \in \mathbb{N}\}$ is simply an indexed collection of random variables.

From one probability space, we can create another with the notion of *conditional probability.*

Definition 2. *Let* A *be an event in* $\mathcal{H}$ *with a non-zero probability, then we can define the probability space* $(\Omega, \mathcal{H}, \Pr(\cdot|A))$ *where the new measure satisfies*

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{1}$$

for $B \in \mathcal{H}$*.*

Intuitively, we fix the event B of occurring and look at the probability of A occurring given that B has happened. For example, if we're rolling a fair die, the probability of getting a "4" is 1/6. However, if we condition on getting an even-valued outcome, the conditional probability of getting a "4" becomes 1/3.

B. Markov Chain

Definition 3. *A stochastic process* $\{X_n : n \in \mathbb{N}\}$ *is a Markov Chain with state-space* Ω *and transition matrix* $P \in \mathbb{R}^{|\Omega|\times|\Omega|}$ *if for* $x, y, z_{t-1}, \ldots, z_1 \in \Omega$*,*

$$\begin{aligned}\Pr(X_{t+1} = y | X_t = x, X_{t-1} = z_{t-1}, \ldots, X_1 = z_1) \\ = \Pr(X_{t+1} = y | X_t = x) = P(x, y)\end{aligned} \tag{2}$$

In plain words, this means that the future of the process is dependent only of the most recent past and not anything before that.

a. Example: Samantha's acorn hunt After a long nap in the winter, Samantha the squirrel found three distinct locations — labeled 1,2, and 3 — each with plentiful acorns. Everyday, Samantha will go to a field, stay for the day, and either stay for another day or move to another field for the next day. Samantha has a few perks when picking which location to go to next: 1) if she were to move, she always goes in the order $1 \to 2 \to 3 \to 1 \to \ldots$, 2) she likes some fields more than the other. If we let $P \in \mathbb{R}^{3\times 3}$ where the (x, y)-th entry denotes the probability of

* txc461@case.edu

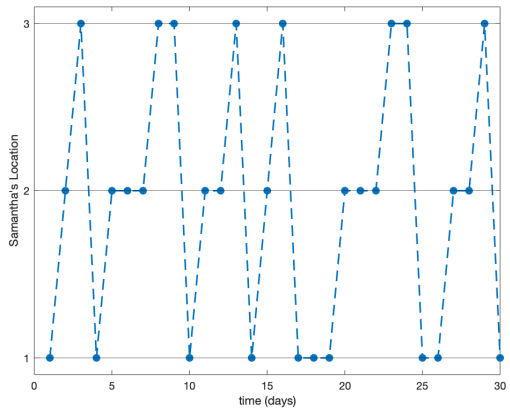

FIG. 1: One possible trajectory of Samantha the squirrel.

transitioning from field x to field y, then P will look like the following.

$$P = \begin{pmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22} & p_{23} \\ p_{31} & 0 & p_{33} \end{pmatrix} \tag{3}$$

and since which field to go next must be probability distribution, the sum of the rows must be one. For example, we can choose the stochastic matrix to be the following:

$$P = \begin{pmatrix} 0.3 & 0.7 & 0 \\ 0 & 0.5 & 0.5 \\ 0.2 & 0 & 0.8 \end{pmatrix} \tag{4}$$

From this stochastic matrix, we can tell that Samantha likes field 3 the most and less so for field 1. We can simulate (which we will talk about in the next section) Samantha's movement as shown in Figure 1.

We can also do this analytically. Let's suppose Samantha is at field 1 the first day, that is, $\Pr(X_1 = 1) = 1$. Then, in the next day, Samantha will be at field 1 with probability p_11 and field 2 with probability p_22 (and field 3 with probability 0). On can also think of this as $\vec{e_1}P$ where $\vec{e_k}$ is a row vector of zeros with 1 on the k-th entry. What about the day after?

$$\Pr(X_3 = i|X_1 = 1) \tag{5}$$

$$= \sum_{j\in\omega} \Pr(X_3 = i, X_2 = j|X_1 = 1) \tag{6}$$

$$= \sum_{j} \Pr(X_3 = i|X_2 = j, X_1 = 1)\Pr(X_2 = j|X_1 = 1) \tag{7}$$

$$= \sum_{j} \Pr(X_3 = i|X_2 = j)\Pr(X_2 = j|X_1 = 1) \tag{8}$$

where the last equality is by the *Markov property*. Those who are a bit sharper will spot that this is the $(1, i)$-th component of $\vec{e_1}P^2$ where $\vec{e_k}$ is a row vector of zeros with 1 on the k-th entry. Straightforwardly by induction, we can see that

$$\Pr(X_t = j|X_s = i) = \left(P^{t-s}\right)(i, j) \tag{9}$$

for any $t > s$. If we start with an initial distribution $\mu_0 = \sum_k q_k\vec{e_k}$, the distribution after t steps satisfies

$$\mu_t = \mu_{t-1}P = \mu_0P^t \tag{10}$$

by linearity of matrix operations. More generally, there is the *Chapman-Kolmogorov equation* that states the following:

$$\Pr(X_t = j|X_1 = i) = \sum_k \Pr(X_t = j|X_s = k)\Pr(X_s = k|X_1 = i) \tag{11}$$

Suppose you haven't been watching Samantha closely for a while and you want to find her now (for whatever reason, perhaps you have an acorn to share). Knowing her preference and how she might go from one field to another, how should you proceed your search? Suppose Samantha can be in any one of the fields to start according to some distribution μ_0, then t days later, we know she might be located with respect to the distribution $\mu_t = \mu_0P^t$. If we take $t \to \infty$, so

$$\pi = \lim_{t\to\infty} \mu_t = \lim_{t\to\infty} \mu_0P^t \tag{12}$$

then, what is π? Well, suppose that the limit exists, we know that

$$\pi = \pi P \tag{13}$$

So, it seems like this π, if it exists in the first place, is 1) independent of μ_0 and 2) an eigenvector of P corresponding to eigenvalue 1. We will formalize this later. But, in terms of Samantha the squirrel, we can numerically probe at what this limit is. Suppose Samantha starts off with the distribution

$$\mu_0 = \begin{pmatrix} 0.4 & 0.5 & 0.1 \end{pmatrix} \tag{14}$$

which was arbitrarily chosen. Then, the evolution of the probabilities is shown in Figure 2.

To simplify notation from here on out, we will introduce the following shorthand notations:

$$\Pr_x(\cdot) \equiv \Pr(\cdot|X_0 = x), \quad \mathbb{E}_x(\cdot) \equiv \mathbb{E}(\cdot|X_0 = x) \tag{15}$$

b. Irreducibility and Aperiodicity There are two important properties that a Markov chain can have.

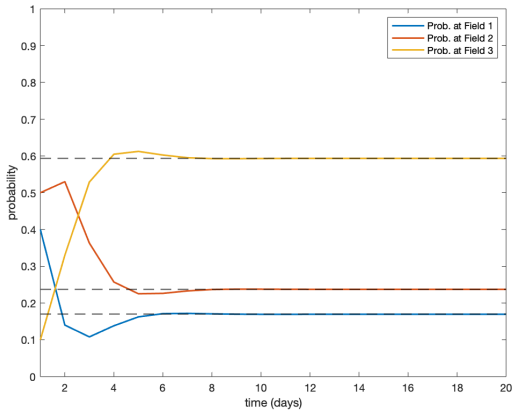

FIG. 2: Samantha was initialized arbitrary distribution and the probability of being at each field is calculated. It seems like the probability converges to some π (in black), that is the left eigenvector of P.

Definition 4. *A chain P is irreducible if for any two states x and y, there exists a $t > 0$ such that $P^t(x, y) > 0$.*

We can think of a Markov Chain as random walk on a graph. Each state is a vertex and an edge connect two vertices if the transition probability is nonzero. So, an irreducible Markov chain will mean that any state is reachable from any other state, which implies strong connectivity.

Definition 5. *For a given chain P, define $\mathcal{T}(x) = \{t > 0 : P^t(x, x) > 0\}$. The period of state x is defined to be the greatest common divisor of the elements in $\mathcal{T}(x)$. The chain is aperiodic if all states of the chain has period 1.*

Intuitively, if a chain is periodic, then for some multiple of times t, $2t$, $3t$, and so on, there is a subset of states the agent can be in (if we take the random walk interpretation). As one can imagine, a periodic Markov chain cannot converge to a single distribution due to its periodicity.

The lemma below will be a good motivation for talking about convergence.

Lemma 1. *If P is aperiodic and irreducible, then there is an integer r such that $P^r(x, y) > 0$ for all $x, y \in \Omega$.*

Proof. Use the following fact from number theory: any set of non-negative integers that is closed under addition and has a greatest common divisor of 1 contains all non-negative integers except for finitely many numbers. Show that for any state x in an aperiodic chain, $\mathcal{T}(x)$ is closed under addition. This means that there exists some time τ for state x where all times after it $P^\tau(x, x) > 0$. Furthermore, by irreducibility, there is some time r for the agent to potentially travel from one state to another. Taking the maximum of all pairs (which exists because we're dealing with finite state-spaces) of $\tau + r$ gives the time until all states have a positive probability. □

Recall that our goal to explain some convergence behavior to the left eigenvector corresponding to eigenvalue 1. So, before anything, we must prove that the eigenvector exists, which is stated as the lemma below.

Lemma 2. *If a chain P is irreducible, then there is a unique distribution π on Ω such that $\pi P = \pi$.*

Proof. First, let $\tau_z = \min\{t > 0 : X_t = z\}$. Use the definition of irreducibility to show that for sufficiently large r, there is an ϵ such that for any $y \in \Omega$,

$$\Pr_z(\tau_y > kr) \leq (1 - \epsilon)\Pr_z(\tau_y > (k - 1)r) \tag{16}$$

Use this fact and the union bound to show that $\mathbb{E}_z(\tau_y) < \infty$.

Then, fix a particular $z \in \Omega$. Define $\tilde{\pi}(y)$ to be the expected number of visits to y before returning to z:

$$\tilde{\pi}(y) = \sum_{t=0}^{\infty} \Pr_z(X_t = y,\ \tau_z > t) \tag{17}$$

And show, by expanding the summation, that $\tilde{pi}(y)$ is stationary.

$$\tilde{\pi} = \tilde{\pi}P \tag{18}$$

Lastly, since we want a probability measure, we get π by normalizing $\tilde{\pi}$.

$$\pi(x) = \frac{\tilde{\pi}(x)}{\sum_{y\in\Omega}\tilde{\pi}(y)} = \frac{\tilde{\pi}(x)}{\mathbb{E}_z(\tau_z)} = \frac{1}{\mathbb{E}_x(\tau_x)} \tag{19}$$

This shows the existence of π. To see the uniqueness, let h be some function. We know if $\pi = \pi P$ exist, there must be an h satisfying $h = Ph$. Show that h must be constant by contradiction. This implies that the eigenspace corresponding to eigenvalue 1 has dimension 1, and the left eigenvector π is unique. □

c. History of Markov Chains Markov chains are invented and developed by Andrey Markov. The study of Markov Chains was motivated by the abundance of assumptions that the random variables are independent. in 1738, Jacob Bernoulli proved the *weak law of large numbers* for independent binary variables. Simeon Poisson, a century later, generalized this to binary variables that are independent,

but not necessarily identically distributed. Pafnuty Chebyshev, Markov's teacher, generalized the law of large numbers of independent random variables with bounded moments in 1867. P.A. Nekrasoc, another mathematician working on the weak law of large numbers at the time, stated in his 1902 paper: "independence is a necessary condition for the law of large numbers." Such a claim motivated Markov's work on dependent random variables, which led to the theory of Markov chains [5].

C. Chain Mixing

Now, we know that an irreducible, aperiodic Markov chain has a stationary distribution. However, we have not characterized how to achieve the stationary distribution. Here, we will discuss the phenomenon that: if we let the chain run for a sufficiently long time, the chain converges towards the stationary distribution. Before that, we must define what "moving towards" mean by defining a metric on the space of distributions.

Definition 6. *The total variation distance between two distribution μ and ν on Ω is defined by*

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| \tag{20}$$

Now, we will show that the total variation distance between the distribution at any time t and the stationary distribution decreases in the limit as $t \to \infty$.

Theroem 1. *Let P be an irreducible, aperiodic Markov chain with stationary distribution π. Then, there are constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t \tag{21}$$

Proof. First, by Lemma 1, we know that there is an r such that $P^r(x, y) > 0$ for all $x, y \in \Omega$. So, for some sufficiently small $\delta > 0$,

$$P^r(x, y) \geq \delta\pi(y) \tag{22}$$

Let $\Pi \in \mathbb{R}^{|\Omega| \times |\Omega|}$ be a matrix with rows that are π, and define $\theta = 1 - \delta$. Then, let Q be a stochastic matrix that satisfies

$$P^r = (1 - \theta)\Pi + \theta Q \tag{23}$$

Then, by induction, show that

$$P^{rk} = (1 - \theta^k)\Pi + \theta^k Q^k \tag{24}$$

Multiply by P^j and rearranging to get the following form:

$$P^{rk+j} - \Pi = \theta^k(Q^k P^j - \Pi) \tag{25}$$

Focus on one particular row x. By set up of Q, we can arrive at the following, which implies the theorem.

$$\|P^{rk+j}(x, \cdot) - \pi\|_{TV} \leq \theta^k \tag{26}$$

□

II. GETTING COMPUTATIONAL

This section will introduce the idea of *Monte Carlo* computations, a class of methods that estimate some desired variable by probabilistically sampling from some distribution. The justification of this is simple: the (strong) law of large numbers!

Theroem 2 ([6])**.** *Let $(X_i)_{i=1}^n$ be a collection of pairwise independent and identically distributed random variables with a finite mean and variance. Then,*

$$\frac{1}{n}\sum_i X_i \xrightarrow{a.s.} \mathbb{E}[X] \tag{27}$$

as $n \to 0$; that is, the average over the samples converges almost surely (with an exception of a set with measure zero) to the expectation.

A. Buffet's needle problem

Georges Louis Leclerc, Comte de Buffon — a French mathematician — first proposed the following problem in 1777: *Suppose there is a board with parallel stripes on in. There is also a needle, whose length is the same as the space between the parallel strips. Drop the needle onto the board with no particular method. What is the probability that the needle will be touching one of the parallel lines?*

a. Analytic Solution We can solve this problem with geometry and basic probability theory. Whether the needle touches the parallel line depends on two things: the distance from the midpoint of the needle to the closest line (X) and the angle formed by (the extended) needle and the line (θ). The needle will intersect the line if the hypotenuse of the triangle — the triangle formed by the (extended) needle, the line, and the altitude from the midpoint to the line — is less than half of the gap between the lines ($L/2$). In mathematical language:

$$X < \frac{L}{2}\cos\theta \tag{28}$$

We will assume the needle is uniformly dropped. So, X is some value between 0 and $L/2$, θ is some value

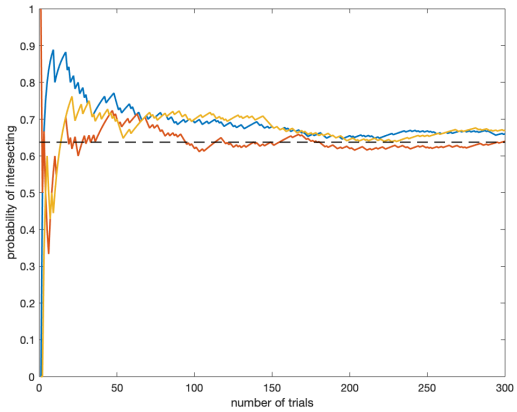

FIG. 3: Simulation of the Buffon's needle problem. The black dashed line is a $\pi/2$, the value we got analytically.

between 0 and $\pi/2$. So, we can carry out the following computation.

$$\Pr\left(X < \frac{L}{2}\cos\theta\right) = \int_0^{\pi/2}\int_0^{L\cos\theta/2} \frac{4}{L\pi} dx d\theta \tag{29}$$

$$= \frac{2}{\pi}\int_0^{\pi/2} \cos\theta d\theta \tag{30}$$

$$= \frac{2}{\pi} \tag{31}$$

b. Sampling Solution The analytic solution is elegant, but what if we're not good at geometry? Intuitively, we can test it out ourselves. Find a needle (there must be one in your house), draw parallel lines that has gaps the same size as the needle, and start dropping it! Keep track of the total number of drops and the number of times you successfully get the needle to intersect the line. Since I'm in college and I don't have a needle handy, I will use a computer to simulate the dropping. The results are shown in Figure 3. We can see that the more times we toss, we converge towards the true value we got from the analytical calculations. In fact, Augustus De Morgan and his student had tried to use this method to estimate the value of π. You can imagine how painful it is without a computer!

B. Monte Carlo methods

Monte Carlo methods was invented by John von Neumann and Stanislaw Ulam during World War II. Stanislaw Ulam initially used it to develop nuclear weapons at Los Almos National Laboratory. At that time, computing using deterministic method is difficult and expensive. That is how Ulam found a way to randomization to perform calculations. He described his thoughts to von Neumann, and they both agreed that it is an incredibly promising approach — it is so good that they shouldn't let other people know! They're colleague Nicholas Metropolis (who we will see later) suggested the name *Monte Carlo*: the casino in Monte Carlo, Monaco where Ulam's uncle would borrow money from his relatives to gamble. Quickly, this method became crucial for the Manhattan project and for many scientific and engineering disciplines after the war [7].

a. Pseudorandomness Before we go dig into Monte Carlo methods, let's step back a bit and think: where do the random numbers come from? We don't know how to generate truly random numbers, so we reside to *pseudorandom number generators*. As the name implies, it is a machine (algorithm) that generates a deterministic sequence of numbers that behaves randomly. This is a topic that lies deeply in the field of *theoretical computer science* and has significant implications in, well, anywhere that requires random computation. Below gives the definition of a pseudorandom number generator.

Definition 7 ([8]). *A pseudorandom number generator is a deterministic function f defined on a subset $U \subseteq \{0,1\}^k$ into $\{0,1\}^k$, where $k < l$, which maps a seed $X \in U$ to a deterministic sequence of l bits:*

$$f(X) = (x_1, x_2, \ldots, x_l) \tag{32}$$

What might this look like? A well-known one is the *Blum Blum Shub generator*. Let n be a product of two primes p and q that are equal to 3 modulo 4 with k-bits. Seeds are ordered $X_0, X_1, \ldots$ and the generation follows the following recurrence:

$$X_{i+1} = X_i^2 \mod n \tag{33}$$

This generates a sequence of 0s and 1s that "looks" random, which we can then use to build other random numbers. However, what "random enough" was, and still is, under debate. Here, we will introduce Andrew Yao's, a Turing award (the Nobel prize equivalent in computer science) winner for his work in pseudorandomness, formulation of what it means to be random.

Theroem 3 ([9, 10]). *Let $\{g_n\}$ be a set of polynomial-time computable family of functions, where $g_n : \{0,1\}^n \to \{0,1\}^m$ and $m = m(n) > n$. Then, a $(\delta(n), s(n))$-pseudorandom generator is a machine that for every probabilistic algorithm A running in time $s(n)$ and for large enough n,*

$$\left|\Pr_{y\in\{0,1\}^m}(A(y) = 1) - \Pr_{x\in\{0,1\}^n}(A(g_n(x)) = 1)\right| \leq \delta(n) \tag{34}$$

Let's break this definition apart term-by-term. The function g is the random number generator. We require it to be computable in polynomial-time: so the generation must be simple enough for anyone's personal computer. We say that the generated number is random enough if, for any algorithm A that needs the random number, the difference in output of A using a truly random string and a pseudorandom string is bounded above by some δ. Basically, if for any practical purposes, we cannot distinguish true random algorithm from pseudorandom ones, then the pseudorandom numbers are random enough.

b. Inverse Sampling Inverse sampling is one of the most simple, but also the most restrictive Monte Carlo algorithms. This will motivate the next section, which deals with a much more powerful paradigm. The idea behind inverse sampling is quite simple. Again, we will detour and talk about transformation of random variables. Let X be some random variable, and define $Y = T(X)$ for some monotone function T. Then, we can express the distribution of Y in terms of the distribution of X.

$$\Pr(Y < y) = \Pr(T(X) < y) = \Pr(X < T^{-1}(y)) \tag{35}$$

where the inverse is well-defined by monotonicity. Now, notice that any random variable X is simply a monotonic function transformation away from the uniform distribution. Namely, consider a uniformly distributed random variable U between $[0, 1]$, which we now know how to generate. And suppose there is a random variable X with a cumulative distribution F_X that we want to sample from. Notice that we can define $X = F_X(U)$:

$$\Pr(X < x) = \Pr(F_X(U) < x) = \Pr(U < F_X^{-1}(x)) \tag{36}$$

and since $F_X : \mathcal{X} \to [0, 1]$ where $\mathcal{X}$ is the support of X. Then, $F_X^{-1} : [0, 1] \to \mathcal{X}$. This gives us an algorithm! To generate samples from X, simply generate a uniform random sample $u \sim U$ and compute $F_X^{-1}(u)$ to get the sample for X.

This method works for any distribution with an easily invertible distribution function, so: any variables defined on discrete state-spaces or a handful of continuous variables. Since we're working with finite state-spaces, let's talk about the first case. If the state-space is relatively small — we're talking perhaps 10^6 states — doing this inversion is not difficult. However, if the state space is large, sometimes even too large to store in our computers (and surprisingly, most discrete problems we care about is like this), and there is no sparsity constraint (so we do not know a priori if there are entries with zero probability), the inverse sampling procedure is simply not feasible.

III. MARKOV CHAIN MONTE CARLO

Markov Chain Monte Carlo, or MCMC for short, is a type of Monte Carlo algorithms that combines the two topics we talked about before. Intuitively, we know that sampling from Markov chains is simple: we can generate new samples each step given the current state. So, we use the property that Markov chains mix into stationary distributions to create a chain that converges towards our desired distribution. Below we give two well-known MCMC algorithms: Metropolis-Hastings and Glauber Dynamics (or sometimes known as Gibbs sampling).

A. Metropolis-Hastings Algorithm

As mentioned before, Metropolis was one of the scientists working in the Manhattan project. A bit later, him, along with Rosenbluth, Teller, and Teller gave the first description of this type of algorithm for a specific distribution in 1953 [11]. Then, Hastings extended his technique to arbitrary distributions, hence the name Metropolis-Hastings [12].

Suppose there is some distribution π that we want to sample from, the goal is to construct a chain P such that π is the stationary distribution of P. We will do this indirectly by constructing another Markov chain Ψ that will help us traverse the state-space. The chain Ψ can be an arbitrary irreducible chain, and we will let P take the following dynamic:

$$P(x, y) = \begin{cases} \Psi(x, y) \left(\frac{\pi(y)\Psi(y,x)}{\pi(x)\Psi(x,y)} \wedge 1 \right) & \text{if } y \neq x \\ 1 - \sum_{z \in \Omega \setminus \{x\}} P(x, z) & \text{if } y = x \end{cases} \tag{37}$$

The chain above describes the *Metropolis chain*. One can think of this as a accept-reject scheme: generate a sample from Ψ, move to the new state with probability defined by the quotient term above, otherwise stay at the current location.

To show that this in fact converges towards the right distribution, we need to show $\pi = \pi P$. We can actually show a stronger condition: that π satisfies *detailed balance*:

$$\pi(x)P(x, y) = \pi(y)P(y, x) \tag{38}$$

for every $x, y \in \Omega$. For the case of $y \neq x$, we have the following:

$$\pi(x)P(x, y) = (\pi(y)\Psi(y, x) \wedge \pi(x)\Psi(x, y)) \tag{39}$$

which quite clearly gives $\pi(y)P(y,x)$. Then, by conservation of probability, the case of $y = x$ must also hold. By satisfying detailed balance, we know that π is stationary. Since Ψ is assumed to be irreducible, we know that π is unique and the chain specified by P will converge towards its stationary distribution, namely, π.

B. Glauber Dynamics

Glauber dynamics, or more commonly called *Gibbs sampling* in the statistics literature, is another common MCMC method. It is usually used for efficient inference over high dimensional spaces, and unlike Metropolis-Hastings algorithm, its formulation can change drastically from one context to another. In honor of its inventor, who are statistical physicists, we will approach the introduction with a more statistical physics setting.

Let V and S be finite sets, and let $\Omega = S^V$. One can visualize this in terms of a graph where V is the set of vertices and S are values each $v \in V$ can take. Let π be some distribution over Ω that we wish to sample over. Furthermore, for $x \in \Omega$ and $v \in V$, define the quantity $\Omega(x, v)$ as the following:

$$\Omega(x, v) = \{y \in \Omega : y(w) = x(w)\ \forall w \neq v\} \tag{40}$$

Then, the *Glauber dynamics* for π obeys the following rule: choose a $v \in V$ uniformly at random, then let the transition from state x to state y occur with probability $\pi(y)/\pi(\Omega(x, v)) \vee 0$. Alternatively, we can write the transition matrix P as below:

$$P(x, y) = \frac{1}{|V|} \sum_{v \in V} \left(\frac{\pi(y)}{\pi(\Omega(x, v))} \vee 0 \right) \tag{41}$$

Again, let's show that it satisfies detailed balance to justify its convergence. First, notice that for all $y \in \Omega(x, v)$, that is, for all y such that $P(x, y)$ is not zero, $\Omega(x, v) = \Omega(y, v)$. So,

$$\pi(x)P(x, y) = \frac{1}{|V|} \sum_{v \in V} \left(\frac{\pi(y)\pi(x)}{\pi(\Omega(x, v))} \vee 0 \right) \tag{42}$$

$$= \frac{1}{|V|} \sum_{v \in V} \left(\frac{\pi(x)\pi(y)}{\pi(\Omega(y, v))} \vee 0 \right) \tag{43}$$

$$= \pi(y)P(y, x) \tag{44}$$

If $\pi(x) > 0$ for all $x \in \Omega$, then the chain is irreducible and we will converge towards the stationary distribution, as guaranteed in Theorem 1.

C. Example: Ising Model

To make things concrete, we will consider the classic *Ising model* from statistical physics. This model was proposed by Wilhelm Lenz in the 1920s for as a simplified version for ferromagnet — the kind of magnetism associated with iron and nickel. Lenz's student Ernst Ising chose to focus on this model for his PhD dissertation and made significant progress. It turns out that 2D Ising models that we will introduce today, though understood now, fueled many other fields in physics, math, and computer science.

The Ising model is a *spin system*. There is a graph $G = (V, E)$, and for each vertex, there is a spin (taking values $\{-1, 1\}$) associated with it. In essence, there is a probability distribution over $\{-1, 1\}^V$. For every state of the system, $\sigma \in \Omega$, there is an energy, or *Hamiltonian*, associated with it defined by the following function:

$$H(\sigma) = -\sum_{(v,w) \in E} \sigma(v)\sigma(w) \tag{45}$$

So, the energy increases if adjacent vertices have opposite spin. Then, the *Gibbs* distribution with respect to Hamiltonian H is defined as the following:

$$\mu(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)} \tag{46}$$

where $Z(\beta)$ is called the *partition function*, which normalizes μ into a valid probability distribution. The parameter β corresponds to the inverse of the temperature of the system. If temperature if infinite ($\beta = 0$), we can see that all terms become 1 and μ is simply the uniform distribution over bit strings. As $\beta \to \infty$, or as temperature tends towards 0 (in the *adiabatic limit*, by approaching 0 infinitely slowly), the distribution approaches the lowest energy configuration. In fact, this technique is called *simulated annealing* and is used to solve many (combinatorial) optimization problems that are difficult to compute.

However, the interesting phenomenon that arises in the Ising model is *phase transition*, which is similar to (perhaps, is exactly) bifurcation in dynamical systems theory. It is when some quantity associated with the system go through an abrupt change, *e.g.* a discontinuity in the function or in its 1st or 2nd derivative. Ising initially studied the 1-dimensional system where each vertex is adjacent to two other vertices in a way that forms a ring, and found no phase transition. However, in 1944, Lars Onsager found a phase transition in the 2-dimensional model. Since then, people have tried to study phase transition in higher dimensional models [13].

Figure 4 shows both simulation technique on the Ising model on a 50×50 square lattice with closed boundary conditions (torus). We can see the transition from something structured to something completely disordered. This type of phase transition. There are a total of 2500 vertices here, meaning that

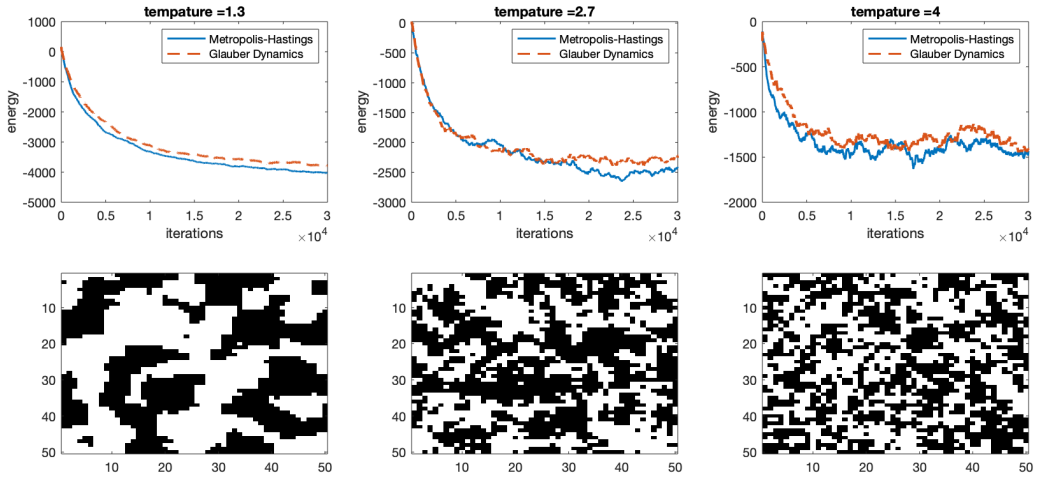

FIG. 4: Simulation of the Ising model using both Metropolis-Hastings and Glauber dyanmics at different temperatures. **(Left)** Low tempature regime with low noise. **(Middle)** Critical regime with noisy structure. **(Right)** High temperature regime with complete disorder.

there are actually 2^{2500} different configurations that can happen. That is literally more than anything in the world! This shows the capability of MCMC methods: it gives us tools to computationally (and mathematically) analyze gigantic systems. This has allowed many calculations that were needed for science and engineering to happen, and is an invaluable tool for both theorists and experimentalists alike.

CODE

All simulations in this project are written by me! You can find the code here: `https://github.com/dannychen0830/RandomCode/blob/main/MATH302demo.m`.

[1] Probability and statistics.
[2] Unfortunately, citation not found.
[3] Andrey nikolayevich kolmogorov.
[4] In fact, we will only outline the proofs rather than proving them. Most full proofs will can be found in *Markov Chain and Mixing Time* by Levin, Wilmer, and Perez [14].
[5] G. P. Basharin, A. N. Langville, and V. A. Naumov, The life and work of aa markov, Linear algebra and its applications **386**, 3 (2004).
[6] E. Cinlar and E. Cnlar, *Probability and stochastics*, Vol. 261 (Springer, 2011).
[7] M. H. Kalos and P. A. Whitlock, *Monte carlo methods* (John Wiley & Sons, 2009).
[8] S. Ballet and R. Rolland, A note on yao's theorem about pseudo-random generators, Cryptography and Communications **3**, 189 (2011).
[9] S. Arora and B. Barak, *Computational complexity: a modern approach* (Cambridge University Press, 2009).
[10] A. C. Yao, Theory and application of trapdoor functions, in *23rd Annual Symposium on Foundations of Computer Science (SFCS 1982)* (IEEE, 1982) pp. 80–91.
[11] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, The journal of chemical physics **21**, 1087 (1953).
[12] W. K. Hastings, Monte carlo sampling methods using markov chains and their applications, (1970).
[13] Glauber's dynamics.
[14] D. A. Levin and Y. Peres, *Markov chains and mixing times*, Vol. 107 (American Mathematical Soc., 2017).