Mixing Times

László Lovász[1] **and Peter Winkler**[2]

Abstract

The critical issue in the complexity of Markov chain sampling techniques has been "mixing time", the number of steps of the chain needed to reach its stationary distribution. It turns out that there are many ways to define mixing time—more than a dozen are considered here—but they fall into a small number of classes. The parameters in each class lie within constant multiples of one another, independent of the chain. Furthermore, there are interesting connections between these classes related to time reversal.

Contents

[1]Dept. of Computer Science, Yale University, New Haven, CT 06510; lovasz@cs.yale.edu
[2]Bell Laboratories, 700 Mountain Ave., Murray Hill, NJ 07974; pw@lucent.com

1 Introduction and preliminaries

In the past ten years there have been numerous applications (see, e.g., [3], [18], [17]) for sampling via finite Markov chains. A Markov chain is constructed from whose stationary distribution one wishes to sample, and then the chain is run for a fixed number of steps after which the distribution of the current state is "nearly" stationary. To determine this fixed number, the "mixing time", is usually the main difficulty in applying the method. Various methods have been developed to estimate the mixing time of a chain: eigenvalues, coupling, coupling from the past, conductance, strong stopping times, etc.

Various applications of this method require different interpretations of the above scheme. First of all, we do not always want to get close to the stationary distribution in the same sense; the measure of distance from the stationary distribution can be chosen in many ways, or rather, the specific needs of the application can lead to different measures here: total variation distance, ℓ_2-distance, χ^2-distance, entropy distance, pointwise filling up etc.

There are other variations of the problem raised by different applications. Perhaps the most basic set-up is when we want to generate a single state from the stationary distribution, starting from some fixed state (determined by the rest of the algorithm). This leads to the definition of the mixing time from a given state. If we don't have more information about the starting state, we have to use the maximum of this over all starting states, getting the basic definition of mixing time.

However, in some applications the starting state is already random, and we only want to use the Markov chain to improve its quality. In this case, one might hope that one can take advantage of the randomness already present. We call this situation the "warm start", and our mixing times will have "warm start" versions.

Quite often, we need to generate several independent (or approximately independent) samples from the stationary distribution. In this case we might start the second run of the Markov chain where the first one stopped, and so the expected time needed for this will be the average, rather than the maximum, of mixing times from individual states. This leads us to the definition of the *reset time*. (Note that this is different from warm start: we cannot make use of the randomness present in the starting state if we want independence.)

As a further example, we may use the Markov chain to find an element from a given, but not directly accessible subset of the state space. The worst expected time needed for this (normalized by the measure of the subset) is the *set hitting time*.

There is no particular reason why a walk must be run for a fixed number of steps; in fact, more general stopping rules, some of which "look where they are going", are capable of achieving the stationary distribution exactly. It turns out to be useful to consider stopping rules that achieve any given distribution, when starting from some other given distribution. From a theoretical standpoint, our stopping rules result in a generalization of the notion of "hitting time" to state-distributions. This measure of distance between distributions behaves quite nicely.

(We put no restriction on the amount of computation needed to implement a stopping rule, making the use of these rules as sampling mechanisms unlikely; but such rules are useful in analysis and, as we shall see, often replaceable by simpler ones without too much loss in efficiency.)

By considering the least expected number of steps required by rules of a particular kind to reach the stationary distribution (or some approximation thereof) we obtain further measures of mixing time. To these we may add some additional parameters, related to eigenvalues and the reverse of the given chain. Altogether we obtain a substantial collection of parameters each of which has mixing time implications.

Our objective is to show that the situation is not so hopelessly complex. The main result is that we can place these numbers into a few equivalence classes, within which numbers differ only by constant factors independent of the chain (see Section 3). For the case of reversible chains, such results were obtained by Aldous [3]; other results in this direction were published in [6, 23, 24]. To prove the main results, we develop a calculus of "exit frequencies", study the behavior of mixing properties under time reversal, and make use of some linear algebraic tools. We hope that these tools shed some light on the mechanism of mixing.

1.1 Preliminaries

Throughout this paper we will assume a fixed irreducible Markov chain with transition matrix $M = \{p_{ij}\}$, with a finite state space V of finite cardinality n (see [6] for extensions to infinite state space). If we say "distribution" without specifying an underlying set, we mean distribution on V. If σ is a distribution on V and $A \subseteq V$, then $\sigma(A) = \sum_{i\in A} \sigma_i$ is the probability of A. If $\sigma(A) > 0$, then we can consider the distribution σ_A defined by

$$(\sigma_A)_i = \begin{cases} \dfrac{\sigma_i}{\sigma(A)}, & \text{if } i \in A, \\ 0, & \text{otherwise.} \end{cases}$$

We denote by π the stationary distribution of the chain. The *density* of a distribution σ is $\max_{i\in V} \sigma_i/\pi_i$.

Most of the time, we use the *total variation metric* to describe the distance of two distributions. This is defined by

$$d(\sigma,\tau) = \max_{A\subseteq V} (\sigma(A) - \tau(A)).$$

It is easy to see that $d(\sigma,\tau) = d(\tau,\sigma)$, and this value is just half of the ℓ_1-norm of $\|\sigma - \tau\|$.

Given a Markov chain with transition probabilities p_{ij}, we define the *reverse chain* as the Markov chain on the same set of states, with transition probabilities $\overleftarrow{p}_{ij} = \pi_j p_{ji}/\pi_i$. We will generally use the reverse arrow over a symbol to indicate that it refers to this reverse chain.

It is trivial but important that the reverse chain has the same stationary distribution as the original. This follows by straightforward substitution. It turns out that there is a very close relationship between mixing properties of a chain and its reverse.

1.2 Examples: stopping rules

In this section we discuss examples which show that "intelligent" stopping rules can sometimes achieve specified distributions in an elegant or surprising manner.

Example 1.1 (Cycle) The following is an interesting fact from folklore. Let G be a cycle of length n and start a random walk on G from a node u. Then the probability that v is the last node visited (i.e., the a random walk visits every other node before hitting v) is the same for each $v \neq u$.

While this is not an efficient way to generate a uniform random points of the cycle, it indicates that there are entirely different ways to use random walks for sampling than walking a given number of steps. This particular method does not generalize; in fact, apart from the complete graph, the cycle is the only graph which enjoys this property (see [21]).

Example 1.2 (Cube) Consider another quite simple graph, the cube, which we view as the graph of vertices and edges of $[0,1]^n$. Let us do a random walk on it as follows: at each vertex, we select a direction (that is, one of the n coordinate indices) at random, then flip a coin. If we get "heads" we walk along the incident edge corresponding to that direction; if "tails" we stay where we are. We stop when we have selected every direction at least once (whether or not we walked along the edge).

It is trivial that after each of the n directions has been selected, the corresponding coordinate will be 0 or 1 with equal probability, independently of the rest of the coordinates. So the vertex we stop at will be uniformly distributed over all vertices.

This method takes about $n \ln n$ coin flips on the average, thus about $n \ln n/2$ actual steps, so it is a quite efficient way to generate a random vertex of the cube (assuming we insist on using random walks; otherwise choosing the coordinates independently is simpler and faster). We will see that it is in fact optimal.

Example 1.3 (Card Shuffling) A classic application of Markov chain mixing is shuffling a deck of playing cards; see e.g. [9] where Bayer and Diaconis argue that seven riffle shuffles are necessary and sufficient to mix a deck of 52 cards. In Aldous and Diaconis [5], the following simple shuffling algorithm is analyzed: a card is removed from the top of an n-card deck and replaced with equal probability in any of the n slots among the remaining $n-1$ cards.

It is not difficult to see that if we note when the card originally at the bottom of the deck has reached the top and perform just one more shuffle, then the deck will be precisely uniformly random. Again, this method takes about $n \ln n$ steps on the average, and also as in the cube case, constitutes a "strong" stopping rule in the sense of [5] (the ending distribution is uniform even when conditioned on the stopping rule having taken some fixed number of steps). However, we will see later that this shuffling rule is *not* optimal.

Without going into details, let us remark that the algorithm of Aldous [4] and Broder [11] can be regarded as a stopping rule for a Markov chain on trees that generates the uniform distribution.

Finally, we note that to stop after a fixed number of steps in the "continuous" time model of Markov chains (where the time needed by a step is an exponentially distributed random variable) can be viewed as choosing a random number T from a Poisson distribution and stopping after T steps. So here a (randomized) stopping rule is considered which, in many respects, has better properties than the "stop after t steps" rule. A similar remark applies to stopping a "lazy" walk as defined, e.g., in [20].

2 Access times and mixing times

2.1 General stopping rules

We denote by V^* be the space of finite "walks" on V, that is, the set of finite strings $w = (w_0, w_1, \ldots, w_t)$, $w_i \in S$. Relative to the Markov chain we then have

$$\mathsf{P}(w) = \sigma_{w_0} \prod_{i=0}^{t-1} p_{w_i, w_{i+1}}\,.$$

The distribution of w_t will be denoted by σ^t, so that $\sigma^0 = \sigma$ and $\sigma^t_i = \sum_{w_t = i} \mathsf{P}((w_0, \ldots, w_t))$.

A *stopping rule* Γ is a rule that observes the walk and tells us whether to stop or not, depending on the walk seen so far (but independently of the continuation of the walk). This decision may be reached using coin flips; so the stopping rule just has to specify, for each walk w, the probability of continuing the walk. Formally, Γ is a map from V^* to $[0,1]$ (it would be enough to define Γ for walks w with $\mathsf{P}(w) > 0$). We interpret $\Gamma(w)$ as the probability of continuing given that w is the walk so far observed, each such stop-or-go decision being made independently. We assume that with probability 1 the walk eventually stops.

We can also regard Γ as a *stopping time*, i.e., a random variable with values in $\{0, 1, \ldots\}$, so that we stop at w_Γ.

The probability of stopping at state j, given starting distribution σ, is

$$\sigma^\Gamma_j := \sum_{w = (w_0, \ldots, w_t = j)} \sigma_{w_0} \left(\prod_{i=0}^{t-1} \Gamma(w_0, \ldots, w_i) p_{w_i, w_{i+1}} \right) (1 - \Gamma(w))\,.$$

We often think of a stopping rule Γ as a means of moving from a starting distribution σ to a given target distribution $\tau = \sigma^\Gamma$; we say then that Γ is a stopping rule from σ to τ.

For any Markov chain and any two distributions σ and τ, there is at least one finite stopping rule Γ from σ to τ; namely, we select a random target state j in accordance with τ and walk until we reach j. We call this the "naive" stopping rule $\Omega_{\sigma,\tau}$.

Note that if $\mathbf{X}$ is a random variable whose values are stopping rules, then $\mathbf{X}$ is equivalent to a stopping rule Γ whose probability of continuing at w is a sum of $\mathbf{X}(w)$ conditioned on $\mathbf{X}$ not having stopped the chain so far. Thus our stopping rules are not less general for insisting on independent randomization at each step. For example, the stopping rule given above in Example 1.2 (Cube) does not at first appear to fit our model, but we may simply walk on the (loopless) cube and stop according to the probability that the "draw and flip" rule would stop, given that it reached the current time and state.

The *mean length* $\mathsf{E}_\sigma \Gamma$ of the stopping rule Γ is its expected duration, starting from the distribution σ. Most often we'll have the distribution σ tacitly included with the stopping rule, and we just write $\mathsf{E}\Gamma$.

If $\mathcal{H}(i,j)$ denotes the expected hitting time (mean number of steps to reach j from i), then the mean length of $\Omega_{\sigma,\tau}$ is given by

$$\mathcal{N}(\sigma, \tau) = \sum_{i,j} \sigma_i \tau_j \mathcal{H}(i,j)\,.$$

When the target is the stationary distribution π for the chain, $\mathsf{E}\Omega_{\sigma,\pi}$ is independent of the starting distribution σ, on account of what we call the "Random Target Identity" (also known as the right averaging principle, e.g. in [1]) which says that there is a constant $\mathcal{N}$ for which

$$\mathcal{N}(\sigma, \pi) = \mathcal{N} \tag{1}$$

for every starting distribution σ. In particular, we have

$$\mathcal{N} = \sum_j \pi_j \mathcal{N}(\pi, j) = \sum_j \pi_j \mathcal{H}(i, j) \tag{2}$$

for every state i.

2.2 Access times

A stopping rule Γ is said to be *mean-optimal* or simply *optimal* (for σ and τ) if $\mathsf{E}\Gamma$ is minimal. The mean length of a mean-optimal stopping rule from σ to τ will be called the *access time* from σ to τ and denoted $\mathcal{H}(\sigma, \tau)$. As suggested by the notation, we think of $\mathcal{H}(\sigma, \tau)$ as a generalized hitting time. It will turn out that $\mathcal{H}(\sigma, \tau)$ leads to the most important mixing measures we are going to study.

The access time between distributions behaves in many respects as a "distance" between the two distributions. Trivially, $\mathcal{H}(\sigma, \tau) = 0$ if and only if $\sigma = \tau$. It is easy to see that the following *triangle inequality* is satisfied for any three distributions α, β and γ:

$$\mathcal{H}(\alpha, \gamma) \leq \mathcal{H}(\alpha, \beta) + \mathcal{H}(\beta, \gamma)\,. \tag{3}$$

(To generate γ from α, we can first use an optimal rule to generate β from γ and then use the state obtained as a starting state for an optimal rule generating γ from β). We should warn the reader, however, that $\mathcal{H}(\sigma, \tau) \neq \mathcal{H}(\tau, \sigma)$ in general.

Thus the access time $\mathcal{H}(\sigma, \tau)$ has the properties of a metric on the space of distributions, except for symmetry; the latter is of course too much to expect since the ordinary hitting time, even for a reversible chain, is not generally symmetric.

If τ is concentrated at state j (for which we write, rather carelessly, "$\tau = j$") then

$$\mathcal{H}(\sigma, j) = \sum_i \sigma_i \mathcal{H}(i, j) \tag{4}$$

since clearly the only optimal stopping rule in this case is $\Omega_{i,j}$, "walk until state j is reached." This is therefore both mean-optimal and max-optimal, although $\max(\Omega_{i,j})$ will generally be infinite. By considering the naive rule Ω_τ, we get the inequality

$$\mathcal{H}(\sigma, \tau) \leq \mathcal{N}(\sigma, \tau) = \sum_{i,j} \sigma_s \tau_j H(i, j)\,. \tag{5}$$

This may be quite far from equality; for example, $\mathcal{H}(\sigma, \sigma) = 0$ for any σ.

The following formula for access times was given in [22]:

Theorem 2.1 *For all distributions σ and τ,*

$$\mathcal{H}(\sigma, \tau) = \max_j \Big(\mathcal{H}(\sigma, j) - \mathcal{H}(\tau, j)\Big)\,.$$

By (4), we can write this as

$$\mathcal{H}(\sigma,\tau) = \max_j \sum_i (\sigma_i - \tau_i)\mathcal{H}(i,j)\,.$$

The inequality $\geq$ in the theorem is trivial by the triangle inequality. The reverse inequality depends on a simple characterization of optimal rules. For any rule $\Gamma:\ \sigma \to \tau$, a state j is called a *halting state*, if it is never exited provided the rule Γ is obeyed. Thus by definition we stop immediately if and when any halting state is entered. (Of course we may stop in other states too, just not all the time.)

Theorem 2.2 *A stopping rule is mean-optimal if and only if it has a halting state.*

The maximum in Theorem 2.1 is achieved precisely when j is a halting state of some optimal rule which stops at τ when started at σ.

In the case of time-reversible chains, we can use the cycle-reversing identity from [14] to obtain the following expression:

$$\mathcal{H}(\sigma,\tau) = \sum_i (\tau_i - \sigma_i)\mathcal{H}(\pi,i) - \min_j \sum_i (\tau_i - \sigma_i)\mathcal{H}(j,i)\,.$$

We will be concerned mostly with the case when $\tau = \pi$ and σ is concentrated at a single state i. In this case this formula can be simplified using the Random Target Identity (1) to get that

$$\mathcal{H}(i,\pi) = \max_j \mathcal{H}(j,i) - \mathcal{H}(\pi,i)\,. \tag{6}$$

Thus the halting state of a reversible chain, in attaining the stationary distribution from a fixed state i, is the state j *from which* the mean time to hit i is greatest.

The quantity $\mathcal{H}(\sigma,j) - \mathcal{H}(\tau,j)$ in Theorem 2.1 turns out to be a very useful measure of the "distance" between σ and τ. Instead of the maximum in j, we can also introduce the average

$$\mathcal{Z}(\sigma,\tau) = \sum_j \pi_j |\mathcal{H}(\sigma,j) - \mathcal{H}(\tau,j)|$$

Note that $\sum_j \pi_j(H(\sigma,j) - H(\tau,j)) = 0$ by the Random Target Identity, and hence we could replace the absolute values by the "surpluses" $\max\{0, H(\sigma,j) - H(\tau,j)\}$, at the cost of a factor of exactly 2. It is obvious from this definition that $\mathcal{Z}(\sigma,\tau) \leq \mathcal{H}(\sigma,\tau)$.

2.3 Different mixing measures: definitions

We can use the notion of the access time $\mathcal{H}(\sigma,\tau)$ between distributions to define a parameter-free notion of mixing time. In fact, as described in the introduction, we have several different, but reasonably well-motived ways to do so, depending on assumptions about the starting distribution, approximation of the target distribution, and possible additional requirements like independence of the starting and ending state.

Notation. Before defining some of these notions of mixing time, it will be useful to establish a reasonably consistent notation scheme. We will use calligraphic letters (like $\mathcal{H}$ or $\mathcal{N}$) to denote mixing measures with different letters often used to distinguish classes of stopping rules. The dependence of a mixing measure on the (fixed) transition matrix of a Markov chain will always be understood.

As before we permit initial and target distributions as parameters, e.g. $\mathcal{H}(\sigma, \tau)$ for the access time from σ to τ, but we will extend the scope of the arguments even further, to sets of distributions. If A and B are two nonempty sets of distributions, we define

$$\mathcal{H}(A, B) := \max_{\alpha \in A} \min_{\beta \in B} \mathcal{H}(\alpha, \beta).$$

(and analogously for other mixing measures).

Thus $\mathcal{H}(A, B) \leq t$ means that for every starting distribution in A there is a stopping rule with mean length at most t that yields a distribution in B. For example, $\mathcal{H}(V, \leq 2\pi)$ would be the minimum expected time needed to reach a distribution $\tau \leq 2\pi$ from a worst single state (equivalently, in this case, from a worst distribution of states). Note that in this notation, the triangle inequality extends to access times between sets:

$$\mathcal{H}(A, C) \leq \mathcal{H}(A, B) + \mathcal{H}(B, C). \tag{7}$$

Often mixing times will be computed from the worst starting distribution among *all* distributions—as in the above case this is usually equivalent to worst *state*—and that will be the assumption if there is only one argument to $\mathcal{H}$. Thus

$$\mathcal{H}(B) := \mathcal{H}(V, B) = \max_{s \in V} \min_{\tau \in B} \mathcal{H}(s, \tau).$$

If there are *no* arguments to $\mathcal{H}$ then the starting distribution is worst case and the target distribution π, as it is in most applications. In symbols:

$$\mathcal{H} := \mathcal{H}(\pi) = \max_{\sigma} \mathcal{H}(\sigma, \pi) = \max_{s \in V} \mathcal{H}(s, \pi) = \mathcal{H}(V, \pi).$$

We bestow upon $\mathcal{H}$ the honor of calling it *the* mixing time of the chain. The state s attaining the maximum in the definition of $\mathcal{H}$ will be called a *pessimal* state.

We have already seen the notation

$$\mathcal{N} := \mathcal{N}(\pi) := \mathcal{N}(\sigma, \pi)$$

for the naive rule and the constant in the Random Target Identity, where σ is any distribution.

Approximate mixing. Often we want to reach only a distribution close to π. Let $d_\varepsilon(\tau)$ denote the ball of radius ε and center τ in the total variation metric. Most often we consider the case $\tau = \pi$, and we set $d_\varepsilon = d_\varepsilon(\pi)$.

We indicate pointwise approximation from above and from below by a bar over or under the subscript, and define

$$D_{\overline{\varepsilon}} = \{\sigma \in D : \ \sigma \leq (1+\varepsilon)\pi\},$$

$$D_{\underline{\varepsilon}} = \{\sigma \in D : \ \sigma \geq (1-\varepsilon)\pi\},$$

and

$$D_\varepsilon = D_{\overline{\varepsilon}} \cap D_{\underline{\varepsilon}} = \{\sigma \in D : \ (1-\varepsilon) \leq \sigma \leq (1+\varepsilon)\pi\}.$$

The set $D_{\overline{1}} = \{\sigma : \sigma \leq 2\pi\}$ will play a particularly important role.

We define

$$\mathcal{H}_\varepsilon := \mathcal{H}(D, d_\varepsilon) = \max_{s} \min_{\tau:\ d(\tau,\pi)<\varepsilon} \mathcal{H}(s, \tau).$$

In terminology, $\mathcal{H}_\varepsilon$ becomes the "approximate mixing time". We also define

$$\mathcal{H}_{\overline{\varepsilon}} := \mathcal{H}(D, \leq (1+\varepsilon)\pi) = \max_{s \in V} \min_{\tau \leq (1+\varepsilon)\pi} \mathcal{H}(s, \tau)$$

and analogously

$$\mathcal{H}_{\underline{\varepsilon}} := \mathcal{H}(D, \geq (1-\varepsilon)\pi).$$

The former will be called the "dispersion time" and the latter the "filling time." At this point, the distinction between mixing in the "filling", "disperse" and "total variation" sense may seem pedantic, but in fact these three mixing measures behave quite differently.

Warm start. We also consider the effect of a "warm start", where the starting distribution σ has bounded density (most often density at most 2, which is an arbitrary but convenient choice). We will indicate a warm start in our mixing time notation by placing a tilde over the calligraphic letter.

We define

$$\tilde{\mathcal{H}} := \mathcal{H}(D_1, \pi) = \max_{\sigma \leq 2\pi} \mathcal{H}(\sigma, \pi)$$

and

$$\tilde{\mathcal{H}}_\varepsilon := \mathcal{H}(D_1, d_\varepsilon) = \max_{\sigma \leq 2\pi} \min_{\substack{\tau \\ d(\tau,\pi) < c}} \mathcal{H}(\sigma, \tau),$$

Another, somewhat different way of handling warm start is to consider a non-empty set $S \subseteq V$ of states, and start from the stationary distribution, but condition on starting in S. We are then interested in, say, $\mathcal{H}(\pi_S, \pi)$. Of course, this quantity increases if S gets smaller, so it is natural to scale and consider

$$\hat{\mathcal{H}} = \max_S \pi(S) H(\pi_S, \pi).$$

We call $\mathcal{H}$ the *mixing time with restricted start*, or briefly *restricted mixing time*. Again, we can define the restricted version of $\mathcal{H}_\varepsilon$ etc.

Blind rules. A general stopping rule, which may make use of a complete knowledge of the chain, and may do an unlimited amount of computation to decide when to stop, is useless from a practical point of view. Therefore we also introduce a type of stopping rule which, while not capable of achieving arbitrary target distributions, and almost never optimal, is easily implementable and will often provide a good approximation of more sophisticated stopping rules. We call a stopping rule Γ *blind* if $\Gamma(w)$ depends only on the length $|w|$ of the walk.

The simplest blind stopping rule is the stopping rule "stop after t steps". We call this the *t-step-rule*. Several other practical methods to generate elements from the stationary distribution (approximately) can also be viewed as blind rules. For example, the *uniform averaging rule* (to be discussed below) of walking u steps and then choosing one of the exited points uniformly can be viewed as a blind rule: after t steps, we stop with probability $1/(u-t)$. *Lazy* random walks have been considered (see Lovász and Simonovits [20]) because they have better convergence to the stationary distribution; the lazy version of a Markov chain is obtained by flipping a coin before each move and staying where we are if we see "heads". Stopping the lazy version of a Markov chain after s steps is equivalent to following the original walk $(w_0, w_1, \ldots, w_u)$ for u steps and then choosing a w_t according to the binomial distribution, i.e. with probability $\binom{u}{t}2^{-u}$. This is again equivalent to a blind rule, where we stop after t steps with probability

$$\binom{u}{t} \Bigg/ \left(\binom{u}{t} + \binom{u}{t+1} + \ldots + \binom{u}{u} \right).$$

It is often more convenient to describe a blind stopping rule by the probabilities a_t that state v_t is selected (not conditioning on not having stopped before). Trivially $a_t \geq 0$ and the finite termination of the rule is equivalent to $\sum_t a_t = 1$. The rule is bounded if and only if the sequence a_t has a finite number of non-0 terms. Thus a blind rule can always be thought of as an averaging, using the distribution (a_t).

One cannot generate any distribution by a blind stopping rule; for example, starting from the stationary distribution, every blind rule generates the stationary distribution itself. Blind stopping rules will be used to generate the stationary distribution, or at least approximations of it.

Unfortunately, blind rules to achieve the stationary distribution exist only for a rather restricted class of chains, as the following (unpublished) result shows, that we state without proof.

Theorem 2.3 *There is a finite blind stopping rule that generates π from any starting state if and only if no eigenvalue of the transition matrix is a positive real number.*

Interestingly, the condition formulated in the theorem is most restrictive for time-reversible chains; then all the eigenvalues are real, and typically many of them are positive. Only very special Markov chains admit a blind rule for generating π which is mean-optimal.

We introduce the letter $\mathcal{B}$ to denote mixing time using blind rules, so that

$$\mathcal{B}(\sigma,\tau) := \min_{\Gamma \text{ blind},\ \sigma^\Gamma=\tau} \mathsf{E}\Gamma$$

and for distribution-sets A and B,

$$\mathcal{B}(A,B) := \max_{\alpha\in A}\min_{\beta\in B} \mathcal{B}(\alpha,\beta)\,.$$

This quantity may well be infinite (for example, when $A = \{\pi\}$ and B consists of a single state different from π). But blind rules will be important in approximate mixing. In particular, we define the "blind approximate mixing time"

$$\mathcal{B}_\varepsilon := \mathcal{B}(d_\varepsilon(\pi)) = \max_{\sigma} \mathcal{B}(\sigma, d_\varepsilon(\pi)).$$

We can further restrict the rules used in the definition of these mixing times, to get closer to practically implementable algorithms. To the letter $\mathcal{B}$, reserved for blind rules, we can add the even more restrictive $\mathcal{U}$ for uniform averaging rules. The exact forms $\mathcal{B}(\sigma,\tau)$ and $\mathcal{U}(\sigma,\tau)$ may be infinite, but the approximate forms, e.g. $\mathcal{U}_\varepsilon$, make sense.

We define the warm *blind* and *uniform* mixing times $\tilde{\mathcal{B}}_\varepsilon$ and $\tilde{\mathcal{U}}_\varepsilon$ in a manner analagous to $\tilde{\mathcal{H}}_\varepsilon$ but for blind and uniform stopping rules.

The maximum number of steps. The *maximum length* $\max(\Gamma)$ of a stopping rule Γ is the maximum length of a walk that has positive probability. Γ is said to be *bounded* if this quantity is finite. The naive rule is in general not bounded, but bounded stopping rules are available when the target distribution τ has sufficiently large support, thus in particular when $\tau = \pi$ (see [22]).

A stopping rule is *max-optimal* (for given σ and τ) if $\max(\Gamma)$ is minimal. The maximum length of a max-optimal stopping rule from σ to τ will be denoted by $\mathcal{M}(\sigma,\tau)$. Thus

$$\mathcal{M}(\sigma,\tau) := \min_{\Gamma:\ \sigma^\Gamma=\tau} \max(\Gamma)\,.$$

Then

$$\mathcal{M} := \mathcal{M}(V,\pi) := \max_\sigma \mathcal{M}(\sigma,\pi)$$

is the related mixing measure, which we cannot resist calling the "maxing time" of the chain.

Independent rules. In many applications repeated independent samples are required from the stationary distribution of a chain. If we use a stopping rule to generate these states, we need to make the following assumption about the rules. A stopping rule is said to be *independent* if the state at which it stops the chain is independent of the one in which the chain was started. If we write

$$\mathcal{I}(\sigma,\tau) := \min\{\mathsf{E}\Gamma :\ \Gamma \text{ independent}\}\ ,$$

then we have

$$\mathcal{I}(\sigma,\tau) = \sum_{i\in V} \sigma_i \mathcal{H}(i,\tau)$$

since an independent rule must produce the target distribution "separately" from each starting state.

To get the time needed to generate an independent sample state, we let both the default initial and target distributions for an independent rule be π:

$$\mathcal{I} := \mathcal{I}(\pi,\pi) = \sum_i \pi_i \mathcal{H}(i,\pi)$$

obtaining what we call the *reset time* of the chain (see [24]). This is only formally similar to Formula (2) for $\mathcal{N}$; in general, $\mathcal{N}$ is much larger. (Instead of π, one might want to generate independent samples from some other distribution. This leads to the notion of the *regeneration time* $\max_\alpha \mathcal{I}(\alpha,\alpha)$, studied in [10]; however, this turns out to be closely related to commute times rather than mixing times, and will not be discussed here.)

We also introduce an approximate version of the independence time. We say that two random variables X and Y (with values from a set V) are *ε-independent*, if for every two sets $A \subseteq V$ and $B \subseteq V$, we have

$$\Big|\mathsf{P}(X\in A, Y\in B) - \mathsf{P}(X\in A)\mathsf{P}(Y\in B)\Big| \le \varepsilon\,.$$

This is a very weak notion of independence, but in some applications of Markov chain techniques to sampling [19], this is exactly what is needed.

Let the *warm reset time* $\tilde{\mathcal{I}}_\varepsilon$ be the smallest t such that for every starting distribution σ with $\sigma \le (1+\varepsilon)\pi$, there exists a stopping rule Γ with $\mathsf{E}\Gamma \le t$ such that if v^0 is from σ then v^Γ is ε-independent of the starting state, and $\sigma^\Gamma \le (1+\varepsilon)\pi$.

Hitting sets. Another very natural extension of hitting time is1 set-hitting time: for $U \subseteq V$, $U \ne \emptyset$, let $\mathcal{H}(s,U)$ be the expected number of steps before hitting U, when starting from s. More generally, if σ is any starting distribution, then we define $\mathcal{H}(\sigma,U)$ as the expected number of steps before hitting U, when starting from a random node drawn from σ; clearly, $H(\sigma,U) = \sum_i \sigma_i H(i,U)$. For two distributions σ and τ, we define

$$\mathcal{S}(\sigma,\tau) = \max_{A\subseteq V} \tau(A)\mathcal{H}(\sigma,A)\ .$$

It is easy to see that if $\sigma' \le A\sigma$ and $\tau' \le B\tau$, then

$$\mathcal{S}(\sigma',\tau') \le AB\mathcal{S}(\sigma,\tau). \tag{8}$$

The *set access time* $\mathcal{S}$ is the mean number of steps required to hit the "toughest" set of states (adjusted by the stationary probability of the set) from the worst start:

$$\mathcal{S} := \max_s \mathcal{S}(s,\pi) = \max_{s\in V,\, U\subseteq V} \pi_U \mathcal{H}(s,U)\,.$$

We can define the warm start version if the set access time simply as $\tilde{\mathcal{S}} = \mathcal{S}(\pi,\pi)$. It is immediate that if $\sigma \le 2\pi$, then $\mathcal{S}(\sigma,\pi) \le 2\tilde{\mathcal{S}}$.

Forgetting where we started. The next mixing measure which we introduce at this stage does not overtly involve the stationary distribution; instead, we look for the *best* target distribution, in the sense of expected access time from a worst starting state for that distribution. We call this mixing measure the *forget time* because it measures, in a sense, the least time it can take to "forget" what state the chain was started in. We denote the forget time by $\mathcal{F}$ and this time no arguments are needed:

$$\mathcal{F} := \min_\tau \max_{s\in V} \mathcal{H}(s,\tau) = \min_\tau \max_\sigma \mathcal{H}(\sigma,\tau)\,.$$

The target distribution τ which minimizes $\max_s \mathcal{H}(s,\tau)$ will be called the "forget distribution" and denoted by φ. It was proved in [24] that this distribution is uniquely determined; see [24] for an explicit formula for μ.

We can define approximate versions of the forget time, by $\mathcal{F}_{\underline{\varepsilon}} := \min_\tau \mathcal{H}(\ge(1-\varepsilon)\tau)$ etc.

Discrepancy. This quantity is quite closely related to mixing measures, although its definition seems somewhat artificial. We define

$$\mathcal{Z} = \max_i \mathcal{Z}(i,\pi) = \max_i \sum_j \pi_j |H(i,j) - H(\pi,j)|\,. \tag{9}$$

and

$$\tilde{\mathcal{Z}} = \mathcal{Z}(D_1,\pi) = \max_{\sigma\le 2\pi} \mathcal{Z}(\sigma,\pi).$$

We call $\mathcal{Z}$ the *discrepancy* of the chain, because it could be considered as a measure of "unevenness" of these hitting times.

Recall that the mixing time can be defined as follows:

$$\mathcal{H} = \max_i \max_j (H(i,j) - H(\pi,j)).$$

So $\mathcal{H}$ can also be considered as the maximum "surplus" of hitting times to a fixed state, over the average hitting time to that state, while $\mathcal{Z}$ is the average "surplus".

We can also average in the other variable, and get a mixing measure introduced when discussing warm start:

$$\hat{\mathcal{H}} = \max_j \sum_i \pi_i |H(i,j) - H(\pi,j)|\,, \tag{10}$$

Finally, we mention one the most used parameters, the "eigenvalue gap". For a time-reversible chain, this is defined as follows. Let $1 = \lambda_1 > \lambda_2 \ge \ldots \ge \lambda_n$ denote the eigenvalues of the transition matrix M. The *eigenvalue gap* is the value $1 - \lambda_2$. The *relaxation time* is defined by

$$\mathcal{L} = \frac{1}{1-\lambda_2}$$

(we use $\mathcal{L}$ to remind us of λ).

Unfortunately, for chains that are not time-reversible, there are several ways to define this number, and these play different roles in mixing. We take the easy route and do not define it at all.

Strong rules. We mention strong stopping rules, introduced by Aldous and Diaconis [5], which have been used very successfully to estimate mixing times. A stopping rule Γ is said to be *strong* if the final state v^Γ is independent of the number of steps Γ. In other words, for any time t with $\mathsf{P}(\Gamma = t) > 0$, the conditional distribution $\{\sigma^t | \Gamma = t\}$ is the same as σ^Γ.

2.4 Different mixing measures: examples

We give some examples of Markov chains and various mixing measures. We state the values without proof (just with some hints), and without any attempts of providing complete analysis. The computation of some of these mixing times is based on the results in Chapter 4.

Example 2.4 (Two States) It turns out that already the simplest nontrivial chain shows some interesting phenomena. Suppose that there are just two states 1 and 2, with transition matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}, \qquad 0 < a \le b < 1.$$

It will be convenient to set $d = a + b$ and $c = a/(a + b)$, so that $a - cd$ and $b = (1 - c)d$.

It is easy to see that $\pi_1 = 1 - c$, $\pi_2 = c$. Hence the chain is time-reversible. It is also clear that $\mathcal{H}(1, 2) = 1/a = 1/(cd)$ and $\mathcal{H}(2, 1) = 1/b = 1/(1 - c)d$.

The naive rule is optimal from any starting state, since the other one will be a halting state. This gives that

$$\mathcal{H} = \mathcal{H}(1, \pi) = \mathcal{H}(2, \pi) = \frac{1}{d}.$$

It follows that $\mathcal{I} = 1/d$ and it is not difficult to compute that $\mathcal{F} = 1/d$. The eigenvalues of the transition matrix are 1 and $1 - d$, and hence $\mathcal{L} = 1/d$ as well.

Approximate mixing times are similar. We only discuss the quantity $\mathcal{H}_\varepsilon$, the time needed to get closer to π than ε in the total variation distance, where $\varepsilon < 1/4$ is given. If we start at 2, then clearly the best target distribution τ, with $d(\tau, \pi) \le \varepsilon$, is $\tau' = (1 - c - \varepsilon, c + \varepsilon)$, and it is easy to compute that

$$\mathcal{H}(2, d_\varepsilon) = \mathcal{H}(2, \tau) = \frac{1 - c - \varepsilon}{(1 - c)d}.$$

The situation is somewhat more complicated if we start at 1. If $\varepsilon \le c$, then a similar computation yields

$$\mathcal{H}(1, d_\varepsilon) = \frac{c - \varepsilon}{cd}.$$

But if $\varepsilon > c$, then already the starting distribution is closer to π than ε, and thus $\mathcal{H}(1, d_\varepsilon) = 0$. It is easy to check that 2 is always the worse starting state, and hence

$$\mathcal{H}_\varepsilon = \frac{1 - c - \varepsilon}{(1 - c)d} > \frac{1}{2d}.$$

So we gain at most a factor of two by allowing only approximate mixing, independently of ε.

Another quantity worth computing is $\tilde{\mathcal{H}}_\varepsilon$, the approximate mixing time from warm start. There are basically two starting distributions to worry about, namely $(1, 0)$ (which we denote by '1') and $\sigma = (1-2c, 2c)$ (all others are convex combinations of these, and it is easy to argue that they are not worse than these two). If $\varepsilon \geq c$ then these distributions are already in d_ε, so in this case $\tilde{\mathcal{H}}_\varepsilon = 0$. If $\varepsilon \leq c$, then we already computed $\mathcal{H}(1, d_\varepsilon)$. A similar argument gives that

$$\mathcal{H}(\sigma, d_\varepsilon) = \frac{c-\varepsilon}{(1-c)d},$$

and so 1 is the worse starting state. Thus we get

$$\tilde{\mathcal{H}}_\varepsilon = \begin{cases} 0, & \text{if } \varepsilon \geq c, \\ (c-\varepsilon)/(cd), & \text{if } \varepsilon \leq c. \end{cases}$$

Note the strange dependence on ε.

The maximum time of any rule from 2 to π is quite different. If we happen to stay at 1 then we cannot stop until the probability of still being at 2 is reduced to its stationary probability. This gives

$$\mathcal{M} \geq \left\lceil \frac{\log c}{\log(1-(1-c)d)} \right\rceil$$

(where in fact equality holds). If, say, $d = 1/2$, then $\mathcal{H} = 2$ while $\mathcal{M} > \log(1/c)$.

Blind rules don't fare well either. Just to give a rough argument, suppose that a blind rule, starting from 2, yields a distribution with density at most 2. If the rule stops after $\mathcal{M} - 2$ or fewer steps, then it stops at 2 with at least four times the stationary probability. Hence at least half of the time it stops after more than $\mathcal{M} - 2$ steps. This shows that

$$\mathcal{B}_{\overline{1}} > \frac{\mathcal{M}-2}{2}.$$

Example 2.5 (Path) Consider the classic case of a random walk on the path with nodes labeled $1, 2, \ldots, n$. The hitting times from endpoints are $\mathcal{H}(1, j+1) = \mathcal{H}(n, n-j) = j^2$ and the stationary distribution is

$$\frac{1}{2n-2}, \frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}, \frac{1}{2n-2},$$

since for any random walk on a graph π_i is proportional to the degree of the node i.

It is clear that halting states from any state to π can only be the endnodes; the two endnodes are also be the pessimal starting states (cf. Corollary 4.12(c)). Suppose that we start at 1. Owing to the special topology of the path, the naive rule has a halting state (namely the state n), and hence it is optimal by Theorem 2.2. Thus we can compute the mixing time easily:

$$\mathcal{H}(1, \pi) = \sum_{k=1}^{n-1} \frac{1}{n-1} k^2 + \frac{1}{2n-2}(n-1)^2 = \frac{2n^2 - 4n + 3}{6}.$$

The naive rule is not bounded, so to get any idea of the maxing time, one has to consider more complicated rules. It can be shown that $\mathcal{M} = O(n^2)$, but the exact value does not seem to be known.

To forget, one can walk to the midpoint of the path and stop there (assume for simplicity that n is odd). It turns out that this is optimal, and hence

$$\mathcal{F} = \frac{(n-1)^2}{4}.$$

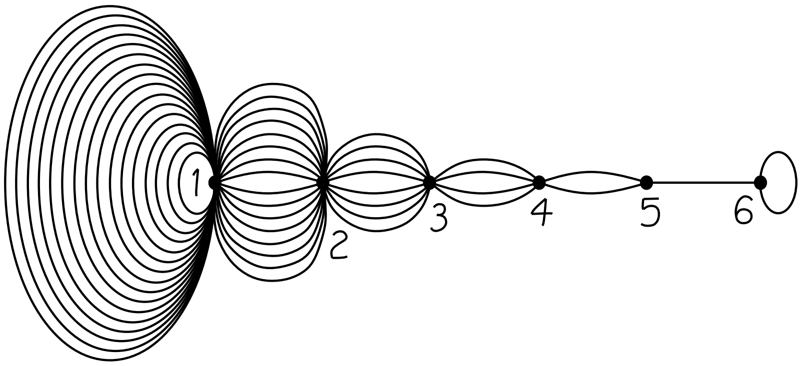

Figure 1: The fat path, for $n = 6$

Thus the forget time is smaller than the mixing time, but has the same order of magnitude. The relaxation time behaves similarly: its value is

$$\mathcal{L} = \frac{1}{1-\cos(\pi/(n+1))} \sim \frac{2n^2}{\pi^2}.$$

Example 2.6 (Fat Path) Let us modify the path to give the walk a large drift. We connect nodes i and $i+1$ by 2^{n-i-1} parallel edges. We also add a (two-way) loop at n and 2^{n-2} loops at 1, just to make the numbers come out nicer (see Fig. 1).

The random walk on the resulting "fat path" is time-reversible. The transition probabilities are $p_{i,i+1} = 1/3$, $p_{i,i-1} = 2/3$ for $2 \le i \le n-1$, $p_{1,2} = 1/3$, $p_{1,1} = 2/3$, $p_{n,n-1} = 1/3$, $p_{n,n} = 2/3$. The stationary probabilities are

$$\frac{1}{2}, \frac{1}{4}, \quad \ldots, \quad \frac{1}{2^{n-1}}, \quad \frac{1}{2^{n-1}}.$$

It can be shown by induction that

$$\mathcal{H}(i,j) = \begin{cases} 3(i-j), & \text{if } j \le i, \\ 3(i-j+2^j-2^i), & \text{if } j \ge i. \end{cases}$$

Again, it follows e.g. from Corollary 4.12(c) that the endnodes are pessimal, and the naive rule is optimal when starting at an endnode, whence elementary computation gives

$$\mathcal{H}(1,\pi) = \mathcal{H}(n,\pi) = 3n - 6 + \frac{3}{2^{n-1}} \sim 3n.$$

Note that it takes exactly the same time to reach π from both endnodes, but through a very different arithmetic!

If we want to get only close to π (in total variation distance), then the two endnodes will behave quite differently. From n, we need to get to the other "most of the time", and hence $\mathcal{H}_\varepsilon \sim 3m$ for every fixed $\varepsilon < 1$. But from 1, we only need to get to the last $\log(1/\varepsilon)$ nodes (the rest has stationary measure less than ε), and hence the time needed for this is only about $\log_2(1/\varepsilon)$.

A similar argument would work to show that the total variation mixing time, with warm start, is only $O(\log(1/\varepsilon))$.

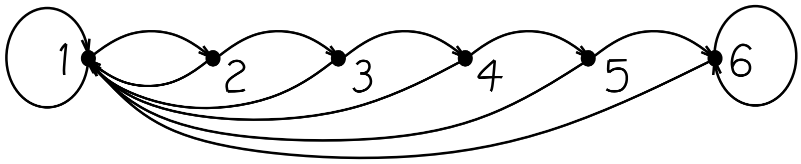

Figure 2: The winning streak, for $n = 6$

Example 2.7 (Cycle) Consider an (undirected) cycle of length n, where (say) n is even. Then of course each node is pessimal, and it is not difficult to compute that

$$\mathcal{H} = \frac{n^2}{12} + \frac{1}{6}.$$

(Compare this with expected time

$$\frac{n-1}{n}\frac{n(n-1)}{2} = \frac{(n-1)^2}{2}$$

for staying at 0 with probability $1/n$ else walking until the last new vertex is hit as in Example 1.1.)

The fact that the cycle has a node-transitive automorphism group makes the determination of the reset and forget times easy. Clearly π is the uniform distribution, $H(i,\pi)$ is the same for all i, and hence $\mathcal{I} = \mathcal{H}$. The uniqueness of the forget distribution implies that it must be just π, and hence $\mathcal{F} = \mathcal{I}$.

The eigenvalues of the cycle are well known, and hence it is easy to compute the relaxation time:

$$\mathcal{L} = \frac{1}{1-\cos(\pi/n)} \sim \frac{2n^2}{\pi^2}.$$

It is also worth taking a look at the directed cycle, which is perhaps the simplest case that is not time-reversible. The walk on this is deterministic, but the mixing times are defined nevertheless. The naive rule is optimal (it has a halting state), and hence $\mathcal{H} = (n-1)/2$. It is also clear that $\mathcal{M} = n-1$. As before, rotational symmetry impies that $\mathcal{F} = \mathcal{I} = \mathcal{H}$.

Example 2.8 (Winning Streak) The following Markov chain, sometimes called the "winning streak" chain, is an interesting example in many respects. Let $V = \{1,\ldots,n\}$, and define the transition probabilities by

$$p_{ij} = \begin{cases} 1/2, & \text{if } j = i+1, \\ 1/2, & \text{if } j = 1 \text{ and } 1 \le i \le n, \\ 1/2, & \text{if } j = i = n, \\ 0, & \text{otherwise.} \end{cases}$$

See Fig. 2. It is easy to check that the stationary distribution is

$$\pi_i = \begin{cases} 2^{-i}, & \text{if } 0 \le i \le n-2, \\ 2^{-(n-1)}, & \text{if } i = n. \end{cases}$$

A very special property of the winning streak chain is that the $(n-1)$-step-rule generates the stationary distribution from every starting state. To see this, note that every walk can be

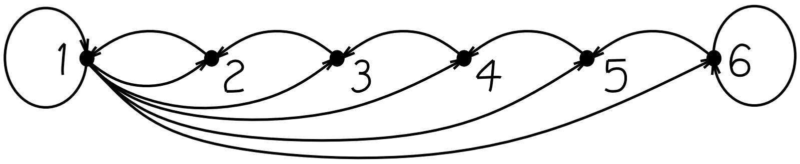

Figure 3: The reverse winning streak, for $n = 6$

described by a sequence $HTTH\ldots$ of independent coin flips, where "head" means to go right (or stay in state n), while "tail" means to fall back to 1. This sequence leads to state i if and only if the sequence ends with exactly $i-1$ heads; the probability that this happens is just 2^{-i} if $i < n$ and $2^{-(n-1)}$, if $i = n$.

The $(n-1)$-step rule is optimal if we start at 1, since n is then a halting state; it is also optimal if we start at n, since $n-1$ is then a halting state; but it is not optimal when starting at any other state, since it is easy to check that it has no halting state. Hence

$$\mathcal{H} = n-1, \qquad \mathcal{I} < n-1.$$

For the reset time, a more tedious calculation shows that

$$\mathcal{I} = n - k + 1 - \frac{n-k}{2^k} - \frac{1}{2^{n-1}},$$

where k is the largest integer with $2^k \leq n - k + 1$ (so $k \approx \log_2 n$; $\mathcal{H}(i, \pi)$ behaves differently for $i < k$ and $i \geq k$). This value is strictly smaller than the mixing time $\mathcal{H}$, but still asymptotically n.

If we want to forget where we started, a very simple strategy is to stop at 1. This happens in an expected number of at most 2 steps, no matter where we start, and hence

$$\mathcal{F} \leq 2.$$

It turns out that it is slightly better if we allow a tiny probability of stopping elsewhere; the exact value of the forget time is

$$\mathcal{F} = 2 - 2^{-(n-2)},$$

achieved by the rule "walk until you hit 0 or make $n-1$ steps, whichever comes first".

Example 2.9 (Reverse Winning Streak) Now look at the reverse chain of the previous example (see Fig. 3). Here the transition probabilities are $p_{i+1,i} = 1$ for $1 \leq i \leq n-1$, $p_{n,n-1} = p_{n,n} = 1/2$, $p_{1,i} = 2^{-i}$ for $0 \leq i \leq n-2$ and $p_{1,n} = 2^{-(n-1)}$ (other transition probabilities are 0). So the transition from any state is deterministic except for the transition from state 1, which leads directly to the stationary distribution. From any other state $i < n$, we can walk to 1 and then make one more move, which takes i moves. This rule is optimal (n is a halting state). We could use the same rule if we start at the last state, but it would not be optimal; instead, the following strategy is optimal: walk for $n-1$ steps.

It follows that the mixing time is $\mathcal{H} = n-1$, while the reset time is

$$\mathcal{I} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \ldots + \frac{1}{2^{n-1}}(n-1) + \frac{1}{2^{n-1}}(n-1) = 2 - \frac{1}{2^{n-1}}.$$

To forget, we can again walk to node 0; in this chain, this will take $n-1$ steps from the worst starting node. It turns out, somewhat surprisingly, that this rule is almost optimal; we have $\mathcal{F} \sim n$.

3 Main results

This section states the main results of this paper. We discuss five groups of mixing times. The relationship between these groups is shown in Fig. 4. Each group contains several mixing measures, which are all within absolute constant factors of each other (for the relaxation group, some these equivalances are only conjectured, and the statements of the results are more complicated for some of the parametric measures; see Theorems 3.5 and 3.6. Mixing measures in a group higher up in the figure are larger (again, up to an absolute constant factor). We will see that parameters in the same group are often quite differently defined and their equivalence is unexpected.

We'll also see that reflecting the figure in the vertical axis corresponds to time reversal. Thus, the Relaxation Group, Mixing Group and Maxing Group are invariant under time reversal (again, up to absolute constants), while the Forget Group and Reset Group are interchanged.

For every chain, either the Forget Group or the Reset Group is within an absolute constant factor of the Mixing group, but generally not both. Thus in every finite Markov chain, either "forgetting is difficult" or "resetting is difficult". We don't know if this dichotomy is reflected in other properties of finite Markov chains. In time-reversible chains, all three groups coincide.

Throughout, we consider the error parameter $0 < \varepsilon < 1/4$ fixed. Below, if we write $\mathcal{A} \approx \mathcal{B}$ for two mixing measures, it means that there are constants $c_1, c_2 > 0$ such that $c_1\mathcal{A} \leq \mathcal{B} \leq c_2\mathcal{A}$. Similarly, $A \ll B$ means that there is a constant $c > 0$ such that $\mathcal{A} \leq c\mathcal{B}$. These constants may depend on the error parameter ε, but not on the chain. It should also be remarked that the constants are moderate (see Section 5 for a detailed statement of these relations).

3.1 Groups of mixing times

The *Mixing Group* is a small group containing the most important mixing measure, the mixing time $\mathcal{H}$. The following theorem shows that the approximate filling time is also in this group. While either one of the forget and reset times may be much smaller than the mixing time, their maximum (or, equivalently, sum) is also in this group.

Theorem 3.1

$$H \approx \mathcal{H}_{\underline{\varepsilon}} \approx \max\{\mathcal{F}, \mathcal{I}\}.$$

The *Forget Group* contains a number of interesting invariants: the forget time $\mathcal{F}$; the set access time $\mathcal{S}$; for any fixed $\varepsilon > 0$, the approximate mixing times $\mathcal{H}_\varepsilon$ and $\mathcal{H}_{\overline{\varepsilon}}$; and both approximate forget times $\mathcal{F}_\varepsilon$ and $\mathcal{F}_{\underline{\varepsilon}}$. Note that the approximate mixing time $\mathcal{H}_{\underline{\varepsilon}}$ is not in this group.

Theorem 3.2

$$\mathcal{F} \approx \mathcal{F}_\varepsilon \approx \mathcal{F}_{\underline{\varepsilon}} \approx \mathcal{H}_\varepsilon \approx \mathcal{H}_{\overline{\varepsilon}} \approx \mathcal{S}.$$

The *Reset Group* contains the reset time $\mathcal{I}$ and the warm mixing time $\tilde{\mathcal{H}}$, as well as the restricted mixing time $\hat{\mathcal{H}}$ and the *blind halving time*

$$\tilde{\mathcal{B}}_{\overline{1/2}} = \mathcal{B}(\leq 2\pi, \leq (3/2)\pi)\,.$$

It will then follow by general results that $\tilde{\mathcal{B}}_{\overline{\varepsilon}}$ is in this group for every fixed ε.

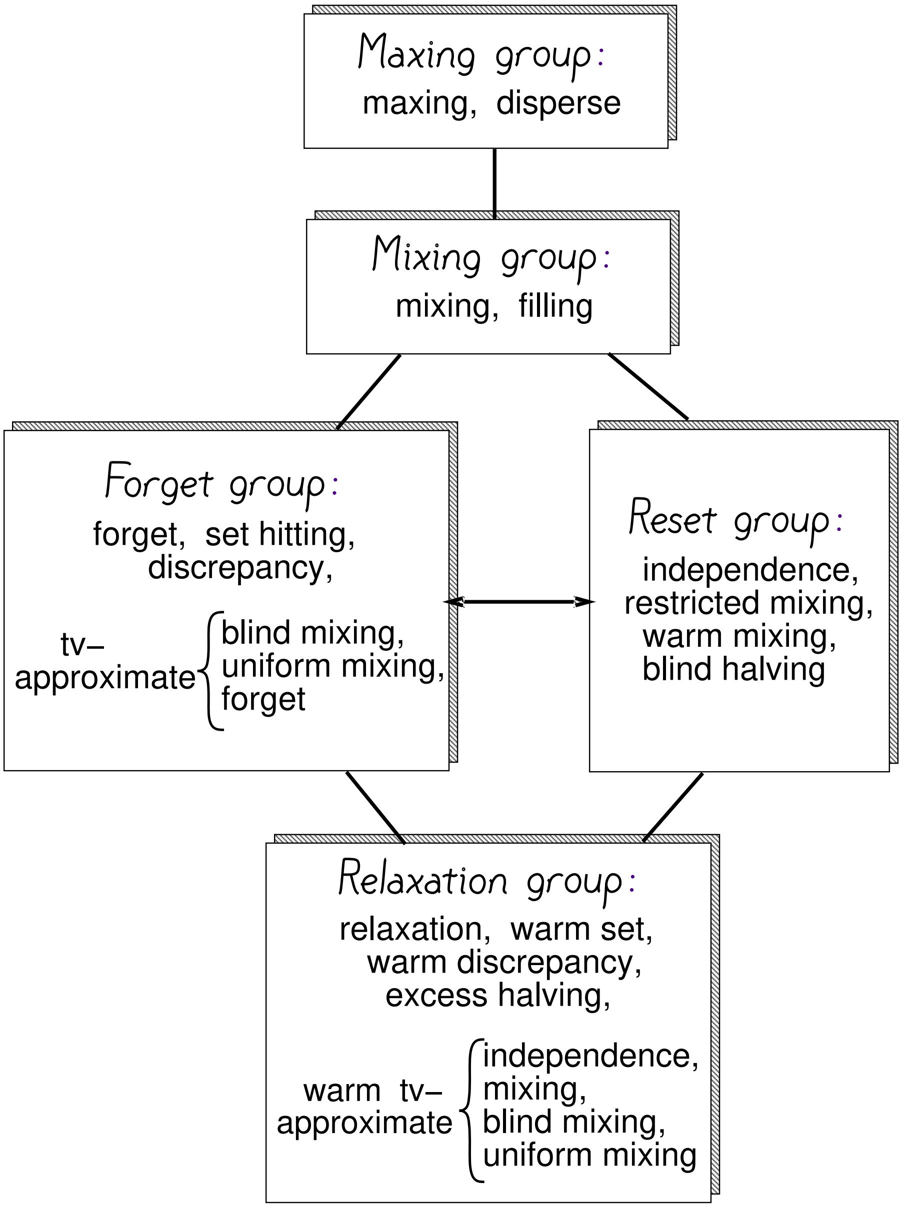

Figure 4: Main groups of mixing times and their relationship

Theorem 3.3

$$\mathcal{I} \approx \tilde{\mathcal{H}} \approx \mathcal{A} \approx \tilde{\mathcal{B}}_{\overline{\varepsilon}}.$$

The *Maxing Group* contains two, seemingly unrelated quantities: the maxing time $\mathcal{M}$ and the blind disperse time $\mathcal{B}_{\overline{\varepsilon}}$.

Theorem 3.4

$$\mathcal{M} \approx \mathcal{B}_{\overline{\varepsilon}}.$$

The *Relaxation Group* contains several important parameters: the warm set access time $\tilde{\mathcal{S}}$, the warm discrepancy $\tilde{\mathcal{Z}}$, the average discrepancy $\mathcal{G}$, the relaxation time $\mathcal{L}$. Among the approximate mixing times, we consider

$$\mathcal{H}^* = \max_{\varepsilon} \mathcal{H}(\leq (1+2\varepsilon)\pi, \leq (1+\varepsilon)\pi),$$

which we call the *excess halving time.*

Theorem 3.5

$$\tilde{\mathcal{S}} \approx \mathcal{H}^* \ll \tilde{\mathcal{Z}}\,.$$

For a reversible chain,

$$\tilde{\mathcal{Z}} \ll \mathcal{L}, \qquad \tilde{\mathcal{S}} \approx \mathcal{L}\,.$$

We list two further interesting mixing measures in this group: the warm independence time $\tilde{\mathcal{I}}_\varepsilon$, and the warm approximate mixing time $\tilde{\mathcal{H}}_\varepsilon$. The following theorem states that these are "essentially equivalent"; but we have to explicitly state the dependence on ε. Recall that in Example 2.4, $\tilde{\mathcal{H}}_\varepsilon$ was 0 for $\varepsilon \geq c$ but positive for $\varepsilon < c$, where $0 < c \leq 1/2$ was an arbitrary parameter of the chain; so no equivalence can be established, say, between $\tilde{\mathcal{H}}_\varepsilon$ and $\tilde{\mathcal{H}}_{2\varepsilon}$.

Theorem 3.6 *(a) For every finite Markov chain and every $\varepsilon > 0$, we have*

$$\varepsilon\tilde{\mathcal{U}}_{2\varepsilon} \leq \tilde{\mathcal{H}}_\varepsilon \leq \tilde{\mathcal{B}}_\varepsilon \leq \tilde{\mathcal{U}}_\varepsilon$$

and

$$\tilde{\mathcal{I}}_{6\varepsilon} \ll \tilde{\mathcal{H}}_\varepsilon \ll \tilde{\mathcal{I}}_{\varepsilon/3}\,.$$

3.2 Relations between groups

The next result shows that the ratio of mixing measures in different groups may be unbounded, but even the ratio quantities in different groups remains relatively small: the reset group and the maxing group are not more than a factor $\ln\hat{\pi}$ apart, where $\hat{\pi} := \min_i\{\pi_i\}$. Considering the reverse chain, we get a similar inequality between the forget and maxing groups. We conjecture that a similar inequality must hold between the relaxation and maxing groups.

Theorem 3.7

$$\mathcal{M} \leq \left(\ln\frac{1}{\hat{\pi}}\right)\tilde{\mathcal{B}}_{\overline{1/2}}.$$

Note that in a typical application in sampling algorithms the state space V is exponentially large (in terms of the natural "size" of the problem), and π is most often uniform, so that $\ln(1/\hat{\pi})$ is polynomial. Thus it follows that in such cases, all mixing measures are within polynomial factors of each other.

The winning streak chain (Example 2.8 shows that the logarithmic factor in the upper bound in the theorem can occur already as the quotient $\mathcal{H}/\mathcal{F}$; in this case of course $\mathcal{M} \approx \mathcal{I} \approx \mathcal{H}$ and $\mathcal{L} \approx \mathcal{F}$. The reverse winning steak chain of course shows a similar situation with $\mathcal{F}$ and $\mathcal{I}$ interchanged. The chain with two states in Example 2.4 shows that this logarithmic factor can also occur as the ratio $\mathcal{M}/\mathcal{H}$, even in the time-reversible case (and then of course $\mathcal{H} \approx \mathcal{F} = \mathcal{I} \approx \mathcal{L}$). Finally, Example 2.6 shows that this logarithmic factor can also occur as $\mathcal{I}/\mathcal{L}$ in a time-reversible chain (in this case of course $\mathcal{M} \approx \mathcal{H} \approx \mathcal{F} = \mathcal{I}$).

We conclude with stating a theorem that shows that the mixing, maxing and relaxation groups are invariant under time reversal, while the forget and reset groups are interchanged.

Theorem 3.8 *For every finite Markov chain,*

(a) $\overleftarrow{\mathcal{H}} = \mathcal{H}$;

(b) $\overleftarrow{\mathcal{B}_{\overline{\varepsilon}}} = \mathcal{B}_{\overline{\varepsilon}}$.

(c) $\overleftarrow{\mathcal{I}} = \mathcal{F}$ *and* $\overleftarrow{\mathcal{F}} = \mathcal{I}$.

(d) $\overleftarrow{\mathcal{A}} = \mathcal{Z}$ *and* $\overleftarrow{\mathcal{Z}} = \mathcal{A}$.

If the chain is time-reversible, then $\mathcal{I} = \mathcal{F}$, and hence by Theorem 3.1, $\mathcal{F} \approx \mathcal{H}$. Thus for time-reversible chains the mixing, forget and reset groups form one group of equivalent mixing measures.

4 Tools

4.1 Exit frequencies

Let us now fix a transition matrix M, an initial distribution σ and a finite stopping rule Γ.

Definition. The *exit frequencies* $x_i = x_i(\Gamma)$ $(i \in V)$ are defined by setting x_i equal to the expected number of times the walk leaves state i before stopping. The *scaled exit frequencies* are defined by $y_i = \frac{1}{\pi_i} x_i$. In the finite case, the use of x_i or y_i is only a matter of taste; in the case of Markov chains with infinite state space, the difference between x_i and y_i is quite significant.

Exit frequencies are a special case of what Pitman [25] calls "pre-T occupation measures" for stopping times T. Indeed, versions of Lemma 4.2 and Theorem 4.4 below are proved in [25] using Pitman's "occupation measure identity".

Exit frequencies for the naive rule are easy to find; see [1]. Several related formulas could be derived using relations to electrical networks, as in [13], [15], or [28].

Lemma 4.1 *The exit frequencies $\tilde{x}$ for the naive stopping rule $\Omega_{i,j}$ from state i to j are given by*

$$\tilde{x}_k = \pi_k \left(\mathcal{H}(i,j) + \mathcal{H}(j,k) - \mathcal{H}(i,k)\right) .$$

More generally, the exit frequencies for the naive stopping rule $\Omega_{\sigma,\tau}$ are given by

$$\tilde{x}_k = \pi_k \sum_{i,j} \sigma_i \tau_j (\mathcal{H}(i,j) + \mathcal{H}(j,k) - \mathcal{H}(i,k)) = \pi_k \left(\sum_{i,j} \sigma_i \tau_j \mathcal{H}(i,j) + \mathcal{H}(\tau,k) - \mathcal{H}(\sigma,k) \right) .$$

Exit frequencies are related to the starting and ending distributions by a simple formula, found in Pitman [25], which can be obtained by counting entrances and exits at a given state:

Lemma 4.2 *The exit frequencies of any stopping rule Γ that reaches distribution τ from distribution σ satisfy the equation*

$$\sum_i p_{i,j} x_i(\Gamma) - x_j(\Gamma) = \tau_j - \sigma_j .$$

One can rewrite this identity as follows.

$$\sum_i p_{i,j} x_i(\Gamma) - \sum_i p_{ji} x_j(\Gamma) = \tau_j - \sigma_j .$$

In other words, the values $p_{ij}x_i(\Gamma)$ can be viewed as values of a flow through the underlying graph, from supply σ to demand τ. This is also easily seen by observing that $p_{ij}x_i(\Gamma)$ is the expected number of passes from i to j while following Γ.

This lemma implies that the exit frequencies of a stopping rule are "almost" determined by the starting and target distributions:

Theorem 4.3 *Fix two distributions σ and τ, and let Γ and Γ' be two finite stopping rules from σ to τ. Then $x(\Gamma') - x(\Gamma) = (\mathsf{E}\Gamma - \mathsf{E}\Gamma')\pi$.*

Conversely if Γ and Γ' are two stopping rules such that when starting from σ, $x(\Gamma)$ and $x(\Gamma')$ differ by a multiple of π, then $\sigma^\Gamma = \sigma^{\Gamma'}$.

It follows from Theorem 4.3 that the exit frequencies of any mean-optimal stopping rule from σ to τ are the same. We denote them by $x_i(\sigma,\tau)$.

Another way of formulating Theorem 4.3 is to say that the values $x_i - \pi_i \mathsf{E}\Gamma$ are the same for *every* stopping rule Γ (optimal or not) from a given σ to a given τ. We call these values the *exit discrepancies* and denote them by $z_i(\sigma,\tau)$. They will be very convenient quantities to use.

In particular, we can compute $z_k(\sigma,\tau)$ by considering the naive rule:

$$\begin{aligned} z_k(\sigma,\tau) &= \pi_k \left(\sum_{i,j} \sigma_i \tau_j \mathcal{H}(i,j) + \mathcal{H}(\tau,k) - \mathcal{H}(\sigma,k) \right) - \pi_k \sum_{i,j} \sigma_i \tau_j \mathcal{H}(i,j) \\ &= \pi_k (\mathcal{H}(\tau,k) - \mathcal{H}(\sigma,k)) . \end{aligned} \tag{11}$$

Although derived from the naive rule, this formula holds for the exit discrepancies of any stopping rule from σ to τ, a fact we shall exploit throughout. Applying it to any optimal rule, we get the following result:

Theorem 4.4 *The scaled exit frequencies of a mean-optimal stopping rule from σ to τ are given by*

$$y_k(\sigma,\tau) = \mathcal{H}(\tau,k) - \mathcal{H}(\sigma,k) + \mathcal{H}(\sigma,\tau).$$

Formula (11) yields immediately that $z_k(\sigma,\tau)$ is a linear function in each of σ and τ, and that it has a number of useful properties:

$$z_k(\sigma,\tau)+z_k(\tau,\rho)=z_k(\sigma,\rho)$$

$$z_k(\sigma,\tau)=-z_k(\tau,\sigma).$$

Time-reversibility is also reflected by the exit discrepancies in a very natural way:

Lemma 4.5 *The Markov chain is reversible if and only if* $\pi_k z_i(k,\pi)=\pi_i z_k(i,\pi)$ *for all* $i,k\in S$.

Exit discrepancies are related to access times by some simple but useful inequalities. Trivially

$$z_i(\sigma,\tau)\geq -\pi_i\mathcal{H}(\sigma,\tau) \tag{12}$$

and hence

$$z_i(\sigma,\tau)\leq \pi_i\mathcal{H}(\tau,\sigma)\,. \tag{13}$$

In terms of $\mathcal{H}(\sigma,\tau)$, we only have the much weaker inequality

$$z_i(\sigma,\tau)\leq (1-\pi_i)\mathcal{H}(\sigma,\tau)\,. \tag{14}$$

To get a stronger inequality, note that since $\sum_i z_i(\sigma,\tau)=0$, we have

$$\max_A\sum_{i\in A} z_i(\sigma,\tau)=\sum_i \max\{0,z_i(\sigma,\tau)\}=\frac{1}{2}\|z(\sigma,\tau)\|\,.$$

Then for the sum of z_i over any subset:

$$\sum_{i\in A} z_i(\sigma,\tau)=\sum_{i\in A} x_i(\sigma,\tau)-\pi(A)\mathcal{H}(\sigma,\tau)\leq \mathcal{H}(\sigma,\tau)-\pi(A)\mathcal{H}(\sigma,\tau)\,. \tag{15}$$

Hence

$$\frac{1}{2}\|z(\sigma,\tau)\|\leq \mathcal{H}(\sigma,\tau)\,. \tag{16}$$

Exit discrepancies are closely related to some other measures introduced before. From (11) we get immediately that

$$\mathcal{Z}=\max_i\|z(i,\pi)\|\,,\qquad \mathcal{G}=\sum_i \pi_i\|z(i,\pi)\|.$$

We can also define the warm start version of $\mathcal{Z}$,

$$\tilde{\mathcal{Z}}:=\max_{\sigma\leq 2\pi}\|z(\sigma,\pi)\|\,.$$

4.2 Optimal stopping rules

We describe two optimal stopping rules whose analysis will be needed in proveing our main results. A more detailed analysis of these and other optimal stopping rules will be deferred to another publication.

Chain rule. Let ρ be a probability distribution on the *subsets* of the state space V, and observe that it provides a stopping rule: "choose a subset U from ρ, and walk until some state in U is hit". The naive rule is of course a special case, with ρ concentrated on singletons.

Theorem 4.6 *For every starting distribution σ and target distribution τ, there exists a unique distribution ρ which is concentrated on a chain of subsets and gives an optimal stopping rule for generating τ.*

Proof. Starting the chain from distribution σ, let α^U be the distribution of the first node in U, for every subset U. For example, $\alpha^V = \sigma$. Let $U_0 = V$, $\mu^0 = \tau$. We define distinct nodes $i_1, \ldots, i_n$, non-negative numbers $\rho_1, \ldots, \rho_n$, and non-negative vectors $\mu^1, \ldots, \mu^n \in \mathbf{R}^V$, by induction. Assume that we have defined v_j, ρ_j and μ_j for $j \leq k$. Let $U_{k+1} = S \setminus \{v_1, \ldots, v_k\}$. Select a node i_{k+1} which minimizes $\mu_i^{k-1}/\alpha_i^{U_k}$, and let

$$\rho_{k+1} = \mu^k_{i_{k+1}}/\alpha^{U_{k+1}}_{i_{k+1}}$$

and

$$\mu_i^{k+1} = \mu_i^k - \rho_{k+1}\alpha_i^{U_{k+1}} .$$

Note that $\rho_1 = \tau_1/\sigma_1$.

It follows by induction and by the choice of i_{k+1} that $\rho_{k+1} \geq 0$, $\mu^{k+1} \geq 0$, and $\nu_i^{k+1} = 0$ for $i \notin U_{k+2}$. Moreover, we have

$$\sum_{k=1}^{n} \rho_k = 1 .$$

□

We call this rule the "chain rule" and denote it by Ξ. A rather neat way to think of this rule is to assign a "price", a real value $r(i) = \sum\{\rho_U : i \in U\}$ to each state i. The rule is then implemented by choosing a random real "budget" b uniformly from $[0,1]$ and walking until a state j with $r(j) \leq b$ (an item that we can buy) is reached.

As an application of the chain rule, we prove a general inequality concerning set hitting times.

Lemma 4.7 *Let σ, τ be two arbitrary distributions and $\varepsilon > 0$. Then*

$$\mathcal{H}(\sigma, \leq (1+\varepsilon)\tau) \leq (1+\varepsilon)\ln(1+1/\varepsilon))\mathcal{S}(\sigma,\tau) ,$$

If $\tau_i \geq d > 0$ for all i, then

$$\mathcal{H}(\sigma,\tau) \leq \ln(1/d)\mathcal{S}(\sigma,\tau) ,$$

Proof. We use the chain rule. By Theorem 4.6, there exists a labelling $\{0, \ldots, n-1\}$ of the nodes and a probability distribution ρ such that setting $S_j = \{j, j+1, \ldots, n-1\}$, selecting a j, $0 \leq j \leq n-1$ according to ρ and then walking until S_j is hit, generates τ. Note that

$$\rho_{S_j} \leq \tau_{S_j}, \tag{17}$$

since if we choose any S_m, $m \geq j$ in the chain rule as our target set, we certainly end up in S_j.

To prove the first inequality, fix the subscript m such that $\rho_{S_{m+1}} \leq 1/(1+\varepsilon) < \rho_{S_m}$. Now modify the chain rule by selecting S_j with probability

$$\rho'_j = \begin{cases} (1+\varepsilon)\rho_j, & \text{if } j \leq m-1, \\ 1-(1+\varepsilon)(\rho_1+\ldots+\rho_{m-1}), & \text{if } j = m, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\rho'_m \leq (1+\varepsilon)\rho_m$ by the choice of m.

Let this new rule generate distribution τ'. Then obviously

$$\tau'_i \leq (1+\varepsilon)\tau_i.$$

The mean length of this new rule is

$$\begin{aligned} \mathcal{H}(\sigma,\tau') &= \sum_{j=1}^{m} \rho'_j \mathcal{H}(\sigma, S_j) \\ &\leq \left(\sum_{j=1}^{m} \frac{\rho'_j}{\tau(S_j)}\right) \mathcal{S}(\sigma,\tau) \\ &\leq \left(\sum_{j=1}^{m} \frac{\rho'_j}{\rho(S_j)}\right) \mathcal{S}(\sigma,\tau) \\ &= (1+\varepsilon)\left(\sum_{j=1}^{m-1} \frac{\rho_j}{\rho(S_j)} + \frac{\rho(S_m)-1+1/(1+\varepsilon)}{\rho(S_m)}\right) \mathcal{S}(\sigma,\tau) \end{aligned}$$

Using the inequality $1-x \leq \ln 1/x$, we get

$$\begin{aligned} \mathcal{H}(\sigma,\tau') &\leq (1+\varepsilon)\left(\sum_{j=1}^{m-1} \ln \frac{\rho(S_j)}{\rho(S_{j+1})} + \ln \frac{\rho(S_m)}{1-1/c}\right) \mathcal{S}(\sigma,\tau) \\ &= (1+\varepsilon) \ln\left(1+\frac{1}{\varepsilon}\right) \mathcal{S}(\sigma,\tau). \end{aligned}$$

This proves the first inequality. The second follows by a similar computation. □

If we know a bound on the ratio σ_i/τ_i, then a sharper inequality can be proved:

Corollary 4.8 *Let σ, τ be two distributions, let $\varepsilon > 0$, and assume that $\sigma \leq (1+2\varepsilon)\tau$. Then*

$$\mathcal{H}(\sigma, \leq (1+\varepsilon)\tau) \leq 2\mathcal{S}(\sigma,\tau),$$

Proof. Let $W = \{i : \sigma_i > (1+\varepsilon)\tau_i\}$ and $U = V \setminus W$. If $W = \emptyset$, then we have nothing to prove, so suppose that $W \neq \emptyset$.

Let $a_i = \max(0, \sigma_i - (1+\varepsilon)\tau_i)$ and $b_i = \max(0, (1+\varepsilon)\tau_i - \sigma_i)$. Let $a = \sum_i a_i$. Then simple computation yields that $\sum_i b_i = a + \varepsilon$. Obviously, $a \leq \varepsilon$. Setting $\alpha_i = a_i/a$, $\beta_i = b_i/(a+\varepsilon)$, we get two distributions α and β.

Now fix a stopping rule Δ from α to some distribution $\beta' \le 2\beta$, such that $\mathsf{E}\Delta = \mathcal{H}(\alpha, \le 2\beta)$. Consider the following rule Γ: if you start at i, then halt right away with probability $(\sigma - a_i)/\sigma_i$; else follow Δ.

First, we show that $\sigma^\Gamma \le (1+\varepsilon)\tau$. Indeed, if $i \in W$ then $\beta_i = 0$ and hence $\beta'_i = 0$, which implies that

$$\sigma_i^\Gamma = \sigma_i - a_i = (1+\varepsilon)\tau_i.$$

If $i \in U$ then

$$\sigma_i^\Gamma = \sigma_i + a\beta'_i \le \sigma_i + 2a\beta_i = \sigma_i + \frac{2a}{a+\varepsilon}((1+\varepsilon)\tau_i - \sigma_i) \le (1+\varepsilon)\tau_i.$$

Second, we show that $\mathsf{E}\Gamma \le 2\mathcal{S}(\sigma,\tau)$:

$$\mathsf{E}\Gamma = a\mathsf{E}\Delta = a\mathcal{H}(\alpha, \le 2\beta)\,.$$

By Lemma 4.7, we have $\mathcal{H}(\alpha, \le 2\beta) \le 2\mathcal{S}(\alpha,\beta)$. But it is easy to see that

$$\alpha \le \frac{\varepsilon}{a(1+2\varepsilon)}\sigma\,,$$

and

$$\beta \le \frac{1+\varepsilon}{\varepsilon}\tau\,,$$

and hence by (8)

$$\mathcal{S}(\alpha,\beta) \le \frac{\varepsilon}{a(1+2\varepsilon)}\frac{1+\varepsilon}{\varepsilon}\mathcal{S}(\sigma,\tau) < \frac{1}{a}\mathcal{S}(\sigma,\tau)\,.$$

Thus

$$\mathsf{E}\Gamma \le 2\mathcal{S}(\sigma,\tau)$$

as claimed. □

Threshold rule. A natural way of generalizing the simplest stopping rule ("walk for k steps") is to have a "threshold" or "release time" $0 \le h_i \le \infty$ associated with each state i. The stopping rule is defined by

$$\Gamma(w_0,\dots,w_k) = \begin{cases} 0 & \text{if } k \ge h_{w_k} \\ 1 & \text{if } k \le h_{w_k} - 1 \\ k - h_{w_k} & \text{otherwise.} \end{cases}$$

Thus if we hit a state past the release time we stop; otherwise we continue, except that if we're within 1, we use the fractional part to randomize.

Theorem 4.9 *For every starting distribution σ and target distribution τ, there exists a threshold function h which gives an optimal stopping rule for generating τ.*

Proof. We define this rule recursively as follows. Let p_i^t be the probability of being at state i after t steps (and thus not having stopped at a prior step); let q_i^t be the probability of stopping at state i in *fewer* than t steps. Let x_i^t be the expected number of exits from state i, again in fewer than t steps. Let x_i be the exit frequency of state i in any optimal rule from σ to τ. We'll maintain that $x_i^t \le x_i$. Then *if we are at state i after step t, we we continue with probability* $\min\{1, (x_i - x_i^t)/p_i^t\}$.

It is clear that this rule gives the recurrence

$$x_i^{t+1} = \min(x_i, x_i^t + p_i^t),$$

and hence $x_i^{t+1} \leq x_i$ remains valid. Let us also observe $q_j^t \leq \tau_j$ remains valid. Indeed, if we stop at j by time t at all, then we have $x^{t+1} = x_j$; but then the expected number of times j is entered by time $t+1$ is

$$\sigma_j + \sum_i p_{ij} x_i^t \leq \sigma_j + \sum_i p_{ij} x_i = \tau_j + x_j = \tau_j + x_j^{t+1},$$

and hence the expected number of times we stop there is at most τ_j.

It is also clear that we stop with probability 1 (since the sum of its exit frequencies is bounded) and does achieve an ending distribution. Thus this distribution must be τ. The rule is optimal, since the state with $x_i = 0$ is a halting state.

To show that this rule is a threshold rule, We define the thresholds h_i as follows. Consider the largest t for which $x_i^t < x_i$, and let

$$h_i = t + \frac{x_i - x_i^t}{p_i^t}.$$

If $x_i^t < x_i$ for all i, then $h_i = \infty$. If t is finite, then $x^{t+1} = x_i \leq x_i^t + p_i^t$, and hence $p_i^t > 0$ and $h_i \leq t+1$). $\square$

The threshold vector may not be uniquely determined by a threshold rule Θ (e.g. all possible thresholds h_i smaller than the time before any possible walk reaches i are equivalent), but by convention we always associate with Θ the vector each of whose coordinates is minimal. Then h is finite if and only if Θ is bounded. In fact more is true,

Theorem 4.10 *For any two distributions σ and τ, the threshold rule is max-optimal. In other words, if h_i $(i \in V)$ are the thresholds for an optimal threshold rule from σ to τ, then*

$$\mathcal{M}(\sigma, \tau) = \max_i \lceil h_i \rceil.$$

Proof. Let, as before, p_i^t be the probability the following the threshold rule Θ, we are at state i after t steps, and let x_i^t be the expected number of exits from state i in fewer than t steps. Let Γ be any stopping rule from σ to τ, and let $p_i^t(\Gamma)$ and $x_i^t(\Gamma)$ be the corresponding quantities for Γ.

We claim that $x_i^t(\Gamma) \leq x_i^t$ for all i and t. This clearly implies that the $\max(\Gamma) \geq \max(\Theta)$ as claimed. Using induction on t, we have

$$x_i^{t+1}(\Gamma) \leq \min(x_i, x_i^t(\Gamma) + p_i^t(\Gamma)) \leq \min(x_i, x_i^t + p_i^t) = x_i^{t+1}.$$

$\square$

4.3 Reverse chains

We defined reverse chains in the introduction and mentioned that they have the same stationary distribution as the "forward" chain. Here we study exit frequencies and access times in reverse chains. The key to the proof of many properties of the reverse chain is the following general "duality formula", proved in [24].

Lemma 4.11 *Let $\alpha, \beta, \gamma, \delta$ be four distributions on the states. Then*

$$\sum_i (\beta_i - \alpha_i) \overleftarrow{y}_i(\gamma, \delta) = \sum_i (\delta_i - \gamma_i) y_i(\alpha, \beta) .$$

We recall some consequences of this lemma.

Corollary 4.12 (a) *For every state j, we have*

$$\mathcal{H}(\pi, j) = \overleftarrow{\mathcal{H}}(\pi, j) .$$

For every two states i and j, we have

$$\overleftarrow{\mathcal{H}}(i, j) = \mathcal{H}(j, i) - \mathcal{H}(\pi, i) + \mathcal{H}(\pi, j) .$$

(b) $\overleftarrow{\mathcal{H}} = \mathcal{H}$; $\overleftarrow{\mathcal{N}} = \mathcal{N}$.

(c) *If i is a pessimal state and j is a halting state from i to π, then j is a pessimal state for the reverse chain and i is a halting state from j to π in the reverse chain.*

In particular, every time-reversible chain has (at least) two pessimal starting states, which are halting states for getting from the other to π.

4.4 Matrix formulas

We describe some useful formulas connecting the transition matrix M with mixing parameters. All matrices below are $n \times n$ (where $n = |V|$), and their rows and columns are indexed by states. H is the matrix of hitting times: $H_{ij} = \mathcal{H}(i, j)$. We denote by R the diagonal matrix with the return time $1/\pi_i$ in the i-th position of the diagonal. The identity matrix is denoted by I, the all-1 matrix, by J.

It is easy to see that these matrices satisfy the equation

$$(I - M)H = J - R. \tag{18}$$

Unfortunately, the matrix $I - M$ is singular, and so (18) does not uniquely determine H. But the only left eigenvector of $I - M$ with eigenvalue 0 is π^{T}, and the only right eigenvector with eigenvalue 0 is $\mathbf{1}$, hence $I - M + \mathbf{1}\pi^{\mathsf{T}} = I - M + JR^{-1}$ is non-singular. Moreover, we have

$$(I - M + JR^{-1})(I - JR^{-1}) = I - M$$

(using that $MJ = J$ and $JR^{-1}J = J$). Hence

$$(I - M + JR^{-1})(I - JR^{-1})H = (I - M)H = J - R ,$$

and thus

$$H = (I - M + JR^{-1})^{-1}(J - R) + JR^{-1}H . \tag{19}$$

For convenience, let us set

$$G := (I - M + JR^{-1})^{-1}(J - R) = J - (I - M + JR^{-1})^{-1} .$$

Equation (19) is still not the right formula for H, since H also occurs on the right hand side. But we can determine $\pi^{\mathsf{T}} H$ by looking at the diagonal: $0 = G_{ii} + (\pi^{\mathsf{T}} H)_i$ and hence

$$-G_{ii} = (\pi^{\mathsf{T}} H)_i = \mathcal{H}(\pi, i)\,.$$

This implies that

$$G_{ij} = \mathcal{H}(i,j) - \mathcal{H}(\pi,j) = \frac{1}{\pi_i} z_i(i,\pi).$$

Some important properties of the matrix G are summarized below:

Lemma 4.13 *(a) The G-matrix of the reverse chain is the transpose of G. In particular, if the chain is time-reversible then G is symmetric.*

The eigenvalues of $R^{-1}G$ are 1 and $1/(1-\lambda_2), \ldots, 1/(1-\lambda_n)$, where $1 = \lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of M.

The first assertion is immediate from Corollary 4.12, the second, from the definition of G.

By Theorem 2.1, we get

$$\mathcal{H}(i,\pi) = \max_j(\mathcal{H}(i,j) - \mathcal{H}(\pi,j)) = \max_j G_{ij}.$$

Hence the mixing time is the largest entry of G.

More generally, the matrix G can be used to express several mixing measures. Obviously,

$$\mathcal{H}(\sigma,\pi) = |G^T\sigma|_\infty.$$

Since

$$z_j(\sigma,\pi) = \pi_j(\mathcal{H}(\sigma,j) - \mathcal{H}(\pi,j)) = (R^{-1}G^T\sigma)_j\,,$$

we get

$$\mathcal{Z} = \max_i \|z(i,\pi)\| = \max_i \|R^{-1}G^T e_i\|.$$

Furthermore,

$$\mathcal{A} = \max_A \pi(A)\mathcal{H}(\pi_A,\pi) = \max_A \max_j \pi_A^T G e_j = \frac{1}{2}\max_j \|R^{-1}Ge_j\|.$$

Let's make a rather trivial but useful remark. Let $w \in \mathbb{R}^V$ be a vector with $\sum_i w_i = 0$. We have considered the ℓ_1-norm

$$\|w\| = \sum_i |w_i| = \max\{w^T x:\ |x|_\infty \le 1\},$$

and used the fact that because of the condition that $\sum_i w_i = 0$, we could as well require that $0 \le x_i \le 2$, which leads to the formula

$$\|w\| = 2\max_{A\subseteq V}\sum_{i\in A} w_i.$$

Another quantity that comes up often is the norm

$$\|w\|' = \max\{w^T x:\ |x|_\infty \le 1,\ \sum_i x_i = 0\}.$$

Using that $\sum_i w_i$, we can also write this as

$$\|w\|' = \max\{w^T x :\ 0 \le x_i \le 2,\ \sum_i \pi_i x_i = 1.\}$$

Here $\sigma_i = \pi_i x_i$ may be viewed as a distribution with $\sigma \le 2\pi$.

Now the two norms are not far from each other:

$$\|w\|' \le \|w\| \le 2\|w\|'. \tag{20}$$

The first inequality is trivial. To see the second, let $A = \{i :\ \ w_i \ge 0\}$. We may assume without loss of generality that $\pi(A) \ge 1/2$ (else, consider $-w$). Then

$$\|w\| = 2\sum_{i\in A} w_i \le 2\pi(A)\sum_{i\in A} \frac{1}{\pi(A)} w_i \le 2\pi(A)\|w\|' \le 2\|w\|'.$$

In particular, we get

$$\begin{aligned}\mathcal{Z}(\sigma,\pi) &= \|z(\sigma,\pi)\| \ge \|z(\sigma,\pi)\|' = \max_{\tau\le 2\pi}\sum_j \frac{\tau_j}{\pi_j} z_j(\sigma,\tau)\\ &= \max_{\tau\le 2\pi} \tau^T G\sigma \ \ge \frac{1}{2}\mathcal{Z}(\sigma,\tau)\,.\end{aligned}$$

Thus we get that

$$\tilde{\mathcal{Z}} = \max_{\sigma\le 2\pi} \|z(\sigma,\pi)\| \ge \max_{\sigma\le 2\pi}\max_{\tau\le 2\pi} \sigma^T G \tau^T \ge \frac{1}{2}\tilde{\mathcal{Z}}. \tag{21}$$

4.5 The filling method

This section describes a folklore procedure to access a distribution, if we can access a positive fraction of it.

Lemma 4.14 *For every two distributions τ and σ, and every $0 \le \varepsilon < 1$,*

$$\mathcal{H}(\sigma,\tau) \le \mathcal{H}(\sigma, \ge (1-\varepsilon)\tau) + \frac{\varepsilon}{1-\varepsilon}\mathcal{H}(\ge (1-\varepsilon)\tau)\,.$$

Proof. Let σ be any starting distribution, and consider the following stopping rule: follow an optimal stopping rule Γ rule from σ to σ^Γ such that $\sigma^\Gamma \ge (1-\varepsilon)\tau$. We get a random state w from distribution σ^Γ. We flip a biased coin and stop with probability $(1-\varepsilon)\tau_w/\sigma^\Gamma_w$ (by our assumption on Γ, this is at most 1). The probability that we stop at state i is

$$\sigma_i^\Gamma \frac{(1-\varepsilon)\tau_i}{\sigma_i^\Gamma} = (1-\varepsilon)\tau_i\,,$$

and hence the probability that we stop at all is $1-\varepsilon$.

If we do not stop, the continuation of the walk can be considered as a walk starting from the distribution $\sigma' = \frac{1}{\varepsilon}\sigma^\Gamma - \frac{1-\varepsilon}{\varepsilon}\tau$. We follow an optimal stopping rule Γ' from σ' to $(\sigma')^{\Gamma'}$, such that $(\sigma')^{\Gamma'} > (1-\varepsilon)\tau$, to get a random state j; there we stop with probability $(1-\varepsilon)\tau_j/(\sigma')^{\Gamma'}_j$, else follow an optimal stopping rule Γ'' from $\sigma'' = \frac{1}{\varepsilon}(\sigma')^{\Gamma'} - \frac{1-\varepsilon}{\varepsilon}\tau$ to get close to τ etc. The probability that we stop (eventually) at state i is

$$(1-\varepsilon)\tau_i + \varepsilon(1-\varepsilon)\tau_i + \varepsilon^2(1-\varepsilon)\tau_i + \ldots = \tau_i\,.$$

These numbers add up to 1, hence we stop with probability 1. The expected number of steps is

$$\mathsf{E}\Gamma + \varepsilon\mathsf{E}\Gamma' + \varepsilon^2\mathsf{E}\Gamma'' + \ldots \le \frac{1}{1-\varepsilon}\mathcal{H}(\ge(1-\varepsilon)\tau)\,.$$

$\square$

Corollary 4.15 *For any target distribution τ and $0 \le \varepsilon < 1$,*

$$\mathcal{H}(\tau) \le \frac{1}{1-\varepsilon}\mathcal{H}(\ge(1-\varepsilon)\tau)\,.$$

In particular,

$$\mathcal{H} \le \frac{1}{1-eps}\mathcal{H}_{\underline{\varepsilon}}.$$

4.6 Elementary properties of access times

As a preliminary remark, we note that making a move independently of where we are does not hurt; more exactly, recalling that d is total variation distance and σ^1 is the state distribution after one step of the chain, we have

Lemma 4.16

$$d(\sigma^1,\pi) \le d(\sigma,\pi); \qquad \mathcal{H}(\sigma^1,\pi) \le \mathcal{H}(\sigma,\pi)\,.$$

Proof. The first inequality is easily checked. To prove the second, consider the following rule from σ to π: make one step, then follow an optimal rule from σ^1 to π. Comparing this with an optimal rule from σ to π using Theorem 4.3, we get

$$\pi_i(\mathcal{H}(\sigma,\pi) - \mathcal{H}(\sigma',\pi)) = x_i(\sigma,\pi) - x_i(\sigma',\pi) + \pi_i - \sigma_i \ge -x_i(\sigma',\pi)$$

by Lemma 4.2. Since there is a state i for which the right hand side is 0, the second inequality also follows. $\square$

As an immediate corollary of Theorem 2.1 we obtain that $\mathcal{H}(\sigma,\tau)$ depends on the difference $\sigma-\tau$ only. In fact a little more follows:

Corollary 4.17 *Let $\alpha,\beta,\gamma,\delta$ be four distributions such that $\alpha-\beta = c(\gamma-\delta)$. Then*

$$\mathcal{H}(\alpha,\beta) = c\mathcal{H}(\gamma,\delta)\,.$$

It is easy to see that the access time $\mathcal{H}(\sigma,\tau)$ is convex in both its arguments.

Next we study how access times behave if the source, or the target, distribution is changed "a little". Theorem 4.14 implies the following bound on the "sensitivity" of the target distribution:

Lemma 4.18 *Let τ and τ' be distributions and $0 < \varepsilon < 1$ such that $\tau' \ge (1-\varepsilon)\tau$. Then for any starting distribution σ,*

$$\mathcal{H}(\sigma,\tau) \le \mathcal{H}(\sigma,\tau') + \frac{\varepsilon}{1-\varepsilon}\mathcal{H}(\tau')\,.$$

In particular,

$$\mathcal{H}(\tau) \le \frac{1}{1-\varepsilon}H(\tau')\,.$$

There is an analogous property of starting distributions:

Lemma 4.19 *Assume that $\sigma' \geq (1-\varepsilon)\sigma$. Then for every target distribution τ,*

$$\mathcal{H}(\sigma', \tau) \leq (1-\varepsilon)\mathcal{H}(\sigma, \tau) + \varepsilon\mathcal{H}(\tau)\,.$$

In particular,

$$\mathcal{H}(\sigma', \sigma) \leq \varepsilon\mathcal{H}(\sigma)\,.$$

Proof. This follows from the decomposition

$$\sigma = (1-\varepsilon)\rho + \varepsilon\left(\frac{1}{\varepsilon}\sigma - \left(\frac{1}{\varepsilon} - 1\right)\rho\right)\,.$$

and the convexity of $\mathcal{H}(\sigma, \tau)$ as a function of σ. □

We also have

$$\|z(\sigma,\rho)\| = d(\sigma,\rho)\|z(\sigma\backslash\rho,\ \rho\backslash\sigma)\| \leq d(\sigma,\rho)\Big(\|z(\sigma\backslash\rho,\ \pi)\| + \|z(\rho\backslash\sigma,\ \pi)\|\Big) \leq 4d(\sigma,\rho)\mathcal{Z}\,. \quad (22)$$

We conclude with another simple lemma:

Lemma 4.20 *For any three distributions σ and $\overline{\sigma}$ and τ there exists a distribution $\overline{\tau}$ such that*

$$d(\tau, \overline{\tau}) \leq d(\sigma, \overline{\sigma})$$

and

$$\mathcal{H}(\overline{\sigma}, \overline{\tau}) \leq \mathcal{H}(\sigma, \tau)\,.$$

Another way of saying this is that if $d(\sigma, \overline{\sigma}) = \varepsilon$ then for every τ,

$$\mathcal{H}_\varepsilon(\overline{\sigma}, \tau) \leq \mathcal{H}(\sigma, \tau)\,.$$

Proof. Let Γ be an optimal stopping rule from σ to τ. Let v^0 be a state from distribution σ and $\overline{v}^0$, a state from distribution $\overline{\sigma}$, coupled so that

$$\mathsf{P}(v^0 = \overline{v}^0) \geq 1 - \varepsilon\,.$$

Define a new stopping rule $\overline{\Gamma}$ by

$$\overline{\Gamma} = \begin{cases} \Gamma, & \text{if } v^0 = \overline{v}^0, \\ 0, & \text{otherwise.} \end{cases}$$

Then $v^{\overline{\Gamma}} = v^\Gamma$ with probability at least $1-\varepsilon$, and so

$$d(\overline{\sigma}^{\overline{\Gamma}}, \sigma^\Gamma) = d(\overline{\sigma}^{\overline{\Gamma}}, \tau) \leq \varepsilon\,.$$

Since trivially $\mathsf{E}\overline{\Gamma} \leq \mathsf{E}\Gamma$, this proves the lemma. □

4.7 Approximation by averaging

If we consider only stopping at some fixed time, even approximating π may take far more time than a mean-optimal rule (for example when the chain is almost periodic). However, there are very simple blind *approximate* stopping rules that are easily implementable and give a good approximation of the stationary distribution, in expected time only a constant factor more than the mixing time.

The *uniform averaging rule* $\Upsilon = \Upsilon_t$ ($t \geq 0$) is defined as follows: choose a random integer Y uniformly from the interval $\{0, \ldots, t-1$, and stop after Y steps. (To describe this as a stopping rule: stop after the u-th step with probability $1/(t-u)$ ($j = 0, \ldots, t-1$).)

We shall give estimates on how close the distribution σ^Υ of the state generated by the averaging rule is to the stationary. To this end, we derive an explicit formula for the distribution of the state produced by the averaging rule.

Lemma 4.21

$$\sigma^\Upsilon = \pi + \frac{1}{t} z(\sigma, \sigma^t).$$

Proof. Let x_i denote the expected number of times state i is exited during the first t steps. Note that $\sigma_i^\Upsilon = x_i/t$. Then the x_i are the exit frequencies of a rule from σ to σ^t, and hence by Theorem 4.3, we have

$$x_i - \pi_i t = z_i(\sigma, \sigma^t),$$

whence the lemma follows. $\square$

With a little more care we can obtain a result that estimates the error of the uniform averaging rule in terms of the approximate mixing time *starting from the same distribution.*

Theorem 4.22 *For any starting distribution σ and $0 \leq \varepsilon \leq 1$,*

$$d(\sigma^\Upsilon, \pi) \leq \varepsilon + \frac{1}{t}\mathcal{H}(\sigma, d_\varepsilon).$$

In particular,

$$d(\sigma^\Upsilon, \pi) \leq \frac{1}{t}\mathcal{H}(\sigma, \pi).$$

Proof. Let τ be a distribution such that $d(\tau, \pi) \leq \varepsilon$ and $\mathcal{H}(\sigma, \tau) = \mathcal{H}(\sigma, d_\varepsilon)$. Then

$$z(\sigma, \sigma^t) = z(\sigma, \tau) + z(\tau, \tau^t) + z(\tau^t, \sigma^t)$$

and hence for every $U \subseteq V$,

$$\begin{aligned}
\sum_{k \in U} (\sigma_k^Y - \pi_k) &= \frac{1}{t} \sum_{k \in U} z_k(\sigma, \sigma^t) \\
&= \frac{1}{t} \sum_{k \in U} z_k(\sigma, \tau) + \frac{1}{t} \sum_{k \in U} z_k(\tau, \tau^t) + \frac{1}{t} \sum_{k \in U} z_k(\tau^t, \sigma^t).
\end{aligned}$$

Putting $\pi_U := \sum_{i \in U} \pi_i$, the first term above is at most $(1 - \pi_U)\mathcal{H}(\sigma, \tau)$ and the last term is at most

$$\pi_U \mathcal{H}(\sigma^t, \tau^t) \leq \pi_U \mathcal{H}(\sigma, \tau)$$

by (13) and (15). For the middle term, we use the rough bound

$$\sum_{k\in U} z_k(\tau,\tau^t) = \sum_{k\in U}\sum_{m=0}^{t-1}(\tau^m - \pi) \le \sum_{m=0}^{t-1} d(\tau^m,\pi) \le td(\tau,\pi) \le t\varepsilon\,.$$

Substituting these bounds, the theorem follows. (See [23] for a simple direct proof.) □

Theorem 4.22 asserts closeness in the total variation distance; approaching π pointwise (as noted, for the reversible case, in [1]) turns out to be somewhat harder in general.

Below we establish results about the pointwise distance of distributions obtained by various averaging rules. However, we shall have to use the worst-case bound on the mixing time.

Thus we cannot generally control the quotient σ_i^Y/π_i by averaging over times of order $\mathcal{H}$. In fact, we have see in Section 3.1 that the time needed to make a two-sided approach to π using averaging is at least the order of the *maximum* length of a max-optimal stopping rule that achieves π exactly.

We can show, however, that in time $O(\mathcal{H})$, the excess of σ^Υ over π can be decreased to an arbitrarily small fraction of its original value.

Theorem 4.23 *Assume that $\sigma \le K\pi$. Then*

$$\sigma^\Upsilon \le \left(1 + \frac{1}{t}K\mathcal{H}\right)\pi\,.$$

Proof. By Lemma 4.21, the triangle inequality, and Lemma 4.16,

$$\begin{aligned}
\sigma_i^Y &= \pi_i + \frac{1}{t}z_i(\sigma,\sigma^t) = \pi_i - \frac{1}{t}z_i(\sigma^t,\sigma)\\
&\le \pi_i + \frac{1}{t}\pi_i\mathcal{H}(\sigma^t,\sigma) \le \pi_i + \frac{1}{t}\pi_i(\mathcal{H}(\sigma^t,\pi)+\mathcal{H}(\pi,\sigma))\\
&\le \pi_i + \frac{1}{t}\pi_i(\mathcal{H}(\sigma,\pi)+\mathcal{H}(\pi,\sigma))\,.
\end{aligned}$$

Here we know that $\pi \ge \frac{1}{K}\sigma$, and so by Lemma 4.19 we get

$$H(\pi,\sigma) \le (K-1)\mathcal{H}\,,$$

and thus

$$\sigma_i^Y \le \left(1 + \frac{1}{t}K\mathcal{H}\right)\pi_i\,.$$

□

Our next goal is to describe a rule giving a point that has a probability of at least $(1-\varepsilon)\pi_i$ of being at state i. We assume that the starting distribution is already close to π in the total variation distance (this can be achieved, by the above, using the averaging rule), and do another averaging. As before, let Y be chosen uniformly from $\{0,\dots,t-1\}$. The main result is the following.

Theorem 4.24 *Let $0 \le \delta \le \varepsilon \le 1$ and assume that $d(\sigma,\pi) = \varepsilon$ and $\sigma \ge (1-\delta)\pi$. Then*

$$\sigma^Y \ge \left(1 - \varepsilon - \frac{\delta}{t}\mathcal{H}\right)\pi\,.$$

Proof. Let $\alpha = \sigma\backslash\pi$ and $\beta = \pi\backslash\sigma$, so that we can write

$$\sigma = \pi + \varepsilon\alpha - \delta\beta \, .$$

Clearly, the supports of α and β are disjoint, and hence $\sigma \geq (1-\delta)\pi$ implies that $\beta \leq (\delta/\varepsilon)\pi$. Clearly

$$\sigma^{\overline{Y}} = \pi^{\overline{Y}} + \varepsilon\alpha^{\overline{Y}} - \varepsilon\beta^{\overline{Y}} = \pi + \varepsilon\alpha^{\overline{Y}} - \varepsilon\beta^{\overline{Y}} \geq \pi - \varepsilon\beta^{\overline{Y}} \, .$$

Now by Theorem 4.23, we have

$$\beta^{\overline{Y}} \leq \left(1 + \frac{\delta}{\varepsilon t}\mathcal{H}\right)\pi \, ,$$

and hence

$$\sigma^{\overline{Y}} \geq \left(1 - \varepsilon - \frac{\delta}{t}\mathcal{H}\right)\pi$$

as claimed. □

It follows that in order to fill π up to a factor $1-\varepsilon$, it suffices to choose two integers Y_1 and Y_2 uniformly and independently from $\{0, \ldots, \mathcal{H}/\varepsilon\}$, and then do $Y_1 + Y_2$ steps; symbolically, $\mathcal{H}_{\underline{\varepsilon}} \leq \mathcal{B}_{\underline{\varepsilon}} \leq \mathcal{H}/\varepsilon$.

This result is not entirely satisfactory, however; one would like to see that the error diminishes exponentially with t or, in other words, that the time needed is proportional to $\log(1/\varepsilon)$ rather than to $1/\varepsilon$. Next we describe another simple averaging rule that achieves this. The result below also follows by adaptation of the "multiplicativity property" in Aldous [1].

Let $M > 0, t = 8\lceil\mathcal{H}\rceil$, and let X be the sum of M independent random variables $Y_1, \ldots, Y_M$, distributed uniformly in $\{0, \ldots, t-1\}$. Stop at v^X.

To analyze this rule, let $X_k = Y_1 + \ldots Y_k$.

Lemma 4.25 *For all $k \geq 1$, we have*

$$d(\sigma^{X_k}, \pi) \leq 2^{-(k+1)}$$

and

$$\sigma^{X_k} \geq \left(1 - 2^{-(k-1)}\right)\pi \, .$$

We prove this inequality by induction on k. For $k = 1$ the inequality follows by Theorem 4.22. Assume that $k > 1$. Then σ^{X_k} can be obtained from $\sigma^{X_{k-1}}$ by the uniform averaging rule, and hence we get by Theorem 4.24 that

$$\sigma^{X_k} \geq \left(1 - 2^{-k} - \frac{1}{8}2^{-(k-2)}\right)\pi > (1 - 2^{-(k-1)})\pi \, . \tag{23}$$

On the other hand, we have by the induction hypothesis and Lemma 4.19 that

$$\mathcal{H}(\sigma^{X_{k-1}}, \pi) \leq 2^{-(k-2)}\mathcal{H} \, ,$$

and hence by Theorem 4.22,

$$d(\sigma^{X_k}, \pi) \leq \frac{1}{8}2^{-(k-2)} = 2^{-(k+1)} \, .$$

This completes the induction. □

Choose $M = \lceil\log(1/\varepsilon)\rceil$, and denote the resulting rule by Γ_ε. Then we have, as an immediate consequence of Lemma 4.25, the following theorem.

Theorem 4.26 *For any starting distribution σ, the rule Γ_ε produces a distribution τ satisfying*

$$\tau \geq (1-\varepsilon)\pi\,,$$

and has mean length $\mathrm{O}(\mathcal{H}\log(1/\varepsilon))$.

The next lemma allows us in some cases to consider only bounded averaging rules.

Lemma 4.27 *Let σ be any starting distribution. Suppose that there exists an averaging rule* Y *and a* $0<\delta<1$ *such that*

$$\sigma^{\mathrm{Y}} \leq (1+\delta)\pi.$$

Then there exists an averaging rule W *such that* $\mathsf{E}\mathrm{W} \leq \mathsf{E}\mathrm{Y}$, $\max \mathrm{W} \leq (2/\delta)\mathsf{E}Y$ *and*

$$\sigma^{\mathrm{W}} \leq (1+3\delta)\pi.$$

Proof. Let t be the smallest integer such that $\mathsf{P}(Y>t)\leq \delta/2$. Define a non-negative integer valued random variable W by

$$\mathsf{P}(\mathrm{W}=k) = \mathsf{P}(Y=k \mid Y\leq t)\,.$$

Then clearly

$$\sigma_i^{\mathrm{W}} = \mathsf{P}(v^{\mathrm{W}}=i) \leq \frac{\sigma_i^Y}{\mathsf{P}(Y\leq t)} \leq \left(\frac{1+\delta}{1-\delta/2}\right)\pi \leq (1+3\delta)\pi\,,$$

and by Markov's inequality,

$$\max(\mathrm{W}) \leq t \leq \frac{2}{\delta}\mathsf{E}\mathrm{Y}\,.$$

□

4.8 Parameter dependence

Some of our mixing measures involve one or two somewhat arbitrary parameters: approximate mixing times depend on the error in the approximation (usually denoted by ε), while quantities defined for warm start depend on the bound on the density of the starting distribution (often set to 2). In this section we give a brief discussion of this dependence.

Of course, when $\varepsilon=0$ all approximate mixing times are the same:

$$\mathcal{H}_{\underline{0}} = \mathcal{H}_{\overline{0}} = \mathcal{H}_0 = \mathcal{H}$$

and

$$\mathcal{F}_{\underline{0}} = \mathcal{F}_{\overline{0}} = \mathcal{F}_0 = \mathcal{F}\,.$$

and in general

$$\mathcal{H}_\varepsilon \leq \mathcal{H}_{\underline{\varepsilon}} \leq \mathcal{H}$$

and

$$\mathcal{F}_\varepsilon \leq \mathcal{F}_{\underline{\varepsilon}} \leq \mathcal{F}\,.$$

Similar inequalities hold for the disperse times and also for the blind rules.

The conventional wisdom is that the error drops exponentially with time, or put it in another way, the mixing time depends logarithmically on the error. This is indeed true in several situations, but not always. Example 2.4 illustrates that the dependence on the error parameter can be quite tricky.

A nice clean result is due to Aldous and Diaconis [5]. Let $e(t) = \max d(\sigma^t, \rho^t)$, where the maximum is taken over all starting distributions σ and ρ. Clearly the maximum is attained when σ and ρ are concentrated on singletons. Note that obviously $d(\sigma^t, \pi) = d(\sigma^t, \pi^t) \le e(t)$ for every σ, and $e(t) \le 2\max_\sigma d(\sigma^t, \pi)$.

Lemma 4.28 *For every* $t_1, t_2 \ge 0$,

$$e(t_1 + t_2) \le e(t_1)e(t_2).$$

Thus if we know, say, that $d(\sigma^t, \pi) \le 1/4$ for every starting distribution σ, then it follows that $e(t) \le 1/2$, and hence $d(\sigma^{kt}, \pi) \le e(kt) \le 1/2^k$.

One can carry over this argument to blind rules, and prove

$$\mathcal{B}_{2\varepsilon_1\varepsilon_2} \le \mathcal{B}_{\varepsilon_1}\mathcal{B}_{\varepsilon_2}.$$

Note that we do not get a similar result for the uniform averaging rule, because there the error in σ decreases only as $1/t$ as the parameter t increases.

For the approximate mixing (filling) time $\mathcal{H}_{\overline{\varepsilon}}$, Theorem 5.9 implies that it is essentially independent of ε. Similar assertion holds for the approximate mixing times $\mathcal{H}_\varepsilon$ and $\mathcal{H}_{\underline{\varepsilon}}$ in the time-reversible case. In the case of general Markov chains, we know from Theorem 5.1 below that

$$\mathcal{H}_{\underline{\varepsilon}} \le H_\varepsilon \le 16(1 + \ln(1/\varepsilon))\mathcal{H}_{1/8},$$

which implies a similar exponential convergence.

In what follows, we study dependence of parameters when we also put restrictions on the starting distribution. Our main interest is in the case of warm start, but we need results about more general restrictions.

Lemma 4.29 *For every* $0 < a, b \le 1$,

$$\mathcal{B}(D_1, D_{ab}) \le \mathcal{B}(D_1, D_a) + \mathcal{B}(D_1, D_b)\,.$$

Proof. Let $\sigma \le 2\pi$ and let Γ be a blind stopping rule such that

$$(1-a)\pi \le \sigma^\Gamma \le (1+a)\pi \qquad \text{and} \qquad \mathsf{E}\Gamma \le \mathcal{B}(D_1, D_a)\,.$$

Let $\rho = (1/a)\sigma^\Gamma + (1 - (1/a))\pi$. Then ρ is a distribution with $\rho \le 2\pi$, and hence there exists a blind stopping rule Δ, independent of Γ, such that

$$(1-b)\pi \le \rho^\Delta \le (1+b)\pi \qquad \text{and} \qquad \mathsf{E}\delta \le \mathcal{B}(D_1, D_b)\,.$$

Substituting the definition of ρ, this implies that

$$(1-ab)\pi \le \sigma^{\Gamma+\Delta} \le (1+ab)\pi.$$

Since

$$\mathsf{E}(\Gamma + \Delta) \le \mathcal{B}(D_1, D_a) + \mathcal{B}(D_1, D_b)\,,$$

this proves the lemma. □

Corollary 4.17 implies the first half of the following lemma; the second half is trivial.

Lemma 4.30 *Let $A, B \subset D_1$. Then*

$$\mathcal{H}(A,B) = \mathcal{H}(2\pi - A, 2\pi - B)\,,$$

and

$$\mathcal{B}(A,B) = \mathcal{B}(2\pi - A, 2\pi - B)\,.$$

We say that a set A of distributions is *closed*, if it is convex and for every $\sigma \in A$, also $\sigma^1 \in A$.

Lemma 4.31 *Let A, B, C be sets of distributions, and assume that B is closed. Then*

$$\mathcal{B}(A, B \cap C) \le \mathcal{B}(A,B) + \mathcal{B}(B,C).$$

and

$$\mathcal{B}(B \cap A, B \cap C) \le \mathcal{B}(A,C).$$

Proof. Let $\sigma \in A$. Apply an optimal blind rule Γ from σ to some $\tau \in B$. Then apply an optimal blind rule Δ from τ to some $\rho \in C$. Since B is closed, it follows that $\rho \in B$. Thus $\mathcal{B}(\sigma, B \cap C) \le \mathsf{E}\Gamma + \mathsf{E}\Delta$, which proves the first inequality. The second inequality is trivial, if we observe that a blind stopping rule from a $\sigma \in A$ to some $\tau \in C$ yields a distribution in $B \cap C$ is σ happens to be in $B \cap A$. □

Corollary 4.32

$$\mathcal{B}(D_1, D_{3/2}) \le 2\mathcal{B}(\le 2\pi, \le (3/2)\pi)\,.$$

For $0 \le c \le 1$, let $W_c : D \to D$ denote the linear map

$$W_c\sigma = c\sigma + (1-c)\pi.$$

Lemma 4.33 *(a) Let A and B be two sets of distributions. Then*

$$\mathcal{H}(W_cA, W_cB) = c\mathcal{H}(A,B)\,,$$

and

$$\mathcal{B}(W_cA, W_cB) = \mathcal{B}(A,B)\,.$$

Note that $\mathcal{H}$ and $\mathcal{B}$ show surprisingly different behavior in terms of c: to get, say, from pointwise relative error ε to relative error $\varepsilon/2$ takes the same time for a blind rule independently of ε, but time proportional to ε for a general rule.

Proof. By Corollary 4.17, we have for every σ and τ,

$$\mathcal{H}(W_c\sigma, W_c\tau) = c\mathcal{H}(\sigma,\tau).$$

This implies the first equation.

For blind rules, it suffices to prove that

$$\mathcal{B}(W_c\sigma, W_c\tau) = \mathcal{B}(\sigma,\tau).$$

for every σ and τ (note that both sides may be infinite). Indeed, let Γ be a blind rule such that $\sigma^\Gamma = \tau$ and $\mathsf{E}\Gamma = \mathcal{B}(\sigma, \tau)$. Then

$$(W_c\sigma)^\Gamma = c\sigma^\Gamma + (1-c)\pi = W_c\tau$$

(here we use that $\pi^\Gamma = \pi$ for a blind rule), and hence

$$\mathcal{B}(W_c\sigma, W_c\tau) \leq \mathsf{E}\Gamma = \mathcal{B}(\sigma, \tau).$$

The reverse inequality follows similarly. □

Lemma 4.34 *For every set A of distributions and every $0 < \varepsilon < 1$,*

$$\mathcal{H}(A, \pi) \leq \frac{1}{1-\varepsilon}\mathcal{H}(A, W_\varepsilon A).$$

(Note that if $\pi \in A$ then trivially $\mathcal{H}(A, \pi) \geq \mathcal{H}(A, W_{1/2}A$.)
Proof. Using (7) and Lemma 4.33, we have

$$\mathcal{H}(A, \pi) \leq \sum_{k=0}^{\infty} \mathcal{H}(W_{\varepsilon^k}A, W_{\varepsilon^{k+1}}A) = \left(\sum_{k=0}^{\infty} \varepsilon^k\right) \mathcal{H}(A, W_\varepsilon A) = \frac{1}{1-\varepsilon}\mathcal{H}(A, W_\varepsilon A)\,.$$

□

The last lemma in this section bounds $\mathcal{H}_\varepsilon(\rho, \pi)$ if we only know that $\rho \leq c\pi$ for some $c > 2$.

Lemma 4.35 *Let ρ be a distribution such that $\rho \leq r\pi$ for some $c \geq 2$. Then*

$$\mathcal{H}_\varepsilon(\rho, \pi) \leq (c-1)\tilde{\mathcal{H}}_{\varepsilon/(2c-3)}\,.$$

Proof. Let

$$h = \frac{\varepsilon}{2c-3} \qquad \text{and} \qquad \alpha = \frac{\rho + (c-2)\pi}{c-1}\,.$$

Then α is a distribution with $\alpha \leq 2\pi$, and hence there exists a distribution τ such that $d(\tau, \pi) = h$ and $\mathcal{H}(\alpha, \tau) \leq \tilde{\mathcal{H}}_h$. Then we can write $\tau = \pi - h\beta + h\gamma$ for some distributions β, γ.

Let

$$u = \frac{1}{1 + h(c-2)}\,,$$

and consider the distributions $\rho' = u\rho + (1-u)\beta$ and $\tau' = u\tau + (1-u)\gamma$. Then

$$\rho' - \tau' = u(\rho - \tau) + (1-u)(\beta - \gamma) = \frac{c-1}{1 + h(c-2)}(\alpha - \tau)\,.$$

Hence

$$\mathcal{H}(\rho', \tau') = \frac{c-1}{1 + h(c-2)}H(\alpha, \tau) < (c-1)\tilde{\mathcal{H}}_h\,.$$

But $d(\rho, \rho') \leq 1 - u < h(c-2)$, and thus by Lemma 4.20, there exists a distribution τ'' such that $d(\tau'', \tau') < h(c-2)$ and $H(\rho, \tau'') \leq H(\rho'\tau')$. Now we have

$$d(\tau'', \pi) \leq d(\tau'', \tau') + d(\tau', \tau) + d(\tau, \pi) \leq h(c-2) + h(c-2) + h = (2c-3)h = \varepsilon\,.$$

Thus

$$\mathcal{H}_\varepsilon(\rho, \pi) \leq \mathcal{H}(\rho, \tau'') \leq \mathcal{H}(\rho', \tau') < (c-1)\tilde{\mathcal{H}}_h\,.$$

□

5 Proofs of the main results

We prove the theorems stated in Chapter 3.1. In fact, we start each proof by stating the theorems with explicit constants; in some cases, we introduce new mixing parameters that facilitate the proof and are perhaps interesting on their own. We treat the groups in a slightly different order, since some partial results from one proof are needed in another.

5.1 The forget group: proof of Theorem 3.2

We start with a more complete statement of the theorem.

Theorem 5.1 *For every finite Markov chain and $0 < \varepsilon < 1/2$, we have the following inequalities.*

$$(1-2\varepsilon)\mathcal{Z} \le \mathcal{F}_\varepsilon \le \mathcal{F}_{\underline{\varepsilon}} \le 2\mathcal{H}_{\varepsilon/2} \le 2\mathcal{H}_{\overline{\varepsilon/2}} \le 4(1+\ln(2/\varepsilon))\mathcal{S} \le 4(1+\ln(2/\varepsilon))\mathcal{Z}\,.$$

In addition,

$$\mathcal{Z} \le \mathcal{F} \le \frac{1}{1-\varepsilon}\mathcal{F}_{\underline{\varepsilon}}\,.$$

Corollary 5.2

$$\mathcal{S} \le \mathcal{Z} \le \mathcal{F} \le 8\mathcal{S}\,.$$

We break down the proof into several steps, one for each inequality claimed. The first step is (up to the constant) a stronger version of (16):

Lemma 5.3 *Let $0 < \varepsilon < 1/2$, then*

$$\mathcal{Z} \le \frac{1}{1-2\varepsilon}\mathcal{F}_\varepsilon\,.$$

Proof. Let ρ be any distribution and let σ be a distribution maximizing $\|z(\sigma,\pi)\|$. By definition, there exists a distribution μ such that $d(\mu,\rho) \le \varepsilon$ and $\mathcal{H}(\sigma,\mu) \le \mathcal{H}_\varepsilon(\rho)$. Then for an appropriate set $A \subseteq V$, we have

$$\mathcal{Z} = \sum_{i\in A} z_i(\sigma,\pi) = \sum_{i\in A} z_i(\sigma,\mu) + \sum_{i\in A} z_i(\mu,\rho)\,.$$

Here by (16),

$$\sum_{i\in A} z_i(\sigma,\mu) = \sum_{i\in V\setminus A} z_i(\mu,\sigma) \le (1-\pi(A))\mathcal{H}(\sigma,\mu) \le (1-\pi(A))\mathcal{H}_\varepsilon(\rho)\,.$$

To estimate the other term, we have by (22)

$$\sum_{i\in A} z_i(\mu,\rho) \le \frac{1}{2}\|z(\mu,\rho)\| \le 2d(\mu,\rho)\mathcal{Z} \le 2\varepsilon\mathcal{Z}\,.$$

Hence

$$\mathcal{Z} \le \mathcal{H}_\varepsilon(\rho) + 2\varepsilon\mathcal{Z}\,,$$

and the lemma follows. □

An immediate consequence of Lemma 4.21 is the following inequality. Choosing $t = (2/\varepsilon)\mathcal{Z}$, the uniform averaging rule yields a distribution which is within variation distance ε of π. Since the averaging rule Υ is a blind rule, we have shown

Corollary 5.4 *For every* $0 \le \varepsilon \le 1$,

$$\mathcal{H}_\varepsilon \le \mathcal{B}_\varepsilon \le \mathcal{U}_\varepsilon \le \frac{2}{\varepsilon}\mathcal{Z}\,.$$

The inequality

$$\mathcal{F}_\varepsilon \le \mathcal{F}_{\underline{\varepsilon}}$$

is trivial.

Lemma 5.5 *For* $\varepsilon < 1/2$,

$$\mathcal{F}_{\underline{2\varepsilon}} \le 2\mathcal{H}_\varepsilon\,.$$

Proof. For each state k, let $\tau(k)$ be a distribution such that $d(\tau(k),\pi) \le \varepsilon$ and $\mathcal{H}(k,\tau(k)) \le \mathcal{H}_\varepsilon$. Let

$$f_i = \min_\mu \sum_k \tau(k)_i \mu_k\,,$$

where μ ranges over all distributions with $d(\mu,\pi) \le \varepsilon$. Let $\mu(i)$ be the distribution achieving this minimum. Write $\mu(i)_k = \pi_k + \alpha(i)_k$ and $\tau(k)_i = \pi_i + \beta(k)_i$. Let $a(i)_k = \min(0,\alpha(i)_k)$ and $b(k)_i = \min(0,\beta(k)_i)$. Then

$$\tau(k)_i\mu(i)_k \ge (\pi_i + a(i)_k)(\pi_k + b(k)_i) \ge \pi_i\pi_k + \pi_i a(i)_k + \pi_k b(k)_i$$

and hence

$$\sum_i f_i = \sum_{i,k} \tau(k)_i\mu(i)_k \ge 1 + \sum_i \pi_i \sum_k a(i)_k + \sum_k \pi_k \sum_i b(k)_i\,.$$

Now here, for any fixed i,

$$\sum_k a(i)_k = -d(\pi,\mu(i)) \ge -\varepsilon\,,$$

and similarly

$$\sum_i b(k)_i = -d(\pi,\tau(k)) \ge -\varepsilon\,.$$

Hence

$$\sum_i f_i \ge 1 - \sum_i \pi_i\varepsilon - \sum_k \pi_k\varepsilon = 1 - 2\varepsilon\,.$$

Consider the distribution

$$\rho_i = f_i \Big/ \sum_j f_j\,,$$

and the following stopping rule: starting at state k, follow an optimal rule from k to $\tau(k)$; if you end at j, then follow an optimal rule from j to $\tau(j)$. Let $\theta(k)$ be the distribution produced. Then

$$\theta(k)_i = \sum_j \tau(k)_j\tau(j)_i \ge f_i \ge (1-2\varepsilon)\rho_i\,.$$

Thus $\mathcal{H}_{\underline{2\varepsilon}}(\rho) \le 2\mathcal{H}_\varepsilon$. □

Lemma 4.7 implies that

Corollary 5.6 *For every* $0 < \varepsilon < 1$,

$$\mathcal{H}_{\overline{\varepsilon}} \leq (1+\varepsilon)\ln(1+1/\varepsilon))\mathcal{S}\,.$$

Lemma 5.7

$$\mathcal{S} \leq \mathcal{Z}/2.$$

Proof. Let σ be any distribution and let $A \subseteq V$. Let ρ be the distribution of the first element in A, starting the chain from σ. Note that $\mathcal{H}(\sigma,\rho) = \mathcal{H}(\sigma, A)$ is just the expected number of steps before hitting A, starting from σ. Also note that trivially $x_i(\sigma,\rho) = 0$ for $i \in A$, and hence

$$\pi_i \mathcal{H}(\sigma,\rho) = -z_i(\sigma,\rho) = z_i(\rho,\sigma)\,.$$

Hence

$$\pi(A)\mathcal{H}(\sigma, A) = \sum_{i\in A} z_i(\rho,\sigma) \leq \frac{1}{2}\mathcal{Z}.$$

□

Corollary 5.8

$$\mathcal{S} \leq \mathcal{Z}\,.$$

To complete the proof of Theorem 5.1, it suffices to notice that letting $\varepsilon \to 0$ in Lemma 5.3 we get

$$\mathcal{Z} \leq \mathcal{F}\,,$$

while Corollary 4.15 implies that

$$\mathcal{F} \leq \mathcal{H}(\rho) \leq \frac{1}{1-\varepsilon}\mathcal{H}_{\underline{\varepsilon}}(\rho) = \frac{1}{1-\varepsilon}\mathcal{F}_{\underline{\varepsilon}}\,.$$

5.2 The mixing group: proof of Theorem 3.1

We state and prove two theorems separately, which together imply Theorem 3.1.

Theorem 5.9

$$\mathcal{H}_{\underline{\varepsilon}} \leq \mathcal{H} \leq \frac{1}{1-\varepsilon}\mathcal{H}_{\underline{\varepsilon}}\,.$$

Proof. The first inequality is trivial; the second follows immediately from Corollary 4.15. □

Theorem 5.10

$$\mathcal{H} \leq 2\Big(\mathcal{S} + \mathcal{I}\Big)\,.$$

Proof. Let $S := \{j :\ \mathcal{H}(j,\pi) \leq 2\mathcal{I}\}$. Then

$$\mathcal{I} = \sum_i \pi_i \mathcal{H}(i,\pi) \geq \sum_{i\in V\setminus S} \pi_i \mathcal{H}(i,\pi) \geq \pi(V\setminus S)(2\mathcal{I}),$$

and hence $\pi(S) \geq 1/2$. Now for any starting state s, we can walk from s until we hit some state $j \in S$, and then follow an optimal rule from j to π. The expected length of this walk is at most

$$\mathcal{H}(s,S) + 2\mathcal{I} \leq \frac{1}{\pi(S)}\mathcal{S} + 2\mathcal{I} \leq 2\mathcal{S} + 2\mathcal{I}.$$

$\square$

Using the inequality $\mathcal{S} \le \mathcal{I}$ (see Corollary 5.2 below), we get

$$\mathcal{H} \le 2(\mathcal{F} + \mathcal{I}) \le 4\max\{\mathcal{F}, \mathcal{I}\}.$$

Since trivially $\max\{\mathcal{F}, \mathcal{I}\} \le \mathcal{H}$, this completes the proof of Theorem 3.1.

5.3 The reset group: proof of Theorem 5.11

Besides the reset time $\mathcal{I}$ and the warm mixing time $\tilde{\mathcal{H}}$, this group also contains the mixing measures in the forget group, defined for the reverse chain. Of these, our proof uses the discrepancy of the reverse chain,

$$\overleftarrow{\mathcal{Z}} = \max_{\substack{v\in V\\ S\subseteq V}} \sum_{i\in S} \overleftarrow{z}_i(\pi, v)$$

Theorem 5.11

$$\frac{1}{16}\mathcal{I} \le \overleftarrow{\mathcal{Z}} \le \tilde{\mathcal{H}} \le 2\mathcal{I}\,.$$

The first inequality is immediate by Corollary 5.2 and Theorem 3.8. To prove the second, let $u, v \in V$ and $A \subset V$ attain the maximum in the definition of $\overleftarrow{Z}$. Since

$$\overleftarrow{z}_i(v, \pi) = \frac{\pi_i}{\pi_v} z_v(i, \pi) = \pi_i(\mathcal{H}(i, v) - \mathcal{H}(\pi, v)),$$

this implies that

$$A = \{i \in V:\ \mathcal{H}(i, v) \ge \mathcal{H}(\pi, v)\}.$$

Case 1. Assume that $\pi(A) \ge 1/2$. Then π_A is a distribution with $\pi_A \le 2\pi$. Moreover,

$$\begin{aligned}
\mathcal{H}(\pi_A, \pi) &\ge \mathcal{H}(\pi_A, v) - \mathcal{H}(\pi, v) = \sum_{i\in A} \frac{\pi_i}{\pi(A)}(\mathcal{H}(i, v) - \mathcal{H}(\pi, v)) \\
&= \frac{1}{\pi(A)} \sum_{i\in A} \frac{\pi_i}{\pi_v} z_v(i, \pi) = \frac{1}{\pi(A)} \sum_{i\in A} \overleftarrow{z}_i(v, \pi) = \frac{1}{\pi(A)} \overleftarrow{\mathcal{Z}} > \overleftarrow{\mathcal{Z}}\,.
\end{aligned}$$

Case 2. Assume that $\pi(A) < 1/2$. Set $t = (1 - 2\pi(A))/(1 - \pi(A))$ and define

$$\rho_i = \begin{cases} 2\pi_i, & \text{if } i \in A, \\ t\pi_i, & \text{otherwise.} \end{cases}$$

Then again ρ is a distribution with $\rho \le 2\pi$. Moreover,

$$\begin{aligned}
\mathcal{H}(\rho, \pi) &\ge \mathcal{H}(\rho, v) - \mathcal{H}(\pi, v) \\
&= \sum_{i\in A} 2\pi_i(\mathcal{H}(i, v) - \mathcal{H}(\pi, v)) + \sum_{i\in V\setminus A} t\pi_i(\mathcal{H}(i, v) - \mathcal{H}(\pi, v)) \\
&= (2 - t) \sum_{i\in A} \pi_i(\mathcal{H}(i, v) - \mathcal{H}(\pi, v)) = \frac{1}{1 - \pi(A)} \sum_{i\in A} \frac{\pi_i}{\pi_v} z_v(i, \pi) \\
&= \frac{1}{1 - \pi(A)} \sum_{i\in A} \overleftarrow{z}_i(v, \pi) = \frac{1}{1 - \pi(A)} \overleftarrow{\mathcal{Z}} > \overleftarrow{\mathcal{Z}}\,.
\end{aligned}$$

This proves the second inequality.

Finally, the third inequality is easy. Assume that $\sigma \le 2\pi$, and consider the following rule from σ to π: look at the starting state and follow an optimal rule from that state. Hence

$$\mathcal{H}(\sigma,\pi) \le \sum_i \sigma_i \mathcal{H}(i,\pi) \le 2\sum_i \pi_i \mathcal{H}(i,\pi) = \mathcal{I}\,.$$

5.4 The maxing group: proof of Theorem 3.4

Lemma 5.12 *Let σ and τ be any two distributions. If $x_i(\sigma,\tau) \le \sigma_i$ for every state i, then $\mathcal{M}(\sigma,\tau) \le 1$.*

Proof. Consider the rule "if you are at i, make one step with probability $x_i(\sigma,\tau)/\sigma_i$ and stop, else stop right away". This has the right exit frequencies, and hence it yields τ. □

Lemma 5.13 *Suppose that there exists a blind rule Δ such that for any initial distribution σ,*

$$4\pi/5 \le \sigma^\Delta \le 5\pi/4\,.$$

Then for every σ there is a (non-blind) stopping rule Γ with $\sigma^\Gamma = \pi$ and $\max(\Gamma) \le 2\max(\Delta)$.

Proof. Let $\max(\Delta) = t$ and $a_j = \mathsf{P}(\Delta = j)$ for $j = 0, 1, \ldots, t$. Replacing the transition matrix M by $\sum_{j=0}^{t} a_j M^j$ preserves the stationary distribution, while the condition of the theorem becomes

$$4\pi/5 \le \sigma^1 \le 5\pi/4\,.$$

for all σ. It now suffices to show that $\mathcal{M}(\sigma^1,\pi) \le 1$.

To do this we bound the exit frequencies $x_i(\sigma^1,\pi)$. Note that for our new fast-mixing Markov chain, we have $\underline{\mathcal{H}_{1/5}} \le 1$, and hence by Corollary 4.15, we have $\mathcal{H} \le 5/4$. Thus Corollary 4.18 gives

$$\mathcal{H}(\sigma^1) \le \frac{5}{4}\mathcal{H} \le \frac{25}{16}\,.$$

Now we can apply Lemma 4.19 twice to get

$$\mathcal{H}(\sigma^1,\pi) \le \frac{1}{5}\mathcal{H} \le \frac{1}{4}$$

and

$$\mathcal{H}(\pi,\sigma^1) \le \frac{1}{5}\mathcal{H}(\sigma^1) \le \frac{5}{16}\,.$$

It follows from (13) that for each state i,

$$x_i(\sigma^1,\pi) = z_i(\sigma^1,\pi) + \pi_i\mathcal{H}(\sigma^1,\pi) \le \pi_i\mathcal{H}(\pi,\sigma^1) + \pi_i\mathcal{H}(\sigma^1,\pi) \le \frac{9}{16}\pi_i \ \le \sigma^1.$$

By Lemma 5.12, this implies that $\mathcal{M}(\sigma^1,\pi) \le 1$ and hence $\mathcal{M}(\sigma,\pi) \le 2$ as claimed. □

Lemma 5.14 *Let $\Phi : \sigma \to \tau$ and $\Psi : \sigma \to \rho$ be two stopping rules such that $\Phi \le \Psi$ for any walk. Then*

$$H(\tau,\rho) \le \mathsf{E}\Psi - \mathsf{E}\Phi\,.$$

Proof. Define a stopping rule from τ to ρ as follows: the starting state from distribution τ may as well be generated by starting from σ and following Φ. But then we can just consider our walk as a continuation and stop it following the rule Ψ. This way we get a rule from τ to ρ with mean length $\mathsf{E}\Psi - \mathsf{E}\Phi$. $\square$

Lemma 5.15 *Assume that $t = \mathcal{M}(\sigma, \tau)$ is finite. Then*

$$\mathcal{H}(\sigma, \tau) + \mathcal{H}(\tau, \sigma^t) \le t\,.$$

Proof. Since $\mathcal{H}(\sigma, \tau)$ is the mean time of the threshold rule $\Theta_{\sigma,\tau}$, t is the mean time of the rule "walk t steps", and $\Theta_{\sigma,\tau} \le t$, Lemma 5.14 implies that $\mathcal{H}(\tau, \sigma^t) \le t - \mathcal{H}(\tau, \sigma^t)$. $\square$

Lemma 5.16 *Let $t = \mathcal{M}(\sigma, \pi)$. Choose a random starting state from σ. Choose a random integer Y uniformly from the interval $[t, 2t-1]$, and walk for Y steps. Then the probability of being at state i is at most $2\pi_i$.*

Proof. The probability that we stop at state i is

$$\frac{1}{t}\sum_{k=t}^{2t-1} \sigma_i^k = \pi_i + \frac{1}{t} z_i(\sigma^t, \sigma^{2t}) = \pi_i - \frac{1}{t} z_i(\pi, \sigma^t) - \frac{1}{t} z_i(\sigma^{2t}, \pi)\,.$$

Here we use that

$$z_i(\pi, \sigma^t) = x_i(\pi, \sigma^t) - \pi_i \mathcal{H}(\pi, \sigma^t) \ge -\pi_i \mathcal{H}(\pi, \sigma^t) \ge \pi_i(\mathcal{H}(\sigma, \pi) - t)$$

(by Lemma 5.15), and

$$z_i(\sigma^{2t}, \pi) = x_i(\sigma^{2t}, \pi) - \pi_i \mathcal{H}(\sigma^{2t}, \pi) \ge -\pi_i \mathcal{H}(\sigma^{2t}, \pi) \ge -\pi_i \mathcal{H}(\sigma, \pi)$$

(by Lemma 4.16). Hence we get that

$$\frac{1}{t}\sum_{k=t}^{2t-1} \sigma_i^k \le \pi_i - \frac{1}{t}\pi_i(\mathcal{H}(\sigma, \pi) - t) + \frac{1}{t}\pi_i \mathcal{H}(\sigma, \pi) = 2\pi_i\,.$$

This proves that $\mathcal{B}_1 \le \frac{3}{2}t \le \frac{3}{2}\mathcal{M}$. $\square$

5.5 The relaxation group: proofs

Lemma 5.17

$$\frac{1}{4}H^* \le \tilde{\mathcal{S}} \le 2H^*\,.$$

Proof. The first inequality is a special case of Corollary 4.8. To prove the second, let A be a set such that $\pi(A)\mathcal{H}(\pi, A) = \tilde{\mathcal{S}}$. Let $\varepsilon = \pi(A)/2(1 - \pi(A))$ and let $B = V \setminus A$. Then $\pi_B \le (1 + 2\varepsilon)\pi$. Let $\tau \le (1 + \varepsilon)\pi$ be a distribution such that

$$\mathcal{H}(\pi_B, \tau) \le \mathcal{H}(\le (1 + 2\varepsilon)\pi, \le (1 + \varepsilon)\pi)\,.$$

Let

$$\sigma_i = \begin{cases} \frac{1}{\tau(A)}((1 + 2\varepsilon)\pi_i - \tau_i), & \text{if } i \in B, \\ 0, & \text{otherwise.} \end{cases}$$

Then it is easy to check that σ is a distribution, and that $\sigma_i \geq (\varepsilon/\tau(A))\pi_i$ for $i \in B$. Hence by Corollary 4.17 and Inequality 8,

$$\begin{aligned}\mathcal{H}(\pi_B, \tau) &= \tau(A)\mathcal{H}(\sigma, \tau_A) \geq \tau(A)\mathcal{H}(\sigma, A) \\ &\geq \varepsilon\mathcal{H}(\pi, A) = \frac{1}{2(1-\pi(A))}\pi(A)\mathcal{H}(\pi, A) \geq \frac{1}{2}\tilde{\mathcal{S}}.\end{aligned}$$

This shows that $\mathcal{H}(\leq (1+2\varepsilon)\pi, \leq (1+\varepsilon)\pi) \geq (1/2)\tilde{\mathcal{S}}$. □

Next we prove inequalities for reversible chains. The next lemma was proved by Aldous and Fill [1]:

Lemma 5.18 *For every time-reversible Markov chain,*

$$\tilde{\mathcal{S}} \leq \mathcal{L}\,.$$

We have a similar inequality for $\tilde{\mathcal{Z}}$:

Lemma 5.19 *For every time-reversible Markov chain,*

$$\tilde{\mathcal{Z}} \leq 8\mathcal{L}.$$

Proof. By (21), we have

$$\tilde{\mathcal{Z}} \leq 2 \max_{\sigma,\tau \leq 2\pi} \sigma^T G\tau.$$

Here we have $\sigma^T G\tau = (R^{1/2}\sigma)^T(R^{-1/2}GR^{-1/2})(R^{1/2}\tau)$. From $\sigma \leq 2\pi$ we get that

$$|R^{1/2}\sigma| = \left(\sum_i \frac{\sigma_i^2}{\pi_i}\right)^{1/2} \leq \left(\sum_i 4\pi_i\right)^{1/2} = 2,$$

and hence

$$\sigma^T G\tau \leq 4\|R^{-1/2}GR^{-1/2}\|,$$

where $\|.\|$ denote spectral norm of a matrix. Now the matrix $R^{-1/2}GR^{-1/2}$ has the same eigenvalues as $R^{-1}G$, and hence its largest eigenvalue is $1/(1-\lambda_2) = \mathcal{L}$. □

Next, let us give a more complete statement of Theorem 3.6.

Theorem 5.20

$$2\varepsilon\tilde{\mathcal{I}}_{4\varepsilon} \leq 2\varepsilon\tilde{\mathcal{U}}_{2\varepsilon} \leq \tilde{\mathcal{H}}_\varepsilon \leq \tilde{\mathcal{B}}_\varepsilon \leq \tilde{\mathcal{U}}_\varepsilon \leq \frac{1}{\varepsilon}\tilde{\mathcal{Z}} \tag{24}$$

and

$$\tilde{\mathcal{H}}_\varepsilon \leq 2\tilde{\mathcal{I}}_{\varepsilon/3}. \tag{25}$$

If the chain is time-reversible, then

$$\tilde{\mathcal{U}}_\varepsilon \leq (\ln \varepsilon)\mathcal{L}\,. \tag{26}$$

The proof takes two lemmas.

Lemma 5.21 *Let $0 < \varepsilon < 1$ and $t \geq 2\tilde{\mathcal{U}}_\varepsilon$. Let v^0 be chosen from a starting distribution $\sigma \leq 2\pi$ and let Y be chosen uniformly from $\{0, \ldots, t-1\}$. Then v^0 and v^Y are (2ε)-independent.*

Note that Theorem 4.22 implies that the distribution of v^Y is closer than 2ε to π in total variation distance. Also, if $\sigma \le (1+\varepsilon)\pi$ then clearly $\sigma^Y \le (1+\varepsilon)\pi$. This lemma is proved (essentially) in [19].

Proof. We have to show that

$$\left|\mathsf{P}(v^0 \in A, v^Y \in B) - \mathsf{P}(v^0 \in A)\mathsf{P}(v^Y \in B)\right| < \varepsilon$$

for every two sets A and B of states. We have

$$\begin{aligned}&\left|\mathsf{P}(v^0 \in A, v^Y \in B) - \mathsf{P}(v^0 \in A)\mathsf{P}(v^Y \in B)\right| = \sigma(A)\left|\mathsf{P}(w^Y \in B \mid w^0 \in A) - \mathsf{P}(w^Y \in B)\right| \\ &= \sigma(A)\left|\sigma_A^Y(B) - \sigma^Y(B)\right| \le \sigma(A) d(\sigma_A^Y, \sigma^Y).\end{aligned}$$

Notice that $\sigma_A \le (2/\sigma(A))\pi$. By the choice of t, we have

$$d(\sigma^Y, \pi) \le \varepsilon.$$

Furthermore, consider the distribution

$$\alpha = \frac{\sigma(A)}{2-\sigma(A)}\sigma_A + \frac{2-2\sigma(A)}{2-\sigma(A)}\pi\,.$$

Then $\alpha \le 2\pi$, and hence by the choice of t, $d(\alpha^Y, \pi) \le \varepsilon$. But

$$\alpha^Y - \pi = \frac{\sigma(A)}{2-\sigma(A)}(\sigma_A^Y - \pi)\,,$$

and hence

$$d(\sigma_A^Y, \pi) \le \frac{2-\sigma(A)}{\sigma(A)} d(\alpha^Y, \pi) \le \frac{2-\sigma(A)}{\sigma(A)}\varepsilon\,.$$

Thus

$$d(\sigma_A^Y, \sigma^Y) \le \varepsilon + \frac{2-\sigma(A)}{\sigma(A)}\varepsilon = \frac{2}{\sigma(A)}\varepsilon\,,$$

and thus

$$\left|\mathsf{P}(v^0 \in A, v^Y \in B) - \mathsf{P}(v^0 \in A)\mathsf{P}(v^Y \in B)\right| \le \sigma(A) d(\sigma_A^Y, \sigma^Y) \le \sigma(A)\frac{2}{\sigma(A)}\varepsilon = 2\varepsilon.$$

□

Lemma 5.22 *Let Γ be a stopping rule such that $d(\pi^\Gamma, \pi) < \varepsilon$ and if v^0 is drawn according to π, then v^0 and v^Γ are ε-independent. Then $\tilde{\mathcal{H}}_{3\varepsilon} \le 2\mathsf{E}_\pi\Gamma$.*

Proof. Consider any starting distribution $\sigma < 2\pi$, and the result of stopping rule Γ starting from σ. Trivially

$$\mathsf{E}_\sigma\Gamma \le 2\mathsf{E}_\pi\Gamma.$$

Thus it suffices to prove that $d(\sigma^\Gamma, \pi) < 3\varepsilon$. Indeed, we have

$$d(\sigma^\Gamma, \pi) \le d(\sigma^\Gamma, \pi^\Gamma) + d(\pi^\Gamma, \pi) \le d(\sigma^\Gamma, \pi^\Gamma) + \varepsilon.$$

To estimate the first term, we write

$$\sigma = \sum_{k=1}^{m} a_k \pi_{A_k},$$

where $a_k > 0$ and $\emptyset \subset A_1 \subset \ldots \subset A_m \subseteq V$. It is easy to see that such a decomposition exists. Looking at an element $i \in A_1$, we see that $\sum_k \frac{a_k}{\pi(A_k)} \leq 2$. Further,

$$\sigma^\Gamma = \sum_k a_k \pi^\Gamma_{A_k}.$$

Hence for an appropriate set $B \subseteq V$,

$$d(\sigma^\Gamma, \pi^\Gamma) = \sum_{i \in B} (\sigma^\Gamma_i - \pi^\Gamma_i) = \sum_k a_k \sum_{i \in B} (\pi^\Gamma_{A_k})_i.$$

Here we have

$$\sum_{i \in B} (\pi^\Gamma_{A_k})_i - \pi^\Gamma_i = \mathsf{P}(v^\Gamma \in B \mid v^0 \in A_k) - \mathsf{P}(v^\Gamma \in B)$$

$$= \frac{\mathsf{P}(v^0 \in A_k,\ v^\Gamma \in B) - \mathsf{P}(v^0 \in A_k)\mathsf{P}(v^\Gamma \in B)}{\mathsf{P}(v^0 \in A_k)} \leq \frac{\varepsilon}{\pi(A_k)}.$$

Thus

$$d(\sigma^\Gamma, \pi^\Gamma) \leq \varepsilon \left(\sum_{k=1}^{m} \frac{a_k}{\pi(A_k)} \right) \leq 2\varepsilon.$$

This proves the lemma. □

Now the first inequality in (24) follows from Lemma 5.21, the second, from Theorem 4.22; the next two are trivial, while the last follows easily from Lemma 4.21. Lemma 5.22 easily implies (25).

This completes the proof of Theorem 5.20. □

5.6 Relations between classes: proofs

Proof of Theorem 3.7. By Lemma 4.33, we have for every $k \geq 1$,

$$\mathcal{B}(\leq (1+2k)\pi, \leq (1+k)\pi) = \mathcal{B}([1-(1/k)..2]\pi, [1-(1/k)..3/2]\pi).$$

Furthermore, by Lemma 4.31,

$$\mathcal{B}([1-(1/k)..2]\pi, [1-(1/k)..3/2]\pi) \leq \mathcal{B}(\leq 2\pi, \leq (3/2)\pi) = \tilde{\mathcal{B}}_{\overline{1/2}}.$$

Let $k = \lceil \ln(1/\hat{\pi} - 1) \rceil$, then by the triangle inequality for $\mathcal{B}(\sigma, \tau)$, we have

$$\mathcal{B}(D, D_1) = \mathcal{B}(\leq (1+2^k)\pi, \leq 2\pi) = \sum_{j=0}^{k-1} \mathcal{B}(\leq (1+2^{j+1})\pi, \leq (1+2^j)\pi) \leq k\tilde{\mathcal{B}}_{\overline{1/2}}.$$

This proves the corollary. □

Proof of Theorem 3.8. Parts (a) and (c) were proved in [24]. Part (b) is strightforward linear algebra. For part (d), note that Corollary 4.12 implies that $\mathcal{H}(i,j) - \mathcal{H}(\pi,j) = \overleftarrow{\mathcal{H}}(j,i) - \overleftarrow{\mathcal{H}}(\pi,i)$, and hence the identity follows from formulas (9) and (10).

References

[1] D.J. Aldous and J. Fill, *Reversible Markov Chains and Random Walks on Graphs* (book), to appear. URL for draft at http://www.stat.Berkeley.EDU/users/aldous/book.html.

[2] D.J. Aldous, Some inequalities for reversible Markov chains, *J. London Math. Soc.* **25** (1982), 564–576.

[3] D.J. Aldous, Applications of random walks on graphs, preprint (1989).

[4] D.J. Aldous, The random walk construction for spanning trees and uniform labelled trees, *SIAM J. Discrete Math.* **3** (1990), 450–465.

[5] D.J. Aldous and P. Diaconis, Shuffling cards and stopping times, *Amer. Math. Monthly* **93** #5 (1986), 333–348.

[6] D.J. Aldous, L. Lovász and P. Winkler, Mixing times for uniformly ergodic Markov chains, *Stochastic Processes and their Applications*, to appear.

[7] S. Asmussen, P.W. Glynn and H. Thorisson, Stationary detection in the initial transient problem, *ACM Transactions on Modeling and Computer Simulation* **2** (1992), 130–157.

[8] J.R. Baxter and R.V. Chacon, Stopping times for recurrent Markov processes, *Illinois J. Math.* **20** (1976), 467–475.

[9] D. Bayer and P. Diaconis, Trailing the dovetail shuffle to its lair, *Ann. Appl. Probab.* **2** (1992), 294–313.

[10] A. Beveridge and L. Lovász, Random walks and the regeneration time, *J. Graph Theory* (to appear)

[11] A. Broder, Generating random spanning trees, *Proc. 30th Annual Symp. on Found. of Computer Science,* IEEE Computer Soc. (1989), 442–447.

[12] R.V. Chacon and D.S. Ornstein, A general ergodic theorem, *Illinois J. Math.* **4** (1960), 153–160.

[13] A.K. Chandra, P. Raghavan, W.L. Ruzzo, R. Smolensky, and P. Tiwari, The Electrical Resistance of a Graph Captures its Commute and Cover Times, *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, May (1989).

[14] D. Coppersmith, P. Tetali, and P. Winkler, Collisions among Random Walks on a Graph, *SIAM J. on Discrete Mathematics* **6** #3 (1993), 363–374.

[15] P.G. Doyle and J.L. Snell, *Random Walks and Electric Networks*, Mathematical Assoc. of America, Washington, DC 1984.

[16] L.E. Dubins, On a theorem of Skorokhod, *Ann. Math. Statist.* **39** (1968), 2094–2097.

[17] M. Dyer, A. Frieze and R. Kannan, A random polynomial time algorithm for estimating volumes of convex bodies, *Proc. 21st Annual ACM Symposium on the Theory of Computing* (1989), 375–381.

[18] M. Jerrum and A. Sinclair, Conductance and the rapid mixing property for Markov chains: the approximation of the permanent resolved, *Proc. 20nd Annual ACM Symposium on Theory of Computing* (1988), 235–243.

[19] R. Kannan, L. Lovász and M. Simonovits: Random walks and an $O^*(n^5)$ volume algorithm, *Random Structures and Algorithms* (1997).

[20] L. Lovász and M. Simonovits, Random walks in a convex body and an improved volume algorithm, *Random Structures and Alg.* **4** (1993), 359–412.

[21] L. Lovász and P. Winkler, A note on the last new vertex visited by a random walk, *J. Graph Theory* **17** (1993), 593–596.

[22] L. Lovász and P. Winkler, Exact mixing in an unknown Markov chain, *Electronic J. Comb.* **2**, (1995), Paper R15.

[23] L. Lovász and P. Winkler, Mixing of random walks and other diffusions on a graph, *Surveys in Combinatorics, 1995*, P. Rowlinson, ed., London Math. Soc. Lecture Note Series 218, Cambridge U. Press (1995), 119–154.

[24] L. Lovász and P. Winkler, Reversal of Markov chains and the forget time, *Combinatorics, Probability and Computing*, to appear.

[25] J.W. Pitman, Occupation measures for Markov chains, *Adv. Appl. Prob.* **9** (1977), 69–86.

[26] D.H. Root, The existence of certain stopping times on Brownian motion, *Ann. Math. Statist.* **40** (1969), 715–718.

[27] A. Skorokhod, *Studies in the Theory of Random Processes*, orig. pub. Addison-Wesley (1965), 2nd ed. Dover, New York 1982.

[28] P. Tetali, Random walks and effective resistance of networks, *J. Theoretical Prob.* #1 (1991), 101–109.