Proof. Let $X_0 = i$ be arbitrary, and choose Y_0 according to the stationary distribution π of $\mathcal{M}$. Fix $\varepsilon \in (0,1]$ and let $t \geq t(\varepsilon)$. Then for any set of states A:

$$\begin{aligned} p^{(i,t)}(A) &= \Pr(X_t \in A) \\ &\geq \Pr(Y_t \in A \wedge X_t = Y_t) \\ &\geq 1 - \Pr(Y_t \notin A) - \Pr(X_t \neq Y_t) \\ &\geq \Pr(Y_t \in A) - \varepsilon \\ &= \pi(A) - \varepsilon, \end{aligned}$$

and similarly for the set $\bar{A} = S \setminus A$. Thus

$$|p^{(i,t)}(A) - \pi(A)| \leq \varepsilon \quad \forall\, t \geq t(\varepsilon),$$

and because A was chosen arbitrarily, also

$$d_V(p^{(i,t)}, \pi) = \max_{A \subseteq S} |p^{(i,t)}(A) - \pi(A)| \leq \varepsilon \quad \forall\, t \geq t(\varepsilon).$$

Thus $\tau(\varepsilon) \leq t(\varepsilon)$. □

Example 3.5 *Gibbs sampler for graph colourings.* Let $G = (V,E)$ be an undirected graph with maximum node degree Δ. Without loss of generality assume that $V = \{1,\ldots,n\}$. A *q-colouring* of G is a map $\sigma : V \to \{1,\ldots,q\} = Q$ such that

$$(i,j) \in E \;\Rightarrow\; \sigma(i) \neq \sigma(j).$$

According to so called Brooks' Theorem, G has a q-colouring for any $q \geq \Delta + 1$. (In fact, already for $q \geq \Delta$ unless G contains a $(\Delta+1)$-clique $K_{\Delta+1}$ as a component.)

For $q \geq \Delta + 2$, one can set up the following Gibbs sampler Markov chain $\mathcal{M}$ to sample q-colourings of G asymptotically uniformly at random (cf. Example 2.2, p. 24):

Given a colouring $\sigma \in Q^V$:

(i) select a node $i \in V$ uniformly at random;

(ii) select a legal colour c for i uniformly at random (c is legal for i if $c \neq \sigma(j)\ \forall\ j \in \Gamma(i)$);

(iii) recolour i with colour c (i.e. move from σ to σ', where $\sigma'(i) = c$ and $\sigma'(j) = \sigma(j)$ for $j \neq i$).

Let us verify that $\mathcal{M}$ is regular for $q \geq \Delta + 2$:

1. Irreducibility: Any colouring can be reached from any other by recolouring the nodes in increasing order; because $q \geq \Delta + 2$ one can avoid conflicts by if necessary first adjusting the colours at higher-numbered neighbours of the present node.

2. Aperiodicity: Each colouring has a nonzero self-loop probability, so aperiodicity follows from regularity.

It is easy to verify that by reversibility $\mathcal{M}$ has as its stationary distribution π the uniform distribution over the set of legal colourings $S \subseteq Q^V$.

Let us then consider how quickly the chain $\mathcal{M}$ converges to π, in terms of the d_V distance. To introduce the ideas, consider first the trivial case $E = \varnothing$ ($\Rightarrow\ S = Q^V$).

In this case one can effect a coupling between two copies of $\mathcal{M}$ as follows: in a transition $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$:

(i) select a node $i \in V$ uniformly at random;

(ii) select a colour $c \in Q$ uniformly at random and recolour i with colour c in both X_t and Y_t; let the resulting colourings be X_{t+1} and Y_{t+1}.

Now clearly (X_t) and (Y_t) are both faithful copies of $\mathcal{M}$, i.e. the marginal transition probabilities work out OK:

$$\begin{aligned} \Pr(X_{t+1} = \sigma' \mid X_t = \sigma, Y_t = \eta) &= \Pr(\sigma, \sigma'), \\ \Pr(Y_{t+1} = \eta' \mid X_t = \sigma, Y_t = \eta) &= \Pr(\eta, \eta'). \end{aligned}$$

On the other hand, it is clear that the time required for the chains (X_t) and (Y_t) to coalesce is not very much larger than n, because at each step of the coupled chain, a randomly chosen node is coloured similarly in both (X_t) and (Y_t).

More precisely, introduce the random variable

$$D_t = \#\{i \in V | X_t(i) \neq Y_t(i)\}.$$

Thus $D_t = 0 \ \Leftrightarrow\ X_t = Y_t \Leftrightarrow\ T \leq t$.

Furthermore,

$$E(D_{t+1} \mid D_t) = \frac{D_t}{n} \cdot (D_t - 1) + \frac{n - D_t}{n} \cdot D_t = \left(1 - \frac{1}{n}\right) \cdot D_t$$

$$\Rightarrow \quad E(D_t \mid D_0) = \left(1 - \frac{1}{n}\right)^t \cdot D_0$$

$$\overset{\text{(Markov)}}{\Rightarrow} \Pr(D_t > 0 \mid D_0) \leq E(D_t \mid D_0) \leq \left(1 - \frac{1}{n}\right)^t \cdot n \leq ne^{-t/n}.$$

Thus, choosing $t \geq n \ln \frac{n}{\varepsilon}$ suffices to guarantee that $\Pr(X_t \neq Y_t) \leq \varepsilon$, which by Lemma 3.12 implies that the mixing time satisfies $\tau(\varepsilon) \leq n \ln \frac{n}{\varepsilon}$.

For the general case we need a more complicated coupling, in order to take into account the constraints on colour choice caused by the edges in E.

We observe that by a simple construction, it is possible to combine two finite state sets A and B to a single state set C so that there are random variables X_A and X_B such that

$$\begin{aligned} &\text{(i)} \quad \Pr(X_A = x) = \begin{cases} 1/|A|, & x \in A, \\ 0, & x \notin A; \end{cases} \\ &\qquad \Pr(X_B = x) = \begin{cases} 1/|B|, & x \in B, \\ 0, & x \notin B; \end{cases} \\ &\text{(ii)} \quad \Pr(X_A = X_B) = \frac{|A \cap B|}{\max\{|A|, |B|\}}. \end{aligned} \tag{11}$$

Denote $\Gamma(i) = \{j \in V \mid (i,j) \in E\}$, $X_t(i)$ = colour of node i in colouring X_t, and $X_t(U) = \{X_t(i) \mid i \in U\}$.

Consider the following coupling $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$:

(i) select a node $i \in V$ uniformly at random;

(ii) select colours $c_X \in Q \setminus X_t(\Gamma(i))$, $c_Y \in Q \setminus Y_t(\Gamma(i))$ uniformly (but not independently) at random, using the joint sample space indicated in (11);

(iii) recolour node i with colour c_X in X_t to yield X_{t+1}; similarly with colour c_Y in Y_t to yield Y_{t+1}.

Denote $A = A_t = \{i \in V \mid X_t(i) = Y_t(i)\}$. Thus $D_t = |\overline{A}| = |V \setminus A|$.

Now clearly $D_{t+1} \in \{D_t + 1, D_t, D_t - 1\}$. Let us compute the probabilities $P(D_{t+1} \mid D_t)$ for each of these cases:

(i) $D_{t+1} = D_t + 1$. In this event the chosen $i \in A$, and $c_X \neq c_Y$.

Denote by $\xi = |Q \setminus X_t(\Gamma(i))|$, $\eta = |Q \setminus Y_t(\Gamma(i))|$, $\zeta = |Q \setminus (X_t(\Gamma(i)) \cup Y_t(\Gamma(i)))|$ the number of legal values for c_X, c_Y, and their overlap, respectively. Thus, the probability that the same colour is chosen for i in both X_{t+1} and Y_{t+1} is $\zeta / \max\{\xi, \eta\}$. Denote $d'(i) = |\Gamma(i) \setminus A|$ (recall that $i \in A$). Then

$$q - \Delta \leq \xi, \eta \leq \zeta + d'(i).$$

Hence:

$$\begin{aligned} \Pr(c_X = c_Y) &= \frac{\zeta}{\max\{\xi,\eta\}} \geq \frac{\max\{\xi,\eta\} - d'(i)}{\max\{\xi,\eta\}} \\ &\geq 1 - \frac{d'(i)}{q-\Delta} \end{aligned}$$

and consequently:

$$\Pr(D_{t+1} = D_t + 1) \leq \frac{1}{n} \sum_{i \in A} \frac{d'(i)}{q-\Delta} = \frac{m'}{(q-\Delta)n},$$

where $m' = \sum_{i \in A} d'(i)$.

(ii) $D_{t+1} = D_t - 1$. In this event the chosen $i \in \bar{A}$, and $c_X = c_Y$.

Denote ξ, η, ζ as in case (i), and $d''(i) = |\Gamma(i) \cap A|$. Now

$$q - \Delta \leq \xi, \eta \leq \zeta + (\Delta - d''(i)).$$

As in case (i), we obtain:

$$\begin{aligned} \Pr(c_X = c_Y) &= \frac{\zeta}{\max\{\xi,\eta\}} \geq \frac{\max\{\xi,\eta\} - (\Delta - d''(i))}{\max\{\xi,\eta\}} \\ &\geq 1 - \frac{\Delta - d''(i)}{q-\Delta} = \frac{q - 2\Delta + d''(i)}{q-\Delta} \end{aligned}$$

and consequently:

$$\begin{aligned} \Pr(D_{t+1} = D_t - 1) &\geq \frac{1}{n} \sum_{i \in \bar{A}} \left(\frac{q-2\Delta}{q-\Delta} + \frac{d''(i)}{q-\Delta} \right) \\ &= \frac{q-2\Delta}{(q-\Delta)n} D_t + \frac{m'}{(q-\Delta)n}, \end{aligned}$$

where $m' = \sum_{i \in \bar{A}} d''(i) = \sum_{i \in A} d'(i)$.

Denoting for brevity

$$a = \frac{q-2\Delta}{(q-\Delta)n}, \quad b = b(m') = \frac{m'}{(q-\Delta)n},$$

we see that

$$\Pr(D_{t+1} = D_t + 1) \le b, \quad \Pr(D_{t+1} = D_t - 1) \ge aD_t + b.$$

Assume that $a > 0$, i.e. that $q > 2\Delta$. Then

$$\begin{aligned} E(D_{t+1}|D_t) &\le b(D_t+1) + (aD_t+b)(D_t-1) + (1-aD_t-2b)D_t \\ &= (1-a)D_t. \end{aligned}$$

Thus, $E(D_t) \le (1-a)^t D_0 \le (1-a)^t n$, and hence by Markov's inequality

$$\Pr(D_t > 0) \le (1-a)^t n \le ne^{-at}.$$

Thus $\Pr(X_t \ne Y_t) \le \varepsilon$ for $t \ge \frac{1}{a}\ln\frac{n}{\varepsilon}$, and so by Lemma 3.12, the mixing time of the chain satisfies

$$\tau(\varepsilon) \le \frac{q-\Delta}{q-2\Delta} \cdot n \ln\frac{n}{\varepsilon} \le (\Delta+1)n\ln\frac{n}{\varepsilon}$$

for $q > 2\Delta$.

4 Exact Sampling with Coupled Markov Chains

In 1996 J. Propp and D. Wilson introduced an intriguing method for producing samples from a Markov chain *exactly* according to its stationary distribution. This *exact sampling* (or "coupling from the past") technique eliminates the need to compute Markov chain convergence rates for quality control: when the algorithm stops, it is guaranteed to produce a perfect sample. However for slowly converging chains stopping will take a long time, so convergence rates are still of importance from the point of view of algorithm efficiency. (There are also some other efficiency caveats in the method besides slow convergence of the simulated chain. These are discussed below.)

Let $\mathcal{M}$ be a regular reversible Markov chain with state set $S = \{1, \ldots, n\}$, transition probability matrix $P = (p_{ij})$, and stationary distribution π.

Consider an explicit simulation of $\mathcal{M}$ by the following method: at each step t, a uniformly distributed random number $R_t \in [0,1)$ is chosen, and the state transition of $\mathcal{M}$ is determined as $X_{t+1} = s(X_t, R_t)$, where

$$s(i,r) = \begin{cases} 1, & \text{if } r \in [0, p_{i1}), \\ 2, & \text{if } r \in [p_{i1}, p_{i1} + p_{i2}), \\ \vdots & \\ n, & \text{if } r \in [p_{i1} + \ldots + p_{i(n-1)}, 1). \end{cases}$$

It is clear that transition probabilities according to the chain $\mathcal{M}$ can equivalently be computed with respect to sequences (R_t) and the above deterministic transition rule, e.g.

$$P_{ij}^{(t)} = \Pr(X_t = j | X_0 = i) = \Pr_{\vec{R}}(s^{(t)}(i, \vec{R}) = j),$$

where

$$s^{(t)}(i, \langle r_0, r_1, \ldots, r_{t-1} \rangle) = \underbrace{s(s(\cdots s(s}_{t}(i, r_0), r_1) \cdots), r_{t-1}).$$

Now let us consider the following curious simulation method for the chain $\mathcal{M}$, from further and further away in the *past* ($t = -T,\ T = 1, 2, 4, 8, \ldots$) to the present ($t = 0$):

Algorithm PW (Propp-Wilson):

set $T \leftarrow 1$
generate random numbers $r_{-T}, \ldots, r_{-1} \in [0,1)$ uniformly at random;
(1) simulate the chain $\mathcal{M}$ as above, using the random numbers
    $r_{-T}, \ldots, r_{-1}$, from every possible initial state $X_{-T} \in S$;
if all the simulations lead to the same state $X_0 = i_0$, then output i_0
    and stop;
otherwise generate T more random numbers $r_{-2T}, \ldots, r_{-T-1} \in [0,1)$
    uniformly at random;
set $T \leftarrow 2T$; go to (1).

For a three-state chain, a run of the PW algorithm might look as illustrated in Figure 14. Here the algorithm has required two restarts, but the third run from $T = -4$ has resulted in all the simulated realisations of the chain coalescing, with common result $i_0 = 2$.

In the following, we shall assume that the PW algorithm always converges with probability 1. Ensuring this may require some care in verifying that the deterministic update rule $s(i,r)$, and the chosen numbering of the Markov chain states do not interact in a bad way.

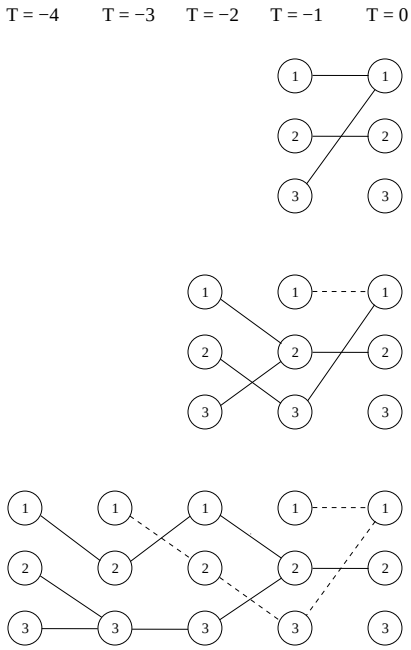

Figure 14: A Propp-Wilson simulation of a 3-state Markov chain.

Theorem 4.1 *Let Y be a random variable indicating the eventual output state of the PW algorithm, under the above assumptions and notations. Then*

$$\Pr_R(Y = i) = \pi_i, \quad \forall\, i \in S.$$

Proof. Fix some value $i \in S$. To prove the Theorem, it suffices to show that for any $\varepsilon > 0$

$$|\Pr_R(Y = i) - \pi_i| \leq \varepsilon.$$

So fix an arbitrary $\varepsilon > 0$. Since we assume that the PW algorithm terminates with probability 1, there is some value of T such that

$$\Pr_R(\text{PW simulation converges for chains of length } T) \geq 1 - \varepsilon. \tag{12}$$

Now consider running the actual chain from time $-T$ to time 0, starting with the stationary distribution:

$$\Pr(X_{-T} = i) = \pi_i.$$

In this case, of course also the variable X_0 is distributed according to the stationary distribution:

$$\Pr_R(X_0 = i) = \pi_i.$$

However, if the coalescence event (12) occurs for a given sequence R of random numbers, then $X_0 = Y$, and so $\Pr_R(X_0 \neq Y) \leq \varepsilon$. Thus,

$$\begin{aligned}\Pr(Y = i) - \pi_i &= \Pr(Y = i) - \Pr(X_0 = i) \\ &\leq \Pr(Y = i, X_0 \neq i) \\ &\leq \varepsilon,\end{aligned}$$

and by a similar argument

$$\pi_i - \Pr(Y = i) \leq \varepsilon.$$

Thus, $|\Pr(Y = i) - \pi_i| \leq \varepsilon$, and the claim is proved. □

Note that the PW algorithm cannot be "simplified" by simulating the chains forwards from time $T = 0$ until they coalesce. This yields biased samples.

The PW algorithm as described above still has two shortcomings:

1. The need to store long sequences of random numbers for reuse (can be a serious problem in long simulations); and
2. The need to simulate the chains starting from all possible initial states (infeasible in many applications where the number of system states is exponential in the size of the system itself).

Problem (1) has been addressed in a recent (2000) modification to the algorithm ("CFTP with read once randomness") by D. Wilson.

For problem (2), Propp & Wilson (1996) proposed a solution that can be applied when the states of the system have a suitable partial order $\sqsubseteq$ respected by the update rule.

Specifically, assume that the states of the system to be simulated form a partial order $(S = \{\sigma_1, \ldots, \sigma_n\}, \sqsubseteq)$ with a unique largest element $\top$ ("top") and unique smallest element $\bot$ ("bottom"), and satisfying the condition

$$\sigma \sqsubseteq \sigma' \;\Rightarrow\; s(\sigma, r) \sqsubseteq s(\sigma', r), \quad \forall\, \sigma, \sigma' \in S \text{ and } r \in [0, 1). \tag{13}$$

Then it suffices to simulate the "top" and "bottom" chains until they couple, since their coupling implies the coalescence of all the other chains as well (cf. Figure 15).

This is of course a huge improvement: reducing the simulation of, say, 2^n parallel chains to just 2.

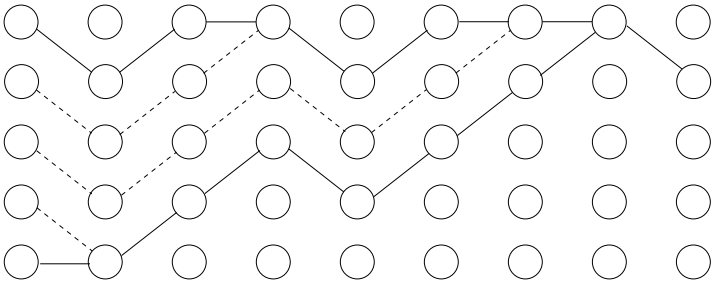

Figure 15: Coalescence of an ordered Propp-Wilson simulation.

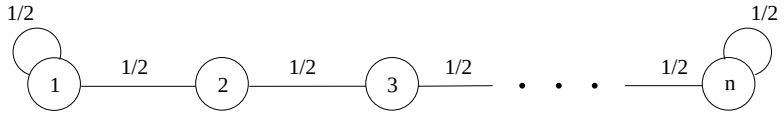

Figure 16: A one-dimensional random walk with semi-reflecting barriers.

So what systems admit this simplification?

A simple example would be a one-dimensional random walk on the state set $S = \{1, \ldots, n\}$ with, say, semi-reflecting barriers to ensure regularity of the chain (Figure 16). Assume the state transition rule is:

$$s(i,r) = \begin{cases} \max\{i-1,1\}, & \text{if } 0 \le r < \frac{1}{2}, \\ \min\{i+1,n\}, & \text{if } \frac{1}{2} \le r < 1. \end{cases}$$

The the natural ordering of states fulfills the condition (13):

$$i \le j \;\Rightarrow\; s(i,r) \le s(j,r) \qquad \forall\, i,j = 1,\ldots,n,\; r \in [0,1).$$

Interestingly, also complicated systems such as the Ising spin glass model admit such orderings. In the case of the Ising model, the order between states $\sigma, \sigma' \in \{-1,+1\}^n$ is determined simply by

$$\sigma \sqsubseteq \sigma' \quad \text{if} \quad \sigma_i \le \sigma'_i \;\forall\, i = 1,\ldots,n.$$

Clearly $\bot = (-1,\ldots,-1)$ and $\top = (1,\ldots,1)$ with respect to $\sqsubseteq$, and also condition (13) can be verified.