MARKOV CHAINS AND COUPLING FROM THE PAST

DYLAN CORDARO

Abstract. We aim to explore Coupling from the Past (CFTP), an algorithm designed to obtain a perfect sampling from the stationary distribution of a Markov chain. We introduce the fundamentals of probability, Markov chains, and coupling to establish a foundation for CFTP. We then briefly study the hardcore and Ising model to gain an overview of CFTP. Specifically, we will place rigorous bounds on the run time of monotone CFTP, give a brief overview on a non-monotone CFTP method designed for the hardcore model, and sketch an algorithm for using CFTP for an arbitrary Markov chain.

Contents

1. Introduction

This work aims to be an expository paper that builds the fundamentals of Markov chains and CFTP primarily by the proofs of theorems, but in the case of CFTP especially, by examples. We begin with a terse overview of the basics of probability theory (section 3) and proceed with a more detailed run-through of Markov chains (section 4.) Specifically, convergence and coupling are highlighted for the former's importance to the analysis of Markov chains, and the latter's more-than-titular relation to CFTP: thus we prove convergence in terms of coupling in section 6. In section 6, we give another example of a Markov chain, Glauber dynamics, a process that in itself is a Markov chain, and give a specific example of the Glauber dynamics, the hardcore model. In section 7, we finally introduce CFTP (primarily monotone CFTP) by means of the Ising model and

a general schematic. In particular, several theorems of importance in regards to monotone CFTP are realized (mainly, its expected run-time and output distribution), and several examples involving monotone CFTP and non-monotone CFTP (for example, the hardcore model and general Markov chains), are given. Since this work is primarily expository, the majority of proofs were taken from the referenced works.

2. Background

We recall some standard definitions of probability. A more detailed examination can be found in [7].

Definition 2.1. A **Probability Space** is an ordered tuple composed of a **state space** Ω, a **σ-algebra** of some sets from Ω, and a **probability measure** P.

Definition 2.2. The **state space** Ω is comprised of outcomes of a certain process we would like to measure, though in general, we can consider it to be a set.

Definition 2.3. The **σ-algebra** is a collection of subsets of Ω that contains Ω and is closed under countable unions of its elements and taking complements. Elements of this algebra are called events. In the case of a finite Ω, the σ-algebra is generally the power set of Ω.

Definition 2.4. The **probability measure P** (in our finite case, a distribution) is a nonnegative function that maps the σ-algebra to $[0, 1]$, or in other words, takes an event from the sample space, and assigns a number from $[0, 1]$ meaning how probable this event is to occur. In intuitive terms, for an event A, if we repeat an experiment a large number of times (consider flipping a coin, and letting A be the number of heads,) $\mathbf{P}\{A\}$ would return the number of times A occurred over the total number of trials. $\mathbf{P}$ has two properties:

(1) $\mathbf{P}\{\Omega\} = 1$
(2) For any disjoint countable union of events, $\mathbf{P}\{\bigcup A_i\} = \sum \mathbf{P}\{A_i\}$.

For the finite Ω we consider, for any event $A \subset \Omega$, we denote $\mathbf{P}\{A\} = \sum_{x \in A} \mathbf{P}\{x\}$.

We then turn to the main focus of probability, random variables and expectations. Random variables, simply put, offer a way to equate events to numbers, and thus are a connection between probabilities and real-world objects. Expectations are more or less a proxy for the average value of a random variable. Both of these concepts are used extensively in probability.

Definition 2.5. A **random variable** is a function from Ω to $\mathbb{R}$. Thus, for some output x of the random variable X, we define the probability distribution $\mu(x)$ of the random variable X by $\mu(x) = \mathbf{P}\{X = x\} = \mathbf{P}\{e \in \Omega \mid e = X^{-1}(x)\}$. Similarly, we write $\{X \in A\}$ as a shorthand for the set $\{e \in \Omega : X(e) \in A\} = X^{-1}(A)$. Examples of random variables range from the number of heads in a coin flip, money won from gambling, to the number of coupons one obtains from a lottery: for an exploration of the former, see Example 2.10.

Definition 2.6. The **expectation**, or expected value, of a random variable X is a function that returns a real number: $\mathbf{E}(X) = \sum_{x \in Im(X)} x\mathbf{P}\{X = x\}$. Note that the expectation is linear: $\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$.

Definition 2.7. For random variables X, Y and values $x, y \in \mathbb{R}$, the **conditional probability** of x occurring given y is denoted by $\mathbf{P}\{X = x|Y = y\}$.

Definition 2.8. Given a probability space and probability measure $\mathbf{P}$, 2 events are **independent** if $\mathbf{P}\{A, B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$, or, in a more intuitive sense, if $\mathbf{P}\{A|B\} = \mathbf{P}\{A\}$ the probability of A occurring given that B has occurred is equal to the probability of A occurring. Random variables X, Y are independent if for all events $A, B, \cdots \in \mathbb{R}$, the events $\{X \in A\}, \{Y \in B\}$ are independent.

Proposition 2.9. *If events A, B are independent, then the complementary events $A^C := \Omega \setminus A$, B^C are independent as well.*

Proof. We prove that if A and B are independent, then A^C and B are independent, and apply this result twice. Note that $B = A^C \cap B \bigcup A \cap B$. Then $\mathbf{P}\{B\} = \mathbf{P}\{A^C, B\} + \mathbf{P}\{A, B\}$. This implies

$$\begin{aligned}\mathbf{P}\{A^C, B\} &= \mathbf{P}\{B\} - \mathbf{P}\{A, B\} \\ &= \mathbf{P}\{B\} - \mathbf{P}\{A\}\mathbf{P}\{B\} \\ &= \mathbf{P}\{B\}(1 - \mathbf{P}\{A\}) = \mathbf{P}\{B\}\mathbf{P}\{A\},\end{aligned}$$

where the second and third lines follow from the independence of A and B and the second property of a probability measure. Apply this result to finish the proof. □

Example 2.10. Suppose our experiment consists of flipping an unfair coin 2 times. Suppose the probability $\mathbf{P}\{H\}$ of flipping heads is 0.60, and the probability $\mathbf{P}\{T\}$ of landing on tails is 0.40. Our sample space $\Omega = \{HH, HT, TH, TT\}$ is all the combinations of heads and tails for 2 flips, and the sigma algebra is the power set of Ω.

Let X, a random variable, denote the number of heads. If we assume that each flip is independent, then the probability of 2 heads, or $\mathbf{P}\{X = 2\} = 0.60^2$. The expected value of X, is equal to $\sum_{x\in\{0,1,2\}} x\mathbf{P}\{X = x\} = 0 + 1(2 * 0.40 * 0.60) + 2(0.60^2) = 1.2$. If one wants further review, look in [6].

3. Markov Chains

We'll first start off with an example, though the reader is welcome to skip to the formal definition and return to this example later: imagine a fair coin with a biased flipper. Suppose every coin flip starts with the outcome of the last flip (heads or tails.) It turns out that the outcome of the flip is slightly more likely to be the side of the coin that you started with. (See [9] for a more detailed look.) Now, suppose (unrealistically) that the coin has a 60% chance of landing on the side it started on.

We make the following intuitive observations: each flip only depends on the last flip, or the current state of the coin: even if we knew all the flips beforehand, our guess would not change.

Now suppose we start the coin at heads, we are blindfolded, and we flip the coin a large number of times. At some point, if we are asked to guess how likely the next flip of the coin is heads or tails, we would guess $50-50$, and gripe about how no matter how many more times we flip the coin, this distribution will not change: our initial information on the initial state of the coin is practically useless.

We formalize the preceding intuitions as follows: we represent the transition probabilities of the coin by a matrix P. For notation, let 0 represent the coin being heads, and 1 represent a tails. Then let P have the property that the i, j th entry corresponds to the probability that the coin, starting on state i, when flipped, lands on state j. We denote this probability by $P(i, j)$. Then

$$P = \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix} = \begin{pmatrix} 0.60 & 0.40 \\ 0.40 & 0.60 \end{pmatrix}.$$

We now make the following statements about P and the intuitions expressed above. If we represent an initial probability distribution of the coin's state as a row vector $p = (\cdot, \cdot)$ corresponding to the probability that the coin is heads or tails respectively, pP^t corresponds to the probability distribution that after t units of time (or flips), the coin is heads or tails, respectively. (If this is not clear, start the coin on tails and multiply by the matrix a few times.)

The rows of P will always add to 1, while it is not guaranteed that the columns will (change the transition probabilities starting from tails to be $(0.70, 0.30)$ to see why.)

After a long period of time, pP^t converges to some probability distribution π, which must have the property $\pi = \pi P$. The reader can determine, for this specific example, that $\pi = (0.50, 0.50)$ by solving the equation formed by the above relation and proving that for each time step, the distribution pP^t approaches π. These ideas will be formalized below. Another example of a Markov chain, and the proofs in the following two sections can be found in [6]

Definition 3.1. The above is an example of a **Markov Chain**: a sequence of random variables $(X_0, X_1, \dots)$ and transition matrix $P \in M_{|\Omega|}(\mathbb{R})$ that satisfy the Markov property:

$$\mathbf{P}\{X_{t+1} = x \mid X_0 = x_0 \cap \dots X_t = y\} = \mathbf{P}\{X_{t+1} = x \mid X_t = y\} = P(x, y)$$

The Markov property conveys the notion that the probability of the outcome of the coin flip depends only on the coin's most recent state.

The transition matrix P satisfies 2 properties:

(1) The entry in the x-th row and y-th column of P, $P(x, y)$, is defined as the probability of transitioning to state y given that one is currently in state x, as seen above in the Markov property.
(2) Given that the x-th row of P is the distribution $P(x, \cdot)$, P is stochastic: its entries are nonnegative, and its rows sums are 1. Formally, for all $x \in \Omega$,

$$\sum_{y \in \Omega} P(x, y) = 1.$$

Remark. For a given probability distribution μ and time t, we denote the Markov chain with transition matrix P started from an initial distribution μ as μP^t. We define $\mathbf{P}_\mu\{\cdot\}, \mathbf{E}_\mu(\cdot)$ to be the probability distributions and expectations when the starting distribution is μ. For shorthand, we define $\mathbf{P}_x\{X_t = y\} = P^t(x, y)$ to be the distribution of the Markov chain with transition matrix P started from an initial distribution of 0 (if $y \neq x$), and 1 in the x-th index.

Definition 3.2. The **stationary distribution** π of a Markov chain with transition matrix P is a probability distribution that satisfies $\pi = \pi P$, or equivalently, for each state $x \in \Omega$,

$$\pi(x) = \sum_{y \in \Omega} \pi(y) P(y, x).$$

An easy way to check if a distribution is stationary is to see if a probability distribution π on Ω satisfies the detailed balance equations:

$$\pi(x) P(x, y) = \pi(y) P(y, x)$$

If those equations hold, then

$$\sum_{y \in \Omega} \pi(y) P(y, x) = \sum_{y \in \Omega} \pi(x) P(x, y) = \pi(x).$$

Definition 3.3. A Markov chain (X_t) is called **reversible** if there exists a probability distribution π such that the detailed balance equations hold. If a chain is reversible, then

$$\pi(x_0)P(x_0, x_1)\dots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1})\dots P(x_1, x_0)$$
$$\mathbf{P}_\pi\{X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\} = \mathbf{P}_\pi\{X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0\}$$

In other words, if a chain (X_t) satisfies the detailed balance equations and has a stationary initial distribution, then any sequence of states $x_0, \dots, x_n$ has the same distribution as $x_n, x_{n-1}, \dots, x_0$.

In the example above, we intuited that the probability distribution of the Markov chain with the coin converged to a stationary distribution over a long period of time. The following theorems below characterize the transition matrices of Markov chains and when the chain converges to a unique stationary distribution. For the reader's sake, these definitions and proofs follow [6].

Definition 3.4. A chain P is called **irreducible** if for all $x, y \in \Omega$ there exists a time $t \in \mathbb{Z}$ such that $P^t(x, y) > 0$. This means that for any states x and y, there is a way to get from x to y in t transitions with positive probability.

Definition 3.5. Let $\mathcal{T}(x) := \{t \geq 1 : P^t(x, x) > 0\}$ be the set of times where it is possible for the chain P to return to a starting position x. The **period** of state x is defined as $gcd(\mathcal{T}(x))$. If for all states in a Markov chain, the period is 1, that Markov chain is called **aperiodic**. A **periodic** Markov chain is a Markov chain that is not aperiodic.

We now move onto a proposition that, while not used for a theorem in this paper, establishes that for irreducible chains, the period does not depend on the state.

Proposition 3.6. *If P is irreducible, then for any states x, y, the period of x equals the period of y, or $gcd(\mathcal{T}(x)) = gcd(\mathcal{T}(y)$*

Proof. Since P is irreducible, we know that there exist times $r, s \geq 0$ such that $P^r(x, y) > 0$ and $P^s(y, x) > 0$. Let $m = r + s$. Then $m \in \mathcal{T}(x) \cap \mathcal{T}(y)$. Since for any $t_y \in \mathcal{T}(y)$ we know that $P^r(x, y)P^{t_y}(y, y)P^s(y, x) > 0$, $gcd(\mathcal{T}(x))|r + t_y + s$ and consequently that $gcd(\mathcal{T}(x)) \mid t_y$. Thus $gcd(\mathcal{T}(x)) \geq gcd(\mathcal{T}(y))$ This can be repeated for $gcd(\mathcal{T}(y))$. □

I will briefly sketch a lemma that, given an irreducible Markov chain, allows one to establish a universal time for which all states can be reached from each other. This lemma is essential to our proof of convergence in section 4, but we skim the number-theoretic fact because the idea is fairly intuitive. The full proof is in [6].

Lemma 3.7. *If P is aperiodic and irreducible, then there is an integer r such that for all states $x, y \in \Omega$, $P^r(x, y) > 0$.*

Proof. **(Sketch)** We use the following number theoretic fact: for any set of positive integers that is closed under addition and has a gcd of 1, that set contains all but finitely many positive integers. Since P is aperiodic, and $\mathcal{T}(x)$ for any state x is closed under addition, we can apply the fact to $\mathcal{T}(x)$. From there, for each state x, we can use P's irreducibility and the number theoretic fact above to establish a minimum time t_x such that for any time $t > t_x$ and any state y, $P^{t_x}(x, y) > 0$. Take the maximum of the t_x. Then for any time t greater than that maximum, $P^t(x, y) > 0$. □

For the remainder of this section, we prove the existence of the unique stationary distribution using the characterization of the stationary distribution as the number of times one visits a state on average, prove the uniqueness through an argument from the definition. In the next section, we

prove that aperiodic and irreducible Markov chains converge to this distribution using coupling to segue into Coupling from the Past. The link `http://www.cc.gatech.edu/~vigoda/MCMC_Course/MC-basics.pdf` provides a proof of all the above using coupling, if one so wishes.

Definition 3.8. For $x \in \Omega$, define the **hitting time** of x to be $\tau_x := min\{t \geq 0 : X_t = x\}$ the minimum amount of time it takes a Markov chain to reach the state x. We define τ_x^+ to be the positive hitting time wherein $(t \geq 1)$. When $X_0 = x$, we call τ_x^+ to be the first return time.

Our goal for the remainder of this section is to prove that, the stationary distribution π exists, is unique, and that asymptotically for each state x, $\pi(x) = \frac{number\ of\ times\ x\ visited}{visits\ to\ all\ states}$. We wish to prove this by noting that the expected value of the hitting time for any state in an irreducible chain is finite. We then prove that the stationary distribution of the chain takes the form above.

Lemma 3.9. *For any states x, y in an irreducible Markov Chain P, $\boldsymbol{E}_x(\tau_y^+) < \infty$*

Proof. Note that $E_x(\tau_y^+) = \sum_{t=0} t\mathbf{P}\{\tau_y^+ = t\}$. We want to change this into an easier to understand form: a probability in terms of $\mathbf{P}\{\tau_y^+ > t\}$, and possibly get rid of the sum over t. We accomplish this by noting that for any non-negative integer-valued random variable (in this case, $X_t = t$),

$$\mathbf{E}(Y) = \sum_{t \geq 0} \mathbf{P}\{Y > t\}.$$

Note that:

$$\begin{aligned}
\mathbf{E}(Y) &= \sum_{t=0} t\mathbf{P}\{Y = y\} \\
&= 0\mathbf{P}\{Y = 0\} + 1\mathbf{P}\{Y = 1\} + 2\mathbf{P}\{Y = 2\} + 3\mathbf{P}\{Y = 3\} + \dots \\
&= \begin{array}{ccccccccccc}
\mathbf{P}\{Y = 1\} & + & & & & & & & & & \\
\mathbf{P}\{Y = 2\} & + & \mathbf{P}\{Y = 2\} & + & & & & & & & \\
\mathbf{P}\{Y = 3\} & + & \mathbf{P}\{Y = 3\} & + & \mathbf{P}\{Y = 3\} & + & & & & & \\
\mathbf{P}\{Y = 4\} & + & \mathbf{P}\{Y = 4\} & + & \mathbf{P}\{Y = 4\} & + & \mathbf{P}\{Y = 4\} & + & & & \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \ddots
\end{array} \\
&= \sum_{t \geq 0} \mathbf{P}\{Y > t\};
\end{aligned}$$

where the last line follows from summing the columns. To get a bound on the probability associated with this, we use the irreducibility of P to find an r such that for all states x, y, $P^r(x, y) = \epsilon > 0$. We know that for all positive real t, $\mathbf{P}_x\{\tau_y^+ > t\}$ is a decreasing function of t. So, we know that for an integer k, and for an state j that the Markov chain arrived at

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq \mathbf{P}_j\{\tau_y^+ > r\}\mathbf{P}_x\{\tau_y^+ > (k-1)r\} \leq (1-\epsilon)\mathbf{P}_x\{\tau_y^+ > (k-1)r\}.$$

By repeated iteration, we see that

$$\mathbf{P}_x\{\tau_y^+ > kr\} \leq (1-\epsilon)^k.$$

Thus,

$$\mathbf{E}_x(\tau_y^+) = \sum_{t \geq 0} \mathbf{P}_x\{\tau_y^+ > t\} \leq \sum_{k=0} r\mathbf{P}_x\{\tau_y^+ > kr\} \leq r\sum_{k=0}(1-\epsilon)^k = \frac{r}{\epsilon} < \infty.$$

because both ϵ and r are positive and fixed. □

Now, we proceed to the main part of the proof, which constructs the stationary distribution.

Proposition 3.10. *Let P be the transition matrix of an irreducible Markov chain. Then there exists a probability distribution on Ω such that $\pi = \pi P$ and*

$$\pi(x) = \frac{1}{\boldsymbol{E}_x(\tau_x^+)} \mid \textit{for all states } x \in \Omega,\ \pi(x) > 0$$

Proof. Let $z \in \Omega$ be an arbitrary state of the Markov chain: we want to see how many times the chain hits other states before it returns to z.

Thus, define $\bar{\pi}(y)$ to be

$$\mathbf{E}_z(\text{the expected number of visits to } y \text{ before returning to } z) = \sum_{t=0} \mathbf{P}_z\{X_t = y, t < \tau_z^+\}$$

We know that for any state y,

$$\bar{\pi}(y) = \sum_{t=0} \mathbf{P}\{X_t = y, t < \tau_z\} \leq \sum_{t=0} \mathbf{P}\{t < \tau_z^+\} = \mathbf{E}_z(\tau_z^+) < \infty,$$

so we now want to know if $\bar{\pi}$ is stationary for P (noting that $\bar{\pi}$ is clearly not a probability distribution, but can be normalized.) We start from the definition. Note that

$$\begin{aligned}
\sum_{x\in\Omega} \bar{\pi}(x)P(x,y) &= \sum_{x\in\Omega}\sum_{t=0} \mathbf{P}_z\{X_t = x, t \leq \tau_z^+\}P(x,y) \\
&= \sum_{x\in\Omega}\sum_{t=0} \mathbf{P}_z\{X_{t+1} = y, X_t = x, t+1 \leq \tau_z^+\} \\
&= \sum_{t=0}\sum_{x\in\Omega} \mathbf{P}_z\{X_{t+1} = y, X_t = x, t+1 \leq \tau_z^+\} \\
&= \sum_{t=0} \mathbf{P}_z\{X_{t+1} = y, t+1 \leq \tau_z^+\} \\
&= \sum_{t=1} \mathbf{P}_z\{X_t = y, t \leq \tau_z^+\} = \bar{\pi}(y).
\end{aligned}$$

The simplification from the first to second line follows from the dependence of X_{t+1} on the state of X_t, and that τ_z^+ depends only on $X_1, \ldots, X_t$. The third line follows from noting that the order of summation does not change the value of the sum.

We rearrange the result into a form that looks a lot like our desired result of $\bar{\pi}(y)$.

$$\begin{aligned}
\bar{\pi}(y) &= \sum_{t=1} \mathbf{P}_z\{X_t = y, t \leq \tau_z^+\} \\
&= \sum_{t=0} \mathbf{P}_z\{X_t = y, t < \tau_z^+\} + \sum_{t=1} \mathbf{P}_z\{X_t = y, t = \tau_z^+\} - \mathbf{P}_z\{X_0 = y, 0 < \tau_z^+\} \\
&= \bar{\pi}(y) + \sum_{t=1} \mathbf{P}_z\{X_t = y, t = \tau_z^+\} - \mathbf{P}_z\{X_0 = y\}
\end{aligned}$$

We just need to prove that the latter two terms are 0 to finish the proof.

(1) Suppose $y = z$. Then clearly, at the minimum return time, $X_{\tau_x^+} = z$, and at any other time, 0, while $X_0 = z$ is clearly true. Then the latter two terms equal 1, and subtract out.
(2) Suppose $y \neq z$. Clearly, both of the latter terms are 0. Since the first term of the sum is $\bar{\pi}(y)$, this proves $\bar{\pi}$ is a stationary distribution.

Thus, we can define $\pi(x) = \frac{\bar{\pi}(x)}{\sum_{y\in\Omega}\bar{\pi}(y)} = \frac{\bar{\pi}(x)}{\mathbf{E}_z(\tau_z^+)}$, where the denominator is the total number of visits to all other states over time. Clearly, this is a probability distribution. An alternative way to express this is $\pi(x) = \frac{1}{\mathbf{E}_x(\tau_x^+)}$, where this fraction literally represents the number of visits to x over the total number of visits to other states before the first return time. □

However, we have a fairly alarming possibility: what if there are two distinct stationary distributions? It turns out that there is a fairly elementary proof to determine that this is not the case.

Proposition 3.11. *Let P be an irreducible transition matrix of a Markov chain. The stationary distribution π of P is unique.*

Proof. Suppose $\mu \neq \upsilon$ are distinct stationary probability distributions on P. We know the following two facts:

(1) Since P is irreducible, all states communicate with each other, so for each state x, $\mu(x) > 0$ and $\upsilon(x) > 0$. If, for contradiction, some state $z \in \Omega$, $\mu(z) = 0$, then for all times t, $\mu P^t(z) = \mu(z) = 0$ because μ is stationary. However, we also know that for some state y where $\mu(y) > 0$ and for some time t, $0 < \mu(y)P^t(y,z) \leq \mu P^t(z)$ because P is irreducible, and μ is a positive probability distribution. This is a contradiction, so for all states y, $\mu(y) > 0$.

(2) Because μ and υ are stationary, we know that any nontrivial linear combination $a\mu + b\upsilon$ is a stationary distribution on P, because

$$\sum_{y\in\Omega} a\mu(y,x) + b\upsilon(y,x) = a\sum_{y\in\Omega}\mu(y,x) + b\sum_{y\in\Omega}\upsilon(y,x) = a\mu(x) + b\upsilon(x).$$

However, we can find a stationary distribution $\rho = \mu - b\upsilon$ for some positive b with the property that there exists a state $x \in \Omega$ such that $\rho(x) = 0$, and for all other states y, $\rho(y)$ is nonnegative. Because $\mu \neq \upsilon$, and that for each state x, the function $f_x(t) := \mu(x) - t\upsilon(x)$ is nonconstant because for all states z, $\upsilon(z) \neq 0$, and attains a 0 at $t_x = \frac{\mu(x)}{\upsilon(x)}$. Set b to be the minimum t_x, normalize ρ, and we will have a stationary distribution ρ on an irreducible chain that has a state with probability 0, which is a contradiction by the first fact of stationary distributions in this proof. □

Thus, for all aperiodic and irreducible chains, there exists a unique stationary distribution. However, we have no way of knowing if chains converge to this distribution, let alone how fast. We also have not addressed a way to classify chains in terms of their states or of visualizing periodic and irreducible chains: such concerns are discussed in [6].

4. Convergence and Coupling

As stated before, all irreducible and aperiodic Markov chains have some stationary distribution: but now, we want to be able to say how fast a Markov chain converges to its stationary distribution, if it even converges at all. We first accomplish this by giving a measure with which to compare two distributions, define a coupling, and prove that irreducible Markov chains do converge to their stationary distribution.

Definition 4.1. The **total variation distance** between two probability distributions μ, υ, is denoted by $\|\mu - \upsilon\|_{TV} = \max_{A\subset\Omega} |\mu(A) - \upsilon(A)|$ the maximum difference in the probabilities of one event. Areas I and II in Figure 1 below are equal to $\|\mu - \upsilon\|_{TV}$ (if one overlooks the lack of a scale.)

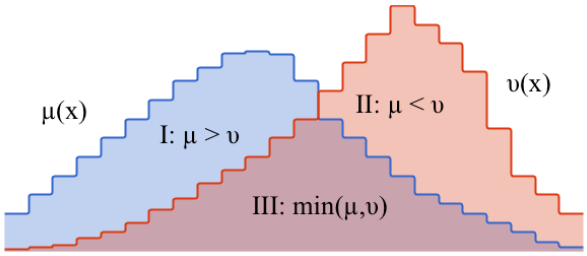


FIGURE 1. Total variation distance

There are two main characterizations of the total variation distance that we shall use in this paper: one of which is a summing over the sample space, and the other a coupling.

Proposition 4.2. *For probability distributions* μ, υ, $\|\mu - \upsilon\|_{TV} = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \upsilon(x)|$.

Proof. Figure 1 pretty much describes all we need: the total variation distance is the difference in area between μ and υ. Let $B = \{x : \mu(x) > \upsilon(x)\}$. For any set $A \subset \Omega$,

$$\mu(A) - \upsilon(A) = \mu(A\cap B) + \mu(A\cap B^C) - \upsilon(A\cap B) - \upsilon(A\cap B^C) \leq \mu(A\cap B) - \upsilon(A\cap B) \leq \mu(B) - \upsilon(B).$$

The fact that $\upsilon(A) - \mu(A) \leq \upsilon(B^C) - \mu(B^C) = \mu(B) - \upsilon(B)$ follows similarly by considering the probability of the complement. Then, clearly, we can let $A = B$ or $A = B^C$ and attain our upper bound. Thus

$$\|\mu - \upsilon\|_{TV} = \sum_{x\in B} |\mu(x) - \upsilon(x)| = \sum_{y\in B^C} |\mu(y) - \upsilon(y)|.$$

But since $B \cup B^C = \Omega$, $\|\mu - \upsilon\|_{TV} = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \upsilon(x)|$. □

Remark. Note that the proof of Proposition 4.2 shows $\|\mu - \upsilon\|_{TV} = \sum_{x\in\Omega, \mu(x)\geq\upsilon(x)} [\mu(x) - \upsilon(x)]$.

We hold off on proving the equivalence of the second definition of total variation distance until we explain what coupling is. In informal terms, a coupling is an attempt to simulate two probability distributions by means of random variables sharing randomness. We can represent this shared distribution in the terms of the random variables by a joint distribution. Since random variables are easier to work with than distributions, we can make our lives much easier. We now define coupling and then prove the convergence theorem.

Definition 4.3. A **coupling** of two probability distributions μ and υ is a pairing of two random variables X and Y, which we denote (X, Y), defined on a single probability space with the property that the marginal distribution of X is μ, and the marginal distribution of Y is υ. In other words, $\mu(x) = \mathbf{P}\{X = x\}$ and $\upsilon(y) = \mathbf{P}\{y = y\}$.

An example of coupling can be expressed with coins: let μ, υ be the "fair coin" measuring weight giving a weight of 0.50 to the elements of $\{0, 1\}$. We can couple μ and υ in 2 ways:

(1) Let X be the result of the first coin, and yet Y be the result of the second. Then for all flips, $\mathbf{P}\{X = x, Y = y\} = \frac{1}{4}$.
(2) Let X be the result of the first coin, and let $Y = X$. Then $\mathbf{P}\{X = Y\} = 1, \mathbf{P}\{X \neq Y\} = 0$, and $\mathbf{P}\{X = 1\} = \mathbf{P}\{X = 0\} = 0.50$.

We define a joint distribution q of (X, Y) on $\Omega \times \Omega$ to be $q(x, y) = \mathbf{P}\{X = x, Y = y\}$. In matrix form, for two random variables with two possible values, q would look like

$$\begin{array}{ccc} & x & y & \mu \\ x & q(x,x) & q(x,y) & \mu(x) \\ y & q(y,x) & q(y,y) & \mu(y) \\ \upsilon & \upsilon(x) & \upsilon(y) & \end{array}$$

where in this form, x, y are the random variables, the marginal distribution corresponds to the sums of the rows and columns, appropriately: $\mu(x) = \sum_{y \in \Omega} q(x, y)$ and vice versa. The reader can check that the two couplings for the coins have the same marginal distributions, though their joint distributions are different.

We now ask a question: how is the total variation distance between probability distributions related to their couplings? For the second part of the above example, the answer is evident: if both random variables are identical, $\mathbf{P}\{X \neq Y\} = \|\mu - \upsilon\|_{TV} = 0$. But what if the distributions are different, and we cannot set $X = Y$? Regions I and II in Figure 1 are the answer.

Proposition 4.4. *Let μ and υ be two probability distributions on Ω. Then $\|\mu - \upsilon\|_{TV} = \inf(\boldsymbol{P}\{X \neq Y\})$ for all couplings (X, Y), and there is a coupling (X, Y) that attains this bound.*

Proof. Our proof will be structured around finding a coupling (X, Y) that differs for a state x only when $\mu(x) \neq \upsilon(x)$, or in other words, setting $X \neq Y$ if x is in region III in the picture. We first note that for any event $A \subset \Omega$ and coupling (X, Y) of μ and υ,

$$\begin{aligned} \mu(A) - \upsilon(A) &= \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} \\ &= \mathbf{P}\{X \in A, Y \notin A\} + \mathbf{P}\{X \in A, Y \in A\} - [\mathbf{P}\{Y \in A, X \in A\} + \mathbf{P}\{Y \in A, X \notin A\}] \\ &\leq \mathbf{P}\{X \in A, Y \notin A\} \\ &\leq \mathbf{P}\{X \neq Y\}. \end{aligned}$$

It follows from the definition that $\|\mu - \upsilon\|_{TV} \leq \inf P(X \neq Y)$ Now, from the diagram above, we aim to construct a coupling (X, Y) that is as equal as can be. Informally, our coupling can be likened to choosing a point in the union of the regions I, II, III: when we land in region III, set $X = Y$, and if we land in region I or II, set $X \neq Y$. The formal coupling argument simply involves transforming this intuition to X and Y.

Let $p = \sum_{x \in \Omega} \min(\mu(x), \upsilon(x)$. This is the area of region III. We wish to determine the exact value of p. Note that

$$\begin{aligned} p &= \sum_{x \in \Omega} \min(\mu(x), \upsilon(x)) \\ &= \sum_{x \in \Omega, \mu(x) > \upsilon(x)} \upsilon(x) + \sum_{x \in \Omega, \mu(x) \leq \upsilon(x)} \mu(x) \\ &= \sum_{x \in \Omega, \mu(x) > \upsilon(x)} \upsilon(x) + [1 - \sum_{x \in \Omega, \mu(x) > \upsilon(x)} \mu(x)] \\ &= 1 - \sum_{x \in \Omega, \mu(x) > \upsilon(x)} \mu(x) - \upsilon(x)] \\ &= 1 - \|\mu - \upsilon\|_{TV} \end{aligned}$$

where the last line follows as a straightforward consequence of the proof of proposition 4.2.

Now imagine flipping a coin with a probability of heads equal to p. We have two cases for constructing our coupling.

(1) The coin is heads (we landed in region III): choose a value Z according to the probability distribution
$$\gamma_{III}(x) = \frac{min\big(\mu(x), \upsilon(x)\big)}{p}$$
Then let $X = Y = Z$.

(2) The coin is tails: (we landed in region I or II.) Choose X according to the probability distribution
$$\gamma_I(x) = \begin{cases} \frac{\mu(x)-\upsilon(x)}{1-p} & \text{if } \mu(x) > \upsilon(x), \\ 0 & \text{otherwise.} \end{cases}$$
Note that the first case for γ_I is equivalent to the point landing in region I, and the second to landing in region II. Independently choose Y according to the probability distribution
$$\gamma_{II}(x) = \begin{cases} \frac{\upsilon(x)-\mu(x)}{1-p} & \text{if } \mu(x) < \upsilon(x), \\ 0 & \text{otherwise.} \end{cases}$$

It is evident from the picture that
$$p\gamma_{III} + (1-p)\gamma_I = \mu,$$
$$p\gamma_{III} + (1-p)\gamma_{II} = \upsilon.$$
Thus the distribution of X is μ, and the distribution of Y is υ. The only time $X \neq Y$ is when the coin is tails: thus $\mathbf{P}\{X \neq Y\} = 1 - p = \|\mu - \upsilon\|_{TV}$. □

While constructing a coupling of two distributions is pretty useful, as we did before, the concept of coupling Markov chains is fairly related to CFTP.

Definition 4.5. A **coupling of Markov chains** with transition matrix P is a process $(X_t, Y_t)_{t=0}^{\infty}$ in which both (X_t) and (Y_t) are Markov chains with transition matrix P.

Remark. We can construct a coupling of Markov chains with the property that if $X_s = Y_s$ then for all times $t \geq s$, $X_t = Y_t$ by running each Markov chain independently according to P until the first time X_t and Y_t simultaneously reach the same state, and then setting $X_t = Y_t$ afterwards. This is equivalent to running both chains by the original coupling (the joint distribution) until both chains meet.

We now bound the total variation distance from the stationary distribution in order to prove the convergence theorem, and thus allow us to obtain bounds on mixing times. The following definition quantifies the distance from the stationary distribution π.

Definition 4.6. Let π be the stationary distribution of P, and t be an arbitrary positive time. We define
$$d(t) := \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_{TV}$$
$$\bar{d}(t) := \max_{x,y\in\Omega} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV}$$
$d(t)$ quantifies the total variation distance from the origin, while $\bar{d}(t)$ is used because we often can bound $\bar{d}(t)$ uniformly over all pairs of states (x, y).

With $d(t)$ defined, we can now quantify how "close" a Markov chain is to the stationary distribution π for a given time.

Definition 4.7. The **mixing time** of a Markov chain is the minimum time t such that for a given nonnegative real ϵ ,

$$t_{mix}(\epsilon) := min\{t : d(t) \leq \epsilon\}$$

We define T_{mix} to be $t_{mix}(\frac{1}{e})$.

We use the following proposition fairly often to bound the mixing time.

Proposition 4.8. *If $d(t)$ and $\bar{d}(t)$ are defined as in Definition 4.6, then*

$$d(t) \leq \bar{d}(t) \leq 2d(t).$$

We refer the reader to page 54 of [6].

Proposition 4.9. *Let $\{(X_t, Y_t)\}$ be a coupling satisfying the remark above, for which $X_0 = x$ and $Y_0 = y$. Let $\tau_c := min\{t : X_t = Y_t\}$, i.e., the first time the chains meet. Then*

$$\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \leq \boldsymbol{P}_{x,y}\{\tau_c > t\}.$$

Proof. Note that $P^t(x,z) = \mathbf{P}_{x,y}\{X_t = z\}$ and $P^t(y,z) = \mathbf{P}_{x,y}\{Y_t = z\}$. This implies that the marginal distributions of X_t and Y_t are $P^t(x,\cdot)$ and $P^t(y,\cdot)$ respectively. Thus (X_t, Y_t) is a coupling of $P^t(x,\cdot)$ and $P^t(y,\cdot)$, and by Proposition 4.4,

$$\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \leq \inf(\mathbf{P}\{X \neq Y\}) \leq \mathbf{P}\{X_t \neq Y_t\}$$

But $\mathbf{P}\{X_t \neq Y_t\} = \mathbf{P}\{\tau_c > t\}$ because X_t and Y_t were constructed to run together after they meet. □

Corollary 4.10. *Suppose for every $x, y \in \Omega$ there exists a coupling (X_t, Y_t) with $X_0 = x, Y_0 = y$. Then $d(t) \leq \max_{x,y\in\Omega} \boldsymbol{P}_{x,y}\{\tau_c > t\}$*

Proof. This proof follows from the definition of $\bar{d}(t)$ and Proposition 4.9. □

Now, at last, we get to the Convergence Theorem, which we will prove with coupling.

Proposition 4.11. *(Convergence Theorem) Suppose P is irreducible and aperiodic with stationary distribution π. Then there exist a constant $\epsilon \in (0,1)$ such that $d(t) \leq (1-\epsilon^2)^t$.*

Proof. Let (X_t, Y_t) be a coupling of μ, υ. This means X_0 is a random variable that has values picked in accordance to μ, and Y_0 is a random variable that has values picked in accordance to υ. We denote this by $X_0 \sim \mu$, and $Y_0 \sim \upsilon$. Since the marginal distributions of X_t and Y_t are μP^t and υP^t, by the same reasoning in Proposition 4.9, and by Proposition 4.4,

$$\|\mu P^t - \upsilon P^t\|_{TV} \leq \mathbf{P}_{x,y}\{X_t \neq Y_t\} = \mathbf{P}\{\tau_c > t\}.$$

Now note that we can take $\pi = \upsilon$, so that the above inequality bounds the distance from the stationary distribution. We only need to prove that there is a coupling (X_t, Y_t), where X_t and Y_t move from state to state according to P, are independent of each other but structured to run together when they meet, and meet after a finite amount of time.

Since P is irreducible, by Lemma 3.7, there exists a positive integer r and a positive real ϵ such that for all $x, y \in \Omega$, $P^r(x,y) > \epsilon > 0$. Now consider the probability that for some x_0, $X_t \neq x_0$ and $Y_t \neq x_0$, or just that $P^t(X_t \neq Y_t)$. By complementary probabilities and independence, we know

$$P^t(X_t \neq Y_t) = 1 - P^t(X_t = x_0, Y_t = x_0) = 1 - P^t(X_t = x_0)P^t(Y_t = x_0) \leq 1 - \epsilon^2.$$

Now note that $X_{kr} \neq Y_{kr}$ is roughly equal to the instance that for k times, we iterated the coupling r times from arbitrary states and found that $X_r \neq Y_r$. It follows that $P^{kr}(X_t \neq Y_t) \leq$

$(1-\epsilon^2)^k$. Thus, from the first part of the proof, for our independent coupling (X_t, Y_t), $t = kr > 0$, as $k \to \infty$,

$$\|\mu P^t - \pi\|_{TV} \leq \mathbf{P}\{\tau_c > t\} = \mathbf{P}_{x,y}\{X_t \neq Y_t\} \leq (1-\epsilon^2)^k \to 0.$$

□

5. Glauber dynamics and the hardcore configuration

For our examples of CFTP, we use the Glauber dynamics, also known as the single-site heat bath algorithm, extensively along with the notion of a grand coupling. One such example is the hardcore model. In this section, we introduce Glauber dynamics and grand coupling in the context of bounding the mixing time of the hardcore model to give an explicit example of the material addressed in section 4.

The hardcore model can be imagined as a finite, undirected graph G with vertices in the vertex set V either possessing a particle (state 1), or not possessing a particle (state 0.) States, or configurations, are subsets of the vertex set V of G. For notation, for a given configuration of particles σ and vertex v, let $\sigma(v)$ return v's state (in this case, 0 or 1.) We call vertices of state 1 occupied, and we call vertices of state 0 vacant. Legal states are those vertex subsets of G with no two occupied vertices being adjacent to each other, i.e., for any vertices v, w connected by an edge, $\sigma(v)\sigma(w) = 0$. This Markov chain is iterated in the following manner:

(1) Choose a vertex v uniformly at random.
(2) If there is a particle adjacent to v, set v to state 0, while if v has no occupied neighbor, place a particle on v with probability 0.50.

The model below is a more generalized version of the hardcore model above.

Definition 5.1. The **hardcore model** with **fugacity** λ is the probability distribution π defined by

$$\pi(\sigma) = \begin{cases} \lambda^{\sum_{v\in V}\sigma(v)}/Z(\lambda) & \text{if for all } \{v,w\} \in E,\ (\sigma(v)\sigma(w) = 0) \\ 0 & \text{otherwise.} \end{cases}$$

As before, $Z(\lambda) = \sum_{\sigma\in\Omega} \lambda^{\sum_{v\in V}\sigma(v)}$ is a normalizing factor to ensure that π is a probability distribution. The single site heat bath algorithm for this model updates a configuration X_t as follows:

(1) Choose a vertex v uniformly at random.
(2) If v has no occupied neighbors, set $X_{t+1}(v) = 1$ with probability $\frac{\lambda}{1+\lambda}$, and set v to 0 with probability $\frac{1}{1+\lambda}$.
    If v has an occupied neighbor, set $X_{t+1}(v) = 0$.
(3) For all other vertices $w \in V$, set $X_{t+1}(w) = X_t(w)$.

The hardcore model with fugacity λ is an example of Glauber dynamics. We now introduce the general definition of the Glauber dynamics for a chain.

Definition 5.2. Let V be a finite set of vertices. Let S be a finite set of values each vertex can take. Suppose that the sample space Ω is a subset of S^V (the configurations of vertices with the values of S.) Let π be a probability distribution on the sample space. The **Glauber dynamics** for π is a reversible Markov chain with state space Ω, stationary distribution π, and transition probabilities from an arbitrary legal state x as below.

(1) Pick a vertex v uniformly at random from V.

(2) Choose a new state according to π conditioned on the set of states equal to $x \in \Omega$ at all vertices except possibly v: we denote this set by

$$\Omega(x, v) = \{y \in \Omega : y(w) = x(w) \text{ for all } w \neq v\}.$$

The distribution π conditioned on $\Omega(x, v)$ is defined as

$$\pi^{x,v}(y) = \pi(y|\Omega(x, v)) = \begin{cases} \frac{\pi(y)}{\pi(\Omega(x,v))} & \text{if } y \in \Omega(x, v), \\ 0 & \text{if } y \notin \Omega(x, v). \end{cases}$$

Note that π is stationary and reversible for the single-site heat bath algorithm. It is a worthwhile exercise to prove this fact (the problem is in [6],) but for the purposes of this paper, we do not need to dig that deeply.

Now, we aim to prove that the mixing time of the hardcore model with a fugacity λ small enough and n vertices is of the order $n \log(n)$. We accomplish this by considering a grand coupling:

Definition 5.3. A **grand coupling** is a collection of random variables $\{X_t^x \mid x \in \Omega, t = 0, 1, 2, \dots\}$ started from each state $x \in \Omega$, governed by the same independent identically distributed random variables $(Z_1, Z_2, \dots)$, such that the sequence $(X_t^x)_{t=0}^{\infty}$ is a Markov chain started from the state x, with transition matrix P. This is accomplished with a **random mapping representation** of P: a way of representing a transition matrix as a function $f : \Omega \times Q \mapsto \Omega$ that takes in a state $x \in \Omega$ and a Q-valued random variable Z, which for our purposes, will most likely be distributed on $[0, 1]$, that satisfies

$$\mathbf{P}\{f(x, Z) = y\} = P(x, y).$$

We proceed to the proof of the bound on the mixing time of the hardcore model with fugacity λ to show an example of a grand coupling.

Theorem 5.4. *Let $c_H(\lambda) = \frac{1+\lambda(1-\Delta)}{1+\lambda}$. For the hardcore model with fugacity λ, n vertices, and maximum degree Δ (the **degree** of a vertex v is the number of edges with v at one end,) if $\lambda < \frac{1}{\Delta-1}$, then*

$$t_{mix}(\epsilon) \leq \frac{n}{c_H(\lambda)}[\log n + log\frac{1}{\epsilon}].$$

Proof. We construct a grand coupling that first chooses a vertex v uniformly at random, and then flips a coin with probability $\frac{\lambda}{1+\lambda}$ of heads. Every hardcore configuration in Ω is then updated at v according to the results of the flip. If heads, a particle is placed at v if there is no neighboring occupied vertex, and if tails, any present particle at v is removed.

Let x, y be arbitrary hardcore configurations. Now let $\rho(x, y) = \sum_{v \in V} 1_{x(v) \neq y(v)}$ the number of sites in x and y that disagree. It is evident that ρ is a metric, i.e., that $\rho(x, y)$ is nonnegative, $\rho(x, y) = 0$ if and only if $x = y$, ρ is symmetric, and for any states x, y, z, $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (subadditive).

Suppose x and y satisfy $\rho(x, y) = 1$, which means that they differ at a vertex v_0. WLOG, assume $x(v_0) = 1, y(v_0) = 0$. We have 3 cases.

(1) If v_0 is selected, then the next time step in the chains started from states x and y satisfies $\rho(X_1^x, X_1^y) = 0$.
(2) If a neighbor w of v_0 is selected and heads is flipped, then w will have a particle removed in x, and filled with a particle in y. Thus $\rho(X_1^x, X_1^y) = 2$.
(3) If a neighbor w of v_0 is selected and tails is flipped, or any vertex other than v_0 or its neighbors is selected. Then $\rho(X_1^x, X_1^y) = 1$.

We now plan to work with $\mathbf{E}(\rho(X_t^x, X_t^y))$, the expectation of the differences between the Markov chains, to establish a bound on the mixing time, because ρ is quite similar to the total variation distance between the distributions of the chains. Using Markov's inequality later can turn this expectation into a bound on the probability that $X_t^x \neq X_t^y$, which from Corollary 4.10, is a bound on the mixing time. We start with the case above and generalize to larger time steps.

Note from the cases above that when $\rho(x,y) = 1$, $\mathbf{P}\{\rho(X_1^x, X_1^y) = 0\} = \frac{1}{n}$, $\mathbf{P}\{\rho(X_1^x, X_1^y) = 2\} \leq \frac{\Delta}{n}\frac{\lambda}{1+\lambda}$, and $\mathbf{P}\{(\rho(X_1^x, X_1^y) = 1\} =$ something. We can remove this lack of information by considering

$$\mathbf{E}(\rho(X_1^x, X_1^y) - 1) = \sum_{k=0}^{2}(k-1)\mathbf{P}\{\rho(X_1^x, X_1^y) = k\} \leq (-1)\frac{1}{n} + (1)\frac{\Delta}{n}\frac{\lambda}{1+\lambda}.$$

From the linearity of expectations, we can add a 1 to both sides,

$$\mathbf{E}(\rho(X_1^x, X_1^y)) \leq 1 - \frac{1}{n} + \frac{\Delta}{n}\frac{\lambda}{1+\lambda} = 1 - \frac{1+\lambda-\Delta\lambda}{n(1+\lambda)} = 1 - \frac{c_H(\lambda)}{n}.$$

If $\lambda < \frac{1}{\Delta-1}$, then $c_H(\lambda) > 0$, which means we can bound the expectation by the taylor series of e^{-x}:

$$\mathbf{E}(\rho(X_1^x, X_1^y)) \leq 1 - \frac{c_H(\lambda)}{n} \leq e^{-c_H(\lambda)/n}.$$

We now proceed to consider any two configurations x, y that satisfy $\rho(x,y) = r$. Since there exists a sequence of states $x = x_0, x_1, x_2, \ldots, x_{r-1}, x_r = y$ with the property that $\rho(x_{i-1}, x_i) = 1$, and that ρ is subadditive, we know that

$$\mathbf{E}(\rho(X_1^x, X_1^y)) \leq \sum_{k=1}^{r}\mathbf{E}(\rho(X_1^{x_{k-1}}, X_1^{x_k})) \leq re^{-c_H(\lambda)/n}.$$

We then consider the conditional expectation $\mathbf{E}(\rho(X_t^x, X_t^y) \mid X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1})$. We know that this expectation is equal to $\mathbf{E}(\rho(X_1^{x_{t-1}}, X_1^{y_{t-1}}))$, something that we can bound, and that for any two discrete random variables X, Y, $\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X)$. A fairly short proof of this can be found in [7] or [8], the Law of total expectation (we need only consider the finite case.) Thus,

$$\begin{aligned}\mathbf{E}(\rho(X_t^x, X_t^y) \mid X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1}) &= \mathbf{E}(\rho(X_1^{x_{t-1}}, X_1^{y_{t-1}})) \\ &\leq \rho(x_{t-1}, y_{t-1})e^{-c_H(\lambda)/n}.\end{aligned}$$

By taking the expectation of both sides, we obtain

$$\mathbf{E}(\rho(X_t^x, X_t^y)) \leq \mathbf{E}(\rho(X_1^{x_{t-1}}, X_1^{y_{t-1}}))e^{-c_H(\lambda)/n} = \mathbf{E}(\rho(X_{t-1}^x, X_{t-1}^y))e^{-c_H(\lambda)/n}.$$

We can iterate over this expression to obtain

$$\mathbf{E}(\rho(X_t^x, X_t^y)) \leq \rho(x,y)(e^{-c_H(\lambda)/n})^t \leq ne^{-tc_H(\lambda)/n}.$$

Since we know that $\rho(x,y) \geq 1$ when $x \neq y$, we use Markov's inequality to conclude

$$\mathbf{P}\{X_t^x \neq X_t^y\} = \mathbf{P}\{\rho(X_t^x, X_t^y) \geq 1\} \leq \mathbf{E}(\rho(X_t^x, X_t^y)) \leq ne^{-tc_H(\lambda)/n}.$$

From Corollary 4.10 and the above inequality, we know that $d(t) \leq \max_{x,y}(\mathbf{P}\{X_t^x \neq P(X_t^y)\}) \leq ne^{-tc_H(\lambda)/n}$. It is a matter of straightforward algebra to conclude that if

$$t > \frac{n}{c_H(\lambda)}[\log n - \log \epsilon],$$

then $d(t) < \epsilon$, which completes the proof. □

What knowing the mixing time tells us, in other words, is that if we start at some arbitrary starting configuration and run the Markov chain forward, we can bound the total variation distance from the stationary distribution in a guaranteed number of steps, and that the time this takes is roughly proportional to $n \log n$. But we still have the issue of initialization bias: depending on where we started, our sample from the stationary distribution π could be slightly different from π in a way we can predict. The next section gives us another way to sample from π, in a way that does not need us to prove any bound on the mixing time in order return a sample from π.

6. Coupling from the Past

We now reach what we want to discuss: Coupling from the Past (CFTP). In a nutshell, Markov chain studies allow us to sample from a potentially unknown stationary distribution π. The most common method of doing so is running the Markov chain forward: prove a bound on the mixing time, start the chain from a randomly generated distribution, and after a certain amount of time, we will know that the total variation distance from the stationary distribution is small enough to discount. To run the algorithm requires nothing more than a transition matrix P (in the case of Glauber dynamics, not even that,) and time: but ascertaining the minimum amount of time to avoid biased samples can be fairly difficult.

CFTP works in a different direction: we start from a time where the Markov chain has converged to π, go back for some amount of time $-T$, and run the same randomizing operations on each state until all the states have converged to the same single state. The intuition lies in the fact that if the chain was already at its stationary distribution in the present time, then the operations we performed must have led us to the stationary distribution.

What separates CFTP from running the chain forward is that its running time is random and self-determined, and that CFTP generates a perfect sample from π. CFTP determines its own running time by stopping if all the states have been mapped to the same state, or continuing going into the past. The generation of a perfect sample is exactly what it sounds like, but this property is more significant than it sounds: running the chain forward may not ever sample exactly from π, or at least in a reasonable amount of time. We first briefly explain the Ising model, CFTP, and then monotone CFTP.

6.1. The Ising Model and CFTP.

The Ising system is primarily a model for ferromagnetism. Picture a configuration of magnets on some sort of grid, lattice, or graph with the possible presence of an external field. Each magnet has a spin which we label spin up ($\uparrow$) or spin down ($\downarrow$). We call a **configuration** an arrangement of $\uparrow$ or $\downarrow$ magnets. Magnets close to each other want to have the same spin, and magnets want to be aligned with the external field. The energy of the magnet configuration (σ) is defined by

$$H(\sigma) = \sum_{i<j} \alpha_{i,j}\sigma(i)\sigma(j) + \sum_i B_i\sigma_i$$

where each i, j corresponds to a vertex, or magnet of the system, $\sigma(i)$ is 1 if vertex i is $\uparrow$, -1 if $\downarrow$, $\alpha_{i,j} > 0$ representing the strength of interaction, or the inverse of the distance between i and j, and B_i representing the strength of the external field as measured at site i.

The Ising model acts on a sample space Ω of all the possible configurations given some number of magnets. The probability the chain being in a given state σ is given by $\frac{e^{-\beta H(\sigma)}}{Z}$, where $\beta \geq 0$ is the inverse of temperature (as β increases, only low energy states are likely: if $\beta = 0$, all states are likely), and $Z = \sum_{\sigma\in\Omega} e^{-\beta H(\sigma)}$ is a normalizing constant designed to ensure the probabilities

of each state sum to 1. This Markov chain proceeds by the single-site heat bath algorithm, another name for Glauber dynamics.

(1) Pick a vertex i uniformly at random from the current state, and a uniformly random number $u \in [0, 1]$ to choose whether the vertex i should be changed to an up or down state. Denote this decision by the pair of numbers (i, u).
(2) Keeping all other vertices the same, change the vertex to an up spin ($\sigma_\uparrow$) or to a down spin ($\sigma_\downarrow$). Letting $Pr[\sigma_\uparrow], Pr[\sigma_\downarrow]$ denote the probability of i being changed to an $\uparrow$ and $\downarrow$ respectively, if $\frac{Pr[\sigma_\uparrow]}{Pr[\sigma_\uparrow]+Pr[\sigma_\downarrow]} > u$, change the state to $\sigma_\uparrow$: if not, change the state to $\sigma_\downarrow$.
(3) Repeat for as long as necessary, picking a new (i, u) for the state just generated.

This Markov chain is ergodic and thus converges to a unique stationary distribution $\pi(x) = \frac{e^{-\beta H(\sigma)}}{Z}$ by the convergence theorem. The chain is irreducible because for any given states σ, τ, there is some finite number of differences between them. We have a positive probability of picking a vertex v such that $\sigma(v) \neq \tau(v)$, and a positive probability of updating $\sigma(v)$ to be equal to $\tau(v)$, (as the u we pick in step (2) has a nonzero chance of being greater or less than the nonzero number u is compared to.) Since each vertex that differs between the two states has a positive chance of being selected and updated, the event of picking and updating all of those vertices consecutively has a positive probability. The chain's aperiodicity is proved with a similar argument.

We now turn to how CFTP functions on the chain. We've assumed that the randomizing process was running the whole time, and that we've stored all the randomizing operations (the site that was updated and the number u used to update the spin), of the heat bath algorithm up until the present. We want to determine the state of the Markov chain in the present to sample from π. So, consider a picture: in it, we have all the states of the Markov chain. Starting from some time $-T$ in the past, we apply T randomization operations to every state, and look at the present state of the chain (the chain from time 0.) If the randomization operations map every state to one state, we're done. If they do not, we go back a step in time to time $-T-1$, apply a new set of randomization operations to each state, and then apply the same T randomization operations we had before to the chain. It is not difficult to see that in this example, the first run of CFTP does not map every state to one state, while the second run maps every state to s_1, and that even if we extended the algorithm further back in time, the output would not change.

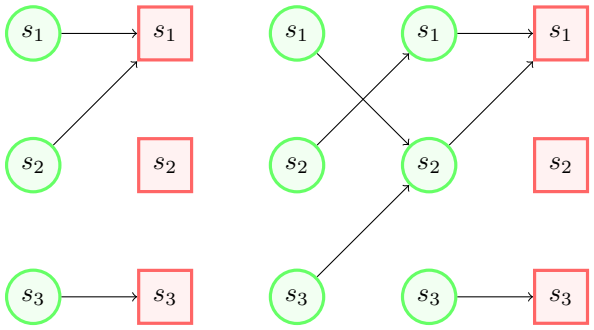


FIGURE 2. One run of CFTP, with s_i representing state i. In the first run, (time $T = -1$), s_3 was mapped to itself, while the other states were mapped to s_1. In the second run, (time $T = -2$), applied a new set of randomization operations (mapping s_1, s_3 to s_2, s_2 to s_1) and then applied the same operations from before.

This picture, in a nutshell, is CFTP. The main issue with this procedure is that there are many possible states of a Markov chain, which would take up an inordinate amount of memory and time. Monotone CFTP is a way to reduce the number of states one has to keep track of to 2.

Monotone CFTP refers to the existence of a partial ordering $\preceq$ between states that is preserved through the randomizing operation. This ordering can be used to obtain an upper and lower bound on all the states of the chain. For the Ising system, a natural ordering $\preceq$ between two states σ, τ is if $\sigma \preceq \tau$ then all the spin up states in σ are spin up in τ. Clearly, $\preceq$ is reflexive, transitive, and antisymmetric (if $\sigma \preceq \tau$ and $\tau \preceq \sigma$ then $\sigma = \tau$), but is not a total ordering on Ω because two states may both have some spin up vertex that is not spin up in the other. We observe two important properties of this order.

(1) There exists a maximum state $\hat{1}$ of all spin up states and a minimum state $\hat{0}$ of all spin down states wherein for any state σ, $\hat{0} \preceq \sigma \preceq \hat{1}$
(2) For any 2 states with the property that $\sigma \preceq \tau$, if the result of the randomization operation (i, u) is applied to both states, the updated states σ', τ' satisfy $\sigma' \preceq \tau'$ because $\frac{Pr[\sigma_\uparrow]}{Pr[\sigma_\uparrow]+Pr[\sigma_\downarrow]} \leq \frac{Pr[\tau_\uparrow]}{Pr[\tau_\uparrow]+Pr[\tau_\downarrow])}$, or that each spin up site in σ is spin up in r and each $\alpha_{i,j}$ is nonnegative. One can see this by noting $\frac{Pr[\sigma_\uparrow]}{Pr[\sigma_\downarrow]} \leq \frac{Pr[\tau_\uparrow]}{Pr[\tau_\downarrow]}$ for the reasons above.

Once we know these properties, the convergence of the chain is evident, because for all states σ, $\hat{0} \preceq \sigma \preceq \hat{1}$, we can apply a sequence of randomization operations on both $\hat{0}$ and $\hat{1}$, and get an upper and lower bound on the probability of every configuration for times $-T$ to 0. Since the chain is ergodic, we know that the chain starting at $\hat{1}$ has a positive chance of reaching $\hat{0}$ for some large enough $-T$, and thus that the lower bounds and upper bounds will meet in one state. With this example, we provide an algorithm to apply monotone CFTP to an arbitrary Markov chain a partial ordering that preserves states in the way described.

$T \leftarrow 1$
repeat
    upper $\leftarrow \hat{1}$, lower $\leftarrow \hat{0}$
    for $t \leftarrow -T$ to -1
        upper $\leftarrow \varphi(\text{upper}, U_t)$
        lower $\leftarrow \varphi(\text{lower}, U_t)$
    $T \leftarrow 2T$
until (upper $\leftarrow$ lower)
return upper

FIGURE 3. Monotone CFTP: U_t represents the random variable u we used in our randomization operations, and φ is a deterministic procedure that takes a state and a random number/variable (our u) and returns an updated state.

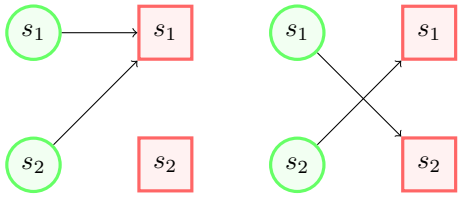
6.2. **Design choices of CFTP.** Now, armed with an algorithm for monotone CFTP, we need to address two important design decisions of the algorithm by analyzing the following two *incorrect* variants of CFTP: not reusing the same U_t and coupling chains into the future. We will run these

algorithms on the Markov chain with the transition matrix $\begin{pmatrix} .5 & .5 \\ 1 & 0 \end{pmatrix}$. The first row of the matrix corresponds to state 1, the second to state 2, which we denote s_1, s_2 respectively. The stationary distribution π for this chain is $(\frac{2}{3}, \frac{1}{3})$. After this analysis, which can be found in [4] we briefly touch on something we took for granted, the update function.

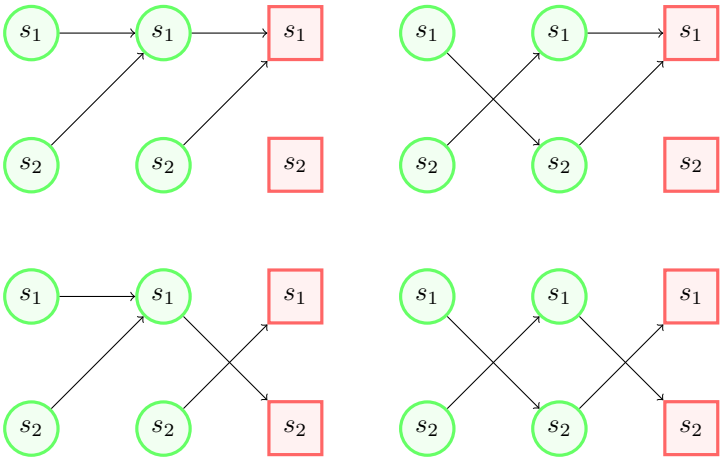
(1) If we do not apply the same randomizing operations each time, the resulting CFTP algorithm will not work properly. Since we have assumed that we have reached a stationary distribution and are running a process from the past, we need to save the results of each U_t. If we do not reuse our U_t, and generate new ones for each trial, then as we go further back in time, our result would change with every trial. We can see this by examining the chain above. Let the random variable M be the largest time m for which this faulty CFTP algorithm decides to run the chains, starting from time $-m$. Let Y be the output of the chain. Note that from the definition of conditional probability,

$$\begin{aligned}\mathbf{P}\{Y = s_1\} &= \sum_{m=1} \mathbf{P}\{M = m, Y = s_1\} \\ &\geq \mathbf{P}\{M = 1, Y = s_1\} + \mathbf{P}\{M = 2, Y = s_1\} \\ &= \mathbf{P}\{M = 1\}\mathbf{P}\{Y = s_1 \mid M = 1\} + \mathbf{P}\{M = 2\}\mathbf{P}\{Y = s_1 \mid M = 2\}.\end{aligned}$$

The two equally likely possibilities when $M = 1$ are



From the picture, $\mathbf{P}\{M = 1\} = 0.5$, and the probability that $Y = s_1$ given $M = 1$ is 1. Thus, with probability of $0.5 = \mathbf{P}\{M \neq 1\}$, we decrement the starting time to -2 and run the chains to time 0. The four possible outcomes of the incorrect algorithm (which, as a reminder, does not fix the U_t for each iteration of the loop, which means all these outcomes are valid,) are listed below:



All but the bottom right image correspond to a run of the incorrect CFTP algorithm with $M = 2$. Thus, the probability that $M = 2$ is $\mathbf{P}\{M \neq 1\} \times \mathbf{P}\{\text{chains couple}\} = \frac{1}{2} \times \frac{3}{4} = \frac{3}{8}$, and the probability that $Y = s_1$ given $M = 2$ is $\frac{2}{3}$. These conclusions imply that

$$\frac{2}{3} = P(Y = s_1) \geq \frac{1}{2} \times 1 + \frac{3}{8} \times \frac{2}{3} = \frac{3}{4}$$

where the $\frac{2}{3}$ follows from our knowledge that $\mathbf{P}\{Y = s_1\}$ must take on the stationary distribution if our faulty algorithm is correct (which it is not.) Obviously, this is a contradiction.

(2) If we couple into the future, i.e., start two Markov chains from time 0 and run the chains into future, the unique state that CFTP converges to would change for each step forward we take in time: given that the intuition here is that we were at a stationary distribution and looked at the moves before, it is clear that this variant of CFTP fails. If we couple the Markov chain we see that coupling into the future always returns the first state with a probability of 1. Suppose the "coupling into the future" algorithm converged to a single state at time t. Then at time $t - 1$, the chains starting from s_1, s_2 must have been on different states. However, the only state the chains could meet in one time step is at s_1.

Remark. We also need to have a valid update function for the specific Markov chain: in our case, we used the update functions for the Ising model and Hardcore model (the single-site heat bath), but choosing the wrong update function could lead to CFTP never ending. See [4] for an example. For most of the proofs past this point, we assume the update function is valid.

6.3. **Monotone CFTP Formalized and Bounded.** Now, given a fair amount of intuition for why the algorithm works the way it does, we give the general procedure one follows for CFTP before proving some general results about monotone CFTP. We need three things for CFTP: a way to generate random maps from the state space Ω back to Ω, a way of composing the maps, a way to determine if total coalescence has occurred, meaning that all the states have converged to one, which in the Ising example, means that the upper bound (the chain starting from $\hat{1}$) equals the lower bound (the chain starting from $\hat{0}$.)

Generating the random maps is known as the random map procedure: we model it as an oracle that returns independently and identically distributed functions $f_{-1}, f_{-2}, \ldots, f_{-T} : \Omega \mapsto \Omega$ governed by some probability distribution P that for our purposes is associated with the transition matrix P of the Markov chain we wish to model. We can represent these maps explicitly: if we have a function $Markov()$ that maps a state σ to another state τ, where the probability of observing $Markov(\sigma) = \tau$ is $P(\sigma, \tau)$, and where repeated calls to $Markov()$ are independent, then we can define $f_{-t}(\sigma)$ by $f_{-t+1}(Markov(\sigma))$ for each state σ.

Most importantly, for our algorithm to end, some time $-T$ must have the property that the composite map (however one implements it) has the property that the composite map from $-T$ to 0 is collapsing, or that

$$F_{-T}^{0} := f_{-1} \circ f_{-2} \circ f_{-3} \circ \cdots \circ f_{-T}$$

is the same state for all $x \in \Omega$. This T is determined by the algorithm itself, given that we have a way to test for coalescence: with monotone CFTP, this is fairly simple, but with applications that are not monotone, several optimizations and workarounds need to be used to accomplish this. In

pseudocode, our general CFTP algorithm looks like

$T \leftarrow 1$
while $f_{-1} \circ f_{-2} \circ f_{-3} \circ \cdots \circ f_{-T}$ has not coalesced
    increase T (doubling, incrementing, etc...)
return $F^0_{-T}(x)$ for any arbitrary $x \in \Omega$.

As long as the probability distribution P governing the random maps has some relation to π, CFTP will return a state distributed from π. Formally, this means that if a random state x is chosen in accordance with π and a random map f is chosen in accordance with P, then the state $f(x)$ will be distributed in accordance with π: i.e., the random maps preserve π. In addition, however one increases $-T$, it has no effect on the distribution CFTP outputs. It turns out that Propp and Wilson proved that doubling T every iteration is close to optimal in [1].

We now go through a few proofs to prove the correctness of CFTP. The first theorem establishes that CFTP terminates with probability 1 and returns a sample from the stationary distribution π. The second proof establishes the property that the random maps must have in order for this to occur and cements the representation of the random map procedure as a deterministic function as seen above. The third proof bounds the coupling time of monotone CFTP in terms of the mixing time of the chain, which implies that if a Markov chain is rapidly mixing, then monotone CFTP is rapidly mixing. Although we do not give explicit examples of the usage of the theorem (proving the mixing time of the Ising model would be fairly involved and would require the introduction of more vocabulary, see [6]) knowing the relationship between CFTP and the mixing time is, if nothing else, nice to know.

For the proofs below, which follow [1], we assume that the chain is monotone, aperiodic, and irreducible:

Definition 6.1. A Markov chain is **monotone** if there exists a partial ordering $\preceq$ and an update function φ that has the property that if the states $\sigma \preceq \tau$, then $\varphi(\sigma, U_0) \preceq \varphi(\tau, U_0)$.

Theorem 6.2. *For a monotone Markov chain, with probability 1, CFTP returns a value which is distributed according to the stationary distribution π of the chain.*

Proof. Since the chain is ergodic, we intuitively know that it will converge to a unique stationary distribution π, and that there exists a time r such that for any states $\sigma, \tau \in \Omega$, there is a positive probability of travelling from σ to τ in r steps. Then we also know that for every time t there is a positive chance that the composite map $f_{-t+1} \circ f_{-t+2} \circ f_{-t+3} \circ \cdots \circ f_{-t-r} = F^t_{t-r}$ is a constant, i.e., that there is a chance that every state gets mapped to one state via the random map representation given by $Markov()$. Since each of the maps $F^0_{-r}, F^{-r}_{-2r}, \ldots$ has a positive chance of being constant, and since each of these maps are independent from the other by $Markov()$'s independence, it is guaranteed that one such map is constant. But by construction, this means that for all sufficiently large times M, F^0_{-M} is constant.

To prove that this returned value is distributed according to π, note that if we run CFTP, it outputs a value we call $\bar{F}^0_{-\infty}$. This value is the probability distribution of the composition of all of the random maps. However, we also know that $\bar{F}^0_{-\infty}$ is just $Markov()$ applied once to $\bar{F}^{-1}_{-\infty}$, and that $\bar{F}^0_{-\infty}$ and $\bar{F}^{-1}_{-\infty}$ must have the same probability distribution (simply apply a coupling argument separated by a timestep of 1.) The only possible candidate is the stationary distribution π, which we know is unique. □

The point of the above proof, in addition to establishing the correctness of the algorithm, serves a dual purpose in thinking of CFTP in another way, in terms of the nondeterministic function $Markov()$. However, we do not exactly need the assumptions of independence we had in the proof. One surely does not need independence to prove coalescence when a different type of coupling could be used. In addition, when updating the coupled chain from time t to $t+1$, the only thing that matters is that the decision does not depend on the state of the chain, and that the randomness used in making that decision is independent from one decision to the next. Thus, as we saw in the Ising model, we can use a deterministic function $\phi(\sigma, U_t)$ that takes in a state and a random variable to represent $f_t(\sigma)$. As long as $\{\dots, U_{-2}, U_{-1}\}$ are independent and identically distributed, the theorem above holds.

Now, we transition to the exact property ϕ needs to preserve the stationary distribution π. The reader may notice that this property resembles the stationarity criterion for individual distributions. We note that this does not just apply to ϕ, but to any update function with this property, which means we can have multiple updating rules for each space, and simply cycle between the update functions provided they all preserve π.

Theorem 6.3. *Let π be a stationary distribution on an arbitrary Markov chain. Let $\dots, U_{-2}, U_{-1}$ be independently and identically distributed random variables, and let $\phi_t(\cdot,\cdot)$ be a sequence of deterministic functions such that for all times t and states $x \in \Omega$,*

$$\sum_{y\in\Omega} \pi(y)\boldsymbol{P}\{\phi_t(y, U_t) = x\} = \pi(x)$$

Now define $f_t(x) = \phi_t(x, U_t)$ and $F_t^0 = f_{-1} \circ f_{-2} \circ \dots f_{-t}$. Assume that with probability 1, there exists a time t for which the map F_t^0 is constant, with a constant value that we may denote by $\phi(\dots, U_{-2}, U_{-1})$. Then the random variable $\phi(\dots, U_{-2}, U_{-1})$ which is defined with probability 1 by assumption, has distribution equal to the stationary distribution π.

Proof. Let X be a random variable on the state space Ω on the Markov chain, and let Y_t be the random variable $F_t^0(X)$. Each Y_t has the distribution π, as each ϕ_t preserves π. Thus $\phi(\dots, U_{-2}, U_{-1}) = Y_m$ for some integer m, which means it has the distribution π. □

Knowing this, we can proceed to bound the running time of a monotone CFTP chain by means of its mixing time. We define T_* as the smallest time t that monotone CFTP takes to coalesce, or when $F_{-t}^0(\bar{0}) = F_{-t}^0(\bar{1})$, and T^* as the smallest time t such that $F_0^t(\bar{0}) = F_0^t(\bar{1})$. We accomplish this by bounding the probability $P(T_* > t)$ of the convalescence time being greater than an arbitrary t, proving said probability is submultiplicative, and then bounding $\mathbf{E}(T_*)$ by means of Markov's inequality. We use T^* because it is simpler, even if the running time is linear in T_*. See [1].

Theorem 6.4. *Let l be the length of the longest chain (a totally ordered subset) in the partially ordered state space Ω. Then*

$$\frac{\boldsymbol{P}\{T^* > t\}}{l} \leq \bar{d}(t) \leq \boldsymbol{P}\{T^* > t\}$$

where T^* is defined $\min\{t | F_0^t(\bar{0}) = F_0^t(\bar{1})\}$, as above.

Proof. Note that $\mathbf{P}\{T^* > t\} = \mathbf{P}\{T_* > t\}$. Now suppose σ is an element of Ω, and let $h(\sigma)$ be the length of the longest chain whose top element is σ. Then let the random variables X_0^t, X_1^t denote the states of the Markov chain after t steps when starting from $\hat{0}$ and $\hat{1}$, respectively. Let ρ_0, ρ_1, denote the distributions on the state space Ω that assign probability 1 to states $\hat{0}, \hat{1}$, respectively, and let ρ_0^t, ρ_1^t denote the probability distribution of the Markov chain started from ρ_0, ρ_1 at time

t. It follows that if $X_0^t \neq X_1^t$, then $h(X_0^t) + 1 \leq h(X_1^t)$, because the state of X_1^t is greater than the state of X_0^t by our partial ordering, so one can take the chain for X_0^t and put X_1^t on top of it. Then

$$\begin{aligned}
\mathbf{P}\{T^* > t\} &= \mathbf{P}\{X_0^t \neq X_1^t\} \\
&= \mathbf{P}\{h(X_1^t) - h(X_0^t) \geq 1\} \\
&\leq \mathbf{E}(h(X_1^t) - h(X_0^t)) \\
&= |\mathbf{E}(h(X_1^t)) - \mathbf{E}(h(X_0^t))| \\
&= |\mathbf{E}_{\rho_1^t}(h(X)) - \mathbf{E}_{\rho_0^t}(h(X))| \\
&= |\sum_{i=0}^{l} i\Big[\mathbf{P}_{\rho_1^t}\{h(X) = i\} - \mathbf{P}_{\rho_0^t}\{h(X) = i\}\Big]|,
\end{aligned}$$

where the first inequality follows from Markov's inequality. Now note that the maximum of any sum is when all the terms are the same sign. In the last line, this is determined by the quantity in the large square brackets. In addition, $l \geq i$, so in the sum of terms that are the same sign, if we multiply each probability term by l, the sum can only increase. However, the sum is a difference between the probabilities of ρ_1^t and ρ_0^t, so by the definition of total variation distance (Definition 4.1),

$$\begin{aligned}
&= |\sum_{i=0}^{l} i\Big[\mathbf{P}_{\rho_1^t}\{h(X) = i\} - \mathbf{P}_{\rho_0^t}\{h(X) = i\}\Big]| \\
&\leq \|\rho_1^t - \rho_0^t\|_{TV} \cdot l \\
&\leq \bar{d}(t)l.
\end{aligned}$$

The last inequality follows from the definition of $\bar{d}(t)$ (Definition 4.6.)

The second inequality of the theorem follows by noting that all the states are bounded by the states $\hat{1}, \hat{0}$. By monotonicity, the probability a coupling starting from any two distributions has coalesced is at least $\mathbf{P}\{T_* \leq t\}$. Thus, for any two starting distributions μ, υ, the total variation distance between μP^t and υP^t is less than or equal to the probability $\mathbf{P}\{T_* > t\}$ that the coupling starting from the largest and smallest state did not converge by time t. □

Now, we have a way to relate the distance from stationarity to the probability of not coupling, but to go further, we prove the submultiplicativity of $\mathbf{P}\{T^* > t\}$ and then bound the expected value of T^* by the mixing time.

Theorem 6.5. *Let T_1, T_2 be nonnegative integer random variables, which can be constant. Then*

$$\boldsymbol{P}\{T^* > T_1 + T_2\} \leq \boldsymbol{P}\{T^* > T_1\}\boldsymbol{P}\{T^* > T_2\}$$

Proof. The event that $F_0^{T_1}$ is constant and the event that $F_{T_1}^{T_1+T_2}$ is constant are independent. If either event is true, then $F_0^{T_1+T_2}$ is constant. Thus,

$$\mathbf{P}\{F_0^{T_1} \text{ is constant} \bigcup F_{T_1}^{T_1+T_2} \text{ is constant}\} \leq \mathbf{P}\{T^* \leq T_1 + T_2\}.$$

By complementary probabilities and the independence of complements (Proposition 2.9),

$$\begin{aligned}
\mathbf{P}\{T^* > T_1 + T_2\} &\leq \mathbf{P}\{F_0^{T_1} \text{ is not constant}, F_{T_1}^{T_1+T_2} \text{ is not constant}\} \\
&= \mathbf{P}\{T^* > T_1\}\mathbf{P}\{F_{T_1}^{T_1+T_2} \text{ is not constant}\}.
\end{aligned}$$

The conclusion follows from noting that $\mathbf{P}\{F_{T_1}^{T_1+T_2} \text{ is not constant}\} \leq \mathbf{P}\{F_0^{T_2} \text{ is not constant}\}$, because if $F_0^{T_2}$ is constant, then $F_{T_1}^{T_1+T_2}$ is. □

The last step is to estimate the tail probabilities for T^* in terms of the expected value of T^*.

Lemma 6.6. $t\boldsymbol{P}\{T^* > t\} \leq \boldsymbol{E}(T^*) \leq \frac{t}{\boldsymbol{P}\{T^* \leq t\}}$.

Proof. The first inequality follows from Markov's inequality. The second inequality follows from noting that if we let $\epsilon = \mathbf{P}\{T^* > t\}$, then by submultiplicativity, $\mathbf{P}\{T^* > kt\} \leq \epsilon^k$. Thus,

$$\mathbf{E}(T^*) = \sum_{i=1} it\mathbf{P}\{T^* > it\} \leq t\sum_{i=1} i\epsilon^i = \frac{t}{1-\epsilon} = \frac{t}{\mathbf{P}\{T^* \leq t\}}$$

□

Now, we can bound the expected run time by the mixing time. As a reminder, for shorthand, we define T_{mix} as the smallest time t when $\bar{d}(t) \leq \frac{1}{e}$. Let l be the length of the longest chain. Now define $k := T_{mix}(1 + \ln(l))$. From the submultiplicativity of $\bar{d}(t)$, which is derived in [6],

$$\bar{d}(k) = \bar{d}(T_{mix}(1 + \ln(l))) \leq \bar{d}(T_{mix})\bar{d}(T_{mix}\ln(l)) \leq \frac{1}{e}\frac{1}{e}^{\ln(l)} = \frac{1}{el}.$$

So, by Theorem 6.5,

$$P(T^* > k) \leq \frac{1}{el}l = \frac{1}{e},$$

and Lemma 6.6,

$$\mathbf{E}(T^*) \leq \frac{k}{1-\frac{1}{e}} \leq 2k \leq 2T_{mix}(1 + \ln(l)).$$

Thus, we know that if a Markov chain is rapidly mixing (the mixing grows in polynomial time in log(number of states in chain)), it is rapidly coupling. The only difference is that we have no control over how long each run CFTP takes.

6.4. **Bounding the Mixing Time of a Random Walk.** As a slight interlude, we note that one can bound the mixing time on a random walk by taking some samples of the coupling time T^*. Imagine we have 10 independent samples of the coupling time random variable $T_1, \ldots, T_{10}$, and obtain $10T_{est} = T_1 + \cdots + T_{10} \leq 1000$. Then suppose we run a Markov chain for $10T_{est}$ time, treating T_{est} as a random variable; we claim that we can bound $d(t)$ by 2^{-10}. Since $\mathbf{P}\{T^* > T_i\} \leq \frac{1}{2}$ because for any two independently and identically distributed random variables,

$$2\mathbf{P}\{T^* > T_i\} = \mathbf{P}\{T^* > T_i\} + \mathbf{P}\{T^* < T_i\} \leq 1,$$

which follows from symmetry. By Theorem 6.5 and Proposition 4.9,

$$d(10T_{est}) \leq \bar{d}(10T_{est}) \leq \mathbf{P}\{T^* > T_1 + \cdots + T_{10}\} \leq 2^{-10}$$

In [1], where the above proof comes from, Propp and Wilson give another method of bounding the mixing time of the random walk.

6.5. **The hardcore model: why we can use CFTP even though it's not monotone.** It turns out that there is an interesting way to use CFTP on the hardcore model with fugacity $\lambda > 0$, even though there is no monotonicity in general, except in the case of bipartite graphs. Instead of associating a 1 or 0 to each vertex v in a state, we associate a 1, 0, or a '?' to a vertex v in a set of states. A '1' means that every state has a particle at v, a '0' means that every state is vacant at v, and a '?' means that it is possible that some states have a particle at v, and some other states do not. We then start from the all '?' state, which we denote by x, and apply the single-site bath algorithm to x until we reach a state τ with no ?'s: it is evident that applying the same randomization procedures to any other state σ would map that state to τ, because x represents the set of all possible states. The single-site heat bath algorithm works as follows: pick a vertex v uniformly at random. Flip a coin with the probability of heads being $\frac{\lambda}{1+\lambda}$. We have a few cases:

(1) The coin is tails, and any particle is removed from v. Thus, for all states σ, $\sigma(v) = 0$.
(2) The coin is heads, and there is a chance of v being occupied in some state σ. In other words, a neighbor w of v satisfies $x(w) =?$. Then place a '?' at v, or let $x(v) =?$.
(3) The coin is heads, and all the neighbors w of v are vacant in x. Then let $x(v) = 1$.

Running this chain tells us when all of the states have coalesced, but bounding the running time of this chain is not the same as bounding the mixing time of the hardcore model. This issue arises from the fact that the hardcore model is **anti-monotone**: there is a partial ordering $\preceq$ where if $\sigma \preceq \tau$ and a randomization operation $\varphi(v, U_t)$ is applied, the expected states satisfy $\varphi(\tau, U_t) \preceq \varphi(\sigma, U_t)$, but no such monotone partial ordering.

However, one can bound the run time of CFTP for the hardcore model by proving that the number of '?'s in the above chain that starts at state x shrinks to 0 exponentially fast provided $\lambda < \frac{1}{\Delta}$. (This is fairly nontrivial, see [6] for details.) As compared to our theorem that the hardcore model was rapidly mixing when $\lambda < \frac{1}{\Delta - 1}$, it turns out that shrinking the number of '?' to 0 is a stronger condition than our theorem involving rapid mixing. We are unaware if for Markov chains that exhibit this exponential decay of '?'s for some bound on λ, the condition on λ implies the Markov chain is rapidly mixing. For a deeper look at anti-monotonicity, see [5]

6.6. **CFTP to sample from an Unknown Markov Chain.** The final step for this paper is to briefly describe CFTP for an unknown Markov chain, because the algorithm itself is fairly interesting. By an "unknown" Markov chain, we mean that the transition probabilities of the Markov chain are hidden, but that we have a black box that simulates transitions from state to state for us. For our purposes, we can imagine this black box as the $Markov()$ function from section 6.3. We imagine the Markov chain is a biased random walk on some graph G whose edges are labelled with weights, and whose vertex set we denote by Ω. (In other words, moving from vertex to vertex is like moving from state to state.) G has the property that the probability of moving from vertex u to v is proportional to the weight of the edge connecting u and v, relative to the sum of the weights of the edges connecting u to its neighbors. We denote the stationary distribution of the chain by π.

In other words, we are missing the first requirement of CFTP: a random map procedure that maps states σ to some other state τ, does not suffer from initialization bias, and works in a reasonable amount of time. We first address the problem of initialization bias before outlining the algorithm. Put simply, if our goal is trying to sample from the stationary distribution, 'incorrect' initialization conditions result in sampling from a probability distribution that is different from π. We call these incorrect initialization conditions **initialization bias.** For example, if one runs the hardcore model forward from a random state, but for a time much less than the mixing time, one would be

sampling from a distribution that is decidedly not the stationary distribution. In another instance, for a chain with a partial ordering $\preceq$, starting your chain at the state $\hat{1}$ would be a pretty bad idea. For conventional sampling algorithms (running the chain forward,) we try to avoid initialization bias mainly by running the chain for a long time and starting the chain from a distribution representative of π (though other more sophisticated methods exist.) I do not have a good reference for avoiding initialization bias, though [2] has some commentary on it.

Put bluntly, avoiding other concerns, simply starting at some vertex v, running the chain (moving from vertex to vertex) for some large number T steps, and mapping all the states that were visited on the walk to the state σ the chain arrived at after T steps is not sufficient: σ is subject to initialization bias.

It turns out that the procedure to generate unbiased random maps involves multiple steps:

(1) Initialization: Start the Markov chain at some arbitrary state. Run the random walk until we reach state 1 (an arbitrary state the experimenter can choose, or not.) Then, from that point, record how long it takes the chain to visit all of the states (vertices in our case), letting T denote the time it takes to do so. Then wait until the Markov chain visits state 1 again.
(2) Map Construction: Set an alarm clock that goes off between 1 and $4T$ steps uniformly at random. Start visiting states σ: when the alarm goes off, map all of the visited states to the current state σ.

This $RandomMap()$ procedure satisfies two vital properties.

(1) The distribution of each T we set in the Initialization step of $RandomMap()$ should satisfy the property that the expected time in which the walk terminates is not dependent on the current state or how one got there. Since this time T was set randomly by the Initialization step of $RandomMap()$, we are okay on that front. More formally, $RandomMap()$ should preserve π, or in other words, satisfy the property of the ϕ_t mentioned in Theorem 6.3 [2].
(2) The time it takes to walk is not so short as to not map many states, but not so long as to never end, or in other words, annoy some experimenters into aborting the algorithm, which would, for this specific algorithm, introduce bias. Propp and Wilson's algorithm [2] (outlined above) accomplishes construction of a $RandomMap()$ function in around 15 times the **cover time**, the amount of time a random walk takes to visit every vertex.

Our pseudocode for CFTP for this type of problem is the same as the our general algorithm before:

$t \leftarrow 0$
$F^0_{-T} \leftarrow$ identity map
while F^0_{-T} is not collapsing
    $t \leftarrow t - 1$
    $f_t \leftarrow RandomMap()$
    $F^0_{-T} \leftarrow F^0_{-T} \circ f_t$
return value that F^0_{-T} collapses Ω to

where $RandomMap()$ is defined as above.

7. Further Steps in CFTP

I would be fairly remiss in saying I covered everything: I did not give a formal definition of antimonotonicity, even though it is very related to monotonicity, and is a key step in CFTP for the hardcore model (section 6.5.) However, the papers I cited are for the most part readable with an understanding of the concepts above (I did, I just did not want to write down more proofs), and the concepts are easy enough to understand, though the nitty-gritty can be painful for certain models (the Ising.) There are also other schemes of CFTP that I did not formally mention, like Fill's Algorithm, which can be interrupted without introducing bias. See [6] for some more details and a more thorough treatment of Markov chains.

Acknowledgments. It is a pleasure to thank my mentor, Jonathan Dewitt, for introducing me to this topic, helping me with some tricky proofs, and giving feedback on my early drafts. In addition, I would also like to thank Peter May for proofreading this paper, and especially for organizing the REU and this opportunity to research a topic. Lastly, thanks to Professor Babai for leading the Apprentice Program.

References

[1] Wilson, D. B., and Propp, J. G. (1996). Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. https://www.stat.berkeley.edu/~aldous/206-RWG/RWGpapers/propp_wilson.pdf

[2] Wilson, D. B., and Propp, J. G. (1998). How to Get a Perfectly Random Sample from a Generic Markov Chain and Generate a Random Spanning Tree of a Directed Graph [Scholarly project]. https://www2.stat.duke.edu/~scs/Projects/Trees/Theory/ProppWilson1998.pdf

[3] Haggstrom, O., and Nelander, K. (1997). Exact Sampling from anti-monotone systems. http://uosis.mif.vu.lt/~stepanauskas/MK/Haggstrom200.%20Finite%20Markov%20chains%20and%20algorithmic%20applications%20%28CUP,%202002%29%28125s%29.pdf

[4] Haggstrom, O. (2008). Finite Markov chains and algorithmic applications. Cambridge: Cambridge University Press.Levin, D. A., Peres, Y., Wilmer, E. L., Propp, J., and Wilson, D. B. (2017). Markov chains and mixing times. Providence, RI: American Mathematical Society. http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf

[5] Häggström, O., and Karin Nelander. (1999). On Exact Simulation of Markov Random Fields Using Coupling from the Past. Scandinavian Journal of Statistics, 26(3), 395-411. http://www.jstor.org/stable/4616564

[6] Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). Markov Chains and Mixing Times. S.l.: American Mathematical Society.

[7] Rosenthal, J. S. (2013). A First Look at Rigorous Probability Theory (2nd ed.). New Jersey, NJ: World Scientific.

[8] Law of total expectation. (2017, August 12). Retrieved August 24, 2017, from https://en.wikipedia.org/wiki/Law_of_total_expectation

[9] The Coin Flip: A Fundamentally Unfair Proposition? (2009, March 29). Retrieved August 13, 2017, from http://econ.ucsb.edu/~doug/240a/Coin%20Flip.htm