Lab Research Project

Analysis of the Propp-Wilson algorithm and its application to the Ising model

Milan Stanković

Supervised by Katharina Eichinger

Center for Applied Mathematics - École Polytechnique

June, 2024

Abstract

Designing an irreducible, aperiodic Markov chain with a given stationary distribution over a given set, and running the chain until its state follows 'approximately' the stationary distribution, is at the core of Markov chain Monte Carlo algorithms. The main challenge is that, in general, we do not know for how many steps to run the Markov chain to obtain samples that resemble the stationary distribution well enough. In 1996, James Gary Propp and David Bruce Wilson designed an algorithm [1] that tackles this challenge, by providing exact samples from the stationary distribution and determining automatically when to stop. In this paper, we give a detailed analysis of the algorithm, prove that it terminates with probability 1 if and only if certain conditions are met - otherwise it terminates with probability 0, and we present a classical application to the Ising model.

1 Introduction

Many scientific endeavours involve sampling elements from a given set according to some probability distribution. The goal might be to study the elements of the set, the set as a whole, the given distribution, or some combination of the three. Obtaining samples is not an issue when the set is of reasonable size and when we can enumerate all of its elements. However, in many cases, the set is too large or its size is unknown, making it practically impossible to enumerate all its elements. In these cases, in order to sample from this set, we can resort to Markov chain Monte Carlo (MCMC) algorithms, which simulate the evolution of an irreducible, aperiodic Markov chain defined on this set, with stationary distribution equal to the distribution we want to study. The distribution of the state of this Markov chain will approach the stationary distribution when we run the chain for long enough. The main challenge is that, in general, we do not know for how many steps to run the Markov chain to obtain satisfactory samples that resemble the stationary distribution well enough.

One way to overcome this challenge is to use an algorithm developed by James Gary Propp and David Bruce Wilson in 1996 [1], which builds on the notions of MCMC algorithms. We call this algorithm by the names of its inventors: the Propp-Wilson algorithm (although, it is also known as the 'Coupling from the past' algorithm). By trying earlier starting times at every iteration, it provides exact samples from the stationary distribution and automatically determines when to stop. Our goal is to analyze the Propp-Wilson algorithm in details, providing rigorous proofs of its main properties. In order to quantitatively define and represent many of the ideas that we have mentioned so far, we start by introducing Markov chains and their basic properties. We then proceed to describe some basic methods for simulating evolutions of Markov chains, which

lay the foundation for implementing Markov chain Monte Carlo algorithms, and the Propp-Wilson algorithm in particular. Perhaps the most important concept to remember from these sections is the notion of 'grand couplings' evolving according to some update function and some random variables. One of the main results of the paper, *the 0-1 law*, is based on these grand couplings and it allows us to prove that the Propp-Wilson algorithm terminates with probability equal to either 0 or 1 (hence the 0-1 law), depending on whether certain conditions are satisfied or not. Only then do we properly introduce the Propp-Wilson algorithm. We give rigorous proofs about the distribution of the output of this algorithm and the bounds on its running time. Finally, we apply the algorithm to the famous model from statistical physics, the Ising model.

2 Markov chains

Markov chains are used to model a large class of processes that are of major importance in various fields, like physics, economics, biology, computer science etc. The common property of all these processes is that the probability distribution of the next state depends only on the current state of the process. In this section, we present without proof some fundamental results about Markov chains from books [2] and [3] that we use in the latter sections.

Definition 2.1. *Let χ be a finite or a countable set. Let $(X_t)_{t\geq 0}$ be a sequence of χ-valued random variables. Then, $(X_t)_{t\geq 0}$ is a homogeneous Markov chain if, for all $t \geq 1$ and for all $(s_0, \ldots, s_t) \in \chi^{t+1}$:*

$$\mathbb{P}(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = \mathbb{P}(X_t = s_t \mid X_{t-1} = s_{t-1}).$$

We call χ the state space of the Markov chain $(X_t)_{t\geq 0}$. Conversely, we call $(X_t)_{t\geq 0}$ a Markov chain over χ.

Remark 2.2. The choice of 0 as the smallest index for the Markov chain is completely arbitrary. We can choose any integer instead.

Remark 2.3. Throughout this paper, we will work only with homogeneous Markov chains, so from now on, whenever we mention a Markov chain, we assume it is homogeneous.

We will focus on the case where the state space χ is finite, although some of the notions we cover can be extended to the infinite case. Finite state space is convenient because we can represent the probabilities of transitioning from one state to another (also called *the transition probabilities*) in a matrix:

Definition 2.4. *Let $N \in \mathbb{N}_{>0}$ and let $\chi = \{s_1, \ldots, s_N\}$ be a finite set. Let $(X_t)_{t\geq 0}$ be a Markov chain over χ. Then, the transition matrix of the chain $(X_t)_{t\geq 0}$ is a $N \times N$ matrix P, whose entries are:*

$$P_{i,j} = \mathbb{P}(X_{t+1} = s_j \mid X_t = s_i)$$

for any $i, j \in \{1, \ldots, N\}$.

Remark 2.5. By Definition 2.1, $\mathbb{P}(X_{t+1} = s_j \mid X_t = s_i)$ does not depend on t, so the Definition 2.4 is valid.

The following theorem [2, Theorem 2.1.] illustrates why the transition matrix can be so useful.

Theorem 2.6. *Let $(X_t)_{t\geq 0}$ be a Markov chain over state space χ with transition matrix P. For all $t \geq 0$, define the row vector $\mu^t = (\mu_1^t, \ldots, \mu_N^t)$ to be the probability distribution of X_t, i.e. for all $i \in \{1, \ldots, N\}$, $\mathbb{P}(X_t = s_i) = \mu_i^t$. Then, for all $t \geq 0$, we have*

$$\mu^t = \mu^0 P^t.$$

Then we have the following corollary [2, Problem 2.5.], which is proved by a simple induction on m.

Corollary 2.7. *Let $(X_t)_{t\geq 0}$ be a Markov chain over state space χ with transition matrix P. For all $m, n \in \mathbb{N}$ and all $i, j \in \{1, \ldots, N\}$, we have*

$$\mathbb{P}(X_{n+m} = s_j \mid X_n = s_i) = P_{i,j}^m.$$

2.1 Irreducibility and aperiodicity

We are interested in Markov chains that satisfy certain assumptions, since many results that we use hold only for Markov chains satisfying this assumptions. For the rest of the section we will consider a Markov chain $(X_t)_{t\geq 0}$ over state space χ with the transition matrix P, unless stated otherwise.

First important assumption is irreducibility, which states that starting from any state in χ, the chain can reach any other state in χ in a finite number of steps.

Definition 2.8. *A Markov chain $(X_t)_{t\geq 0}$ over state space χ is irreducible if for all $i, j \in \{1, \dots, N\}$ there exist $m \in \mathbb{N}$ such that*

$$\mathbb{P}(X_{n+m} = s_j \mid X_n = s_i) > 0.$$

Remark 2.9. According to Corollary 2.7, $\mathbb{P}(X_{n+m} = s_j \mid X_n = s_i)$ does not depend on n, and we could have defined a chain $(X_t)_{t\geq 0}$ with transition matrix P to be irreducible if for all $i, j \in \{1, \dots, N\}$ there exist $m \in \mathbb{N}_{>0}$ such that $P^m_{i,j} > 0$.

Then, we define aperiodicity. First we define the period of a state.

Definition 2.10. *Let $s_i \in \chi$. The period of the state s_i is*

$$g(s_i) := gcd\{n > 1 \mid P^n_{i,i} > 0\}.$$

In other words, it is the greatest common divisor of the times at which it is possible for the chain to return to state s_i, assuming that it started from s_i.

Next, we can define aperiodicity for each state, and for the whole chain.

Definition 2.11. *A state $s_i \in \chi$ is aperiodic if $g(s_i) = 1$. Markov chain $(X_t)_{t\geq 0}$ is aperiodic if all of its states are aperiodic.*

2.2 Total variation distance and coupling

For a moment, we take a step back from Markov chains and focus on probability distributions. We will consider probability distributions over a finite set χ, which can represent a state space for some Markov chain. Our main interest is the *total variation distance* between two distributions, and we will give three equivalent definitions of this property.

Definition 2.12. *Let μ, ν be two probability distributions over $\chi = \{s_1, \dots, s_N\}$. For all $i \in \{1, \dots, N\}$, we write $\mu_i := \mu(s_i)$ and $\nu_i := \nu(s_i)$. The total variation distance between μ and ν is given by*

$$d_{TV}(\mu, \nu) := \frac{1}{2} \sum_{i=1}^{N} \mid \mu_i - \nu_i \mid .$$

One can check that $0 \leq d_{TV}(\mu, \nu) \leq 1$ for any distributions μ and ν (this is why we introduce the constant $\frac{1}{2}$ in front of the sum). The next proposition gives an alternative definition of the total variation distance [3, Proposition 4.2.].

Proposition 2.13. *Let μ, ν be two probability distributions over χ. Then, we have*

$$d_{TV}(\mu, \nu) = \max_{A \subset \chi} \mid \mu(A) - \nu(A) \mid .$$

Both of these definitions give some intuition on what the total variation distance represents: it allows us to quantify how much two distributions differ between each other. It is not hard to see that $d_{TV}(\mu, \nu) = 0$ if and only if $\mu = \nu$ (i.e. $\mu_i = \nu_i$ for all i). The third alternative definition is slightly more abstract, but it is the one that we will use in some proofs in the latter sections. First, we introduce the notion of *coupling*.

Definition 2.14. *Let X, Y be two random variables taking values in χ. Then, (X, Y) is a coupling of distributions μ and ν if $\mathbb{P}(X = s_i) = \mu_i$ and $\mathbb{P}(Y = s_i) = \nu_i$ for all $i \in \{1, \dots, N\}$. If this is the case, we write $(X, Y) \sim (\mu, \nu)$.*

Using coupling, we can give another definition of the total variation distance, shown in the following proposition [3, Proposition 4.7.].

Proposition 2.15. *Let μ, ν be two probability distributions over χ. Then, we have*

$$d_{TV}(\mu, \nu) = \inf\{\mathbb{P}(X \neq Y) \mid (X, Y) \sim (\mu, \nu)\}.$$

2.3 Stationary distributions and distance from stationarity

We now have the tools to present one of the crucial properties of Markov chains, without which none of the algorithms that we introduce later would work. First, we define a *stationary distribution* of a Markov chain.

Definition 2.16. *Let π be a probability distribution over χ. π is a stationary distribution of the Markov chain $(X_t)_{t\geq 0}$ over χ with transition matrix P if*

$$\pi = \pi P.$$

The following proposition states the existence of a stationary distribution for irreducible, aperiodic chains [2, Theorem 5.1.].

Proposition 2.17. *Any irreducible, aperiodic Markov chain has a stationary distribution.*

And now comes the most important result, stated in the theorem below [2, Theorem 4.9.] or [3, Theorem 5.2.]. It tells us that, if we run an irreducible, aperiodic Markov chain for long enough, its current state will be distributed 'almost' according to the stationary distribution.

Theorem 2.18 (Convergence theorem)**.** *Let $(X_t)_{t\geq 0}$ be an irreducible, aperiodic Markov chain over χ, with transition matrix P. Assume π is its stationary distribution. Assume that X_0 follows some distribution ν^0. Recall that ν^t denotes the distribution of X_t. We have*

$$\lim_{t\to\infty} d_{TV}(\pi, \nu^t) = 0.$$

From this theorem, it is possible to obtain a result on unicity of the stationary distribution[2, Theorem 5.3.].

Corollary 2.19. *Any irreducible, aperiodic Markov chain has a unique stationary distribution.*

Notice that the quantity $d_{TV}(\pi, \nu^t)$ from Theorem 2.18 indicates how 'close' the chain is to its stationary distribution. However, this quantity depends on the initial distribution ν^0. It is useful to have a quantity that denotes how 'close' the chain is to its stationary distribution but which depends only on time t, and not on the initial distribution. For that purpose, we introduce two notions of distance from stationarity. We denote by $\mathcal{D}(\chi)$ the set of all probability distributions over χ.

Definition 2.20. *Let $(X_t)_{t\geq 0}$ be an irreducible, aperiodic Markov chain with transition matrix P. Let π be its stationary distribution. We define*

$$d(t) = sup_{\nu\in\mathcal{D}(\chi)}(d_{TV}(\nu P^t, \pi))$$

and

$$\bar{d(t)} = sup_{\nu,\mu\in\mathcal{D}(\chi)^2}(d_{TV}(\nu P^t, \mu P^t)).$$

We denote by $\rho_{(i)}$ the probability distribution which puts all the probability on state s_i, i.e. for all $j \in \{1, \ldots, N\}$ $\rho_{(i)j} = \delta_{i,j}$. Notice that $\rho_{(i)}P^t$ is just the i^{th} row of P^t, denoting the distribution at time t of the chain started from state s_i. It is possible to show that

$$d(t) = \max_{1\leq i\leq N}(d_{TV}(\rho_{(i)}P^t, \pi))$$

and

$$\bar{d}(t) = \max_{1\leq i,j\leq N}(d_{TV}(\rho_{(i)}P^t, \rho_{(j)}P^t))$$

[3, exercise 4.1.].

We now present two properties of $d(t)$ and $\bar{d}(t)$ [3, Lemma 4.10. and 4.11.], that will be useful later.

Proposition 2.21. *Let $d(t)$, $\bar{d}(t)$ as defined in 2.20. Then*

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

Proposition 2.22 (Submultiplicativity of $\bar{d}$)**.** *For all integers $t, s \geq 0$, we have $\bar{d}(t+s) \leq \bar{d}(t)\bar{d}(s)$.*

Finally, we define the mixing time of a Markov chain.

Definition 2.23. *For any $\epsilon > 0$, we define the mixing time of an irreducible aperiodic Markov chain $(X_t)_{t\geq 0}$*

$$t_{mix}(\epsilon) = \min\{t \geq 0 \mid \bar{d}(t) \leq \epsilon\}.$$

By default, we take $\epsilon = 1/4$ and write $t_{mix} = \min\{t \geq 0 \mid \bar{d}(t) \leq 1/4\}$.

3 Simulation

In order to exploit some of the useful properties of Markov chains that we defined earlier, we need to be able to simulate a Markov chain, given its transition probabilities (i.e. the transition matrix). We introduce a relatively simple but quite powerful method [2, Section 3] that is useful not only for simulation but also for proving some properties of the algorithms that we present in later sections.

Assume that we have access to i.i.d. random variables $(U_t)_t$ following the uniform distribution on interval $[0, 1]$. It is worth noting that, in practice, this assumption is almost never satisfied. This is because our computers are usually only able to generate 'pseudo random' numbers, meaning that the numbers are a product of a deterministic algorithm. However, these 'pseudo random' numbers tend to behave well for our purposes (meaning that they resemble truly random numbers quite well), so we will stick to our assumption.

3.1 Update function

Let $\chi = \{s_1, \ldots, s_N\}$ be a finite set and let $(X_t)_{t\geq 0}$ be a Markov chain over χ with a transition matrix P. Assume that X_0 is fixed ($X_0 = s_i$ for some $s_i \in \chi$). Let $(U_t)_{t\geq 1}$ be a sequence of i.i.d. uniform random variables. In order to simulate the evolution of $(X_t)_{t\geq 0}$ using $(U_t)_{t\geq 1}$, 'as a source of randomness', we need what is called an update function.

Definition 3.1. *Let $\phi : \chi \times [0, 1] \to \chi$ be a function. Then, ϕ is a valid update function for the chain $(X_t)_{t\geq 0}$ if, for all $i, j \in \{1, \ldots, N\}$, we have*

$$\mathbb{P}(\phi(s_i, U_t) = s_j) = P_{i,j}.$$

Remark 3.2. $\mathbb{P}(\phi(s_i, U_t) = s_j)$ does not depend on t because $(U_t)_t$ are i.i.d.

Proposition 3.3. *The sequence $(Y_t)_{t\geq 0}$, defined as follows:*
1) fix $Y_0 = s_k$ for some $s_k \in \chi$
2) set $Y_t = \phi(Y_{t-1}, U_t)$ for all $t \geq 1$
is a Markov chain with transition matrix P.

Proof. Indeed, for all $t \geq 1$ and all $i_1, \ldots, i_t \in \{1, \ldots N\}$ we have:

$$\begin{aligned}
&\mathbb{P}(Y_t = s_{i_t} \mid Y_{t-1} = s_{i_{t-1}}, \ldots, Y_1 = s_{i_1}, Y_0 = s_k) \\
&= \mathbb{P}(\phi(s_{i_{t-1}}, U_t) = s_{i_t} \mid \phi(s_{i_{t-2}}, U_{t-1}) = s_{i_{t-1}}, \ldots, \phi(s_k, U_1) = s_{i_1}, Y_0 = s_k) \\
&= \mathbb{P}(\phi(s_{i_{t-1}}, U_t) = s_{i_t}) \\
&= P_{i_{t-1}, i_t}.
\end{aligned}$$

□

We say that the Markov chain $(Y_t)_{t\geq 0}$ evolves according to the update function ϕ and i.i.d. uniform random variables $(U_t)_{t\geq 1}$. Furthermore, notice that Y_t is obtained by applying ϕ to s_k t times, with $U_1, \ldots, U_t$ as the second arguments. Thus, Y_t is a function of $s_k, U_1, \ldots, U_t$. We can write this as

$$Y_t = \phi(Y_{t-1}, U_t) = \cdots = \phi(\phi(\ldots(\phi(s_k, U_1), U_2)\ldots), U_t).$$

However, it is convenient to introduce a tidier notation.

Definition 3.4. *For all $t \geq 1$, define $F_t : \chi \times [0,1]^t \to \chi$ as follows:*
1) if $t = 1$, then $F_1(s_i, U_1) = \phi(s_i, U_1)$ for any $s_i \in \chi$ and any $U_1 \in [0,1]$
2) if $t \geq 2$, then $F_t(s_i, U_1, \ldots, U_t) = F_{t-1}(\phi(s_i, U_1), U_2, \ldots, U_t)$ for any $s_i \in \chi$ and any $U_1, \ldots, U_t \in [0,1]$.

If $(Y_t)_t$ is defined as above, it is clear from the definition of F_t that $Y_t = F_t(s_k, U_1, \ldots, U_t)$.

3.2 Grand couplings

It is often useful to run multiple instances of a Markov chain with transition matrix P, possibly starting from different initial states, as we shall see in Section 6. Thus, we define a notion of a coupling of Markov chains.

Definition 3.5. *The sequence $(X_t, Y_t)_{t\geq 0}$ is a coupling of Markov chains with transition matrix P if $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are both Markov chains with transition matrix P.*

Remark 3.6. Consider a coupling of Markov chains with transition matrix P: $(X_t, Y_t)_{t\geq 0}$. If μ^0 and ν^0 are two probability distributions and X_0 is chosen according to μ^0 and Y_0 is chosen according to ν^0, then for all $t \geq 0$, (X_t, Y_t) is a coupling of distributions $\mu^t = \mu^0 P^t$ and $\nu^t = \nu^0 P^t$ ($(X_t, Y_t) \sim (\mu^0 P^t, \nu^0 P^t)$).

We will denote by $(X_t^{(k)})_{t\geq 0}$ a Markov chain started from s_k, i.e. if $\mathbb{P}(X_0^{(k)} = s_k) = 1$. The Propp-Wilson algorithm relies on running N Markov chains (where $N = \mid \chi \mid$), one starting from each state. This is why we need to define grand couplings.

Definition 3.7. *Let $\chi = \{s_1, \ldots, s_N\}$ be a finite set. Then, $(X_t^{(1)}, \ldots, X_t^{(N)})_{t\geq 0}$ is a grand coupling over χ with transition matrix P if, for all $k \in \{1, \ldots, N\}$ we have $\mathbb{P}(X_0^{(k)} = s_k) = 1$, and $(X_t^{(k)})_{t\geq 0}$ is a Markov chain over χ with transition matrix P.*

In practice, we will simulate all of the chains from a grand coupling using the same update function ϕ and the same sequence of i.i.d. random variables $(U_t)_t$. This means that for all $t \geq 1$ and all $k \in \{1, \ldots, N\}$ we have $X_t^{(k)} = F_t(s_k, U_1, \ldots, U_t)$. By Proposition 3.3, $(F_t(s_1, U_1, \ldots, U_t), \ldots, F_t(s_N, U_1, \ldots, U_t))_{t\geq 0}$ is indeed a grand coupling over χ with transition matrix P. The main upside of this approach is that if two chains collide at some point, they will continue to run together. Indeed, if we have $X_t^{(i)} = X_t^{(j)}$ for some $t \geq 0$ and $i, j \in \{1, \ldots, N\}$, then, for all $t_0 \geq 0$, we have

$$\begin{aligned}
X_{t+t_0}^{(i)} &= F_{t+t_0}(s_i, U_1, \ldots, U_{t+t_0}) \\
&= F_{t_0}(F_t(s_i, U_1, \ldots, U_t), U_{t+1}, \ldots, U_{t+t_0}) \\
&= F_{t_0}(X_t^{(i)}, U_{t+1}, \ldots, U_{t+t_0}) \\
&= F_{t_0}(X_t^{(j)}, U_{t+1}, \ldots, U_{t+t_0}) \\
&= F_{t_0}(F_t(s_j, U_1, \ldots, U_t), U_{t+1}, \ldots, U_{t+t_0}) \\
&= X_{t+t_0}^{(j)}.
\end{aligned}$$

This means that if we have $X_t^{(1)} = \cdots = X_t^{(N)}$ for some $t \geq 0$, then we have $X_{t+t_0}^{(1)} = \cdots = X_{t+t_0}^{(N)}$ for all $t_0 \geq 0$, and we say that the grand coupling has *coalesced*.

If, for all $k \in \{1, \ldots, N\}$, $X_t^{(k)}$ is defined as above, we say that the grand coupling $(X_t^{(1)}, \ldots, X_t^{(N)})_{t\geq 0}$ evolves according to the same update function ϕ and the same source of randomness $(U_t)_{t\geq 1}$.

4 Markov chain Monte Carlo algorithms

Having introduced the main properties of Markov chains and some basic tools for simulating them, we can now present a wide class of algorithms, called Markov chain Monte Carlo algorithms (MCMC). These algorithm rely on the convergence property of Markov chains (see Theorem 2.18).

Suppose we have a set $\chi = \{s_1, \ldots, s_N\}$, and a probability distribution π over χ. Our goal is to draw samples from π. Here is one simple way to do this. Take a $[0,1]$-uniform random variable U, and define $\psi : [0,1] \to \chi$ as follows:

$$\psi(u) = \begin{cases} s_1 \text{ if } 0 \leq u < \pi_1 \\ \ldots \\ s_i \text{ if } \sum_{j=1}^{i-1} \pi_j \leq u < \sum_{j=1}^{i} \pi_j \\ \ldots \\ s_N \text{ if } \sum_{j=1}^{N-1} \pi_j \leq u \leq 1 \end{cases} .$$

Then, $\mathbb{P}(\psi(U) = s_i) = \mathbb{P}(\sum_{j=1}^{i-1} \pi_j \leq U < \sum_{j=1}^{i} \pi_j) = \pi_i$, so $\psi(U)$ is distributed according to π.

However, if N is very large, this is not a feasible method in practice. One approach is to slightly alleviate our requirement: instead of drawing samples from π exactly, we are satisfied with samples from some distribution π' which is similar to π, i.e. $d_{TV}(\pi, \pi')$ is smaller then some threshold. Markov chain Monte Carlo methods achieve this by constructing an irreducible, aperiodic Markov chain $(X_t)_{t \geq 0}$ whose stationary distribution is π and then running the chain for some (large, not necessarily deterministic) number of steps n. If we denote by μ^n the distribution of X_n, we know from the Convergence Theorem, that $d_{TV}(\mu^n, \pi) \to 0$ as $n \to \infty$, for any choice of the initial state. Thus, by running the chain for long enough, we will hopefully obtain a sample from approximately the stationary distribution.

It is reasonable to ask how we can construct such chain $(X_t)_t$ and simulate it efficiently, if we are unable to efficiently draw samples from a distribution on χ. We present one example from statistical physics in Section 7: the Ising model. For more examples, see [2, section 7].

Another, more legitimate concern, mentioned in the introduction, is posed by the fact that we usually do not know $d_{TV}(\mu^n, \pi)$, so we do not know for how long to run the chain to obtain satisfactory samples. As we already mentioned, the Propp-Wilson algorithm [1] represents one potential solution to this issue. It also relies on the ability to construct an irreducible, aperiodic Markov chain whose stationary distribution is the desired distribution, so one could say that it is itself a MCMC algorithm. However, it gives exact samples from the stationary distribution π, and it determines automatically when to stop. A detailed analysis is provided in the Section 6.

5 The 0-1 law

For a grand coupling $(X_t^{(1)}, \ldots, X_t^{(N)})_{t \geq 0}$, we define the coalescence time to be the random variable

$$\tau = \min\{t > 0 \mid X_t^{(1)} = \cdots = X_t^{(N)}\}.$$

Equivalently, if the coupling evolves according to the i.i.d. random variables $(U_t)_{t \geq 1}$ and update function ϕ, we can define the coalescence time as

$$\tau = \min\{t > 0 \mid F_t(\cdot, U_1, \ldots, U_t) = const\}.$$

$F_t(\cdot, U_1, \ldots, U_t) = const$ simply denotes that the function $s \in \chi \mapsto F_t(s, U_1, \ldots, U_t) \in \chi$ is constant. Coalescence time is a crucial property for studying the Propp-Wilson algorithm. In this section, we show that the probability of τ being finite is either 0 or 1, depending on specific conditions on the Markov chain $(X_t)_t$ (more precisely, on the grand coupling $(X_t^{(1)}, \ldots, X_t^{(N)})_{t \geq 0}$). But first, we state and prove two auxiliary propositions.

Proposition 5.1. *Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of events, over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We have the following:*

i) if $\mathbb{P}(A_n) = 0$ for all $n \in \mathbb{N}$, then $\mathbb{P}(\bigcup_{n \in \mathbb{N}} A_n) = 0$

ii) if $\mathbb{P}(A_n) = 1$ for all $n \in \mathbb{N}$, then $\mathbb{P}(\bigcap_{n \in \mathbb{N}} A_n) = 1$

Proof. First, we show that $\mathbb{P}(\bigcup_{n \in \mathbb{N}} A_n) \leq \sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$, for any sequence of events $(A_n)_{n \in \mathbb{N}}$. Construct the sequence of events $(B_n)_{n \in \mathbb{N}}$ such that $B_0 = A_0$, and $B_n = A_n \setminus \bigcup_{i<n} A_i$ for all $n \geq 1$. Then, $(B_n)_{n \in \mathbb{N}}$ is a sequence of pairwise-disjoint events, by construction. Moreover, we have

$\bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n$, again, by construction of $(B_n)_{n\in\mathbb{N}}$. Finally, for all $n \in \mathbb{N}$, $B_n \subset A_n$, so $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$. This yields the desired inequality:

$$\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \mathbb{P}(\bigcup_{n\in\mathbb{N}} B_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(B_n) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

i) If $\mathbb{P}(A_n) = 0$ for all $n \in \mathbb{N}$, then we obtain $0 \leq \mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = 0$ so $\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = 0$.

ii) Assume $\mathbb{P}(A_n) = 1$ for all $n \in \mathbb{N}$. Then $\mathbb{P}((A_n)^c) = 0$ for all $n \in \mathbb{N}$. By *i)*, we have $\mathbb{P}(\bigcap_{n\in\mathbb{N}} A_n) = 1 - \mathbb{P}((\bigcap_{n\in\mathbb{N}} A_n)^c) = 1 - \mathbb{P}(\bigcup_{n\in\mathbb{N}}(A_n)^c) = 1$. □

The next proposition presents the so called 'memoryless property', applied to couplings of Markov chains. It is analogous to Corollary 2.7.

Proposition 5.2. *Let $(X_t)_{t\geq 0}$ be an irreducible, aperiodic Markov chain over a finite state-space $\chi = \{s_1, s_2, \ldots, s_N\}$, with a transition matrix P. Assume that ϕ is valid update function for this chain. Let $(i,j) \in \{1,\ldots,N\}^2$ and let $(X_t^{(i)}, X_t^{(j)})_{t\geq 0}$ be a coupling of two P-chains, started from (s_i, s_j), evolving according to ϕ and uniform random variables $(U_t)_{t\geq 1}$. Let $T \geq 0$, and define $Y_t = X_{T+t}^{(i)}$ and $Z_t = X_{T+t}^{(j)}$. for all $t \geq 0$. Then, for any $(l,m) \in \{1,\ldots,N\}^2$ such that $\mathbb{P}((X_T^{(i)}, X_T^{(j)}) = (s_l, s_m)) > 0$ and for all $t \geq 0$, the joint distribution of (Y_t, Z_t) conditional on $(Y_0, Z_0) = (s_l, s_m)$ (or, equivalently, on $(X_T^{(i)}, X_T^{(j)}) = (s_l, s_m)$) is the same as the joint distribution of $(X_t^{(l)}, X_t^{(m)})$, where $(X_t^{(l)}, X_t^{(m)})_{t\geq 0}$ is a coupling started from (s_l, s_m), evolving according to ϕ and $(U_t)_{t\geq 1}$.*

Proof. Let $(l,m) \in \{1,\ldots,N\}^2$. We proceed by induction on t. For $t = 0$, we have $\mathbb{P}[(Y_0, Z_0) = (s_l, s_m) \mid (Y_0, Z_0) = (s_l, s_m)] = 1$ and $\mathbb{P}[(X_0^{(l)}, X_0^{(m)}) = (s_l, s_m)] = 1$. Now, let $t \geq 0$ and assume that $\mathbb{P}[(Y_t, Z_t) = (s_h, s_k) \mid (Y_0, Z_0) = (s_l, s_m)] = \mathbb{P}[(X_t^{(l)}, X_t^{(m)}) = (s_h, s_k)]$ for all $(h,k) \in \{1,\ldots,N\}^2$. Then, for any $(p,q) \in \{1,\ldots,N\}^2$:

$$\begin{aligned}
&\mathbb{P}[(Y_{t+1}, Z_{t+1}) = (s_p, s_q) \mid (Y_0, Z_0) = (s_l, s_m)] = \\
&= \sum_{(h,k)\in\{1,\ldots,N\}^2} \mathbb{P}[(Y_{t+1}, Z_{t+1}) = (s_p, s_q) \mid (Y_t, Z_t) = (s_h, s_k), (Y_0, Z_0) = (s_l, s_m)] \times \\
&\qquad\qquad \mathbb{P}[(Y_t, Z_t) = (s_h, s_k) \mid (Y_0, Z_0) = (s_l, s_m)] \\
&= \sum_{(h,k)\in\{1,\ldots,N\}^2} \mathbb{P}[\phi(s_h, U_{T+t+1}) = s_p, \phi(s_k, U_{T+t+1}) = s_q] \times \mathbb{P}[(X_t^{(l)}, X_t^{(m)}) = (s_h, s_k)] \\
&= \sum_{(h,k)\in\{1,\ldots,N\}^2} \mathbb{P}[\phi(s_h, U_{t+1}) = s_p, \phi(s_k, U_{t+1}) = s_q] \times \mathbb{P}[(X_t^{(l)}, X_t^{(m)}) = (s_h, s_k)] \\
&= \sum_{(h,k)\in\{1,\ldots,N\}^2} \mathbb{P}[(X_{t+1}^{(l)}, X_{t+1}^{(m)}) = (s_p, s_q) \mid (X_t^{(l)}, X_t^{(m)}) = (s_h, s_k)] \times \mathbb{P}[(X_t^{(l)}, X_t^{(m)}) = (s_h, s_k)] \\
&= \mathbb{P}[(X_{t+1}^{(l)}, X_{t+1}^{(m)}) = (s_p, s_q)].
\end{aligned}$$

This concludes our proof by induction. □

Theorem 5.3 (The 0-1 law)**.** *Let $(X_t)_t$ be an irreducible, aperiodic Markov chain over a finite state-space $\chi = \{s_1, s_2, \ldots, s_N\}$, with a transition matrix P. Assume ϕ is a valid update function for the chain. Consider a grand coupling $(X_t^{(1)}, \ldots, X_t^{(N)})$ evolving according to ϕ and uniform random variables $(U_t)_{t\geq 1}$. Let $\tau = min\{t > 0 \mid X_t^{(1)} = \cdots = X_t^{(N)}\}$ be the coalescence time of this grand coupling. If there exists an integer $T > 0$ such that for all $(i,j) \in \{1,\ldots,N\}^2$, $\mathbb{P}(X_T^{(i)} = X_T^{(j)}) > 0$, then $\mathbb{P}(\tau < \infty) = 1$. Otherwise, $\mathbb{P}(\tau < \infty) = 0$*

Proof. Assume there exists an integer $T > 0$ such that for all $(i,j) \in \{1,\ldots,N\}^2$, $\mathbb{P}(X_T^{(i)} = X_T^{(j)}) > 0$. Let $\delta = \min_{1\leq i,j\leq N} \mathbb{P}(X_T^{(i)} = X_T^{(j)})$. By assumption, $\delta > 0$. Now we define for all $i = 1,\ldots,N$ the coalescence time between $X_t^{(i)}$ and $X_t^{(N)}$: $\tau_i = \min\{t \geq 0 \mid X_t^{(i)} = X_t^{(N)}\}$ (note that $\tau_N = 0$).

First we prove that for all $i = 1, \dots, N$, for all $k \in \mathbb{N}^*$, we have $\mathbb{P}(\tau_i > kT) \leq (1-\delta)^k$. The result clearly holds for $i = N$, so assume $i \leq N-1$. We proceed by induction on k:

1) For $k = 1$, $\mathbb{P}(\tau_i > T) = \mathbb{P}(X_T^{(i)} \neq X_T^{(N)}) \leq 1 - \delta$.

2) Assume the result holds for some $k \in \mathbb{N}^*$. Then

$$\begin{aligned}
\mathbb{P}(\tau_i > (k+1)T) &= \mathbb{P}(\tau_i > (k+1)T \mid \tau_i > kT)\mathbb{P}(\tau_i > kT) \\
&\leq \mathbb{P}(\tau_i > (k+1)T \mid \tau_i > kT)(1-\delta)^k \qquad (5.1) \\
&= \mathbb{P}(X_{(k+1)T}^{(i)} \neq X_{(k+1)T}^{(N)} \mid X_{kT}^{(i)} \neq X_{kT}^{(N)})(1-\delta)^k
\end{aligned}$$

where inequality (5.1) follows from the induction hypothesis. We can then define $Y_t = X_{kT+t}^{(i)}$ and $Z_t = X_{kT+t}^{(N)}$ for all $t \geq 0$. Moreover, let $S^+ = \{(j,k) \in \{1, \dots N\}^2 \mid \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)] > 0\}$. Then, by Proposition 5.2, the joint distribution of (Y_t, Z_t) conditional on $(Y_0, Z_0) = (s_j, s_k)$ is the same as the joint distribution of $(X_t^{(j)}, X_t^{(k)})$, for any $(j,k) \in S^+$ and for all $t \geq 0$. This leads to the following result:

$$\begin{aligned}
\mathbb{P}(Y_T \neq Z_T \mid Y_0 \neq Z_0) &= \mathbb{P}[Y_T \neq Z_T \mid \bigcup_{j,k \in \{1,\dots,N\}, j \neq k} \{(Y_0, Z_0) = (s_j, s_k)\}] \\
&= \frac{\mathbb{P}[\{Y_T \neq Z_T\} \cap (\bigcup_{j,k \in \{1,\dots,N\}, j \neq k} \{(Y_0, Z_0) = (s_j, s_k)\})]}{\mathbb{P}[\bigcup_{j,k \in \{1,\dots,N\}, j \neq k} \{(Y_0, Z_0) = (s_j, s_k)\}]} \\
&= \frac{\mathbb{P}[\bigcup_{j,k \in \{1,\dots,N\}, j \neq k} (\{Y_T \neq Z_T\} \cap \{(Y_0, Z_0) = (s_j, s_k)\})]}{\sum_{j,k \in \{1,\dots,N\}, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]} \\
&= \frac{\sum_{j,k \in \{1,\dots,N\}, j \neq k} \mathbb{P}[\{Y_T \neq Z_T\} \cap \{(Y_0, Z_0) = (s_j, s_k)\}]}{\sum_{j,k \in \{1,\dots,N\}, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]} \\
&= \frac{\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[Y_T \neq Z_T \mid (Y_0, Z_0) = (s_j, s_k)]\mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]}{\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]} \\
&= \frac{\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[X_T^{(j)} \neq X_T^{(k)}]\mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]}{\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]} \\
&\leq \frac{(1-\delta)\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]}{\sum_{(j,k) \in S^+, j \neq k} \mathbb{P}[(Y_0, Z_0) = (s_j, s_k)]} = 1 - \delta.
\end{aligned}$$

Therefore $\mathbb{P}(X_{(k+1)T}^{(i)} \neq X_{(k+1)T}^{(N)} \mid X_{kT}^{(i)} \neq X_{kT}^{(N)}) = \mathbb{P}(Y_T \neq Z_T \mid Y_0 \neq Z_0) \leq 1 - \delta$, so $\mathbb{P}(\tau_i > (k+1)T) \leq (1-\delta)^{k+1}$. This concludes the induction proof that $\mathbb{P}(\tau_i > kT) \leq (1-\delta)^k$ for all starting states s_i and for all positive integers k.

We can now deduce that the chain started at state s_N will coalesce with any other chain from the coupling with probability equal to 1. Let $i \in \{1, \dots, N\}$. We have, by continuity from above and by the fact that $\mathbb{P}(\tau_i > kT) \leq (1-\delta)^k$:

$$0 \leq \mathbb{P}(\tau_i = \infty) = \mathbb{P}(\bigcap_{k=1}^{\infty} \{\tau_i > kT\}) = \lim_{k \to \infty} \mathbb{P}(\tau_i > kT) \leq \lim_{k \to \infty} (1-\delta)^k = 0,$$

so $\mathbb{P}(\tau_i < \infty) = 1 - \mathbb{P}(\tau_i = \infty) = 1$.

Finally, observe that $\tau = \max_{1 \leq i \leq N} \tau_i$. This comes from the fact that two chains from the coupling continue to evolve identically when they collide. Thus, once the chain $(X_t^{(N)})_{t \geq 0}$ coalesces with all the other chains from the coupling, it means that the whole coupling has coalesced. Since τ is a maximum over a finite set, we can write $\mathbb{P}(\tau < \infty) = \mathbb{P}(\bigcap_{i=1}^N \{\tau_i < \infty\})$. Moreover, since $\mathbb{P}(\tau_i < \infty) = 1$ for all $i = 1, \dots, N$, then $\mathbb{P}(\tau < \infty) = \mathbb{P}(\bigcap_{i=1}^N \{\tau_i < \infty\}) = 1$, by Proposition 5.1.

Now, assume that for all $T > 0$, there exist $(i_T, j_T) \in \{1, \ldots, N\}^2$ such that $\mathbb{P}(X_T^{(i_T)} = X_T^{(j_T)}) = 0$.. Then, for all $T > 0$, $\mathbb{P}(\tau \leq T) = \mathbb{P}(X_T^{(1)} = \cdots = X_T^{(N)}) \leq \mathbb{P}(X_T^{(i_T)} = X_T^{(j_T)}) = 0$. Thus, $\mathbb{P}(\tau > T) = 1$ for all T, so we conclude that $\mathbb{P}(\tau = \infty) = \mathbb{P}(\bigcap_{T=1}^{\infty}\{\tau > T\}) = 1$, by Proposition 5.1. □

6 The Propp-Wilson algorithm

Let $(X_t)_t$ be an irreducible, aperiodic Markov chain over a finite state-space $\chi = \{s_1, s_2, \ldots, s_N\}$, with a transition matrix P, and let π be its unique stationary distribution. The Propp-Wilson algorithm [1] allows us to draw samples from π, grand couplings of $(X_t)_t$ chains. Here is how it works.

Assume that we have access to a sequence of i.i.d random variables $(U_t)_{t\leq 0}$ uniformly distributed over $[0, 1]$. Further, assume that $\phi : \chi \times [0, 1] \to \chi$ is a valid update function for the transition matrix P. Let $(T_i)_{i\geq 0}$ be a strictly increasing sequence of natural numbers. These will denote our starting times. Then, for each starting time T_i, we run a grand coupling $(X_t^{(1)}, \ldots, X_t^{(N)})_{-T_i\leq t\leq 0}$ evolving according to ϕ and $(U_t)_{-T_i+1\leq t\leq 0}$. This simply means that, for all $1 \leq k \leq N$ and all $-T_i + 1 \leq t \leq 0$, we have

$$\begin{aligned} X_{-T_i}^{(k)} &= s_k \text{ and} \\ X_t^{(k)} &= \phi(X_{t-1}^{(k)}, U_t) = F_{T_i-t}(s_k, U_{-T_i+1}, \ldots, U_t). \end{aligned}$$

If $X_0^{(1)} = \cdots = X_0^{(N)}$ (or, equivalently $F_{T_i}(\cdot, U_{-T_i+1}, \ldots, U_0) = const$), then we return $X_0^{(1)}$ (or any other $X_0^{(k)}$). If not, proceed to the next starting time T_{i+1} and repeat. The pseudocode of the algorithm is given in 1.

Algorithm 1 Propp-Wilson algorithm

Require: i.i.d. uniform random variables $(U_t)_{t\leq 0}$; strictly increasing starting times $(T_i)_{i\geq 0}$

```
j ← 0
coalesced ← False
while not coalesced do
    for k = 1, 2, ..., N do                ▷ Initialize N chains, one starting from each state
        X^(k) ← s_k
    end for
    for t = −T_j + 1, −T_j + 2, ..., 0 do
        for k = 1, 2, ..., N do
            X^(k) ← φ(X^(k), U_t)          ▷ Update the chains
        end for
    end for
    if X^(1) = ··· = X^(N) then            ▷ If the chains have coalesced, we have an output
        result ← X^(1)
        coalesced ← True
    end if
    j ← j + 1
end while
j ← j − 1
return result, j
```

Note that we also return the index of the earliest starting time that we try, stored inside the variable j. This allows us to calculate the running time of our algorithm, which we define in a later section.

From the description, it is not obvious whether or not this algorithm terminates. The 0-1 law tells us that, in general, it either terminates with probability 1 if certain conditions are met, or

it does not terminate at all (i.e. it terminates with probability 0). To see this more clearly, it is useful to define the "backward coalescence time"

$$\bar{\tau} := \min\{t \geq 1 \mid F_t(\cdot, U_{-t+1}, \ldots, U_0) = const\}.$$

Notice that for all $t \geq \bar{\tau}$ we have $F_t(\cdot, U_{-t+1}, \ldots, U_0) = const$. Indeed, for any $k, l \in \{1, \ldots, N\}$:

$$\begin{aligned} F_t(s_k, U_{-t+1}, \ldots, U_0) &= F_{\bar{\tau}}(F_{t-\bar{\tau}}(s_k, U_{-t+1}, \ldots, U_{-\bar{\tau}}), U_{-\bar{\tau}+1}, \ldots, U_0) \\ &= F_{\bar{\tau}}(F_{t-\bar{\tau}}(s_l, U_{-t+1}, \ldots, U_{-\bar{\tau}}), U_{-\bar{\tau}+1}, \ldots, U_0) \\ &= F_t(s_l, U_{-t+1}, \ldots, U_0). \end{aligned}$$

And conversely, if $F_t(\cdot, U_{-t+1}, \ldots, U_0) = const$ then $t \geq \bar{\tau}$, by definition of $\bar{\tau}$.

The crucial observation, summarized in Proposition 6.1 below, is that the coalescence time τ and the backward coalescence time $\bar{\tau}$ are identically distributed.

Proposition 6.1. *Let $(U_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. uniform random variables. Let ϕ be an update function for an irreducible, aperiodic Markov chain. Then, the coalescence time $\tau = \min\{t > 0 \mid F_t(\cdot, U_1, \ldots, U_t) = const\}$ and the backward coalescence time $\bar{\tau} := \min\{t \geq 1 \mid F_t(\cdot, U_{-t+1}, \ldots, U_0) = const\}$ are identically distributed.*

Proof. For any $t \geq 0$, $U_{-t+1}, \ldots, U_t$, are i.d.d, so we have

$$\mathbb{P}(F_t(\cdot, U_1, \ldots, U_t) = const) = \mathbb{P}(F_t(\cdot, U_{-t+1}, \ldots, U_0) = const).$$

Therefore,

$$\begin{aligned} \mathbb{P}(\tau \leq t) &= \mathbb{P}(F_t(\cdot, U_1, \ldots, U_t) = const) \\ &= \mathbb{P}(F_t(\cdot, U_{-t+1}, \ldots, U_0) = const) = \mathbb{P}(\bar{\tau} \leq t), \end{aligned}$$

so τ and $\bar{\tau}$ are identically distributed. □

Since $F_t(\cdot, U_{-t+1}, \ldots, U_0) = const$ for all $t \geq \bar{\tau}$, it is clear that the algorithm will stop as soon as it tries the first starting time $T_i \geq \bar{\tau}$ (i.e. when it runs the coupling $(X_t^{(1)}, \ldots, X_t^{(N)})_{t\geq -T_i}$). Thus, the algorithm terminates if and only if $\bar{\tau}$ is finite. Since $\bar{\tau}$ and τ are distributed identically, the 0-1 law tells us that the Propp-Wilson algorithm terminates if and only if there exits $T \geq 0$ such that, for all $k, l \in \{1, \ldots, N\}$, $\mathbb{P}(X_T^{(k)} = X_T^{(l)} > 0)$, where $X_T^{(k)} = F_T(s_k, U_1, \ldots, U_T)$ and $X_T^{(l)} = F_T(s_l, U_1, \ldots, U_T)$, for some i.i.d. $[0, 1]$-uniform random variables $(U_t)_{t\geq 1}$.

It is up to the user of the algorithm to check if the Markov chain and the update function that they want to use satisfy this hypothesis of the 0-1 law. Now we proceed to prove that the output is indeed distributed according to π, under the assumption that the algorithm terminates.

Theorem 6.2. *[2, Theorem 10.1.] Let $(X_t)_t$ be an irreducible, aperiodic Markov chain over a finite state-space χ, with a transition matrix P and let π be its unique stationary distribution. Assume that the Propp-Wilson algorithm for this chain terminates with probability 1, and denote by S, J its output. Then the distribution of S is given by π.*

The proof of the theorem illustrates the main idea of the algorithm: by starting sufficiently early in the past (i.e. from initial time $-T_i$, where T_i is possibly quite large), we can ensure that the N chains we started from every possible initial state will all coalesce with a hypothetical chain that has been running "forever" (since time $-\infty$), and thus follows the stationary distribution π. Before we proceed to the proof, we state and prove a simple lemma:

Lemma 6.3. *Consider a discrete probability space $(\Omega, \mathbb{P})$. Let $A, B \subset \Omega$ be two events. Then $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$.*

Proof. We have

$$1 \geq \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

so

$$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + (\mathbb{P}(B) - 1) = \mathbb{P}(A) - \mathbb{P}(B^c).$$

□

Proof of the theorem 6.2. We will show that $\mid \mathbb{P}(S = s_k) - \pi_k \mid \leq \epsilon$, for all $\epsilon > 0$ and for all $k \in \{1, \ldots, N\}$. It then follows that $\mathbb{P}(S = s_k) = \pi_k$ for all $k \in \{1, \ldots, N\}$, meaning that S is distributed according to π.

Fix $\epsilon > 0$. Denote by T^* the earliest starting time:

$$T^* = T_J.$$

Recall that we obtain J from the algorithm, as the index of the largest starting time that we try. By assumption, $\mathbb{P}(T^* < \infty) = 1$. Then, we can obtain a limit: $\mathbb{P}(T^* < \infty) = \mathbb{P}(\bigcup_{i=0}^{\infty}\{T^* \leq T_i\}) = \lim_{i\to\infty}\mathbb{P}(T^* \leq T_i) = 1$, where the first equality comes from the fact that $(T_i)_{i\geq 0}$ are strictly increasing, and the second equality comes from the "continuity from below" property of countable unions. This means that there exists $i \geq 0$ such that $\mathbb{P}(T^* \leq T_i) \geq 1 - \epsilon$.

Let $(Y_t)_{-T_i \leq t \leq 0}$ be a P Markov chain started from time $-T_i$, which evolves according to the same update function ϕ and the same random variables $(U_t)_{-T_i+1\leq t\leq 0}$ as the chains $(X_t^{(1)})_{-T_i\leq t\leq 0}$, $\ldots$, $(X_t^{(N)})_{-T_i\leq t\leq 0}$. However, assume that its initial state Y_{-T_i} is chosen according to the stationary distribution π, independently of $(U_t)_t$. Then, for all $-T_i \leq t \leq 0$, Y_t has the distribution π. Notice that, no matter the initial state of Y_t, there will always be exactly one of the chains $(X_t^{(1)})_{-T_i\leq t\leq 0}$, $\ldots$, $(X_t^{(N)})_{-T_i\leq t\leq 0}$ that started from the same initial state as Y_t. Since they evolve according to the same update function ϕ and the same random variables $(U_t)_{-T_i+1\leq t\leq 0}$, these two chains will continue to evolve identically. This suggests that, if $(X_t^{(1)})_{-T_i\leq t\leq 0}$, $\ldots$, $(X_t^{(N)})_{-T_i\leq t\leq 0}$ all coalesce (meaning that $T^* \leq T_i$), then we have $S = Y_0$. Now we shall prove that this is indeed the case.

$$\begin{aligned}
\mathbb{P}(S = Y_0 \mid T^* \leq T_i) &= \sum_{1\leq k\leq N, \pi_k > 0} \mathbb{P}(S = Y_0 \mid Y_{-T_i} = s_k, T^* \leq T_i)\mathbb{P}(Y_{-T_i} = s_k \mid T^* \leq T_i) \\
&= \sum_{1\leq k\leq N, \pi_k > 0} \mathbb{P}(X_0^{(k)} = Y_0 \mid Y_{-T_i} = s_k, T^* \leq T_i)\mathbb{P}(Y_{-T_i} = s_k) \qquad (6.1) \\
&= \sum_{1\leq k\leq N, \pi_k > 0} \mathbb{P}[F_{T_i}(s_k, U_{-T_i+1}, \ldots, U_0) = F_{T_i}(Y_{-T_i}, U_{-T_i+1}, \ldots, U_0) \mid Y_{-T_i} = s_k, T^* \leq T_i]\pi_k \\
&= \sum_{1\leq k\leq N, \pi_k > 0} \mathbb{P}[F_{T_i}(s_k, U_{-T_i+1}, \ldots, U_0) = F_{T_i}(s_k, U_{-T_i+1}, \ldots, U_0) \mid T^* \leq T_i]\pi_k \\
&= \sum_{1\leq k\leq N, \pi_k > 0} \pi_k = 1.
\end{aligned}$$

Recall that π_k is the probability assigned to state s_k by the distribution π. Since Y_{-T_i} follows the distribution π, we have $\mathbb{P}(Y_{-T_i} = s_k) = \pi_k$, for all $k \in \{1, \ldots, N\}$. Furthermore, if $T^* \leq T_i$, then $X_0^{(1)} = \cdots = X_0^{(N)} = S$. This and the fact that Y_{-T_i} is chosen independently of any other random variables give us the equality (6.1). Then, Y_0 is obtained by applying the update function T_i times to Y_{-T_i}, using random variables $U_{-T_i+1}, \ldots, U_0$, so $Y_0 = F_{T_i}(Y_{-T_i}, U_{-T_i+1}, \ldots, U_0)$. Similarly, since $X_{-T_i}^{(k)} = s_k$, $X_0^{(k)} = F_{T_i}(s_k, U_{-T_i+1}, \ldots, U_0)$.

We finally obtain $\mathbb{P}(S = Y_0 \mid T^* \leq T_i) = 1$.

Now, we are ready to prove that $\mid \mathbb{P}(S = s_k) - \pi_k \mid \leq \epsilon$, for all $\epsilon > 0$ and for all $k \in \{1, \ldots, N\}$. First, we have

$$\mathbb{P}(S = Y_0) \geq \mathbb{P}(S = Y_0, T^* \leq T_i) = \mathbb{P}(S = Y_0 \mid T^* \leq T_i)\mathbb{P}(T^* \leq T_i) = \mathbb{P}(T^* \leq T_i) \geq 1 - \epsilon,$$

so $\mathbb{P}(S \neq Y_0) \leq \epsilon$. Then, for any $k \in \{1, \ldots N\}$ we have

$$\begin{aligned}
\epsilon &\geq \mathbb{P}(S \neq Y_0) \\
&\geq \mathbb{P}(S = s_k, Y_0 \neq s_k) \\
&\geq \mathbb{P}(S = s_k) - \mathbb{P}(Y_0 = s_k) \qquad (6.2) \\
&= \mathbb{P}(S = s_k) - \pi_k
\end{aligned}$$

and, similarly:

$$\begin{aligned}\epsilon &\geq \mathbb{P}(S \neq Y_0) \\ &\geq \mathbb{P}(S \neq s_k, Y_0 = s_k) \\ &\geq \mathbb{P}(Y_0 = s_k) - \mathbb{P}(S = s_k) \\ &= \pi_k - \mathbb{P}(S = s_k).\end{aligned} \tag{6.3}$$

Inequalities (6.2) and (6.3) are a consequence of the Lemma 6.3. Thus, for all $k \in \{1, \ldots, N\}$,

$$\mathbb{P}(S = s_k) = \pi_k, .$$

meaning that S is distributed according to π. □

Going back to the Section 4, we recall that the main purpose of Markov chain Monte Carlo algorithms, and the Propp-Wilson algorithm in particular, is to sample from a distribution over a large set χ. It is clear that the Propp-Wilson algorithm, as presented above, is hardly applicable to this problem, as it requires us to run $N =\mid \chi \mid$ chains. However, in some cases, it is possible to infer that coalescence has occurred by running only two chains. To be able to do this, we need to impose some sort of monotonicity of our Markov chains.

6.1 Monotonicity

[2, Section 11], [1]

Let $(X_t)_t$ be an irreducible, aperiodic Markov chain over a finite state-space χ, with a transition matrix P. Assume that there exists some partial ordering $\leq$ on the set χ, and a minimal and a maximal element for that ordering, denoted by $\hat{0}$ and $\hat{1}$, respectively. This is to say that, for all $s \in \chi$, we have $\hat{0} \leq s \leq \hat{1}$. Furthermore, suppose that there exists a valid update function for our chain $\phi : \chi \times [0,1] \to \chi$ which preserves this ordering:

$$\forall s_1, s_2 \in \chi, \forall U \in [0,1] : s_1 \leq s_2 \implies \phi(s_1, U) \leq \phi(s_2, U).$$

If these conditions are satisfied, than we do not need to simulate N chains (1 for each state) to obtain an output of the Propp-Wilson algorithm. In fact, it is enough to simulate only 2 chains, one starting from $\hat{0}$ and another one starting from $\hat{1}$. To see this, consider a grand coupling $(X_t^{(\hat{0})}, \ldots, X_t^{(\hat{1})})_{t \geq 0}$, evolving according to the update function ϕ and random variables $(U_t)_{t \geq 1}$. By an easy induction on t, we can show that for all $s \in \chi$ and for all $t \geq 0$, we have $X_t^{(\hat{0})} \leq X_t^{(s)} \leq X_t^{(\hat{1})}$ (with probability 1):

1) for $t = 0$ we have $X_0^{(s)} = s$ for all $s \in \chi$. Therefore, for all $s \in \chi$:

$$X_0^{(\hat{0})} = \hat{0} \leq X_0^{(s)} \leq \hat{1} = X_t^{(\hat{1})}$$

2) let $t \geq 0$ and assume that $X_t^{(\hat{0})} \leq X_t^{(s)} \leq X_t^{(\hat{1})}$ for all $s \in \chi$. Then, by the order-preserving property of ϕ, we get

$$X_{t+1}^{(\hat{0})} = \phi(X_t^{(\hat{0})}, U_{t+1}) \leq \phi(X_t^{(s)}, U_{t+1}) = X_{t+1}^{(s)} \leq \phi(X_t^{(\hat{1})}, U_{t+1}) = X_{t+1}^{(\hat{1})}$$

again, for all $s \in \chi$. This completes the induction proof.

We immediately get the following result: if, for some $t \geq 0$, $X_t^{(\hat{0})} = X_t^{(\hat{1})}$, then $X_t^{(s)} = X_t^{(\hat{0})} (= X_t^{(\hat{1})})$ for all $s \in \chi$. Indeed, since $X_t^{(\hat{0})} \leq X_t^{(s)}$ and $X_t^{(s)} \leq X_t^{(\hat{1})} = X_t^{(\hat{0})}$ then by the antisymmetry of $\leq$, it must be $X_t^{(\hat{0})} = X_t^{(s)}$. Therefore, the grand coupling has coalesced if and only if $X_t^{(\hat{0})}$ and $X_t^{(\hat{1})}$ have coalesced, meaning that it is enough to run the chains $X_t^{(\hat{0})}$ and $X_t^{(\hat{0})}$ (but from $-T_i$ to 0, for $i = 0, 1, 2, \ldots$) and return $X_0^{(\hat{0})}$ as soon as $X_0^{(\hat{0})} = X_0^{(\hat{1})}$.

6.2 Running time

We define the running time τ^* of our algorithm to be the number of calls we make to the update function. It is not hard to see that the running time is closely related to the backward coalescence time $\bar{\tau}$, defined in one of the previous sections. Indeed, if $(T_i)_{i\geq 0}$ is our sequence of (strictly increasing) starting times, then $\tau^* = n\sum_{i=0}^{J} T_i$, where $J = \min\{k \geq 0 : T_k \geq \bar{\tau}\}$ is the second argument of the output of our algorithm. The factor n in front of the sum corresponds to the number of chains from the grand coupling that we actually simulate. In the general setting, we have to run one chain for every possible state, so $n = |\chi|$ in that case. However, if the monotonicity assumptions are satisfied, we only need to run two chains: one starting from the bottom state $\hat{0}$ and another starting from the top state $\hat{1}$. We will work under the assumption that monotonicity is satisfied, since the Propp-Wilson algorithm is hardly applicable in practice if this is not the case.

Therefore, any bound on $\bar{\tau}$ can be translated into a bound on τ^*. Moreover, since the backward coalescence time $\bar{\tau}$ follows the same distribution as the coalescence time τ, we can focus on bounding τ.

Recall the notion of distance from stationarity $\bar{d}(t)$ at time t, defined in 2.20.

Proposition 6.4. *[1] Let l be the size of the largest totally ordered subset of χ. Then, for all $t \geq 0$*

$$\frac{\mathbb{P}(\tau > t)}{l} \leq \bar{d}(t) \leq \mathbb{P}(\tau > t).$$

Proof. Recall that $\rho_{(0)}$ (resp. $\rho_{(1)}$) is the probability distribution on χ which puts probability 1 on the state $\hat{0}$ (resp. $\hat{1}$). Furthermore, for all $s \in \chi$, let $h(s)$ denote the size of the largest totally ordered subset of χ that has s as its maximum. Consider our coupling $(X_t^{(\hat{0})}, X_t^{(\hat{1})})_{t\geq 0}$. By the order-preserving property of our update function, we have $X_t^{(\hat{0})} \leq X_t^{(\hat{1})}$ for all $t \geq 0$. Thus, we have $h(X_t^{(\hat{0})}) \leq h(X_t^{(\hat{1})})$ for all $t \geq 0$, with equality if and only if $X_t^{(\hat{0})} = X_t^{(\hat{1})}$. This allows us to define τ equivalently as $\tau = \min\{t \geq 0 \mid h(X_t^{(\hat{0})}) = h(X_t^{(\hat{1})})\}$.

Let us prove the first inequality of the proposition. Let $t \geq 0$. We have:

$$\begin{aligned}
\mathbb{P}(\tau > t) &= \mathbb{P}[h(X_t^{(\hat{0})}) \neq h(X_t^{(\hat{1})})] \\
&= \mathbb{P}[h(X_t^{(\hat{1})}) - h(X_t^{(\hat{0})}) \geq 1] \\
&= \sum_{k=1}^{\infty} \mathbb{P}[h(X_t^{(\hat{1})}) - h(X_t^{(\hat{0})}) = k] \\
&\leq \sum_{k=1}^{\infty} k\mathbb{P}[h(X_t^{(\hat{1})}) - h(X_t^{(\hat{0})}) = k] \\
&= \mathbb{E}[h(X_t^{(\hat{1})}) - h(X_t^{(\hat{0})})] = \mathbb{E}[h(X_t^{(\hat{1})})] - \mathbb{E}[h(X_t^{(\hat{0})})].
\end{aligned}$$

Notice that $X_t^{(\hat{1})}$ follows the distribution $\rho_{(1)}^t$ and $X_t^{(\hat{0})}$ follows $\rho_{(0)}^t$. Now, let (Y, Z) be any coupling of the distributions $\rho_{(0)}^t$ and $\rho_{(1)}^t$. Then Y is distributed identically as $X_t^{(\hat{0})}$, and Z is distributed identically as $X_t^{(\hat{1})}$, so we have $\mathbb{E}[h(X_t^{(\hat{1})})] = \mathbb{E}[h(Z)]$ and $\mathbb{E}[h(X_t^{(\hat{0})})] = \mathbb{E}[h(Y)]$. Thus, we have

$$\begin{aligned}
\mathbb{E}[h(X_t^{(\hat{1})})] - \mathbb{E}[h(X_t^{(\hat{0})})] &= \mathbb{E}[h(Z)] - \mathbb{E}[h(Y)] \\
&= \mathbb{E}[h(Z) - h(Y)] \\
&= 0\mathbb{P}[Z = Y] + \sum_{s,w\in\chi, s\neq w} \mathbb{P}[(Y, Z) = (s, w)](h(w) - h(s)) \\
&\leq \max_{s\in\chi} h(s) \sum_{s,w\in\chi, s\neq w} \mathbb{P}[(Y, Z) = (s, w)] \\
&= \max_{s\in\chi} h(s)\mathbb{P}[Z \neq Y] = l\mathbb{P}[Z \neq Y].
\end{aligned}$$

Since this holds for any coupling (Y, Z) of $\rho_{(0)}^t$ and $\rho_{(1)}^t$, then, by 2.15,

$$\mathbb{E}[h(X_t^{(\hat{1})})] - \mathbb{E}[h(X_t^{(\hat{0})})] \leq l \inf_{Y,Z \sim \rho_0,\rho_1} \mathbb{P}[Z \neq Y] = l d_{TV}(\rho_{(0)}^t, \rho_{(1)}^t)$$

Finally, we obtain $\mathbb{P}(\tau > t) \leq l d_{TV}(\rho_{(0)}^t, \rho_{(1)}^t) \leq l\bar{d}(t)$.

Now, we prove the second inequality. Let μ, ν be two probability distributions over χ. Let $(Y_t, Z_t)_{t\geq 0}$ be a coupling of chains evolving according to the same update function ϕ and random variables $(U_t)_{t\geq 0}$ as our coupling $(X_t^{(\hat{0})}, X_t^{(\hat{1})})_{t\geq 0}$. Assume that Y_0 is chosen according to μ and Z_0 is chosen according to ν, independently from each other. Then (Y_t, Z_t) is a coupling of the distributions μP^t and νP^t. Moreover, it is not hard to see that $\mathbb{P}(\tau > t \mid Y_t \neq Z_t) = 1$:

$$\begin{aligned}
\mathbb{P}(\tau > t \mid Y_t \neq Z_t) &= \sum_{j,k\in\{1,\ldots,N\},\mu_j\neq 0,\nu_k\neq 0} \mathbb{P}[\tau > t \mid Y_t \neq Z_t, (Y_0, Z_0) = (s_j, s_k)]\mu_j\nu_k \\
&= \sum_{j,k\in\{1,\ldots,N\},\mu_j\neq 0,\nu_k\neq 0} \mathbb{P}[\tau > t \mid F(s_j, U_1, \ldots, U_t) \neq F(s_k, U_1, \ldots, U_t)]\mu_j\nu_k \\
&= \sum_{j,k\in\{1,\ldots,N\},\mu_j\neq 0,\nu_k\neq 0} \mathbb{P}[\tau > t \mid X_t^{(j)} \neq X_t^{(k)}]\mu_j\nu_k \\
&= \sum_{j,k\in\{1,\ldots,N\},\mu_j\neq 0,\nu_k\neq 0} \mu_j\nu_k = 1
\end{aligned}$$

This yields

$$\mathbb{P}(\tau > t) \geq \mathbb{P}(\tau > t \mid Y_t \neq Z_t)\mathbb{P}(Y_t \neq Z_t) = \mathbb{P}(Y_t \neq Z_t) \geq d_{TV}(\mu P^t, \nu P^t).$$

For any two distributions μ and ν over χ, $\mathbb{P}(\tau > t) \geq d_{TV}(\mu^t, \nu^t)$. Thus

$$\mathbb{P}(\tau > t) \geq \sup_{(\mu,\nu)\in\mathcal{D}(\chi)^2} d_{TV}(\mu P^t, \nu P^t) = \bar{d}(t).$$

□

Next, we show that $\mathbb{P}(\tau > t)$ is sub-multiplicative.

Proposition 6.5. *[1] Let $t_1, t_2 \in \mathbb{N}$. Then*

$$\mathbb{P}(\tau > t_1 + t_2) \leq \mathbb{P}(\tau > t_1)\mathbb{P}(\tau > t_1)$$

Proof. We have $\mathbb{P}(\tau > t_1 + t_2) = \mathbb{P}(\tau > t_1 + t_2 \text{ and } \tau > t_1) = \mathbb{P}(\tau > t_1 + t_2 \mid \tau > t_1)\mathbb{P}(\tau > t_1)$. Now we prove that $\mathbb{P}(\tau > t_1 + t_2 \mid \tau > t_1) \leq \mathbb{P}(\tau > t_2)$.

First, recall that τ is defined as the time of coalescence of a grand coupling $(X_t^{(\hat{0})}, \ldots X_t^{(\hat{1})})_{t\geq 0}$. Recall also that the grand coupling has coalesced if and only if the coupling $(X_t^{(\hat{0})}, X_t^{(\hat{1})})_{t\geq 0}$ has coalesced, so we may equivalently define τ as the coalescence time of the coupling $(X_t^{(\hat{0})}, X_t^{(\hat{1})})_{t\geq 0}$. This coupling evolves according to an order-preserving update function ϕ and a sequence of random variables $(U_t)_{t\geq 1}$. Let $(Y_t^{(\hat{0})}, Y_t^{(\hat{1})})_{t\geq 0}$ be another Markovian coupling started from $(\hat{0}, \hat{1})$ but this one evolving according to $(U_t)_{t\geq t_1+1}$ (and the same update function ϕ). Note that

$$\hat{0} = Y_0^{(\hat{0})} \leq X_{t_1}^{(\hat{0})} \leq X_{t_1}^{(\hat{1})} \leq Y_0^{(\hat{1})} = \hat{1}.$$

Moreover, $Y_t^{(\hat{0})}, Y_t^{(\hat{1})}, X_{t_1+t}^{(\hat{0})}, X_{t_1+t}^{(\hat{1})}$ are updated using a common random variable U_{t_1+t+1} (and the same update function ϕ), for all $t \geq 0$. So, by the order-preserving property of ϕ, we have

$$Y_t^{(\hat{0})} \leq X_{t_1+t}^{(\hat{0})} \leq X_{t_1+t}^{(\hat{1})} \leq Y_t^{(\hat{1})} \tag{6.4}$$

for all $t \geq 0$. This means that "$Y_t^{(\hat{0})} = Y_t^{(\hat{1})}$" implies "$X_{t_1+t}^{(\hat{0})} = X_{t_1+t}^{(\hat{1})}$". Conceptually, one could think of $Y_t^{(\hat{0})}$ and $Y_t^{(\hat{1})}$ as of two chains started at time t_1 in order to 'control' the chains $X_t^{(\hat{0})}$ and $X_t^{(\hat{1})}$ after time t_1. This is useful because it gives us some information about $X_{t_1+t}^{(\hat{0})}$ and $X_{t_1+t}^{(\hat{1})}$, which will soon become clearer.

Let $\tau' = \min\{t \geq 0 : Y_t^{(\hat{0})} = Y_t^{(\hat{1})}\}$ be the time of coalescence of $Y_t^{(\hat{0})}$ and $Y_t^{(\hat{1})}$. Then τ' and τ are distributed identically. By definition of τ and τ' we get the equivalences

$$\tau > t \iff X_t^{(\hat{0})} \neq X_t^{(\hat{1})} \tag{6.5}$$

$$\tau' > t \iff Y_t^{(\hat{0})} \neq Y_t^{(\hat{1})}. \tag{6.6}$$

for all $t \geq 1$. Moreover, by (6.4) we have the implication: $X_{t_1+t}^{(\hat{0})} \neq X_{t_1+t}^{(\hat{1})} \implies Y_t^{(\hat{0})} \neq Y_t^{(\hat{1})}$. Putting all these together, we get

$$\tau > t_1 + t_2 \implies \tau' > t_2$$

meaning that $\mathbb{P}(\tau > t_1 + t_2 \mid \tau > t_1) \leq \mathbb{P}(\tau' > t_2 \mid \tau > t_1)$.

By equivalences (6.5) and (6.6), $\mathbb{P}(\tau' > t_2 \mid \tau > t_1) = \mathbb{P}(Y_{t_2}^{(\hat{0})} \neq Y_{t_2}^{(\hat{1})} \mid X_{t_1}^{(\hat{0})} \neq X_{t_1}^{(\hat{1})})$. However, $Y_{t_2}^{(\hat{0})}$ and $Y_{t_2}^{(\hat{1})}$ are fully determined by $(U_t)_{t_1+1 \leq t \leq t_1+t_2}$:

$$Y_{t_2}^{(\hat{0})} = F_{t_2}(\hat{0}, U_{t_1+1}, \ldots, U_{t_1+t_2})$$

$$Y_{t_2}^{(\hat{1})} = F_{t_2}(\hat{1}, U_{t_1+1}, \ldots, U_{t_1+t_2}).$$

Likewise, $X_{t_1}^{(\hat{0})}$ and $X_{t_1}^{(\hat{1})}$ are fully determined by $(U_t)_{1 \leq t \leq t_1}$:

$$X_{t_1}^{(\hat{0})} = F_{t_1}(\hat{0}, U_1, \ldots, U_{t_1})$$

$$X_{t_1}^{(\hat{1})} = F_{t_1}(\hat{1}, U_1, \ldots, U_{t_1}).$$

Since $(U_t)_{t \geq 0}$ are all independent, so are the events $\{Y_{t_2}^{(\hat{0})} \neq Y_{t_2}^{(\hat{1})}\}$ and $\{X_{t_1}^{(\hat{0})} \neq X_{t_1}^{(\hat{1})}\}$. Thus,

$$\begin{aligned}
\mathbb{P}(\tau > t_1 + t_2 \mid \tau > t_1) &\leq \mathbb{P}(\tau' > t_2 \mid \tau > t_1) \\
&= \mathbb{P}(Y_{t_2}^{(\hat{0})} \neq Y_{t_2}^{(\hat{1})} \mid X_{t_1}^{(\hat{0})} \neq X_{t_1}^{(\hat{1})}) \\
&= \mathbb{P}(Y_{t_2}^{(\hat{0})} \neq Y_{t_2}^{(\hat{1})}) \\
&= \mathbb{P}(\tau' > t_2) = \mathbb{P}(\tau > t_2).
\end{aligned}$$

The desired inequality follows immediately. □

Remark 6.6. We have, for any $t, k \in \mathbb{N}$

$$\mathbb{P}(\tau > tk) \leq \mathbb{P}(\tau > t)^k.$$

This follows by a simple induction on k.

The following lemma gives a bound the expectation of τ.

Lemma 6.7. *[1] For any integer t large enough so that $\mathbb{P}(\tau \leq t) > 0$,*

$$t\mathbb{P}(\tau > t) \leq \mathbb{E}(\tau) \leq \frac{t}{\mathbb{P}(\tau \leq t)}.$$

Proof. Since τ takes values in $\mathbb{N}$, we have

$$\mathbb{E}(\tau) = \sum_{k=1}^{\infty} k\mathbb{P}(\tau = k) \geq \sum_{k=t}^{\infty} k\mathbb{P}(\tau = k) \geq t\sum_{k=t}^{\infty} \mathbb{P}(\tau = k) \geq t\mathbb{P}(\tau > t)$$

For the second inequality, since τ takes values in $\mathbb{N}$, we can write the expected value of τ as $\sum_{k=0}^{\infty} \mathbb{P}(\tau > k)$. Thus,

$$\mathbb{E}(\tau) = \sum_{k=0}^{\infty} \mathbb{P}(\tau > k) = \sum_{k=0}^{\infty}\sum_{j=0}^{t-1} \mathbb{P}(\tau > tk + j) \leq \sum_{k=0}^{\infty}\sum_{j=0}^{t-1} \mathbb{P}(\tau > tk) = t\sum_{k=0}^{\infty} \mathbb{P}(\tau > tk).$$

By submultiplicativity, $\mathbb{P}(\tau > tk) \leq \mathbb{P}(\tau > t)^k$ for all $k \geq 0$. Let $\alpha = \mathbb{P}(\tau > t)$ (by assumption $\alpha < 1$). Then

$$\mathbb{E}(\tau) \leq t\sum_{k=0}^{\infty} \alpha^k = \frac{t}{1-\alpha} = \frac{t}{\mathbb{P}(\tau \leq t)}.$$

□

The final result is a useful relation between the expected coalescence time and the mixing time of a Markov chain. Recall the definition of the mixing time: $t_{mix} = \min\{t \geq 0 \mid \bar{d}(t) \leq 1/4\}$.

Theorem 6.8. *[1] Let l be the size of the largest totally ordered subset of our state space χ. Then*

$$\mathbb{E}(\tau) \leq (3 + \log l)t_{mix}.$$

Proof. Let $t = t_{mix}\lceil\frac{1+\log l}{2}\rceil$. By submultiplicativity of $\bar{d}$ (from Proposition 2.22, we have

$$\bar{d}(t) \leq \bar{d}(t_{mix})^{\lceil\frac{1+\log l}{2}\rceil} \leq \left(\frac{1}{4}\right)^{\frac{1+\log l}{2}} = \frac{1}{2l}.$$

Then, by proposition, 6.4 $\mathbb{P}(\tau > t) \leq l\bar{d}(t) = \frac{1}{2}$, so $\mathbb{P}(\tau \leq t) \geq \frac{1}{2}$. Finally, by lemma 6.7, we obtain

$$\mathbb{E}(\tau) \leq \frac{t}{\mathbb{P}(\tau \leq t)} \leq 2t = 2\left\lceil\frac{1+\log l}{2}\right\rceil t_{mix} \leq (3 + \log l)t_{mix}.$$

□

In practice, we usually choose starting times $T_i = 2^i$ for $i \geq 0$. This way, the running time τ^* is at most a constant times the backward coalescence time $\bar{\tau}$. Indeed, we have $\tau^* = 2\sum_{i=0}^{J} T_i = 2\sum_{i=0}^{J} 2^i = 2(2^{J+1} - 1) \leq 2^{J+2}$, where $J = \min\{i \geq 0 \mid T_i \geq \bar{\tau}\} = \min\{i \geq 0 \mid 2^i \geq \bar{\tau}\}$, so $2^{J-1} < \bar{\tau}$. Therefore, $\tau^* \leq 2^{J+2} \leq 8\bar{\tau}$.

7 Ising model

Consider a graph $G = (V, E)$. The Ising model on G represents a probability distribution over the set $\chi = \{-1, 1\}^V$ (mappings from V to $\{-1, 1\}$). From a physical interpretation, vertices V can represent atoms in magnetic field, and edges E connect atoms that are close together. The atoms can have spin -1 or $+1$.

Elements of χ are called *configurations*. For any configuration $\eta \in \chi$, we define the *energy* of η:

$$H(\eta) = -\sum_{(v,w)\in E} \eta(v)\eta(w).$$

When atoms that are close to each other have different spins, the energy of the configuration increases. Then, for a given parameter $\beta \geq 0$, we define $\pi^{(\beta)}$ to be the Gibbs distribution over χ:

$$\pi^{(\beta)}(\eta) = \frac{e^{-\beta H(\eta)}}{Z(\beta)}.$$

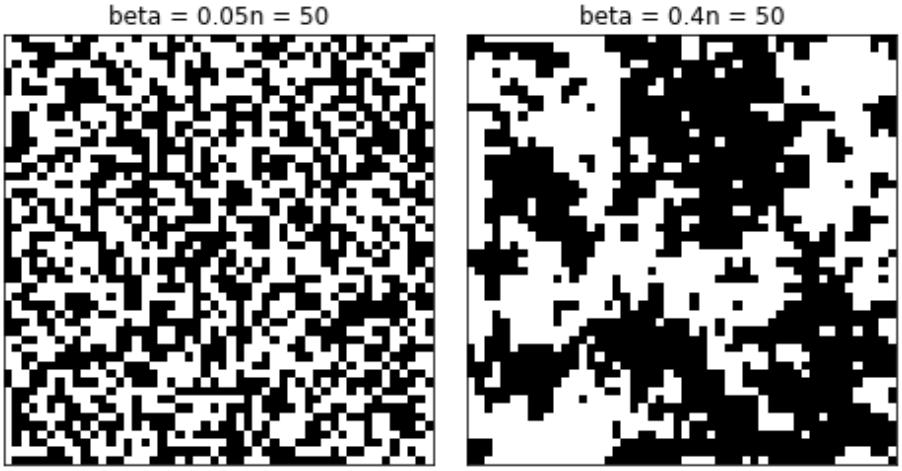

Figure 1: Configurations drawn from the Gibbs distribution on a 50×50 square lattice, at inverse temperature 0.05 (left) and 0.4 (right), using our implementation [5] of the Propp-Wilson algorithm

$\pi^{(\beta)}(\eta)$ is the probability $\pi^{(\beta)}$ puts on η, for any $\eta \in \chi$. $Z(\beta)$ is just a normalizing constant: $Z(\beta) = \sum_{\eta \in \chi} e^{-\beta H(\eta)}$. Parameter β is called the *inverse temperature*. For $\beta > 0$, the Gibbs distribution puts more probability on states with lower energy. However, for $\beta = 0$ (corresponding to 'infinite' temperature) the Gibbs distribution is uniform and energy plays no role. Figure 1 illustrates this: $\beta = 0.05$ corresponds to a very high temperature and the spins seem to be uniformly distributed, while $\beta = 0.4$ is a lower temperature, and some areas of constant spin tend to form. For a more detailed physics interpretation, see, for example [4].

We now focus on finding an efficient way of sampling from the Gibbs distribution. First, notice that the size of the set χ grows exponentially with the number of vertices in our Ising model:

$$| \chi | = 2^{|V|}.$$

Thus, the size of χ gets quite large even for moderate sizes of V, meaning that, to sample from a distribution over χ, one might want to use Markov chain Monte Carlo. For that purpose, one constructs a Markov chain $(X_t)_{t \geq 0}$ over χ whose stationary distribution is the Gibbs distribution. For clarity, we first specify how we simulate such a chain, and then we specify its exact transition probabilities (i.e. transition matrix), since they are less informative.

For the initial state, we can choose any configuration from χ, since the convergence result does not depend on the initial state. For example, set $X_0 = \xi$, where $\xi \in \chi$ is the configuration where all vertices get the value 1 (in practice, there might be some 'better' choices for the initial state, but this will suffice for us). Then, assume we have access to a sequence $(v_t)_{t \geq 1}$ of vertices, such that each v_t is chosen uniformly at random from V, independently from all the previous choices (i.e. $(v_t)_{t \geq 1}$ are i.i.d. random variables). Moreover, assume we have access to i.i.d. uniform random variables $(W_t)_{t \geq 0}$, on interval $[0, 1]$, independent from $(v_t)_{t \geq 1}$. Keep in mind that $(W_t)_{t \geq 0}$ are not equivalent to the uniform random variables we generally use to simulate transitions (usually denoted by $(U_t)_t$) by passing them to an update function. Indeed, in this case, $(W_t)_{t \geq 0}$ will not be the only source of randomness we use in our simulation, unlike in the most general setting. The chain will run as follows:

For all $t \geq 0$:

Set

$$X_{t+1}(v_{t+1}) = \begin{cases} 1 & \text{if } W_{t+1} < \frac{e^{2\beta(k_+(v_{t+1}, X_t) - k_-(v_{t+1}, X_t))}}{e^{2\beta(k_+(v_{t+1}, X_t) - k_-(v_{t+1}, X_t))} + 1} \\ -1 & \text{otherwise} \end{cases}$$

$$X_{t+1}(w) = X_t(w) \text{ for all } w \in V \setminus \{v_{t+1}\}$$

where $k_+(v,\eta) =\mid \{w \in V \mid (v,w) \in E, \eta(w) = 1\} \mid$ and $k_-(v,\eta) =\mid \{w \in V \mid (v,w) \in E, \eta(w) = -1\} \mid$ denote the number of neighbours of a vertex v with $+1$ and -1 spin respectively, in a configuration $\eta \in \chi$.

We formalize this transition procedure by defining

$$\Phi : \chi \times V \times [0,1] \to \chi,$$

$$\Phi(\eta, v_t, W_t)(v) = \begin{cases} \eta(v) & \text{if } v \neq v_t \\ 1 & \text{if } v = v_t \text{ and } W_t < \frac{e^{2\beta(k_+(v_t,\eta)-k_-(v_t,\eta))}}{e^{2\beta(k_+(v_t,\eta)-k_-(v_t,\eta))}+1} \\ -1 & \text{if } v = v_t \text{ and } W_t \geq \frac{e^{2\beta(k_+(v_t,\eta)-k_-(v_t,\eta))}}{e^{2\beta(k_+(v_t,\eta)-k_-(v_t,\eta))}+1} \end{cases} \tag{7.1}$$

Then, we simply set $X_{t+1} = \Phi(X_t, v_{t+1}, W_{t+1})$ for all $t \geq 0$.

Since X_{t+1} is a function of X_t, W_{t+1}, and v_{t+1}; and W_{t+1} and v_{t+1} are independent of any previous variables $W_t, \ldots, W_1$ and $v_t, \ldots, v_1$, then we can show, analogously to Proposition 3.3, that $(X_t)_{t\geq 0}$ is a Markov chain.

Now, we specify the transition probabilities. For any $\eta, \epsilon \in \chi$, define

$$P_{\eta,\epsilon} := \mathbb{P}(X_{t+1} = \epsilon \mid X_t = \eta)$$

$$c(\eta, \epsilon) =\mid \{v \in V \mid \eta(v) \neq \epsilon(v)\} \mid .$$

$P_{\eta,\epsilon}$ is the probability of transitioning from η to ϵ (analogous to an entry of the transition matrix) and $c(\eta, \epsilon)$ is the number of vertices in which η and ϵ disagree.

The following proposition specifies the transition probabilities of a chain defined the procedure (7.1).

Proposition 7.1. *Assume that $(X_t)_{t\geq 0}$ is a Markov chain evolving according to transition procedure* (7.1)*. Then, its transition probabilities are given by*

$$P_{\eta,\epsilon} = \begin{cases} \frac{1}{|V|} \sum_{v\in V} \frac{e^{\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} & \textit{if } c(\eta,\epsilon) = 0 \\ \frac{1}{|V|} \frac{e^{-\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} & \textit{if } c(\eta,\epsilon) = 1 \textit{ and } \eta(v) \neq \epsilon(v) \\ 0 \textit{ otherwise.} \end{cases}$$

Proof. Notice first that at every step, we modify at most one vertex, so $P_{\eta,\epsilon} = 0$ if $c(\eta, \epsilon) \geq 2$. Assume now that $c(\eta, \epsilon) = 1$. let v be the vertex in which they disagree: $\eta(v) \neq \epsilon(v)$. Then, if $\eta(v) = 1$, we have:

$$\begin{aligned} P_{\eta,\epsilon} &= \mathbb{P}(X_{t+1} = \epsilon \mid X_t = \eta) \\ &= \mathbb{P}(X_{t+1}(v_{t+1}) = -1 \mid v_{t+1} = v, X_t = \eta)\mathbb{P}(v_{t+1} = v \mid X_t = \eta) \\ &= \frac{1}{\mid V \mid}(1 - \frac{e^{2\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{2\beta(k_+(v,\eta)-k_-(v,\eta))} + 1}) \\ &= \frac{1}{\mid V \mid} \frac{1}{e^{2\beta(k_+(v,\eta)-k_-(v,\eta))} + 1} \\ &= \frac{1}{\mid V \mid} \frac{e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))} + e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} \\ &= \frac{1}{\mid V \mid} \frac{e^{-\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))} + e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}}. \end{aligned}$$

If $\eta(v) = -1$, one computes in a similar way

$$\begin{aligned} P_{\eta,\epsilon} &= \mathbb{P}(X_{t+1}(v_{t+1}) = 1 \mid v_{t+1} = v, X_t = \eta)\mathbb{P}(v_{t+1} = v \mid X_t = \eta) \\ &= \frac{1}{\mid V \mid} \frac{e^{2\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{2\beta(k_+(v,\eta)-k_-(v,\eta))}+1} \\ &= \frac{1}{\mid V \mid} \frac{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} \\ &= \frac{1}{\mid V \mid} \frac{e^{-\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} \end{aligned}$$

so we conclude that $P_{\eta,\epsilon} = \frac{1}{|V|} \frac{e^{-\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}}$ if $c(\eta,\epsilon) = 1$ and $\eta(v) \neq \epsilon(v)$. Finally, assume that $c(\eta,\epsilon) = 0$, meaning that $\epsilon = \eta$, so $P_{\eta,\epsilon} = P_{\eta,\eta}$. Then

$$\begin{aligned} P_{\eta,\eta} &= \mathbb{P}(X_{t+1} = \eta \mid X_t = \eta) \\ &= \sum_{v \in V} \mathbb{P}(X_{t+1}(v) = \eta(v) \mid v_{t+1} = v, X_t = \eta)\mathbb{P}(v_{t+1} = v \mid X_t = \eta) \\ &= \frac{1}{\mid V \mid} \sum_{v \in V} \mathbb{P}(X_{t+1}(v) = \eta(v) \mid v_{t+1} = v, X_t = \eta). \end{aligned}$$

Consider $\mathbb{P}(X_{t+1}(v) = \eta(v) \mid v_{t+1} = v, X_t = \eta)$ for some $v \in V$. Considering the cases $\eta(v) = 1$ and $\eta(v) = -1$, and performing similar computations as above, we obtain

$$\mathbb{P}(X_{t+1}(v) = \eta(v) \mid v_{t+1} = v, X_t = \eta) = \frac{e^{\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}}$$

. This leads to

$$\begin{aligned} P_{\eta,\eta} &= \frac{1}{\mid V \mid} \sum_{v \in V} \mathbb{P}(X_{t+1}(v) = \eta(v) \mid v_{t+1} = v, X_t = \eta) \\ &= \frac{1}{\mid V \mid} \sum_{v \in V} \frac{e^{\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}}. \end{aligned}$$

□

The following proposition (see [3, page 45, 3.12] or [2, Example 11.2. and Problem 11.3.]), which we state without proof, tells us that the Markov chain we presented above indeed has the Gibbs distribution as its stationary distribution.

Proposition 7.2. *Let $(X_t)_{t\geq 0}$ be a Markov chain over $\chi = \{-1,1\}^V$, with transition probabilities*

$$P_{\eta,\epsilon} = \begin{cases} \frac{1}{|V|} \sum_{v \in V} \frac{e^{\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} & \text{if } c(\eta,\epsilon) = 0 \\ \frac{1}{|V|} \frac{e^{-\eta(v)\beta(k_+(v,\eta)-k_-(v,\eta))}}{e^{\beta(k_+(v,\eta)-k_-(v,\eta))}+e^{-\beta(k_+(v,\eta)-k_-(v,\eta))}} & \text{if } c(\eta,\epsilon) = 1 \text{ and } \eta(v) \neq \epsilon(v) \\ 0 & \text{otherwise} \end{cases}$$

for $\eta, \epsilon \in \chi$. Then, $(X_t)_{t\geq 0}$ is irreducible, aperiodic, and its stationary distribution is the Gibbs distribution $\pi^{(\beta)}$. We call $(X_t)_{t\geq 0}$ the Glauber dynamics for $\pi^{(\beta)}$.

Thus, by 2.18, the distribution of X_t converges (in total variation distance) to the Gibbs distribution $\pi^{(\beta)}$ as $t \to \infty$. We can, therefore, use the chain $(X_t)_{t\geq 0}$ in a MCMC algorithm to draw approximate samples from $\pi^{(\beta)}$. Moreover, we can use the Propp-Wilson algorithm to obtain exact samples from $\pi^{(\beta)}$. For that, we need to define some partial ordering $\leq$ on χ, and show that our transition procedure Φ, which can be seen as a special instance of an update function, preserves this ordering. The ordering we define is quite simple: for all $\epsilon, \eta \in \chi$, $\epsilon \leq \eta$ if and only

if $\epsilon(v) \leq \eta(v)$ for all $v \in V$. We define $\hat{1}$ to be the configuration with all vertices set to 1 (i.e. $\hat{1}(v) = 1$ for all $v \in V$) and $\hat{0}$ to be the configuration with all vertices set to -1(i.e. $\hat{0}(v) = -1$ for all $v \in V$). Then, it is clear that for all $\eta \in \chi$, we have

$$\hat{0} \leq \eta \leq \hat{1},$$

so $\hat{0}$ is our minimal state and $\hat{1}$ is our maximal state. Next, we show that the transition procedure described in Φ preserves this order.

Proposition 7.3. *Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ be two Markov chains over χ, both evolving according to Φ, with uniform random variables $(W_t)_{t\geq 1}$ and uniform random vertices $(v_t)_{t\geq 1}$ (with possibly different starting states). For all $t \geq 0$,*

$$\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t) = 1.$$

Proof. We have

$$\begin{aligned}\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t) &= \sum_{w\in V} \mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w)\mathbb{P}(v_{t+1} = w) \\ &= \frac{1}{|V|} \sum_{w\in V} \mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w).\end{aligned}$$

Let $w \in V$ and consider $\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w)$. We have

$$\begin{aligned}&\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w) = \\ &= \mathbb{P}(X_{t+1}(w) \leq Y_{t+1}(w) \mid X_t \leq Y_t, v_{t+1} = w) + \sum_{v\in V\setminus\{w\}} \mathbb{P}(X_{t+1}(v) \leq Y_{t+1}(v) \mid X_t \leq Y_t, v_{t+1} = w)\end{aligned}$$

$$= \mathbb{P}(X_{t+1}(w) \leq Y_{t+1}(w) \mid X_t \leq Y_t, v_{t+1} = w)$$

$$\begin{aligned}&= \mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = 1, 1 \mid X_t \leq Y_t, v_{t+1} = w) \\ &+ \mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, 1 \mid X_t \leq Y_t, v_{t+1} = w) \\ &+ \mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, -1 \mid X_t \leq Y_t, v_{t+1} = w).\end{aligned}$$

Notice that, if $X_t \leq Y_t$, then, by definition of k_+ and k_- in the procedure (7.1), for all $v \in V$, we have $k_+(v, X_t) \leq k_+(v, Y_t)$ and $k_-(v, X_t) \geq k_-(v, Y_t)$, so

$$\frac{1}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1} \geq \frac{1}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1}$$

and

$$\frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1} \leq \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1}.$$

Then, we have

$$\begin{aligned}&\mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = 1, 1 \mid X_t \leq Y_t, v_{t+1} = w) = \\ &= \mathbb{P}(W_{t+1} < \min\left(\frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1}, \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1}\right) \mid X_t \leq Y_t, v_{t+1} = w) \\ &= \mathbb{P}(W_{t+1} < \frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1} \mid X_t \leq Y_t) \qquad (7.2) \\ &= \frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1}.\end{aligned}$$

Similarly,

$$\mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, 1 \mid X_t \leq Y_t, v_{t+1} = w) =$$

$$= \mathbb{P}(\frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1} \leq W_{t+1} < \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1} \mid X_t \leq Y_t, v_{t+1} = w)$$
$$= \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1} - \frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1}. \tag{7.3}$$

And, finally,

$$\mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, -1 \mid X_t \leq Y_t, v_{t+1} = w) =$$

$$= \mathbb{P}(W_{t+1} \geq \max\left(\frac{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}}{e^{2\beta(k_+(w,X_t)-k_-(w,X_t))}+1}, \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1}\right) \mid X_t \leq Y_t, v_{t+1} = w)$$
$$= \mathbb{P}(W_{t+1} \geq \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1} \mid X_t \leq Y_t) \tag{7.4}$$
$$= 1 - \frac{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}}{e^{2\beta(k_+(w,Y_t)-k_-(w,Y_t))}+1}.$$

In (7.2), (7.3) and (7.4), we use the fact that $(W_t)_t$ are independent from $(v_t)_t$.

This yields

$$\begin{aligned}
&\mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = 1, 1 \mid X_t \leq Y_t, v_{t+1} = w) \\
&+ \mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, 1 \mid X_t \leq Y_t, v_{t+1} = w) \\
&+ \mathbb{P}(X_{t+1}(w), Y_{t+1}(w) = -1, -1 \mid X_t \leq Y_t, v_{t+1} = w) \\
&= 1.
\end{aligned}$$

Thus, we conclude that $\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w) = 1$. Since this holds for any $w \in V$, we get

$$\mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t) = \frac{1}{|V|} \sum_{w \in V} \mathbb{P}(X_{t+1} \leq Y_{t+1} \mid X_t \leq Y_t, v_{t+1} = w) = 1.$$

Thus, the transition procedure Φ preserves ordering. □

By Proposition 7.3 we have monotonicity, so we can apply the Propp-Wilson algorithm to the Ising model, by Section 6.1. Here is its pseudo code

However, to make sure that the algorithm terminates, we need to check that our transition procedure verifies the conditions of the 0-1 law.

7.1 Ising model and 0-1 law

Let $G = (V, E)$ be a graph. Consider an Ising model at inverse temperature β on G. For any configuration $\eta \in \{-1, 1\}^V$, let $(X_t^{(\eta)})_{t \geq 0}$ be a Glauber dynamics started from η (i.e. $X_0^{(\eta)} = \eta$), generated using the transition procedure Φ (7.1). The following theorem states that a grand coupling over $\{-1, 1\}^V$ will always coalesce into one chain if we run it for sufficiently long.

Theorem 7.4. *Let $N = |\{-1, 1\}^V|$. Consider a grand coupling $(X_t^{(\eta_1)}, \ldots, X_t^{(\eta_N)})$ over $\{-1, 1\}^V$, evolving according to transition procedure Φ, uniform random variables $(W_t)_{t \geq 1}$ and uniform random vertices $(v_t)_{t \geq 1}$. Let $\tau = min\{t > 0 \mid X_t^{(\eta_1)} = \cdots = X_t^{(\eta_N)}\}$ be its coalescence time. Then $\mathbb{P}(\tau < \infty) = 1$.*

Algorithm 2 Propp-Wilson algorithm for the Ising model

Require: i.i.d. uniform random variables $(W_t)_{t\leq 0}$ on interval $[0,1]$; vertices $(v_t)_{t\leq 0}$ chosen uniformly at random from the set V, each independent of the previous choices and of $(W_t)_{t\leq 0}$; strictly increasing starting times $(T_i)_{i\geq 0}$

$j \leftarrow 0$
$coalesced \leftarrow False$
while not *coalesced* **do**
  $X^{(\hat{1})} \leftarrow \hat{1}$
  $X^{(\hat{0})} \leftarrow \hat{0}$
  for $t = -T_j + 1, -T_j + 2, \ldots, 0$ **do**
    $X^{(\hat{1})} \leftarrow \Phi(X^{(\hat{1})}, v_t, W_t)$
    $X^{(\hat{0})} \leftarrow \Phi(X^{(\hat{0})}, v_t, W_t)$
  end for
  if $X^{(\hat{0})} = X^{(\hat{1})}$ **then**
    $result \leftarrow X^{(\hat{1})}$
    $coalesced \leftarrow True$
  end if
  $j \leftarrow j + 1$
end while
$j \leftarrow j - 1$
return $result, j$

Proof. Let $n = |V|$. We prove that for any $\eta, \epsilon \in \{-1,1\}^V$, we have $\mathbb{P}(X_n^{(\eta)} = X_n^{(\epsilon)}) > 0$. The result then follows directly from the 0-1 law (5.3). Let $\eta, \epsilon \in \{-1,1\}^V$. We recall that $X_t^{(\eta)}$ and $X_t^{(\epsilon)}$ are updated according to the following rule:
For all $t \geq 0$:
Set

$$X_{t+1}^{(\eta)}(v_{t+1}) = \begin{cases} 1 & \text{if } W_{t+1} < \frac{e^{2\beta(k_+(v_{t+1}, X_t^{(\eta)}) - k_-(v_{t+1}, X_t^{(\eta)}))}}{e^{2\beta(k_+(v_{t+1}, X_t^{(\eta)}) - k_-(v_{t+1}, X_t^{(\eta)}))} + 1} \\ -1 & \text{otherwise} \end{cases}$$

$$X_{t+1}^{(\epsilon)}(v_{t+1}) = \begin{cases} 1 & \text{if } W_{t+1} < \frac{e^{2\beta(k_+(v_{t+1}, X_t^{(\epsilon)}) - k_-(v_{t+1}, X_t^{(\epsilon)}))}}{e^{2\beta(k_+(v_{t+1}, X_t^{(\epsilon)}) - k_-(v_{t+1}, X_t^{(\epsilon)}))} + 1} \\ -1 & \text{otherwise} \end{cases}$$

$$X_{t+1}^{(\eta)}(v) = X_t^{(\eta)}(v) \text{ and } X_{t+1}^{(\epsilon)}(v) = X_t(v)^{(\epsilon)} \text{ for all } v \neq v_{t+1}.$$

Now, we define random variables $\rho_t^{(\eta)}$ and $\rho_t^{(\epsilon)}$ representing the thresholds for updating $X_t^{(\eta)}$ and $X_t^{(\epsilon)}$:

$$\rho_t^{(\eta)} = \frac{e^{2\beta(k_+(v_{t+1}, X_t^{(\eta)}) - k_-(v_{t+1}, X_t^{(\eta)}))}}{e^{2\beta(k_+(v_{t+1}, X_t^{(\eta)}) - k_-(v_{t+1}, X_t^{(\eta)}))} + 1}$$

$$\rho_t^{(\epsilon)} = \frac{e^{2\beta(k_+(v_{t+1}, X_t^{(\epsilon)}) - k_-(v_{t+1}, X_t^{(\epsilon)}))}}{e^{2\beta(k_+(v_{t+1}, X_t^{(\epsilon)}) - k_-(v_{t+1}, X_t^{(\epsilon)}))} + 1}.$$

Notice that $X_{t+1}^{(\eta)}(v_{t+1}) = X_{t+1}^{(\epsilon)}(v_{t+1})$ if $W_{t+1} \geq \max\{\rho_t^{(\eta)}, \rho_t^{(\epsilon)}\}$ or $W_{t+1} \leq \min\{\rho_t^{(\eta)}, \rho_t^{(\epsilon)}\}$. Thus, a (constant) bound on $\rho_t^{(\eta)}$ and $\rho_t^{(\epsilon)}$ would be quite useful for bounding the probability of matching the spin of a vertex v_{t+1} in $X_{t+1}^{(\eta)}$ and $X_{t+1}^{(\epsilon)}$. Let Δ be the maximum degree of G. Then, for all $v \in V$ and for all $t \geq 0$, we have $-\Delta \leq k_+(v, X_t^{(\eta)}) - k_-(v, X_t^{(\eta)}) \leq \Delta$ and $-\Delta \leq k_+(v, X_t^{(\epsilon)}) - k_-(v, X_t^{(\epsilon)}) \leq \Delta$. Since $x \in \mathbb{R} \mapsto \frac{e^{2\beta x}}{e^{2\beta x}+1}$ is an increasing function, we obtain the following bound:

$$0 < \frac{e^{-2\beta\Delta}}{e^{-2\beta\Delta}+1} \leq \rho_t^{(\eta)}, \rho_t^{(\epsilon)} \leq \frac{e^{2\beta\Delta}}{e^{2\beta\Delta}+1} < 1.$$

We are now ready to prove that $\mathbb{P}(X_n^{(\eta)} = X_n^{(\epsilon)}) > 0$. Notice that one way of achieving $X_n^{(\eta)} = X_n^{(\epsilon)}$ in n steps is by choosing a new vertex in every step (this is for $t = 1, \ldots, n$), and setting the spins of that vertex in $X_t^{(\eta)}$ and $X_t^{(\epsilon)}$ to be equal. This leads to the following inequalities:

$$\begin{aligned}
\mathbb{P}(X_n^{(\eta)} = X_n^{(\epsilon)}) &\geq \frac{(n-1)!}{n^{n-1}} \mathbb{P}(W_1 \in [0, \min\{\rho_1^{(\eta)}, \rho_1^{(\epsilon)}\}] \cup [\max\{\rho_t^{(\eta)}, \rho_t^{(\epsilon)}\}, 1], \ldots, \\
&\qquad W_n \in [0, \min\{\rho_n^{(\eta)}, \rho_n^{(\epsilon)}\}] \cup [\max\{\rho_n^{(\eta)}, \rho_n^{(\epsilon)}\}, 1]) \\
&\geq \frac{(n-1)!}{n^{n-1}} \mathbb{P}(W_1 \in [0, \frac{e^{-2\beta\Delta}}{e^{-2\beta\Delta}+1}] \cup [\frac{e^{2\beta\Delta}}{e^{2\beta\Delta}+1}, 1], \ldots, \\
&\qquad W_n \in [0, \frac{e^{-2\beta\Delta}}{e^{-2\beta\Delta}+1}] \cup [\frac{e^{2\beta\Delta}}{e^{2\beta\Delta}+1}, 1]) \\
&= \frac{(n-1)!}{n^{n-1}} (\frac{e^{-2\beta\Delta}}{e^{-2\beta\Delta}+1} + 1 - \frac{e^{2\beta\Delta}}{e^{2\beta\Delta}+1})^n \\
&> 0
\end{aligned}$$

Note that the probability of choosing a new vertex every time, n times in a row, is $\frac{(n-1)!}{n^{n-1}}$.

Thus, we have shown that for any two states $\eta, \epsilon \in \{-1, 1\}^V$ there is a non-zero probability that the chains started at η and ϵ will coalesce by the time $T = n$. This verifies the hypothesis in the 0-1 law, so we have $\mathbb{P}(\tau < \infty) = 1$. □

8 Summary

We have demonstrated how we can exploit some properties of Markov chains to devise algorithms that allow us to draw random samples from very large and complex sets, where we cannot enumerate all the elements. The generic Markov chain Monte Carlo algorithms allow us to draw samples from approximately the desired distribution, but they suffer a major issue because, in general, we do not know for how long we need to run them. We then demonstrated how the Propp-Wilson algorithm mitigates this issue, by giving samples from the exact distribution while determining automatically when to stop.

It is important to emphasize that the general Propp-Wilson algorithm, which runs as many chains as there are elements in the state spaces, is not applicable in practice for an obvious reason: the original motivation for using it is to sample from a state space that is extremely large! Thus, we need additional assumptions on our Markov chain (and update function) in order to be able to use the Propp-Wilson algorithm. In particular, in Section 6.1, we saw that if we impose some kind of monotonicity, we might need to run only two chains simultaneously, which much much better than running as many chains as there are states in the statespace. This makes the Propp-Wilson algorithm applicable to a more narrow class of problems compared to the generic MCMC. Nevertheless, we are still able to apply it to the Ising model, and draw exact samples from the Gibbs distribution.

A careful reader might have spotted one major practical drawback of the Propp-Wilson algorithm. In order to implement it correctly, we need to store all the random variables U_t that we use. Indeed, to simulate the run from $-T_i$ to 0, we need to reread the values of $U_0, U_{-1}, \ldots, U_{-T_{i-1}+1}$ that we used to simulate the run from $-T_{i-1}$ to 0. Thus we need to store all the variables U_t that we use. This essentially means that the required memory is proportional to the running time of the algorithm. Thus, if the algorithm takes too long to terminate, it might consume all of the available memory resources, forcing us to stop the simulation before obtaining a sample from the desired distribution. One might suggest that we simply ignore the failed outputs and only consider the ones that were produced before running out of memory (or, equivalently, those for which the running time was shorter). However, since the output is not independent from the running time, these outputs will be biased we will not obtain (unbiased) samples from the desired distribution.

A solution to this problem was proposed by Wilson, one of the creators of the original algorithm, in 2000 [6]. It uses read-once randomness, meaning that each U_t is read only once, so there is no need to store them. It is a modification of the original Propp-Wilson algorithm, the main idea being that we run a grand coupling from time 0 until coalescence, and then continue running it for a random number of steps. Determining this random number of steps is an elaborate procedure.

We implemented both the original Propp-Wilson algorithm and the version with read-once randomness. The code can be found at the link [5], in a file ising.py.

The next step of this project would be to perform a similar detailed analysis of this modified version of the Propp-Wilson algorithm. We could then compare it to the original algorithm, in terms of expected running time and computational complexity, parallelizability etc.

References

[1] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252, 1996.

[2] Olle Häggström. *Finite Markov chains and algorithmic applications*, volume 52. Cambridge University Press, 2002.

[3] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[4] Gordon F Newell and Elliott W Montroll. On the theory of the ising model of ferromagnetism. *Reviews of Modern Physics*, 25(2):353, 1953.

[5] Implementation of the propp-wilson algorithm. `https://github.com/pirke010/PRL`.

[6] David Bruce Wilson. How to couple from the past using a read-once source of randomness. *Random Structures & Algorithms*, 16(1):85–113, 2000.