# Chapter 9
# A Second Look at General Markov Chains

David Aldous
Department of Statistics
University of California
Berkeley, CA 94720

James Allen Fill
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218-2692

April 21, 1995

In the spirit of Chapter 2, this is an unsystematic treatment of scattered topics which are related to topics discussed for reversible chains, but where reversibility plays no essential role. Section 1 treats constructions of stopping times with various optimality properties. Section 2 discusses random spanning trees associated with Markov chains, the probabilistic elaboration of "the matrix-tree theorem". Section 3 discusses self-verifying algorithms for sampling from a stationary distribution. Section 4 discusses "reversiblizations" of irreversible chains. Section 5 gives an example to show that the nonasymptotic interpretation of relaxation time, so useful in the reversible setting, may fail completely in the general case. At first sight these topics may seem entirely unrelated, but we shall see a few subtle connections.

Throughout the chapter, our setting is a finite irreducible discrete-time Markov chain $(X_n)$ with transition matrix $\mathbf{P} = (p_{ij})$.

## 1 Minimal constructions and mixing times

Chapter 4 Theorem yyy involved three mixing time parameters; $\tau_1$ related to variation distance to stationarity, $\tau_1^{(1)}$ related to "separation" from stationarity, and $\tau_1^{(2)}$ related to stationary times (see below). In Chapter 4 these parameters were defined under worst-case initial distributions, and our focus was on "equivalence" of these parameters for reversible chains. Here we discuss underlying "exact" results. Fix an initial distribution $\mu$. Then associated with each notion of mixing, there is a corresponding construction of a minimal random time $T$, stated in Theorems 1 - 3 below.

xxx randomized stopping times

Call a stopping time $T$ a *strong stationary time* if

$$P_\mu(X_t = j, T = t) = \pi_j P_\mu(T = t) \text{ for all } j, t \tag{1}$$

i.e. if $X_T$ has distribution $\pi$ and is independent of $T$. Call a stopping time $T$ a *stationary time* if

$$P_\mu(X_T = j) = \pi_j \text{ for all } j. \tag{2}$$

Call a random time $T$ a *coupling time* if we can construct a joint distribution $((X_t, Y_t); t \geq 0)$ such that $(X_t)$ is the chain with initial distribution $\mu$, $(Y_t)$ is the stationary chain, and $X_t = Y_t, t \geq T$. (A coupling time need not be a stopping time, even w.r.t. the joint process; this is the almost the only instance of a random time which is not a stopping time that we encounter in this book.)

Recall from yyy the notion of *separation* of $\theta$ from $\pi$:

$$\mathrm{sep}(\theta) \equiv \min\{u : \theta_j \geq (1-u)\pi_j \ \forall j\}.$$

Write $\mathrm{sep}_\mu(t)$ for the separation at time $t$ when the initial distribution was $\mu$:

$$\mathrm{sep}_\mu(t) = \min\{u : P_\mu(X_t = j) \geq (1-u)\pi_j \ \forall j\}.$$

Similarly write $\mathrm{vd}_\mu(t)$ for the variation distance from stationarity at time $t$:

$$\mathrm{vd}_\mu(t) = \frac{1}{2}\sum_j |P_\mu(X_t = j) - \pi_j|.$$

**Theorem 1** *Let $T$ be any strong stationary time for the $\mu$-chain. Then*

$$\mathrm{sep}_\mu(t) \leq P_\mu(T > t) \textit{ for all } t \geq 0. \tag{3}$$

*Moreover there exists a* minimal *strong stationary time $T$ for which*

$$\mathrm{sep}_\mu(t) = P_\mu(T > t) \textit{ for all } t \geq 0. \tag{4}$$

**Theorem 2** *For any coupling time $T$,*

$$\mathrm{vd}_\mu(t) \leq P_\mu(T > t) \textit{ for all } t \geq 0.$$

*Moreover there exists a* minimal *coupling time $T$ for which*

$$\mathrm{vd}_\mu(t) = P_\mu(T > t) \textit{ for all } t \geq 0.$$

**Theorem 3** *For any stationary time $T$,*

$$E_\mu T \geq \max_j (E_\mu T_j - E_\pi T_j). \tag{5}$$

*Moreover there exist* mean-minimal *stationary times $T$ for which*

$$E_\mu T = \max_j (E_\mu T_j - E_\pi T_j). \tag{6}$$

In each case, the first assertion is immediate from the definitions, and the issue is to carry out a construction of the required $T$. Despite the similar appearance of the results, attempts to place them all in a common framework have not been fruitful. We will prove Theorems 1 and 3 below, and illustrate with examples. These two proofs involve only rather simple "greedy" constructions. We won't give the proof of Theorem 2 (the construction is usually called the *maximal coupling*: see Lindvall [25]) because the construction is a little more elaborate and the existence of the minimal coupling time is seldom useful, but on the other hand the *coupling inequality* in Theorem 2 will be used extensively in Chapter 14. In the context of Theorems 1 and 2 the minimal times $T$ are clearly unique in distribution, but in Theorem 3 there will generically be many mean-minimal stationary times $T$ with different distributions.

## 1.1 Strong stationary times

For any stopping time $T$, define

$$\theta_j(t) = P_\mu(X_t = j, T \geq t), \qquad \sigma_j(t) = P_\mu(X_t = j, T = t). \tag{7}$$

Clearly these vectors satisfy

$$0 \leq \sigma(t) \leq \theta(t), \ \ (\theta(t) - \sigma(t))\mathbf{P} = \theta(t+1) \ \forall t; \ \ \theta_0 = \mu. \tag{8}$$

Conversely, given $(\theta(t), \sigma(t); t \geq 0)$ satisfying (8), we can construct a randomized stopping time $T$ satisfying (7) by declaring that $P(T = t | X_t = j, T \geq t, X_s, s < t) = \sigma_j(t)/\theta_j(t)$. The proofs of Theorems 1 and 3 use different definitions of vectors satisfying (8).

*Proof of Theorem 1.* A particular sequence $(\theta(t), \sigma(t); t \geq 0)$ can be specified inductively by (8) and

$$\sigma(t) = r_t \pi \ , \quad \text{where } r_t = \min_j \theta_j(t)/\pi_j. \tag{9}$$

The associated stopping time satisfies

$$P_\mu(X_t = j, T = t) = \sigma_j(t) = r_t \pi_j$$

and so is a strong stationary time with $P_\mu(T = t) = r_t$. One can now verify inductively that

$$P_\mu(X_t \in \cdot) = \theta(t) + P_\mu(T \leq t-1) \ \cdot pi$$

and so the separation is

$$\mathrm{sep}_\mu(t) = 1 - \min_j \frac{P_\mu(X_t = j)}{\pi_j} = P_\mu(T \geq t) - r_t = P_\mu(T > t).$$

### 1.2 Stopping times attaining a specified distribution

For comparison with the other two results, we stated Theorem 3 in terms of stopping times at which the stationary distribution is attained, but the underlying result (amplified as Theorem 4) holds for an arbitrary target distribution $\rho$. So fix $\rho$ as well as the initial distribution $\mu$. Call a stopping time $T$ *admissible* if $P_\mu(X_T \in \cdot) = \rho$. Write $\bar{t}(\mu, \sigma)$ for the *inf* of $E_\mu T$ over all admissible stopping times $T$.

**Theorem 4** *(a)* $\bar{t}(\mu, \sigma) = \max_j (E_\mu T_j - E_\rho T_j)$.

*(b) The "filling scheme" below defines an admissible stopping time such that* $E_\mu T = \bar{t}(\mu, \sigma)$.

*(c) Any admissible stopping time $T$ with the property*

$$\exists\ k \textit{ such that } P_\mu(T \leq T_k) = 1. \tag{10}$$

*satisfies* $E_\mu T = \bar{t}(\mu, \sigma)$.

Part (c) is rather remarkable, and can be rephrased as follows. Call a state $k$ with property (10) a *halting state* for the stopping time $T$. In words, the chain must stop if and when it hits a halting state. Then part (c) asserts that, to verify that an admissible time $T$ attains the minimum $\bar{t}(\mu, \rho)$, it suffices to show that there exists some halting state. In the next section we shall see this is very useful in simple examples.

*Proof.* The greedy construction used here is called a *filling scheme.* Recall from (7) the definitions

$$\theta_j(t) = P_\mu(X_t = j, T \geq t), \qquad \sigma_j(t) = P_\mu(X_t = j, T = t).$$

Write also $\Sigma_j(t) = P_\mu(X_T = j, T \leq t)$. We now define $(\theta(t), \sigma(t); t \geq 0)$ and the associated stopping time $\bar{T}$ inductively via (8) and

$$\begin{aligned} \sigma_j(t) &= 0 \text{ if } \Sigma_j(t-1) = \rho_j \\ &= \theta_t \text{ if } \Sigma_j(t-1) + \theta_j(t) \leq \rho_j \\ &= \rho_j - \Sigma_j(t-1) \text{ otherwise.} \end{aligned}$$

In words, we stop at the current state ($j$, say) provided our "quota" $\rho_j$ for the chance of stopping at $j$ has not yet been filled. Clearly

$$\Sigma_j(t) \leq \rho_j \ \forall j \ \forall t. \tag{11}$$

We now claim that $\bar{T}$ satisfies property (10). To see this, consider

$$t_j \equiv \min\{t : \Sigma_j(t) = \rho_j\} \leq \infty.$$

Then (10) holds by construction for any $k$ such that $t_k = \max_j t_j \leq \infty$. In particular, $\bar{T} \leq T_k < \infty$ a.s. and then by (11) $P_\mu(X_{\bar{T}} \in \cdot) = \lim t \to \infty \Sigma(t) = \rho$. So $\bar{T}$ is an admissible stopping time.

*Remark.* Generically we expect $t_j = \infty$ for exactly one state $j$, though other possibilities may occur, e.g. in the presence of symmetry.

Now consider an arbitrary admissible stopping time $T$, and consider the associated *occupation measure* $\mathbf{x} = (x_j)$:

$$x_j \equiv E_\mu(\text{number of visits to } j \text{ during times } 0, 1, \ldots, T-1).$$

We shall show

$$x_j + \rho_j = \mu_j + \sum_i x_i p_{ij} \ \forall j. \tag{12}$$

Indeed, by counting the number of visits during $0, 1, \ldots, T-1, T$ in two ways,

$$x_j + \rho_j = \mu_j + E_\mu(\text{number of visits to } j \text{ during } 1, 2, \ldots, T).$$

Chapter 2 Lemma yyy showed the (intuitively obvious) fact

$$x_i p_{ij} = E_\mu(\text{ number of transitions } i \to j \text{ starting before time } T).$$

So summing over $i$,

$$\sum_i x_i p_{ij} = E_\mu(\text{number of visits to } j \text{ during } 1, 2, \ldots, T)$$

and (12) follows.

Write $\bar{\mathbf{x}}$ for the occupation measure associated with the stopping time $\bar{T}$ produced by the filling scheme. By (10), $\min_k \bar{x}_k = 0$. If $\mathbf{x}$ and $\mathbf{x}'$ are solutions of (12) then the difference $\mathbf{d} = \mathbf{x} - \mathbf{x}'$ satisfies $\mathbf{d} = \mathbf{dP}$ and so is a multiple of the stationary distribution $\pi$. In particular, if $\mathbf{x}$ is the occupation measure for some arbitrary admissible time $T$, then

$$\mathbf{x} \geq \bar{\mathbf{x}}, \text{ with equality iff } \min_k x_k = 0.$$

Since $E_\mu T = \sum_i x_i$, we have established parts (b) and (c) of the theorem, and

$$\bar{t}(\mu, \sigma) = \sum_i \bar{x}_i.$$

To prove (a), choose a state $k$ such that $\bar{x}_k = 0$, that is such that $\bar{T} \leq T_k$. Then $E_\mu T_k = E_\mu \bar{T} + E_\rho T_k$ and hence $\bar{t}(\mu, \sigma) \leq \max_j (E_\mu T_j - E_\rho T_j)$. But for any admissible stopping time $T$ and any state $j$

$$E_\mu T_j \leq E_\mu T + E_\rho T_j$$

giving the reverse inequality $\bar{t}(\mu, \sigma) \geq \max_j (E_\mu T_j - E_\rho T_j)$. □

**Corollary 5** *The minimal strong stationary time has mean $\bar{t}(\mu, \pi)$, i.e. is mean-minimal amongst all not-necessarily-strong stationary times, iff there exists a state $k$ such that*

$$P_\mu(X_t = k)/\pi_k = \min_j P_\mu(X_t = j)/\pi_j \ \forall t.$$

*Proof.* From the construction of the minimal strong stationary time, this is the condition for $k$ to be a halting state.

## 1.3 Examples

**Example 6** *Patterns in coin-tossing.*

Recall Chapter 2 Example yyy: $(X_t)$ is the chain on the set $\{H, T\}^n$ of $n$-tuples $i = (i_1, \ldots, i_n)$. Start at some arbitrary initial state $j = (j_1, \ldots, j_n)$. Here the deterministic stopping time "$T = n$" is a strong stationary time. Now a state $k = (k_1, \ldots, k_n)$ will be a halting state provided it does not overlap $j$, that is provided there is no $1 \leq u \leq n$ such that $(j_u, \ldots, j_n) = (k_1, \ldots, k_{n+u-1})$. But the number of overlapping states is at most $1 = 2 + 2^2 + \ldots + 2^{n-1} = 2^n - 1$, so there exists a non-overlapping state, i.e. a halting state. So $ET$ attains the minimum $(= n)$ of $\bar{t}(j, \pi)$ over all stationary times (and not just over all *strong* stationary times).

**Example 7** *Top-to-random card shuffling.*

Consider the following scheme for shuffling an $n$-card deck: the top card is removed, and inserted in one of the $n$ possible positions, chosen uniformly at random. Start in some arbitrary order. Let $T$ be the first time that the card which was originally second-from-bottom has reached the top of the deck. Then it is not hard to show (Diaconis [15] p. 177) that $T+1$ is a strong stationary time. Now any configuration in which the originally-bottom card is the top card will be a halting state, and so $T+1$ is mean-minimal over all stationary times. Here $E(T+1) = 1 + \sum_{m=2}^{n-1} \frac{n}{m} = n(h_n - 1)$.

**Example 8** *The winning streak chain.*

In a series of games which you win or lose independently with chance $0 < c < 1$, let $\hat{X}_t$ be your current "winning streak", i.e. the number of games won since your last loss. For fixed $n$, the truncated process $X_t = \min(X_t, n-1)$ is the Markov chain on states $\{0, 1, 2, \ldots, n-1\}$ with transition probabilities

$$p(i,0) = 1-c, \quad p(i, \min(i+1, n-1)) = c; \ 0 \le i \le n-1$$

and stationary distribution

$$\pi_i = (1-c)c^i, \ 0 \le i \le n-2; \quad \pi_{n-1} = c^{n-1}.$$

We present this chain, started at 0, as an example where it is easy to see there are different mean-minimal stationary times $T$. We'll leave the simplest construction until last – can you guess it now? First consider $T_J$, where $J$ has the stationary distribution. This is a stationary time, and $n-1$ is a halting state, so it is mean-minimal. Now it is easy to show

$$E_0 T_j = \frac{1}{(1-c)c^j} - \frac{1}{1-c}, 1 \le j \le n-1.$$

(Slick proof: in the not-truncated chain, Chapter 2 Lemma yyy says $1 = E_j($ number of visits to $j$ before $T_0) = \pi_j(E_j T_0 + E_0 T_j) = \pi_j(1/(1-c) + E_0 T_j)$.) So

$$\bar{t}(0,\pi) = E_0 T_J = \sum_{j \ge 1} \pi_j E_0 T_j = n - 2 + \frac{\pi_{n-1}}{(1-c)c^{n-1}} - \frac{1-\pi_0}{1-c} = n-1.$$

Here is another stopping time $T$ which is easily checked to attain the stationary distribution, for the chain started at 0. With chance $1-c$ stop at time

0. Otherwise, run the chain until either hitting $n-1$ (in which case, stop) or returning to 0. In the latter case, the return to 0 occurs as a transition to 0 from some state $M \geq 0$. Continue until first hitting $M+1$, then stop. Again $n-1$ is a halting state, so this stationary time also is mean-minimal. Of course, the simplest construction is the deterministic time $T = n-1$. This is a strong stationary time (the winning streak chain is a function of the patterns in coin tossing chain), and again $n-1$ is clearly a halting state. Thus $\bar{t}(0,\pi) = n-1$ without needing the calculation above.

*Remark.* One could alternatively use Corollary 5 to show that the strong stationary times in Examples 6 and 7 are mean-minimal stationary times. The previous examples are atypical: here is a more typical example in which the hypothesis of Corollary 5 is not satisfied and so no mean-optimal stationary time is a strong stationary time.

**Example 9** *xxx needs a name!*

Chapter 2 Example yyy can be rewritten as follows. Let $(U_t)$ be independent uniform on $\{0,1,\ldots,n-1\}$ and let $(A_t)$ be independent events with $P(A_t) = a$. Define a chain $X$ on $\{0,1,\ldots,n-1\}$ by

$$\begin{aligned} X_{t+1} &= U_{t+1} \text{ on } A_t^c \\ &= X_t + 1 \bmod n \text{ on } A_t. \end{aligned}$$

The stationary distribution is the uniform distribution. Take $X_0 = 0$. Clearly $T \equiv \min\{t : A_t \text{ occurs }\}$ is a strong stationary time, and $ET = 1/(1-a)$, and it is easy to see that $T$ is the minimal strong stationary time. But $T$ is not a mean-minimal stationary time. The occupation measure $\mathbf{x}$ associated with $T$ is such that $\min_j x_j = x_{n-1} = a^{n-1} + a^{2n-1} + \ldots = a^{n-1}/(1-a^n)$, and so the occupation measure $\bar{\mathbf{x}}$ associated with a mean-minimal stationary time is $\bar{\mathbf{x}} = \mathbf{x} - \frac{a^{n-1}}{1-a^n}\pi$, and so $\bar{t}(0,\pi) = \frac{1}{1-a} - \frac{a^{n-1}}{1-a^n}$.

# 2 Markov chains and spanning trees

## 2.1 General Chains and Directed Weighted Graphs

Let's jump into the details and defer the discussion until later. Consider a finite irreducible discrete-time Markov chain $(X_n)$ with transition matrix $\mathbf{P} = (p_{vw})$, and note we are not assuming reversibility. We can identify $\mathbf{P}$ with a weighted directed graph, which has (for each $(v,w)$ with $p_{vw} > 0$)

a directed edge $(v, w)$ with weight $p_{vw}$. A *directed spanning tree* $\mathbf{t}$ is a spanning tree with one vertex distinguished as the root, and with each edge $e = (v, w)$ of $\mathbf{t}$ regarded as being directed towards the root. Write $\mathcal{T}$ for the set of directed spanning trees. For $\mathbf{t} \in \mathcal{T}$ define

$$\bar{\rho}(\mathbf{t}) \equiv \prod_{(v,w)\in\mathbf{t}} p_{vw}.$$

Normalizing gives a probability distribution $\rho$ on $\mathcal{T}$:

$$\rho(\mathbf{t}) \equiv \frac{\bar{\rho}(\mathbf{t})}{\sum_{\mathbf{t}'\in\mathcal{T}} \bar{\rho}(\mathbf{t}')}.$$

Now fix $n$ and consider the stationary Markov chain $(X_m : -\infty < m \leq n)$ run from time minus infinity to time $n$. We now use the chain to construct a random directed spanning tree $\mathbf{T}_n$. The root of $\mathbf{T}_n$ is $X_n$. For each $v \neq X_n$ there was a final time, $L_v$ say, before $n$ that the chain visited $v$:

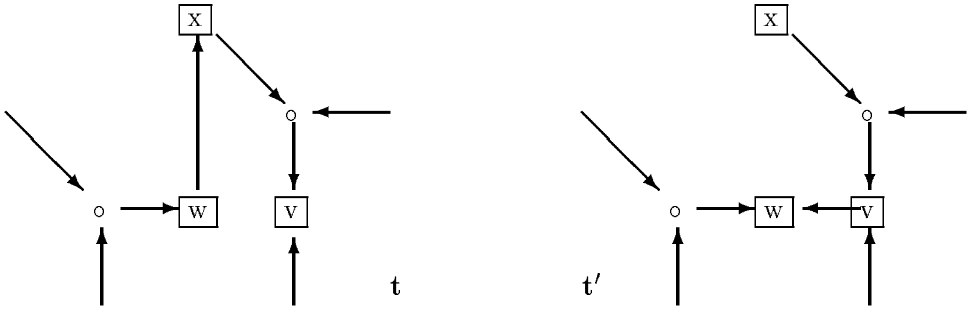
$$L_v \equiv \max\{m \leq n : X_m = v\}.$$

Define $\mathbf{T}_n$ to consist of the directed edges

$$(v = X_{L_v}, X_{L_v+1}),\ v \neq X_n.$$

So the edges of $\mathbf{T}_n$ are the last-exit edges from each vertex (other than the root $X_n$). It is easy to check that $\mathbf{T}_n$ is a directed spanning tree.

Now consider what happens as $n$ changes. Clearly the process $(\mathbf{T}_n : -\infty < n < \infty)$ is a stationary Markov chain on $\mathcal{T}$, with a certain transition matrix $\mathbf{Q} = (q(\mathbf{t}, \mathbf{t}'))$, say. The figure below indicates a typical transition $\mathbf{t} \to \mathbf{t}'$. Here $\mathbf{t}$ was constructed by the chain finishing at its root $v$, and $\mathbf{t}'$ is the new tree obtained when the chain makes a transition $v \to w$.



**Theorem 10 (The Markov chain tree theorem)** *The stationary distribution of* $(\mathbf{T}_n)$ *is* $\rho$.

*Proof.* Fix a directed spanning tree $\mathbf{t}'$. We have to verify

$$\sum_{\mathbf{t}} \bar{\rho}(\mathbf{t}) q(\mathbf{t}, \mathbf{t}') = \bar{\rho}(\mathbf{t}'). \tag{13}$$

Write $w$ for the root of $\mathbf{t}'$. For each vertex $x \neq w$ there is a tree $\mathbf{t}_x$ constructed from $\mathbf{t}'$ by adding an edge $(w, x)$ and then deleting from the resulting cycle the edge $(v, w)$ (say, for some $v = v(x)$) leading into $w$. For $x = w$ set $v(x) = x$. It is easy to see that the only possible transitions into $\mathbf{t}'$ are from the trees $\mathbf{t}_x$, and that

$$\frac{\bar{\rho}(\mathbf{t}_x)}{\bar{\rho}(\mathbf{t}')} = \frac{p_{wx}}{p_{vw}}; \qquad q(\mathbf{t}_x, \mathbf{t}') = p_{vw}.$$

Thus the left side of (13) becomes

$$\sum_x \bar{\rho}(\mathbf{t}_x) q(\mathbf{t}_x, \mathbf{t}') = \bar{\rho}(\mathbf{t}') \sum_x p_{wx} = \bar{\rho}(\mathbf{t}').$$

□

The underlying chain $X_n$ can be recovered from the tree-valued chain $\mathbf{T}_n$ via $X_n = \mathrm{root}(\mathbf{T}_n)$, so we can recover the stationary distribution of $X$ from the stationary distribution of $T$, as follows.

**Corollary 11 (The Markov chain tree formula)** *For each vertex $v$ define*

$$\bar{\pi}(v) \equiv \sum_{\mathbf{t}:\ v = root(\mathbf{t})} \bar{\rho}(\mathbf{t}).$$

$$\pi(v) \equiv \frac{\bar{\pi}(v)}{\sum_w \bar{\pi}(w)}.$$

*Then $\pi$ is the stationary distribution of the original chain $(X_n)$.*

See the Notes for comments on this classical result.

Theorem 10 and the definition of $\mathbf{T}_0$ come close to specifying an algorithm for constructing a random spanning tree with distribution $\rho$. Of course the notion of running the chain from time $-\infty$ until time 0 doesn't sound very algorithmic, but we can rephrase this notion using time-reversal. Regarding the stationary distribution $\pi$ as known, the time-reversed chain $X^*$ has transition matrix $p^*_{vw} \equiv \pi_w p_{wv} / \pi_v$. Here is the restatement of Theorem 10 in terms of the time-reversed chain.

**Corollary 12** *Let $(X^*_m : 0 \leq m \leq C)$ be the time-reversed chain, run until the cover time $C$. Define $\mathbf{T}$ to be the directed spanning tree with root $X_0$ and with edges $(v = X_{T_v}, X_{T_v - 1})$, $v \neq X_0$. If $X_0$ has distribution $\pi$ then $\mathbf{T}$ has distribution $\rho$. If $X_0$ is deterministically $v_0$, say, then $T$ has distribution $\rho$ conditioned on being rooted at $v_0$.*

Thus $\mathbf{T}$ consists of the edges by which each vertex is *first* visited, directed backwards.

For a reversible chain, we can of course use the chain itself in Corollary 12 above, in place of the time-reversed chain. If the chain is random walk on a unweighted graph $G$, then

$$\bar{\rho}(\mathbf{t}) = d(\mathrm{root}(\mathbf{t})) \prod_v \frac{1}{d(v)}$$

where $d(v)$ is the degree of $v$ in $G$. So $\bar{\rho}$, restricted to the set of spanning trees with specified root $v_0$, is *uniform* on that set. In this setting, Corollary 12 specializes as follows.

**Corollary 13** *Let $(X_m : 0 \leq m \leq C)$ be random walk on an unweighted graph $G$, started at $v_0$ and run until the cover time $C$. Define $\mathbf{T}$ to be the directed spanning tree with root $v_0$ and with edges $(v = X_{T_v}, X_{T_v - 1})$, $v \neq v_0$. Then $\mathbf{T}$ is uniform on the set of all directed spanning trees of $G$ rooted at $v_0$.*

We can rephrase this. If we just want "plain" spanning trees without a root and directions, then the $\mathbf{T}$ above, regarded as a plain spanning tree, is uniform on the set of all plain spanning trees. On the other hand, if we want a rooted spanning tree which is uniform on all such trees without prespecified root, the simplest procedure is to construct $\mathbf{T}$ as in Corollary 13 with deterministic start $v_0$, and at the end re-root $\mathbf{T}$ at a uniform random vertex. (This is slightly subtle – we could alternatively start with $X_0$ uniform, which is typically *not* the stationary distribution $\pi$.)

Using the bounds on cover time developed in Chapter 6, we now have an algorithm for generating a uniform spanning tree of a $n$-vertex graph in $O(n^3)$ steps (and $O(n^2)$ steps on a regular graph). No other known algorithm achieves these bounds.

## 2.2 Electrical network theory

The ideas in this subsection (and much more) are treated in a long but very readable survey paper by Pemantle [30], which we encourage the interested reader to consult. As observed above, in the reversible setting we have the obvious simplification that we can construct uniform spanning trees using the chain itself. Deeper results can be found using the electrical network analogy. Consider random walk on a weighted graph $G$. The random spanning tree $\mathbf{T}$ constructed by Corollary 12, interpreted as a "plain" spanning tree, has distribution

$$\rho(\mathbf{t}) = c \prod_{e \in \mathbf{t}} w_e$$

where $c$ is the normalizing constant. If an edge $e$ is essential, it must be in every spanning tree, so $P(e \in \mathbf{T}) = 1$. If the edge is inessential, the probability will be strictly between 0 and 1. Intuitively, $P(e \in \mathbf{T})$ should provide a measure of "how nearly essential $e$ is". This should remind the reader of the inessential edge inequality (yyy). Interpreting the weighted graph as an electrical network where an edge $e = (v, x)$ has resistance $1/w_e$, the effective resistance $r_{vx}$ between $v$ and $x$ satisfies

$$r_{vx} \leq 1/w_{vx} \text{ with equality iff } (v, x) \text{ is essential}$$

**Proposition 14** *For each edge $(v, x)$,*

$$P((v, x) \in \mathbf{T}) = w_{vx} r_{vx}.$$

Note that in a $n$-vertex graph, $\mathbf{T}$ has exactly $n - 1$ edges, so Proposition 14 implies Foster's theorem (Chapter 3 yyy)

$$\sum_{\text{edges } (v,x)} w_{vx} r_{vx} = n - 1.$$

*Proof.* Consider the random walk started at $v$ and run until the time $U$ of the first return to $v$ after the first visit to $x$. Let $p$ be the chance that $X_{U-1} = x$, i.e. that the return to $x$ is along the edge $(x, v)$. We can calculate $p$ in two ways. In terms of random walk started at $x$, $p$ is the chance that the first visit to $v$ is from $x$, and so by Corollary 12 (applied to the walk started at $x$) $p = P((x, v) \in \mathbf{T})$. On the other hand, consider the walk started at $v$ and let $S$ be the first time that the walk traverses $(x, v)$ in that direction. Then

$$ES = EU/p.$$

But by yyy and yyy

$$ES = w/w_{vx}, \ EU = w r_{vx}$$

and hence $p = w_{vx} r_{vx}$ as required. □

The next result indicates the usefulness of the electrical network analogy.

**Proposition 15** *For any two edges $e_1 \neq e_2$,*

$$P(e_1 \in \mathbf{T}, e_2 \in \mathbf{T}) \leq P(e_1 \in \mathbf{T}) P(e_2 \in \mathbf{T}).$$

*Proof.* Consider the "shorted" graph $G^{\mathrm{short}}$ in which the end-vertices $(x_1, x_2)$ of $e_1$ are shorted into a single vertex $x$, with edge-weights $w_{xv} = w_{x_1 v} + w_{x_2 v}$. The natural $1-1$ correspondence $\mathbf{t} \leftrightarrow \mathbf{t} \cup \{e_1\}$ between spanning trees of $G^{\mathrm{short}}$ and spanning trees of $G$ containing $e_1$ maps the distribution $\rho^{\mathrm{short}}$ to the conditional distribution $\rho(\cdot | e_1 \in \mathbf{T})$. So, writing $\mathbf{T}^{\mathrm{short}}$ for the random spanning tree associated with $G^{\mathrm{short}}$,

$$P(e_2 \in \mathbf{T}^{\mathrm{short}}) = P(e_2 \in \mathbf{T} | e_1 \in \mathbf{T}).$$

But, setting $e_2 = (z_1, z_2)$, Proposition 14 shows

$$P(e_2 \in \mathbf{T}^{\mathrm{short}}) = w_{z_1 z_2} r^{\mathrm{short}}_{z_1 z_2}, \quad P(e_2 \in \mathbf{T}) = w_{z_1 z_2} r_{z_1 z_2}.$$

By Rayleigh's monotonicity principle, $r^{\mathrm{short}}_{z_1 z_2} \leq r_{z_1 z-2}$, and the result follows.

## 3 Self-verifying algorithms for sampling from a stationary distribution

To start with an analogy, we can in principle compute a mean hitting time $E_i T_j$ from the transition matrix $\mathbf{P}$, but we could alternatively estimate $E_i T_j$ by "pure simulation": simulate $m$ times the chain started at $i$ and run until hitting $j$, and then (roughly speaking) the empirical average of these $m$ hitting times will be $(1 \pm O(m^{-1/2})) E_i T_j$. In particular, for fixed $\varepsilon$ we can (roughly speaking) estimate $E_i T_j$ to within a factor $(1 \pm \varepsilon)$ in $O(E_i T_j)$ steps. Analogously, consider some notion of mixing time $\tau$ (say $\tau_1$ or $\tau_2$, in the reversible setting). The focus in this book has been on theoretical methods for bounding $\tau$ in terms of $\mathbf{P}$, and of theoretical consequences of such bounds. But can we bound $\tau$ by pure simulation? More importantly, in the practical context of Markov chain Monte Carlo, can we devise a "self-verifying" algorithm which produces an approximately-stationary sample from a chain in $O(\tau)$ steps without having prior knowledge of $\tau$?

xxx tie up with MCMC discussion.

To say things a little more carefully, a "pure simulation" algorithm is one in which the transition matrix $\mathbf{P}$ is unknown to the algorithm. Instead, there is a list of the states, and at each step the algorithm can obtain, for any state $i$, a sample from the jump distribution $p(i,\cdot)$, independent of previous samples.

In the MCMC context we typically have an exponentially large state space and seek polynomial-time estimates. The next lemma (which we leave to the reader to state and prove more precisely) shows that no pure simulation algorithm can guarantee to do this.

**Lemma 16** *Consider a pure simulation algorithm which, given any irreducible $n$-state chain, eventually outputs a random state whose distribution is guaranteed to be within $\varepsilon$ of the stationary distribution in variation distance. Then the algorithm must take $\Omega(n)$ steps for every $\mathbf{P}$.*

*Outline of proof.* If there is a state $k$ with the property that $1-p(k,k)$ is extremely small, then the stationary distribution will be almost concentrated on $k$; an algorithm which has some chance of terminating without sampling a step from every state cannot possibly guarantee that no unvisited state $k$ has this property. □

### 3.1 Exact sampling via the Markov chain tree theorem

Lovasz and Winkler [27] observed that the Markov chain tree theorem (Theorem 10) could be used to give a "pure simulation" algorithm for generating exactly from the stationary distribution of an arbitrary $n$-state chain. The algorithm takes

$$O(\tau_1^* \; n^2 \log n) \tag{14}$$

steps, where $\tau_1^*$ is the mixing time parameter defined as the smallest $t$ such that

$$P_i(X_{U_\sigma} = j) \geq \frac{1}{2}\pi_j \text{ for all } i,j \in I, \ \sigma \geq t \tag{15}$$

where $U_\sigma$ denotes a random time uniform on $\{0,1,\ldots,\sigma-1\}$, independent of the chain.

xxx tie up with Chapter 4 discussion and [26].

The following two facts are the mathematical ingredients of the algorithm. We quote as Lemma 17(a) a result of Ross [33] (see also [10] Theorem XIV.37); part (b) is an immediate consequence.

**Lemma 17** *(a) Let $\pi$ be a probability distribution on $I$ and let $(F_i; i \in I)$ be independent with distribution $\pi$. Fix $j$, and consider the digraph with edges $\{(i,F_i) : i \neq j\}$. Then with probability (exactly) $\pi_j$, the digraph is a tree with edges directed toward the root $j$.*

*(b) So if $j$ is first chosen uniformly at random from $I$, then the probability above is exactly $1/n$.*

As the second ingredient, observe that the Markov chain tree formula (Corollary 11) can be rephrased as follows.

**Corollary 18** *Let $\pi$ be the stationary distribution for a transition matrix* $\mathbf{P}$ *on $I$. Let $J$ be random, uniform on $I$. Let $(\xi_i; i \in I)$ be independent, with $P(\xi_i = j) = p_{ij}$. Consider the digraph with edges $\{(i, \xi_i) : i \neq J\}$. Then, conditional on the digraph being a tree with edges directed toward the root $J$, the probability that $J = j$ equals $\pi_j$.*

So consider the special case of a chain with the property

$$p^*_{ij} \geq (1/2)^{1/n} \pi_j \ \forall i, j. \tag{16}$$

The probability of getting any particular digraph under the procedure of Corollary 18 is at least 1/2 the probability of getting that digraph under the procedure of Lemma 17, and so the probability of getting some tree is at least $1/2n$, by Lemma 17(b). So if the procedure of Corollary 18 is repeated $r = \lceil 2n \log 4 \rceil$ times, the chance that some repetition produces a tree is at least $1 - (1 - 1/2n)^{2n \log 4} = 3/4$, and then the root $J$ of the tree has distribution exactly $\pi$.

Now for any chain, fix $\sigma > \tau_1^*$. The submultiplicativity (yyy) property of separation, applied to the chain with transition probabilities $\tilde{p}_{ij} = P_i(X_{U_\sigma} = j)$, shows that if $V$ denotes the sum of $m$ independent copies of $U_\sigma$, and $\xi_i$ is the state reached after $V$ steps of the chain started at $i$, then

$$P(\xi_i = j) \equiv P_i(X_V = j) \geq (1 - 2^{-m}) \pi_j \ \forall i, j.$$

So putting $m = -\log_2(1 - (1/2)^{1/n}) = \Theta(\log n)$, the set of probabilities $(P(\xi_i = j))$ satisfy (16).

Combining these procedures, we have (for fixed $\sigma > \tau_1^*$) an algorithm which, in a mean number $nm\sigma r = O(\sigma n^2 \log n)$ of steps, has chance $\geq 3/4$ to produce an output, and (if so) the output has distribution exactly $\pi$. Of course we initially don't know the right $\sigma$ to use, but we simply try $n, 2n, 4n, 8n, \ldots$ in turn until some output appears, and the mean total number of steps will satisfy the asserted bound (14).

### 3.2 Approximate sampling via coalescing paths

A second approach involves the parameter $\tau_0 = \sum_j \pi_j E_i T_j$ arising in the random target lemma (Chapter 2 yyy). Aldous [3] gives an algorithm which, given $\mathbf{P}$ and $\varepsilon > 0$, outputs a random state $\xi$ for which $||P(\xi \in \cdot) - \pi|| \leq \varepsilon$, and such that the mean number of steps is at most

$$81\tau_0/\varepsilon^2. \tag{17}$$

The details are messy, so let us just outline the (simple) underlying idea. Suppose we can define a procedure which terminates in some random number $Y$ of steps, where $Y$ is an estimate of $\tau_0$: precisely, suppose that for any $\mathbf{P}$

$$P(Y \leq \tau_0) \leq \varepsilon; \quad EY \leq K\tau_0 \tag{18}$$

where $K$ is an absolute constant. We can then define an algorithm as follows.

> Simulate $Y$; then run the chain for $U_{Y/\varepsilon}$ steps and output the final state $\xi$

where as above $U_\sigma$ denotes a random time uniform on $\{0, 1, \ldots, \sigma - 1\}$, independent of the chain. This works because, arguing as at xxx,

$$||P(X_{U_\sigma} \in \cdot) - \pi|| \leq \tau_0/\sigma$$

and so

$$||P(\xi \in \cdot) - \pi|| \ \leq E \max(1, \frac{\tau_0}{Y/\varepsilon}) \leq 2\varepsilon.$$

And the mean number of steps is $(1 + \frac{1}{2\varepsilon})EY$.

So the issue is to define a procedure terminating in $Y$ steps, where $Y$ satisfies (18). Label the states $\{1, 2, \ldots, n\}$ and consider the following *coalescing paths* routine.

(i) Pick a uniform random state $J$.

(ii) Start the chain at state 1, run until hitting state $J$, and write $A_1$ for the set of states visited along the path.

(iii) Restart the chain at state $\min\{j : j \not\in A_1\}$, run until hitting some state in $A_1$, and write $A_2$ for the union of $A_1$ and the set of states visited by this second path.

(iiii) Restart the chain at state $\min\{j : j \not\in A_2\}$, and continue this procedure until every state has been visited. Let $Y$ be the total number of steps.

The random target lemma says that the mean number of steps in (ii) equals $\tau_0$, making this $Y$ a plausible candidate for a quantity satisfying (18). A slightly more complicated algorithm is in fact needed – see [3].

### 3.3 Exact sampling via backwards coupling

Write $U$ for a r.v. uniform on $[0,1]$, and $(U_t)$ for an independent sequence of copies of $U$. Given a probability distribution on $I$, we can find a (far from unique!) function $f : [0,1] \to I$ such that $f(U)$ has the prescribed distribution. So given a transition matrix $\mathbf{P}$ we can find a function $f : I \times [0,1] \to I$ such that $P(f(i,U) = j) = p_{ij}$. Fix such a function. Simultaneously for each state $i$, define

$$X_0^{(i)} = i; \quad X_t^{(i)} = f(X_{t-1}^{(i)}, U_t), t = 1, 2, \ldots.$$

xxx tie up with coupling treatment

Consider the (forwards) coupling time

$$C^* = \min\{t : X_t^{(i)} = X_t^{(j)} \ \forall i, j\} \leq \infty.$$

By considering an initial state $j$ chosen according to the stationary distribution $\pi$,

$$\max_i ||P_i(X_t \in \cdot) - \pi|| \leq P(C > t).$$

This can be used as the basis for an approximate sampling algorithm. As a simple implementation, repeat $k$ times the procedure defining $C^*$, suppose we get finite values $C_1^*, \ldots, C_k^*$ each time, then run the chain from an arbitrary initial start for $\max_{1 \leq j \leq k} C_j^*$ steps and output the final state $\xi$. Then the error $||P(\xi \in \cdot) - \pi||$ is bounded by a function $\delta(k)$ such that $\delta(k) \to 0$ as $k \to \infty$.

Propp and Wilson [31] observed that by using instead a *backwards coupling* method (which has been exploited in other contexts – see Notes) one could make an exact sampling algorithm. Regard our i.i.d. sequence $(U_t)$ as defined for $-\infty < t \leq 0$. For each state $i$ and each time $s < 0$ define

$$X_s^{(i,s)} = i; \quad X_t^{(i,s)} = f(X_{t-1}^{(i,s)}, U_t), t = s+1, s+2, \ldots, 0.$$

Consider the *backwards* coupling time

$$C = \max\{t : X_0^{(i,t)} = X_0^{(j,t)} \ \forall i, j\} \geq -\infty.$$

**Lemma 19 (Backwards coupling lemma)** *If $S$ is a random time such that $-\infty < S \leq C$ a.s. then the random variable $X^{(i,S)}$ does not depend on $i$ and has the stationary distribution $\pi$.*

xxx describe algorithm

xxx poset story

xxx analysis in general setting and in poset setting.

xxx compare the 3 methods

# 4 Making reversible chains from irreversible chains

Let $\mathbf{P}$ be an irreducible transition matrix on $I$ with stationary distribution $\pi$. The following straightforward lemma records several general ways in which to construct from $\mathbf{P}$ a transition matrix $\mathbf{Q}$ for which the associated chain still has stationary distribution $\pi$ but is *reversible*. These methods all involve the time-reversed matrix $\mathbf{P}^*$

$$\pi_i p_{ij} = \pi_j p^*_{ji}$$

and so in practice can only be used when we know $\pi$ explicitly (as we have observed several times previously, in general we cannot write down a useful explicit expression for $\pi$ in the irreversible setting).

**Lemma 20** *The following definitions each give a transition matrix* $\mathbf{Q}$ *which is reversible with respect to* $\pi$.

*The additive reversiblization:* $\mathbf{Q}^{(1)} = \frac{1}{2}(\mathbf{P} + \mathbf{P}^*)$

*The multiplicative reversiblization:* $\mathbf{Q}^{(2)} = \mathbf{P}\mathbf{P}^*$

*The Metropolis reversiblization;* $\mathbf{Q}^{(3)}_{i,j} = \min(p_{i,j}, p^*_{j,i}),\ j \neq i.$

Of these three construction, only $\mathbf{Q}^{(1)}$ is automatically irreducible. Consider for instance the "patterns in coin tossing" example (Chapter 2 Example yyy). Here are the distributions of a step of the chains from state $(i_1, \ldots, i_n)$.

($\mathbf{Q}^{(1)}$). To $(i_2, \ldots, i_n, 0)$ or $(i_2, \ldots, i_n, 1)$ or $(0, i_1, \ldots, i_{n-1})$ or $(1, i_1, \ldots, i_{n-1})$, with probability 1/4 each.

($\mathbf{Q}^{(2)}$). To $(0, i_2, \ldots, i_n)$ or $(1, i_2, \ldots, i_n)$, with probability 1/2 each. So the state space decomposes into 2-element classes.

($\mathbf{Q}^{(3)}$). Here a "typical" $i$ is isolated.

We shall discuss two aspects of the relationship between irreversible chains and their reversibilizations.

## 4.1 Mixing times

Because the theory of $L^2$ convergence to stationarity is nicer for reversible chains, a natural strategy to study an irreversible chain (transition matrix $\mathbf{P}$) would be to first study a reversibilization $\mathbf{Q}$ and then seek some general result relating properties of the $\mathbf{P}$-chain to properties of the $\mathbf{Q}$-chain. There are (see Notes) general results relating spectra, but we don't pursue these because (cf. section 5) there seems no useful way to derive finite-time results for irreversible chains from spectral gap estimates.

xxx Persi, Fill etc stuff

### 4.2 Hitting times

Here are a matrix-theoretic result and conjecture, whose probabilistic significance (loosely relating to mean hitting times and reversiblization) will be discussed below. As usual $\mathbf{Z}$ is the fundamental matrix associated with $\mathbf{P}$, and $\mathbf{P}^*$ is the time-reversal.

**Proposition 21** *trace* $\mathbf{Z}(\mathbf{P}^* - \mathbf{P}) \geq 0$.

**Conjecture 22** *trace* $\mathbf{Z}^2(\mathbf{P}^* - \mathbf{P}) \geq 0$.

Proposition 21 is essentially due to Fiedler et al [18]. In fact, what is proved in ([18], p. 91) is that, for a positive matrix $\mathbf{V}$ with largest eigenvalue $< 1$,

$$\text{trace } (\sum_{m=1}^{\infty} \mathbf{V}^m)(\mathbf{V} - \mathbf{V}^T) \leq 0. \tag{19}$$

Applying this to $v_{ij} = s\pi_i^{1/2} p_{ij} \pi_j^{-1/2}$ for $s < 1$ gives

$$\text{trace } (\sum_{m=0}^{\infty} s^m (p_{ij}^{(m)} - \pi_j))(\mathbf{P} - \mathbf{P}^*) = \text{trace } (\sum_{m=0}^{\infty} s^m \mathbf{P}^{(m)})(\mathbf{P} - \mathbf{P}^*)$$

$$= s^{-1}\text{trace } (\sum_{m=0}^{\infty} \mathbf{V}^m)(\mathbf{V} - \mathbf{V}^T) \leq 0.$$

Letting $s \uparrow 1$ gives the Proposition as stated. □

The proof in [18] of (19) has no simple probabilistic interpretation, and it would be interesting to find a probabilistic proof. It is not clear to me whether Conjecture 22 could be proved in a similar way.

Here is the probabilistic interpretation of Proposition 21. Recall the elementary result (yyy) that in a $n$-state chain

$$\sum_a \sum_b \pi_a p_{ab} E_b T_a = n - 1. \tag{20}$$

The next result shows that replacing $E_b T_a$ by $E_a T_b$ gives an inequality. This arose as an ingredient in work of Tetali [37] discussed at xxx.

**Corollary 23** $\sum_a \sum_b \pi_a p_{ab} E_a T_b \leq n - 1$.

*Proof.* We argue backwards. By (20), the issue is to prove

$$\sum_a \sum_b \pi_a p_{ab} (E_b T_a - E_a T_b) \geq 0.$$

Using Lemma yyy, the quantity in question equals

$$\sum_a \sum_b \pi_a p_{ab} \left( \frac{Z_{aa}}{\pi_a} - \frac{Z_{ba}}{\pi_a} - \frac{Z_{bb}}{\pi_b} + \frac{Z_{ab}}{\pi_b} \right)$$

$$= \text{trace } \mathbf{Z} - \text{trace } \mathbf{PZ} - \text{trace } \mathbf{Z} + \text{trace } \mathbf{P}^* \mathbf{Z} = \text{trace } (\mathbf{P}^* - \mathbf{P}) \mathbf{Z} \geq 0.$$

□

Here is the motivation for Conjecture 22. For $0 \leq \lambda \leq 1$ let $\mathbf{P}(\lambda) = (1-\lambda)\mathbf{P} + \lambda \mathbf{P}^*$, so that $\mathbf{P}(1/2)$ is the "additive reversiblization" in Lemma 20. Consider the average hitting time parameters $\tau_0 = \tau_0(\lambda)$ from Chapter 4.

**Corollary 24** *Assuming Conjecture 22 is true,* $\tau_0(\lambda) \leq \tau_0(1/2)$ *for all* $0 \leq \lambda \leq 1$.

In other words, making the chain "more reversible" tends to increase mean hitting times.

*Proof.* This depends on results about differentiating with respect to the transition matrix, which we present as slightly informal calculations. Introduce a "perturbation" matrix $\mathbf{Q}$ such that

$$\sum_j q_{ij} = 0 \ \forall i; \quad q_{ij} = 0 \text{ whenever } p_{ij} = 0. \qquad (21)$$

Then $\mathbf{P} + \theta \mathbf{Q}$ is a transition matrix, for $\theta$ is some neighborhood of 0. Write $\frac{d}{d\theta}$ for the derivative at $\theta = 0$. Then, writing $N_i(t)$ for the number of visits to $i$ before time $t$,

$$\frac{d}{d\theta} \, E_a T_b = \sum_i E_a N_i(T_b) \, \sum_j q_{ij} E_j T_b.$$

This holds because the $\sum_j$ term gives the effect on $ET_b$ of a $\mathbf{Q}$-step from $i$. Using general identities from Chapter 2 yyy, and (21), this becomes

$$\frac{d}{d\theta} \, E_a T_b = \sum_i \left( \frac{\pi_i (z_{ab} - z_{bb})}{\pi_b} + z_{bi} - z_{ai} \right) \, \sum_j q_{ij} z_{jb} / \pi_b.$$

Now specialize to the case where $\pi$ is the stationary distribution for each $\mathbf{P} + \theta\mathbf{Q}$, that is where

$$\sum_i \pi_i q_{ij} = 0 \ \forall j.$$

Then the expression above simplifies to

$$\frac{d}{d\theta} \ E_a T_b = \sum_i (z_{bi} - z_{ai}) \ \sum_j q_{ij} z_{jb}/\pi_b.$$

Averaging over $a$, using $\sum_a \pi_a z_{ai} = 0$,

$$\frac{d}{d\theta} \ E_\pi T_b = \sum_i \sum_j z_{bi} q_{ij} z_{jb}/\pi_b$$

and then averaging over $b$,

$$\frac{d}{d\theta} \ \tau_0 = \text{ trace } \mathbf{ZQZ} = \text{ trace } \mathbf{Z}^2\mathbf{Q}.$$

So consider $\lambda < 1/2$ in Corollary 24. Then

$$\begin{aligned} \frac{d}{d\lambda}\tau_0(\lambda) &= \text{ trace } \mathbf{Z}^2(\lambda)(\mathbf{P}^* - \mathbf{P}) \\ &= (1-2\lambda)^{-1} \text{ trace } \mathbf{Z}^2(\lambda)(\mathbf{P}^*(\lambda) - \mathbf{P}(\lambda)) \end{aligned}$$

and Conjecture 22 would imply this is $\geq 0$, implying the conclusion of Corollary 24.

# 5 An example concerning eigenvalues and mixing times

Here is an example, adapted from Aldous [1]. Let $(\lambda_u : 1 \leq u \leq n)$ be the eigenvalues of $\mathbf{P}$ with $\lambda_1 = 1$, and let

$$\beta = \max\{|\lambda_u| : 2 \leq u \leq n\}.$$

A weak quantification of "mixing" is provided by

$$\alpha(t) \equiv \max_{A,B} |P_\pi(X_0 \in A, X_t \in B) - \pi(A)\pi(B)|.$$

By definition, $\alpha(t)$ is less than the *maximal correlation* $\rho(t)$ discussed in Chapter 4 yyy, and so by yyy

$$\alpha(t) \leq \beta^t \text{ for a reversible chain.} \tag{22}$$

The convergence theorem (Chapter 2 yyy) says that $\alpha(t) \to 0$ as $t \to \infty$ provided $\beta < 1$. So one might expect some analog of (22) to hold in general. But this is dramatically false: Example 26 shows

**Lemma 25** *There exists a family of $n$-state chains, with uniform stationary distributions, such that* $\sup_n \beta_n < 1$ *while* $\inf_n \alpha_n(n) > 0$.

Loosely, this implies there is no reasonable hypothesis on the spectrum of a $n$-state chain which implies an $o(n)$ mixing time. There is a time-asymptotic result

$$\alpha(t) \leq \rho(t) \leq C\beta^t \ \forall t,$$

for some $C$ depending on the chain. But implicit claims in the literature that bounding the spectrum of a general chain has some consequence for finite-time behavior should be treated with extreme skepticism!

**Example 26** Let $(Y_t)$ be independent r.v.'s taking values in $\{0, 1, \ldots, n-1\}$ with distribution specified by

$$P(Y \leq j) = \frac{j+1}{j+2}, \ 0 \leq j \leq n-2.$$

Define a Markov chain $(X_t)$ on $\{0, 1, \ldots, n-1\}$ by

$$X_t = \max(X_{t-1} - 1, Y_t).$$

This chain has the property (cf. the "patterns in coin-tossing" chain) of attaining the stationary distribution in finite time. Precisely: for any initial distribution $\sigma$, the distribution of $X_{n-1}$ is uniform, and hence $X_t$ is uniform for all $t \geq n-1$. To prove this, we simply observe that for $0 \leq j \leq n-1$,

$$\begin{aligned} P_\sigma(X_{n-1} \leq j) &= P(Y_{n-1} \leq j, Y_{n-2} \leq j+1, \ldots, Y_0 \leq j+n-1) \\ &= \frac{j+1}{j+2} \times \frac{j+2}{j+3} \times \ldots \times \frac{n-1}{n} \times \ 1 \times \ldots 1 \\ &= \frac{j+1}{n}. \end{aligned}$$

If $X_0$ is either 0 or 1 then $X_1$ is distributed as $Y_1$, implying that the vector $v$ with $v_i = 1_{(i=0)} - 1_{(i=1)}$ is an eigenvector of $\mathbf{P}$ with eigenvalue 0. By soft "duality" arguments it can be shown [1] that this is the largest eigenvalue, in the sense that

$$\mathcal{R}(\lambda_u) \leq 0 \text{ for all } 2 \leq u \leq n. \tag{23}$$

I believe it is true that

$$\beta_n = \max\{|\lambda_u| : 2 \leq u \leq n\}$$

is bounded away from 1, but we can avoid proving this by considering the "lazy" chain $\hat{X}_t$ with transition matrix $\hat{\mathbf{P}} = (\mathbf{I} + \mathbf{P})/2$, for which by (23)

$$\hat{\beta}_n \leq \sup\{|(1+\lambda)/2| : |\lambda| \leq 1, \mathcal{R}(\lambda) \leq 0\} = \sqrt{1/2}.$$

So the family of lazy chains has the eigenvalue property asserted in Lemma 25. But by construction, $X_t \geq X_0 - t$, and so $P(X_0 > 3n/4, X_{n/2} < n/4) = 0$. For the lazy chains we get

$$P_\pi(X_0 > 3n/4, X_n < n/4) \to 0 \text{ as } n \to \infty$$

establishing the (non)-mixing property asserted in the lemma.

# 6 Miscellany

## 6.1 Mixing times for irreversible chains

In Chapter 4 yyy we discussed equivalences between different definitions of "mixing time" in the $\tau_1$ family. Lovasz and Winkler [26] give a detailed treatment of analogous results in the non-reversible case.

xxx state some of this ?

## 6.2 Balanced directed graphs

Any Markov chain can be viewed as random walk on a weighted directed graph, but even on unweighted digraphs it is hard to relate properties on the walk to graph-theoretic properties, because (as we have often observed) it is in general hard to get useful information about the stationary distribution. An exception is the case of a *balanced* digraph, i.e. when the in-degree equals the out-degree ($= r_v$, say) at each vertex $v$. Random walk on a balanced digraph clearly retains the "undirected" property that the stationary probabilities $\pi_v$ are proportional to $r_v$. Now the proofs of Theorems yyy and yyy in Chapter 6 extend unchanged to the balanced digraph setting, showing that the cover-and-return time $C^+$ satisfies

$$\max_v E_v C^+ \leq n^3 \text{ in general; } \max_v E_v C^+ \leq 6n^2 \text{ on a regular balanced digraph.}$$

(The proofs rely on the edge-commute inequality (Chapter 3 yyy), rather than any "resistance" property).

### 6.3 An absorption time problem

Consider a Markov chain on states $\{1, 2, \ldots, n\}$ for which the only possible transitions are downward, i.e. for $i \geq 2$ we have

$$p(i,j) = 0, \ j \geq i$$

and $p(1,1) = 1$. The chain is ultimately absorbed in state 1. A question posed by Gil Kalai is whether there is a bound on the mean absorption time involving a parameter similar to that appearing in Cheeger's inequality. For each proper subset $A$ of $\{1, \ldots, n\}$ with $1 \notin A$ define

$$c(A) = \frac{|A||A^c|}{n \ \sum_{i \in A} \sum_{j \in A^c} p(i,j)}$$

and then define

$$\kappa = \max_A c(A).$$

**Open Problem 27** *Prove that* $\max_i E_i T_1$ *is bounded by a polynomial function of* $\kappa \log n$.

## 7 Notes on Chapter 9

*Section 1.* The idea of a maximal coupling goes back to Goldstein [21]: see Lindvall [25] for further history. Strong stationary times were studied in detail by Diaconis - Fill [17, 16] and Fill [19, 20], with particular attention to the case of one-dimensional stochastically monotone chains where there is some interesting "duality" theory. The special case of random walks on groups had previously been studied in Aldous - Diaconis [4, 5], and the idea is implicit in the regenerative approach to time-asymptotics for general state space chains, discussed at xxx. The theory surrounding Theorem 4 goes back to Rost [34]. This is normally regarded as part of the potential theory of

Markov chains, which emphasizes analogous results in the transient setting, and the recurrent case is rather a sideline in that setting. See Revuz [32] sec. 2.5 or Dellacherie - Meyer [14] Chapter 9 sec. 3 for textbook treatments in the general-space setting. The observation that the theory applied in simple finite examples such as those in section 1.3 was made in Lovasz - Winkler [26], from whom we borrowed the phrase *halting state*. Monotonicity properties like that in the statement of Corollary 5 were studied in detail by Brown [12] from the viewpoint of approximate exponentiality of hitting times.

*Section 2.* A slightly more sophisticated and extensive textbook treatment of these topics is in Lyons [28]. The nomenclature reflects my taste: Theorem 10 is "the underlying theorem" which implies "the formula" for the stationary distribution in terms of weighted spanning trees. Different textbooks (e.g. [22] p. 340 xxx more refs) give rather different historical citations for the Markov chain tree formula, and in talks I often call it "the most often rediscovered result in probability theory": it would be an interesting project to track down the earliest explicit statement. Of course it can be viewed as part of a circle of ideas (including the matrix-tree theorem for the number of spanning trees in a graph) which is often traced back to Kirchoff. The fact that Theorem 10 underlies the formula was undoubtably folklore for many years (Diaconis attributes it to Peter Doyle, and indeed it appears in an undergraduate thesis [36] of one of his students), but was apparently not published until the paper of Anantharam and Tsoucas [7]. The fact that the Markov chain tree theorem can be interpreted as an algorithm for generating uniform random spanning trees was observed by Aldous [2] and Broder [11], both deriving from conversations with Diaconis. [2] initiated study of theoretical properties of uniform random spanning trees, proving e.g. the following bounds on the diameter $\Delta$ of the random tree in a regular $n$-vertex graph.

$$\frac{n^{1/2}}{K_1 \tau_2 \log n} \leq E\Delta \leq K_2 \tau_2^{1/2} n^{1/2} \log n \tag{24}$$

where $K_1$ and $K_2$ are absolute constants. Loosely, "in an expander, a random spanning tree has diameter $n^{1/2 \pm o(1)}$". Results on asymptotic Poisson distribution for the degrees in a random spanning tree are given in Aldous [2], Pemantle [30] and Pemantle and Burton [13]. Pemantle [29] discusses the analog of uniform random spanning trees on the *infinite* $d$-dimensional lattice, and Aldous and Larget [6] give simulation results on quantitative behavior on the $d$-dimensional torus.

*Section 2.2.* As described in Pemantle [30] and Burton and Pemantle [13], the key to deeper study of random spanning trees is

**Theorem 28 (Transfer-impedance theorem)** *Fix a graph $G$. There is a symmetric function $H(e_1, e_2)$ on pairs of edges in $G$ such that for any edges $(e_1, \ldots, e_r)$*

$$P(e_i \in \mathbf{T} \textit{ for all } 1 \leq i \leq r) = \det\ M(e_1, \ldots, e_r)$$

*where $M(e_1, \ldots, e_r)$ is the matrix with entries $H(e_i, e_j), 1 \leq i, j \leq r$.*

*Section 3.* The first "pure simulation" algorithm for sampling exactly from the stationary distribution was given by Asmussen et al [8], using a quite different idea, and lacking explicit time bounds.

*Section 3.1.* In our discussion of these algorithms, we are assuming that we have a list of all states. Lovasz - Winkler [27] gave the argument in a slightly different setting, where the algorithm can only "address" a single state, and their bound involved $\max_{ij} E_i T_j$ in place of $\tau_1^*$.

*Section 3.3.* Letac [23] gives a survey of the "backwards coupling" method for establishing convergence of continuous-space chains: it suffices to show there exists a r.x. $X^{-\infty}$ such that $X_0^{(x,s)} \to X^{-\infty}$ a.s. as $s \to -\infty$, for each state $x$. This method is especially useful in treating matrix-valued chains of the form $X_t = A_t X_{t-1} + B_t$, where $(A_t, B_t), t \geq 1$ are i.i.d. random matrices. See Barnsley and Elton [9] for a popular application.

*Section 4.1.* One result on spectra and reversibilizations is the following. For a transition matrix $\mathbf{P}$ write

$$\tau(\mathbf{P}) = \sup\{\frac{1}{1 - |\lambda|} : \lambda \neq 1 \text{ an eigenvalue of } \mathbf{P}\}.$$

Then for the additive reversibilization $\mathbf{Q}^{(1)} = \frac{1}{2}(\mathbf{P} + \mathbf{P}^*)$ we have (e.g. [35] Proposition 1)

$$\tau(\mathbf{P}) \leq 2\tau(\mathbf{Q}^{(1)}).$$

# References

[1] D.J. Aldous. Finite-time implications of relaxation times for stochastically monotone processes. *Probab. Th. Rel. Fields*, 77:137–145, 1988.

[2] D.J. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM J. Discrete Math.*, 3:450–465, 1990.

[3] D.J. Aldous. On simulating a Markov chain stationary distribution when transition probabilities are unknown. In D.J. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, editors, *Discrete Probability and Algorithms*, volume 72 of *IMA Volumes in Mathematics and its Applications*, pages 1–9. Springer-Verlag, 1995.

[4] D.J. Aldous and P. Diaconis. Shuffling cards and stopping times. *Amer. Math. Monthly*, 93:333–348, 1986.

[5] D.J. Aldous and P. Diaconis. Strong uniform times and finite random walks. *Adv. in Appl. Math.*, 8:69–97, 1987.

[6] D.J. Aldous and B. Larget. A tree-based scaling exponent for random cluster models. *J. Phys. A: Math. Gen.*, 25:L1065–L1069, 1992.

[7] V. Anantharam and P. Tsoucas. A proof of the Markov chain tree theorem. *Stat. Probab. Letters*, 8:189–192, 1989.

[8] S. Asmussen, P.W. Glynn, and H. Thorisson. Stationarity detection in the initial transient problem. *ACM Trans. Modeling and Computer Sim.*, 2:130–157, 1992.

[9] M.F. Barnsley and J.H. Elton. A new class of Markov processes for image encoding. *Adv. in Appl. Probab.*, 20:14–32, 1988.

[10] B. Bollobás. *Random Graphs.* Academic Press, London, 1985.

[11] A. Broder. Generating random spanning trees. In *Proc. 30'th IEEE Symp. Found. Comp. Sci.*, pages 442–447, 1989.

[12] M. Brown. Consequences of monotonicity for Markov transition functions. Technical report, City College, CUNY, 1990.

[13] R. Burton and R. Pemantle. Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances. *Ann. Probab.*, 21:1329–1371, 1993.

[14] C. Dellacherie and P.-A. Meyer. *Probabilités et Potentiel: Théorie Discrète du Potentiel.* Hermann, Paris, 1983.

[15] P. Diaconis. *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics, Hayward CA, 1988.

[16] P. Diaconis and J.A. Fill. Examples for the theory of strong stationary duality with countable state spaces. *Prob. Engineering Inform. Sci.*, 4:157–180, 1990.

[17] P. Diaconis and J.A. Fill. Strong stationary times via a new form of duality. *Ann. Probab.*, 18:1483–1522, 1990.

[18] M. Fiedler, C.R. Johnson, T.L. Markham, and M. Neumann. A trace inequality for $M$-matrices and the symmetrizability of a real matrix by a positive diagonal matrix. *Linear Alg. Appl.*, 71:81–94, 1985.

[19] J.A. Fill. Time to stationarity for a continuous-time Markov chain. *Prob. Engineering Inform. Sci.*, 5:45–70, 1991.

[20] J.A. Fill. Strong stationary duality for continuous-time Markov chains. part I: Theory. *J. Theoretical Probab.*, 5:45–70, 1992.

[21] S. Goldstein. Maximal coupling. *Z. Wahrsch. Verw. Gebiete*, 46:193–204, 1979.

[22] H. Haken. *Synergetics.* Springer-Verlag, 1978.

[23] G. Letac. A contraction principle for certain Markov chains and its applications. In *Random Matrices and Their Applications*, volume 50 of *Contemp. Math.*, pages 263–273. American Math. Soc., 1986.

[24] T.M. Liggett. *Interacting Particle Systems.* Springer-Verlag, 1985.

[25] T. Lindvall. *Lectures on the Coupling Method.* Wiley, 1992.

[26] L. Lovász and P. Winkler. Efficient stopping rules for Markov chains. In *Proc. 27th ACM Symp. Theory of Computing*, page xxx, 1995.

[27] L. Lovász and P. Winkler. Exact mixing in an unknown Markov chain. *Electronic J. Combinatorics*, 2:#R15, 1995.

[28] R. Lyons. Probability and trees. Book in preparation, 1995.

[29] R. Pemantle. Choosing a spanning tree for the integer lattice uniformly. *Ann. Probab.*, 19:1559–1574, 1991.

[30] R. Pemantle. Uniform random spanning trees. In J. Laurie Snell, editor, *Topics in Contemporary Probability*, page xxx, Boca Raton, FL, 1994. CRC Press.

[31] J. Propp and D. Wilson. Exact sampling with coupled Markov chains. In preparation, 1995.

[32] D. Revuz. *Markov Chains.* North-Holland, second edition, 1984.

[33] S. M. Ross. A random graph. *J. Appl. Probab.*, 18:309–315, 1981.

[34] H. Rost. The stopping distributions of a Markov process. *Inventiones Math.*, 14:1–16, 1971.

[35] W.G. Sullivan. $L^2$ spectral gap and jump processes. *Z. Wahrsch. Verw. Gebiete*, 67:387–398, 1984.

[36] D.E. Symer. Expanded ergodic Markov chains and cycling systems. Senior thesis, Dartmouth College, 1984.

[37] P. Tetali. Design of on-line algorithms using hitting times. Bell Labs, 1994.

Reversibility doesn't help with this obstruction. A different approach is to seek to use coupling ideas.

zzz tie up with coupling discussion.

If we can specify a Markov coupling, then for any pair $(\mu, \nu)$ of initial distributions we can simulate the coupled processes and estimate the coupling time, and then the coupling inequality provides a self-verifying bound on the time $t$ taken for $||P_\mu(X_t \in \cdot) - P_\nu(X_t \in \cdot)||$ to become small. Unfortunately, in the context where we cannot simulate directly from the stationary distribution $\pi$, we would in general need to simulate coupled processes started from every state. But let us describe a very special setting where we can get away with coupling only two initial distributions. (This is motivated by a result of Propp - Wilson [31] described later).

Suppose we have a finite poset (*partially ordered* set) $I$ with partial order $\preceq$. Call $\mathbf{P}$ *algorithmically monotone* if we can explicitly find an (easily computable) function $f : I \times [0,1] \to I$ such that, writing $U$ for a r.v. uniformly distributed on $[0,1]$,

$$P(f(i,U) = j) = p_{ij} \tag{25}$$

$$\text{if } i \preceq j \text{ then } f(i,u) \preceq f(j,u) \ \forall u. \tag{26}$$

(This is closely related to the standard theoretical notion of a *monotone* $\mathbf{P}$: see Notes). For such a chain, we can use a sequence $(U_t; t \geq 1)$ of independent uniform r.v.'s to define, simultaneously for each $i$, versions $X_t^{(i)}; t \geq 0$ of the chain started at $i$:

$$X_0^{(i)} = i; \quad X_t^{(i)} = f(X_{t-1}^{(i)}, U_t), t \geq 1. \tag{27}$$

Now suppose that $I$ has a minimum element $i_*$ and a maximum element $i^*$, so that $i_* \preceq i \preceq i^*$ for all $i$. Write $X^*$ and $X_*$ for the chains $X^{(i^*)}$ and $X^{(i_*)}$ constructed above, and write

$$C_*^* = \min\{t : X^*(t) = X_*(t)\}.$$

By construction $X_*(t) \preceq X^{(i)}(t) \preceq X^*(t)$, and similarly $X_*(t) \preceq X^{(J)}(t) \preceq X^*(t)$, where $J$ denotes a random initial state picked according to the stationary distribution. So

$$X^{(i)}(t) = X^{(J)}(t) \text{ for all } t \geq C_*^*.$$

In other words, in terms of the definition $d(t)$ of variation distance from stationarity (Chapter 2 yyy),

$$d(t) \leq P(C_*^* > t). \tag{28}$$

It should now be clear that, in the setting of an algorithmically monotone chain with minimum and maximum states, we can use (28) to make an algorithm for sampling approximately from the stationary distribution. The point is that $C^*_*$ is the first meeting time for the joint process $(X^*(t), X_*(t))$,and so can be simulated. Here is one implementation.

*Step 1.* Simulate $k$ values of $C^*_*$, and let $\hat{\tau}$ be the maximum of these $k$ values.

*Step 2.* Start from an arbitrary state, simulate $2\hat{\tau}$ steps of the chain, output the final state. Repeat $k$ times, outputting $(Y_1, \ldots, Y_k)$.

This gives $k$ samples whose joint distribution is close to the joint distribution $\pi^{k*}$ of $k$ independent samples from $\pi$: precisely,

$$||\text{dist}(Y_1, \ldots, Y_k) - \pi^{k*}|| \leq \frac{1 + \log^2 k}{k}. \qquad (29)$$

zzz give argument

The key point is that this procedure requires no prior bound on mixing times: it's a "self-verifying" procedure. It is important to understand why this doesn't contradict Lemma 16: instead of just simulating the $\mathbf{P}$-chain, the procedure for obtaining $\hat{\tau}$ involves simulating the coupled chain whose definition involves the detailed structure of $\mathbf{P}$.

zzz strong hypothesis; examples in c-t

zzz gap between lemma and s-m case. Three results.

zzz comment on relationship between results, cost of exact vs approx.

zzz 3 results have same structure

*Notes* The partial order $\preceq$ on $I$ induces a partial order, say $\preceq^*$, on the set of probability distributions on $I$: this has several equivalent definitions, of which the "probabilistic" definition is

$$\mu_1 \preceq^* \mu_2 \text{ iff there exist } (X_{\mu_1}, X_{\mu_2}) \text{ with } \text{dist}(X_{\mu_u}) = \mu_u \text{ and } P(X_{\mu_1} \preceq X_{\mu_2}) = 1.$$

A Markov chain taking values in the poset $I$ is called *monotone* if the transition probabilities $p(i, \cdot)$ satisfy

$$\text{if } i \preceq j \text{ then } p(i, \cdot) \preceq^* p(j, \cdot).$$

There is a standard theory of "monotone couplings" for monotone chains: see Lindvall [25] Chapter IV. Our set-up differs in one or two ways from the standard theory. First, existence of a function $f$ satisfying (25,26) is equivalent to existence of a joint distribution $(Z_i; i \in I)$ with $P(Z_i = k) = p_{ik}$ and

$$\text{if } i \preceq j \text{ then } Z_i \preceq Z_j. \qquad (30)$$

But of course for algorithmic purposes we need an explicit $f$, rather than just existence of $f$. Secondly, monotonicity implies that we can achieve (30) for any specified pair $i \preceq j$, but it is not clear (to me) whether it implies existence of a whole family $(Z_i; i \in I)$ satisfying (30) for each pair $i \preceq j$.

Use of (28) is central to the analysis of *attractive spin systems*: see Liggett [24] Chapter 3.